Rough Table of Contents for notes that accompany the lectures

<table>
<thead>
<tr>
<th>Topic</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple linear regression (SLR)</td>
<td>1-32</td>
</tr>
<tr>
<td>Multiple linear regression</td>
<td>33</td>
</tr>
<tr>
<td>Matrix formulation for MLR</td>
<td>36-40</td>
</tr>
<tr>
<td>Variable selection in MLR</td>
<td>41-52</td>
</tr>
<tr>
<td>Partial and seq. sums of squares</td>
<td>47-50</td>
</tr>
<tr>
<td>ANOVA as regression</td>
<td>69</td>
</tr>
<tr>
<td>General linear model</td>
<td>72</td>
</tr>
<tr>
<td>ANCOVA</td>
<td>82-88</td>
</tr>
<tr>
<td>Lack-of-fit</td>
<td>89-91</td>
</tr>
<tr>
<td>One-way ANOVA, contrasts</td>
<td>92-101</td>
</tr>
<tr>
<td>Multiple comparisons</td>
<td>102-108</td>
</tr>
<tr>
<td>Expected mean squares, part I</td>
<td>109-110</td>
</tr>
<tr>
<td>Sample size computation</td>
<td>111-115</td>
</tr>
<tr>
<td>Orthogonal polynomial contrasts</td>
<td>116-120</td>
</tr>
<tr>
<td>Multi-factorial expts</td>
<td>121</td>
</tr>
<tr>
<td>2 × 2 expts</td>
<td>121-132</td>
</tr>
<tr>
<td>More than 2 levels</td>
<td>133</td>
</tr>
<tr>
<td>More than 2 factors</td>
<td>140</td>
</tr>
<tr>
<td>Unbalanced example</td>
<td>148-154</td>
</tr>
<tr>
<td>Blocking</td>
<td>155-164</td>
</tr>
<tr>
<td>Latin squares</td>
<td>165-174</td>
</tr>
<tr>
<td>One-way random effects</td>
<td>175-188</td>
</tr>
<tr>
<td>Mixed models</td>
<td>189-235</td>
</tr>
<tr>
<td>Split-plot expts</td>
<td>236-255</td>
</tr>
</tbody>
</table>
ST 512: Exptl Stats for Biol. Sciences II

Topic 1 Simple linear regression

Reading: Rao, Ch. 10

An example: The association between corn yield and rainfall

Yields $y$ (in bushels/acre) on corn raised in six midwestern states from 1890 to 1927 recorded with rainfall $x$ (inches/yr).

$$y_1, \ldots, y_{38} \quad \text{and} \quad x_1, \ldots, x_{38}.$$
A scatterplot provides some indication of an association between $y$ and $x$. In particular, yields seem to increase with rainfall.

Some questions:

- How can we describe the association between yield and rainfall? Does it appear linear?
- How can we measure the strength of the linear association?
- To what degree is the variability in yield described or explained by its association with rainfall?
- How can we use this association to estimate average yield, given a certain level of rainfall?
- How can we use this association to predict future yield, if we have an idea about what the rainfall will be?
Correlation

Definition: The sample correlation coefficient $r_{xy}$ of the paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

is defined by

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/(n - 1)}{\sqrt{\sum (x_i - \bar{x})^2/(n - 1) \times \sum (y_i - \bar{y})^2/(n - 1)}} = \frac{s_{xy}}{s_x s_y}$$

$s_{xy}$ is called the sample covariance of $x$ and $y$:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}.$$

Some properties of $r_{xy}$

- $r_{xy}$ is a measure of the linear assn. between $x$ and $y$ in a dataset.
- correlation coefficients always between -1 and 1:

$$-1 \leq r_{xy} \leq 1$$

- The closer $r_{xy}$ is to 1, the stronger the positive linear association between $x$ and $y$
- The closer $r_{xy}$ is to -1, the stronger the negative linear association ($y$ tends to be smaller than avg. when $x$ bigger than avg.)
- The bigger $|r_{xy}|$, the stronger the linear association
- If $|r_{xy}| = 1$, then $x$ and $y$ are said to be perfectly correlated.
Summary statistics for corn yields data:

\[ \bar{x} = 10.8, \quad s_x^2 = 5.13 \quad s_x = 2.27 \]

\[ \bar{y} = 31.9, \quad s_y^2 = 19.0 \quad s_y = 4.44 \]

\[ s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{147.3}{38 - 1} = 3.98 \]

Applying the formula for \( r_{xy} \), we get

\[ r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{3.98}{\sqrt{5.13 \times 19.0}} = \frac{3.98}{9.87} = 0.40 \]

The population correlation coefficient \( \rho \)

Just as \( \bar{x} \) provides empirical information about a population mean \( \mu_x \), \( r_{xy} \) can be used for inference about the population correlation coefficient \( \rho_{xy} \). This parameter refers to the correlation among \( x \) and \( y \) in the population of interest:

\[ \rho_{XY} = E \left[ \frac{(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y} \right] = \rho. \]

Here, \( E(\cdot) \) denotes mathematical expectation. This concept underlies all of the statistical modelling to be discussed in this course.
A test statistic useful for inference about $\rho$ is

$$Z(\rho) = \left(\frac{1}{2}\sqrt{n-3}\right) \left(\log \frac{1+R}{1-R} - \log \frac{1+\rho}{1-\rho}\right).$$

Asymptotically, $Z$ has the standard normal ($N(0,1)$) distribution so that it can be used to derive methods of inference for $\rho$ (testing and confidence intervals).

Under $H_0 : \rho = 0$,

$$\left(\frac{1}{2}\sqrt{n-3}\right) \log \frac{1+R}{1-R} \sim N(0,1).$$

A large-sample test of $H_0$ with level $\alpha$ then rejects $H_0$ whenever

$$\frac{1}{2}\sqrt{n-3}\log \left(\frac{1+R}{1-R}\right) > z_{\alpha/2}$$

or

$$\frac{1}{2}\sqrt{n-3}\log \left(\frac{1+R}{1-R}\right) < z_{1-\alpha/2}$$

where $z_\alpha$ satisfies $\alpha = \Pr(Z > z_\alpha)$ with $Z \sim N(0,1)$.

An approximate 100(1-$\alpha$)% confidence interval for $\rho$ can be obtained by inverting the Fisher transformation:

$$\psi = \frac{1}{2} \log \left(\frac{1+\rho}{1-\rho}\right).$$

The probability statement

$$1-\alpha = \Pr(z_{1-\alpha/2} < \sqrt{n-3}\frac{1}{2} \left(\log \left(\frac{1+R}{1-R}\right) - \log \left(\frac{1+\rho}{1-\rho}\right)\right) < z_{\alpha/2})$$

can be rearranged to yield an approximate 100(1-$\alpha$)% confidence interval for $\psi$:

$$\frac{1}{2} \log \left(\frac{1+R}{1-R}\right) \pm \frac{z_{\alpha/2}}{\sqrt{n-3}}.$$
Note that $\rho$ and $\psi$ are related by

$$\rho = \frac{e^{2\psi} - 1}{e^{2\psi} + 1}$$

Evaluating $\rho$ at the limits for $\psi$ leads to the interval

$$\left( \frac{1+R}{1-R} e^{-2z/\sqrt{n-3}} - 1, \frac{1+R}{1-R} e^{2z/\sqrt{n-3}} - 1 \right)$$

$$\left( \frac{1+R}{1-R} e^{-2z/\sqrt{n-3}} + 1, \frac{1+R}{1-R} e^{2z/\sqrt{n-3}} + 1 \right).$$

For the corn yields data, $r = 0.4$ and $n = 38$, and a 95% interval is given by

$$(0.09 < \rho < 0.64).$$

(There is a one-to-one correspondence between testing and interval estimation here, so that $H_0 : \rho = 0$ would be rejected at $\alpha = 0.05$.)

Exercise:

1. Examine the butterfat and temperature data plotted on the next page. Is there evidence of linear association? The sample correlation coefficient is $r = -0.45$ based on randomly sampled days. Carry out an appropriate test. Obtain a 95% confidence interval for the population correlation coefficient describing the linear association between % butterfat and temperature.

2. Suppose that two variables $X$ and $Y$ have correlation $\rho = 0.6$. What is the probability that a random sample of $n = 30$ observations from this bivariate population will yield a sample correlation coefficient of 0.7 or higher?

Hint for # 1.:

```
proc corr data=butterfat FISHER(BIASADJ=NO);
  var butterfat temp;
run;
```
Some example scatterplots ($r_1 = 0.04$ and $r_2 = -0.45$)
An exercise/activity:
Label the four plots below with the four sample correlation coefficients:

1. \( r = 0.3 \)
2. \( r = 0.7 \)
3. \( r = 0.1 \)
4. \( r = -0.6 \)
Correlation does not imply causation

Famous examples of *spurious correlations*:

- A study finds a high positive correlation between coffee drinking and coronary heart disease. Newspaper reports say the fragrant essence of the roasted beans of *Coffea arabica* are a menace to public health.

- In a city, if you were to observe the amount of damage and the number of fire engines for enough recent fires, you would likely see a positive and significant correlation among these variables. Obviously, it would be erroneous to conclude that fire engines cause damage.

- *Lurking variable* - a third variable that is responsible for a correlation between two others. (A.k.a. confounding factor.) An example would be to assess the association between say the reading skills of children and other measurements taken on them, such as shoe size. There may be a statistically significant association between shoe size and reading skills, but that doesn’t imply that one causes the other. Rather, both are positively associated with a third variable, *age*.

- Among 50 countries examined in a dietary study, high positive correlation among fat intake and cancer (see figure, next page). This example is taken from from *Statistics* by Freedman, Pisani and Purves.
In countries where people eat lots of fat like the United States rates of breast cancer and colon cancer are high. This correlation is often used to argue that fat in the diet causes cancer. How good is the evidence?

Discussion. If fat in the diet causes cancer, then the points in the diagram should slope up, other things being equal. So the diagram is some evidence for the theory. But the evidence is quite weak, because other things aren’t equal. For example, the countries with lots of fat in the diet also have lots of sugar. A plot of colon cancer rates against sugar consumption would look just like figure 8, and nobody thinks that sugar causes colon cancer. As it turns out, fat and sugar are relatively expensive. In rich countries, people can afford to eat fat and sugar rather than starchier grain products. Some aspects of the diet in these countries, or other factors in the life-style, probably do cause certain kinds of cancer and protect against other kinds. So far, epidemiologists can identify only a few of these factors with any real confidence. Fat is not among them.

(p. 152, *Statistics* by Friedman, Pisani, Purves and Adhikari)
Figure 8. Cancer rates plotted against fat in the diet for a sample of countries

A linear model for regression

Observe \( n \) independent pairs \((y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\)

A probabilistic model for \( Y \) conditional on \( X = x \):

\[
Y_i = \beta_0 + \beta_1 x_i + E_i
\]

where \( E_1, \ldots, E_n \) are independent and identically distributed normal random variables with mean 0 and variance \( \sigma^2 \).

(Write \( E_i \sim \text{iid} \sim N(0, \sigma^2) \).)

Note that this implies

1. \( \mu(x) = E(Y|X = x) = \beta_0 + \beta_1 x \)
2. \( \text{Var}(Y|X = x) = \sigma^2 \)

Definitions:

- response or dependent variable \( Y \) (left side of regression equation)
- independent variable or predictor variable \( X \) (right side)
- intercept term \( \beta_0 = E(Y|X = 0) \) (where \( \mu(x) \) crosses \( y \)-axis.)
- slope term \( \beta_1 \), average change in \( E(Y) \) per unit increase in \( x \)
- error variance \( \sigma^2 \)

\( \beta_0, \beta_1 \) and \( \sigma^2 \) are modelled as fixed, unknown parameters which can be estimated from the data using simple linear regression.

Nonlinear regression: other models for \( E(Y|X = x) \) such as

\[
\mu(x) = \beta_0 x^{\beta_1}.
\]
Fitting a linear model

- Choose “best” values for $\beta_0, \beta_1$.

Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that

\[
SS[E] = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2
\]

is minimized. These are “least squares” (LS) estimates.

Definitions:

- Predicted value of response $Y_i$ given $X = x_i$:
  
  \[
  \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.
  \]

- Residual for the $i^{th}$ observation:
  
  \[
  e_i = y_i - \hat{y}_i
  \]

Elementary calculus can show that $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize the sum of squared residuals $SS[E]$ are given by

\[
\hat{\beta}_1 = \frac{\sum (x_j - \bar{x})(y_j - \bar{y})}{\sum (x_j - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad (\text{OL notation, p. 542, Rao notation, p. 396})
\]

\[
= \frac{s_{xy}}{s_x^2} \quad (\text{sample covariance } \div \text{ sample variance } )
\]

\[
= r_{xy} \frac{s_y}{s_x}
\]

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]

An unbiased estimate of $\sigma^2$ is given by

\[
\hat{\sigma}^2 = MS[E] = \frac{SS[E]}{n-2}.
\]
Definition: The line satisfying the equation
\[ y = \hat{\beta}_0 + \hat{\beta}_1 x \]
is called the linear regression of \( y \) on \( x \).
(It is also called the least squares line.)

For the corn yield data, recall that
\[
\bar{x} = 10.8 \text{ inches} \quad \bar{y} = 31.9 \text{ bushels per acre} \\
\text{s}_x^2 = 5.1, \text{s}_y^2 = 19.0, \text{s}_{xy} = 3.98, r_{xy} = 0.40
\]
so that
\[
\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{3.98}{5.13} = r_{xy} \frac{s_y}{s_x}
\]
\[
= (0.40) \sqrt{\frac{19.0}{5.1}}
\]
\[
= 0.776(\text{ bushels per acre} \div \text{ inches per year})
\]
\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]
\[
= 31.9 - 0.776(10.8)
\]
\[
= 23.5 \text{ bushels per acre}
\]
yielding the least squares line of
\[
\hat{y} = 23.5 + (0.776)x.
\]

Note that
1. \( \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0 \)
2. \( \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \) is minimized
The ANOVA table from simple linear regression

Observed variability in the response $Y$ is measured by the total sum of squares $SS[TOT]$ and can be partitioned into independent components: the sum of squares due to regression, $SS[R]$ and the sum of squares due to error, $SS[E]$.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>$SS[R]$</td>
<td>1</td>
<td>$MS[R]$</td>
<td>$MS[R]/MS[E]$</td>
</tr>
<tr>
<td>Error</td>
<td>$SS[E]$</td>
<td>$n-2$</td>
<td>$MS[E]$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$SS[TOT]$</td>
<td>$n-1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The sums of squares are defined by

\[ SS[TOT] = \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

\[ SS[R] = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \]

\[ = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}))^2 \]

\[ = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

\[ = \hat{\beta}_1^2 S_{xx} \]

\[ SS[E] = \sum e_i^2 \]

\[ = \sum (y_i - \hat{y}_i)^2 \]

\[ = SS[TOT] - SS[R] \]
- The $F$ ratio can be used to test for “significance of regression” or to test the null hypothesis that the slope parameter, $\beta_1$ is zero:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

at level $\alpha$. The critical value for $F$ is the upper $\alpha$ percentile from the $F$ distribution with 1 numerator and $n - 2$ denominator degrees of freedom. This $F$ test is equivalent to a $T$ test based on the statistic we’re about to discuss:

$$T = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \quad \text{and} \quad T^2 = F = \frac{MS[R]}{MS[E]}$$

- The mean square for error, $MS[E]$, is an unbiased estimator for $\sigma^2$, the common variance of the response variable $Y$ conditional on an observed independent variable $X$. ($E(MS[E]) = \sigma^2 = \text{Var}(Y|X)$). (Here, conditional means for those elements in the population with independent variable $X = x$.) As such, it can be used to construct confidence intervals for $\beta_0$ and $\beta_1$. It is based on $n - 2$ degrees of freedom.

- The ratio of $SS[R]$ to $SS[TOT]$ is called the coefficient of determination, or sometimes, simply “r-square”. It represents the proportion of variation observed in the response variable $y$ which can be “explained” by its linear association with $x$. In simple linear regression, “r-square” is in fact equal to $r_{xy}^2$. (But this isn’t the case in multiple regression.) It is also equal to the squared correlation between $y_i$ and $\hat{y}_i$. (This is the case in multiple regression.)
Confidence intervals for $\beta_0, \beta_1$

Important results for sampling distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ (given $x_1, \ldots, x_n$)

- unbiasedness:

$$E(\hat{\beta}_1|x_1, \ldots, x_n) = \beta_1 \text{ and } E(\hat{\beta}_0|x_1, \ldots, x_n) = \beta_0$$

- for normal data, $\bar{Y} \perp \hat{\beta}_1$ which leads to

$$\begin{align*}
\text{Var}(\hat{\beta}_1|x_1, \ldots, x_n) &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\
\text{Var}(\hat{\beta}_0|x_1, \ldots, x_n) &= \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)
\end{align*}$$

Take $\sqrt{}$ and substitute $MS[E]$ for $\sigma^2$ to get estimated standard errors:

$$\begin{align*}
\hat{SE}(\hat{\beta}_1) &= \sqrt{\frac{MS[E]}{S_{xx}}} \\
\hat{SE}(\hat{\beta}_0) &= \sqrt{MS[E]} \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)
\end{align*}$$

100(1 - $\alpha$)% confidence intervals for $\beta_1$ and $\beta_0$ are given by

$$\begin{align*}
\hat{\beta}_1 &\pm t(n-2, \alpha/2) \sqrt{\frac{MS[E]}{S_{xx}}} \\
\hat{\beta}_0 &\pm t(n-2, \alpha/2) \sqrt{MS[E]} \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)
\end{align*}$$
Any hypothetical slope, like $H_0 : \beta_1 = \text{slope}_0$ may be tested using the $T$-statistic below with $df = n - 2$:

$$T = \frac{\hat{\beta}_1 - \text{slope}_0}{\widehat{SE}(\hat{\beta}_1)}$$

Confidence interval for $E(Y|X = x_0)$

The conditional mean $E(Y|X = x_0)$ can be estimated by evaluating the regression function $\mu(x_0)$ at the estimates $\hat{\beta}_0, \hat{\beta}_1$. The conditional variance of the expression isn’t too difficult:

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0|X = x_0) = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

This yields a confidence interval of the form

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t(n - 2, \alpha/2) \sqrt{MS[E] \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

Exercise: derive these variances.
The yield on corn by rainfall example

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>114</td>
<td>1</td>
<td>114</td>
<td>6.95</td>
</tr>
<tr>
<td>Error</td>
<td>591</td>
<td>36</td>
<td>16.4</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>705</td>
<td>37</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The $\alpha = 0.05$ critical value is $F(1, 36, 0.05) = 4.11$ Therefore, there is a significant, positive linear association between yield and rainfall.

For 95% CIs, use $t(26, 0.025) = 2.028$. For $\beta_1$, note that

$$S_{xx} = (n-1)s_x^2 = 5.13(38-1) = 189.8$$

so that a c.i. is given by

$$0.776 \pm (2.028)\sqrt{\frac{16.4}{189.8}}$$

or

$$0.776 \pm (2.028)(0.294) \text{ or } 0.776 \pm 0.596$$

For $\hat{\beta}_0$,

$$23.5 \pm (2.028)\sqrt{16.4 \left(\frac{1}{38} + \frac{(10.8)^2}{189.8}\right)}$$

or

$$23.5 \pm (2.028)(3.24) \text{ or } 23.5 \pm 6.57$$
Prediction

Often, prediction of the response variable $Y$ for a given value, say $x_0$, of the independent variable is of interest. In order to make statements about future values of $Y$, we need to take into account

- the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$
- the randomness of a future value $Y$.

We’ve seen that the predicted value of $Y$ based on the linear regression is given by $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

In order to form a 95% prediction interval take

$$\hat{Y}_0 \pm t(n - 2, \alpha/2) \sqrt{MS[E]\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$ 

Example: Suppose that one year rainfall is $x_0 = 14$ inches, but that yield $Y_0$ from the six states hasn’t been measured. Obtain a 95% prediction interval for $Y_0$, using the model.

$$23.5 + 0.776(14) \pm 2.028 \sqrt{16.4 \left(1 + \frac{1}{38} + \frac{(14 - 10.8)^2}{189.8}\right)}$$

or

$$34.4 \pm 8.5 \quad \text{or} \quad (25.9, 42.9)$$
The 95% prediction interval is (25.9, 42.9).

A 95% confidence interval for $E(Y|X = 14)$ is given by

$$23.5 + 0.776(14) \pm 2.028 \sqrt{16.4 \left( \frac{1}{38} + \frac{(14 - 10.8)^2}{189.8} \right)}$$

or

$$34.4 \pm 2.32 \quad \text{or} \quad (32.1, 36.7)$$

What is the difference?

Exercise taken from Dickey’s notes:
An industrial quality control expert takes 200 hourly measurements on an industrial furnace which is under control and finds that a 95% confidence interval for the mean temperature is (500.35, 531.36). As a result he tells management that the process should be declared out of control whenever hourly measurements fall outside this interval and, of course, is later fired for incompetence. (Why and what should he have done?)
A note of caution:
Mark Twain, in *Life on the Mississippi*:

In the space of 176 years the Lower Mississippi has shortened itself 252 miles. That is an average of a trifle more than one mile and a third per year. Therefore any calm person, who is not blind or idiotic, can see that in 742 years from now, the lower Mississippi will be one mile and three quarters long, and Cairo, Ill., and New Orleans will have joined their streets together.

It is not safe to extrapolate the results of a linear regression for the purposes of prediction beyond the range of observed independent variables.
1. Butterfat by temperature in cows:
   - $y_j$: average “percent butterfat” for 10 cows on date $j$
   - $x_j$: temperature on date $j$
   - $n = 20$ successive days

<table>
<thead>
<tr>
<th>Date</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_j$</td>
<td>64</td>
<td>65</td>
<td>65</td>
<td>64</td>
<td>61</td>
<td>55</td>
<td>39</td>
<td>41</td>
<td>46</td>
<td>59</td>
</tr>
<tr>
<td>$y_j$</td>
<td>4.65</td>
<td>4.58</td>
<td>4.67</td>
<td>4.60</td>
<td>4.83</td>
<td>4.55</td>
<td>5.14</td>
<td>4.71</td>
<td>4.69</td>
<td>4.65</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_j$</td>
<td>56</td>
<td>56</td>
<td>62</td>
<td>37</td>
<td>37</td>
<td>45</td>
<td>57</td>
<td>58</td>
<td>60</td>
<td>55</td>
</tr>
<tr>
<td>$y_j$</td>
<td>4.36</td>
<td>4.82</td>
<td>4.65</td>
<td>4.66</td>
<td>4.95</td>
<td>4.60</td>
<td>4.68</td>
<td>4.65</td>
<td>4.6</td>
<td>4.46</td>
</tr>
</tbody>
</table>

2. Hybrid duck data:
   - Mallard and Pintail ducks were crossed yielding $n = 11$ second generation males with attributes as given in the table
   - $y_j$: Behavioral index
   - $x_j$: Plumage index
   - A 0 corresponds to a purely mallard phenotype and a 15 corresponds to a purely pintail phenotype.
   - The same scoring is used to quantify duck behavioral traits.

   | $x_j$ | 7   | 13  | 14  | 6   | 14  | 15  | 4   | 8   | 7   | 9   | 14  |
   | $y_j$ | 3   | 10  | 11  | 5   | 15  | 15  | 7   | 10  | 4   | 9   | 11  |

3. Cricket Data
   - $y_j$: Chirps per second
   - $x_j$: Temperature ($^\circ F$)
   - Striped ground cricket.

   | $y_j$ | 20  | 16  | 19.8| 18.4| 17.1| 15.5| 14.7| 17.1| 15.4| 16.2|
   | $x_j$ | 88.6| 71.6| 93.3| 84.3| 80.6| 75.2| 69.7| 82  | 69.4| 83.3|

   | $y_j$ | 15  | 17.2| 16  | 17  | 14.4|
   | $x_j$ | 79.6| 82.6| 80.6| 83.5| 76.3|
A classical dataset:
The association between height of adults and their parents
Consider a statistical model for these data, randomly sampled from some population of interest. In particular, choose a model which accounts for the apparent linear dependence of the mean height of sons on midparent height \( X \). Let \( Y_1, \ldots, Y_n \) denote the sons’ heights. Given \( X = x_i \),

\[
Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{for } i = 1, \ldots, n (n = 928).
\]

where \( E_1, \ldots, E_n \) are

- independent,
- identically and
- normally distributed random variables with mean 0 and error variance \( \sigma^2 \).

(Write \( E_i \overset{iid}{\sim} N(0, \sigma^2) \).)

This implies

1. \( \mu(x) = E(Y|X = x) = \beta_0 + \beta_1 x \)
2. \( \text{Var}(Y|X = x) = \sigma^2 \) (Three unknown parameters \( \beta_0, \beta_1, \sigma^2 \) quantify the whole population of interest.)

Question: Suppose we ignore midparent height \( x \). Consider estimating the mean \( E(Y) \). Propose a model. Propose a method for obtaining a confidence interval for the mean height of the sons in the population from which these data were randomly sampled. Use summary statistics on page 6 to complete this naive analysis.
Many questions to answer using regression analysis:

1. What is the meaning, in words, of $\beta_1$?

2. True/false: (a) $\beta_1$ is a statistic (b) $\beta_1$ is a parameter (c) $\beta_1$ is unknown.

3. What is the observed value of $\hat{\beta}_1$?

4. True/false: (a) $\hat{\beta}_1$ is a statistic (b) $\hat{\beta}_1$ is a parameter (c) $\hat{\beta}_1$ is unknown.

5. Is $\hat{\beta}_1 = \beta_1$?

6. How much does $\hat{\beta}_1$ vary about $\beta_1$ from sample to sample? (Provide an estimate of the standard error, as well as an expression indicating how it was computed.)

7. What is a region of plausible values for $\beta_1$ suggested by the data?

8. What is the line that best fits these data, using the criterion that smallest sum of squared residuals is “best?”

9. How much of the observed variation in the heights of sons (the $y$-axis) is explained by this “best” line?

10. What is the estimated average height of sons whose midparent height is $x = 68$?

11. Is this the true average height in the whole population of sons whose midparent height is $x = 68$?

12. Under the model, what is the true average height of sons with midparent height $x = 68$?
13. What is the estimated standard deviation among the population of sons whose parents have midparent height $x = 68$? Would you call this standard deviation a “standard error?”

14. What is the estimated standard deviation among the population of sons whose parents have midparent height $x = 72$? Bigger, smaller, or the same as that for $x = 68$? Is your answer obviously supported or refuted by inspection of the scatterplot?

15. What is the estimated standard error of the estimated average for sons with midparent height $x = 68$, $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1$? Provide an expression for this standard error.

16. Is the estimated standard error of $\hat{\mu}(72)$ bigger, smaller, or the same as that for $\hat{\mu}(68)$?

17. Is the observed linear association between son’s height and midparent height strong? Report a test statistic.

18. What quantity can you use to describe or characterize the linear association between height and midparent height in the whole population? Is this a parameter or a statistic?

19. Let $Y$ denote the height of a male randomly sampled from this population and $X$ his midparent height. Is it true that the population correlation coefficient $\rho$ satisfies

$$\rho = E\left[\left(\frac{Y - \mu_Y}{\sigma_Y}\right) \times \left(\frac{X - \mu_X}{\sigma_X}\right)\right]?$$

20. Define $\mu_Y, \sigma_Y, \mu_X, \sigma_X, \rho$. Parameters or statistics?

21. What are plausible values for $\rho$ suggested by the data?

22. Is $E_1, \ldots, E_{928} \overset{iid}{\sim} N(0, \sigma^2)$ a reasonable assumption?
options ls=75 nodate;

data Galton;
    array cdata(14);
    if _n_ = 1 then input cdata1-cdata14 @;
    retain cdata1-cdata14; drop cdata1-cdata14 i;
    input parent @;
    do i = 1 to 14; input count @; son=cdata(i);
    output;
    end;
cards;
    61.7 62.2 63.2 64.2 65.2 66.2 67.2 68.2 69.2 70.2 71.2 72.2 73.2 73.7
    73.0 0 0 0 0 0 0 0 0 0 0 0 0 0
    72.5 0 0 0 0 0 0 0 1 2 1 2 7 2 4
    71.5 0 0 0 0 1 3 4 3 5 10 4 9 2 2
    70.5 1 0 1 1 1 1 3 12 18 14 7 4 3 3
    69.5 0 0 1 16 4 17 27 20 33 25 20 11 4 5
    68.5 1 0 7 11 16 25 31 34 48 21 18 4 3 0
    67.5 0 3 5 14 15 36 38 28 38 19 11 4 0 0
    66.5 0 3 3 5 2 17 17 14 13 4 0 0 0 0
    65.5 1 0 9 5 7 11 11 7 7 5 2 1 0 0
    64.5 1 1 4 4 1 5 5 0 2 0 0 0 0 0
    64.0 1 0 2 4 1 2 2 1 1 0 0 0 0 0
    ;
proc print data=Galton(obs=100);
run;

data big; set galton; drop j count;
    do j=1 to count; output; end;
proc print data=big(obs=20);
proc means; var son parent;

data questions;
    /* these values used for prediction or estimation at x=68,x=72 */
    input parent son;
    cards;
    68 .
    72 .
    ;
run;

data big;
    set big questions;
run;
proc reg;
    model son=parent/clb;
    output out=out1 residual=r p=yhat ucl=pihigh lcl=pilow uclm=cihigh lclm=cilow stdp=stdmean;

data questions; set out1;
    if son=.;
proc print;
    title "questions regarding prediction, estimation when x=68, x=72";
run;
data fisherz;
  n=928;
  r=sqrt(0.2105);
  rratio=(1+r)/(1-r);
  z=probit(0.975);
  expon=probit(0.975)/sqrt(n-3);
  rlow=(rratio*exp(-2*expon)-1)/(rratio*exp(-2*expon)+1);
  rhigh=(rratio*exp(2*expon)-1)/(rratio*exp(2*expon)+1);
run;

proc print;run;

*goptions dev=pslepsf colors=(black);

symbol1 i=rl
  value=dot;

proc gplot;
  plot son*parent;
run;
quit;

proc univariate data=out1 normal plot;
  title "residual analysis";
  var r;
run;
*/

The SAS System

The MEANS Procedure

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>son</td>
<td>928</td>
<td>68.0884698</td>
<td>2.5179414</td>
<td>61.7000000</td>
<td>73.7000000</td>
</tr>
<tr>
<td>parent</td>
<td>928</td>
<td>68.3081897</td>
<td>1.7873334</td>
<td>64.0000000</td>
<td>73.0000000</td>
</tr>
</tbody>
</table>
The REG Procedure
Dependent Variable: son

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>1236.93401</td>
<td>1236.93401</td>
<td>246.84</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>926</td>
<td>4640.27261</td>
<td>5.01109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>927</td>
<td>5877.20663</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.23855
R-Square 0.2105
Dependent Mean 68.08847
Adj R-Sq 0.2096

Coeff Var 3.28770

Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>23.94153</td>
<td>2.81088</td>
<td>8.52</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>parent</td>
<td>1</td>
<td>0.64629</td>
<td>0.04114</td>
<td>15.71</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>18.42510 29.45796</td>
</tr>
<tr>
<td>parent</td>
<td>1</td>
<td>0.56556 0.72702</td>
</tr>
</tbody>
</table>

Questions regarding prediction, estimation when $x=68, x=72$

<table>
<thead>
<tr>
<th>Obs</th>
<th>parent</th>
<th>son</th>
<th>yhat</th>
<th>stdmean</th>
<th>cilow</th>
<th>cihigh</th>
<th>pilow</th>
<th>pihigh</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>.</td>
<td>67.8893</td>
<td>0.07457</td>
<td>67.7429</td>
<td>68.0356</td>
<td>63.4936</td>
<td>72.2849</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>72</td>
<td>.</td>
<td>70.4745</td>
<td>0.16871</td>
<td>70.1434</td>
<td>70.8056</td>
<td>66.0688</td>
<td>74.8801</td>
<td>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>n</th>
<th>r</th>
<th>rratio</th>
<th>z</th>
<th>expon</th>
<th>rlow</th>
<th>rhigh</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>928</td>
<td>0.45880</td>
<td>2.69551</td>
<td>1.95996</td>
<td>0.064443</td>
<td>0.40645</td>
<td>0.50815</td>
</tr>
</tbody>
</table>
Answers to questions from simple linear regression analysis of Galton’s height data

1. Change in average son’s height (inches) per one inch increase in midparent height (in the whole population.)

2. $\beta_1$ is an unknown parameter.

3. $\hat{\beta}_1 = 0.65$ son inches/midparent inch (from output.)

4. $\hat{\beta}_1 = 0.65$ is an observed value of a statistic.

5. $\beta_1$ is the slope of the population mean, $\hat{\beta}_1$ is the slope from the SLR of the observed data. $\hat{\beta}_1 = \beta_1$ is unlikely.

6. $\hat{SE}(\hat{\beta}_1) = \sqrt{MS[E]/S_{xx}} = 0.04$ (from output.)

7. Add and subtract about 2 SE to get $(0.57, 0.73)$

8. $y = 23.9 + 0.65x$

9. $r^2 = 21\%$

10. $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1 = 67.9$ (from output also.)

11. Not sure, as $\mu(68) = \beta_0 + 68\beta_1$ is unknown.

12. $\mu(68) = \beta_0 + 68\beta_1$. 
13. $\sqrt{MS[E]} = 2.24$. Not a SE.

14. $\sqrt{MS[E]} = 2.24$. (Assume homoscedasticity.)

15. $SE(\hat{\beta}_0 + 68\hat{\beta}_1) = 0.07$. Expressions given by

$$
\widehat{SE}(\mu(68)) = \sqrt{MS[E]\left(\frac{1}{n} + \frac{(68 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}
$$

$$
= \sqrt{(1, 68)'MS[E](X'X)^{-1}(1, 68)}
$$

$X$ a $(928 \times 2)$ design matrix.

16. $\widehat{SE}(\mu(72)) > \widehat{SE}(\mu(68))$

17. $r = \sqrt{0.21} = 0.46$, moderate, positive.

18. $\rho$ is a population correlation coefficient.

19. True.

20. These parameters describe the bivariate population of son and midparent heights.

21. Using the complicated expression in Rao and in notes, the confidence interval is

$$
\left(\frac{1+r e^{-2z/\sqrt{n-3}}}{1-r} - 1, \frac{1+r e^{2z/\sqrt{n-3}}}{1-r} - 1\right)
$$

or $0.41 < \rho < 0.51$.

22. Residuals reasonably symmetric, no heavy tails.
**Topic:** Multiple linear regression (MLR)

**Reading**  
Rao, Ch. 11

Multiple linear regression - an example

A random sample of students taking the same exam:

<table>
<thead>
<tr>
<th>IQ</th>
<th>Study TIME</th>
<th>GRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>105</td>
<td>10</td>
<td>75</td>
</tr>
<tr>
<td>110</td>
<td>12</td>
<td>79</td>
</tr>
<tr>
<td>120</td>
<td>6</td>
<td>68</td>
</tr>
<tr>
<td>116</td>
<td>13</td>
<td>85</td>
</tr>
<tr>
<td>122</td>
<td>16</td>
<td>91</td>
</tr>
<tr>
<td>130</td>
<td>8</td>
<td>79</td>
</tr>
<tr>
<td>114</td>
<td>20</td>
<td>98</td>
</tr>
<tr>
<td>102</td>
<td>15</td>
<td>76</td>
</tr>
</tbody>
</table>

Consider a regression model for the GRADE of subject \( i \), \( Y_i \), in which the mean of \( Y_i \) is a linear function of two predictor variables \( X_{i1} = IQ \) and \( X_{i2} = \text{Study TIME} \) for subjects \( i = 1, \ldots, 8 \):

\[
Y = \beta_0 + \beta_1 IQ + \beta_2 \text{TIME} + \text{error}
\]

or

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + E_i
\]

or

\[
\begin{align*}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + E_1 \\
Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + E_2 \\
\vdots &= \vdots \\
Y_8 &= \beta_0 + \beta_1 X_{81} + \beta_2 X_{82} + E_8
\end{align*}
\]
GRADE AND STUDY TIME EXAMPLE FROM ST 512 NOTES

Plot of STUDY*IQ. Symbol used is ‘*’.

<table>
<thead>
<tr>
<th>STUDY</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>*(98)</td>
</tr>
<tr>
<td>19</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>*(91)</td>
</tr>
<tr>
<td>15</td>
<td>*(76)</td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>*(85)</td>
</tr>
<tr>
<td>12</td>
<td>*(79)</td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>*(75)</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>*(79)</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>*(68)</td>
</tr>
</tbody>
</table>

---

100 105 110 115 120 125 130

IQ
A multiple linear regression (MLR) model w/ \( p \) independent variables

Let \( p \) independent variables be denoted by \( x_1, \ldots, x_p \).

- Observed values of \( p \) independent variables for \( i^{th} \) subject from sample denoted by \( x_{i1}, x_{i2}, \ldots, x_{ip} \)
- response variable for \( i^{th} \) subject denoted by \( Y_i \)
- For \( i = 1, \ldots, n \), MLR model for \( Y_i \):
  \[
  Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + E_i.
  \]

- As in SLR, \( E_1, \ldots, E_n \uparrow id \sim N(0, \sigma^2) \),

Least squares estimates of regression parameters minimize \( SS[E] \):

\[
SS[E] = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2
\]

\[
\hat{\sigma}^2 = \frac{SS[E]}{n-p-1}
\]

Interpretations of regression parameters:

- \( \sigma^2 \) is unknown error variance parameter.
- \( \beta_0, \beta_1, \ldots, \beta_p \) are \( p + 1 \) unknown regression parameters:
  - \( \beta_0 \) : average response when \( x_1 = x_2 = \ldots = x_p = 0 \)
  - \( \beta_i \) is called a partial slope for \( x_i \). Represents mean change in \( y \) per unit increase in \( x_i \) with all other independent variables held fixed.

For this example, with \( p = 2 \) and \( n = 8 \),

\[
\hat{\beta}_0 = 0.74, \quad \hat{\beta}_1 = 0.47, \quad \hat{\beta}_2 = 2.1
\]

What is the uncertainty associated with these parameter estimates?
Matrix formulation of MLR

Let a \((1 \times (p + 1))\) vector for \(p\) observed independent variables for individual \(i\) be defined by

\[ x_i = (1, x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip}). \]

The MLR model for \(Y_1, \ldots, Y_n\) is given by

\[
\begin{align*}
Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + E_1 \\
Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + E_2 \\
\vdots &= \vdots \\
Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + E_n
\end{align*}
\]

This system of \(n\) equations can be expressed using matrices:

\[
Y = X\beta + E
\]

where

- \(Y\) denotes a response vector \((n \times 1)\)
- \(X\) denotes a design matrix \((n \times (p + 1))\)
- \(\beta\) denotes a vector of regression parameters \(((p + 1) \times 1)\)
- \(E\) denotes an error vector \((n \times 1)\)

Here, the error vector \(E\) is assumed to follow a multivariate normal distribution with variance-covariance matrix \(\sigma^2 I_n\)

For individual \(i\),

\[ Y_i = x_{i.}\beta + E_i \]
Some simplified expressions: \((a\) is a known \(p \times 1\) vector)\

\[
\begin{align*}
\hat{\beta} &= (X'X)^{-1}X'Y \\
\text{Var}(\hat{\beta}) &= \sigma^2(X'X)^{-1} \\
&= \Sigma \\
\hat{\text{Var}}(\hat{\beta}) &= MS[E](X'X)^{-1} \\
&= \hat{\Sigma} \\
\hat{\text{Var}}(a'\hat{\beta}) &= a'\hat{\Sigma}a
\end{align*}
\]

(What are the dimensions of each of these quantities?)

- \((X'X)^{-1}\) may be verbalized as “x prime x inverse”
- \(\hat{\Sigma}\) is the estimated variance-covariance matrix for the estimate of the regression parameter vector \(\hat{\beta}\)
- \(X\) is assumed to be of full rank

Some more simplified expressions:

\[
\begin{align*}
\hat{Y} &= X\hat{\beta} \\
&= X(X'X)^{-1}X'y \\
&= Hy \\
e &= Y - \hat{Y} \\
&= Y - X\hat{\beta} \\
&= (I - H)Y
\end{align*}
\]

- \(\hat{Y}\) is called the vector of **fitted** or predicted values
- \(H = X(X'X)^{-1}X\) is called the **hat matrix**
- \(e\) is the vector of **residuals**
For the IQ, Study TIME example, with $p = 2$ independent variables and $n = 8$ observations, consider $X, Y, (X'X)^{-1}, (X'X)^{-1}X; Y, X(X'X)^{-1}X'Y:\n$

\[
X = \begin{pmatrix}
1 & 105 & 10 \\
1 & 110 & 12 \\
1 & 120 & 6 \\
1 & 116 & 13 \\
1 & 122 & 16 \\
1 & 130 & 8 \\
1 & 114 & 20 \\
1 & 102 & 15
\end{pmatrix}
\]

and

\[
X'X = \begin{pmatrix}
8 & 919 & 100 \\
919 & 106165 & 11400 \\
100 & 11400 & 1394
\end{pmatrix}
\]

\[
(X'X)^{-1} = \begin{pmatrix}
28.90 & -0.23 & -0.22 \\
-0.23 & 0.0018 & 0.0011 \\
-0.22 & 0.0011 & 0.0076
\end{pmatrix}
\]

\[
(X'X)^{-1}X'Y = \begin{pmatrix}
0.74 \\
0.47 \\
2.10
\end{pmatrix}
\]

\[
SS[E] = e'e = (Y - \hat{Y})'(Y - \hat{Y}) = 45.8, \quad e'e/df = 9.15 =?
\]

\[
\hat{\Sigma} = MS[E](X'X)^{-1} = \begin{pmatrix}
264.45 & -2.07 & -2.05 \\
-2.07 & 0.017 & 0.010 \\
-2.05 & 0.010 & 0.070
\end{pmatrix}
\]
Some questions - use preceding page

1. What is the estimate for $\beta_1$? Interpretation?

2. What is the standard error of $\hat{\beta}_1$?

3. Is $\beta_1 = 0$ plausible, while controlling for possible linear associations between Test Score and Study time? ($t(0.025, 5) = 2.57$)

4. Estimate the mean grade among the population of ALL students with $IQ = 113$ who study $TIME = 14$ hours.

5. Report a standard error

6. Report a 95% confidence interval

Some answers

1. $\hat{\beta}_1 = 0.47$ (second element of $(X'X)^{-1}X'Y$, exam points per IQ point for students studying the same amount)

2. $\sqrt{0.017} = 0.13$ (square root of middle element of $\hat{\Sigma}$)

3. $H_0 : \beta_1 = 0$, T-statistic: $t = (\hat{\beta}_1 - 0)/SE(\hat{\beta}_1)$
   Observed value is $t = .47/\sqrt{.017} = .47/.13 = 3.6 > 2.57$, ("$\hat{\beta}_1$ differs significantly from 0.")

4. Unknown population mean: $\theta = \beta_0 + \beta_1(113) + \beta_1(14)$
   Estimate: $\hat{\theta} = (1, 113, 14) \ast \hat{\beta} = 83.6$

5. $\text{Var}((1, 113, 14) \ast \hat{\beta}) = (1, 113, 14) \text{Var}(\hat{\beta})(1, 113, 14)'$
   or $(1, 113, 14) \hat{\Sigma}(1, 113, 14)' = 1.3$ or $SE(\hat{\theta}) = \sqrt{1.3} = 1.14$

6. $\hat{\theta} \pm t(0.025, 5)SE(\hat{\theta})$ or $83.6 \pm 2.57(1.14)$ or $(80.7, 86.6)$
DATA GRADES; INPUT IQ STUDY GRADE $; CARDS;
105 10 75 110 12 79 120 6 68 116 13 85 122 16 91 130 8 79 114 20 98 102 15 76
DATA EXTRA; INPUT IQ STUDY GRADE; CARDS;
113 14 .
DATA BOTH; SET GRADES EXTRA;
PROC REG; MODEL GRADE = IQ STUDY/P CLM XPX INV COVB;

The SAS System
The REG Procedure

Model Crossproducts X'X X'Y Y'Y

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intercept</th>
<th>IQ</th>
<th>STUDY</th>
<th>GRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>8</td>
<td>919</td>
<td>100</td>
<td>651</td>
</tr>
<tr>
<td>IQ</td>
<td>919</td>
<td>106165</td>
<td>11400</td>
<td>74881</td>
</tr>
<tr>
<td>STUDY</td>
<td>100</td>
<td>11400</td>
<td>1394</td>
<td>8399</td>
</tr>
<tr>
<td>GRADE</td>
<td>651</td>
<td>74881</td>
<td>8399</td>
<td>53617</td>
</tr>
</tbody>
</table>

X'X Inverse, Parameter Estimates, and SSE

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intercept</th>
<th>IQ</th>
<th>STUDY</th>
<th>GRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>28.898526711</td>
<td>-0.226082693</td>
<td>-0.224182192</td>
<td>0.7365546771</td>
</tr>
<tr>
<td>IQ</td>
<td>-0.226082693</td>
<td>0.00118460178</td>
<td>0.0011217122</td>
<td>0.473083715</td>
</tr>
<tr>
<td>STUDY</td>
<td>-0.224182192</td>
<td>0.0011217122</td>
<td>0.0076260404</td>
<td>2.1034362851</td>
</tr>
<tr>
<td>GRADE</td>
<td>0.7365546771</td>
<td>0.473083715</td>
<td>2.1034362851</td>
<td>45.759884688</td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>596.11512</td>
<td>298.05756</td>
<td>32.57</td>
<td>0.0014</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>45.75988</td>
<td>9.15198</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>7</td>
<td>641.87500</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimates

| Variable | DF | Estimate | Error | t Value | Pr > |t| |
|----------|----|----------|-------|---------|------|
| Intercept| 1  | 0.73655  | 16.2628 | 0.05    | 0.9656 |
| IQ       | 1  | 0.47308  | 0.12998 | 3.64    | 0.0149 |
| STUDY    | 1  | 2.10344  | 0.26418 | 7.96    | 0.0005 |

Covariance of Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intercept</th>
<th>IQ</th>
<th>STUDY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>264.47864999</td>
<td>-2.069103589</td>
<td>-2.051710248</td>
</tr>
<tr>
<td>IQ</td>
<td>-2.069103589</td>
<td>0.016894712</td>
<td>0.010265884</td>
</tr>
<tr>
<td>STUDY</td>
<td>-2.051710248</td>
<td>0.010265884</td>
<td>0.0697933458</td>
</tr>
</tbody>
</table>

Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Variable</th>
<th>Predicted</th>
<th>Std Error</th>
<th>95% CL Mean</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75.0000</td>
<td>71.4447</td>
<td>1.9325</td>
<td>66.4770</td>
<td>76.4124</td>
</tr>
<tr>
<td>8</td>
<td>76.0000</td>
<td>80.5426</td>
<td>1.9287</td>
<td>75.5847</td>
<td>85.5005</td>
</tr>
<tr>
<td>9</td>
<td>.</td>
<td>83.6431</td>
<td>1.1414</td>
<td>80.7092</td>
<td>86.5771</td>
</tr>
</tbody>
</table>

(abbreviated)
Model Selection

$x_1, x_2, x_3$ denote $p$ independent variables. Consider several models:

1. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1$
2. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2$
3. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_3 x_3$
4. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
5. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$
6. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
7. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2 + \beta_3 x_3$

$A$ is nested in $B$ means model $A$ can be obtained by restricting (e.g. setting to 0) parameter values in model $B$.

True or false:

- Model 1 nested in Model 4
- Model 2 nested in Model 4
- Model 3 nested in Model 4
- Model 1 nested in Model 5
- Model 4 nested in Model 1
- Model 5 nested in Model 4

$A$ nested in $B \rightarrow A$ called reduced, $B$ called full.

$p$ - number of regression parameters in full model
$q$ - number of regression parameters in reduced model
$p - q$ - number of regression parameters being tested.

Recall:

$$SS[R] = \sum (\hat{Y}_i - \bar{Y})^2$$
$$SS[E] = \sum (\hat{Y}_i - Y_i)^2$$
$$SS[Tot] = \sum (Y_i - \bar{Y})^2$$
Model Selection - concepts

In comparing two models, suppose
\( \beta_1, \ldots, \beta_q \) in reduced model (A)
\( \beta_1, \ldots, \beta_q, \beta_{q+1}, \ldots, \beta_p \) in full model (B).

Comparison of models A and B amounts to testing

- \( H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0 \) (model A ok)
- \( H_1 : \beta_{q+1}, \beta_{q+2}, \cdots, \beta_p \) not all 0 (need model B)

Let

\[
F = \frac{(SS[E]_r - SS[E]_f) / (p - q)}{MS(E)_f} = \frac{MS[H_0]}{MS[E]}
\]

\((r\) and \(f\) abbreviate reduced and full, respectively.)

Difference in the numerator called an extra regression sum of squares:

\[
R(\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p | \beta_0, \beta_1, \beta_2, \ldots, \beta_q) = SS[R]_f - SS[R]_r.
\]

(ok to supress \( \beta_0 \) in these extra SS terms.)

Theory gives that if \( H_0 \) holds (model A is appropriate) \( F \) behaves according to the \( F \) distribution with \( p - q \) numerator and \( n - p - 1 \) denominator degrees of freedom.

Extra SS terms for comparing some of the nested models on preceding page:

- Model 1 in model 4: \( R(\beta_2, \beta_3 | \beta_1) \)
- Model 2 in model 4 ?
- Model 3 in model 4 ?
- Model 1 in model 5: \( R(\beta_3 | \beta_1) \)
- Model 5 in model 4: ?
An example: How to measure body fat?

For each of $n = 20$ healthy individuals, the following measurements were made: bodyfat percentage $y_i$, triceps skinfold thickness, $x_1$, thigh circumference $x_2$, midarm circumference $x_3$

\[
\begin{array}{cccc}
 x_1 & x_2 & x_3 & y \\
 19.5 & 43.1 & 29.1 & 11.9 \\
 24.7 & 49.8 & 28.2 & 22.8 \\
 30.7 & 51.9 & 37.0 & 18.7 \\
 29.8 & 54.3 & 31.1 & 20.1 \\
 19.1 & 42.2 & 30.9 & 12.9 \\
 25.6 & 53.9 & 23.7 & 21.7 \\
 31.4 & 58.5 & 27.6 & 27.1 \\
 27.9 & 52.1 & 30.6 & 25.4 \\
 22.1 & 49.9 & 23.2 & 21.3 \\
 25.5 & 53.5 & 24.8 & 19.3 \\
 31.1 & 56.6 & 30.0 & 25.4 \\
 30.4 & 56.7 & 28.3 & 27.2 \\
 18.7 & 46.5 & 23.0 & 11.7 \\
 19.7 & 44.2 & 28.6 & 17.8 \\
 14.6 & 42.7 & 21.3 & 12.8 \\
 29.5 & 54.4 & 30.1 & 23.9 \\
 27.7 & 55.3 & 25.7 & 22.6 \\
 30.2 & 58.6 & 24.6 & 25.4 \\
 22.7 & 48.2 & 27.1 & 14.8 \\
 26.2 & 51.0 & 27.5 & 21.1 \\
\end{array}
\]

Summary statistics:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Variable</th>
<th>mean</th>
<th>st. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>Body fat</td>
<td>20.2</td>
<td>5.1</td>
</tr>
<tr>
<td>$x_1$</td>
<td>Triceps</td>
<td>25.3</td>
<td>5.0</td>
</tr>
<tr>
<td>$x_2$</td>
<td>Thigh Circ.</td>
<td>51.2</td>
<td>5.2</td>
</tr>
<tr>
<td>$x_3$</td>
<td>Midarm Circ.</td>
<td>27.6</td>
<td>3.6</td>
</tr>
</tbody>
</table>
Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

<table>
<thead>
<tr>
<th></th>
<th>y</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1.00000</td>
<td>0.84327</td>
<td>0.87809</td>
<td>0.14244</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.5491</td>
</tr>
<tr>
<td>x1</td>
<td>0.84327</td>
<td>1.00000</td>
<td>0.92384</td>
<td>0.45778</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.0424</td>
</tr>
<tr>
<td>x2</td>
<td>0.87809</td>
<td>0.92384</td>
<td>1.00000</td>
<td>0.08467</td>
</tr>
<tr>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.7227</td>
</tr>
<tr>
<td>x3</td>
<td>0.14244</td>
<td>0.45778</td>
<td>0.08467</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.5491</td>
<td>0.0424</td>
<td>0.7227</td>
<td></td>
</tr>
</tbody>
</table>

Marginal associations between $y$ and $x_1$ and between $y$ and $x_2$ are highly significant, providing evidence of a strong $r \approx 0.85$ linear association between average bodyfat and triceps skinfold and between average bodyfat and thigh circumference.

**Multicollinearity:** linear associations among the independent variables; causes problems such as inflated sampling variances for $\hat{\beta}$. 
data bodyfat;
  input x1 x2 x3 y;
cards;
19.5 43.1 29.1 11.9
24.7 49.8 28.2 22.8
(data abbreviated)
22.7 48.2 27.1 14.8
25.2 51.0 27.5 21.1
;

proc reg data=bodyfat;
  model y=x1 x2 x3;
  model y=x1;
  model y=x2;
  model y=x3;
  *model y=x1 x2 x3/xpx i covb corrb;

Yields the following (abbreviated) output

The SAS System
The REG Procedure

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|----------|----------------|---------|------|
| Intercept| 1  | 117.08469| 99.78240       | 1.17    | 0.2578 |
| x1       | 1  | 4.33409  | 3.01551        | 1.44    | 0.1699 |
| x2       | 1  | -2.85685 | 2.58202        | -1.11   | 0.2849 |
| x3       | 1  | -2.18606 | 1.59550        | -1.37   | 0.1896 |

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|----------|----------------|---------|------|
| Intercept| 1  | -1.49610 | 3.31923        | -0.45   | 0.6576 |
| x1       | 1  | 0.85719  | 0.12878        | 6.66    | <.0001 |

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|----------|----------------|---------|------|
| Intercept| 1  | -23.63449| 5.65741        | -4.18   | 0.0006 |
| x2       | 1  | 0.85655  | 0.11002        | 7.79    | <.0001 |

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|----------|----------------|---------|------|
| Intercept| 1  | 14.68678 | 9.09593        | 1.61    | 0.1238 |
| x3       | 1  | 0.19943  | 0.32663        | 0.61    | 0.5491 |
Model Selection - examples

In the bodyfat data, consider comparing the simple model that \( Y \) depends only on \( x_1 \) (triceps) and not on \( x_2 \) (thigh) or \( x_3 \) (midarm) after accounting for \( x_1 \) versus the full model that it depends on all three.

Model \( A : \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 \)

Model \( B : \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \)

or the null hypothesis

\[ H_0 : \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_1 : \beta_2, \beta_3 \text{ not both 0} \]

after accounting for \( x_1 \).

\[ F = \frac{(396.9 - 352.3)/2}{6.15} = \frac{22.3}{6.15} = 3.64 \]

How many df? The 95\(^{th}\) percentile is \( F(0.05, \ , \ ) = 3.63. \)

Q: Conclusion from this comparison of nested models?

After accounting for variation in bodyfat explained by linear dependence on triceps, there is still some linear association between mean bodyfat and at least one of \( x_2, x_3 \) (thigh,midarm).

To get this \( F \)-ratio in SAS, try

```sas
proc reg data=bodyfat;
   model y=x1 x2 x3;
   test x2=0,x3=0;
run;
```

Adding \( x_2, x_3 \) leads to a model that is too complex for it’s own good.
**PROC GLM** can replace **PROC REG** to get the **SUMS OF SQUARES** for use in model selection as in the following output:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>396.9846118</td>
<td>132.3282039</td>
<td>21.52</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>98.4048882</td>
<td>6.1503055</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>495.3895000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1</td>
<td>352.2697968</td>
<td>352.2697968</td>
<td>57.28</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>x2</td>
<td>1</td>
<td>33.1689128</td>
<td>33.1689128</td>
<td>5.39</td>
<td>0.0337</td>
</tr>
<tr>
<td>x3</td>
<td>1</td>
<td>11.5459022</td>
<td>11.5459022</td>
<td>1.88</td>
<td>0.1896</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1</td>
<td>12.70489278</td>
<td>12.70489278</td>
<td>2.07</td>
<td>0.1699</td>
</tr>
<tr>
<td>x2</td>
<td>1</td>
<td>7.52927788</td>
<td>7.52927788</td>
<td>1.22</td>
<td>0.2849</td>
</tr>
<tr>
<td>x3</td>
<td>1</td>
<td>11.54590217</td>
<td>11.54590217</td>
<td>1.88</td>
<td>0.1896</td>
</tr>
</tbody>
</table>

| Parameter       | Estimate | Error         | t Value | Pr > |t| |
|-----------------|----------|---------------|---------|------|------|
| Intercept       | 117.0846948 | 99.78240295   | 1.17    | 0.2578|
| x1              | 4.3340920  | 3.01551136    | 1.44    | 0.1699|
| x2              | -2.8568479 | 2.58201527    | -1.11   | 0.2849|
| x3              | -2.1860603 | 1.59549900    | -1.37   | 0.1896|

Note agreement between \( p \)—values from Type III \( F \) tests and \( p \)—values from \( t \) tests from parameter estimates from MLR.
Type I sums of squares - sequential (order of selection matters)
Type III sums of squares - partial
Type II sums of squares - partial (change in SSE due to adding term A to model with all other terms not ‘containing’ A)

In the output on the preceding page,

\[ R(\beta_1|\beta_0) = 352.3 \]
\[ R(\beta_2|\beta_0, \beta_1) = 33.2 \]
\[ R(\beta_3|\beta_0, \beta_1, \beta_2) = 11.5 \]
\[ R(\beta_1|\beta_0, \beta_2, \beta_3) = 12.7 \]
\[ R(\beta_2|\beta_0, \beta_1, \beta_3) = 7.5 \]

Type III test for \( \beta_j \) - test of partial association between \( y \) and \( x_j \) after accounting for all other \( x_i \)

Type III \( F \)-ratios from bodyfat data for \( x_1, x_2, x_3 \), respectively:

\[ F = \frac{12.7/1}{6.15} = 2.07, \quad F = \frac{7.5/1}{6.15} = 1.22, \quad F = \frac{11.5/1}{6.15} = 1.88. \]

(Partial) effects significant? (Use \( F(0.95, 1, 16) = 4.49 \).)

Exercise: specify the comparison of nested models that corresponds to each of these \( F \)-ratios.
In GLM output, which models are the type I tests comparing?

1. Type I SS for $x_1$ from PROC GLM appropriate for SLR of $y$ on $x_1$.
2. Type I SS for $x_2$ from PROC GLM appropriate for test of association between $y$ and $x_2$ after accounting for $x_1$.
3. Type I test for $x_3$ from PROC GLM same as type III test for $x_3$.

In all three of these tests, $MS[E]$ computed from full model (#4).

Some model comparison examples

1. Compare models 1 and 6
2. Compare models 2 and 6

For 1. use $R(\beta_2|\beta_0, \beta_1)$ in the $F$ ratio:

\[
F = \frac{R(\beta_2|\beta_0, \beta_1)}{MS[E]_6} = \frac{33.2}{(SS[Tot] - R(\beta_1, \beta_2|\beta_0))/(20 - 2 - 1)} = \frac{(495.4 - 352.3 - 33.2)/(20 - 2 - 1)}{33.2} = \frac{109.9/17}{33.2} = 5.1
\]

Note that

$SS[E]_f = (SS[Tot] - SS[R]_f)$ and $SS[R]_f = SS[R]_r + R(\beta_2|\beta_0, \beta_1)$

$F(0.05, 1, 17) = 4.45$ : nested model 1 is rejected in favor of model 6: there is evidence ($p = 0.037$) of association between $y$ and $x_2$ after accounting for dependence on $x_1$. 
To compare models 2 and 6, we need $SS[R]_r = R(\beta_2|\beta_0) = 382.0$ which cannot be gleaned from preceding output. You could also get it from $r^2_{yx_2} \times SS[Tot]$ or from running something like

```plaintext
proc reg;
  model y=x1 x2/ss1 ss2;
run;
```

The REG Procedure

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>385.43871</td>
<td>192.71935</td>
<td>29.80</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>17</td>
<td>109.95079</td>
<td>6.46769</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>495.38950</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Variable | DF | Estimate | Error | t Value | Pr > |t|   | Type I SS | Type II SS |
|----------|----|----------|-------|---------|-------|-----|-----------|------------|
| Intercept| 1  |-19.17425 | 8.36064 | -2.29   | 0.0348 | 8156.76050 | 34.01785   |
| x1       | 1  | 0.22235  | 0.30344 | 0.73    | 0.4737 | 352.26980  | 3.47289    |
| x2       | 1  | 0.65942  | 0.29119 | 2.26    | 0.0369 | 33.16891   | 33.16891   |

\[
F = \frac{R(\beta_1|\beta_0, \beta_2)/(\Delta df)}{MS[E]_f} \\
= \frac{(SS[R]_f - SS[R]_r)/1}{6.5} \\
= \frac{352.3 + 33.2 - 382.0}{6.5} \\
= \frac{3.4}{6.5} \approx 0.5
\]

Conclusions?

- $x_2$ gives you a little when you add it to model with $x_1$
- $x_1$ gives you nothing when you add it to model with $x_2$
- Take model with $x_2$. (Has higher $r^2$ too.)

Note that all of these comparisons of nested models are easy to carry out using the TEST statement in PROC REG.
Another example, revisiting test scores and study times
Consider this sequence of analyses:

1. Regress GRADE on IQ.
2. Regress GRADE on IQ and TIME.
3. Regress GRADE on TIME IQ TI where TI = TIME*IQ.

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ</td>
<td>1</td>
<td>15.9393</td>
<td>15.9393</td>
<td>0.153</td>
<td>0.71</td>
</tr>
<tr>
<td>Error</td>
<td>6</td>
<td>625.935</td>
<td>104.32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It appears that IQ has nothing to do with grade, but we did not look at study time. Looking at the *multiple* regression we get

The REG Procedure

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>596.11512</td>
<td>298.05756</td>
<td>32.57</td>
<td>0.0014</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>45.75988</td>
<td>9.15198</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>7</td>
<td>641.87500</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Variable | DF | Estimate | Error | t Value | Pr > |t| |
|----------|----|----------|-------|---------|-------|
| Intercept | 1  | 0.73655  | 16.26280 | 0.05 | 0.9656 |
| IQ       | 1  | 0.47308  | 0.12998 | 3.64 | 0.0149 |
| study    | 1  | 2.10344  | 0.26418 | 7.96 | 0.0005 |

Now the test for dependence on IQ is significant $p = 0.0149$. Why?
The interaction model

The SAS System
The REG Procedure

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>610.81033</td>
<td>203.60344</td>
<td>26.22</td>
<td>0.0043</td>
</tr>
<tr>
<td>Error</td>
<td>4</td>
<td>31.06467</td>
<td>7.76617</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>7</td>
<td>641.87500</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimates

| Variable  | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| | Type I SS | Type II SS |
|-----------|----|--------------------|----------------|---------|------|---------|-----------|------------|
| Intercept | 1  | 72.20608           | 54.07278       | 1.34    | 0.2527| 52975   | 13.84832  |
| IQ        | 1  | -0.13117           | 0.45530        | -0.29   | 0.7876| 15.93930| 0.64459   |
| study     | 1  | -4.11107           | 4.52430        | -0.91   | 0.4149| 580.17582| 6.41230   |
| IQ_study  | 1  | 0.05307            | 0.03858        | 1.38    | 0.2410| 14.69521| 14.69521  |

Discussion of the interaction model. We call the product \( I*S = IQ*STUDY \) an ”interaction” term. Our model is

\[
\hat{G} = 72.21 - 0.13 * I - 4.11 * S + 0.0531(I * S)
\]

Now if IQ = 100 we get

\[
\hat{G} = (72.21 - 13.1) + (-4.11 + 5.31)S
\]

and if IQ 120 we get

\[
\hat{G} = (72.21 - 15.7) + (-4.11 + 6.37)S.
\]

Thus we expect an extra hour of study to increase the grade by 1.20 points for someone with IQ = 100 and by 2.26 points for someone with IQ = 120 if we use this interaction model. Since the interaction is not significant, we may want to go back to the simpler “main effects” model. (This example taken from Dickey's ST512 notes.)
Some questions about design matrices

Recall three models under consideration for the bodyfat data

\[ M_1 : \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 \]
\[ M_2 : \mu(x_1, x_2, x_3) = \beta_0 + \beta_2 x_2 \]
\[ M_6 : \mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \]

Q: \( MS[E]_{M_6} < MS[E]_{M_1} \) and \( MS[E]_{M_6} < MS[E]_{M_2} \) but the partial slopes have larger standard errors in \( M_6 \). Why?

Design matrices

\[
X_{M_6} = \begin{pmatrix}
1 & 19.5 & 43.1 \\
1 & 24.7 & 49.8 \\
\vdots & \vdots & \vdots \\
1 & 22.7 & 48.2 \\
1 & 25.2 & 51.0 \\
\end{pmatrix}
\]
\[
X_{M_1} = \begin{pmatrix}
1 & 19.5 \\
\vdots & \vdots \\
1 & 22.7 \\
1 & 25.2 \\
\end{pmatrix}
\]

(similarly for \( X_{M_2} \)).

\[
(X'X)_{M_6} = \begin{pmatrix}
? & 506.1 & 1023.4 \\
13386.3 & 26358.7 \\
52888.0 \\
\end{pmatrix}
\]
\[
(X'X)_{M_1} = \begin{pmatrix}
? & ? \\
? & ? \\
\end{pmatrix}
\]
\[
(X'X)^{-1}_{M_1} = \begin{pmatrix}
1.39 & -0.053 \\
-0.002 \\
\end{pmatrix}
\]
\[
(X'X)_{M_2} = \begin{pmatrix}
? & ? \\
? & ? \\
\end{pmatrix}
\]
\[
(X'X)^{-1}_{M_2} = \begin{pmatrix}
5.08 & -0.098 \\
0.0019 \\
\end{pmatrix}
\]
\[
(X'X)^{-1}_{M_6} = \begin{pmatrix}
10.8 & 0.29 & -0.35 \\
0.014 & -0.012 & 0.013 \\
\end{pmatrix}
\]

Q: Why is \( \text{Var}(\hat{\beta}_0) \) bigger in \( M_2 \) than in \( M_1 \)?
Recall the Resolution Run 5k race data

<table>
<thead>
<tr>
<th>Obs</th>
<th>age</th>
<th>sex</th>
<th>race</th>
<th>pace</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>M</td>
<td>16.6833</td>
<td>5.38333</td>
</tr>
<tr>
<td>2</td>
<td>39</td>
<td>M</td>
<td>16.9500</td>
<td>5.46667</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>M</td>
<td>17.1333</td>
<td>5.51667</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>M</td>
<td>17.4000</td>
<td>5.61667</td>
</tr>
<tr>
<td>157</td>
<td>52</td>
<td>F</td>
<td>46.8833</td>
<td>15.1000</td>
</tr>
<tr>
<td>158</td>
<td>10</td>
<td>F</td>
<td>53.6000</td>
<td>17.2667</td>
</tr>
<tr>
<td>159</td>
<td>10</td>
<td>F</td>
<td>53.6167</td>
<td>17.2667</td>
</tr>
<tr>
<td>160</td>
<td>81</td>
<td>M</td>
<td>54.3167</td>
<td>17.5000</td>
</tr>
</tbody>
</table>

(abbreviated)

Summary statistics ($n = 160$):

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Variable</th>
<th>mean</th>
<th>st. dev.</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>Pace</td>
<td>9.1</td>
<td>2.2</td>
<td>5.0</td>
</tr>
<tr>
<td>$x$</td>
<td>Age</td>
<td>35.1</td>
<td>14.7</td>
<td>216.5</td>
</tr>
</tbody>
</table>
Quadratic model for pace ($Y$) as a function of age ($x$):

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i \quad \text{for } i = 1, \ldots, 160$$

where $E_i \overset{iid}{\sim} N(0, \sigma^2)$.

- $\beta = (\beta_0, \beta_1, \beta_2)'$ is a vector of unknown regression parameters
- $\sigma^2$ is the unknown error variance of paces given age $x$.

Compare this model with the (previously discarded) SLR model

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{for } i = 1, \ldots, 160$$

Q1: Does $\beta_1$ have the same interpretation in both models?

Q2: How can we compare the two models?

A2: Using $F$-ratios to compare nested models (see output next page).

$$F = \frac{R(\beta_2 | \beta_0, \beta_1)}{MS[E]_{full}}$$

$$= \frac{(SS[R]_{full} - SS[R]_{red})/1}{MS[E]_{full}}$$

$$= \frac{(113.6 - 1.1)/1}{4.3}$$

$$= \frac{(SS[E]_{red} - SS[E]_{full})/1}{MS[E]_{full}}$$

$$= \frac{(787.0 - 674.4)/1}{4.3}$$

$$= 26.2$$

$$= \left( \frac{\hat{\beta}_2}{SE} \right)^2$$

with $F(0.05, 1, 157) = 3.90$. Since $26.2 >> 3.9$, the linear model is implausible when compared to the quadratic model.
/* age2 defined in data step as age*age */
PROC REG;
  MODEL pace=age;
  MODEL pace=age age2/ss1; /* ss1 generates sequential sums of squares */
  /* only the 2nd model statement really necessary */
RUN;

The SAS System
The REG Procedure
Model: MODEL1

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>R-Square</th>
<th>Adj R-Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>1.09650</td>
<td>1.09650</td>
<td>0.22</td>
<td>0.6396</td>
<td>0.0014</td>
<td>-0.0049</td>
</tr>
<tr>
<td>Error</td>
<td>158</td>
<td>786.99821</td>
<td>4.98100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>159</td>
<td>788.09472</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.23182 R-Square 0.0014
Dependent Mean 9.12063 Adj R-Sq -0.0049

Parameter Standard

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th>Type I SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>8.92271</td>
<td>0.45724</td>
<td>19.51</td>
<td>&lt;.0001</td>
<td>13310</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>0.00564</td>
<td>0.01203</td>
<td>0.47</td>
<td>0.6396</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model: MODEL2

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
<th>R-Square</th>
<th>Adj R-Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>113.64500</td>
<td>56.82250</td>
<td>13.23</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>157</td>
<td>674.44972</td>
<td>4.29586</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>159</td>
<td>788.09472</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.07265 R-Square 0.1442
Dependent Mean 9.12063 Adj R-Sq 0.1333

Parameter Standard

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th>Type I SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>11.78503</td>
<td>0.70216</td>
<td>16.78</td>
<td>&lt;.0001</td>
<td>13310</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>-0.19699</td>
<td>0.04113</td>
<td>-4.79</td>
<td>&lt;.0001</td>
<td>1.09650</td>
<td></td>
</tr>
<tr>
<td>age2</td>
<td>1</td>
<td>0.00294</td>
<td>0.00057380</td>
<td>5.12</td>
<td>&lt;.0001</td>
<td>112.54850</td>
<td></td>
</tr>
</tbody>
</table>
Fitted model is

\[ \hat{\mu}(x) = 11.785 - 0.197x + 0.00294x^2 \]

or

\[ \hat{\mu}(\text{age}) = 11.785 - 0.197\text{age} + 0.00294\text{age}^2. \]
Inference for response $Y$ given predictor $x_i$.

A random sample of $n = 31$ trees is drawn from a population of trees. On each tree, indexed by $i$, three variables are measured:

- $x_{i1}$: “girth”, tree diameter in inches
- $x_{i2}$: “height” (in feet)
- $Y_i$: volume of timber, in cubic feet.

Given $x_1$ and $x_2$, a MLR model for these data is given by

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + E_i \text{ for } i = 1, \ldots, n$$

where errors are assumed iid normal w/ constant variance $\sigma^2$.

For trees with $x_1, x_2$ the model for mean volume is

$$\mu(x_1, x_2) = E(Y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$
Some questions involving linear combinations of regression coefficients
Consider all trees with girth $x_{01} = 15$ in and height $x_{02} = 80$ ft.

- Estimate the mean volume among these trees, along with a standard error and 95% confidence interval.

- Obtain a 95% prediction interval of $y_0$, the volume from an individual tree sampled from this population of 80 footers, with girth 15 inches.

SAS generates $\hat{\beta}$ and $\widehat{Var}(\hat{\beta}) = MSE \times (X'X)^{-1}$

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>DF</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>7684.16251</td>
<td>28</td>
<td>3842.08126</td>
<td>254.97</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>421.92136</td>
<td>28</td>
<td>15.06862</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>8106.08387</td>
<td>30</td>
<td>8106.08387</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 3.88183  R-Square 0.9480
Dependent Mean 30.17097  Adj R-Sq 0.9442
Coef Var 12.86612

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|
| Intercept| 1  | -57.98766          | 8.63823        | -6.71   | <.0001|
| Girth    | 1  | 4.70816            | 0.26426        | 17.82   | <.0001|
| Height   | 1  | 0.33925            | 0.13015        | 2.61    | 0.0145|

Covariance of Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intercept</th>
<th>Girth</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>74.6189461</td>
<td>0.4321713812</td>
<td>-1.050768886</td>
</tr>
<tr>
<td>Girth</td>
<td>0.4321713812</td>
<td>0.0698357838</td>
<td>-0.017860301</td>
</tr>
<tr>
<td>Height</td>
<td>-1.050768886</td>
<td>-0.017860301</td>
<td>0.0169393298</td>
</tr>
</tbody>
</table>
Moments of linear combinations of random vectors (Appendix B)

Let $W$ denote a $p \times 1$ random vector with mean $\mu_W$ and covariance matrix $\Sigma_W$. Suppose $a$ is a $p \times 1$ (fixed) vector of coefficients. Then

$$
E(a'W) = a'\mu_W
$$
$$
\text{Var}(a'W) = a'\Sigma_W a.
$$

(See [http://www.stat.ncsu.edu/people/dickey/courses/st512/crsnotes/notes_1.htm](http://www.stat.ncsu.edu/people/dickey/courses/st512/crsnotes/notes_1.htm)

Inference for the mean response in MLR

Consider all trees with Girth 15 and Height 80 To estimate mean volume among these trees, along with an estimated standard error, take $x'_0 = (1, 15, 80)$ and consider $\hat{\mu}(x_0) = x'_0\hat{\beta}$.

$$
E(x'_0\hat{\beta}) = x'_0\beta
$$
$$
\text{Var}(x'_0\hat{\beta}) = x'_0\hat{\Sigma}x_0
$$

Substitution of $\hat{\beta}$ and $\hat{\Sigma} = MSE(X'X)^{-1}$ gives the estimates:

$$
\hat{\mu}(x_0) = (1, 15, 80) \begin{pmatrix} -58.0 \\ 4.71 \\ 0.34 \end{pmatrix} = 39.8
$$
$$
\hat{\text{Var}}(\hat{\mu}(x_0)) = (1, 15, 80) \begin{pmatrix} 74.62 & 0.43 & -1.05 \\ 0.43 & 0.070 & -0.018 \\ -1.05 & -0.018 & 0.017 \end{pmatrix} \begin{pmatrix} 1 \\ 15 \\ 80 \end{pmatrix} = 0.72
$$
$$
\hat{SE}(\hat{\mu}(x_0)) = \sqrt{0.72} = 0.849
$$

which can be obtained using PROC REG and the missing y trick:

<table>
<thead>
<tr>
<th>Obs</th>
<th>treenumber</th>
<th>Girth</th>
<th>Height</th>
<th>Volume</th>
<th>p sepred</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>100</td>
<td>15</td>
<td>80</td>
<td>.</td>
<td>39.7748</td>
</tr>
</tbody>
</table>

95% Prediction limits? Use $\pm t(.025, 28)\sqrt{0.72 + MSE}$.
**Partial correlations**

The partial correlation coefficient for $x_1$ in the MLR

$$E(Y|x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

is defined as the correlation coefficient between the residuals computed from the two regressions below:

$$Y = \beta_0 + \beta_2 x_2 + \cdots + \beta_p x_p + E$$
$$X_1 = \beta_0 + \beta_2 x_2 + \cdots + \beta_p x_p + E$$

Call these sets of residuals $e_{y, 2, 3, \ldots, p}$ and $e_{1, 2, 3, \ldots, p}$ respectively. The *partial correlation* between $y$ and $x_1$ after accounting for the linear association between $y$ and $x_2, x_3, \ldots, x_p$ is defined as

$$r_{y1, 2, 3, \ldots, p} = \text{correlation between } e_{y, 2, 3, \ldots, p} \text{ and } e_{1, 2, 3, \ldots, p}.$$  

The *partial coeff. of determination* is $r_{y1, 2, 3, \ldots, p}^2$.

Note also (see Figure 11.7) that

$$r_{y1, 2, \ldots, p}^2 = \frac{R(\beta_1|\beta_0, \beta_2, \ldots, \beta_p)}{SS[Tot] - R(\beta_2, \beta_3, \ldots, \beta_p|\beta_0)}.$$
Bodyfat data, compare models 1, 2 and 6 (ignore $x_3$.)

### bodyfat data

<table>
<thead>
<tr>
<th>Obs</th>
<th>x1</th>
<th>x2</th>
<th>y</th>
<th>$py_1$</th>
<th>ey_1</th>
<th>e2_1</th>
<th>$py_2$</th>
<th>ey_2</th>
<th>e1_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.5</td>
<td>43.1</td>
<td>11.9</td>
<td>15.2190</td>
<td>-3.31903</td>
<td>13.2827</td>
<td>-1.38267</td>
<td>1.34939</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>24.7</td>
<td>49.8</td>
<td>22.8</td>
<td>19.6764</td>
<td>3.12360</td>
<td>-0.78756</td>
<td>3.77847</td>
<td>0.60956</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30.7</td>
<td>51.9</td>
<td>18.7</td>
<td>24.8195</td>
<td>-6.11952</td>
<td>20.8203</td>
<td>-2.12028</td>
<td>4.74782</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>29.8</td>
<td>54.3</td>
<td>20.1</td>
<td>24.0481</td>
<td>-3.94805</td>
<td>22.8760</td>
<td>-2.77599</td>
<td>1.72013</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>19.1</td>
<td>42.2</td>
<td>12.9</td>
<td>14.8762</td>
<td>-6.11952</td>
<td>12.5118</td>
<td>0.38822</td>
<td>1.74728</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>25.2</td>
<td>51.0</td>
<td>21.1</td>
<td>20.1050</td>
<td>0.99500</td>
<td>-0.06892</td>
<td>20.0494</td>
<td>1.05061</td>
<td>0.04571</td>
</tr>
</tbody>
</table>

The partial correlation coefficient between $y$ and $x_1$ after accounting for $x_2$ is $r_{y1\cdot 2} = 0.17$ and the partial for $x_2$ after accounting for $x_1$ is $r_{y2\cdot 1} = 0.48$. The partial coefficients of determination are

$$r_{y1\cdot 2}^2 = 0.03062 \text{ and } r_{y2\cdot 1}^2 = 0.23176.$$

**Q:** If you had to choose one variable or the other from $x_1$ and $x_2$, which would it be?

**Q:** Anything wrong with throwing both $x_1$ and $x_2$ in the final model?

**Q:** Write the coefficients of determination in terms of extra sums of squares, using $R(\cdot | \cdot)$ notation.

Note: partial correlations obtained in SAS using `PCORR2` option:

<table>
<thead>
<tr>
<th>Squared Partial Corr Type II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable DF Corr Type II</td>
</tr>
<tr>
<td>Intercept 1 .</td>
</tr>
<tr>
<td>x1 1 0.03062</td>
</tr>
<tr>
<td>x2 1 0.23176</td>
</tr>
</tbody>
</table>
Partial regression plots

A plot of the residuals from the regression

\[ Y = \beta_0 + \beta_2 x_2 + \cdots + \beta_3 x_p + E \]

versus the residuals from the regression

\[ X_1 = \beta_0 + \beta_2 x_2 + \cdots + \beta_3 x_p + E \]

is called a partial regression plot for \( x_1 \) or a partial leverage plot of \( x_1 \) in the MLR. They can be generated

- in SAS/INSIGHT by clicking
  - Analyze • Fit XY • Output tab • Plots: (Partial Leverage).

- using the \texttt{PARTIAL} command in the MODEL statement of PROC REG

\begin{figure}
\centering
\includegraphics[width=\textwidth]{screencapture.png}
\caption{Partial Regression Residual Plot}
\end{figure}

\begin{verbatim}
SAS System
The REG Procedure
Partial Regression Residual Plot
\end{verbatim}
Q: What can these plots tell us?

A1: They convey info. about linear associations between $y$ and a candidate independent variable $x_i$ after accounting for linear associations between $y$ and other independent variables $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$.

A2: They can convey info about nonlinear associations between $y$ and $x_i$ after accounting for linear associations with other variables.

A3: They can illuminate possible outliers.
Some exercises (hint: use matrix algebra or SAS).

1. Suppose you are a local 32 yr-old male runner. Regarding these data as randomly sampled from the population of all local runners, fit a quadratic regression function and use it to obtain an estimate of the mean 5k pace in your cohort of all 32 yr-old male runners. Report a standard error and 95% confidence interval.

2. Obtain a 95% prediction interval for your time if you are about to run the race.

3. Explain the difference between the two intervals in questions 1 and 2.

4. At what rate is $\mu(x)$ changing with age? Estimate the appropriate function.

5. Estimate $\theta$, the peak age to run a 5k in the fastest time. Is $\theta$ a linear function of regression parameters? Can you obtain an unbiased estimate of the standard error of $\theta$?

**Residual diagnostics**

- Residuals can be plotted against independent variables to check for model inadequacy.

- Residuals can be plotted against predicted values to look for inhomogeneity of variance (heteroscedasticity).

- The sorted residuals can be plotted against the normal inverse of the empirical CDF of the residuals in a normal plot to assess the normal distributional assumption. A nonlinear association in such a q-q plot indicates nonnormality.
1. Obtain the observed quantiles by ordering the residuals:

\[ e(1) \leq e(2) \leq \cdots \leq e(n). \]

2. For each \( i = 1, \ldots, n \) compute the expected quantile from

\[ q(i) = z\left(1 - \frac{i}{n + 1}\right). \]

3. Plot the (ordered) residuals on the vertical axis versus the (ordered) theoretical quantiles on the horizontal axis.

As an illustration, we’ll obtain the \( e_{ij} \) and the \( q(i) \) for the data in a table. To do this, we’ll need the ranks of the residuals

\[ \text{Rank of } e(i) = i. \]

The *empirical cumulative probability* associated with \( e(i) \) is

\[ p_i = \frac{\text{Rank of } e(i)}{N + 1}. \]

These can be used to obtain the corresponding theoretical quantiles via

\[ q(i) = z(1 - p(i)). \]
ods listing close;
proc reg data=running;
   model pace=sexf age age2;
   output out=resids p=yhat r=resid;
run;
proc rank data=resids out=resids2;
   ranks rankresid;
   var resid;
run;
data resids2;
   set resids2;
   ecdf=rankresid/(160+1); *160 runners;
   q=probit(ecdf);
run;
ods listing;
proc print data=resids2;
   var age pace yhat resid rankresid ecdf q;
run;
proc gplot data=resids;
   plot resid*q;
run;

The SAS System 1

Obs age pace yhat resid rankresid ecdf q
1  28  5.3833  7.5837 -2.20040  14.0  0.08696 -1.35974
2  39  5.4667  7.7671 -2.30046  10.0  0.06211 -1.53728
3  41  5.5167  7.8735 -2.35681   6.0  0.03727 -1.78332
4  42  5.6167  7.9351 -2.31841   9.0  0.05590 -1.59015
5  40  5.9333  7.8175 -1.88416  18.0  0.11180 -1.21700
6  23  6.0000  7.7250 -1.72501  25.0  0.15527 -1.01405
7  40  6.0167  7.8175 -1.80083  23.0  0.14286 -1.06757
8  42  6.0667  7.9351 -1.86841  19.0  0.11801 -1.18498
9  28  6.0833  7.5837 -1.50040  27.0  0.16770 -0.96329
10 18  6.1833  8.0067 -1.82335  22.0  0.13665 -1.09551
11 32  6.4333  7.5718 -1.13847  42.0  0.26087 -0.64067
   ...
An exercise: Match up letters a,b,c,d with the model violation

1. Heteroscedasticity
2. Nonlinearity
3. Nonnormality
4. Model fits
ST 512
Topic: The general linear model
Reading Ch. 8, 9, 12

ANOVA revisited:
Following data come from study investigating binding fraction for several antibiotics using $n = 20$ bovine serum samples:

<table>
<thead>
<tr>
<th>Antibiotic</th>
<th>Binding Percentage</th>
<th>Sample mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penicillin G</td>
<td>29.6 24.3 28.5 32</td>
<td>28.6</td>
</tr>
<tr>
<td>Tetracyclin</td>
<td>27.3 32.6 30.8 34.8</td>
<td>31.4</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>5.8 6.2 11 8.3</td>
<td>7.8</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>21.6 17.4 18.3 19</td>
<td>19.1</td>
</tr>
<tr>
<td>Chloramphenicol</td>
<td>29.2 32.8 25 24.2</td>
<td>27.8</td>
</tr>
</tbody>
</table>

A completely randomized design (CRD) was used.
Q: Are the population means for these 5 treatments plausibly equal?
Q: Do these (sample) treatment means differ significantly?
Modelling the binding fraction expt

One model parameterizes antibiotic effects as differences from mean:

\[ Y_{ij} = \mu + \tau_i + E_{ij} \]

for \( i = 1, \ldots, 5 \) and \( j = 1, \ldots, 4 \), where \( E_{ij} \) are i.i.d. \( N(0, \sigma^2) \) errors.

Unknown parameters

- \( \mu \) - overall population mean (avg of 5 treatment population means)
- \( \tau_i \) - difference between (population) mean for treatment \( i \) and \( \mu \)
- \( \sigma^2 \) - (population) variance of bf for a given antibiotic

To test \( H_0 : \tau_1 = \tau_2 = \ldots = \tau_5 = 0 \), we just carry out one-way ANOVA:

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>4</td>
<td>1481</td>
<td>370</td>
<td>41</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>136</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>1617</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion? (Use \( F(0.05, 4, 15) = 3.06 \))

Parameter estimates \( \hat{\mu}, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4, \hat{\tau}_5 \)? \( (\bar{y}_{++} = 22.935) \)?

Standard errors of parameter estimates?
Table for balanced one-way ANOVA

Y_{ij} denotes \( j^{th} \) observation receiving level \( i \) of treatment factor with \( t \) levels, for a total of \( N \) observations.

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>( t-1 )</td>
<td>( SS[T] )</td>
<td>( MS[T] = \frac{SS[T]}{(t-1)} )</td>
<td>( F = \frac{MS[T]}{MS[E]} )</td>
</tr>
<tr>
<td>Error</td>
<td>( N-t )</td>
<td>( SS[E] )</td>
<td>( MS[E] = \frac{SS[E]}{(N-t)} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( N-1 )</td>
<td>( SS[TOT] )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where

\[
SS[T] = \sum \sum (\bar{y}_{i+} - \bar{y}_{++})^2
\]
\[
SS[E] = \sum \sum (y_{ij} - \bar{y}_{i+})^2
\]
\[
SS[TOT] = \sum \sum (y_{ij} - \bar{y}_{++})^2
\]

The linear model \( \mu_{ij} = E(Y_{ij}) = \mu + \tau_i \) could be fit using MLR with 5 indicator variables \( x_1, \ldots, x_5 \) for the 5 antibiotics. Let

\[
x_{ij} = \begin{cases} 
1 & \text{if treatment } j \\
0 & \text{else}
\end{cases}
\]

The MLR model is

\[
Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + E_i \quad i = 1, \ldots, 20
\]

where a design matrix \( X \) of 1’s and 0’s of dimension \( (20 \times 6) \) could be specified. Note that \( \beta_0 = \mu \) and \( \beta_j = \tau_j \).

- problem: \( (X'X)^{-1} \) does not exist
- standard errors for parameter estimates \( \hat{\beta} \) can’t be obtained
- model is overparameterized (6 parameters, 5 means)
A general linear model

Models which parameterize the effects of classification factors this way are general linear models. One-way ANOVA and linear regression models are general linear models. The linearity pertains to the parameters, not the explanatory variables.

Here, reparameterizing using $5 - 1$ indicator variables leads to a general linear model. Define $x_1, x_2, x_3, x_4$ as before. Then the MLR model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + E_i \quad i = 1, \ldots, 20$$

where $E_i \overset{iid}{\sim} N(0, \sigma^2)$. The $X$ matrix looks like

$$X = \begin{pmatrix}
1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0
\end{pmatrix}$$
Remarks:

- $(X'X)^{-1}$ exists
- continuous covariates (as opposed to indicators) can be added and it is still a general linear model

For the one-way ANOVA,

$$
\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix}
27.8 \\
0.8 \\
3.6 \\
-20.0 \\
-8.7
\end{pmatrix}
$$

Estimates for the five treatment means obtained by substitution of $\hat{\beta}$ into $\mu(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

$$
\hat{\mu}(1, 0, 0, 0) = \hat{\beta}_0 + \hat{\beta}_1 = 28.6 \\
\hat{\mu}(0, 1, 0, 0) = \hat{\beta}_0 + \hat{\beta}_2 = 31.4 \\
\hat{\mu}(0, 0, 1, 0) = \hat{\beta}_0 + \hat{\beta}_3 = 7.8 \\
\hat{\mu}(0, 0, 0, 1) = \hat{\beta}_0 + \hat{\beta}_4 = 19.1 \\
\hat{\mu}(0, 0, 0, 0) = \hat{\beta}_0 = 27.8
$$

(Compare with page 69.)
For standard errors, use $\hat{\Sigma}$:

$$\hat{\Sigma} = MS[E](X'X)^{-1} = \begin{pmatrix} 2.3 & -2.3 & -2.3 & -2.3 & -2.3 \\ 4.5 & 2.3 & 2.3 & 2.3 \\ 4.5 & 2.3 & 2.3 \\ 4.5 & 2.3 \end{pmatrix}$$

Let $a, b, c, d$ be defined by

$a' = (1, 1, 0, 0, 0), b' = (1, 0, 1, 0, 0), c' = (1, 0, 0, 1, 0), d' = (1, 0, 0, 0, 1)$.

Then

$$\mu(1, 0, 0, 0) = \hat{\beta}_0 + \hat{\beta}_1 = a' \hat{\beta}$$
$$\mu(0, 1, 0, 0) = \hat{\beta}_0 + \hat{\beta}_2 = b' \hat{\beta}$$
$$\mu(0, 0, 1, 0) = \hat{\beta}_0 + \hat{\beta}_3 = c' \hat{\beta}$$
$$\mu(0, 0, 0, 1) = \hat{\beta}_0 + \hat{\beta}_4 = d' \hat{\beta}$$
$$\mu(0, 0, 0, 0) = \hat{\beta}_0 = \hat{\beta}_0$$

and

$$a' \hat{\Sigma} a = b' \hat{\Sigma} b = c' \hat{\Sigma} c = d' \hat{\Sigma} d = \hat{\Sigma}_{11} = 2.3 = \hat{\text{Var}}(\hat{\beta}_0) = \hat{\text{Var}}(\hat{\beta}_0 + \hat{\beta}_j)$$

so the estimated SE for any sample treatment mean is $\sqrt{2.3} = 1.5$.

Recall from one-way ANOVA that

$$\hat{SE}(\bar{y}_{i+}) = \sqrt{\frac{MS[E]}{n_i}} = \sqrt{\frac{9.1}{4}} = \sqrt{2.3} = 1.5$$
A general linear model for 5k times of men AND women  
(Resolution Run, Jan. 1, 2004, Centennial Campus)  
Quadratic model $\mu(x) = \beta_0 + \beta_1 x + \beta_1 x^2$ was used for the association  
between mean pace for male runners and age $x$. Consider modelling  
female race times as well. How could the model be extended to  
incorporate sex differences? Let $x_2 = x^2$ and let an indicator variable  
$x_3$ be defined by  

$$x_3 = \begin{cases} 
1 & \text{female} \\
0 & \text{male} 
\end{cases}$$  

Some candidate models:  

$$\begin{align*} 
\mu(x_1, x_2, x_3) &= \beta_0 \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_3 x_3 \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 
\end{align*}$$  

Resolution Run (5k), 1/1/2004
data race5k;
set race5k;
sexf=(sex="F");
age2=age*age; agef=age*sexf; age2f=age2*sexf;
run;
proc reg data=one;
model pace=;
model pace=sexf; /* equivalent to two-sample t-test */
model pace=age age2;
model pace=sexf age age2;
model pace=sexf age age2 agef age2f;
test agef=0, age2f=0;
run;

The REG Procedure
Model: MODEL1

                 Sum of Mean
Source            DF  Squares  Square  F Value  Pr > F
Model             0   0       .        .         .
Error             159 788.09472 4.95657
Corrected Total   159 788.09472

Root MSE         2.22634  R-Square  0.0000

Parameter Standard
Variable  DF  Estimate  Error  t Value  Pr > |t|
Intercept  1   9.12063  0.17601 51.82  <.0001

Model: MODEL2

                 Sum of Mean
Source            DF  Squares  Square  F Value  Pr > F
Model             1 170.74137 170.74137 43.70  <.0001
Error             158 617.35335 3.90730
Corrected Total   159 788.09472

Root MSE         1.97669  R-Square  0.2167

Parameter Standard
Variable  DF  Estimate  Error  t Value  Pr > |t|
Intercept  1   8.26614  0.20280 40.76  <.0001
sexf      1   2.10335  0.31819   6.61  <.0001
(For MODEL3 output, see linear and quadratic fits from multiple regression notes)

Model: MODEL4

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>113.64500</td>
<td>56.82250</td>
<td>13.23</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>157</td>
<td>674.44972</td>
<td>4.29586</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>159</td>
<td>788.09472</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.07265  R-Square 0.1442

| Variable | DF | Estimate | Error  | t Value | Pr > |t| |
|----------|----|----------|--------|---------|-------|----|
| Intercept| 1  | 11.78503 | 0.70216 | 16.78   | <.0001|
| age      | 1  | -0.19699 | 0.04113 | -4.79   | <.0001|
| age2     | 1  | 0.00294  | 0.00057380 | 5.12  | <.0001|

Model: MODEL5

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>290.34851</td>
<td>96.78284</td>
<td>30.33</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>156</td>
<td>497.74621</td>
<td>3.19068</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>159</td>
<td>788.09472</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 1.78625  R-Square 0.3684

| Variable | DF | Estimate | Error  | t Value | Pr > |t| |
|----------|----|----------|--------|---------|-------|----|
| Intercept| 1  | 10.18317 | 0.64228 | 15.85   | <.0001|
| sexf     | 1  | 2.19792  | 0.29535 | 7.44    | <.0001|
| age      | 1  | -0.17146 | 0.03562 | -4.81   | <.0001|
| age2     | 1  | 0.00281  | 0.00049481 | 5.67  | <.0001|

Fitted model is

\[
\mu(x) = \begin{cases} 
\beta_0 + \beta_1 x + \beta_2 x^2 \\
= 10.18 - 0.17x + 0.0028x^2 \\
\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \\
= 10.18 + 2.20 - 0.17x + 0.0028x^2
\end{cases}
\]

for men

for women
Model: MODEL6

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>293.52828</td>
<td>58.70566</td>
<td>18.28</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>154</td>
<td>494.56644</td>
<td>3.21147</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>159</td>
<td>788.09472</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 1.79206 R-Square 0.3725

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|
| Intercept| 1  | 10.60848           | 0.88641        | 11.97   | <.0001|
| sexf     | 1  | 1.25728            | 1.23237        | 1.02    | 0.3092|
| age      | 1  | -0.19986           | 0.04842        | -4.13   | <.0001|
| age2     | 1  | 0.00321            | 0.00064628     | 4.96    | <.0001|
| agef     | 1  | 0.06882            | 0.07298        | 0.94    | 0.3471|
| age2f    | 1  | -0.00103           | 0.00103        | -0.99   | 0.3217|

\[
\mu(x) = \begin{cases} 
\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3(0) + \beta_4(0) + \beta_5(0) & \text{men} \\
10.61 - 0.20x + 0.0032x^2 & \\
\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3(1) + \beta_4(x) + \beta_5(x^2) & \text{women} \\
\beta_0 + \beta_3 + (\beta_1 + \beta_4)x + (\beta_2 + \beta_5)x^2 & \\
10.61 + 1.25 + (-0.20 + 0.07)x + (0.0032 - 0.0010)x^2 & \\
11.86 - 0.13x + 0.0022x^2 & 
\end{cases}
\]
Which model is “better”? What do we mean by “better?” Is there a test we can use to compare these models?
Comparison of models 5 and 6

reduced: $\mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

full: $\mu(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3$

Extra sum of squares:

$$R(\beta_4, \beta_5|\beta_0, \beta_1, \beta_2) = SS[R]_f - SS[R]_r = 293.5 - 290.3 = 3.0$$

The $F$-ratio

$$F = \frac{R(\beta_4, \beta_5|\beta_0, \beta_1, \beta_2, \beta_3)/(5 - 3)}{MS[E]_f} = \frac{3.2/2}{3.21} = \frac{1.6}{3.21} = 0.5$$

The observed $F$-ratio is not significant on $df = 2, 154$.

In SAS, you could use

```sas
proc reg;
    model pace=age age2 sexf agef age2f;
    test agef=0, age2f=0;
run;
```

to get the following model selection $F$-ratio in the output:

```
The REG Procedure
Model: MODEL6

Test 1 Results for Dependent Variable pace

                      Mean
Source       DF    Square    F Value    Pr > F
Numerator      2    1.58988    0.50    0.6105
Denominator  154    3.21147
```

Which model do we choose at this point?
Nonlinear functions of parameters
Estimate the “peak” running age for men and for women. Is it different? \( \theta_M \) and \( \theta_W \) denote peak running ages for men and women respectively. Using calculus on the model 6 regression,

\[
\theta_M = \frac{-\beta_1}{2\beta_2} \quad \theta_W = \frac{-\beta_1 - \beta_1}{2(\beta_2 + \beta_5)}
\]

These are nonlinear functions of regression parameters. Note that acceptance of any model but 6 implies equality of these peak ages.

\[
\hat{\theta}_W = \begin{cases} 
30.5 & \text{different intercepts model (5)} \\
30.1 & \text{full model (6)} 
\end{cases}
\]

\[
\hat{\theta}_M = \begin{cases} 
30.5 & \text{different intercepts model (5)} \\
31.1 & \text{full model (6)} 
\end{cases}
\]

Things to ponder:

Q: Which of these estimates is “better”?

Q: Which of these estimates are closest to the true peak(s)?

Q: What criterion can we use to assess the estimation?
Analysis of covariance, ANCOVA

*Covariates* are predictive responses. Associations between covariates $z$ and the main response variable of interest $y$ can be used to reduce unexplained variation $\sigma^2$.

**An nutrition example**

A nutrition scientist conducted an experiment to evaluate the effects of four vitamin supplements on the weight gain of laboratory animals. The experiment was conducted in a completely randomized design with $N = 20$ animals randomized to $a = 4$ supplement groups, each with sample size $n = 5$. The response variable of interest is weight gain, but calorie intake $z$ was measured concomitantly.

<table>
<thead>
<tr>
<th>Diet</th>
<th>$y(g)$</th>
<th>Diet</th>
<th>$y$</th>
<th>Diet</th>
<th>$y$</th>
<th>Diet</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>2</td>
<td>65</td>
<td>3</td>
<td>79</td>
<td>4</td>
<td>59</td>
</tr>
<tr>
<td>1</td>
<td>67</td>
<td>2</td>
<td>49</td>
<td>3</td>
<td>52</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>1</td>
<td>78</td>
<td>2</td>
<td>37</td>
<td>3</td>
<td>63</td>
<td>4</td>
<td>59</td>
</tr>
<tr>
<td>1</td>
<td>69</td>
<td>2</td>
<td>75</td>
<td>3</td>
<td>65</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>1</td>
<td>53</td>
<td>2</td>
<td>63</td>
<td>3</td>
<td>67</td>
<td>4</td>
<td>34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\bar{y}_1+ = 63$</th>
<th>$\bar{y}_2+ = 57.8$</th>
<th>$\bar{y}_3+ = 65.2$</th>
<th>$\bar{y}_4+ = 48.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$s_1 = 12.3$</td>
<td>$s_2 = 14.9$</td>
<td>$s_3 = 9.7$</td>
<td>$s_4 = 10.9$</td>
</tr>
</tbody>
</table>

Q: Is there evidence of a vitamin supplement effect?

**The GLM Procedure**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>797.800000</td>
<td>265.933333</td>
<td>1.82</td>
<td>0.1836</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>2334.400000</td>
<td>145.900000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>3132.200000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
But calorie intake $z$ was measured concomitantly:

<table>
<thead>
<tr>
<th>Diet</th>
<th>$y$</th>
<th>$z$</th>
<th>Diet</th>
<th>$y$</th>
<th>$z$</th>
<th>Diet</th>
<th>$y$</th>
<th>$z$</th>
<th>Diet</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>350</td>
<td>2</td>
<td>65</td>
<td>400</td>
<td>3</td>
<td>79</td>
<td>510</td>
<td>4</td>
<td>59</td>
<td>530</td>
</tr>
<tr>
<td>1</td>
<td>67</td>
<td>440</td>
<td>2</td>
<td>49</td>
<td>450</td>
<td>3</td>
<td>52</td>
<td>410</td>
<td>4</td>
<td>50</td>
<td>520</td>
</tr>
<tr>
<td>1</td>
<td>78</td>
<td>440</td>
<td>2</td>
<td>37</td>
<td>370</td>
<td>3</td>
<td>63</td>
<td>470</td>
<td>4</td>
<td>59</td>
<td>520</td>
</tr>
<tr>
<td>1</td>
<td>69</td>
<td>510</td>
<td>2</td>
<td>73</td>
<td>530</td>
<td>3</td>
<td>65</td>
<td>470</td>
<td>4</td>
<td>42</td>
<td>510</td>
</tr>
<tr>
<td>1</td>
<td>53</td>
<td>470</td>
<td>2</td>
<td>63</td>
<td>420</td>
<td>3</td>
<td>67</td>
<td>480</td>
<td>4</td>
<td>34</td>
<td>430</td>
</tr>
</tbody>
</table>

Q: How and why could these new data be incorporated into analysis?
A: ANCOVA can be used to reduce unexplained variation.

Model, given $z_i$,

$$ Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_z z_i + E_i \quad \text{for } i = 1, \ldots, 20 $$

where $x_{ij}$ is an indicator variable for subject $i$ receiving vitamin supplement $j$:

$$ x_{ij} = \begin{cases} 
1 & \text{subject } i \text{ receives supplement } j \\
0 & \text{else}
\end{cases} $$

and errors $E_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Exercise: specify the parametric mean weight gain for the first subject in each treatment group, conditional on their caloric intakes.
Proceeding with MLR analysis of this general linear model:

The GLM Procedure
Class Level Information
Class Levels Values
diet 4 1 2 3 4

Dependent Variable: y

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>1951.680373</td>
<td>487.920093</td>
<td>6.20</td>
<td>0.0038</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>1180.519627</td>
<td>78.701308</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>3132.200000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square Coeff Var Root MSE y Mean
0.623102 15.11308 8.871376 58.70000

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>diet</td>
<td>3</td>
<td>797.800000</td>
<td>265.933333</td>
<td>3.38</td>
<td>0.0463</td>
</tr>
<tr>
<td>z</td>
<td>1</td>
<td>1153.880373</td>
<td>1153.880373</td>
<td>14.66</td>
<td>0.0016</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>diet</td>
<td>3</td>
<td>1537.071659</td>
<td>512.357220</td>
<td>6.51</td>
<td>0.0049</td>
</tr>
<tr>
<td>z</td>
<td>1</td>
<td>1153.880373</td>
<td>1153.880373</td>
<td>14.66</td>
<td>0.0016</td>
</tr>
</tbody>
</table>

To test for a diet effect: \( H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \), use the type III \( F \)-ratio, on 3 and 15 numerator and denominator degrees of freedom. (Note that this is a comparison of nested models.)

Q: Conclusion?

FYI: this model was fit with the following code:

```latex
proc glm;
  class diet;
  model y=diet z;
  means diet;
  lsmeans diet/stderr;
run;
```

NOTE: the drop in \( \sqrt{\text{MSE}} \) (was \( \hat{\sigma} \approx 12g \) is \( \hat{\sigma} \approx 9g \))
Adjusted and unadjusted means

Recall the sample mean weight gains for the four diets (generated by the `means diet;` statement in `proc glm`):

<table>
<thead>
<tr>
<th>Level of diet</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>63.000000</td>
<td>12.2678441</td>
<td>442.000000</td>
<td>58.9067059</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>57.800000</td>
<td>14.8727940</td>
<td>434.000000</td>
<td>61.0737259</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>65.200000</td>
<td>9.6540147</td>
<td>468.000000</td>
<td>36.3318042</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>48.800000</td>
<td>10.8949530</td>
<td>502.000000</td>
<td>40.8656335</td>
</tr>
</tbody>
</table>

These means \( y \) are computed without taking \( z \) into account, so they are called *unadjusted* means.

Unadjusted means do not make any adjustment for the facts that

1. caloric intake may vary by diet (presumably by chance, not because of diet)
2. weight gain depends on caloric intake

*Adjusted* means are estimated mean weight gains at a common reference value (sample mean, \( \bar{z} \)) of the covariate, \( z \).

Here, \( \bar{z} = (442 + 434 + 468 + 502)/4 = 461.5 \). The adjusted means are then just

\[
\bar{y}_{1,a} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_z(461.5) \\
\bar{y}_{2,a} = \hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_z(461.5) \\
\bar{y}_{3,a} = \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_z(461.5) \\
\bar{y}_{4,a} = \hat{\beta}_0 + \hat{\beta}_z(461.5)
\]
To get SAS to report the estimated regression parameter vector $\hat{\beta}$, use the solution option in the model statement. The default parameterization is the one we’ve adopted here where $\beta_0$ is the mean of the last level of the classification treatment factor:

| Parameter | Estimate       | Standard Error | t Value | Pr > |t| |
|----------|----------------|----------------|---------|-------|---|
| Intercept| $-35.66310108$ | $22.41252629$ | -1.59   | 0.1324|
| diet 1   | $24.29519136$  | $6.19932022$  | 3.92    | 0.0014|
| diet 2   | $20.44121688$  | $6.35678835$  | 3.22    | 0.0058|
| diet 3   | $22.12060844$  | $5.80625371$  | 3.81    | 0.0017|
| diet 4   | $0.00000000$   | .              |         |       |
| z        | $0.16825319$   | $0.04394140$  | 3.83    | 0.0016|

NOTE: The X’X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Substitution of $\hat{\beta}$ into the expressions for adjusted means yields

$$\bar{y}_{1,a} = -35.7 + 24.3 + 0.17(461.5) = 66.3$$
$$\bar{y}_{2,a} = -35.7 + 20.4 + 0.17(461.5) = 62.4$$
$$\bar{y}_{3,a} = -35.7 + 22.1 + 0.17(461.5) = 64.1$$
$$\bar{y}_{4,a} = -35.7 + +0.17(461.5) = 42.0$$

Standard errors of $\bar{y}_{j,a}$

Consider $\bar{y}_{2,a}$. What vector $c$ is needed so that $c'\hat{\beta} = \bar{y}_{2,a}$?

What is the standard error of $c'\hat{\beta}$?
To get SAS to produce the adjusted means and estimated standard errors, use an `lsmeans` statement for the factor `diet`

```
To the GLM Procedure
Least Squares Means

| diet | y LSMEAN | Standard Error | Pr > |t|
|------|----------|----------------|------|
| 1    | 66.2809372 | 4.0588750 | <.0001 |
| 2    | 62.4269627 | 4.1473443 | <.0001 |
| 3    | 64.1063543 | 3.9776677 | <.0001 |
| 4    | 41.9857458 | 4.3482563 | <.0001 |
```

Concerns:
Aside from the usual residual-based checks for model adequacy, does treatment affect the covariate? To check this, one could carry out a one-way ANOVA treating $z$ as a response variable and check for a diet effect on the mean of $z$:

```
To the GLM Procedure
Dependent Variable: z

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (diet)</td>
<td>3</td>
<td>14095.00000</td>
<td>4698.33333</td>
<td>1.84</td>
<td>0.1798</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>40760.00000</td>
<td>2547.50000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>54855.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

A: No evidence that treatment affects covariate.

Q: Among the diets, which we’ve concluded are different, what are the differences? (Look at the means, have a guess.)

Q: If you are a lab animal and you want to gain weight, which diet(s) would you choose?

Q: Why are the standard errors for the adjusted means different?
Q: Which adjusted means require the most adjustment?
vitsupp.dat <- read.table("vitsupp.txt",header=TRUE)
vitsupp.dat$z <- 10*vitsupp.dat$z
pdf(file="vitsupp1.pdf")
par(cex=1.2)
attach(vitsupp.dat)
plot(z,y,pch=Diet,main="Vitamin supplement ANCOVA",xlab="z: calories",ylab="y: grams",xlim=c(330,550),ylim=c(30,90))
legend(330,90,legend=c("Diet 1","Diet 2","Diet 3","Diet 4"),pch=1:4,lty=1:4,lwd=2)
vitsupp.fit <- lm(y~as.factor(Diet)+z)
betahat <- coef(vitsupp.fit)
abline(betahat[1],betahat[5],lwd=2)
abline(sum(betahat[1:2]),betahat[5],lwd=2,lty=2)
abline(sum(betahat[c(1,3)]),betahat[5],lwd=2,lty=3)
abline(sum(betahat[c(1,4)]),betahat[5],lwd=2,lty=4)
dev.off()
Lack-of-fit of a SLR model (supplementary to textbook)

Hiking example: completely randomized experiment involving alpine meadows in the White Mountains of New Hampshire. \( N = 20 \) lanes of dimension \( 0.5m \times 1.5m \) randomized to 5 trampling treatments:

<table>
<thead>
<tr>
<th>( i ) : trt group</th>
<th>( x ) : Number of passes</th>
<th>( y_{ij} ) : Height(cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>20.7 15.9 17.8 17.6</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>12.9 13.4 12.7 9.0</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>11.8 12.6 11.4 12.1</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>7.6 9.5 9.9 9.0</td>
</tr>
<tr>
<td>5</td>
<td>500</td>
<td>7.8 9.0 8.5 6.7</td>
</tr>
</tbody>
</table>

Two models for mean plant height:

\[
\text{SLR model : } \mu(x) = \beta_0 + \beta_1 x \\
\text{one-factor ANOVA model : } \mu_{ij} = \mu + \tau_i
\]
proc reg data=one;
  model y=numpass;
run;

proc glm data=one;
  class cnumpass;
  model y=numpass cnumpass;
run;

The SAS System

The REG Procedure
Dependent Variable: y height(cm)
Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>141.29532</td>
<td>141.29532</td>
<td>19.15</td>
<td>0.0004</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>132.79418</td>
<td>7.37745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>274.08950</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.71615 R-Square 0.5155
Dependent Mean 11.79500 Adj R-Sq 0.4886

The GLM Procedure
Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>cnumpass</td>
<td>5</td>
<td>0 25 75 200 500</td>
</tr>
</tbody>
</table>

Dependent Variable: y height(cm)

The REG Procedure
Dependent Variable: y height(cm)
Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>243.1620000</td>
<td>60.7905000</td>
<td>29.48</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>30.9275000</td>
<td>2.0618333</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>274.0895000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square 0.887163 Coeff Var 12.17387 Root MSE 1.435909 Mean 11.79500

The GLM Procedure

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>numpass</td>
<td>1</td>
<td>141.2953228</td>
<td>141.2953228</td>
<td>68.53</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>cnumpass</td>
<td>3</td>
<td>101.8666772</td>
<td>33.9555591</td>
<td>16.47</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>numpass</td>
<td>0</td>
<td>0.0000000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cnumpass</td>
<td>3</td>
<td>101.8666772</td>
<td>33.9555591</td>
<td>16.47</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
When the $t$ treatments have an interval scale, the SLR model, and all polynomials of degree $p \leq t - 2$, are nested in one-factor ANOVA model with $t$ treatment means.

**$F$-ratio for lack-of-fit**

To test for lack-of-fit of a polynomial \textit{(reduced)} model of degree $p$, use extra sum-of-squares $F$-ratio on $t - 1 - p$ and $N - t$ df:

$$F = \frac{SS[\text{lack of fit}]/(t - 1 - p)}{MS[\text{pure error}]}$$

where

$$MS[\text{pure error}] = MS[E]_{\text{full}}$$

and

$$SS[\text{lack-of-fit}] = SS[T_{rt}] - SS[R]_{poly}$$

$$= SS[E]_{poly} - SS[E]_{full}$$

$$= SS[E]_{poly} - SS[\text{pure error}]$$

In a simple linear ($p = 1$) model for the meadows data,

$$SS[\text{lack of fit}] = 243.163 - 141.295 = 101.867 \text{ on } t - 1 - p = 3df$$

and the sum of squares for pure error is $SS[E]_{full} = 30.93$ yielding

$$F = \frac{101.867/3}{30.93/15} \approx \frac{34}{2.1} = 16.5.$$  

(highly significant since $F(0.01, 3, 15) = 5.42.$)

\[
\boxed{\Rightarrow \text{model misspecified: SLR model suffers from lack of fit.}}
\]

Next step: either go with the one-factor ANOVA model or specify some other model, such as quadratic.
ST 512

**Topic:** Completely randomized factorial designs

**Reading** Rao, Ch. 9,13

- This packet
  - Introduction, notation, jargon:
    Terms: factors, levels, treatments, treatment combinations, main effects, interaction effects, crossed factors, nested factors, contrasts, orthogonal contrasts, expected mean squares, multiplicity of comparisons, familywise or experimentwise error rates, power.
  - Specific topics:
    * multiple comparisons,
    * expected mean squares
    * power computations

- Next packet
  - $2 \times 2$ experiments
  - $a \times b$ experiments
  - three-factor ANOVA
  - nested vs. crossed designs
Comparisons (contrasts) among means

Definition: In the one-way ANOVA layout:

\[ Y_{ij} = \mu_i + E_{ij}, i = 1, 2, \ldots, t, \text{ and } j = 1, 2, \ldots, n_i \]

with \( E_{ij} \overset{iid}{\sim} N(0, \sigma^2) \),
a linear function of the group means of the form

\[ \theta = c_1 \mu_1 + c_2 \mu_2 + \cdots + c_t \mu_t \]

is called a [linear combination] of the treatment means.

Definition: The \( c_i \)s are the [coefficients] of the linear combination.

If

\[ c_1 + c_2 + \cdots + c_t = \sum_{j=1}^{t} c_j = 0, \]

the linear combo is called a [contrast].

Definition: Contrasts in which only two of the coefficients are nonzero are called [simple] or [pairwise] contrasts.

Definition: Contrasts in with more than two nonzero coefficients are called [complex] contrasts.

Result: The best estimator for a contrast of interest can be obtained by substituting treatment group sample means \( \bar{y}_{i+} \) for treatment population means \( \mu_i \) in the contrast \( \theta \):

\[ \hat{\theta} = c_1 \bar{Y}_{1+} + c_2 \bar{Y}_{2+} + \cdots + c_t \bar{Y}_{t+}. \]
Example For the binding fraction data, consider the pairwise contrast comparing penicillin (population) mean to Tetracyclin mean:

$$\theta = \mu_1 - \mu_2 = (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 + (0)\mu_4 + (0)\mu_5$$

Using the result, point estimator of $\theta$ is

$$\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{Y}_{1+} - \bar{Y}_{2+}$$

Recall the binding fraction data and ANOVA table:

<table>
<thead>
<tr>
<th>Antibiotic</th>
<th>Binding Percentage</th>
<th>Sample mean</th>
<th>Sample variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penicillin G</td>
<td>29.6 24.3 28.5</td>
<td>28.6</td>
<td>10.4</td>
</tr>
<tr>
<td>Tetracyclin</td>
<td>27.3 32.6 30.8</td>
<td>31.4</td>
<td>10.1</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>5.8 6.2 11</td>
<td>7.8</td>
<td>5.7</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>21.6 17.4 18.3</td>
<td>19.1</td>
<td>3.3</td>
</tr>
<tr>
<td>Chloramphenicol</td>
<td>29.2 32.8 25</td>
<td>27.8</td>
<td>15.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>4</td>
<td>1481</td>
<td>370</td>
<td>41</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>136</td>
<td>9.05</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>1617</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Substitution of $\bar{y}_{1+}$ and $\bar{y}_{2+}$ yields $\hat{\theta} = 28.6 - 31.4 = -2.8$.

Q: How good is this estimate?
Sampling distribution of $\hat{\theta}$

Q: What is the sampling distribution of $\hat{\theta}$?

Q’: That is, what are $E(\hat{\theta})$, $SE(\hat{\theta})$ and shape of distribution of $\hat{\theta}$?

$$\hat{\theta} \sim N(\theta, \text{Var}(\hat{\theta}))$$

Normality follows because $\hat{\theta}$ is a linear function of normal data $Y_{ij}$.

Standard error:

$$SE(\hat{\theta}) = \sqrt{\frac{c_1^2}{n_1}\sigma^2 + \frac{c_2^2}{n_2}\sigma^2 + \cdots + \frac{c_t^2}{n_t}\sigma^2} = \sqrt{\sigma^2 \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}},$$

estimated by

$$\hat{SE}(\hat{\theta}) = \sqrt{\text{MS}[E] \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}}$$

To test $H_0 : \theta = \theta_0$ (often 0) versus $H_1 : \theta \neq \theta_0$ a t use $t$-test:

$$t = \frac{\text{est} - \text{null}}{\hat{SE}} = \frac{\hat{\theta} - \theta_0}{\hat{SE}(\hat{\theta})} \sim t_{N-t}.$$ 

At level $\alpha$, the critical value for this test is $t(N-t, \alpha/2)$.

100$(1 - \alpha)$% confidence interval for a contrast $\theta = \sum c_i \mu_i$ given by

$$\sum c_i \bar{y}_{i+} \pm t(N-t, \alpha/2)\sqrt{\text{MS}[E] \sum \frac{c_i^2}{n_i}}$$
Here,

$$\hat{SE}(\hat{\theta}) = \sqrt{\left(\frac{1^2}{n_1} + \frac{(-1)^2}{n_2}\right)} (9.05) = \sqrt{\frac{9.05}{2}} = 2.127$$

So that the \( t \) statistic becomes

$$\frac{-2.8}{2.127} = -1.32$$

which is not in the critical region, so that the sample mean binding fractions for Penicillin G and Tetracyclin do not differ significantly.

A 95% confidence interval is given by

$$-2.8 \pm 2.13(2.127) \text{ or } (-7.3, 1.7)$$

Code (next page) estimates all pairwise contrasts involving Pen. G:

- \( \theta_1 = a'\mu = (1, -1, 0, 0, 0)\mu \)
- \( \theta_2 = b'\mu =? \)
- \( \theta_3 = c'\mu =? \)
- \( \theta_4 = d'\mu =? \)

along with the complex contrast comparing Pen G. with mean of other four antibiotics:

$$\theta_5 = (, , , , , )\mu$$

Here \( \mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)' \).
proc glm data=one;
  class drug;
  model y=drug/clparm;
  estimate "theta1" drug 1 -1;
  estimate "theta2" drug 1 0 -1;
  estimate "theta3" drug 1 0 0 -1;
  estimate "theta4" drug 1 0 0 0 -1;
  estimate "theta5" drug 4 -1 -1 -1 -1/divisor=4;
run;

The GLM Procedure
Class Level Information

Class     Levels Values
drug          5    1 2 3 4 5

Sum of
Source    DF  Squares   Mean Square  F Value  Pr > F
Model     4 1480.823000 370.205750 40.88   <.0001
Error    15 135.822500  9.054833
Corrected Total 19 1616.645500

R-Square       Coeff Var   Root MSE     y Mean
0.915985      13.12023    3.009125    22.93500

Source    DF  Type III SS   Mean Square  F Value  Pr > F
drug    4 1480.823000 370.205750 40.88   <.0001

Standard
Parameter  Estimate    Error   t Value  Pr > |t|
theta1   -2.7750000  2.12777270 -1.30  0.2118
theta2    20.7750000  2.12777270  9.76 <.0001
theta3    9.5250000  2.12777270  4.48  0.0004
theta4   0.8000000  2.12777270  0.38  0.7122
theta5   7.0812500  1.68215202  4.21  0.0008

Parameter  95% Confidence Limits
theta1      -7.3102402  1.7602402
theta2     16.2397598  25.3102402
theta3     4.9897598  14.0602402
theta4    -3.7352402  5.3352402
theta5     3.4958278  10.6666722
Orthogonal contrasts

In the same way the $SS[TOT]$ can be partitioned into independent components $SS[Trt]$ and $SS[E]$, the sum of squares for treatments, $SS[Trt]$ can be partitioned into $t - 1$ independent components.

Let two contrasts $\theta_1$ and $\theta_2$ be given by

$$\theta_1 = c_1\mu_1 + \cdots + c_t\mu_t \quad \text{and} \quad \theta_2 = d_1\mu_1 + \cdots + d_t\mu_t$$

or

$$\theta_1 = \sum_{i=1}^{t} c_i\mu_i \quad \text{and} \quad \theta_2 = \sum_{i=1}^{t} d_i\mu_i$$

Definition: The two contrasts $\theta_1$ and $\theta_2$ are mutually orthogonal if the products of their coefficients sum to zero:

$$c_1d_1 + \cdots + c_tc_t = \sum_{i=1}^{t} c_id_i = 0$$

Consider several contrasts, say $k$ of them: $\theta_1, \ldots, \theta_k$. The set is mutually orthogonal if all pairs are mutually orthogonal.

Examples:

$(-1, 1, 0, 0, 0)$ and $(0, 0, -1, 1, 0)$ orthogonal?

$(1, -1/2, -1/2, 0, 0)$ and $(0, 0, 0, -1, 1)$ orthogonal?

$(-1, 1, 0, 0, 0)$ and $(0, -1, 1, 0, 0)$ orthogonal?

$\theta_i$ and $\theta_j$ orthogonal $\implies \hat{\theta}_i$ and $\hat{\theta}_j$ are statistically independent.
Sums of squares for contrasts
In the same way $SS[Trt]$ was obtained for a treatment effect, a sum of squares term can be obtained for a contrast:

$$SS[\hat{\theta}_1] = \frac{\hat{\theta}_1^2}{\frac{c_1^2}{n_1} + \cdots + \frac{c_t^2}{n_t}}$$

If $\theta_1, \ldots, \theta_{t-1}$ are $t-1$ mutually orthogonal contrasts, then

$$SS[Trt] = SS(\hat{\theta}_1) + SS(\hat{\theta}_2) + \cdots + SS(\hat{\theta}_{t-1})$$

There is one $df$ associated with a sum of squares for an individual contrast $\hat{\theta}_j$ and if $\theta_j = 0$, then it can be shown that

$$E(SS[\hat{\theta}_j]) = \sigma^2.$$ 

To test $H_0 : \theta_j = 0$ versus $H_1 : \theta_j \neq 0$ use

$$F = \frac{SS[\hat{\theta}_j]}{MS[E]}$$
on 1 numerator degree of freedom and $N - t$ denominator degrees of freedom. For $\theta_1 = \mu_1 - \mu_2$ in the binding fractions,

$$F = \frac{(-2.8)^2}{MS[E] \left( \frac{1}{4} + \frac{(-1)^2}{4} + 0 + 0 + 0 \right)} = 1.73.$$ 

(Using $F(0.05, 1, 15) = 4.54$, is the value $\theta_1 = 0$ plausible?)
A new dataset:

Number of contaminants in IV fluids made by \( t = 3 \) pharmaceutical companies

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments (or pharmacies)</td>
<td>2</td>
<td>113646</td>
<td>56823</td>
<td>5.81</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>146753</td>
<td>9784</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>260400</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consider the following 2 contrasts:

\[
\theta_1 = \mu_M - \mu_A \quad \text{and} \quad \theta_2 = \mu_C - \frac{\mu_M + \mu_A}{2}
\]

Q: Are these contrasts orthogonal?
Q: Are the estimated contrasts \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) independent?
Exercise: Compute \( SS[\hat{\theta}_1] \) and \( SS[\hat{\theta}_2] \). Add em up.
proc format;
   value firmfmt 1="Cutter" 2="Abbott" 3="McGaw";
run;

data one;
   infile "pharmfirm.dat";
   input firm con;
   format firm firmfmt.;
run;

proc glm order=formatted;
   title "contaminant particles in IV fluids";
   class firm;
   model con=firm;
   contrast 'C - avg of M and A' firm -0.5 1 -0.5;
   contrast 'McGaw - Abbott' firm -1 0 1;
   estimate 'C - avg of M and A' firm -0.5 1 -0.5;
   estimate 'McGaw - Abbott' firm -1 0 1;
run;

contaminant particles in IV fluids
1
The GLM Procedure

Class Levels Values
firm 3 Abbott Cutter McGaw

Sum of
Source            DF    Squares Mean Square  F Value  Pr > F
Model             2 113646.3333  56823.1667  5.81  0.0136
Error            15 146753.6667  9783.5778
Corrected Total  17 260400.0000

R-Square      Coeff Var    Root MSE     con Mean
  0.436430       33.91268    98.91197    291.6667

Contrast               DF   Contrast SS  Mean Square  F Value  Pr > F
C - avg of M and A     1  2862.2500  2862.2500        0.29  0.5965
McGaw - Abbott         1 110784.0833 110784.0833      11.32  0.0043

Parameter          Estimate Error   t Value  Pr > |t|
C - avg of M and A  -26.750000  49.4559849 -0.54  0.5965
McGaw - Abbott      192.166667  57.1068524   3.37  0.0043
Multiple Comparisons

• Can’t go carrying out many many tests of significance willy-nilly
• e.g. consider the case with $t = 5$ (antibiotic treatments): all simple (pairwise) contrasts of the form $\theta = \mu_i - \mu_j$

• $\binom{5}{2} = 10$ tests of significance each at level $\alpha = 0.05$

• probability of committing at least one type I error?

Definition: When testing $k$ contrasts, the experimentwise error rate (or familywise) is

$$f_{we} = \Pr(\text{at least one type I error})$$

Methods for simultaneous inference for multiple contrasts include

• Scheffé
• Bonferroni
• Tukey

FDR

When the number of comparisons is in the hundreds or thousands, and FWE control is hopeless, more manageable type I error rate is the False Discovery Rate (FDR):

$$FDR = E\left( \frac{\text{Falsely rejected null hypotheses}}{\text{Number of rejected null hypotheses}} \right)$$

See `qvalue()` in R and

http://www4.stat.ncsu.edu/~jaosborn/research/microarray/software/qvalues.sas
A context in which multiplicity is a big issue: Microarray experiments, which may involve thousands of genes and tests

(Data courtesy of Cassi Myburg)
Bonferroni

Suppose interest lies in exactly \( k \) contrasts. The Bonferroni adjustment to \( \alpha \) which controls \( fwe \) is

\[
\alpha' = \frac{\alpha}{k}
\]

Simultaneous 95\% confidence intervals for the \( k \) contrasts given by

\[
a_1 \bar{Y}_{1+} + a_2 \bar{Y}_{2+} + \cdots + a_t \bar{Y}_{t+} \pm t\left(\frac{\alpha'}{2}, \nu\right) \sqrt{MS[E] \sum a_j^2 n_j} \]

and

\[
b_1 \bar{Y}_{1+} + b_2 \bar{Y}_{2+} + \cdots + b_t \bar{Y}_{t+} \pm t\left(\frac{\alpha'}{2}, \nu\right) \sqrt{MS[E] \sum b_j^2 n_j} \]

\[ \vdots \]

\[
k_1 \bar{Y}_{1+} + k_2 \bar{Y}_{2+} + \cdots + k_t \bar{Y}_{t+} \pm t\left(\frac{\alpha'}{2}, \nu\right) \sqrt{MS[E] \sum k_j^2 n_j} \]

where \( \nu \) denotes \( df \) for error. \( t(\frac{\alpha'}{2}, \nu) \) might have to be obtained using software.

For the binding fraction example, consider only pairwise comparisons with Penicillin:

\[
\theta_1 = \mu_1 - \mu_2, \theta_2 = \mu_1 - \mu_3, \theta_3 = \mu_1 - \mu_4, \theta_4 = \mu_1 - \mu_5
\]

We have \( k = 4, \alpha' = 0.05/k = 0.0125 \), and \( t(\frac{\alpha'}{2}, 15) = 2.84 \).
Substitution leads to
\[ t(\alpha', 15) \sqrt{MS[E] \left( \frac{(-1)^2}{4} + \frac{(-1)^2}{4} + \frac{0^2}{4} + \cdots \frac{0^2}{4} \right)} = 2.84 \sqrt{(9.05)^2} = 6.0 \]
so that simultaneous 95% confidence intervals for \( \theta_1, \theta_2, \theta_3, \theta_4 \) take the form
\[ \bar{y}_1 - \bar{y}_i \pm 6.0 \]

In SAS, an adjustment for \( k = 4 \) can be achieved with care:

```
proc glm data=one;
   title "Bonferroni correction for 4 contrasts";
   class drug;
   model y=drug/clparm alpha=.0125;
   estimate "theta1" drug -1 1;
   estimate "theta2" drug -1 0 1;
   estimate "theta3" drug -1 0 0 1;
   estimate "theta4" drug -1 0 0 0 1;
run;
```

```
Bonferroni correction for 4 contrasts
The GLM Procedure
Class Level Information

Class    Levels    Values
drug      5        1 2 3 4 5

Standard

| Parameter | Estimate | Error   | t Value | Pr > |t| |
|-----------|----------|---------|---------|------|---|
| theta1    | 2.7750000| 2.12777270 | 1.30    | 0.2118 |
| theta2    | -20.7750000| 2.12777270 | -9.76   | <.0001 |
| theta3    | -9.5250000| 2.12777270 | -4.48   | 0.0004 |
| theta4    | -0.8000000| 2.12777270 | -0.38   | 0.7122 |

Parameter 98.75% Confidence Limits
theta1 -3.2606985 8.8106985
theta2 -26.8106985 -14.7393015
theta3 -15.5606985 -3.4893015
theta4 -6.8356985 5.2356985

(actually simultaneous 95% confidence intervals)
Another method: Scheffé

For simultaneous 95% confidence intervals for ALL contrasts, use

$$
\sum_{i=1}^{t} c_i \bar{y}_i \pm \sqrt{(t - 1)(F^*)MS[E] \sum_{i=1}^{t} c_i^2 n_i^n}
$$

where, $F^* = F(\alpha, t - 1, N - t)$. For a pairwise comparisons of means, $\mu_j$ and $\mu_k$, this yields

$$
\bar{y}_{j+} - \bar{y}_{k+} \pm \sqrt{(t - 1)(F^*)MS[E](1/n_j + 1/n_k)}
$$

Using $\alpha = 0.05$, need to specify

- $t$ (from the design)
- $F^*$ (same critical value as for $H_0 : \tau_i \equiv 0$).
- $MS[E]$ (from the data)
- $\bar{y}_{i+}, \bar{y}_{j+}$
- $n_i, n_j$ (from the data)

For binding fraction data,

$$
\sqrt{(t - 1)(F^*)MS[E] \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} = \sqrt{(5 - 1)(3.06)9.05 \left( \frac{1}{4} + \frac{1}{4} \right)} = 7.44
$$

If any two sample means differ by more than 7.44, they differ significantly.

For IV fluids,

$$
\bar{y}_{A} = 204.5, \quad \bar{y}_{M} = 396.67 \quad \bar{y}_{C} = 273.83
$$

and

$$
\sqrt{(t - 1)(F^*)MS[E] \left( \frac{1}{n_j} + \frac{1}{n_k} \right)} = \sqrt{(3 - 1)(3.68)9784(\frac{1}{6} + \frac{1}{6})} = 154.9
$$

conclusion about pairwise contrasts? (compare w/ Bonferroni)
Tukey

Tukey’s method is better than Scheffé’s method when making all pairwise comparisons in balanced designs \( n = n_1 = n_2 = \cdots = n_t \). It is conservative, controlling the experimentwise error rate, and has a lower type II error rate in these cases than Scheffé. (It is more powerful.)

For simple contrasts of the form

\[
\theta = \mu_j - \mu_k
\]

to test

\[
H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0
\]

reject \( H_0 \) at level \( \alpha \) if

\[
|\hat{\theta}| > q(t, N - t, \alpha) \sqrt{\frac{MS[E]}{n}}
\]

where \( q(t, N - t, \alpha) \) denotes \( \alpha \) level studentized range for \( t \) means and \( N - t \) degrees of freedom. These studentized ranges can be found in Table C.11 of Rao.

For the IV data, \( q(3, 15, 0.05) = 3.67 \). Tukey’s 95% honestly significant difference (HSD) for pairwise comparisons of treatment means in this balanced design are

\[
3.67 \sqrt{\frac{9784}{6}} = 148.3
\]
proc glm;
   class firm;
   model con=firm;
   means firm/scheffe tukey;
run;

Tukey’s Studentized Range (HSD) Test for con

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha 0.05
Error Degrees of Freedom 15
Error Mean Square 9783.578
Critical Value of Studentized Range 3.67338
Minimum Significant Difference 148.33

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>Tukey Grouping</th>
<th>Mean</th>
<th>N</th>
<th>firm</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>396.67</td>
<td>6</td>
<td>McGaw</td>
</tr>
<tr>
<td>B A</td>
<td>273.83</td>
<td>6</td>
<td>Cutter</td>
</tr>
<tr>
<td>B</td>
<td>204.50</td>
<td>6</td>
<td>Abbott</td>
</tr>
</tbody>
</table>

The GLM Procedure

Scheffe’s Test for con

NOTE: This test controls the Type I experimentwise error rate.

Alpha 0.05
Error Degrees of Freedom 15
Error Mean Square 9783.578
Critical Value of F 3.68232
Minimum Significant Difference 154.98

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>Scheffe Grouping</th>
<th>Mean</th>
<th>N</th>
<th>firm</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>396.67</td>
<td>6</td>
<td>McGaw</td>
</tr>
<tr>
<td>B A</td>
<td>273.83</td>
<td>6</td>
<td>Cutter</td>
</tr>
<tr>
<td>B</td>
<td>204.50</td>
<td>6</td>
<td>Abbott</td>
</tr>
</tbody>
</table>
Expected mean squares

Definition: The treatment mean square is given by

\[ MS[Trt] = \frac{SS[Trt]}{t-1} = \frac{1}{t-1} \sum_i \sum_j (\bar{y}_{ij} - \bar{y}_{++})^2 \]

\[
\left(\bar{y}_{++} = \bar{y}_{++} \quad \bar{y}_{i+} = \frac{1}{n_i} \sum_{i=1}^{n_i} y_{ij}\right)
\]

Q: Why are there \( t - 1 \) degrees of freedom associated with \( MS[Trt] \)?

A: Note that the terms in the \( \Sigma \Sigma \) above do not depend on \( j \), so it is a sum of squares from \( t \) independent sample treatment means, leaving \( t - 1 \) df for assessing variability.

It can be shown that

\[
E[MS[Trt]; H_1] = E[SS[Trt]/(t - 1); H_1]
\]

\[
= \sigma^2 + \frac{1}{t-1} \sum n_i(\mu_i - \mu)^2
\]

\[
= \sigma^2 + n\frac{1}{t-1} \sum (\mu_i - \mu)^2 \quad \text{(balanced case)}
\]

\[
= \sigma^2 + n\psi_T^2
\]

where

\[
\psi_T^2 = \frac{1}{t-1} \sum (\mu_i - \mu)^2.
\]

Note that under \( H_0 : \mu_i \equiv \mu \) and \( \psi_T^2 = 0 \) so that

\[
E[MS[Trt]; H_0] = E[SS[Trt]/(t - 1); H_0]
\]

\[
= \sigma^2
\]
Definition: The error mean square is given by

\[ MS[E] = \frac{SS[E]}{N - t} \]

This is just a generalization of the pooled variance \( S_p^2 \) to the case of more than \( t = 2 \) groups:

\[
MS[E] = \frac{SS[E]}{N - t} = \frac{1}{N - t} \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \\
= \frac{1}{N - t} \sum_{i=1}^{t} (n_i - 1) s_i^2 \\
= \left( \frac{n_1 - 1}{N - t} \right) s_1^2 + \left( \frac{n_2 - 1}{N - t} \right) s_2^2 + \cdots + \left( \frac{n_t - 1}{N - t} \right) s_t^2 \\
= "S_p^2"
\]

Since \( E(S_i^2) = \sigma^2 \), \( MS[E] \) is unbiased for \( \sigma^2 \) regardless of \( H_0 \) or \( H_1 \):

\[
E(S_i^2) = \sigma^2 \\
\implies E[MS[E]] = \left( \frac{n_1 - 1}{N - t} \right) \sigma^2 + \left( \frac{n_2 - 1}{N - t} \right) \sigma^2 + \cdots + \left( \frac{n_t - 1}{N - t} \right) \sigma^2 \\
= \sigma^2 \\
\therefore E[MS[E]] = \sigma^2
\]
Sample size computations for one-way ANOVA

Now consider the null hypothesis in a balanced experiment using one-way ANOVA to compare $t$ treatment means and $\alpha = 0.05$:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_t = \mu$$

versus the alternative

$$H_1: \mu_i \neq \mu_j \text{ for some } i \neq j$$

Q: Suppose that we intend to use a balanced design. How big does our sample size $n_1 = n_2 = \ldots = n_t = n$ need to be?

Of course, the answer depends on lots of things, namely, $\sigma^2$ and how many treatment groups $t$ we have and how much of a difference among the means we hope to be able to detect, and with how big a probability.

Given $\alpha$, $\mu_1, \ldots, \mu_t$, and $\sigma^2$, we can choose $n$ to ensure a power of at least $\beta$ using the noncentral $F$ distribution.

Recall that the critical region for the statistic $F = MS[T]/MS[E]$ is everything bigger than $F(\alpha, t - 1, N - t) = F^*$. The power of the $F$-test conducted using $\alpha = 0.05$ to reject $H_0$ under this alternative is given by

$$1 - \beta = \Pr(MS[T]/MS[E] > F^*; H_1 \text{ is true}).$$

(1)
Let $\tau_i = \mu_i - \mu$ for each treatment $i$ so that

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

When some $H_1$ is true and the sample size $n$ is used in each group, it can be shown that the $F$ ratio has the noncentral $F$ distribution with noncentrality parameter

$$\gamma = \sum_{j=1}^{t} n_j \left( \frac{\tau_j}{\sigma} \right)^2 = n \sum_{j=1}^{t} \left( \frac{\tau_j}{\sigma} \right)^2$$

This is the parameterization for the $F$ distribution used in both SAS and S+.

One way to obtain an adequate sample size is trial and error. Software packages can be used to get probabilities of the form (1) for various values of $n$. Russ Lenth’s website is also terrific and helpful:

http://www.stat.uiowa.edu/~rlenth/Power/

We’ll write SAS code to do some computations but one could also use the procedure PROC POWER or other software such as R.
An example: suppose that a balanced completely randomized design (CRD) is to be used to test for a difference in the number of contaminant particles in IV fluid for three pharmaceutical companies. It is believed that the standard deviation on a given observation is about 100 particles for each company. In order to test \( H_0 : \mu_1 = \mu_2 = \mu_3 \) at level \( \alpha = 0.05 \), how large does the common sample size, \( n \), need to be?

Q: “What alternative to \( H_0 \) would be meaningful? What is \( \sigma \)?”

A: The alternative \( H_1 : \mu_1 = \mu_2 = (\mu - 30) = 230, \mu_3 = \mu + 60 = 320 \) would be meaningful. Assume \( \sigma \approx 100 \).

Q: “What is an acceptable type II error rate, or what kind of power are we looking for?”

A: Suppose that \( 1 - \beta = 0.8 \) should be good enough.

To obtain probabilities of the form (1) we need the noncentrality parameter \( \gamma \):
\[
\gamma = n \left[ \left( \frac{T_1}{\sigma} \right)^2 + \left( \frac{T_2}{\sigma} \right)^2 + \left( \frac{T_3}{\sigma} \right)^2 \right] = n \left[ -30^2 + -30^2 + (60)^2 \right] / 100^2 = 0.54n
\]
The \( \alpha = 0.05 \) critical value for \( H_0 \) is given by
\[
F^* = F(3 - 1, 3(n - 1), 0.05).
\]
We need the area to the right of \( F^* \) for the noncentral \( F \) distribution with degrees of freedom 2 and \( 3(n - 1) \) and noncentrality parameter \( \gamma = 0.54n \). The following printout suggests the sufficiency of \( n = 19 \) for power of \( 1 - \beta = 0.8 \).
data one;
  do n=3 to 25; output; end;
run;

data one;
  set one;
  t=3;
  nu1=t-1;
  nu2=t*(n-1);
  sumtau2=(-30)**2 + (-30)**2 + 60**2;
  sigma2=10000;
  *sigma2u=(2/3)*var(100,100,190);
  *ncp=t*sigma2u/(2*sigma2);
  ncp=n*sumtau2/sigma2;
  qf=finv(0.95,nu1,nu2);
  pf=probf(qf,nu1,nu2,ncp);
  power=1-pf;
run;

proc print;run;

OBS  N  T  NU1  NU2  SUMTAU2  SIGMA2  SIGMA2U  NCP     QF     PF     POWER
     1  3  3  2   6  5400   10000   1800  1.62  5.14325  0.86663  0.13337
     2  4  3  2   9  5400   10000   1800  2.16  4.25649  0.81604  0.18396
     3  5  3  2  12  5400   10000   1800  2.70  3.88529  0.76402  0.23598
     4  6  3  2  15  5400   10000   1800  3.24  3.68232  0.71152  0.28848
     5  7  3  2  18  5400   10000   1800  3.78  3.55456  0.65935  0.34065
     6  8  3  2  21  5400   10000   1800  4.32  3.46680  0.60817  0.39183
     7  9  3  2  24  5400   10000   1800  4.86  3.40283  0.55853  0.44147
     8 10  3  2  27  5400   10000   1800  5.40  3.35413  0.51085  0.49115
     9 11  3  2  30  5400   10000   1800  5.94  3.31583  0.46546  0.53454
    10 12  3  2  33  5400   10000   1800  6.48  3.28492  0.42257  0.57743
    11 13  3  2  36  5400   10000   1800  7.02  3.25945  0.38233  0.61767
    12 14  3  2  39  5400   10000   1800  7.56  3.23810  0.34481  0.65519
    13 15  3  2  42  5400   10000   1800  8.10  3.21994  0.31002  0.68998
    14 16  3  2  45  5400   10000   1800  8.64  3.20432  0.27795  0.72205
    15 17  3  2  48  5400   10000   1800  9.18  3.19073  0.24850  0.75150
    16 18  3  2  51  5400   10000   1800  9.72  3.17880  0.22160  0.77840
    17 19  3  2  54  5400   10000   1800 10.26  3.16825  0.19712  0.80288
    18 20  3  2  57  5400   10000   1800 10.80  3.15884  0.17493  0.82507
    19 21  3  2  60  5400   10000   1800 11.34  3.15041  0.15489  0.84511
    20 22  3  2  63  5400   10000   1800 11.88  3.14281  0.13684  0.86316
    21 23  3  2  66  5400   10000   1800 12.42  3.13592  0.12065  0.87935
    22 24  3  2  69  5400   10000   1800 12.96  3.12964  0.10616  0.89384
    23 25  3  2  72  5400   10000   1800 13.50  3.12391  0.09324  0.90676
Another example: suppose we want to test equal mean binding fractions among antibiotics against the alternative

\[ H_1 : \mu_P = \mu + 3, \mu_T = \mu + 3, \mu_S = \mu - 6, \mu_E = \mu, \mu_C = \mu \]

so that

\[ \tau_1 = 3, \tau_2 = 3, \tau_3 = -6, \tau_4 = \tau_5 = 0. \]

Assume \( \sigma = 3 \) and we need to use \( \alpha = \beta = 0.05 \).

The noncentrality parameter is given by

\[ \gamma = n\left(\frac{3}{3} \right)^2 + \left(\frac{3}{3} \right)^2 + \left(\frac{-6}{3} \right)^2 \].

The following code should do the trick

```
data one;
  do n=2 to 10; output; end;
run;
data one;
  set one;
  t=5; nu1=t-1; nu2=t*(n-1);
  sumtau2=3**2+3**2+(-6)**2;
  sigma2=9;
  ncp=n*sumtau2/sigma2;
  qf=finv(0.95,nu1,nu2);
  pf=probf(qf,nu1,nu2,ncp);
  power=1-pf;
run;
proc print;run;
```

<table>
<thead>
<tr>
<th>OBS</th>
<th>N</th>
<th>T</th>
<th>NU1</th>
<th>NU2</th>
<th>SUMTAU2</th>
<th>SIGMA2</th>
<th>NCP</th>
<th>QF</th>
<th>PF</th>
<th>POWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>54</td>
<td>9</td>
<td>12</td>
<td>5.19217</td>
<td>0.59246</td>
<td>0.40754</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>10</td>
<td>54</td>
<td>9</td>
<td>18</td>
<td>3.47805</td>
<td>0.22465</td>
<td>0.77535</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>15</td>
<td>54</td>
<td>9</td>
<td>24</td>
<td>3.05557</td>
<td>0.06437</td>
<td>0.93563</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>20</td>
<td>54</td>
<td>9</td>
<td>30</td>
<td>2.86608</td>
<td>0.01533</td>
<td>0.98467</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>25</td>
<td>54</td>
<td>9</td>
<td>36</td>
<td>2.75871</td>
<td>0.00319</td>
<td>0.99681</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>30</td>
<td>54</td>
<td>9</td>
<td>42</td>
<td>2.68963</td>
<td>0.00060</td>
<td>0.99940</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>35</td>
<td>54</td>
<td>9</td>
<td>48</td>
<td>2.64147</td>
<td>0.00010</td>
<td>0.99990</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>40</td>
<td>54</td>
<td>9</td>
<td>54</td>
<td>2.60597</td>
<td>0.00002</td>
<td>0.99998</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>45</td>
<td>54</td>
<td>9</td>
<td>60</td>
<td>2.57874</td>
<td>0.00000</td>
<td>1.00000</td>
</tr>
</tbody>
</table>
Orthogonal polynomial contrasts

Example: poultry science experiment measures bodyweights of chickens from \( a = 4 \) diet groups, characterized by protein concentration in diet.

- \( Y \): 21-day bodyweights of chickens
- completely randomized design with one factor, protein in diet, with four \textit{equally spaced} levels.
- thanks to P. Plumstead for data.
- \( n = 18 \) pens, \( N = 72 \).

<table>
<thead>
<tr>
<th>diet group</th>
<th>( x ): level of protein</th>
<th>( \bar{y}_{i+} )</th>
<th>( s_i )</th>
<th>Tukey grouping</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.8</td>
<td>993</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>23.5</td>
<td>1003</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>25.2</td>
<td>1022</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>26.9</td>
<td>1050</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

One-way ANOVA table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>34311.7666</td>
<td>11437.2555</td>
<td>9.57</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>68</td>
<td>81279.4678</td>
<td>1195.2863</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>71</td>
<td>115591.2344</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square Coeff Var Root MSE AMBW21D Mean
0.296837 3.399254 34.57291 1017.073
Some omitted exam questions:

1. Sketch a plot of mean bodyweight at 21 days against protein content.

2. Consider the following three sample contrasts

   \[ \hat{\theta}_1 = -3\bar{y}_1 + \bar{y}_2 + \bar{y}_3 + 3\bar{y}_4 \]
   \[ \hat{\theta}_2 = \bar{y}_1 + \bar{y}_2 - \bar{y}_3 + \bar{y}_4 \]
   \[ \hat{\theta}_3 = -\bar{y}_1 + 3\bar{y}_2 + 3\bar{y}_3 - \bar{y}_4 \]

   (a) True/false: These estimated contrasts are orthogonal.
   (b) True/false: If contrast sums of squares are obtained, then
   \[ SS[\hat{\theta}_1] + SS[\hat{\theta}_2] + SS[\hat{\theta}_3] = SS[Trt]? \]
   (c) Report \( \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3. \)
   (d) Provide an expression for the standard error of \( \hat{\theta}_2. \)
   (e) Estimate the standard error of \( \hat{\theta}_2. \)
   (f) Report \( SS(\hat{\theta}_1). \)

3. Fitting the SLR model leads to \( SS[Reg] = 32742 \) and \( SS[Tot] = 115591. \)

   (a) Report the \( F\)-ratio for testing for a lack-of-fit of the linear model.
   (b) The appropriate critical value for a test with level \( \alpha = 0.05 \) is \( F^* = 3.13. \) Draw a conclusion about the adequacy of the linear model using \( \alpha = 0.05: \) is there evidence that the linear model is inadequate?
The contrasts in problem 2 are called orthogonal polynomial contrasts. The table below gives coefficients for orthogonal polynomial contrasts for balanced single-factor experiments with 3, 4, or 5 equally spaced levels.

<table>
<thead>
<tr>
<th>Factor levels</th>
<th>Poly. Degree</th>
<th>Coefficients for $\bar{y}<em>{1+}$ $\bar{y}</em>{2+}$ $\bar{y}<em>{3+}$ $\bar{y}</em>{4+}$ $\bar{y}_{5+}$</th>
<th>$SS(\hat{\theta}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>$\hat{\theta}_1$ -1 0 1</td>
<td>$R(\beta_1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$\hat{\theta}_2$ 1 -2 1</td>
<td>$R(\beta_2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>$\theta_1$ -3 -1 1 3</td>
<td>$R(\beta_1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$\hat{\theta}_2$ 1 -1 -1 1</td>
<td>$R(\beta_2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>$\hat{\theta}_3$ -1 3 -3 1</td>
<td>$R(\beta_3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>$\theta_1$ -2 -1 0 1 2</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$\hat{\theta}_2$ 2 -1 -2 -1 2</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>$\hat{\theta}_3$ -1 2 0 -2 1</td>
<td>?</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$\hat{\theta}_4$ 1 -4 6 -4 1</td>
<td>?</td>
</tr>
</tbody>
</table>

Rightmost column indicates extra SS in MLR of the form

$$\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots.$$ 

The contrast corresponding to a polynomial of degree $p$ can be used to test for a $p^{th}$ degree association:

- large $|\hat{\theta}_1|$ indicates linear association between $y$ and $x$.
- large $|\hat{\theta}_2|$ indicates quadratic association between $y$ and $x$.
- large $|\hat{\theta}_3|$ indicates cubic association between $y$ and $x$.

This is computationally (and otherwise) easier than fitting polynomial regressions of various degrees.
Proc glm;
  title "protein concentration and chicken weights";
  class cp;
  MODEL AMBW21D=cp;
  contrast 'cp linear' cp -3 -1 1 3;
  contrast 'CP quadratic' CP 1 -1 -1 1;
  contrast 'CP cubic' CP -1 3 -3 1;
  contrast 'all three' CP -3 -1 1 3,
    cp 1 -1 -1 1,
    cp -1 3 -3 1;
  /* 'all three' tests that the 3-vector of contrasts is (0,0,0)' */
  estimate 'cp linear' cp -3 -1 1 3;
  estimate 'CP quadratic' CP 1 -1 -1 1;
  estimate 'CP cubic' CP -1 3 -3 1;
RUN;
proc glm; /* no class statement will fit regression model*/
  model ambw21d=cp cp*cp cp*cp*cp;
run;
proc glm;
  model ambw21d=cp;
run;

protein concentration and chicken weights

The GLM Procedure

Class Levels Values
CP 4 21.8 23.5 25.2 26.9

Sum of
Source DF Squares Mean Square F Value Pr > F
Model 3 34311.76658 11437.25553 9.57 <.0001
Error 68 81279.4678 1195.2863
Corrected Total 71 115591.2344

Contrast DF Contrast SS Mean Square F Value Pr > F
  cp linear 1 32741.55648 32741.55648 27.39 <.0001
  CP quadratic 1 1568.66674 1568.66674 1.31 0.2560
  CP cubic 1 1.54337 1.54337 0.00 0.9714
  all three 3 34311.76658 11437.25553 9.57 <.0001

Standard
Parameter Estimate Error t Value Pr > |t|
  cp linear 190.734127 36.4430498 5.23 <.0001
  CP quadratic 18.670635 16.2978273 1.15 0.2560
  CP cubic 1.309524 36.4430498 0.04 0.9714
protein concentration and chicken weights 3

The GLM Procedure

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>34311.7666</td>
<td>11437.2555</td>
<td>9.57</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>68</td>
<td>81279.4678</td>
<td>1195.2863</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>71</td>
<td>115591.2344</td>
<td>115591.2344</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP</td>
<td>1</td>
<td>32741.5565</td>
<td>32741.5565</td>
<td>27.39</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>CP*CP</td>
<td>1</td>
<td>1568.6667</td>
<td>1568.6667</td>
<td>1.31</td>
<td>0.2560</td>
</tr>
<tr>
<td>CP<em>CP</em>CP</td>
<td>1</td>
<td>1.54337</td>
<td>1.54337</td>
<td>0.00</td>
<td>0.9714</td>
</tr>
</tbody>
</table>

| Parameter  | Estimate | Error    | t Value | Pr > |t| |
|------------|----------|----------|---------|------|---|
| Intercept  | 1060.706 | 17690.23 | 0.06    | 0.9524|
| CP         | 11.320   | 2192.81  | 0.01    | 0.9959|
| CP*CP      | -1.630   | 90.321   | -0.02   | 0.9857|
| CP*CP*CP   | 0.044    | 1.236    | 0.04    | 0.9714|

protein concentration and chicken weights 6

The GLM Procedure

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>32741.556</td>
<td>32741.556</td>
<td>27.66</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>70</td>
<td>82849.6779</td>
<td>1183.5667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>71</td>
<td>115591.2344</td>
<td>115591.2344</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Parameter  | Estimate | Error    | t Value | Pr > |t| |
|------------|----------|----------|---------|------|---|
| Intercept  | 743.8748 | 52.1007  | 14.28   | <.0001|
| CP         | 11.2196 | 2.1331   | 5.26    | <.0001|

Note MSE. Linear regression on $x$ preferred to one-factor model for Plumstead’s data. Multiple comparisons among treatment means might be unnecessary.
**Topic:** Multi-factor ANOVA  
**Reading:** Rao, Ch. 13

- $2 \times 2$ experiments  
- $a \times b$ experiments  
- three-factor ANOVA  
- nested vs. crossed designs (not described in packet)

An example of a $2 \times 2$ study

Cholesterol measurements for random samples of $n_j \equiv 7$ people from four populations are given in the table below. The groups (cohorts) are defined as follows:

I  The population of women younger than 50  
II  The population of men younger than 50  
III The population of women 50 years or older  
IV The population of men 50 years or older

<table>
<thead>
<tr>
<th>Group</th>
<th>Cholesterol level</th>
<th>avg</th>
<th>std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>221 213 202 183 185 197 162</td>
<td>$\bar{y}_I = 194.7$</td>
<td>$s = 20$</td>
</tr>
<tr>
<td>II</td>
<td>271 192 189 209 227 236 142</td>
<td>$\bar{y}_{II} = 209.4$</td>
<td>$s = 41$</td>
</tr>
<tr>
<td>III</td>
<td>262 193 224 201 161 178 265</td>
<td>$\bar{y}_{III} = 212.0$</td>
<td>$s = 40$</td>
</tr>
<tr>
<td>IV</td>
<td>192 253 248 278 232 267 289</td>
<td>$\bar{y}_{IV} = 251.3$</td>
<td>$s = 32$</td>
</tr>
</tbody>
</table>
One-way ANOVA Model:

\[ Y_{ij} = \mu_i + E_{ij} \]
\[ = \mu + \tau_i + E_{ij} \]

\( i = 1, 2, 3, 4 \quad j = 1, 2, \ldots, 7 \) and \( E_{ij} \) i.i.d. \( N(0, \sigma^2) \)

Parameters: \( \mu, \tau_1, \tau_2, \tau_3, \tau_4, \sigma^2 \), with \( \sum_1^4 \tau_i = 0 \) constraint.

One-way ANOVA table:

The GLM Procedure

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>cohort</td>
<td>4</td>
<td>I II III IV</td>
</tr>
</tbody>
</table>

Number of observations 28

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>12280.85714</td>
<td>4093.61905</td>
<td>3.46</td>
<td>0.0323</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>28434.57143</td>
<td>1184.77381</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>27</td>
<td>40715.42857</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square 0.301627  Coeff Var 15.87245  Root MSE 34.42054  y Mean 216.8571

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>cohort</td>
<td>3</td>
<td>12280.85714</td>
<td>4093.61905</td>
<td>3.46</td>
<td>0.0323</td>
</tr>
</tbody>
</table>

Conclusion so far is the cohort means, \( \mu_i \) or \( \mu + \tau_i \), are not plausibly equal (using \( \alpha = 0.05 \)).
Some terminology

**Definition:** A [factor](#) in an experiment or study is a variable whose effect on the response is of primary interest. The values that a factor takes in the experiment are called factor [levels](#) or [treatments](#).

**Definition:** In [completely randomized designs](#) experimental units are randomly assigned to factor levels, or treatment groups.

**Note:** The cholesterol study is NOT a completely randomized design, as randomization of subjects to different levels of AGE and GENDER isn’t possible.

**Definition:** When the same number of units are used for each treatment, the design is [balanced](#).

In one-way analysis of cholesterol data, COHORT is the only factor. This factor can be broken down into two factors in a two-way analysis: AGE (factor A) and GENDER (factor B).

**Definition:** If there are observations at all combinations of all factors, the design is [complete](#), otherwise it is [incomplete](#).
Exercise

1. Estimate the mean difference in cholesterol between young men and young women.

2. Estimate the mean difference between old men and old women.

3. Estimate the mean difference between men and women.

4. Estimate the mean difference between older and younger folks.

5. Estimate the mean difference between the differences estimated in 1. and 2.

6. Provide standard errors for all of these estimated contrasts.

7. Specify the vectors defining these contrasts. For example, the first contrast of cohort means can be written

\[
\theta_1 = (-1, 1, 0, 0)' \left( \begin{array}{c} 
\mu_1 \\
\mu_2 \\
\mu_3 \\
\mu_4 
\end{array} \right) = \mu_2 - \mu_1
\]
Consider the following contrasts of the cohort cholesterol means in the population:

\[ \theta_3 = (-1, 1, -1, 1)' \mu \]
\[ \theta_4 = (-1, -1, 1, 1)' \mu \]
\[ \theta_5 = (-1, 1, 1, -1)' \mu \]

Q: Are these contrasts orthogonal?

Q: True/False: \( SS(\hat{\theta}_3) + SS(\hat{\theta}_4) + SS(\hat{\theta}_5) = SS[Trt] \)

Another exercise:

1. Compute the sums of squares for the estimated contrasts in 3., 4. and 5. using the exercise just completed and the fact that if \( \hat{\theta} = \sum c_i \bar{y}_{i+} \) then

\[ SS[\hat{\theta}] = \frac{\hat{\theta}^2}{\sum \frac{c_i^2}{n_i}}. \]

2. Formulate a test of \( H_0 : \theta_i = 0 \) for each of these three contrasts. Obtain the \( F \)-ratio for each of these tests.

3. Obtain the \( \alpha = 0.05 \) critical region for each test. Compare the observed \( F \)-ratios to critical value and draw conclusions about

(a) an age effect
(b) a gender effect
(c) an age \( \times \) gender interaction
Types of effects

Two-way ANOVA model for the cholesterol measurements:

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk} \]

\( i = 1, 2 = a \) and \( j = 1, 2 = b \) and \( k = 1, 2, \ldots, 7 = n. \)

\( E_{ijk} \overset{iid}{\sim} N(0, \sigma^2). \) Parameter constraints: \( \sum_i \alpha_i = \sum_j \beta_j = 0 \) and \( \sum_i (\alpha \beta)_{ij} = 0 \) for each \( j \) and \( \sum_j (\alpha \beta)_{ij} = 0 \) for each \( i. \)

Factor A: AGE has \( a = 2 \) levels - \( A_1 \) : younger and \( A_2 \) : older

Factor B: GENDER has \( b = 2 \) levels - \( B_1 \) : female and \( B_2 \) : male

Three kinds of effects in \( 2 \times 2 \) designs:

1. **Simple** effects are simple contrasts.
   - \( \mu(A_1B) = \mu_{II} - \mu_I \) - simple effect of gender for young folks.
   - \( \mu(AB_1) = \mu_{III} - \mu_I \) - simple effect of age for women

2. **Interaction** effects are differences of simple effects:
   \[ \mu(AB) = \mu(AB_2) - \mu(AB_1) = (\mu_{IV} - \mu_{II} - (\mu_{III} - \mu_I)) \]
   - difference between simple age effects for men and women
   - difference between simple gender effects for old and young folks
   - interaction effect of AGE and GENDER.

3. **Main** effects are averages or sums of simple effects
   \[ \mu(A) = \frac{1}{2} (\mu(AB_1) + \mu(AB_2)) \]
   \[ \mu(B) = \frac{1}{2} (\mu(A_1B) + \mu(A_2B)) \]

Exercise: Classify the contrasts in the last exercise as simple, interaction or main effects.
Partitioning the treatment $SS$ into $t - 1$ orthogonal components


- $(a - 1)(b - 1)$ df for $AB$ interaction
- $(a - 1)$ df for main effect of $A$
- $(b - 1)$ df for main effect of $B$

$F$ test for interaction effect

To test for interaction,

$$H_0: (\alpha \beta)_{11} = (\alpha \beta)_{12} = (\alpha \beta)_{21} = (\alpha \beta)_{22} = 0$$

vs.

$$H_1: (\alpha \beta)_{ij} \neq 0 \text{ for some } i, j$$

use $\theta_5 = \mu(AB)$ and

$$F = \frac{SS(\hat{\theta})/((a - 1)(b - 1))}{MS[E]}$$

on 1 and $28 - 4 = 24$ numerator, denominator df. For cholesterol data, the estimated interaction effect is

$$\hat{\theta}_5 = \hat{\mu}(AB) = (251.3 - 209.4) - (212 - 194.7) = 41.9 - 17.3 = 24.6$$

the associated sum of squares is

$$SS(\hat{\theta}_5) = \frac{(24.6)^2}{\frac{1}{7} + \frac{(-1)^2}{7} + \frac{(-1)^2}{7} + \frac{1}{7}} = \frac{(24.6)^2}{\frac{4}{7}} = 1056$$

and

$$F = \frac{1056}{1185} = 0.9$$

which isn’t significant at $\alpha = 0.05$ on 1,24 df.
To test for main effects of A: AGE

\( H_0 : \alpha_1 = \alpha_2 = 0 \) vs. \( H_1 : \alpha_1 \neq 0 \) or \( \alpha_2 \neq 0 \)

use \( \theta_4 = \mu(A) \) and

\[ F = \frac{SS(\hat{\theta}_4)}{MS[E]} \]

on 1, 24 \( df \). The estimated main effect of AGE is

\[ \hat{\mu}(A) = \frac{(251.3 - 209.4)}{2} + \frac{(212 - 194.7)}{2} = \frac{59.2}{2} = 29.6 \]

the associated sum of squares is

\[ SS(\hat{\theta}_4) = \frac{(29.6)^2}{(\frac{1}{2})^2 + (\frac{1}{2})^2 + (\frac{1}{2})^2 + (\frac{1}{2})^2} = \frac{(29.6)^2}{\frac{1}{7}} = 6121 \]

and

\[ F = \frac{6121}{1185} = 5.2 \]

since \( F(0.05, 1, 24) = 4.26 \) AGE effect significant at \( \alpha = 0.05 \).

Similarly for the main effect of B: gender

\( H_0 : \beta_1 = \beta_2 = 0 \) vs. \( H_1 : \beta_1 \neq 0 \) or \( \beta_2 \neq 0 \)

use \( \theta_3 = \mu(B) \) on 1 and 24 \( df \)

\[ \hat{\mu}(B) = \frac{209.4 - 194.7}{2} + \frac{251.3 - 212}{2} = 27 \]

\[ SS(\hat{\theta}_3) = \frac{(27)^2}{(\frac{1}{2})^2 + (\frac{1}{2})^2 + (\frac{1}{2})^2 + (\frac{1}{2})^2} = \frac{(27)^2}{\frac{1}{7}} = 5103 \]

and

\[ F = \frac{5103}{1185} = 4.3 \]

since \( F(0.05, 1, 24) = 4.26 \) GENDER effect significant at \( \alpha = 0.05 \).
Confidence intervals for effects

If $\theta = c'\mu$, $100(1 - \alpha)\%$ confidence interval given by

$$\hat{\theta} \pm t(\alpha/2, N - t)\sqrt{MS[E]\sum \frac{c_i^2}{n_i}}$$

For the cholesterol data, with $t(0.025, 24) = 2.06$ we have a 95% confidence interval for the AGE $\times$ GENDER interaction effect:

$$24.6 \pm 2.06 \sqrt{\frac{4}{7}1185} \quad \text{or} \quad 24.6 \pm 2.06(26.0) \quad \text{or} \quad (-29, 78)$$

a 95% confidence interval for the AGE effect:

$$29.6 \pm 2.06 \sqrt{\frac{1}{7}1185} \quad \text{or} \quad 29.6 \pm 2.06(13.0) \quad \text{or} \quad (2.7, 56.4)$$

and a 95% confidence interval for the GENDER effect:

$$27.0 \pm 2.06 \sqrt{\frac{1}{7}1185} \quad \text{or} \quad 27.0 \pm 2.06(13.0) \quad \text{or} \quad (0.15, 53.9).$$

The term under the $\sqrt{}$ is the estimated standard error of the estimated contrast:

$$\hat{SE}(\sum c_i\bar{y}_{i+}) = \sqrt{MS[E]\sum \frac{c_i^2}{n_i}}$$
SAS code for cholesterol problem

data one;
  input cohort $ @;
  do subj=1 to 7;
    input y @;
    if cohort="I" then do; gender="W"; age="y"; end;
    else if cohort="II" then do; gender="M"; age="y";end;
    else if cohort="III" then do; gender="W"; age="o";end;
    else if cohort="IV" then do; gender="M"; age="o";end;
    output;
  end;
cards;
I  221 213 202 183 185 197 162
II 271 192 189 209 227 236 142
III 262 193 224 201 161 178 265
IV  192 253 248 278 232 267 289
;
run;
proc glm;
  class cohort;
  model y=cohort/clparm;
  constrast "main effect of age " cohort -1 -1 1 1;
  constrast "main effect of gender" cohort -1 1 -1 1;
  constrast "interaction effect " cohort -1 1 1 -1;
  estimate "main effect of age " cohort -1 -1 1 1/divisor=2;
  estimate "main effect of gender" cohort -1 1 -1 1/divisor=2;
  estimate "interaction effect " cohort -1 1 1 -1;
run;

proc glm;
  class gender age;
  model y=age|gender;
run;

(SAS will overlook misspelling of contrast.)
**SAS output (abbreviated) for cholesterol problem**

The SAS System
The GLM Procedure

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>cohort</td>
<td>4</td>
<td>I II III IV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sum of Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>12280.85714</td>
<td>4093.61905</td>
<td>3.46</td>
<td>0.0323</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>28434.57143</td>
<td>1184.77381</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>27</td>
<td>40715.42857</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>Coeff Var</th>
<th>Root MSE</th>
<th>y Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.301627</td>
<td>15.87245</td>
<td>34.42054</td>
<td>216.8571</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Contrast SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>main effect of age</td>
<td>1</td>
<td>6121.285714</td>
<td>6121.285714</td>
<td>5.17</td>
<td>0.0323</td>
</tr>
<tr>
<td>main effect of gender</td>
<td>1</td>
<td>5103.000000</td>
<td>5103.000000</td>
<td>4.31</td>
<td>0.0488</td>
</tr>
<tr>
<td>interaction effect</td>
<td>1</td>
<td>1056.571429</td>
<td>1056.571429</td>
<td>0.89</td>
<td>0.3544</td>
</tr>
</tbody>
</table>

| Standard Parameter | Estimate | Error | t Value | Pr > |t| |
|--------------------|----------|-------|---------|------|---|
| main effect of age | 29.5714286 | 13.0097426 | 2.27 | 0.0323 |
| main effect of gender | 27.0000000 | 13.0097426 | 2.08 | 0.0488 |
| interaction effect | -24.5714286 | 26.0194851 | -0.94 | 0.3544 |

<table>
<thead>
<tr>
<th>Parameter</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>main effect of age</td>
<td>2.7206396 to 56.4222175</td>
</tr>
<tr>
<td>main effect of gender</td>
<td>0.1492111 to 53.8507889</td>
</tr>
<tr>
<td>interaction effect</td>
<td>-78.2730065 to 29.1301493</td>
</tr>
</tbody>
</table>
The GLM Procedure

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>2</td>
<td>M, W</td>
</tr>
<tr>
<td>age</td>
<td>2</td>
<td>o, y</td>
</tr>
</tbody>
</table>

Sum of

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>12280.85714</td>
<td>4093.61905</td>
<td>3.46</td>
<td>0.0323</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>28434.57143</td>
<td>1184.77381</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>27</td>
<td>40715.42857</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>Coeff Var</th>
<th>Root MSE</th>
<th>y Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.301627</td>
<td>15.87245</td>
<td>34.42054</td>
<td>216.8571</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1</td>
<td>6121.285714</td>
<td>6121.285714</td>
<td>5.17</td>
<td>0.0323</td>
</tr>
<tr>
<td>gender</td>
<td>1</td>
<td>5103.000000</td>
<td>5103.000000</td>
<td>4.31</td>
<td>0.0488</td>
</tr>
<tr>
<td>gender*age</td>
<td>1</td>
<td>1056.571429</td>
<td>1056.571429</td>
<td>0.89</td>
<td>0.3544</td>
</tr>
</tbody>
</table>

Exercise:

1. Express the effects below in terms of model parameters $\alpha_i, \beta_j, (\alpha\beta)_{ij}$:
   (a) $\mu(AB_1)$
   (b) $\mu(AB_2)$
   (c) $\mu(A_1B)$
   (d) $\mu(A)$
   (e) $\mu(B)$
   (f) $\mu(AB)$

2. Estimate these effects

<table>
<thead>
<tr>
<th>GENDER</th>
<th>female ($j = 1$)</th>
<th>male ($j = 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>younger ($i = 1$)</td>
<td>194.7</td>
</tr>
<tr>
<td></td>
<td>older ($i = 2$)</td>
<td>212.0</td>
</tr>
</tbody>
</table>
\[ a \times b \text{ designs} \]

An example: Entomologist records energy expended \((y)\) by \(N = 27\) honeybees at \(a = 3\) temperature \((A)\) levels \((20, 30, 40^\circ C)\) consuming liquids with \(b = 3\) levels of sucrose concentration \((B)\) \((20\%, 40\%, 60\%)\) in a balanced, completely randomized \(3 \times 3\) design.

<table>
<thead>
<tr>
<th>Temp</th>
<th>Suc</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>3.1</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>5.5</td>
</tr>
<tr>
<td>20</td>
<td>60</td>
<td>7.9</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>6.6</td>
</tr>
<tr>
<td>30</td>
<td>40</td>
<td>11.5</td>
</tr>
<tr>
<td>30</td>
<td>60</td>
<td>17.5</td>
</tr>
<tr>
<td>40</td>
<td>20</td>
<td>7.7</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>15.7</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
<td>19.1</td>
</tr>
</tbody>
</table>

The SAS System
The GLM Procedure

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEMP</td>
<td>3</td>
<td>20 30 40</td>
</tr>
<tr>
<td>SUC</td>
<td>3</td>
<td>20 40 60</td>
</tr>
</tbody>
</table>

Sum of

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>8</td>
<td>630.2474074</td>
<td>78.7809259</td>
<td>87.07</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>16.2866667</td>
<td>0.9048148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>26</td>
<td>646.5340741</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>Coeff Var</th>
<th>Root MSE</th>
<th>y Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.974809</td>
<td>8.795505</td>
<td>0.951218</td>
<td>10.81481</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEMP</td>
<td>2</td>
<td>293.1585185</td>
<td>146.5792593</td>
<td>162.00</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>SUC</td>
<td>2</td>
<td>309.9585185</td>
<td>154.9792593</td>
<td>171.28</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>TEMP*SUC</td>
<td>4</td>
<td>27.1303704</td>
<td>6.7825926</td>
<td>7.50</td>
<td>0.0010</td>
</tr>
</tbody>
</table>
3 × 3 honeybee example continued

Unlike 2 × 2 study, not possible to express interaction between factors A: TEMP and B: SUCROSE using a single number (w/ 1 df).

<table>
<thead>
<tr>
<th>Level of TEMP</th>
<th>Level of SUC</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>3</td>
<td>3.8333333</td>
<td>0.80829038</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>3</td>
<td>6.5000000</td>
<td>0.91651514</td>
</tr>
<tr>
<td>20</td>
<td>60</td>
<td>3</td>
<td>8.8000000</td>
<td>0.78102497</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>3</td>
<td>6.8000000</td>
<td>0.75498344</td>
</tr>
<tr>
<td>30</td>
<td>40</td>
<td>3</td>
<td>12.6000000</td>
<td>0.98488578</td>
</tr>
<tr>
<td>30</td>
<td>60</td>
<td>3</td>
<td>16.0000000</td>
<td>1.41067360</td>
</tr>
<tr>
<td>40</td>
<td>20</td>
<td>3</td>
<td>8.5000000</td>
<td>0.91651514</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>3</td>
<td>15.3000000</td>
<td>0.87177979</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
<td>3</td>
<td>19.0000000</td>
<td>0.95393920</td>
</tr>
</tbody>
</table>

The plot above is called an interaction plot.

Exercise: Obtain an interaction plot for the cholesterol data.
Partitioning $SS[Tot]$ in $a \times b$ design

Two-way ANOVA Model:

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk} \]

\((i = 1, 2, \ldots, a \text{ and } j = 1, 2, \ldots, b \text{ and } k = 1, 2, \ldots, n)\)

Deviations:

total: \(y_{ijk} - \bar{y}_{+++}\)

due to level \(i\) of factor A: \(\bar{y}_{i++} - \bar{y}_{+++}\)

due to level \(j\) of factor B: \(\bar{y}_{+j} - \bar{y}_{+++}\)

due to levels \(i\) of factor A and \(j\) of factor B after subtracting main effects:

\[ \bar{y}_{ij} - \bar{y}_{+++} - (\bar{y}_{i+} - \bar{y}_{+++}) - (\bar{y}_{+j} - \bar{y}_{+++}) = \bar{y}_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{+++} \]

\[
SS[Tot] = \sum_{i} \sum_{j} \sum_{k} (y_{ijk} - \bar{y}_{+++})^2
\]

\[
SS[A] = \sum_{i} \sum_{j} \sum_{k} (\bar{y}_{i+} - \bar{y}_{+++})^2
\]

\[
SS[B] = \sum_{i} \sum_{j} \sum_{k} (\bar{y}_{+j} - \bar{y}_{+++})^2
\]

\[
SS[AB] = \sum_{i} \sum_{j} \sum_{k} (\bar{y}_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{+++})^2
\]

\[
SS[E] = \sum_{i} \sum_{j} \sum_{k} (y_{ijk} - \bar{y}_{ij+})^2
\]

\[ y_{ijk} = \bar{y}_{+++} + (\bar{y}_{i+} - \bar{y}_{+++}) + (\bar{y}_{+j} - \bar{y}_{+++}) + (\bar{y}_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{+++}) + y_{ijk} - \bar{y}_{ij+} \]

\[ y_{ijk} - \bar{y}_{+++} = (\bar{y}_{i+} - \bar{y}_{+++}) + (\bar{y}_{+j} - \bar{y}_{+++}) + (\bar{y}_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{+++}) + y_{ijk} - \bar{y}_{ij+} \]

Square both sides, \(\times\)-products vanish. (See output, p. 133, note that SS terms add up to model SS. Also, Type I and III SS equal.)
Test for interaction effect generalizes from p.127:

\[ H_0 : (\alpha \beta)_{ij} = 0 \text{ vs. } H_1 : (\alpha \beta)_{ij} \neq 0 \text{ for some } i, j \]

\[ F = \frac{MS[AB]}{MS[E]} \]

on \((a - 1)(b - 1)\) and \(N - ab\) numerator, denominator \(df\).

For honeybee data,

\[ SS[AB] = n \sum_{i=1}^{3} \sum_{j=1}^{3} (\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2 = 27.1 \]

\[ F = \frac{27.1/4}{0.9} = 7.5 \]

which is highly significant \((p = 0.001)\) on 4,18 \(df\).

We could proceed to test for main effects, but we won’t.

Q: Why not?

A: Because effect of one factor depends on the level of the other factor, it doesn’t make sense to talk about main effects.

If one insists on main effects, the appropriate \(F\)-ratios are

\[ F_A = \frac{SS[A]/(a - 1)}{MS[E]} \text{ on } a - 1, N - ab \text{ df} \]

\[ F_B = \frac{SS[B]/(b - 1)}{MS[E]} \text{ on } b - 1, N - ab \text{ df} \]
Another $a \times b$ design - no interaction

Yields on 36 tomato crops from balanced, complete, crossed design with $a = 3$ varieties ($A$) at $b = 4$ planting densities ($B$):

<table>
<thead>
<tr>
<th>Variety</th>
<th>Density k/hectare</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>7.9 9.2 10.5</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>8.1 8.6 10.1</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>15.3 16.1 17.5</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>11.2 12.8 13.3</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>11.5 12.7 13.7</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>16.6 18.5 19.2</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>12.1 12.6 14.0</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>13.7 14.4 15.4</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>18.0 20.8 21.0</td>
</tr>
<tr>
<td>1</td>
<td>40</td>
<td>9.1 10.8 12.5</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>11.3 12.5 14.5</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>17.2 18.4 18.9</td>
</tr>
</tbody>
</table>

ANOVA table

```
The SAS System
The GLM Procedure

Class Levels Values
a 3 1 2 3
b 4 10 20 30 40

Number of observations  36

Sum of Sources   DF   Squares  Mean Square  F Value  Pr > F

Model           11 422.3155556 38.3923232 24.22 <.0001
Error           24  38.0400000  1.5850000
Corrected Total 35 460.3555556

R-Square Coeff Var Root MSE  y Mean
0.917368 9.064568 1.258968 13.88889

Source   DF  Type I SS  Mean Square  F Value  Pr > F
a        2 327.5972222 163.7986111 103.34 <.0001
b        3  86.6866667  28.8955556  18.23 <.0001
a*b      6  8.0316667  1.3386111  0.84  0.5484
```
Analysis of replicated two (or more) factor designs often proceed according to the following directions:

1. Check for interaction
2. If no interaction, analyze main effects
3. If interaction, analyze simple effects

Since there is no evidence of interaction, we proceed to analyze main effects. The $F$-ratios for factors A and B are each highly significant ($p < 0.0001$).

<table>
<thead>
<tr>
<th>Level of</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>11.3333333</td>
<td>1.88309867</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>12.2083333</td>
<td>2.34887142</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>18.1250000</td>
<td>1.73369023</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level of</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>11.4777778</td>
<td>3.75458978</td>
</tr>
<tr>
<td>20</td>
<td>9</td>
<td>14.3888889</td>
<td>2.96835158</td>
</tr>
<tr>
<td>30</td>
<td>9</td>
<td>15.7777778</td>
<td>3.36480972</td>
</tr>
<tr>
<td>40</td>
<td>9</td>
<td>13.9111111</td>
<td>3.53250777</td>
</tr>
</tbody>
</table>
A conventional look at main effects is just to make pairwise comparisons among marginal means, after averaging over other factors. Pairwise comparisons of density means using Tukey’s procedure with $\alpha = 0.05$ are given below.

(Use `means b/tukey;` to obtain the output.)

The GLM Procedure

Tukey’s Studentized Range (HSD) Test for $y$

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

<table>
<thead>
<tr>
<th>Alpha</th>
<th>0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error Degrees of Freedom</td>
<td>24</td>
</tr>
<tr>
<td>Error Mean Square</td>
<td>1.585</td>
</tr>
<tr>
<td>Critical Value of Studentized Range</td>
<td>3.90126</td>
</tr>
<tr>
<td>Minimum Significant Difference</td>
<td>1.6372</td>
</tr>
</tbody>
</table>

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>Tukey Grouping</th>
<th>Mean</th>
<th>N</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15.778</td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td>A</td>
<td>14.389</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>13.911</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td>C</td>
<td>11.478</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>
A three-factor example

In a balanced, complete, crossed design, \( N = 36 \) shrimp were randomized to \( abc = 12 \) treatment combinations from the factors below:

A1: Temperature at 25\(^o\) C
A2: Temperature at 35\(^o\) C
B1: Density of shrimp population at 80 shrimp/40 l
B2: Density of shrimp population at 160 shrimp/40 l
C1: Salinity at 10 units
C2: Salinity at 25 units
C3: Salinity at 40 units

The response variable of interest is weight gain \( Y_{ijkl} \) after four weeks.

Three-way ANOVA Model:

\[
Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + E_{ijkl}
\]

\[
i = 1, 2
\]
\[
j = 1, 2
\]
\[
k = 1, 2, 3
\]
\[
l = 1, 2, 3
\]
\[
E_{ijkl} \overset{iid}{\sim} N(0, \sigma^2)
\]
The GLM Procedure

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>11</td>
<td>467636.3333</td>
<td>42512.3939</td>
<td>14.64</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>69690.6667</td>
<td>2903.7778</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>35</td>
<td>537327.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>15376.0000</td>
<td>15376.0000</td>
<td>5.30</td>
<td>0.0304</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>21218.7778</td>
<td>21218.7778</td>
<td>7.31</td>
<td>0.0124</td>
</tr>
<tr>
<td>a*b</td>
<td>1</td>
<td>8711.1111</td>
<td>8711.1111</td>
<td>3.00</td>
<td>0.0961</td>
</tr>
<tr>
<td>c</td>
<td>2</td>
<td>96762.5000</td>
<td>48381.2500</td>
<td>16.66</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>a*c</td>
<td>2</td>
<td>300855.1667</td>
<td>150427.5833</td>
<td>51.80</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>a<em>b</em>c</td>
<td>2</td>
<td>24038.3889</td>
<td>12019.1944</td>
<td>4.14</td>
<td>0.0285</td>
</tr>
</tbody>
</table>

R-Square: 0.870301
Coeff Var: 19.30270
Root MSE: 53.88671
y Mean: 279.1667

![Graph showing mean over levels of a, b, and c]
<table>
<thead>
<tr>
<th>Level of</th>
<th>Level of</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>25</td>
<td>80</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>160</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>80</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>160</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>25</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>40</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>40</td>
<td>6</td>
</tr>
<tr>
<td>b</td>
<td>c</td>
<td>80</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>40</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>160</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>160</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>160</td>
<td>40</td>
<td>6</td>
</tr>
<tr>
<td>a</td>
<td>b</td>
<td>25</td>
<td>80</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>80</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>80</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>160</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>160</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>160</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>80</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>80</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>80</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>160</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>160</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35</td>
<td>160</td>
<td>40</td>
</tr>
</tbody>
</table>
Interpretation of second order interaction

1st order interaction is between two factors
2nd order interaction is between three factors

Consider the $AB$ interaction at each of three levels, $C_1, C_2, C_3$.

To do this, look at three $2 \times 2$ tables as follows:

\[
\begin{array}{c|cc}
   & B1 & B2 \\
\hline
C = 1 & A1 & 70 & 71 \\
   & A2 & 408 & 331 \\
\hline
C = 2 & A1 & 466 & 333 \\
   & A2 & 275 & 312 \\
\hline
C = 3 & A1 & 359.0 & 252 \\
   & A2 & 243 & 231 \\
\end{array}
\]

Q: How is the $ABC$ interaction manifested here?

A: We could compute $\hat{\mu}(ABC_1), \hat{\mu}(ABC_2), \hat{\mu}(ABC_3)$ and see if these first order interactions, with $C$ fixed, are the same. (We know they are not by the $F_{ABC}$ ratio and $p$-value.)

\[
\hat{\mu}(ABC_1) = 408 - 70 - (331 - 71) \approx 77
\]

Exercise: Obtain $\hat{\mu}(ABC_2)$ and $\hat{\mu}(ABC_3)$ as well as AB interaction plots for $C = 1$, $C = 2$ and $C = 3$. Interpret the plots.
Getting interaction contrasts using the ESTIMATE statement in GLM

To get SAS to estimate the interaction like $\mu(ABC_1)$, the AB interaction at $C = 1$, you must specify the parameters involved.

We saw on the last page that

$$\hat{\mu}(ABC_1) = \frac{\bar{y}_{211+} - \bar{y}_{111+}}{\text{effect of } A \text{ at } B_1C_1} - \frac{\bar{y}_{221+} - \bar{y}_{121+}}{\text{effect of } A \text{ at } B_2C_1}$$

Using the model to specify parameters, we can write

$$E(\bar{y}_{211+}) = \mu + \alpha_2 + \beta_1 + \gamma_1 + (\alpha\beta)_{21} + (\alpha\gamma)_{21} + (\beta\gamma)_{11} + (\alpha\beta\gamma)_{211}$$

$$E(-\bar{y}_{111+}) = -\mu - \alpha_1 - \beta_1 - \gamma_1 - (\alpha\beta)_{11} - (\alpha\gamma)_{11} - (\beta\gamma)_{11} - (\alpha\beta\gamma)_{111}$$

$$E(-\bar{y}_{221+}) = -\mu - \alpha_2 - \beta_2 - \gamma_1 - (\alpha\beta)_{22} - (\alpha\gamma)_{21} - (\beta\gamma)_{21} - (\alpha\beta\gamma)_{221}$$

$$E(\bar{y}_{121+}) = \mu + \alpha_1 + \beta_2 + \gamma_1 + (\alpha\beta)_{12} + (\alpha\gamma)_{11} + (\beta\gamma)_{21} + (\alpha\beta\gamma)_{121}$$

Add these all up to get the contrast we’re interested in, $\mu(ABC_1)$.

Note that all terms vanish except the second order parameters and first order $(\alpha\beta)_{ij}$ parameters:

$$\mu(ABC_1) = (\alpha\beta)_{21} - (\alpha\beta)_{11} - (\alpha\beta)_{22} + (\alpha\beta)_{12} + (\alpha\beta\gamma)_{211} - (\alpha\beta\gamma)_{111} - (\alpha\beta\gamma)_{221} + (\alpha\beta\gamma)_{121}$$

These can be rearranged so that they agree with the ordering of the treatment combinations employed by the ESTIMATE statement:

$$\mu(ABC_1) = -(\alpha\beta)_{11} + (\alpha\beta)_{12} + (\alpha\beta)_{21} - (\alpha\beta)_{22} - (\alpha\beta\gamma)_{111} + (\alpha\beta\gamma)_{121} + (\alpha\beta\gamma)_{211} - (\alpha\beta\gamma)_{221}$$
ESTIMATE statement with two two-level factors

If \( A \) appears before \( B \) in the CLASS statement, SAS uses the following ordering for \((\alpha \beta)\) terms when specifying contrasts in ESTIMATE (or CONTRAST) statements:

\[
A1B1, A1B2, A2B1, A2B2
\]

ESTIMATE statement with two two-level and a three level factor

Similarly, with a \texttt{CLASS a b c;} statement, the following order is used for second-order interaction parameters

\[
\]

\[
\]

If \((\alpha \beta)\) is a vector of \( AB \) interaction effects with the default ordering:

\[
(\alpha \beta) = \begin{pmatrix}
(\alpha \beta)_{11} \\
(\alpha \beta)_{12} \\
(\alpha \beta)_{21} \\
(\alpha \beta)_{22}
\end{pmatrix}
\]

and likewise for \((\alpha \beta \gamma)\), then the contrast \( \mu[ABC_1] \) on p. 25 can be written

\[
\mu[ABC_1] = (-1, 1, 1, -1)(\alpha \beta) + (-1, 0, 0, 1, 0, 1, 0, 0, -1, 0, 0)(\alpha \beta \gamma).
\]

Similarly for \( \mu[ABC_2] \) and \( \mu[ABC_3] \):

\[
\mu[ABC_2] = (-1, 1, 1, -1)(\alpha \beta) + (0, -1, 0, 0, 1, 0, 0, 1, 0, -1, 0)(\alpha \beta \gamma).
\]

\[
\mu[ABC_3] = (-1, 1, 1, -1)(\alpha \beta) + (0, 0, -1, 0, 0, 1, 0, 1, 0, 1, 0, -1)(\alpha \beta \gamma).
\]
proc glm;
  class a b c;
  model y=a|b|c;
  estimate "theta1: ABC1" a*b -1 1 1 -1
    a*b*c -1 0 0 1 0 0 1 0 0 -1 0 0;
  estimate "theta2: ABC2" a*b -1 1 1 -1
    a*b*c 0 -1 0 0 1 0 0 1 0 0 -1 0;
  estimate "theta3: ABC3" a*b -1 1 1 -1
    a*b*c 0 0 -1 0 0 1 0 0 1 0 0 -1;
  estimate "t1-av(t2+t3)" a*b*c -2 1 1 2 -1 -1 2 -1 -1 -2 1 1/divisor=2;
  means a|b|c;
run;

| Parameter         | Estimate | Standard Error | t Value | Pr > |t| |
|-------------------|----------|----------------|---------|-------|---------|
| theta1: ABC1      | 77.333333 | 62.2230159     | 1.24    | 0.2259|
| theta2: ABC2      | -169.666667 | 62.2230159    | -2.73   | 0.0118|
| theta3: ABC3      | -94.333333 | 62.2230159     | -1.52   | 0.1426|
| t1-av(t2+t3)      | 209.333333 | 76.2073196     | 2.75    | 0.0112|
| t2-av(t1+t3)      | -161.166667 | 76.2073196    | -2.11   | 0.0450|

The $F_{ABC}$ ratio indicates the three contrasts estimated above are not plausibly ($p = 0.05$) equal. ($\hat{\mu}(ABC_1), \hat{\mu}(ABC_2)$ and $\hat{\mu}(ABC_3)$ differ significantly). Interaction plots from p.24 suggest the comparison

$$\theta = \mu(ABC_1) - \frac{1}{2}(\mu(ABC_2) + \mu(ABC_3))$$

Adding up all the coefficients in this combination yields the contrast below to use with an ESTIMATE statement:

$$(-1, \frac{1}{2}, \frac{1}{2}, 1, -\frac{1}{2}, -\frac{1}{2}, 1, -\frac{1}{2}, -\frac{1}{2}, -1, \frac{1}{2}, \frac{1}{2})$$

PS: Three-factor interactions are not easily interpreted. Effects can sometimes be made additive through a transformation of the response.
Activity w/ ESTIMATE statement
A linear function \( \theta(\beta) \) of parameters is \textit{estimable} if and only if there is a linear combination of \( Y \) whose expected value is \( \theta \).

Exercise: identify the estimable contrasts in each of the ESTIMATE statements in the correspondence below, which pertains to a 3 \( \times \) 2 study with factors and levels

\[
\begin{array}{cc}
\text{Factor} & \text{Levels} \\
A : \text{additive} & \text{acetic, nothing, sorbate} \\
B : \text{uv} & 0, 1. \\
\end{array}
\]

To: osborne@stat.ncsu.edu
Subject: non estimatable estimate statements

I am still having trouble with the estimate statements, the only ones that work for the additive*uv interaction are where we contrast the same additive over the uv, can anything be done about this??

```plaintext
proc glm;
  class additive uv;
  model ycount=additive uv uv*additive;
  estimate 'acetic uv=0 vs acetic uv=1' uv 1 -1 uv*additive 1 -1 0 0 0 0;
  estimate 'acetic uv=0 vs nothing uv=0' uv 1 -1 uv*additive 1 0 -1 0 0 0;
  estimate 'acetic uv=0 vs nothing uv=1' uv 1 -1 uv*additive 1 0 0 -1 0 0;
  estimate 'acetic uv=0 vs sorbate uv=0' uv 1 -1 uv*additive 1 0 0 0 -1 0;
  estimate 'acetic uv=0 vs sorbate uv=1' uv 1 -1 uv*additive 1 0 0 0 0 -1;
  estimate 'acetic uv=1 vs nothing uv=0' uv 1 -1 uv*additive 0 1 -1 0 0 0;
  estimate 'acetic uv=1 vs nothing uv=1' uv 1 -1 uv*additive 0 1 0 -1 0 0;
  estimate 'acetic uv=1 vs sorbate uv=0' uv 1 -1 uv*additive 0 1 0 0 -1 0;
  estimate 'acetic uv=1 vs sorbate uv=1' uv 1 -1 uv*additive 0 1 0 0 0 -1;
  estimate 'nothing uv=0 vs nothing uv=1' uv 1 -1 uv*additive 0 0 1 -1 0 0;
  estimate 'nothing uv=0 vs sorbate uv=0' uv 1 -1 uv*additive 0 0 1 0 -1 0;
  estimate 'nothing uv=0 vs sorbate uv=1' uv 1 -1 uv*additive 0 0 1 0 0 -1;
  estimate 'nothing uv=1 vs sorbate uv=0' uv 1 -1 uv*additive 0 0 0 1 -1 0;
  estimate 'nothing uv=1 vs sorbate uv=1' uv 1 -1 uv*additive 0 0 0 1 0 -1;
  estimate 'sorbate uv=0 vs sorbate uv=1' uv 1 -1 uv*additive 0 0 0 0 1 -1;
  estimate 'uv=0 vs uv=1' uv 1 -1;
  estimate 'acetic vs nothing' additive 1 -1;
  estimate 'acetic vs sorbate' additive 1 0 -1;
  estimate 'nothing vs sorbate' additive 0 1 -1;
```
Recall the $2 \times 2$ cholesterol study. Suppose the study is unbalanced and the data are given by

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
<th>Marginal mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>271,192,189,209,</td>
<td>162</td>
<td>$\bar{y}_{1++} = 212.3$</td>
</tr>
<tr>
<td>young</td>
<td></td>
<td>227,236</td>
<td></td>
<td></td>
</tr>
<tr>
<td>old</td>
<td></td>
<td>289</td>
<td>262,193,224,201</td>
<td>$\bar{y}_{2++} = 221.6$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>161,178,265</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\bar{y}_{+1+} = 230.4$</td>
<td>$\bar{y}_{+2+} = 205.8$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Marginal sample means are not real useful in this unbalanced study.

Q: How are group population means estimated then?
A: **Least squares means** (what would be estimated by marginal means if design were balanced).
Parametric expressions for all the population means of interest are given below for the additive model:

<table>
<thead>
<tr>
<th>Population group</th>
<th>effect of interest</th>
<th>estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young folks</td>
<td>$\mu + \alpha_1 + \frac{1}{2}(\beta_1 + \beta_2)$</td>
<td>188.03</td>
</tr>
<tr>
<td>Older folks</td>
<td>$\mu + \alpha_2 + \frac{1}{2}(\beta_1 + \beta_2)$</td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>$\mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \beta_1$</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>$\mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \beta_2$</td>
<td></td>
</tr>
<tr>
<td>Young men</td>
<td>$\mu + \alpha_1 + \beta_1$</td>
<td></td>
</tr>
<tr>
<td>Older men</td>
<td>$\mu + \alpha_2 + \beta_1$</td>
<td></td>
</tr>
<tr>
<td>Young women</td>
<td>$\mu + \alpha_1 + \beta_2$</td>
<td></td>
</tr>
<tr>
<td>Older women</td>
<td>$\mu + \alpha_2 + \beta_2$</td>
<td></td>
</tr>
</tbody>
</table>

Invoking the command `lsmeans age gender;` will report least squares estimates for the first four means above.

```
The GLM Procedure
Least Squares Means

| gender | y LSMEAN | Error  | Pr > |t| |
|--------|----------|--------|------|---|
| m      | 251.525773 | 16.233482 | <.0001 |
| w      | 183.597938  | 15.842256 | <.0001 |
```

```
The GLM Procedure
Least Squares Means

| age   | y LSMEAN | Error  | Pr > |t| |
|-------|----------|--------|------|---|
| jr    | 188.025773 | 16.233482 | <.0001 |
| sr    | 247.097938  | 15.842256 | <.0001 |
```

All of these quantities are estimated using linear combinations of the treatment means of the form:

$$\hat{\theta} = c_{11}\bar{y}_{11+} + c_{12}\bar{y}_{12+} + c_{21}\bar{y}_{21+} + c_{22}\bar{y}_{22+}.$$ 

The coefficients are chosen so that $E(\hat{\theta}) = \theta$ and $\sum \frac{c_{ij}^2}{n_{ij}}$ is minimized.
Example: What are the coefficients for the contrast which estimates the population mean for young folks, $\mu + \alpha_1 + \frac{1}{2}(\beta_1 + \beta_2)$ with minimum variance?

$$
c_{11} + c_{12} = 1 \text{ (coeff for } \alpha_1) \\
c_{21} + c_{22} = 0 \text{ (coeff for } \alpha_2) \\
c_{11} + c_{21} = \frac{1}{2} \text{ (coeff for } \beta_1) \\
c_{12} + c_{22} = \frac{1}{2} \text{ (coeff for } \beta_2)
$$

Variance is then proportional to

$$
\frac{c_{11}^2}{n_{11}} + \frac{c_{12}^2}{n_{12}} + \frac{c_{21}^2}{n_{21}} + \frac{c_{22}^2}{n_{22}} = \frac{c_{11}^2}{n_{11}} + \frac{(1 - c_{11})^2}{n_{12}} + \frac{(\frac{1}{2} - c_{11})^2}{n_{21}} + \frac{(c_{11} - \frac{1}{2})^2}{n_{22}}
$$

which is minimized at $c_{11} = \frac{66}{97}$ by setting the derivative to zero and solving. The least squares mean is then

$$
\hat{\theta} = \frac{66}{97}\bar{y}_{11} + (1 - \frac{66}{97})\bar{y}_{12} + (\frac{1}{2} - \frac{66}{97})\bar{y}_{21} + (\frac{66}{97} - \frac{1}{2})\bar{y}_{22} = 188.03.
$$

Similarly for old folks, the contrast with minimum variance has

$$
c_{11} = 18/97 = -c_{12}
$$

and

$$
c_{21} = \frac{1}{2} - 18/97, \quad c_{22} = \frac{1}{2} + 18/97
$$

so that the estimate for the old folks mean is

$$
\frac{18}{97}\bar{y}_{11} - \frac{18}{97}\bar{y}_{12} + \left(\frac{1}{2} - \frac{18}{97}\right)\bar{y}_{21} + \left(\frac{1}{2} + \frac{18}{97}\right)\bar{y}_{22} = 247.1.
$$

Exercise: obtain least squares estimators and estimates of marginal means for men and women as well as for each age $\times$ gender combination.
Q: Is there an age effect? Should we base our conclusion on
\[
\sum_i \sum_j \sum_k (\bar{y}_{i++} - \bar{y}_{+++})^2 = 325.6?
\]

A: Might not be a good idea if factor B has an effect.
(This is the type I SS for Age if Age is the first factor entered into
the model, or the so-called unadjusted sum of squares for age.)

Alternatively, consider the contrast
\[
\theta_{age} = \alpha_1 - \alpha_2.
\]

We can obtain the coefficients of the LS estimate of this contrast and
then use them to get \(SS[\hat{\theta}_{age}]\), which is the sum of squares for the
age effect adjusted for gender, or type II sum of squares for age:
\[
\hat{\theta}_{age} = \hat{\alpha}_1 - \hat{\alpha}_2 = c_{11}\bar{y}_{11} + c_{12}\bar{y}_{12} + c_{21}\bar{y}_{21} + c_{22}\bar{y}_{22}
\]

where
\[
\begin{align*}
c_{11} + c_{12} &= 1 (\text{coeff for } \alpha_1) \\
c_{21} + c_{22} &= -1 (\text{coeff for } \alpha_2) \\
c_{11} + c_{21} &= 0 (\text{coeff for } \beta_1) \\
c_{12} + c_{22} &= 0 (\text{coeff for } \beta_2)
\end{align*}
\]

\(\text{Var}(\hat{\theta}_{age})\) is minimized when \(c_{11} = \frac{48}{97}\) which leads to
\[
\hat{\theta}_{age} = -59.07
\]

with
\[
SS[\hat{\theta}_{age}] = \frac{(-59.07)^2}{\left(\frac{48}{97}\right)^2} + \frac{(-1-\frac{48}{97})^2}{1} + \frac{(-\frac{48}{97})^2}{1} + \frac{(-1+\frac{48}{97})^2}{7} = 6044.
\]
/*
I  221  213  202  183  185  197  162
II  271  192  189  209  227  236  142
III  262  193  224  201  161  178  265
IV  192  253  248  278  232  267  289
*/

options ls=75;
data one;
in  put gender $ age $2. @@;
do i=1 to 7;
in  put y @@;
   output;
end;
cards;
w  jr  .     .     .     .     .     .     162
m  jr  271  192  189  209  227  236  .
w  sr  262  193  224  201  161  178  265
m  sr  .     .     .     .     .     .     289
;
run;

proc glm;
class age gender;
model y=age gender/solution;
lsmeans gender age/stderr;
estimate "lsmean for young folks" intercept 2 age 2 0 gender 1 1/divisor=2;
estimate "lsmean for older folks" intercept 2 age 0 2 gender 1 1/divisor=2;
estimate "lsmean for men" intercept 2 age 1 1 gender 2 0/divisor=2;
estimate "lsmean for women" intercept 2 age 1 1 gender 0 2/divisor=2;
estimate "lsmean for young men" intercept 1 age 1 0 gender 1 0;
estimate "lsmean for young women" intercept 1 age 1 0 gender 0 1;
estimate "lsmean for old men" intercept 1 age 0 1 gender 1 0;
estimate "lsmean for old women" intercept 1 age 0 1 gender 0 1;
contrast "age effect" age 1 -1;
estimate "age effect" age 1 -1;
contrast "gender effect" gender 1 -1;
means gender age;
run;
The SAS System
The GLM Procedure
Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>2</td>
<td>jr sr</td>
</tr>
<tr>
<td>gender</td>
<td>2</td>
<td>m w</td>
</tr>
</tbody>
</table>

Number of observations 28

NOTE: Due to missing values, only 15 observations can be used in this analysis.

<table>
<thead>
<tr>
<th>Sum of</th>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>8318.06735</td>
<td>4159.03368</td>
<td>3.42</td>
<td>0.0669</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>14606.86598</td>
<td>1217.23883</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>14</td>
<td>22924.93333</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>Coeff Var</th>
<th>Root MSE</th>
<th>y Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.362839</td>
<td>16.05812</td>
<td>34.8895</td>
<td>217.2667</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1</td>
<td>325.629762</td>
<td>325.629762</td>
<td>0.27</td>
<td>0.6144</td>
</tr>
<tr>
<td>gender</td>
<td>1</td>
<td>7992.437592</td>
<td>7992.437592</td>
<td>6.57</td>
<td>0.0249</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1</td>
<td>6044.348306</td>
<td>6044.348306</td>
<td>4.97</td>
<td>0.0457</td>
</tr>
<tr>
<td>gender</td>
<td>1</td>
<td>7992.437592</td>
<td>7992.437592</td>
<td>6.57</td>
<td>0.0249</td>
</tr>
</tbody>
</table>

| Standard Parameter | Estimate | Error | t Value | Pr > |t| |
|--------------------|----------|-------|---------|------|----|
| Intercept          | 213.1340206 B | 12.77243521 | 16.69 | <.0001 |
| age jr             | -59.0721649 B | 26.50916484 | -2.23 | 0.0457 |
| age sr             | 0.00000000 B  | .         | .      |      |
| gender m           | 67.9278351 B  | 26.50916484 | 2.56  | 0.0249 |
| gender w           | 0.00000000 B  | .         | .      |      |
Least Squares Means

| gender | y LSMEAN | Error | Pr > |t| |
|--------|----------|-------|------|---|
| m      | 251.525773 | 16.233482 | <.0001 |
| w      | 183.597938  | 15.842256  | <.0001 |

| age    | y LSMEAN   | Error   | Pr > |t| |
|--------|------------|---------|------|---|
| jr     | 188.025773 | 16.233482 | <.0001 |
| sr     | 247.097938 | 15.842256 | <.0001 |

Contrast

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Contrast SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>age effect</td>
<td>1</td>
<td>6044.348306</td>
<td>6044.348306</td>
<td>4.97</td>
<td>0.0457</td>
</tr>
<tr>
<td>gender effect</td>
<td>1</td>
<td>7992.437592</td>
<td>7992.437592</td>
<td>6.57</td>
<td>0.0249</td>
</tr>
</tbody>
</table>

| Parameter          | Estimate | Error | t Value | Pr > |t| |
|--------------------|----------|-------|---------|------|---|
| lsmean for young folks | 188.025773 | 16.233481 | 11.58 | <.0001 |
| lsmean for older folks | 247.097938 | 15.842256 | 15.60 | <.0001 |
| lsmean for men      | 251.525773 | 16.233481 | 15.49 | <.0001 |
| lsmean for women    | 183.597938 | 15.842256 | 11.59 | <.0001 |
| lsmean for young men | 221.989691 | 13.719796 | 16.18 | <.0001 |
| lsmean for old men  | 281.061856 | 26.271409 | 10.70 | <.0001 |
| lsmean for old women | 213.134021 | 12.772435 | 16.69 | <.0001 |
| age effect          | -59.072165 | 26.509164 | -2.23 | 0.0457 |

(These notes were adapted from one of Dr. Dickey’s lectures:)

http://www.stat.ncsu.edu/people/dickey/courses/st512/crsnotes/rnotes_unbal.htm
Block Designs

**Reading** Ch. 15.3, 15.4, 15.6

Motivation - sometimes the variability of responses among experimental units is large, making detection of differences among treatment means $\mu_1, \mu_2, \ldots, \mu_t$ difficult.

In a randomized block design (RBD),

1. matched sets of experimental units are formed, each consisting of $t$ units. Goal is reduced variability of response within a block. That is, the units within a block are homogeneous. Variance between blocks is ok.

2. Blocks are randomly assigned to each of the $t$ treatments. (Restricted randomization, as opposed to a completely randomized design.)
Acrophobia can be treated in several ways.

- “Contact desensitization” - activity/task demonstrated then walked through while a therapist is in constant contact with the subject.
- “Demonstration participation” - therapist talks subject through task without any contact.
- “Live Modeling” - subject simply watches completion of task

Severity of acrophobia measured by HAT (Height Avoidance Test) scores. The point of the study is to investigate the effectiveness of the three therapies. The study will measure HAT scores before and after therapy. There is considerable heterogeneity in the degree to which acrophobia afflicts subjects. So \( N = 15 \) subjects will be put into blocks according to their original HAT score, then one from each block will be randomly assigned to a therapy: Let \( Y_{ij} \) denote the change in HAT score for subject in block \( j \) assigned to treatment \( i \). (Bigger score means bigger reduction in fear.)

<table>
<thead>
<tr>
<th>Block ( j )</th>
<th>Therapy</th>
<th>( \bar{y}_{+j} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contact Desensitization</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>( y_{12} = 11 )</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>Avg ( \bar{y}_{i+} )</td>
<td>13.6</td>
<td>9</td>
</tr>
</tbody>
</table>
### RBD example

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Therapies</td>
<td>260.9</td>
<td>2</td>
<td>130.5</td>
<td>15.3</td>
</tr>
<tr>
<td>B: Blocks</td>
<td>438</td>
<td>4</td>
<td>109.5</td>
<td>12.8</td>
</tr>
<tr>
<td>Error</td>
<td>68.4</td>
<td>8</td>
<td>8.6</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>767.3</td>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Data taken from Larsen and Marx, 1986)

\[
\]

\[
SS[Tot] = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{++})^2
\]

\[
SS[A] = \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{i+} - \bar{y}_{++})^2 = b \sum_{i=1}^{a} (\bar{y}_{i+} - \bar{y}_{++})^2
\]

\[
SS[B] = \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{+j} - \bar{y}_{++})^2 = a \sum_{j=1}^{b} (\bar{y}_{+j} - \bar{y}_{++})^2
\]

\[
SS[E] = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i+} - \bar{y}_{j} + \bar{y}_{++})^2
\]

Note that

\[
y_{ij} - \bar{y}_{++} = (\bar{y}_{i+} - \bar{y}_{++}) + (\bar{y}_{+j} - \bar{y}_{++}) + (y_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{++})
\]

therapy effect + block effect + error
**F-tests in the RBD**

A model for RBD with fixed treatment (therapy) effects is

\[ Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij} \]

where \( i = 1, \ldots, a \) \( j = 1, \ldots, b \) and \( E_{ij} \overset{iid}{\sim} N(0, \sigma^2) \)

Mean squares obtained by dividing \( SS \) by \( df \):

\[
MS[A] = \frac{SS[A]}{a-1} \\
MS[B] = \frac{SS[B]}{b-1} \\
MS[E] = \frac{SS[E]}{N-a-b+1}
\]

The primary hypothesis of interest is for a therapy effect:

\[ H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad \text{vs} \quad H_1 : \text{not all equal.} \]

Using level \( \alpha \), reject \( H_0 \) if

\[ F = \frac{MS[A]}{MS[E]} > F(\alpha, a-1, N-a-b+1) \]

The \( EMS \) for error is \( E(MS[E]) = \sigma^2 \), but only under the *additivity* assumption that there is no block-trt interaction. This assumption is required for inference about treatment effects in the absence of replication, common to block designs.

For the HAT scores, \( F_A = MS[A]/MS[E] = 130.5/8.6 = 15.3 \)

which has \( p < 0.01 \) on 2, 8 df, providing strong evidence of a therapy effect. Inference, including MCPs, for CONTRASTS involving fixed effects is the same in the complete RBD as it is for other factorial experiments with fixed effects. (E.g. \( \widehat{SE}(\bar{Y}_{i+}) = \sqrt{MS[E]/b} \))
Multiple comparisons among means in the RBD

Scheffé simultaneous 95% confidence intervals for contrasts like

\[ c_1 \mu_1 + c_2 \mu_2 + \cdots + c_a \mu_a \]

look like

\[ c_1 \bar{y}_{1+} + c_2 \bar{y}_{2+} + \cdots + c_a \bar{y}_{a+} \pm \sqrt{(a - 1)(F^*)MS[E] \sum \frac{c_i^2}{b}} \]

where \( F^* = F(0.05, a-1, N-a-b+1) \). For simultaneous pairwise differences, these look like

\[ \bar{y}_{i+} - \bar{y}_{j+} \pm \sqrt{(a - 1)(F^*)MS[E] \frac{2}{b}} \]

“minimum significant difference”

For the HAT scores,

\[ \bar{y}_{1+} = 13.6, \quad \bar{y}_{2+} = 9, \quad \bar{y}_{3+} = 3.4 \]

and

\[ \sqrt{(a - 1)(F^*) (MS[E])(1/5 + 1/5)} = \sqrt{(3 - 1)(4.46)(8.6)(2/5)} = 5.5 \]

with \( \bar{y}_{LM+} \) significantly different from the other two. (LM brings about significantly less improvement than the other two therapies.)

The Tukey minimum significant difference term in the RBD is

\[ q(a, n - a - b + 1, \alpha) \sqrt{MS[E] \frac{1}{b}} \]

For the acrophobia RBD, the term is \( 4.04 \times \sqrt{\frac{8.6}{5}} = 5.3 \)

means therapy/scheffe tukey; will get the job done in SAS.
Tukey’s Studentized Range (HSD) Test for variable: DIFF

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

\[ \text{Alpha} = 0.05 \quad \text{df} = 8 \quad \text{MSE} = 8.55 \]
\[ \text{Critical Value of Studentized Range} = 4.041 \]
\[ \text{Minimum Significant Difference} = 5.2843 \]

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>Tukey Grouping</th>
<th>Mean</th>
<th>N</th>
<th>TREAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>13.600</td>
<td>5</td>
<td>Contact Desensit</td>
</tr>
<tr>
<td>A</td>
<td>9.000</td>
<td>5</td>
<td>Demonstration Pa</td>
</tr>
<tr>
<td>B</td>
<td>3.400</td>
<td>5</td>
<td>Live Modelling</td>
</tr>
</tbody>
</table>

Scheffe’s test for variable: DIFF

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than REGWF for all pairwise comparisons.

\[ \text{Alpha} = 0.05 \quad \text{df} = 8 \quad \text{MSE} = 8.55 \]
\[ \text{Critical Value of F} = 4.45897 \]
\[ \text{Minimum Significant Difference} = 5.5226 \]

<table>
<thead>
<tr>
<th>Scheffe Grouping</th>
<th>Mean</th>
<th>N</th>
<th>TREAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>13.600</td>
<td>5</td>
<td>Contact Desensit</td>
</tr>
<tr>
<td>A</td>
<td>9.000</td>
<td>5</td>
<td>Demonstration Pa</td>
</tr>
<tr>
<td>B</td>
<td>3.400</td>
<td>5</td>
<td>Live Modelling</td>
</tr>
</tbody>
</table>
Another example, blocks are random
(this material to be covered after random effects have been introduced)

A study investigates the efficiency of four different unit-dose injection systems. For each system, an individual subject (pharmacist or nurse) measures the average time it takes to remove a unit of each system from its outer package, assemble it, and simulate an injection. (Data from Larsen and Marx, 1986.)

<table>
<thead>
<tr>
<th>Subject</th>
<th>Standard</th>
<th>Vari-Ject</th>
<th>Unimatic</th>
<th>Tubex</th>
<th>$\bar{y}_{+j}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.6</td>
<td>17.3</td>
<td>24.4</td>
<td>25.0</td>
<td>25.6</td>
</tr>
<tr>
<td>2</td>
<td>31.3</td>
<td>16.4</td>
<td>22.4</td>
<td>26.0</td>
<td>24.0</td>
</tr>
<tr>
<td>3</td>
<td>36.2</td>
<td>18.1</td>
<td>22.8</td>
<td>25.3</td>
<td>25.6</td>
</tr>
<tr>
<td>4</td>
<td>31.1</td>
<td>17.8</td>
<td>21</td>
<td>24</td>
<td>23.5</td>
</tr>
<tr>
<td>5</td>
<td>39.4</td>
<td>18.8</td>
<td>23.3</td>
<td>24.2</td>
<td>26.4</td>
</tr>
<tr>
<td>6</td>
<td>34.7</td>
<td>17</td>
<td>21.8</td>
<td>26.2</td>
<td>24.9</td>
</tr>
<tr>
<td>7</td>
<td>34.1</td>
<td>14.5</td>
<td>23</td>
<td>24</td>
<td>23.9</td>
</tr>
<tr>
<td>8</td>
<td>36.5</td>
<td>17.9</td>
<td>24.1</td>
<td>20.9</td>
<td>24.9</td>
</tr>
<tr>
<td>9</td>
<td>40.7</td>
<td>16.4</td>
<td>31.3</td>
<td>36.9</td>
<td>31.3</td>
</tr>
<tr>
<td>$\bar{y}_{i+}$</td>
<td>35.5</td>
<td>17.1</td>
<td>23.8</td>
<td>25.8</td>
<td>25.6</td>
</tr>
</tbody>
</table>

Model

$$Y_{ij} = \mu + \alpha_i + B_j + E_{ij}$$

- $i = 1, \ldots, 4 = a$ and $j = 1, \ldots, 9 = b$
- $\alpha_i$ denote fixed system effects
- $B_j \overset{iid}{\sim} N(0, \sigma_B^2)$ and $E_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ denote random subject (block) and error effects ($B \perp E$).
data one;
  input subject system time;
cards;
  1 1 35.6
  2 1 31.3
  ...
  8 4 20.9
  9 4 36.9
;
run;

proc mixed method=type3;
  class system subject;
  model time=system/ddfm=satterth;
  random subject;
  lsmeans system/adj=tukey cl pdiff;
run;

The SAS System
The Mixed Procedure
Model Information

Data Set WORK.ONE
Dependent Variable time
Covariance Structure Variance Components
Estimation Method Type 3
Residual Variance Method Factor
Fixed Effects SE Method Model-Based
Degrees of Freedom Method Satterthwaite

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>system</td>
<td>4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>subject</td>
<td>9</td>
<td>1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

Total Observations 36

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>system</td>
<td>3</td>
<td>1559.202222</td>
<td>519.734074</td>
<td>Var(Residual) + Q(system)</td>
</tr>
<tr>
<td>subject</td>
<td>8</td>
<td>177.405000</td>
<td>22.175625</td>
<td>Var(Residual) + 4 Var(subject)</td>
</tr>
<tr>
<td>Residual</td>
<td>24</td>
<td>148.472778</td>
<td>6.186366</td>
<td>Var(Residual)</td>
</tr>
</tbody>
</table>
Clearly, the injection system effects are highly significant, as is the random block (or subject) effect, which has an estimated variance component of
\[
\hat{\sigma}_B^2 = \frac{1}{a}(MS[B] - MS[E]) = \frac{1}{4}(22.2 - 6.2) = 4(\text{ squared seconds})
\]
**Differences of Least Squares Means**

| Effect | System | _System | Estimate | Error | DF  | t Value | Pr > |t| | Adjustment |
|--------|--------|---------|----------|-------|------|---------|-------|---|----------------|
| system 1 | 2 | | 18.3778 | 1.1725 | 24 | 15.67 | <.0001 | Tukey-Kramer |
| system 1 | 3 | | 11.7222 | 1.1725 | 24 | 10.00 | <.0001 | Tukey-Kramer |
| system 1 | 4 | | 9.6778 | 1.1725 | 24 | 8.25 | <.0001 | Tukey-Kramer |
| system 2 | 3 | | -6.6556 | 1.1725 | 24 | -5.68 | <.0001 | Tukey-Kramer |
| system 2 | 4 | | -8.7000 | 1.1725 | 24 | -7.42 | <.0001 | Tukey-Kramer |
| system 3 | 4 | | -2.0444 | 1.1725 | 24 | -1.74 | 0.0940 | Tukey-Kramer |

**Differences of Least Squares Means**

<table>
<thead>
<tr>
<th>Effect</th>
<th>System</th>
<th>_System</th>
<th>Adj P</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
<th>Adj Lower</th>
<th>Adj Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>system 1</td>
<td>2</td>
<td></td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>15.9579</td>
<td>20.7977</td>
<td>15.1433</td>
<td>21.6122</td>
</tr>
<tr>
<td>system 1</td>
<td>3</td>
<td></td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>9.3023</td>
<td>14.1421</td>
<td>8.4878</td>
<td>14.9567</td>
</tr>
<tr>
<td>system 1</td>
<td>4</td>
<td></td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>7.2579</td>
<td>12.0977</td>
<td>6.4433</td>
<td>12.9122</td>
</tr>
<tr>
<td>system 2</td>
<td>3</td>
<td></td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>-9.0755</td>
<td>-4.2356</td>
<td>-9.8900</td>
<td>-3.4211</td>
</tr>
<tr>
<td>system 2</td>
<td>4</td>
<td></td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>-11.1199</td>
<td>-6.2801</td>
<td>-11.9345</td>
<td>-5.4655</td>
</tr>
<tr>
<td>system 3</td>
<td>4</td>
<td></td>
<td>0.3242</td>
<td>0.05</td>
<td>-4.4644</td>
<td>0.3755</td>
<td>-5.2789</td>
<td>1.1900</td>
</tr>
</tbody>
</table>

Note the \( df \) columns:

- For difference of means, pesky mean random effects wash out
- For means, pesky mean random effects don’t wash out, necessitating a Satterthwaite \( df \) approximation

\[
\bar{Y}_{i1+} = \mu + \alpha_{i1} + \bar{B} + \bar{E}_{i1+} \\
\bar{Y}_{i2+} = \mu + \alpha_{i2} + \bar{B} + \bar{E}_{i2+} \\
\bar{Y}_{i1+} - \bar{Y}_{i2+} = \alpha_{i1} - \alpha_{i2} + \bar{E}_{i1+} - \bar{E}_{i2+} \\
SE(\bar{Y}_{i1+}) = \sqrt{\frac{1}{b} (\sigma_B^2 + \sigma^2)} \\
SE(\bar{Y}_{i1+} - \bar{Y}_{i2+}) = \sqrt{\frac{2}{b} \sigma^2}
\]
Latin squares for experiments with two blocking factors

- Experiment with \( \sim 30 \) plants, 3 fertilizer trts.
- blocks to control for variability in exptl. units
- location on bench a second blocking factor

Non-randomized design (number indicates trt):

\[
\begin{array}{ccc}
1 & 2 & 3 \\
3 & 1 & 2 \\
2 & 3 & 1 \\
\end{array}
\]

Fertilizer trts randomized to row×column (or sunlight × iheight) combinations by randomly permuting the

1. columns
2. rows

Eg, random number generator permutes (1, 2, 3) to get (2, 3, 1). Placing the columns in this order leads to

\[
\begin{array}{ccc}
2 & 3 & 1 \\
1 & 2 & 3 \\
3 & 1 & 2 \\
\end{array}
\]
Another random permutation of \((1, 2, 3)\) is \((3, 2, 1)\). Placing the rows in this order leads to an unreplicated \(3 \times 3 \times 3\) design:

\[
\begin{array}{c|c|c}
2 & 3 & 1 \\
1 & 2 & 3 \\
3 & 1 & 2 \\
\end{array} \quad \rightarrow \quad 
\begin{array}{c|c|c}
3 & 1 & 2 \\
1 & 2 & 3 \\
2 & 3 & 1 \\
\end{array}
\]

Suppose nine rows are available. Three latin squares may be generated, one below the other. The one on top is closest to sunlight, the one on bottom furthest. Columns correspond to initial height blocks.

\[
\begin{array}{c|c|c}
3 & 1 & 2 \\
1 & 2 & 3 \\
2 & 3 & 1 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
2 & 3 & 1 \\
3 & 1 & 2 \\
1 & 2 & 3 \\
\end{array}
\]

\[
\begin{array}{c|c|c}
2 & 1 & 3 \\
3 & 2 & 1 \\
1 & 3 & 2 \\
\end{array}
\]
SAS code to illustrate how such an experiment might go follows. To consider an unreplicated $3 \times 3 \times 3$ Latin square, ignore squares 2 and 3. The fixed effects model generated by the code is:

$$Y_{ijk} = \mu + \tau_k + \rho_i + \kappa_j + E_{ijk}$$

where

- $\mu = 13, (\tau_1, \tau_2, \tau_3) = (-1, 0, 1)$ are trt effects
- $(\rho_1, \rho_2, \rho_3) = (1, 0, -1)$ are sunlight effects.
- $(\kappa_1, \kappa_2, \kappa_3) = (-2, 0, 2)$ are initial ht. effects.
- $E_{ijk} \sim iid \sim N(0, \sigma^2)$ with $\sigma^2 = 1$.

Exercises:

1. specify the theoretical mean in each of the nine cells of the unreplicated design in square 1.
2. specify the marginal means for trt, row and column

Fixed effect model replicated design has extra term:

$$Y_{ijkl} = \mu + \tau_l + \rho_{i(k)} + \kappa_j + \beta_k + E_{ijkl}$$

where

- $(\beta_1, \beta_2, \beta_3) = (3, 0, -3)$ are square effects
- For each $k$, $(\rho_{1(k)}, \rho_{2(k)}, \rho_{3(k)}) = (3, 0, -3)$ are nested row effects
data latinsq;
  array square{3} (3,0,-3);
  array slight1{3} (1,0,-1); /* in square 1 */
  array slight2{3} (1,0,-1); /* in square 2 */
  array slight3{3} (1,0,-1); /* in square 3 */
  array iheight{3} (-2,0,2); /* initial height effects */
  array treatment{3} (-1,0,1); /* fertilizer effects */
  input square row col trt;
  growth=round(growth,0.1);
  sigma=1; /* try various values of sigma to generate the data */
  if square=1 then do;
    growth = 10 + square{square} + slight1{row} + iheight{col}
        + treatment{trt} + sigma*rannor(1234);
  end;
  else if square=2 then do;
    growth = 10 + square{square} + slight2{row} + iheight{col}
        + treatment{trt} + sigma*rannor(1234);
  end;
  else do;
    growth = 10 + square{square} + slight3{row} + iheight{col}
        + treatment{trt} + sigma*rannor(1234);
  end;
  height=col;
cards;
  1 1 1 3
  1 1 2 1
  1 1 3 2
  1 2 1 1
  1 2 2 2
  1 2 3 3
  1 3 1 2
  1 3 2 3
  1 3 3 1
  2 1 1 2
  2 1 2 3
  2 1 3 1
  2 2 1 1
  2 2 2 2
  2 2 3 3
  2 3 1 3
  2 3 2 1
  2 3 3 2
  3 1 1 2
  3 1 2 1
  3 1 3 3
  3 2 1 3
  3 2 2 2
  3 2 3 1
  3 3 1 1
  3 3 2 3
  3 3 3 2 ;
data one; set latinsq; if square=1; run;
proc glm data=one;
   class square row col trt;
   model growth = row col trt;
   lsmeans row col trt;
run;

The SAS System
1
The GLM Procedure

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>square</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>row</td>
<td>3</td>
<td>1 2 3</td>
</tr>
<tr>
<td>col</td>
<td>3</td>
<td>1 2 3</td>
</tr>
<tr>
<td>trt</td>
<td>3</td>
<td>1 2 3</td>
</tr>
</tbody>
</table>

Number of observations 9

<table>
<thead>
<tr>
<th>Sum of Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>31.55333333</td>
<td>5.25888889</td>
<td>10.45</td>
<td>0.0899</td>
</tr>
<tr>
<td>Error</td>
<td>2</td>
<td>1.00666667</td>
<td>0.50333333</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>8</td>
<td>32.56000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square Coeff Var Root MSE growth Mean
0.969083 5.415724 0.709460 13.100000

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>row</td>
<td>2</td>
<td>9.94666667</td>
<td>4.97333333</td>
<td>9.88</td>
<td>0.0919</td>
</tr>
<tr>
<td>col</td>
<td>2</td>
<td>20.16000000</td>
<td>10.08000000</td>
<td>20.03</td>
<td>0.0476</td>
</tr>
<tr>
<td>trt</td>
<td>2</td>
<td>1.44666667</td>
<td>0.72333333</td>
<td>1.44</td>
<td>0.4103</td>
</tr>
</tbody>
</table>

growth
row
LSMEAN
1 14.5000000
2 12.8333333
3 11.9666667

col
LSMEAN
1 14.7000000
2 13.5000000
3 11.1000000

trt
LSMEAN
1 12.5333333
2 13.3666667
3 13.4000000
proc glm data=latinsq;
  class square row col trt;
  *model growth = row*square col trt;
  model growth = square row(square) col trt;
  lsmeans square row(square) col trt;
run;

The SAS System
The GLM Procedure

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>12</td>
<td>301.0822222</td>
<td>25.0901852</td>
<td>27.49</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>14</td>
<td>12.7777778</td>
<td>0.9126984</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>26</td>
<td>313.8600000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>square</td>
<td>2</td>
<td>155.3355556</td>
<td>77.6677778</td>
<td>85.10</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>row(square)</td>
<td>6</td>
<td>33.5777778</td>
<td>5.5962963</td>
<td>6.13</td>
<td>0.0025</td>
</tr>
<tr>
<td>col</td>
<td>2</td>
<td>96.6200000</td>
<td>48.3100000</td>
<td>52.93</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>trt</td>
<td>2</td>
<td>15.5488889</td>
<td>7.7744444</td>
<td>8.52</td>
<td>0.0038</td>
</tr>
</tbody>
</table>

growth

<table>
<thead>
<tr>
<th>square</th>
<th>LSMEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.1000000</td>
</tr>
<tr>
<td>2</td>
<td>9.7555556</td>
</tr>
<tr>
<td>3</td>
<td>7.2444444</td>
</tr>
</tbody>
</table>

growth

<table>
<thead>
<tr>
<th>row</th>
<th>square</th>
<th>LSMEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>14.5000000</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>12.8333333</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>11.9666667</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>11.3333333</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>9.5000000</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>8.4333333</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>8.6333333</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>7.1333333</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>5.9666667</td>
</tr>
</tbody>
</table>

growth

<table>
<thead>
<tr>
<th>col</th>
<th>LSMEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.3666667</td>
</tr>
<tr>
<td>2</td>
<td>10.0000000</td>
</tr>
<tr>
<td>3</td>
<td>7.7333333</td>
</tr>
</tbody>
</table>

growth

<table>
<thead>
<tr>
<th>trt</th>
<th>LSMEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.0444444</td>
</tr>
<tr>
<td>2</td>
<td>10.1666667</td>
</tr>
<tr>
<td>3</td>
<td>10.8888889</td>
</tr>
</tbody>
</table>
Exercise

Let $Y_{ij}$ denote the observation in row $i$ column $j$. Let $\bar{Y}_k$ denote the treatment mean for level $k$ of the treatment factor. For an unrepli-
cated latin square, identify these sums of squares:

$$\sum_{i=1}^{a} \sum_{j=1}^{a} (\bar{y}_{i+} - \bar{y}_{++})^2 = SS[\ ]$$

$$\sum_{i=1}^{a} \sum_{j=1}^{a} (y_{ij} - \bar{y}_{++})^2 = SS[\ ]$$

$$a \sum_{j=1}^{a} (\bar{y}_{+j} - \bar{y}_{++})^2 = SS[\ ]$$

$$\sum_{i=1}^{3} \sum_{j=1}^{3} (y_{ij} - \bar{y}_{i+} - \bar{y}_{+j} - \bar{y}_k + 2\bar{y}_{++})^2 = SS[\ ]$$

$$a \sum_{k=1}^{a} (\bar{y}_k - \bar{y}_{++})^2 = SS[\ ]$$

Note that $\bar{y}_k$ is determined by the $i, j$ combination. In the $3 \times 3 \times 3$ scheme used for our plant heights, fertilizer $k = 3$ was assigned to the first row and column so that in the last sum of squares above, for $i = 1, j = 1$, $\bar{y}_k$ is the third fertilizer treatment mean, $\bar{y}_3 = 13.40$. 

A 4 × 4 × 4 example taken from Ott and Longnecker

• Blocking factors: plot rows, plot columns

• Treatment factor: Fertilizer (4 levels, 2 factors)
  (1=broad A, 2=broad B, 3=band A, 4=band B)

\[
\begin{array}{cccc}
1 & 3 & 4 & 2 \\
2 & 1 & 3 & 4 \\
4 & 2 & 1 & 3 \\
3 & 4 & 2 & 1 \\
\end{array}
\]

data watermelons;
  input row col trt yield;
cards;
1 1 1 1.75
1 2 3 1.43
1 3 4 1.28
1 4 2 1.66
2 1 2 1.7
2 2 1 1.78
2 3 3 1.40
2 4 4 1.31
3 1 4 1.35
3 2 2 1.73
3 3 1 1.69
3 4 3 1.41
4 1 3 1.45
4 2 4 1.36
4 3 2 1.65
4 4 1 1.73
;
proc glm;
  class row col trt;
  model yield = row col trt;
  estimate "micronutrient effect A-B" trt 1 -1 1 -1/divisor=2;
  contrast "micronutrient effect A-B" trt 1 -1 1 -1;
  estimate "placement effect" trt 1 1 -1 -1/divisor=2;
  contrast "placement effect" trt 1 1 -1 -1;
  estimate "placement-x-nutrient interaction " trt 1 -1 -1 1;
  contrast "placement-x-nutrient interaction " trt 1 -1 -1 1;
  lsmeans row col trt; run;
The SAS System
The GLM Procedure

Sum of Squares

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>0.49335000</td>
<td>0.05481667</td>
<td>438.53</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>6</td>
<td>0.00075000</td>
<td>0.00012500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>15</td>
<td>0.49410000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Type I SS

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>row</td>
<td>3</td>
<td>0.00085000</td>
<td>0.00028333</td>
<td>2.27</td>
<td>0.1810</td>
</tr>
<tr>
<td>col</td>
<td>3</td>
<td>0.01235000</td>
<td>0.00411667</td>
<td>32.93</td>
<td>0.0004</td>
</tr>
<tr>
<td>trt</td>
<td>3</td>
<td>0.48015000</td>
<td>0.16005000</td>
<td>1280.40</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Contrast

<table>
<thead>
<tr>
<th>Contrast</th>
<th>DF</th>
<th>Contrast SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>micro.effect A-B</td>
<td>1</td>
<td>0.02250000</td>
<td>0.02250000</td>
<td>180.00</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>place.effect</td>
<td>1</td>
<td>0.45562500</td>
<td>0.45562500</td>
<td>3645.00</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>interaction</td>
<td>1</td>
<td>0.00202500</td>
<td>0.00202500</td>
<td>16.20</td>
<td>0.0069</td>
</tr>
</tbody>
</table>

Standard Error

| Parameter       | Estimate | Error     | t Value | Pr > |t| |
|-----------------|----------|-----------|---------|------|---|
| micro.effect A-B | 0.07500000 | 0.00559017 | 13.42   | <.0001|
| place.effect     | 0.33750000 | 0.00559017 | 60.37   | <.0001|
| interaction      | -0.04500000 | 0.01118034 | -4.02   | 0.0069|

trt   yield LSMEAN
1     1.73750000
2     1.68500000
3     1.42250000
4     1.32500000

Plot of mean_yield*placement. Symbol is value of micronutrient.

mean_yield |
| 1.8 +     |
|          A
|          B
| 1.6 +     |
|          |
|          |
| 1.4 + A   |
| B         |
| 1.2 +     |

-----------------------------
band broadcast
proc glm data=watermelons;
class row col micronutrient placement;
model yield = row col micronutrient|placement;
lsmeans micronutrient|placement;
run;

The SAS System
The GLM Procedure

Class Levels Values
row 4 1 2 3 4
col 4 1 2 3 4
micronutrient 2 A B
placement 2 band broadcast

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>0.49335000</td>
<td>0.05481667</td>
<td>438.53</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>6</td>
<td>0.00075000</td>
<td>0.00012500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>15</td>
<td>0.49410000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square | Coeff Var | Root MSE | yield Mean
---------|-----------|----------|-----------
0.998482 | 0.724819  | 0.011180 | 1.542500 |

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>row</td>
<td>3</td>
<td>0.00085000</td>
<td>0.00028333</td>
<td>2.27</td>
<td>0.1810</td>
</tr>
<tr>
<td>col</td>
<td>3</td>
<td>0.01235000</td>
<td>0.00411667</td>
<td>32.93</td>
<td>0.0004</td>
</tr>
<tr>
<td>micronutrient</td>
<td>1</td>
<td>0.02250000</td>
<td>0.02250000</td>
<td>180.00</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>placement</td>
<td>1</td>
<td>0.45562500</td>
<td>0.45562500</td>
<td>3645.00</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>micronutri*placement</td>
<td>1</td>
<td>0.00202500</td>
<td>0.00202500</td>
<td>16.20</td>
<td>0.0069</td>
</tr>
</tbody>
</table>

micronutrient yield LSMEAN
A 1.58000000
B 1.50500000

placement yield LSMEAN
band 1.37375000
broadcast 1.71125000

micronutrient placement yield LSMEAN
A band 1.42250000
A broadcast 1.73750000
B band 1.32650000
B broadcast 1.68500000
**Topic:** Mixed Models for factorial designs

**Reading:** Rao Ch. 14.1, 14.2, 14.3

- One-way random effects model to study *variances*
- Mixed effects models
- Subsampling
- Expected mean squares for mixed models

**An example using one-way random effects model**

- Genetics study w/ beef animals. Measure birthweight $Y$ (lbs).
- $t = 5$ sires, each mated to a separate group of $n = 8$ dams.
- $N = 40$, completely randomized.

<table>
<thead>
<tr>
<th>Sire #</th>
<th>Level</th>
<th>Sample</th>
<th>$\bar{y}_{i+}$</th>
<th>$s_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>177</td>
<td>1</td>
<td>61 100 56 113 99 103 75 62</td>
<td>83.6</td>
<td>22.6</td>
</tr>
<tr>
<td>200</td>
<td>2</td>
<td>75 102 95 103 98 115 98 94</td>
<td>97.5</td>
<td>11.2</td>
</tr>
<tr>
<td>201</td>
<td>3</td>
<td>58 60 60 57 57 59 54 100</td>
<td>63.1</td>
<td>15.0</td>
</tr>
<tr>
<td>202</td>
<td>4</td>
<td>57 56 67 59 58 121 101 101</td>
<td>77.5</td>
<td>25.9</td>
</tr>
<tr>
<td>203</td>
<td>5</td>
<td>59 46 120 115 115 93 105 75</td>
<td>91.0</td>
<td>28.0</td>
</tr>
</tbody>
</table>

Q: Statistical model for these data?

A: One-way fixed effects model?

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

where $\tau_i$ denotes the difference between the mean birthweight of population of offspring from sire $i$ and $\mu$, mean of whole population.
The one-way random effects model

\[ Y_{ij} = \mu_{\text{fixed}} + T_i + E_{ij} \text{ random } \quad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n \]

with

- \( T_1, T_2, \ldots, T_t \iid N(0, \sigma^2_T) \)
- \( E_{11}, \ldots, E_{tn} \iid N(0, \sigma^2) \)
- \( T_1, T_2, \ldots, T_t \) independent of \( E_{11}, \ldots, E_{tn} \)

Features

- \( T_1, T_2, \ldots \) denote random effects, drawn from some population of interest. That is, \( T_1, T_2, \ldots \) is a random sample!
- \( \sigma^2_T \) and \( \sigma^2 \) are called [variance components]
- conceptually different from one-way fixed effects model

For beef animal genetic study, with \( t = 5 \) and \( n = 8 \), the random effects \( T_1, T_2, \ldots, T_5 \) reflect sire-to-sire variability.

No particular interest in \( \tau_1, \tau_2, \ldots, \tau_5 \) from the fixed effects model:

\[ Y_{ij} = \mu_{\text{fixed}} + \tau_i + E_{ij} \text{ random } \quad \text{for } i = 1, 2, \ldots, t \text{ and } j = 1, \ldots, n \]

with

- \( \tau_1, \tau_2, \ldots, \tau_t \) unknown model parameters
- \( E_{11}, \ldots, E_{tn} \iid N(0, \sigma^2) \)
One-way random effects model continued

Exercise: Using the random effects model, specify

\[ E(Y_{ij}) \text{ and } \text{Var}(Y_{ij}) \]

- Two *components* to variability in data: \( \sigma^2, \sigma^2_T \)
- \( T_1, T_2, T_3, T_4, T_5 \) a random sample of sire effects
- Sire effects is a population in its own right.

Contrast this situation with the binding fractions. Why not model antibiotic effects random? Why fixed? (See Ch. 17 for more discussion.)

Model parameters: \( \sigma^2, \sigma^2_T, \mu \)

Sums of squares and mean squares - same as in one-way fixed effects ANOVA:

\[
\begin{align*}
SS[T] &= \sum_{i} \sum_{j} (\bar{y}_i - \bar{y}_{++})^2 \\
SS[E] &= \sum_{i} \sum_{j} (y_{ij} - \bar{y}_i)^2 \\
SS[Tot] &= \sum_{i} \sum_{j} (y_{ij} - \bar{y}_{++})^2
\end{align*}
\]

The ANOVA table is almost the same, it just has a different expected mean squares column:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Expected MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>SS[T]</td>
<td>( t - 1 )</td>
<td>MS[T]</td>
<td>( \sigma^2 + n\sigma^2_T )</td>
</tr>
<tr>
<td>Error</td>
<td>SS[E]</td>
<td>( N - t )</td>
<td>MS[E]</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>Total</td>
<td>SS[Tot]</td>
<td>( N - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Estimating parameters of one-way random effects model

\[ \hat{\mu} = \bar{y}_{++} \]
\[ \hat{\sigma}^2 = MS[E] \]
\[ \hat{\sigma}_T^2 = \frac{MS[T] - MS[E]}{n} \]

For sires data, \( \bar{y}_{++} = 82.6 \) and

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Expected MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sire</td>
<td>5591</td>
<td>4</td>
<td>1398</td>
<td>( \sigma^2 + 8\sigma_T^2 )</td>
</tr>
<tr>
<td>Error</td>
<td>16233</td>
<td>35</td>
<td>464</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>Total</td>
<td>21824</td>
<td>39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \hat{\mu} = 82.6 \]
\[ \hat{\sigma}^2 = 464 \text{ (lbs}^2\text{)} \]
\[ \hat{\sigma}_T^2 = \frac{1398 - 464}{8} \]
\[ = 117 \text{ (lbs}^2\text{)} \]

Specific questions pertaining to this study:

Consider the birthweight of a randomly sampled calf.

1. What is the estimated variance of such a calf?
2. Estimate how much of this variation is due to the sire effect.
3. Estimate how much of this variation is not due to the sire effect.

General questions:

1. Is it possible for an estimated variance component to be negative?
2. How?
3. What do you do in that case?
Other parameters of interest in random effects models

**Coefficient of variation** (CV):

\[
CV(Y_{ij}) = \frac{\sqrt{\text{Var}(Y_{ij})}}{|E(Y_{ij})|} = \frac{\sqrt{\sigma_T^2 + \sigma^2}}{|\mu|}
\]

Note: this is *not* estimated by **Coeff Var** in PROC GLM output.

**Intraclass correlation coefficient**

\[
\rho_I = \frac{\text{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{Var}(Y_{ij}) \text{Var}(Y_{ik})}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2}
\]

- Interpretation: the correlation between two responses receiving the same level of the random factor.
- Bigger values of \( \rho_I \) correspond to (bigger/smaller?) random treatment effects.

For sires,

\[
\hat{CV} = \frac{\sqrt{117+464}}{82.6} = 0.29
\]

\[
\hat{\rho}_I = \frac{117}{117+464} = 0.20
\]

Interpretations:
- The estimated standard deviation of a birthweight, 24.1 is 29% of the estimated mean birthweight, 82.6.
- The estimated correlation between any two calves with the same sire for a male parent, or the estimated *intrasire* correlation coefficient, is 0.20.
Using PROC GLM for random effects models

data one;
  input sire @;
  do i=1 to 8;
    input bw @; output;
  end;
cards;
  177 61 100 56 113 99 103 75 62
  200 75 102 95 103 98 115 98 94
  201 58 60 60 57 57 59 54 100
  202 57 56 67 59 58 121 101 101
  203 59 46 120 115 115 93 105 75;
run;

proc glm;
  class sire;
  model bw=sire;
  random sire;
run;

The GLM Procedure

Class Levels Values
  sire 5 177 200 201 202 203

Sum of Source DF Squares Mean Square F Value Pr > F
  Model 4 5591.15000 1397.78750 3.01 0.0309
  Error 35 16232.75000 463.79286
  Corrected Total 39 21823.90000

R-Square Coeff Var Root MSE bw Mean
  0.256194 26.08825 21.53585 82.55000

Source Type III Expected Mean Square
  sire Var(Error) + 8 Var(sire)

(\(\sigma^2 = \text{Var(Error)}\) and \(\sigma^2_T = \text{Var(sire)}\).)
Testing a variance component - \( H_0 : \sigma^2_T = 0 \)

Recall that \( \sigma^2_T = \text{Var}(T_i) \), the variance among the population of treatment effects.

\[
F = \frac{MS[T]}{MS[E]}
\]

reject \( H_0 \) at level \( \alpha \) if \( F > F(\alpha, t - 1, N - t) \)

For the sires,

\[
F = \frac{1398}{464} = 3.01 > 2.64 = F(0.05, 4, 35)
\]

so \( H_0 \) is rejected at \( \alpha = 0.05 \). (The \( p \)-value is 0.0309)

Q: “Isn’t this just just like the \( F \)-test for one-way ANOVA with fixed effects?”

A: “Yes.”
Interval Estimation of some model parameters

A 95% confidence interval for $\mu$ derived by consideration of $SE(\bar{Y}_{++})$:

$$\bar{Y}_{++} = \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{n} Y_{ij}$$

$$= \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{n} (\mu + T_i + E_{ij})$$

$$= \mu + \bar{T}_+ + \bar{E}_{++}$$

where $\bar{T}_+ = (T_1 + \cdots + T_t)/t$ and $\bar{E}_{++} = (\sum \sum E_{ij})/N$, so that

$$\text{Var}(\bar{Y}_{++}) = \text{Var}(\bar{T} + \bar{E}_{++})$$

$$= \frac{\sigma_T^2}{t} + \frac{\sigma^2}{nt}$$

$$= \frac{1}{nt} (n\sigma_T^2 + \sigma^2)$$

$$= \frac{1}{nt} E(MS[T]).$$

If the data are normally distributed, then

$$\frac{\bar{Y}_{++} - \mu}{\sqrt{\frac{MS[T]}{nt}}} \sim t_{t-1}$$

and a 95% confidence interval for $\mu$ given by

$$\bar{Y}_{++} \pm t(0.025, t-1)\sqrt{\frac{MS[T]}{nt}}$$

Sires data: $\bar{y}_{++} = 82.6$, $MS[T] = 1398$, $nt = 40$. Critical value $t(0.025, 4) = 2.78$ yields the interval

$$82.6 \pm 2.78(5.91) \text{ or } (66.1, 99.0).$$
Confidence interval for $\rho_I$

A 95% confidence interval for $\rho_I$ can be obtained from the expression

$$\frac{F_{\text{obs}} - F_{\alpha/2}}{F_{\text{obs}} + (n - 1)F_{\alpha/2}} < \rho_I < \frac{F_{\text{obs}} - F_{1-\alpha/2}}{F_{\text{obs}} + (n - 1)F_{1-\alpha/2}}$$

where $F_{\alpha/2} = F(\frac{\alpha}{2}, t - 1, N - t)$ and $F_{\text{obs}}$ is the observed $F$-ratio for treatment effect from the ANOVA table.

For the sires, $F_{\text{obs}} = 3.01$ and $F_{0.025} = 3.179, F_{0.975} = 0.119$. The formula gives $(-0.01, 0.75)$.

Note the asymmetry and disagreement with test of $H_0 : \sigma_T^2 = 0$

These formulas arrived at via some distributional results:

- $(t - 1)\frac{MS[T]}{\sigma^2 + n\sigma_T^2} \sim \chi^2_{t-1}$
- $(N - t)\frac{MS[E]}{\sigma^2} \sim \chi^2_{N-t}$
- $MS[T]$ and $MS[E]$ are independent
- Ratio of independent $\chi^2$ RVs divided by $df$ has an $F$ distribution
- $$\left( \frac{MS[T]}{\sigma^2 + n\sigma_T^2} \right) / \left( \frac{MS[E]}{\sigma^2} \right) \sim F_{t-1,N-t}$$
  (which explains the $F$ test for $H_0 : \sigma_t^2 = 0$)
- Rearranging the probability statement below

$$1-\alpha = \Pr \left( F(1 - \frac{\alpha}{2}, t - 1, N - t) < \frac{MS[T]}{\sigma^2 + n\sigma_T^2} < F(\frac{\alpha}{2}, t - 1, N - t) \right)$$

so that $\rho_I$ gets left in the middle yields the confidence interval yields the c.i. at the top o’ the page.
Using PROC MIXED for random effects models

```sas
proc mixed cl;
class sire;
model bw=;
random sire;
estimate "mean" intercept 1/cl;
run;
```

The SAS System
The Mixed Procedure
Model Information

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>bw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariance Structure</td>
<td>Variance Components</td>
</tr>
<tr>
<td>Estimation Method</td>
<td>REML</td>
</tr>
<tr>
<td>Residual Variance Method</td>
<td>Profile</td>
</tr>
<tr>
<td>Fixed Effects SE Method</td>
<td>Model-Based</td>
</tr>
<tr>
<td>Degrees of Freedom Method</td>
<td>Containment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>sire</td>
<td>5</td>
<td>177 200 201 202 203</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>sire</td>
<td>116.75</td>
<td>0.05</td>
<td>29.9707</td>
<td>7051.37</td>
</tr>
<tr>
<td>Residual</td>
<td>463.79</td>
<td>0.05</td>
<td>305.11</td>
<td>789.17</td>
</tr>
</tbody>
</table>

Estimates

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>82.5500</td>
<td>5.9114</td>
<td>4</td>
<td>13.96</td>
<td>0.0002</td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>66.1373</td>
<td>98.9627</td>
</tr>
</tbody>
</table>
More interval estimation for variance components

The estimated residual variance component for the sire data was \( \hat{\sigma}^2 = MS[E] = 464 \text{ lbs}^2 \).

A 95% confidence interval for this variance component is given by

\[
\left( \frac{(40 - 5)464}{53.2} < \sigma^2 < \frac{(40 - 5)464}{20.6} \right)
\]
or

\[
\left( \frac{35464}{53.2} < \sigma^2 < \frac{35464}{20.6} \right)
\]
or \((305.2, 789.5)\text{ lbs}^2\)

This can be derived using the distributional result

\[
(N - t) \frac{MS[E]}{\sigma^2} \sim \chi^2_{N-t}
\]

setting up the probability statement

\[
1 - \alpha = \Pr \left( \chi^2(1 - \frac{\alpha}{2}, N - t) < (N - t) \frac{MS[E]}{\sigma^2} < \chi^2(\frac{\alpha}{2}, N - t) \right)
\]

Rearranging to get \( \sigma^2 \) in the middle yields the 100(1\( - \alpha \)% confidence interval for \( \sigma^2 \):

\[
\left( \frac{(N - t)MS[E]}{\chi^2_{\alpha/2}}, \frac{(N - t)MS[E]}{\chi^2_{1-\alpha/2}} \right)
\]

Q: What are the mean and variance of the \( \chi^2_{35} \) distribution?
Interval estimation for $\sigma^2_T$

The estimated variance component for the random sire effect was $\hat{\sigma}^2_T = 117$.

Q: How can we get a 95% confidence interval for $\sigma^2_T$?

A: In a similar fashion, but the confidence level based on Satterthwaite’s approximation to the degrees of freedom of the linear combination of $MS$ terms:

$$\begin{pmatrix}
\hat{df}\hat{\sigma}^2_T & \hat{df}\hat{\sigma}^2_T \\
\chi^2_{\alpha/2,\hat{df}} & \chi^2_{1-\alpha/2,\hat{df}}
\end{pmatrix}$$

where

$$\hat{df} = \frac{(n\hat{\sigma}^2_T)^2}{\frac{MS[T]^2}{t-1} + \frac{MS[E]^2}{N-t}}$$

For the sire data,

$$\hat{df} = \frac{(8 \times 117)^2}{1398^2} + \frac{464^2}{35} = 1.76$$

Using the \texttt{CL} option in the \texttt{MIXED} statement will request this confidence interval and will use this approximation to $df$ and will not round to the nearest integer $df$:

$$\chi^2_{0.975,1.76} = 0.029, \quad \chi^2_{0.025,1.76} = 6.87$$

yielding the 95% confidence interval

$$\left(\frac{1.76(117)}{6.87}, \frac{1.76(117)}{0.29}\right)$$

or

$$(30, 7051)$$
Review of one-way random effects ANOVA

The model

\[ Y_{ij} = \mu_{\text{fixed}} + T_{i\text{random}} + E_{ij\text{random}} \]

for \( i = 1, 2, \ldots, t \) and \( j = 1, \ldots, n \)

with

\[ T_1, T_2, \ldots, T_t \sim iid N(0, \sigma_T^2) \]

independent of \( E_{11}, \ldots, E_{tn} \sim iid N(0, \sigma^2) \)

Remarks:

- \((T_1, T_2, \ldots \text{ randomly drawn from pop'n of treatment effects.})\)
- Only three parameters: \( \mu, \sigma, \sigma_T^2 \)
- Several functions of these parameters of interest
  - \( CV(Y) = \frac{\sqrt{\sigma^2 + \sigma_T^2}}{\mu} \)
  - \( \rho_I = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_T^2}{\sigma^2 + \sigma_T^2} \)
- Two observations from same treatment group not independent

Exercise: match up the formulas for confidence intervals below with their targets, \( \rho_I, \sigma^2, \sigma_T^2, \mu \):

\[ \bar{Y}_{++} \pm t(0.025, t - 1) \sqrt{\frac{MS[T]}{nt}} \]

\[ \left( \frac{F_{obs} - F_{1 - \alpha/2}}{F_{obs} + (n - 1)F_{1 - \alpha/2}} \right), \left( \frac{F_{obs} - F_{\alpha/2}}{F_{obs} + (n - 1)F_{\alpha/2}} \right), \left( \frac{\tilde{df} \tilde{\sigma}_T^2}{\chi^2_{\alpha/2, \tilde{df}}} \right), \left( \frac{\tilde{df} \tilde{\sigma}_T^2}{\chi^2_{1 - \alpha/2, \tilde{df}}} \right) \]

\[ \left( \frac{(N-t)MS[E]}{\chi^2_{\alpha/2}} \right), \left( \frac{(N-t)MS[E]}{\chi^2_{1 - \alpha/2}} \right) \]
Modelling factorial effects: fixed, or random?
A guide

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Levels</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- selected from conceptually ∞ popn of collection of levels</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>- finite number of possible levels</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>Another expt</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- would use same levels</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>- would involve new levels sampled from same popn</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>Goal</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- estimate varcomps</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>- estimate longrun means</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>Inference</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- for these levels used in this expt</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>- for the popn of levels</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
**Topic:** Mixed Models for factorial experiments  
**Reading:** Rao, Ch.14

Two-factor designs  
with factors that are fixed/random and nested/crossed

1. Entomologist records energy expended ($y$) by $N = 27$ honeybees  
   - at three TEMPERATURES (20, 30, 40°C)  
   - consuming three levels of SUCROSE (20%, 40%, 60%)

<table>
<thead>
<tr>
<th>Temp</th>
<th>Suc</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>3.1</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>5.5</td>
</tr>
<tr>
<td>20</td>
<td>60</td>
<td>7.9</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>6.0</td>
</tr>
<tr>
<td>30</td>
<td>40</td>
<td>11.5</td>
</tr>
<tr>
<td>30</td>
<td>60</td>
<td>17.5</td>
</tr>
<tr>
<td>40</td>
<td>20</td>
<td>7.7</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>15.7</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
<td>19.1</td>
</tr>
</tbody>
</table>

2. Experiment to study effect of drug and method of administration on fasting blood sugar in a random sample of $N = 18$ diabetic patients. (dataset on website is b1sugar.dat)  
   - First factor is drug: brand I tablet, brand II tablet, insulin injection  
   - Second factor is type of administration (see table)

<table>
<thead>
<tr>
<th>Drug ($i$)</th>
<th>Type of Administration ($j$)</th>
<th>Mean $\bar{y}_{ij}$</th>
<th>Variance $s^2_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>($i = 1$) Brand I tablet</td>
<td>($j = 1$)30mg × 1</td>
<td>15.7</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>($j = 2$)15mg × 2</td>
<td>19.7</td>
<td>9.3</td>
</tr>
<tr>
<td>($i = 2$) Brand II tablet</td>
<td>($j = 1$)20mg × 1</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>($j = 2$)10mg × 2</td>
<td>17.3</td>
<td>6.3</td>
</tr>
<tr>
<td>($i = 3$) Insulin injection</td>
<td>($j = 1$) before breakfast</td>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>($j = 2$) before supper</td>
<td>33</td>
<td>9</td>
</tr>
</tbody>
</table>
3. An experiment is conducted to determine variability among laboratories (interlaboratory differences) in their assessment of bacterial concentration in milk after pasteurization. Milk w/ various degrees of contamination was tested by randomly drawing four samples of milk from a collection of cartons at various stages of spoilage. \( Y \) is colony-forming units/\( \mu l \). Labs think they’re receiving 8 independent samples

<table>
<thead>
<tr>
<th>Lab</th>
<th>Sample 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2200</td>
<td>3000</td>
<td>210</td>
<td>270</td>
</tr>
<tr>
<td></td>
<td>2200</td>
<td>2900</td>
<td>200</td>
<td>260</td>
</tr>
<tr>
<td>2</td>
<td>2600</td>
<td>3600</td>
<td>290</td>
<td>360</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>3500</td>
<td>240</td>
<td>380</td>
</tr>
<tr>
<td>3</td>
<td>1900</td>
<td>2500</td>
<td>160</td>
<td>230</td>
</tr>
<tr>
<td></td>
<td>2100</td>
<td>2200</td>
<td>200</td>
<td>230</td>
</tr>
<tr>
<td>4</td>
<td>2600</td>
<td>2800</td>
<td>330</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>4300</td>
<td>1800</td>
<td>340</td>
<td>290</td>
</tr>
<tr>
<td>5</td>
<td>4000</td>
<td>4800</td>
<td>370</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>3900</td>
<td>4800</td>
<td>340</td>
<td>480</td>
</tr>
</tbody>
</table>

(Data from Oehlert, 2000)

4. An expt measures \textit{Campylobacter} counts in \( N = 120 \) chickens in a processing plant, at four locations, over three days. Means (std) for \( n = 10 \) chickens sampled at each location tabulated below:

<table>
<thead>
<tr>
<th>Day</th>
<th>Location</th>
<th>Before Washer</th>
<th>After Washer</th>
<th>After mic. rinse</th>
<th>After chill tank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>70070.00</td>
<td>48310.00</td>
<td>12020.00</td>
<td>11790.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(79034.49)</td>
<td>(34166.80)</td>
<td>(3807.24)</td>
<td>(7832.05)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>75890.00</td>
<td>52020.00</td>
<td>8090.00</td>
<td>8690.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(74551.32)</td>
<td>(17686.27)</td>
<td>(4848.01)</td>
<td>(5526.19)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>95260.00</td>
<td>33170.00</td>
<td>6200.00</td>
<td>8370.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(03176.00)</td>
<td>(22259.08)</td>
<td>(5028.81)</td>
<td>(5720.15)</td>
</tr>
</tbody>
</table>

Data courtesy of Michael Bashor, General Mills

Transformation?
5. An experiment to assess the variability of a particular acid among plants and among leaves of plants:

<table>
<thead>
<tr>
<th>Plant $i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaf $j$</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$k = 1$</td>
<td>11.2</td>
<td>16.5</td>
<td>18.3</td>
</tr>
<tr>
<td>$k = 2$</td>
<td>11.6</td>
<td>16.8</td>
<td>18.7</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>12.0</td>
<td>16.1</td>
<td>19.0</td>
</tr>
</tbody>
</table>

Data from Neter, et al (1996)

6. Plantheights from 10 pots (not 2!) randomized to 5 treatment combinations. (See Table 14.2 from Rao.)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Dark</th>
<th>Source</th>
<th>Intensity</th>
<th>Pot</th>
<th>Seedling 1</th>
<th>Seedling 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD</td>
<td>1</td>
<td>D</td>
<td>D</td>
<td>1</td>
<td>32.94</td>
<td>35.98</td>
</tr>
<tr>
<td>DD</td>
<td>1</td>
<td>D</td>
<td>D</td>
<td>2</td>
<td>34.76</td>
<td>32.40</td>
</tr>
<tr>
<td>AL</td>
<td>0</td>
<td>A</td>
<td>L</td>
<td>1</td>
<td>30.55</td>
<td>32.64</td>
</tr>
<tr>
<td>AL</td>
<td>0</td>
<td>A</td>
<td>L</td>
<td>2</td>
<td>32.37</td>
<td>32.04</td>
</tr>
<tr>
<td>AH</td>
<td>0</td>
<td>A</td>
<td>H</td>
<td>1</td>
<td>31.23</td>
<td>31.09</td>
</tr>
<tr>
<td>AH</td>
<td>0</td>
<td>A</td>
<td>H</td>
<td>2</td>
<td>30.62</td>
<td>30.42</td>
</tr>
<tr>
<td>BL</td>
<td>0</td>
<td>B</td>
<td>L</td>
<td>1</td>
<td>34.41</td>
<td>34.88</td>
</tr>
<tr>
<td>BL</td>
<td>0</td>
<td>B</td>
<td>L</td>
<td>2</td>
<td>34.07</td>
<td>33.87</td>
</tr>
<tr>
<td>BH</td>
<td>0</td>
<td>B</td>
<td>H</td>
<td>1</td>
<td>35.61</td>
<td>35.00</td>
</tr>
<tr>
<td>BH</td>
<td>0</td>
<td>B</td>
<td>H</td>
<td>2</td>
<td>33.65</td>
<td>32.91</td>
</tr>
</tbody>
</table>
### Six types of two-factor models

Fixed and/or random effects that are either crossed or nested

1. \( Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + E_{ijk} \) | crossed/random
2. \( Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + E_{ijk} \) | nested/fixed
3. \( Y_{ijk} = \mu + A_i + B_{j(i)} + E_{ijk} \) | nested/random
4. \( Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk} \) | crossed/mixed
5. \( Y_{ijk} = \mu + \alpha_i + B_{j(i)} + E_{ijk} \) | nested/mixed
6. \( Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + E_{ijk} \) | crossed/fixed

In the models above, (which are not ordered according to the six prior datasets)

- GREEK symbols parameterize FIXED, unknown treatment means
- CAPITAL letters represent RANDOM effects
- for Model 1, \( \sum \alpha_i = \sum \beta_j = \sum_i (\alpha \beta)_{ij} = \sum_j (\alpha \beta)_{ij} \equiv 0 \)
- for Model 2, \( \sum \alpha_i = \sum_j \beta_{j(i)} \equiv 0 \)
- for Model 3, \( A_i, B_j, (AB)_{ij} \) are all independent
- for Model 4, \( \sum \alpha_i = 0 \) and \( B_j, (\alpha B)_{ij} \) are all independent
- for Model 5, \( A_i, B_{j(i)} \) are all independent
- for Model 6, \( \sum \alpha_i = 0 \)

Recall

- **RANDOM effects are used when it makes sense to think of LEVELS of factor as random sample from a population.**
Identifying the appropriate model for our 6 examples:

1. Energy expended by honeybees.
   - First factor:
   - Second factor:
   - Fixed or random?
   - Crossed or nested?
   - Model:
     \[ Y_{ijk} = \mu + E_{ijk} \]

2. Change in fasting blood sugar for diabetics
   - First factor:
   - Second factor:
   - Fixed or random?
   - Crossed or nested?
   - Model:
     \[ Y_{ijk} = \mu + E_{ijk} \]

3. Measuring bacterial concentration in milk
   - First factor:
   - Second factor:
   - Fixed or random?
   - Crossed or nested?
   - Model:
     \[ Y_{ijk} = \mu + E_{ijk} \]
4. Measuring bacteria counts in chickens at processing plant

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model:

\[ Y_{ijk} = \mu + E_{ijk} \]

5. Acids in leaves of plants

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model:

\[ Y_{ijk} = \mu + E_{ijk} \]

6. Effect of light source and intensity on plant heights (Rao Table 14.2)

- First factor:
- Second factor:
- Fixed or random?
- Crossed or nested?
- Model:

\[ Y_{ijk} = \mu + E_{ijk} \]
Tables of expected means squares (EMS): (see Rao table 14.1) (compare w/ OL tables 17.10, 17.17 and 17.31)

When factors $A$ and $B$ are CROSSED, and no sum-to-zero assumptions are made on random effects, expected means associated with sums of squares are given in the table below:

<table>
<thead>
<tr>
<th>Source</th>
<th>$df$</th>
<th>$A, B$ fixed</th>
<th>$A, B$ random</th>
<th>$A$ fixed $B$ random</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$a - 1$</td>
<td>$\sigma^2 + nb\psi^2_A$</td>
<td>$\sigma^2 + nb\sigma^2_A + n\sigma^2_{AB}$</td>
<td>$\sigma^2 + nb\psi^2_A + n\sigma^2_{AB}$</td>
</tr>
<tr>
<td>$B$</td>
<td>$b - 1$</td>
<td>$\sigma^2 + na\psi^2_B$</td>
<td>$\sigma^2 + na\sigma^2_B + n\sigma^2_{AB}$</td>
<td>$\sigma^2 + na\sigma^2_B + n\sigma^2_{AB}$</td>
</tr>
<tr>
<td>$AB$</td>
<td>$(a - 1) \times (b - 1)$</td>
<td>$\sigma^2 + n\psi^2_{AB}$</td>
<td>$\sigma^2 + n\sigma^2_{AB}$</td>
<td>$\sigma^2 + n\sigma^2_{AB}$</td>
</tr>
<tr>
<td>Error</td>
<td>$ab(n - 1)$</td>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>

When factor $B$ is NESTED in factor $A$, expected means associated with sums of squares are given in the table below:

<table>
<thead>
<tr>
<th>Source</th>
<th>$df$</th>
<th>$A, B$ fixed</th>
<th>$A, B$ random</th>
<th>$A$ fixed $B$ random</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$a - 1$</td>
<td>$\sigma^2 + nb\psi^2_A$</td>
<td>$\sigma^2 + nb\sigma^2_A + n\sigma^2_{B(A)}$</td>
<td>$\sigma^2 + nb\psi^2_A + n\sigma^2_{B(A)}$</td>
</tr>
<tr>
<td>$B(A)$</td>
<td>$a(b - 1)$</td>
<td>$\sigma^2 + n\psi^2_{B(A)}$</td>
<td>$\sigma^2 + n\sigma^2_{B(A)}$</td>
<td>$\sigma^2 + n\sigma^2_{B(A)}$</td>
</tr>
<tr>
<td>Error</td>
<td>$ab(n - 1)$</td>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>

where $\psi^2$ and $\sigma^2$ values are defined on the next page.

Help with computing expected mean squares (without sum-to-zero assumptions on random effects)

1. If a factor $X$ with index $i$ is random then $EMS(X)$ is a linear combo of $\sigma^2$ and varcomps for all random effects containing index $i$. Coefficients for varcomps are limits of indexes NOT listed (summed over) in random effects.

2. If a factor $X$ is fixed. Treat it like it is random and then just replace the varcomp for X with the effect size, $\psi^2_X$. 

\[
\psi^2_A = \frac{1}{a-1} \sum_{i=1}^{a} \alpha_i^2 \quad \text{effect size of factor } A
\]
\[
\psi^2_B = \frac{1}{b-1} \sum_{i=1}^{b} \beta_i^2 \quad \text{effect size of factor } B
\]
\[
\psi^2_{AB} = \frac{1}{(a-1)(b-1)} \sum_{i=1}^{a} \sum_{j=1}^{b} (\alpha \beta)_{ij}^2 \quad \text{effect size of interaction}
\]
\[
\psi^2_{B(A)} = \frac{1}{a(b-1)} \sum_{i=1}^{a} \sum_{j=1}^{b} \beta_{j(i)}^2 \quad \text{effect size of factor } B
\]
\[
\sigma_A^2 = \text{Var}(A_i) \quad \text{variance component for factor } A
\]
\[
\sigma_B^2 = \text{Var}(B_i) \quad \text{variance component for factor } B
\]
\[
\sigma_{AB}^2 = \text{Var}((AB)_{ij}) \quad \text{variance component for interaction}
\]
\[
\sigma_{B(A)}^2 = \text{Var}(B_{j(i)}) \quad \text{variance component for factor } B
\]
\[
\sigma^2 = \text{Var}(E_{ijk}) \quad \text{error variance}
\]

The term *effect size* is often used in power considerations and sometimes involves division by \( \sigma^2 \).
Using expected mean squares to analyze data in mixed models

- EMS tables dictate which $F$-ratios test which effects
- EMS tables yield estimating equations for variance components

**Milk example (p.190):** $F$-tests and estimating variance components.

1. To test for interaction effect, use $F_{AB} = \frac{MS[AB]}{MS[E]}$
2. To test for main effect of A, use $F_A = \frac{MS[A]}{MS[AB]}$
3. To test for main effect of B, use $F_B = \frac{MS[B]}{MS[AB]}$

Note the departure from fixed effects analysis, where $MS[E]$ is always used in the denominator.

The SAS System

The GLM Procedure

Dependent Variable: ly = log(y)

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>19</td>
<td>56.03510844</td>
<td>2.94921623</td>
<td>191.44</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sample</td>
<td>3</td>
<td>53.18978788</td>
<td>17.72992929</td>
<td>1150.89</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>lab</td>
<td>4</td>
<td>2.30248803</td>
<td>0.57562201</td>
<td>37.37</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>sample*lab</td>
<td>12</td>
<td>0.54283253</td>
<td>0.04523604</td>
<td>2.94</td>
<td>0.0161</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>0.30810726</td>
<td>0.01540536</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>39</td>
<td>56.34321569</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The wrong $F$-ratio and $p$-value for testing for random LAB (A) effect:

$$F = \frac{MS[A]}{MS[E]} = \frac{0.5756}{0.0154} = 37.37(p < 0.0001)$$

The correct $F$-ratio and $p$-value for testing for random LAB (A) effect:

$$F = \frac{MS[A]}{MS[AB]} = \frac{0.5756}{0.0452} = 12.72(p = 0.0003)$$
Estimating variance components

The estimated variance components satisfy the following system of equations:

\[
\begin{align*}
MS[E] &= \hat{\sigma}^2 \\
MS[AB] &= \hat{\sigma}^2 + n\hat{\sigma}_{AB}^2 \\
&= \hat{\sigma}^2 + 2\hat{\sigma}_{AB}^2 \\
MS[A] &= \hat{\sigma}^2 + nb\hat{\sigma}_A^2 + n\hat{\sigma}_{AB}^2 \\
&= \hat{\sigma}^2 + 8\hat{\sigma}_A^2 + 2\hat{\sigma}_{AB}^2 \\
MS[B] &= \hat{\sigma}^2 + na\hat{\sigma}_B^2 + n\hat{\sigma}_{AB}^2 \\
&= \hat{\sigma}^2 + 10\hat{\sigma}_B^2 + 2\hat{\sigma}_{AB}^2
\end{align*}
\]

Substitution of

\[
\begin{align*}
MS[E] &= 0.0154 \\
MS[AB] &= 0.0452 \\
MS[A] &= 0.5756 \\
MS[B] &= 17.7299
\end{align*}
\]

into the system of equations yields estimated variance components:

\[
\begin{align*}
\hat{\sigma}^2 &= MS[E] = 0.0154 \\
\hat{\sigma}_{AB}^2 &= \frac{MS[AB] - MS[E]}{n} = \frac{0.0452 - 0.0154}{2} = 0.01492 \\
\hat{\sigma}_A^2 &= \frac{MS[A] - MS[AB]}{nb} = \frac{0.5756 - 0.0452}{8} = 0.0663 \\
\hat{\sigma}_B^2 &= \frac{MS[B] - MS[AB]}{na} = \frac{17.7299 - 0.0452}{10} = 1.768
\end{align*}
\]
data one;
  infile "milk.dat" firstobs=4;
  input sample lab y;
  ly=log(y);
run;

proc glm;
  class lab sample;
  model ly=sample|lab;
  random sample lab sample*lab;
  test h=lab sample e=sample*lab;
  lsmeans sample*lab;
run;

The GLM Procedure

Dependent Variable: ly

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample</td>
<td>3</td>
<td>53.18978788</td>
<td>17.72992929</td>
<td>1150.89</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>lab</td>
<td>4</td>
<td>2.30248803</td>
<td>0.57562201</td>
<td>37.37</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>lab*sample</td>
<td>12</td>
<td>0.54283253</td>
<td>0.04523604</td>
<td>2.94</td>
<td>0.0161</td>
</tr>
</tbody>
</table>

Tests of Hypotheses Using the Type III MS for lab*sample as an Error Term

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>lab</td>
<td>4</td>
<td>2.30248803</td>
<td>0.57562201</td>
<td>12.72</td>
<td>0.0003</td>
</tr>
<tr>
<td>sample</td>
<td>3</td>
<td>53.18978788</td>
<td>17.72992929</td>
<td>391.94</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
proc varcomp;
  class sample lab;
  model y=sample|lab;
run;

Variance Components Estimation Procedure

<table>
<thead>
<tr>
<th>Variance Component</th>
<th>ly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var(sample)</td>
<td>1.76847</td>
</tr>
<tr>
<td>Var(lab)</td>
<td>0.06630</td>
</tr>
<tr>
<td>Var(sample*lab)</td>
<td>0.01492</td>
</tr>
<tr>
<td>Var(Error)</td>
<td>0.01541</td>
</tr>
</tbody>
</table>

Q: At the end of the day, what is the conclusion from the analysis of this crossed, random effects experiment?

- There is evidence of variability due to laboratory $\times$ sample interaction; interlaboratory effects vary by sample.
- The estimated parameters ($\mu +$ variance components) of the model
  \[ Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + E_{ijk} \]
  are
  \[ \hat{\sigma}^2 = 0.0154 \]
  \[ \hat{\sigma}_{AB}^2 = 0.0149 \]
  \[ \hat{\sigma}_A^2 = 0.0663 \]
  \[ \hat{\sigma}_B^2 = 1.7685 \]
  \[ \hat{\mu} = 6.82 \text{(log scale)} \]

- The standard error of $\bar{Y}_{+++}$ can be derived by
  \[ \bar{Y}_{+++} = \mu + \bar{A}_+ + \bar{B}_+ + \overline{(AB)}_{++} + \bar{E}_{+++} \]
  \[ \text{Var}(\bar{Y}_{+++}) = \text{Var}(\bar{A}_+) + \text{Var}(\bar{B}_+) + \text{Var}(\overline{(AB)}_{++}) + \text{Var}(\bar{E}_{+++}) \]
  \[ = \frac{\sigma_A^2}{a} + \frac{\sigma_B^2}{b} + \frac{\sigma_{AB}^2}{ab} + \frac{\sigma^2}{abn} \]
Estimation of standard error and approximation of \( df \)

The standard error

\[
SE(\bar{Y}_{+++}) = \sqrt{\frac{\sigma_A^2}{a} + \frac{\sigma_B^2}{b} + \frac{\sigma_{AB}^2}{ab} + \frac{\sigma^2}{abn}}
\]

can be estimated by substitution of estimated variance components \((\hat{\sigma}^2)\), which leads to

\[
\hat{SE}(\bar{Y}_{+++}) = \sqrt{\frac{\hat{\sigma}_A^2}{a} + \frac{\hat{\sigma}_B^2}{b} + \frac{\hat{\sigma}_{AB}^2}{ab} + \frac{\hat{\sigma}^2}{abn}}
\]

\[= \text{lots of algebra and cancellations} \]

\[= \sqrt{\frac{1}{nab} (MS[A] + MS[B] - MS[AB])} \]

For the milk data, we have

\[
\hat{SE}(\bar{Y}_{+++}) = \sqrt{\frac{1}{40} (0.58 + 17.73 - 0.05)} = 0.6757
\]

For a 95\% confidence interval, we have a problem: we don’t know how many \( df \) are associated with a \( t \) statistic based on this estimated \( SE \).
ST511 Flashback:
Unequal variances independent samples $t$-test:

Example: Suspended particulate matter $Y$ (in micrograms per cubic meter) in homes with smokers ($Y_1$) and without smokers ($Y_2$):

<table>
<thead>
<tr>
<th>smokers</th>
<th>133</th>
<th>128</th>
<th>136</th>
<th>135</th>
<th>131</th>
<th>131</th>
<th>130</th>
<th>131</th>
<th>132</th>
<th>147</th>
</tr>
</thead>
<tbody>
<tr>
<td>no smokers</td>
<td>106</td>
<td>85</td>
<td>84</td>
<td>95</td>
<td>104</td>
<td>79</td>
<td>72</td>
<td>115</td>
<td>95</td>
<td></td>
</tr>
</tbody>
</table>

Summary statistics:

$\bar{y}_1 = 133.2$, $s_1^2 = 26.0$

$\bar{y}_2 = 92.8$, $s_2^2 = 195.4$

$n_1 = 11$, $n_2 = 9$

Assumptions:

- $Y_{11}, \ldots, Y_{1n_1}$ and $Y_{21}, \ldots, Y_{2n_2}$ are independent random samples from normal distributions with unknown $\mu_1, \mu_2, \sigma_1, \sigma_2$ and $\sigma_1^2 \neq \sigma_2^2$. Note the large difference in the sample variances.

$H_0 : \mu_1 - \mu_2 = 0$ v. $H_1 : \mu_1 - \mu_2 \neq 0$

Consider the test statistic

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

For small $n_1, n_2$, this quantity does not have the standard normal distribution. nor does the version where $S_p^2$ is used in the denominator. An approximate solution is to use the student $t$ distribution with $df$ approximated by the Satterthwaite approximation:

$$\hat{df} = \frac{(c_1 MS_1 + c_2 MS_2)^2}{(c_1 MS_1)^2/df_1 + (c_2 MS_2)^2/df_2}$$

where $MS_i = S_i^2$ and $c_i = 1/n_i$. 

ST511 Flashback continued:
For the air pollution in homes with a smoking occupant data, \( c_1 M S_1 = 26/11 = 2.36, c_2 M S_2 = 195.4/9 = 21.71 \) and

\[
\hat{df} = \frac{(2.36 + 21.71)^2}{\frac{2.36^2}{10} + \frac{21.71^2}{8}} = 9.74
\]

The 97.5\(^{th}\) percentile of the \( t \) distribution with \( df = 9.74 \) is \( t(0.025, 9.74) = 2.236 \)

A 95\% confidence interval for the mean difference between homes with and without a smoking occupant, \( \mu_1 - \mu_2 \) is given by

\[
133.2 - 92.8 \pm 2.236\sqrt{26/11 + 195.4/9}
\]

or

\[
40.4 \pm 2.236(4.91)
\]

or

\[
40.4 \pm 10.97
\]

or

\[
(29.4, 51.4)
\]

These data would lead to the rejection of \( H_0 : \mu_1 = \mu_2 = 0 \) versus the two-tailed alternative. The observed test statistic is given by

\[
t_{obs} = \frac{133.2 - 92.8}{\sqrt{26/11 + 195.4/9}} = \frac{40.4}{4.91} = 8.2 \ (p < 0.0001)
\]

This problem aka the Behrens-Fisher problem.
data one;
  infile "smokers.dat" firstobs=2;
  input y smoke;
  label y="suspended particulate matter";
run;

proc ttest;
  class smoke;
  var y;
run;

The SAS System 1
The TTEST Procedure
Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>smoke</th>
<th>N</th>
<th>Lower CL Mean</th>
<th>Upper CL Mean</th>
<th>Lower CL Mean</th>
<th>Upper CL Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>0</td>
<td>9</td>
<td>82.032</td>
<td>92.778</td>
<td>103.52</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>1</td>
<td>11</td>
<td>129.76</td>
<td>133.18</td>
<td>136.6</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>Diff (1-2)</td>
<td>-49.91</td>
<td>-40.4</td>
<td>-30.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>smoke</th>
<th>Std Dev</th>
<th>Std Dev</th>
<th>Std Dev</th>
<th>Std Dev</th>
<th>Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>0</td>
<td>9.443</td>
<td>13.98</td>
<td>26.783</td>
<td>4.66</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>1</td>
<td>3.5603</td>
<td>5.0955</td>
<td>8.9422</td>
<td>1.5363</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>Diff (1-2)</td>
<td>7.6046</td>
<td>10.064</td>
<td>14.883</td>
<td>4.5235</td>
<td></td>
</tr>
</tbody>
</table>

T-Tests

| Variable | Method | Variances | DF | t Value | Pr > |t| |
|----------|--------|-----------|----|---------|------|---|
| y        | Pooled | Equal     | 18 | -8.93   | <.0001|
| y        | Satterthwaite | Unequal | 9.74* | -8.23* | <.0001|

Equality of Variances

<table>
<thead>
<tr>
<th>Variable</th>
<th>Method</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>Folded F</td>
<td>8</td>
<td>10</td>
<td>7.53</td>
<td>0.0045</td>
</tr>
</tbody>
</table>
The two-way random effects model for milk data

Satterthwaite’s approximation (cont’d)

To approximate the \( df \) associated with a \( t \) statistic based on a standard error of the form

\[
\sqrt{c_1 MS_1 + c_2 MS_2 + \cdots + c_k MS_k}
\]

(a linear combination of mean square terms), use the Satterthwaite approximation:

\[
\hat{df} = \frac{(c_1 MS_1 + c_2 MS_2 + \cdots + c_k MS_k)^2}{(c_1 MS_1)^2/df_1 + (c_2 MS_2)^2/df_2 + \cdots + (c_k MS_k)^2/df_k}
\]

Recall that for the milk data, we have

\[
\hat{SE}(\bar{Y}_{+++}) = \sqrt{\frac{1}{40}(MS[A] + MS[B] - MS[AB])}
\]

\[
= \sqrt{\frac{1}{40}(0.58 + 17.73 - 0.05)}
\]

\[
= 0.6757
\]

The degrees of freedom associated with this linear combination is approximated by

\[
\hat{df} = \frac{(0.6757)^4}{\left(\frac{1}{40}(17.73)^2/3 + \left(\frac{1}{40}0.58)^2/4 + \left(\frac{1}{40}0.045)^2/12 \right) \right)} = 3.18
\]

Using \( t(0.025, 3.18) = 3.08 \), a 95% confidence interval for the mean \( \mu \) among the population of all labs and samples is given by

\[
6.82 \pm 3.08(0.6757)
\]

or

\[
6.82 \pm 2.08
\]

(log scale)
data one;
  infile "milk.dat" firstobs=4;
  input sample lab y;
  ly=log(y);
run;

proc mixed cl;
  class sample lab;
  model ly=/s ddfm=satterth cl;
  random sample lab sample*lab;
run;

The SAS System
The Mixed Procedure

Model Information

  Dependent Variable     ly
  Covariance Structure   Variance Components
  Estimation Method      REML
  Residual Variance Method Profile
  Fixed Effects SE Method Model-Based
  Degrees of Freedom Method Satterthwaite

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample</td>
<td>1.7685</td>
<td>0.05</td>
<td>0.5664</td>
<td>24.8486</td>
</tr>
<tr>
<td>lab</td>
<td>0.06630</td>
<td>0.05</td>
<td>0.02233</td>
<td>0.7260</td>
</tr>
<tr>
<td>sample*lab</td>
<td>0.01492</td>
<td>0.05</td>
<td>0.005761</td>
<td>0.09261</td>
</tr>
<tr>
<td>Residual</td>
<td>0.01541</td>
<td>0.05</td>
<td>0.009017</td>
<td>0.03213</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect         | Estimate | Error | DF | t Value | Pr > |t| Alpha |
|----------------|----------|-------|----|---------|-------|-------|
| Intercept      | 6.8156   | 0.6757| 3.18| 10.09   | 0.0016| 0.05  |

<table>
<thead>
<tr>
<th>Effect</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.7325</td>
<td>8.8987</td>
</tr>
</tbody>
</table>


milk.data <- read.table("milk.dat",skip=3,col.names=c("sample","lab","bacteria"))
attach(milk.data)
postscript(file="milkplot1.ps")
par(mfrow=c(2,1)) # A 2x1 template (for two plots in a single column)
plot(x=sample,y=bacteria,pch=lab)
title("untransformed data")
legend(2.5,4400,pch=1:5,legend=c("Lab 1","Lab 2","Lab 3","Lab 4","Lab 5"))

plot(x=sample,y=log(bacteria),pch=lab)
title("log transformed data")
postscript()
Interaction plot for milk - raw counts

Interaction plot for milk - log(counts)
A nested design

Experiment to study effect of drug and method of administration on fasting blood sugar in diabetic patients

- First factor is drug: brand I tablet, brand II tablet, insulin injection
- Second factor is type of administration (see table)

<table>
<thead>
<tr>
<th>Drug Type</th>
<th>Type of Administration</th>
<th>Mean</th>
<th>Variance</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand I tablet</td>
<td>30mg × 1</td>
<td>15.7</td>
<td>6.3</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td>15mg × 2</td>
<td>19.7</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>Brand II tablet</td>
<td>20mg × 1</td>
<td>20</td>
<td>1</td>
<td>18.7</td>
</tr>
<tr>
<td></td>
<td>10mg × 2</td>
<td>17.3</td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>Insulin injection</td>
<td>before breakfast</td>
<td>28</td>
<td>4</td>
<td>30.5</td>
</tr>
<tr>
<td></td>
<td>before supper</td>
<td>33</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

(This is exercise 13.35. Grand mean is $\bar{y}_{+++} = 22.3$.)

Definition: Factor $B$ is [nested] in factor $A$ if there is a new set of levels of factor $B$ for every different level of factor $A$. 
Analysis of variance in nested designs

Consider a two-factor design in which factor $B$ is nested in factor $A$. Let $Y_{ijk}$ denote the $k^{th}$ response at level $j$ of factor $B$ within level $i$ of factor $A$. A model:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + E_{ijk}$$

for $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$, $k = 1, 2, \ldots, n$

$SS[Tot]$ can be broken down into components reflecting variability due to $A$, $B(A)$ and variability not due to either factor ($SS[E]$):


$$SS[A] = \sum_{i} \sum_{j} \sum_{k} (\bar{y}_{i++} - \bar{y}_{+++})^2$$

$$SS[B(A)] = \sum_{i} \sum_{j} \sum_{k} (\bar{y}_{ij+} - \bar{y}_{i+})^2$$

$$SS[E] = \sum_{i} \sum_{j} \sum_{k} (y_{ijk} - \bar{y}_{ij+})^2$$

The ANOVA table looks like

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$a - 1$</td>
<td>$SS[A]$</td>
<td>$MS[A] = \frac{SS[T]}{(a-1)}$</td>
<td>$F_A = \frac{MS[A]}{MS[E]}$</td>
</tr>
<tr>
<td>$B(A)$</td>
<td>$\sum_{i}(b_i - 1)$</td>
<td>$SS[B(A)]$</td>
<td>$MS(B(A)) = \frac{SS[B(A)]}{\sum_{i}(b_i - 1)}$</td>
<td>$F_{B(A)} = \frac{MS[B(A)]}{MS[E]}$</td>
</tr>
<tr>
<td>Error</td>
<td>$N - \sum b_i$</td>
<td>$SS[E]$</td>
<td>$MS[E] = \frac{SS[E]}{(N-t)}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$N - 1$</td>
<td>$SS[TOT]$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If $b_1 = b_2 = \cdots b_a = b$ then $\sum (b_i - 1) = a(b - 1)$ and $df_E = ab(n - 1)$. 
Inference from nested designs

To test \( H_0 : \alpha_i \equiv 0 \), use \( F_A \) on \( a - 1 \) and \( df_E \) degrees of freedom.

To test \( H_0 : \beta_{j(i)} \equiv 0 \), for all \( i, j \), use \( F_{B(A)} \) on \( \sum (b_i - 1) \) and \( df_E \) degrees of freedom.

For the diabetics blood sugar data, with \( \bar{y}_{+++} = 22.3 \) and means

<table>
<thead>
<tr>
<th>Drug (i)</th>
<th>Type of Administration (j)</th>
<th>Mean</th>
<th>Variance</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand I tablet</td>
<td>30mg ( \times 1 )</td>
<td>15.7</td>
<td>6.3</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td>15mg ( \times 2 )</td>
<td>19.7</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>Brand II tablet</td>
<td>20mg ( \times 1 )</td>
<td>20</td>
<td>1</td>
<td>18.7</td>
</tr>
<tr>
<td></td>
<td>10mg ( \times 2 )</td>
<td>17.3</td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>Insulin injection</td>
<td>before breakfast</td>
<td>28</td>
<td>4</td>
<td>30.5</td>
</tr>
<tr>
<td></td>
<td>before supper</td>
<td>33</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

\[
SS[A] = 2(3)[(17.7 - 22.3)^2 + (18.7 - 22.3)^2 + (30.5 - 22.3)^2] = 611.4
\]

\[
SS[B(A)] = 3[(15.7 - 17.7)^2 + (19.7 - 17.7)^2 + (20.0 - 18.7)^2 + (17.3 - 18.7)^2 + (28 - 30.5)^2 + (33 - 30.5)^2] = 72.2
\]

\[
SS[E] = 72
\]

Q1: How many \( df \) associated with \( SS[A] \)?
Q2: How many \( df \) associated with \( SS[B(A)] \)?
Q3: How many \( df \) associated with \( SS[E] \)?
data one;
  infile "blsugar.dat" firstobs=2 dlm='09'x;
  input a b rep y;
  drug=a; admin=b;
run;

proc glm;
  class a b;
  model y=a b(a);
  output out=two p=p r=r;
  means a b(a)/lsd;
  estimate "effect of B within A=1" b(a) -1 1;
  estimate "effect of B within A=2" b(a) 0 0 -1 1;
  estimate "effect of B within A=3" b(a) 0 0 0 -1 1;
  estimate "A=1 mean - A=2 mean" a 1 -1;
  estimate "A=1 mean - A=3 mean" a 1 0 -1;
  estimate "A=2 mean - A=3 mean" a 0 1 -1;
run;

The GLM Procedure

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>683.6111111</td>
<td>136.7222222</td>
<td>22.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>72.0000000</td>
<td>6.0000000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>17</td>
<td>755.6111111</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>Coeff Var</th>
<th>Root MSE</th>
<th>y Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.904713</td>
<td>10.99522</td>
<td>2.449490</td>
<td>22.27778</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>611.4444444</td>
<td>305.7222222</td>
<td>50.95</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>b(a)</td>
<td>3</td>
<td>72.1666667</td>
<td>24.0555556</td>
<td>4.01</td>
<td>0.0344</td>
</tr>
</tbody>
</table>

| Parameter                               | Estimate | Error | t Value | Pr > |t| |
|-----------------------------------------|----------|-------|---------|------|---|
| effect of B within A=1                  | 4.0000000 | 2.0000000 | 2.00   | 0.0687 |
| effect of B within A=2                  | -2.6666667 | 2.0000000 | -1.33  | 0.2072 |
| effect of B within A=3                  | 5.0000000 | 2.0000000 | 2.50   | 0.0279 |
| A=1 mean - A=2 mean                     | -1.0000000 | 1.41421356 | -0.71  | 0.4930 |
| A=1 mean - A=3 mean                     | -12.8333333 | 1.41421356 | -9.07  | <.0001 |
| A=2 mean - A=3 mean                     | -11.8333333 | 1.41421356 | -8.37  | <.0001 |
Conclusions?

• The administration effect $B$ (nested in the type of drug effect $A$) is statistically significant ($p = 0.0344$). This is due mostly to the before breakfast/supper difference, which is estimated to be

$$\bar{y}_{32+} - \bar{y}_{31+} = 5 \text{mg/dl}$$

with an (estimated) standard error of $SE = 2 = ?$.

• The effect of type of drug (factor $A$) is highly significant ($p < 0.0001$). Unadjusted pairwise comparisons indicate that the insulin injections yield greater changes, on average, in blood sugar than either pill and the mean changes brought by the pills don’t differ significantly.

• The following contrasts may be of interest:

$$\theta_1 = \mu_{1(3)} - \frac{1}{4}(\mu_{1(1)} + \mu_{2(1)} + \mu_{1(2)} + \mu_{2(2)})$$

$$\theta_2 = \mu_{2(3)} - \frac{1}{4}(\mu_{1(1)} + \mu_{2(1)} + \mu_{1(2)} + \mu_{2(2)})$$

Exercise: Estimate them and test their significance ($H_0 : \theta_i = 0$).
More Two-factor mixed models (Ch. 14)

- Expt measures *campylobacter* counts in \( N = 120 \) chickens in a processing plant
  - Crossed design with two factors
    * Location (4 levels)
    * Day (3 levels)
  - \( 4 \times 3 \) layout, \( n = 10 \) chickens per combo

<table>
<thead>
<tr>
<th>Day</th>
<th>Location</th>
<th>Before Washer</th>
<th>After Washer</th>
<th>After mic. rinse</th>
<th>After chill tank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Before Washer</td>
<td>70070.00</td>
<td>48310.00</td>
<td>12020.00</td>
<td>11790.00</td>
</tr>
<tr>
<td></td>
<td>(79034.49)</td>
<td>(34166.80)</td>
<td>(3807.24)</td>
<td>(7832.05)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Before Washer</td>
<td>75890.00</td>
<td>52020.00</td>
<td>8090.00</td>
<td>8690.00</td>
</tr>
<tr>
<td></td>
<td>(74551.32)</td>
<td>(17686.27)</td>
<td>(4848.01)</td>
<td>(5526.19)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Before Washer</td>
<td>95260.00</td>
<td>33170.00</td>
<td>6200.00</td>
<td>8370.00</td>
</tr>
<tr>
<td></td>
<td>(03176.00)</td>
<td>(22259.08)</td>
<td>(5028.81)</td>
<td>(5720.15)</td>
<td></td>
</tr>
</tbody>
</table>

Data courtesy of Michael Bashor, General Mills

- An experiment to assess the variability of a particular acid among plants and among leaves of plants:

<table>
<thead>
<tr>
<th>Plant ( i )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaf ( j )</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>( k = 1 )</td>
<td>11.2</td>
<td>16.5</td>
<td>18.3</td>
<td>14.1</td>
</tr>
<tr>
<td>( k = 2 )</td>
<td>11.6</td>
<td>16.8</td>
<td>18.7</td>
<td>13.8</td>
</tr>
<tr>
<td>( k = 3 )</td>
<td>12.0</td>
<td>16.1</td>
<td>19.0</td>
<td>14.2</td>
</tr>
</tbody>
</table>

- Study of light source and intensity on plant height.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Dark</th>
<th>Source</th>
<th>Intensity</th>
<th>Pot</th>
<th>Seedling 1</th>
<th>Seedling 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD</td>
<td>1</td>
<td>D</td>
<td>D</td>
<td>1</td>
<td>32.94</td>
<td>35.98</td>
</tr>
<tr>
<td>DD</td>
<td>1</td>
<td>D</td>
<td>D</td>
<td>2</td>
<td>34.76</td>
<td>32.04</td>
</tr>
<tr>
<td>AL</td>
<td>0</td>
<td>A</td>
<td>L</td>
<td>1</td>
<td>30.55</td>
<td>32.64</td>
</tr>
<tr>
<td>AL</td>
<td>0</td>
<td>A</td>
<td>L</td>
<td>2</td>
<td>32.37</td>
<td>32.04</td>
</tr>
<tr>
<td>AH</td>
<td>0</td>
<td>A</td>
<td>H</td>
<td>1</td>
<td>31.23</td>
<td>31.09</td>
</tr>
<tr>
<td>AH</td>
<td>0</td>
<td>A</td>
<td>H</td>
<td>2</td>
<td>30.62</td>
<td>30.42</td>
</tr>
<tr>
<td>BL</td>
<td>0</td>
<td>B</td>
<td>L</td>
<td>1</td>
<td>34.41</td>
<td>34.88</td>
</tr>
<tr>
<td>BL</td>
<td>0</td>
<td>B</td>
<td>L</td>
<td>2</td>
<td>34.07</td>
<td>33.87</td>
</tr>
<tr>
<td>BH</td>
<td>0</td>
<td>B</td>
<td>H</td>
<td>1</td>
<td>35.61</td>
<td>35.00</td>
</tr>
<tr>
<td>BH</td>
<td>0</td>
<td>B</td>
<td>H</td>
<td>2</td>
<td>33.65</td>
<td>32.91</td>
</tr>
</tbody>
</table>
Analysis of *Campylobacter* counts on chickens data
Residual plots (resid .vs \( \hat{y} \)) for bacteria counts, after fitting two factor fixed effects models (similar plots for mixed models):
data one; /* Bashor data */
   infile "bashor.dat" firstobs=3;
   input day location y; ly=log(y);
run;

proc glm;
   class day location;
   model y ly=location|day;
   output out=two r=residual residual_log p=predicted predicted_log;
run;

/*
symbol1 value=dot color=black; symbol2 value=square color=black;
symbol3 value=triangle color=black; symbol4 value=diamond color=black;

axis1 offset=(1,1) label=(height=3);
axis2 offset=(1,1) label=(height=3 angle=90);
legend1 label=(height=2);

proc gplot data=two;
   title "residuals versus predicted";
   plot residual*predicted=location/haxis=axis1 vaxis=axis2 legend=legend1;
   plot residual_log*predicted_log=location/haxis=axis1 vaxis=axis2 legend=legend1;
run; */

proc mixed method=type3 cl;
   class day location;
   model ly=location/ddfm=satterth outp=predz;
   random day day*location;
   lsmeans location/adj=tukey;
run;

/*proc glm; * the old way of doing things, before PROC MIXED ;
   class day location;
   model ly=day|location;
   random day day*location;
   test h=location e=day*method;
   lsmeans location/pdiff; *wrong;
run;*/

proc mixed method=type3; * to get ANOVA table with EMS terms;
*proc mixed cl; * to get asymmetric confidence intervals ;
class day location;
model ly=location/ddfm=satterth;
random day day*location;
lsmeans location/adj=tukey;
run;
The SAS System
The Mixed Procedure
Model Information

Data Set WORK.ONE
Dependent Variable ly
Covariance Structure Variance Components
Estimation Method Type 3
Residual Variance Method Factor
Fixed Effects SE Method Model-Based
Degrees of Freedom Method Satterthwaite

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>day</td>
<td>3</td>
<td>1 2 3</td>
</tr>
<tr>
<td>location</td>
<td>4</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>3</td>
<td>97.865388</td>
<td>32.621796</td>
<td>Var(Residual) + 10 + Var(day*location) + Q(location)</td>
</tr>
<tr>
<td>day</td>
<td>2</td>
<td>2.787355</td>
<td>1.393677</td>
<td>Var(Residual) + 10 + Var(day*location) + 40 Var(day)</td>
</tr>
<tr>
<td>day*location</td>
<td>6</td>
<td>4.533565</td>
<td>0.755594</td>
<td>Var(Residual) + 10 + Var(day*location)</td>
</tr>
<tr>
<td>Residual</td>
<td>108</td>
<td>59.254946</td>
<td>0.548657</td>
<td>Var(Residual)</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>Error Term</th>
<th>DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>MS(day*location)</td>
<td>6</td>
<td>43.17</td>
<td>0.0002</td>
</tr>
<tr>
<td>day</td>
<td>MS(day*location)</td>
<td>6</td>
<td>1.84</td>
<td>0.2375</td>
</tr>
<tr>
<td>day*location</td>
<td>MS(Residual)</td>
<td>108</td>
<td>1.38</td>
<td>0.2303</td>
</tr>
</tbody>
</table>
(generated by 2nd run of PROC MIXED)

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>day</td>
<td>0.01595</td>
<td>0.05</td>
<td>0.002071</td>
<td>1156981</td>
</tr>
<tr>
<td>day*location</td>
<td>0.02069</td>
<td>0.05</td>
<td>0.002844</td>
<td>145734</td>
</tr>
<tr>
<td>Residual</td>
<td>0.5487</td>
<td>0.05</td>
<td>0.4274</td>
<td>0.7303</td>
</tr>
</tbody>
</table>

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num</th>
<th>Den</th>
<th>DF</th>
<th>DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>3</td>
<td>6</td>
<td>43.17</td>
<td>0.0002</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Least Squares Means

| Effect | location | Estimate | Error | DF | t Value | Pr > |t| |
|--------|----------|----------|-------|----|---------|------|---|
| location | 1 | 10.8870 | 0.1747 | 7.33 | 62.33 | <.0001 |
| location | 2 | 10.4953 | 0.1747 | 7.33 | 60.09 | <.0001 |
| location | 3 | 8.8745 | 0.1747 | 7.33 | 50.81 | <.0001 |
| location | 4 | 8.9394 | 0.1747 | 7.33 | 51.18 | <.0001 |

Differences of Least Squares Means

| Effect | location | _location | Estimate | Error | DF | t Value | Pr > |t| |
|--------|----------|-----------|----------|-------|----|---------|------|---|
| location | 1 | 2 | 0.3917 | 0.2244 | 6 | 1.75 | 0.1316 |
| location | 1 | 3 | 2.0125 | 0.2244 | 6 | 8.97 | 0.0001 |
| location | 1 | 4 | 1.9476 | 0.2244 | 6 | 8.68 | 0.0001 |
| location | 2 | 3 | 1.6208 | 0.2244 | 6 | 7.22 | 0.0004 |
| location | 2 | 4 | 1.5559 | 0.2244 | 6 | 6.93 | 0.0004 |
| location | 3 | 4 | -0.06488 | 0.2244 | 6 | -0.29 | 0.7823 |

Differences of Least Squares Means

<table>
<thead>
<tr>
<th>Effect</th>
<th>location</th>
<th>_location</th>
<th>Adjustment</th>
<th>Adj P</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>1</td>
<td>2</td>
<td>Tukey-Kramer</td>
<td>0.3801</td>
</tr>
<tr>
<td>location</td>
<td>1</td>
<td>3</td>
<td>Tukey-Kramer</td>
<td>0.0004</td>
</tr>
<tr>
<td>location</td>
<td>1</td>
<td>4</td>
<td>Tukey-Kramer</td>
<td>0.0005</td>
</tr>
<tr>
<td>location</td>
<td>2</td>
<td>3</td>
<td>Tukey-Kramer</td>
<td>0.0015</td>
</tr>
<tr>
<td>location</td>
<td>2</td>
<td>4</td>
<td>Tukey-Kramer</td>
<td>0.0018</td>
</tr>
<tr>
<td>location</td>
<td>3</td>
<td>4</td>
<td>Tukey-Kramer</td>
<td>0.9907</td>
</tr>
</tbody>
</table>
Theory for mixed/crossed model used to analyze *Campylobacter* data

Discussion of MIXED output

Model

\[ Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk} \]

w/ variance components \( \sigma_B^2, \sigma_{\alpha B}^2, \sigma^2 \).
Campylobacter analysis, continued

Fixed Factor A: location
Random Factor B: day

To test $H_0 : \sigma^2_{AB} = 0$, use

$$F_{AB} = \frac{MS[AB]}{MS[E]} = \frac{0.76}{0.55} = 1.38$$

on $(a - 1)(b - 1) = 6$ and $ab(n - 1) = 108$ df. The $p$-value is 0.2303, providing no evidence of a random day \times location interaction effect. The variance component for this random effect is estimated by

$$\hat{\sigma}^2_{\alpha B} = \frac{MS[AB] - MS[E]}{n} = \frac{0.76 - 0.55}{10} = 0.021$$

Interpretation: there is no evidence that day-to-day variability varies by location. The estimated variance component is itself very small.

$$\hat{\sigma}^2 = MS[E] = [0.55]$$
$$\hat{\sigma}^2_{AB} = \frac{MS[AB] - MS[E]}{n} = \frac{0.76 - 0.55}{10} = 0.021$$
$$\hat{\sigma}^2_B = \frac{MS[B] - MS[AB]}{na} = \frac{1.39 - 0.76}{40} = 0.016$$
Implied correlation structure

What is the correlation of two observations taken on the same day

- at the same location?
- at different locations?

Recall that \( Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijk} \).

\[
\text{Corr}(Y_{ijk1}, Y_{ijk2}) = \frac{\text{Cov}(Y_{ijk1}, Y_{ijk2})}{\sigma^2 + \sigma_B^2 + \sigma_{\alpha B}^2} = \frac{\text{Cov}(B_i, B_i) + \text{Cov}((\alpha B)_{ij}, (\alpha B)_{ij})}{\sigma^2 + \sigma_B^2 + \sigma_{\alpha B}^2}
\]

\[
\text{Corr}(Y_{1jk1}, Y_{2jk2}) = \frac{\text{Cov}(Y_{1jk1}, Y_{2jk2})}{\sigma^2 + \sigma_{B}^2 + \sigma_{\alpha B}^2} = \frac{\text{Cov}(B_i, B_i)}{\sigma^2 + \sigma_{B}^2 + \sigma_{\alpha B}^2}
\]

Estimates of these correlations are

- \( \frac{0.016 + 0.021}{0.016 + 0.021 + 0.55} = \frac{0.037}{0.587} = 0.063 \)
- \( \frac{0.016}{0.016 + 0.021 + 0.55} = \frac{0.016}{0.587} = 0.027 \)

Which is which?

What about the correlation of two observations on different days?
Some analysis of fixed effects

Consider testing for a fixed effect of location. That is, test the hypothesis that average bacteria counts are constant across the locations,

\[ H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0 \]

\[ F_A = \frac{MS[A]}{MS[AB]} = \frac{32.6}{0.76} = 43.2 \]

on \( a - 1 = 3 \) and \( (a - 1)(b - 1) = 6 \) df, which is significant (\( p = 0.0002 \)).
**Campylobacter** analysis, continued

To estimate the a pairwise comparison among location means, such as, $\alpha_4 - \alpha_3$, consider

$$\hat{\theta} = \bar{y}_{4++} - \bar{y}_{3++} = 8.940 - 8.875 = -0.065$$

Note that

$$\text{Var}(\bar{Y}_{4++} - \bar{Y}_{3++}) \neq \sigma^2 \left( \frac{1}{nb} + \frac{1}{nb} \right)$$

(Why not?)

What is $SE(\hat{\theta})$ and how can it be estimated?

$$\hat{\theta} = \bar{Y}_{2++} - \bar{Y}_{1++}$$

$$= \alpha_2 + \bar{B} + \alpha\bar{B}_2 + \bar{E}_{2++}$$

$$- (\alpha_1 + \bar{B} + \alpha\bar{B}_1 + \bar{E}_{1++})$$

$$= \alpha_2 - \alpha_1 + \alpha\bar{B}_2 - \alpha\bar{B}_1 + \bar{E}_{2++} - \bar{E}_{1++}$$

which has variance

$$\text{Var}(\hat{\theta}) = \text{Var}(\alpha\bar{B}_2) + \text{Var}(\alpha\bar{B}_1) + \text{Var}(\bar{E}_{2++}) + \text{Var}(\bar{E}_{1++})$$

$$= 2\sigma^2_{\alpha B} + 2\sigma^2_{B}$$

$$= \frac{2}{nb}(\sigma^2 + n\sigma^2_{\alpha B})$$

which can be estimated nicely on $(a - 1)(b - 1) = 6df$ by

$$\hat{\text{Var}}(\hat{\theta}) = \frac{2}{nb}MS[AB]$$

for the chickens, where $\bar{y}_{4++} - \bar{y}_{3++} = -0.06$ the $SE$ is

$$\sqrt{\hat{\text{Var}}(\hat{\theta})} = \sqrt{\frac{2}{3*10}0.76} = 0.22$$

Since $t(0.025, 6) = 2.45$, a 95% c.i. for $\theta$ given by $-0.06 \pm 2.45(0.22)$. 
Campylobacter analysis, continued

Reporting standard errors for sample means of levels of fixed factor, like LOCATION means, is a little messier:

$$\bar{Y}_{i++} = \mu + \alpha_i + \bar{B} + \bar{\alpha B}_{i+} + \bar{E}_{i+}$$

$$\text{Var}(\bar{Y}_{i++}) = \text{Var}(\bar{B}) + \text{Var}(\bar{\alpha B}_{i+}) + \text{Var}(\bar{E}_{i+})$$

$$= \frac{\sigma_B^2}{b} + \frac{\sigma_{\alpha B}^2}{b} + \frac{\sigma^2}{nb}$$

$$= \frac{1}{nb} (n\sigma_B^2 + n\sigma_{\alpha B}^2 + \sigma^2)$$

estimated by

$$\widehat{\text{Var}}(\bar{Y}_{i++}) = \frac{1}{nb} (n\hat{\sigma}_B^2 + n\hat{\sigma}_{\alpha B}^2 + \hat{\sigma}^2)$$

algebra yields a linear combo of multiple EMS terms

$$= \frac{1}{nab} \{(a - 1)EMS[AB] + EMS[B]\}$$

The standard error is estimated easily enough:

$$\widehat{SE}(\bar{Y}_{i++}) = \sqrt{\frac{1}{nab} \{(a - 1)MS[AB] + MS[B]\}}$$

$$= \sqrt{\frac{1}{120} \{(4 - 1)0.76 + 1.39\}}$$

$$= \sqrt{0.03} = 0.175$$

but the df must be approximated using the Satterthwaite approach

$$\hat{df} = \frac{0.175^4}{\frac{1}{120^2} \left( \frac{((4-1)0.76)^2}{6} + \frac{1.39^2}{2} \right)} = 7.33$$

with df$_{AB} = 6$, df$_B = 2$. Since $t(0.025, 7.33) = 2.34$, a 95% c.i. for the population mean of location 1, for example, is $10.9 \pm 2.34(0.175)$. 
SAS code to fit two-factor random effects model for plant acid data

Nested or crossed?

options ls=75 nodate;

data one;
   infile "plantacid.dat";
   input y plant leaf rep;
run;

proc mixed cl method=type3;
*proc mixed cl;
   class plant leaf;
   model y=/s cl;
   random plant leaf(plant);
run;

goptions colors=(black) dev=pslepsf;
*goptions colors=(black);

axis1 value=(h=2) offset=(10);

symbol1 value=dot h=1.5;
symbol2 value=diamond h=1.5;
symbol3 value=plus h=1.5;

proc gplot;
   title "plant acids";
   plot y*plant=leaf/haxis=axis1;
run;
The Mixed Procedure
Class Level Information

Class Levels Values
plant 4 1 2 3 4
leaf 3 1 2 3

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>3</td>
<td>343.178889</td>
<td>114.392963</td>
</tr>
<tr>
<td>leaf(plant)</td>
<td>8</td>
<td>187.453333</td>
<td>23.431667</td>
</tr>
<tr>
<td>Residual</td>
<td>24</td>
<td>3.033333</td>
<td>0.126389</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Expected Mean Square</th>
<th>Error Term</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>Var(Residual) + 3 Var(leaf(plant)) + 9 Var(plant)</td>
<td>MS(leaf(plant))</td>
<td>8</td>
</tr>
<tr>
<td>leaf(plant)</td>
<td>Var(Residual) + 3 Var(leaf(plant))</td>
<td>MS(Residual)</td>
<td>24</td>
</tr>
<tr>
<td>Residual</td>
<td>Var(Residual)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>4.88</td>
<td>0.0324</td>
</tr>
<tr>
<td>leaf(plant)</td>
<td>185.39</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>10.1068</td>
<td>0.05</td>
<td>-10.3930</td>
<td>30.6066</td>
</tr>
<tr>
<td>leaf(plant)</td>
<td>7.7684</td>
<td>0.05</td>
<td>0.1142</td>
<td>15.4227</td>
</tr>
<tr>
<td>Residual</td>
<td>0.1264</td>
<td>0.05</td>
<td>0.07706</td>
<td>0.2446</td>
</tr>
</tbody>
</table>

/*Covariance Parameter Estimates*/

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>10.1068</td>
<td>0.05</td>
<td>2.6599</td>
<td>499.70</td>
</tr>
<tr>
<td>leaf(plant)</td>
<td>7.7684</td>
<td>0.05</td>
<td>3.5322</td>
<td>28.7787</td>
</tr>
<tr>
<td>Residual</td>
<td>0.1264</td>
<td>0.05</td>
<td>0.07706</td>
<td>0.2446</td>
</tr>
</tbody>
</table>
Solution for Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>14.2611</td>
<td>1.7826</td>
<td>3</td>
<td>8.00</td>
<td>0.0041</td>
<td></td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>8.5882</td>
<td>19.9341</td>
</tr>
</tbody>
</table>

**plant acids**

![Graph showing plant acids data]
Discussion of MIXED output and analysis of plant acid data

Random, nested model

\[ Y_{ijk} = \mu + A_i + B_{j(i)} + E_{ijk} \]

w/ variance components \( \sigma^2, \sigma_A^2, \sigma_{B(A)}^2 \).

To test for random effect of nested factor \( B \) (leaf), \( H_0 : \sigma_{B(A)}^2 = 0 \),

\[ F = \frac{MS[B(A)]}{MS[E]} = \frac{23.4}{0.13} = 185.4 \]
on \( (b-1)a = 8 \) and \( (n-1)ab = 24 \) df \( (p\text{-value} < 0.0001) \).

To test for random effect of factor \( A \) (plant), \( H_0 : \sigma_A^2 = 0 \),

\[ F = \frac{MS[A]}{MS[B(A)]} = \frac{114.4}{23.4} = 4.88 \]
on \( a-1 = 3 \) and \( (b-1)a = 8df \) with \( p = 0.0324 \).

Reminder: Watch that denominator \( MS! \)

\[
\hat{\sigma}^2 = MS[E] = \boxed{0.13}
\]

\[
\hat{\sigma}_{B(A)}^2 = \frac{MS[B(A)]}{\frac{MS[B(A)] - MS[E]}{n}} = \frac{23.4 - 0.13}{3} = \boxed{7.8}
\]

\[
\hat{\sigma}_A^2 = \frac{MS[A] - MS[B(A)]}{\frac{nb}{9}} = \frac{114.4 - 23.4}{9} = \boxed{10.1}
\]

So there is some evidence of both a random plant effect and a random leaf effect, nested in plant. The magnitudes of these effects are quantified by the estimated variance components. The statistical significance addressed by the \( p \)-values.
Implied correlation structure for plant acids

What is the correlation of two observations taken from the same plant

• and the same leaf?
• and different leaves?

Recall that \( Y_{ijk} = \mu + A_i + B_{j(i)} + E_{ijk} \).

\[
\text{Corr}(Y_{ik1}, Y_{ik2}) = \frac{\text{Cov}(Y_{ik1}, Y_{ik2})}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2} = \frac{\text{Cov}(A_i, A_i) + \text{Cov}(B_{j(i)}, B_{j(i)})}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2}
\]

\[
= \frac{\sigma_A^2 + \sigma_{B(A)}^2}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2}
\]

\[
\text{Corr}(Y_{ijk1}, Y_{ijk2}) = \frac{\text{Cov}(Y_{ijk1}, Y_{ijk2})}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2} = \frac{\text{Cov}(A_i, A_i)}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2}
\]

\[
= \frac{\sigma_A^2}{\sigma^2 + \sigma_A^2 + \sigma_{B(A)}^2}
\]

Estimates of these correlations are

• \( \frac{10.1 + 7.8}{10.1 + 7.8 + 0.13} = \frac{17.9}{18.0} = 0.99 \)
• \( \frac{10.1}{10.1 + 7.8 + 0.13} = \frac{10.1}{18.0} = 0.56 \)

This means that two measurements taken on the same leaf are almost perfectly correlated. Almost all the variation in any measurement can be explained by the leaf and plant effects.
<table>
<thead>
<tr>
<th>Treatment</th>
<th>Dark</th>
<th>Source</th>
<th>Intensity</th>
<th>Pot</th>
<th>Seedling 1</th>
<th>Seedling 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD</td>
<td>1</td>
<td>D</td>
<td>D</td>
<td>1</td>
<td>32.94</td>
<td>35.98</td>
</tr>
<tr>
<td>DD</td>
<td>1</td>
<td>D</td>
<td>D</td>
<td>2</td>
<td>34.76</td>
<td>32.40</td>
</tr>
<tr>
<td>AL</td>
<td>0</td>
<td>A</td>
<td>L</td>
<td>1</td>
<td>30.55</td>
<td>32.64</td>
</tr>
<tr>
<td>AL</td>
<td>0</td>
<td>A</td>
<td>L</td>
<td>2</td>
<td>32.37</td>
<td>32.04</td>
</tr>
<tr>
<td>AH</td>
<td>0</td>
<td>A</td>
<td>H</td>
<td>1</td>
<td>31.23</td>
<td>31.09</td>
</tr>
<tr>
<td>AH</td>
<td>0</td>
<td>A</td>
<td>H</td>
<td>2</td>
<td>30.62</td>
<td>30.42</td>
</tr>
<tr>
<td>BL</td>
<td>0</td>
<td>B</td>
<td>L</td>
<td>1</td>
<td>34.41</td>
<td>34.88</td>
</tr>
<tr>
<td>BL</td>
<td>0</td>
<td>B</td>
<td>L</td>
<td>2</td>
<td>34.07</td>
<td>33.87</td>
</tr>
<tr>
<td>BH</td>
<td>0</td>
<td>B</td>
<td>H</td>
<td>1</td>
<td>35.61</td>
<td>35.00</td>
</tr>
<tr>
<td>BH</td>
<td>0</td>
<td>B</td>
<td>H</td>
<td>2</td>
<td>33.65</td>
<td>32.91</td>
</tr>
</tbody>
</table>

- Response ($y$) is seedling height,
- treatments are light sources, intensities,
- experimental units are 10 pots (points on graph).
Experiment with light treatments on seedlings

\[ Y_{ijk} = \mu + \alpha_i + P_{j(i)} + E_{ijk} \]

\( \alpha_i \) - treatment effects for \( i = 1, 2, 3, 4, 5 \)

\( P_{j(i)} \) - pot effects, nested in treatments, \( j = 1, 2 \) for each \( i \).

\( E_{ijk} \) - seedling/experimental errors, \( k = 1, 2 \)

\[ P_{j(i)} \overset{iid}{\sim} N(0, \sigma_P^2), \quad E_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \quad (P_{j(i)} \perp E_{ijk}) \]

For treatment effects, use \( MS(Pot(treatments)) \) as error term.

For example, for \( H_0 : \alpha_1 = \alpha_2 = \cdots = 0 \), use

\[ F = \frac{MS(treatment)}{MS(Pot(treatment))} \sim F_{5-1,5(2-1)} \text{ or } F_{4,5} \]

Be careful not to use

\[ F = \frac{MS(treatment)}{MS(E)} \]

For these data, we get

\[ F = \frac{10.27}{1.22} = 8.4 (df = 4, 5, p = .0192) \]

providing evidence of a treatment effect on plant heights. SAS code to fixed the mixed effects model (output follows):

```
proc mixed method=type3 cl;
*proc mixed data=planhtts cl;
  class pot treatment;
  model y=treatment;
  random pot(treatment);
  *lsmeans treatment/diffs adj=tukey;
  lsmeans treatment/diffs;
  estimate "main effect of source" treatment 1 1 -1 -1/divisor=2;
  estimate "main effect of intensity" treatment 1 -1 1 -1/divisor=2;
  estimate "interaction " treatment 1 -1 -1 1;
  contrast "main effect of source" treatment 1 1 -1 -1;
  contrast "main effect of intensity" treatment 1 -1 1 -1;
  contrast "interaction " treatment 1 -1 -1 1;
run;
```
The SAS System
The Mixed Procedure

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>pot</td>
<td>2</td>
<td>1 2</td>
</tr>
<tr>
<td>treatment</td>
<td>5</td>
<td>AH AL BH BL DD</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Sum of</th>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>treatment</td>
<td>4</td>
<td>41.080770</td>
<td>10.270192</td>
</tr>
<tr>
<td></td>
<td>pot(treatment)</td>
<td>5</td>
<td>6.112350</td>
<td>1.222470</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>10</td>
<td>10.264200</td>
<td>1.026420</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Error Term</th>
<th>Source</th>
<th>Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>treatment</td>
<td>Var(Residual) + 2 Var(pot(treatment)) + Q(treatment)</td>
</tr>
<tr>
<td></td>
<td>pot(treatment)</td>
<td>Var(Residual) + 2 Var(pot(treatment))</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>Var(Residual)</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Error Term</th>
<th>Source</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>treatment</td>
<td>8.40</td>
<td>0.0192</td>
</tr>
<tr>
<td></td>
<td>pot(treatment)</td>
<td>1.19</td>
<td>0.3793</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>pot(treatment)</td>
<td>0.09802</td>
<td>0.05</td>
<td>-0.7831</td>
<td>0.9792</td>
</tr>
<tr>
<td>Residual</td>
<td>1.0264</td>
<td>0.05</td>
<td>0.5011</td>
<td>3.1612</td>
</tr>
</tbody>
</table>

Estimates

| Label                | Estimate  | Error Term | DF | t Value | Pr > |t| |
|----------------------|-----------|-------------|----|---------|-------|-----|
| main effect of source | -2.9300   | 0.5528      | 5  | -5.30   | 0.0032|
| main effect of intensity | -0.5375 | 0.5528      | 5  | -0.97   | 0.3756|
| interaction           | -1.0450   | 1.1057      | 5  | -0.95   | 0.3880|

Contrasts

<table>
<thead>
<tr>
<th>Label</th>
<th>Num</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>main effect of source</td>
<td>1</td>
<td>5</td>
<td>28.09</td>
<td>0.0032</td>
</tr>
<tr>
<td>main effect of intensity</td>
<td>1</td>
<td>5</td>
<td>0.95</td>
<td>0.3756</td>
</tr>
<tr>
<td>interaction</td>
<td>1</td>
<td>5</td>
<td>0.89</td>
<td>0.3880</td>
</tr>
</tbody>
</table>
Output for plant heights and light sources, cont’d

Least Squares Means

| Effect | treatment | Estimate | Error  | DF  | t Value | Pr > |t| |
|--------|-----------|----------|--------|-----|---------|-------|
| treatment | AH        | 30.8400  | 0.5528 | 5   | 55.79   | <.0001|
| treatment | AL        | 31.9000  | 0.5528 | 5   | 57.70   | <.0001|
| treatment | BH        | 34.2925  | 0.5528 | 5   | 62.03   | <.0001|
| treatment | BL        | 34.3075  | 0.5528 | 5   | 62.06   | <.0001|
| treatment | DD        | 34.0200  | 0.5528 | 5   | 61.54   | <.0001|

Differences of Least Squares Means

| Effect | treatment | _treatment | Estimate | Error  | DF  | t Value | Pr > |t| |
|--------|-----------|------------|----------|--------|-----|---------|-------|
| treatment | AH        | AL         | -1.0600  | 0.7818 | 5   | -1.36   | 0.2332|
| treatment | AH        | BH         | -3.4525  | 0.7818 | 5   | -4.42   | 0.0069|
| treatment | AH        | BL         | -3.4675  | 0.7818 | 5   | -4.44   | 0.0068|
| treatment | AH        | DD         | -3.1800  | 0.7818 | 5   | -4.07   | 0.0097|
| treatment | AL        | BH         | -2.3925  | 0.7818 | 5   | -3.06   | 0.0281|
| treatment | AL        | BL         | -2.4075  | 0.7818 | 5   | -3.08   | 0.0275|
| treatment | AL        | DD         | -2.1200  | 0.7818 | 5   | -2.71   | 0.0422|
| treatment | BH        | BL         | -0.0150  | 0.7818 | 5   | -0.02   | 0.9854|
| treatment | BH        | DD         | 0.2725   | 0.7818 | 5   | 0.35    | 0.7416|
| treatment | BL        | DD         | 0.2875   | 0.7818 | 5   | 0.37    | 0.7281|

(In practice, if all 10 pairwise comparisons were of interest, Tukey’s adjustment can be considered. Here, the output was too wide for the page, and there was greater interest in factorial effects (source, light, interaction), anyway.)
Using nested factorial effects to get SAS to produce appropriate contrast sums of squares for factorial effects analysis of plant height and light source data

```
proc mixed method=type3;
  class pot treatment source intensity dark;
  model y=dark source(dark) intensity(dark) source*intensity(dark) dark ;
  random pot(source*intensity*dark);
  lsmeans dark source(dark) intensity(dark) source*intensity(dark);
run;
```

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>dark</td>
<td>1</td>
<td>4.493520</td>
<td>4.493520</td>
<td>Var(Residual) + 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Var(pot(source<em>intensity</em>dark)) +</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q(dark,source(dark),intensity(dark),source*intensity(dark))</td>
</tr>
<tr>
<td>source(dark)</td>
<td>1</td>
<td>34.339600</td>
<td>34.339600</td>
<td>Var(Residual) + 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Var(pot(source<em>intensity</em>dark)) +</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q(source(dark),source*intensity(dark))</td>
</tr>
<tr>
<td>intensity(dark)</td>
<td>1</td>
<td>1.155625</td>
<td>1.155625</td>
<td>Var(Residual) + 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Var(pot(source<em>intensity</em>dark)) +</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q(intensity(dark),source*intensity(dark))</td>
</tr>
<tr>
<td>source*intensi(dark)</td>
<td>1</td>
<td>1.092025</td>
<td>1.092025</td>
<td>Var(Residual) + 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Var(pot(source<em>intensity</em>dark)) +</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q(source*intensity(dark))</td>
</tr>
<tr>
<td>pot(sour<em>inten</em>dark)</td>
<td>5</td>
<td>6.112350</td>
<td>1.222470</td>
<td>Var(Residual) + 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Var(pot(source<em>intensity</em>dark))</td>
</tr>
<tr>
<td>Residual</td>
<td>10</td>
<td>10.264200</td>
<td>1.026420</td>
<td>Var(Residual)</td>
</tr>
</tbody>
</table>

Error

<table>
<thead>
<tr>
<th>Source</th>
<th>Error Term</th>
<th>DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>dark</td>
<td>MS(pot(source<em>inten</em>dark))</td>
<td>5</td>
<td>3.68</td>
<td>0.1133</td>
</tr>
<tr>
<td>source(dark)</td>
<td>MS(pot(source<em>inten</em>dark))</td>
<td>5</td>
<td>28.09</td>
<td>0.0032</td>
</tr>
<tr>
<td>intensity(dark)</td>
<td>MS(pot(source<em>inten</em>dark))</td>
<td>5</td>
<td>0.95</td>
<td>0.3756</td>
</tr>
<tr>
<td>source*intensi(dark)</td>
<td>MS(pot(source<em>inten</em>dark))</td>
<td>5</td>
<td>0.89</td>
<td>0.3880</td>
</tr>
<tr>
<td>pot(sour<em>inten</em>dark)</td>
<td>MS(Residual)</td>
<td>10</td>
<td>1.19</td>
<td>0.3793</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>pot(sour<em>inten</em>dark)</td>
<td>0.09802</td>
</tr>
<tr>
<td>Residual</td>
<td>1.0264</td>
</tr>
</tbody>
</table>
Inference for light effects

Model for treatment combination “ijk” and pot l, seedling m:

\[ Y_{ijklm} = \mu + \delta_i + \alpha_j(i) + \beta_k(i) + (\alpha\beta)_{jk}(i) + P_{l(ijk)} + E_{ijklm} \]

For treatment effects, use \( MS(Pot(treatments)) \) as the error term.

e.g.: Is intensity effect is constant across light types? (\( H_0 : \gamma_{1jk} \equiv 0 \))

\[
F = \frac{MS(\text{interaction(dark)})}{MS(\text{Pot(dark*source*intensity))}} = \frac{1.09}{1.22} = .89 (p = .3880)
\]

Degrees of freedom: \( (df = ?, ?) \)

Estimation of variance components:

\[
\hat{\sigma}^2 = MS(E) = 1.02 (df = 10)
\]

\[
\hat{\sigma}^2_{P(T)} = \frac{MS(\text{pot(treatment)}) - MS(E)}{2} = \frac{1.22 - 1.02}{2} = 0.098 (df = \hat{df})
\]

Correlation structure? Intrapot correlation?

\[
\hat{\text{Corr}}(Y_{ijklm_1}, Y_{ijklm_2}) = \frac{\hat{\sigma}^2_{P(T)}}{\hat{\sigma}^2 + \hat{\sigma}^2_{P(T)}} = \frac{.098}{.098 + 1.02} = .088
\]
**Topic:** Split-plots: a repeated measures design

**Reading:** Rao, Ch. 16.

---

**Repeated measures models**

Consider an experiment to study effects of irrigation and aerially sprayed pesticide on yields of different varieties of corn.

Factors:

- $A$: Pesticide treatment, $a = 3$ levels
- $B$: Irrigation treatment, $b = 4$ levels (called ‘treatment, trt’)
- Plots, $n = 2$ per level of $A$, total of $na = 6$ plots

For the moment, ignore $B$, or fix the level of $B$.

<table>
<thead>
<tr>
<th>plot</th>
<th>pest</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>53.4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>46.5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>54.3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>57.2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>55.9</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>57.4</td>
</tr>
</tbody>
</table>

One-way ANOVA for $A$ effect and plot effects:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ : Pesticide</td>
<td>$a - 1 = 2$</td>
</tr>
<tr>
<td>Error or “plots”</td>
<td>$(n - 1)a = (2 - 1)3 = 3$</td>
</tr>
<tr>
<td>Total</td>
<td>$an - 1 = 5$</td>
</tr>
</tbody>
</table>
Split-plot design

Levels of factor $B$ are randomly assigned to $b = 4$ subplots within each of the $na = 6$ plots in a split-plot design:

<table>
<thead>
<tr>
<th>pest</th>
<th>plot</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>53.4</td>
<td>53.8</td>
<td>58.2</td>
<td>59.5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>46.5</td>
<td>51.1</td>
<td>49.2</td>
<td>51.3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>54.3</td>
<td>56.3</td>
<td>60.4</td>
<td>64.5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>57.2</td>
<td>56.9</td>
<td>61.6</td>
<td>66.8</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>55.9</td>
<td>58.6</td>
<td>62.4</td>
<td>64.5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>57.4</td>
<td>60.2</td>
<td>57.2</td>
<td>62.7</td>
</tr>
</tbody>
</table>

Each row corresponds to one of $na = 6$ plots. Each plot is divided into $b = 4$ subplots and levels of factor $B$ are assigned to these at random.

The ANOVA table on the preceding page is at the whole plot level. Sources of variation for the split-plot level:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>B: treatments</td>
<td>$b - 1 = 3$</td>
</tr>
<tr>
<td>$A \times B$</td>
<td>$(a - 1)(b - 1) = 6$</td>
</tr>
<tr>
<td>$B \times \text{plot(A)}$</td>
<td>$(b - 1)(n - 1)a = 9$</td>
</tr>
</tbody>
</table>

aka Subplot error

- $A$ - a between plots or between subjects factor
- $B$ - a within plots or within subjects factor
- Plots are ‘subjects’ in repeated measures terminology, where time is often the within subjects factor.

Suggestion: draw a picture of the layout.
<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Pesticide</td>
<td>(a - 1 = 2)</td>
<td>(\sigma^2 + b\sigma_s^2 + bn\psi_{AB}^2)</td>
</tr>
<tr>
<td>Plot((A))</td>
<td>((n - 1)a = (2 - 1)3 = 3)</td>
<td>(\sigma^2 + b\sigma_s^2)</td>
</tr>
<tr>
<td>B: treatments</td>
<td>(b - 1 = 3)</td>
<td>(\sigma^2 + na\psi_B^2)</td>
</tr>
<tr>
<td>A (\times) B</td>
<td>((a - 1)(b - 1) = 6)</td>
<td>(\sigma^2 + n\psi_{AB}^2)</td>
</tr>
<tr>
<td>B (\times) plot(A)</td>
<td>((b - 1)(n - 1)a = 9)</td>
<td>(\sigma^2)</td>
</tr>
<tr>
<td>Subplot error</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>(abn - 1 = 23)</td>
<td></td>
</tr>
</tbody>
</table>

Where variance components and size effects pertain to the model for a completely randomized split-plot design:

\[
Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + S_{k(i)} + E_{ijk}. \]

Here, \(i = 1, \ldots, a\) and \(j = 1, \ldots, b\) and \(k = 1, \ldots, n_i\) where \(n_i\) denotes the number of plots treated with level \(i\) of factor \(a\). If \(n_i\) is constant, call it \(n\).

Random effects and variance components:

\[
S_{k(i)} \overset{iid}{\sim} N(0, \sigma_s^2) \]

\[
E_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \]

Size effects for fixed factors same as in prior 2-factor models. For our example, \(\alpha_i\) denote pesticide effects, \(\beta_j\) denote irrigation effects, \((\alpha\beta)_{ij}\) are interactions. \(F\)-tests for fixed effects guided by EMS column above.
For the corn yields data on p. 2,

<table>
<thead>
<tr>
<th>Source</th>
<th>MS</th>
<th>df</th>
<th>EMS</th>
<th>F</th>
<th>p−value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A : Pesticide</td>
<td>128.1</td>
<td>2</td>
<td>$\sigma^2 + b\sigma^2_s + b n\psi^2_A$</td>
<td>3.9</td>
<td>0.1452</td>
</tr>
<tr>
<td>Whole plot error $MS[S(A)]$</td>
<td>32.6</td>
<td>3</td>
<td>$\sigma^2 + b\sigma^2_s$</td>
<td>10.1</td>
<td>0.0031</td>
</tr>
<tr>
<td>B: treatments</td>
<td>60.2</td>
<td>3</td>
<td>$\sigma^2 + na\psi^2_B$</td>
<td>18.7</td>
<td>0.0003</td>
</tr>
<tr>
<td>$A \times B$</td>
<td>4.1</td>
<td>6</td>
<td>$\sigma^2 + n\psi^2_{AB}$</td>
<td>1.3</td>
<td>0.3607</td>
</tr>
<tr>
<td>$B \times \text{plot}(A)$</td>
<td>$MS[E] = 3.2$</td>
<td>9</td>
<td>$\sigma^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Subplot error)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $MS[S(A)]$ denotes mean square for WHOLE plots (nested in A)
- $MS[E]$ denotes error or subplot mean square

For pesticide effect, on 2, 3 df:

$$F = MS[A]/MS[S(A)] = 128.1/32.6$$

For irrigation effect, on 3, 9 df:

$$F = MS[B]/MS[E] = 60.2/3.2$$

For pesticide by irrigation interaction, on 6, 9 df:

$$F = MS[AB]/MS[E] = 4.1/3.2$$

For random effect of whole plots, on 3, 9 df:

$$F = MS[S(A)]/MS[E] = 32.6/3.2$$

Estimated varcomps:

$$\hat{\sigma}^2 = MS[E] = 3.2 \quad \text{and} \quad \hat{\sigma}_s^2 = (MS[S(A)] - MS[E])/4 = 7.3$$
Pairwise comparisons

Several kinds of pairwise comparisons of treatment means:

1. Main effects of $A$: $\bar{y}_{i1} - \bar{y}_{i2}$
2. Main effects of $B$: $\bar{y}_{j1} - \bar{y}_{j2}$
3. Simple effects of $A$: $\bar{y}_{i1j} - \bar{y}_{i2j}$
4. Simple effects of $B$: $\bar{y}_{ij1} - \bar{y}_{ij2}$
5. Interaction effects: $\bar{y}_{i1j1} - \bar{y}_{i2j2}$

Skipping the algebra, the standard errors for all of these comparisons, save # 3 and #5, can be estimated ‘cleanly.’ That is, with single $MS$ terms and integer $df$. (See table 16.6, careful of errata)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Variance</th>
<th>Estimate</th>
<th>$df$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_{i1} - Y_{i2}$</td>
<td>$\frac{2}{nb}(\sigma^2 + b\sigma^2_s)$</td>
<td>$\frac{2}{nb}MS[S(A)]$</td>
<td>$(n - 1)a$</td>
</tr>
<tr>
<td>$\bar{Y}<em>{j1} - \bar{Y}</em>{j2}$</td>
<td>$\frac{2}{na}\sigma^2$</td>
<td>$\frac{2}{na}MS[E]$</td>
<td>$(n - 1)(b - 1)a$</td>
</tr>
<tr>
<td>$\bar{Y}<em>{i1j} - \bar{Y}</em>{i2j}$</td>
<td>$\frac{2}{n}\sigma^2$</td>
<td>$\frac{2}{n}(\hat{\sigma}^2 + \hat{\sigma}^2_s)$</td>
<td>messy</td>
</tr>
<tr>
<td>$\bar{Y}<em>{ij1} - \bar{Y}</em>{ij2}$</td>
<td>$\frac{2}{n}\sigma^2$</td>
<td>$\frac{2}{n}MS[E]$</td>
<td>$(n - 1)(b - 1)a$</td>
</tr>
<tr>
<td>$\bar{Y}<em>{i1j1} - \bar{Y}</em>{i2j2}$</td>
<td>$\frac{2}{n}\sigma^2$</td>
<td>$\frac{2}{n}(\hat{\sigma}^2 + \hat{\sigma}^2_s)$</td>
<td>messy</td>
</tr>
</tbody>
</table>

To analyze data from a CRSPD in SAS, consider using PROC MIXED instead of PROC GLM:

```sas
proc mixed method=type3;
    class field pest trt irr cv;
    model y=trt|pest/ddfm=satterth;
    random field(pest);
    *parms /nobound;
    lsmeans trt pest/pdiff; /* can use adj=bon; to adjust for multiplicity */
    *lsmeans trt|pest/pdiff; /* if there were interaction */
run;
```

/* parms statement can be used to keep SAS from dropping random effects w/ negative estimated varcomps */
The SAS System
The Mixed Procedure

Model Information

Data Set WORK.ONE
Dependent Variable y
Covariance Structure Variance Components
Estimation Method Type 3
Residual Variance Method Factor
Fixed Effects SE Method Model-Based
Degrees of Freedom Method Satterthwaite

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>field</td>
<td>2</td>
<td>1 2</td>
</tr>
<tr>
<td>pest</td>
<td>3</td>
<td>1 2 3</td>
</tr>
<tr>
<td>irr</td>
<td>2</td>
<td>1 2</td>
</tr>
<tr>
<td>cv</td>
<td>2</td>
<td>1 2</td>
</tr>
<tr>
<td>trt</td>
<td>4</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>3</td>
<td>180.697917</td>
<td>60.232639</td>
<td>Var(Residual) + Q(trt,pest*trt)</td>
</tr>
<tr>
<td>pest</td>
<td>2</td>
<td>256.275833</td>
<td>128.137917</td>
<td>Var(Residual) + 4 Var(field(pest)) + Q(pest,pest*trt)</td>
</tr>
<tr>
<td>pest*trt</td>
<td>6</td>
<td>24.490833</td>
<td>4.081806</td>
<td>Var(Residual) + Q(pest*trt)</td>
</tr>
<tr>
<td>field(pest)</td>
<td>3</td>
<td>97.806250</td>
<td>32.602083</td>
<td>Var(Residual) + 4 Var(field(pest))</td>
</tr>
<tr>
<td>Residual</td>
<td>9</td>
<td>29.058750</td>
<td>3.228750</td>
<td>Var(Residual)</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>Error Term</th>
<th>DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>MS(Residual)</td>
<td>9</td>
<td>18.66</td>
<td>0.0003</td>
</tr>
<tr>
<td>pest</td>
<td>MS(field(pest))</td>
<td>3</td>
<td>3.93</td>
<td>0.1452</td>
</tr>
<tr>
<td>pest*trt</td>
<td>MS(Residual)</td>
<td>9</td>
<td>1.26</td>
<td>0.3607</td>
</tr>
<tr>
<td>field(pest)</td>
<td>MS(Residual)</td>
<td>9</td>
<td>10.10</td>
<td>0.0031</td>
</tr>
</tbody>
</table>
Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>field(pest)</td>
<td>7.3433</td>
</tr>
<tr>
<td>Residual</td>
<td>3.2287</td>
</tr>
</tbody>
</table>

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>3</td>
<td>9</td>
<td>18.66</td>
<td>0.0003</td>
</tr>
<tr>
<td>pest</td>
<td>2</td>
<td>3</td>
<td>3.93</td>
<td>0.1452</td>
</tr>
<tr>
<td>pest*trt</td>
<td>6</td>
<td>9</td>
<td>1.26</td>
<td>0.3607</td>
</tr>
</tbody>
</table>

Least Squares Means

| Effect | pest | trt  | Estimate | Error | DF | t Value | Pr > |t| |
|--------|------|------|----------|-------|----|---------|------|---|
| trt    | 1    | 54.1167 | 1.3274   | 4.9   | 40.77 | <.0001 |
| trt    | 2    | 56.1500 | 1.3274   | 4.9   | 42.30 | <.0001 |
| trt    | 3    | 58.1667 | 1.3274   | 4.9   | 43.82 | <.0001 |
| trt    | 4    | 61.5500 | 1.3274   | 4.9   | 46.37 | <.0001 |
| pest   | 1    | 52.8750 | 2.0187   | 3     | 26.19 | <.0001 |
| pest   | 2    | 59.7500 | 2.0187   | 3     | 29.60 | <.0001 |
| pest   | 3    | 59.8625 | 2.0187   | 3     | 29.65 | <.0001 |

Differences of Least Squares Means

| Effect | pest | trt  | _pest | _trt | Estimate | Error | DF | t Value | Pr > |t| |
|--------|------|------|-------|------|----------|-------|----|---------|------|---|
| trt    | 1    | 2    | 2     | -2.0333 | 1.0374   | 9     | -1.96 | 0.0816 |
| trt    | 1    | 3    | 3     | -4.0500 | 1.0374   | 9     | -3.90 | 0.0036 |
| trt    | 1    | 4    | 4     | -7.4333 | 1.0374   | 9     | -7.17 | <.0001 |
| trt    | 2    | 3    | 3     | -2.0167 | 1.0374   | 9     | -1.94 | 0.0838 |
| trt    | 2    | 4    | 4     | -5.4000 | 1.0374   | 9     | -5.21 | 0.0006 |
| trt    | 3    | 4    | 4     | -3.3833 | 1.0374   | 9     | -3.26 | 0.0098 |
| pest   | 1    | 2    | 2     | -6.8750 | 2.8549   | 3     | -2.41 | 0.0952 |
| pest   | 1    | 3    | 3     | -6.9875 | 2.8549   | 3     | -2.45 | 0.0919 |
| pest   | 2    | 3    | 3     | -0.1125 | 2.8549   | 3     | -0.04 | 0.9710 |
Corn yield, irrigation, pesticide and cultivars continued

So the treatment effect $B$ is highly significant ($p = 0.0003$). Are there particular comparisons among the three treatments that are of interest? There are because real experiment is actually slightly more complicated than previously described. The factor $B$ is really a $2 \times 2$ combination of irrigation and cultivar:

<table>
<thead>
<tr>
<th>$B$</th>
<th>Irr</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>yes</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>yes</td>
<td>2</td>
</tr>
</tbody>
</table>

The 3 $df$ for the within plot factor $B$ can be broken up into three 1 $df$ components due to main effect of irr, main effect of CV and interaction. Same with the $AB$ interaction. The plot below averages over pesticide and field:
proc mixed method=type3;
   class field pest irr cv trt;
   *model y=trt|pest/ddfm=satterth;
   model y=irr|cv|pest/ddfm=satterth;
   random field(pest);
   *parms /nound;
   *lsmeans trt pest/pdiff adj=tukey;
   lsmeans irr cv /pdiff;
   lsmeans irr*cv;
run;

The SAS System
The Mixed Procedure
Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>field</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>pest</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>irr</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>cv</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>trt</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>irr</td>
<td>1</td>
<td>133.953750</td>
<td>133.953750</td>
<td>Var(Residual) + Q(irr, irr<em>cv, pest</em>irr, pest<em>irr</em>cv)</td>
</tr>
<tr>
<td>cv</td>
<td>1</td>
<td>44.010417</td>
<td>44.010417</td>
<td>Var(Residual) + Q(cv, irr<em>cv, pest</em>cv, pest<em>irr</em>cv)</td>
</tr>
<tr>
<td>irr*cv</td>
<td>1</td>
<td>2.733750</td>
<td>2.733750</td>
<td>Var(Residual) + Q(irr<em>cv, pest</em>irr*cv)</td>
</tr>
<tr>
<td>pest</td>
<td>2</td>
<td>256.275833</td>
<td>128.137917</td>
<td>Var(Residual) + 4 Var(field(pest)) + Q(pest, pest<em>irr, pest</em>cv, pest<em>irr</em>cv)</td>
</tr>
<tr>
<td>pest*irr</td>
<td>2</td>
<td>17.747500</td>
<td>8.873750</td>
<td>Var(Residual) + Q(pest<em>irr, pest</em>irr*cv)</td>
</tr>
<tr>
<td>pest*cv</td>
<td>2</td>
<td>1.385833</td>
<td>0.692917</td>
<td>Var(Residual) + Q(pest<em>cv, pest</em>irr*cv)</td>
</tr>
<tr>
<td>pest<em>irr</em>cv</td>
<td>2</td>
<td>5.357500</td>
<td>2.678750</td>
<td>Var(Residual) + Q(pest<em>irr</em>cv)</td>
</tr>
<tr>
<td>field(pest)</td>
<td>3</td>
<td>97.806250</td>
<td>32.602083</td>
<td>Var(Residual) + 4 Var(field(pest))</td>
</tr>
<tr>
<td>Residual</td>
<td>9</td>
<td>29.058750</td>
<td>3.228750</td>
<td>Var(Residual)</td>
</tr>
</tbody>
</table>
Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>Error Term</th>
<th>DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>irr</td>
<td>MS(Residual)</td>
<td>9</td>
<td>41.49</td>
<td>0.0001</td>
</tr>
<tr>
<td>cv</td>
<td>MS(Residual)</td>
<td>9</td>
<td>13.63</td>
<td>0.0050</td>
</tr>
<tr>
<td>irr*cv</td>
<td>MS(Residual)</td>
<td>9</td>
<td>0.85</td>
<td>0.3815</td>
</tr>
<tr>
<td>pest</td>
<td>MS(field(pest))</td>
<td>3</td>
<td>3.93</td>
<td>0.1452</td>
</tr>
<tr>
<td>pest*irr</td>
<td>MS(Residual)</td>
<td>9</td>
<td>2.75</td>
<td>0.1171</td>
</tr>
<tr>
<td>pest*cv</td>
<td>MS(Residual)</td>
<td>9</td>
<td>0.21</td>
<td>0.8109</td>
</tr>
<tr>
<td>pest<em>irr</em>cv</td>
<td>MS(Residual)</td>
<td>9</td>
<td>0.83</td>
<td>0.4670</td>
</tr>
<tr>
<td>field(pest)</td>
<td>MS(Residual)</td>
<td>9</td>
<td>10.10</td>
<td>0.0031</td>
</tr>
<tr>
<td>Residual</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>field(pest)</td>
<td>7.3433</td>
</tr>
<tr>
<td>Residual</td>
<td>3.2287</td>
</tr>
</tbody>
</table>

Least Squares Means

| Effect | irr  | cv  | Estimate | Error | DF  | t Value | Pr > |t| |
|--------|------|-----|----------|-------|-----|---------|-------|
| irr    | 1    | 55.1333 | 1.2219 | 3.61 | 45.12 | <.0001 |
| irr    | 2    | 59.8583 | 1.2219 | 3.61 | 48.99 | <.0001 |
| cv     | 1    | 56.1417 | 1.2219 | 3.61 | 45.95 | <.0001 |
| cv     | 2    | 58.8500 | 1.2219 | 3.61 | 48.16 | <.0001 |
| irr*cv | 1    | 1    | 54.1167 | 1.3274 | 4.9 | 40.77 | <.0001 |
| irr*cv | 1    | 2    | 56.1500 | 1.3274 | 4.9 | 42.30 | <.0001 |
| irr*cv | 2    | 1    | 58.1667 | 1.3274 | 4.9 | 43.82 | <.0001 |
| irr*cv | 2    | 2    | 61.5500 | 1.3274 | 4.9 | 46.37 | <.0001 |

Differences of Least Squares Means

| Effect | irr  | cv  | _irr  | _cv  | Estimate | Error | DF  | t Value | Pr > |t| |
|--------|------|-----|-------|------|----------|-------|-----|---------|-------|
| irr    | 1    | 2   | -4.7250 | 0.7336 | 9 | -6.44 | 0.0001 |
| cv     | 1    | 2   | -2.7083 | 0.7336 | 9 | -3.69 | 0.0050 |
Split-plot in blocks (RCBSPD)

In the randomized block split-plot design, sets of homogeneous plots are be formed and levels of the whole plot factor are assigned to the plots within these sets in a restricted randomization. Assignment of levels of the split-plot factor are as in the CRSPD.

In the split-plot experiment with pesticide as the whole plot factor and irrigation × CV as the split-plot factor, suppose the six plots come from two farms, with three plots in each farm. Suppose that the three pesticide treatments are randomized to plots within farms. Renumbering plots (1,2,1,2,1,2) as (1,2,3,4,5,6) and supposing plots (2,3,6) come from farm 1 and plots (1,4,5) from farm 2, the data are given as

<table>
<thead>
<tr>
<th>Obs</th>
<th>farm</th>
<th>pest</th>
<th>plot</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>53.4</td>
<td>53.8</td>
<td>58.2</td>
<td>59.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>46.5</td>
<td>51.1</td>
<td>49.2</td>
<td>51.3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>54.3</td>
<td>56.3</td>
<td>60.4</td>
<td>64.5</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>57.2</td>
<td>56.9</td>
<td>61.6</td>
<td>66.8</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>55.9</td>
<td>58.6</td>
<td>62.4</td>
<td>64.5</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>57.4</td>
<td>60.2</td>
<td>57.2</td>
<td>62.7</td>
</tr>
</tbody>
</table>

At the whole plot level (ignoring the split-plot factor), the $df$ in an ANOVA for pesticide effects are given by

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>A : Pesticide</td>
<td>2</td>
</tr>
<tr>
<td>Farms</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
</tr>
</tbody>
</table>

so that an $F$-ratio for the pesticide effect is based on $df = 2, 2$
In general, for a RCBSPD with \( a \) levels of a whole-plot level (\( A \)) randomized to \( r \) blocks (for a total of \( ra \) plots) and \( b \) levels of a split-plot factor (\( B \)) within each plot, the model and ANOVA table are given by

\[
Y_{ijk} = \mu + \alpha_i + R_k + \beta_j + (\alpha\beta)_{ij} + (SR)_{ik} + E_{ijk}
\]

where

- \( i \) denotes level of \( A \),
- \( j \) denotes level of \( B \),
- \( k \) denotes block.

\( R_k \overset{iid}{\sim} N(0, \sigma_r^2) \) and \( SR_{ik} \overset{iid}{\sim} N(0, \sigma_{sr}^2) \). All random errors are mutually independent.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>( a - 1 )</td>
<td>( \sigma^2 + b\sigma_{sr}^2 + br\psi_A^2 )</td>
</tr>
<tr>
<td>Blocks</td>
<td>( r - 1 )</td>
<td>( \sigma^2 + b\sigma_{sr}^2 + ab\sigma_r^2 )</td>
</tr>
<tr>
<td>Whole plot error (Block×( A ))</td>
<td>( (r - 1)(a - 1) )</td>
<td>( \sigma^2 + b\sigma_{sr}^2 )</td>
</tr>
<tr>
<td>( B )</td>
<td>( b - 1 )</td>
<td>( \sigma^2 + ar\psi_B^2 )</td>
</tr>
<tr>
<td>( AB )</td>
<td>( (a - 1)(b - 1) )</td>
<td>( \sigma^2 + r\psi_{AB}^2 )</td>
</tr>
<tr>
<td>Error (( B\times Blocks(A) ))</td>
<td>( a(b - 1)(r - 1) )</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>Total</td>
<td>( abr - 1 )</td>
<td></td>
</tr>
</tbody>
</table>
data one; /* 'one' is original dataset */
   set one;
   if plot in (2,3,6) then farm=1;
   else farm=2;
run;
proc mixed method=type3;
   class farm plot pest irr cv trt;
   model y=pest|trt;
   random farm farm*pest;
run;

The Mixed Procedure

Class Levels Values
farm 2 1 2
plot 6 1 2 3 4 5 6
pest 3 1 2 3
trt 4 1 2 3 4

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>pest</td>
<td>2</td>
<td>256.275833</td>
<td>128.137917</td>
<td>Var(Residual) + 4 Var(farm<em>pest) + Q(pest,pest</em>trt)</td>
</tr>
<tr>
<td>trt</td>
<td>3</td>
<td>180.697917</td>
<td>60.232639</td>
<td>Var(Residual) + Q(trt,pest*trt)</td>
</tr>
<tr>
<td>pest*trt</td>
<td>6</td>
<td>24.490833</td>
<td>4.081806</td>
<td>Var(Residual) + Q(pest*trt)</td>
</tr>
<tr>
<td>farm</td>
<td>1</td>
<td>59.220417</td>
<td>59.220417</td>
<td>Var(Residual) + 4 Var(farm*pest) + 12 Var(farm)</td>
</tr>
<tr>
<td>farm*pest</td>
<td>2</td>
<td>38.585833</td>
<td>19.292917</td>
<td>Var(Residual) + 4 Var(farm*pest)</td>
</tr>
<tr>
<td>Residual</td>
<td>9</td>
<td>29.058750</td>
<td>3.228750</td>
<td>Var(Residual)</td>
</tr>
</tbody>
</table>

Cov Parm Estimate

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>farm</td>
<td>3.3273</td>
</tr>
<tr>
<td>farm*pest</td>
<td>4.0160</td>
</tr>
<tr>
<td>Residual</td>
<td>3.2287</td>
</tr>
</tbody>
</table>

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>pest</td>
<td>2</td>
<td>2</td>
<td>6.64</td>
<td>0.1309</td>
</tr>
<tr>
<td>trt</td>
<td>3</td>
<td>9</td>
<td>18.66</td>
<td>0.0003</td>
</tr>
<tr>
<td>pest*trt</td>
<td>6</td>
<td>9</td>
<td>1.26</td>
<td>0.3607</td>
</tr>
</tbody>
</table>
Researchers for an ice cream manufacturer conduct an experiment to study the effects of variety (sweetcharlie, camarosa, and gaviota) and mixing speed (slow, medium and fast) on ice cream quality. One batch of each variety of strawberries is sampled on Monday over four consecutive weeks. Each batch is divided into three parts, which are randomized to the three mixing speeds and three quarts of iced cream are produced, stored for one month, then tested for texture quality, on a scale from 1-100.

Data:

<table>
<thead>
<tr>
<th>Obs</th>
<th>week</th>
<th>variety</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>c</td>
<td>47</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>g</td>
<td>48</td>
<td>51</td>
<td>54</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>s</td>
<td>46</td>
<td>49</td>
<td>51</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>c</td>
<td>43</td>
<td>46</td>
<td>49</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>g</td>
<td>51</td>
<td>55</td>
<td>53</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>s</td>
<td>48</td>
<td>49</td>
<td>48</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>c</td>
<td>47</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>g</td>
<td>50</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>s</td>
<td>44</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>c</td>
<td>44</td>
<td>49</td>
<td>51</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>g</td>
<td>53</td>
<td>53</td>
<td>58</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>s</td>
<td>48</td>
<td>51</td>
<td>52</td>
</tr>
</tbody>
</table>
Model

\[ Y_{ijk} = \mu_{ij} + B_k + (\alpha B)_{ik} + E_{ijk} \]

where

\[ i = 1, 2, 3 = a(\text{ variety}) \]
\[ j = 1, 2, 3 = b(\text{ speed}) \]
\[ k = 1, 2, 3, 4 = r(\text{ week}) \]

ANOVA sketch

<table>
<thead>
<tr>
<th>Source</th>
<th>( df )</th>
<th>Expected MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variety</td>
<td>( a - 1 = 2 )</td>
<td></td>
</tr>
<tr>
<td>Block</td>
<td>( r - 1 = 3 )</td>
<td></td>
</tr>
<tr>
<td>V×B</td>
<td>( (a - 1)(r - 1) = 6 )</td>
<td></td>
</tr>
<tr>
<td>Speed</td>
<td>( b - 1 = 2 )</td>
<td></td>
</tr>
<tr>
<td>V×S</td>
<td>( (a - 1)(b - 1) = 4 )</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>( (b - 1)(r - 1)a = 18 )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( abr - 1 = 35 )</td>
<td></td>
</tr>
</tbody>
</table>
/* SAS code for split-plot in blocks */

data one;
    infile "strawberryice.dat" firstobs=3;
    input week variety $ speed $ tq;
run;
*goptions dev=ps;
axis1 offset=(1 cm,1 cm) label=(height=2 "variety")
    value=(height=2);
axis2 offset=(1 cm,1 cm) label=(height=2 "TQ")
    value=(height=2);

symbol1 value=dot c=black h=1.2;
symbol2 value=plus c=black h=1.2;
symbol3 value=diamond c=black h=1.2;

proc gplot data=one;
    title "Ice cream texture quality";
    plot tq*variety=speed/haxis=axis1 vaxis=axis2;
    *plot tq*speed=variety;
run; quit;
proc mixed data=one method=type3;
    class week variety speed;
    model tq=variety|speed;
    random week week*variety;
    lsmeans variety speed/diff;
    lsmeans variety*speed;
run;
The SAS System
The Mixed Procedure

Model Information

Data Set WORK.ONE
Dependent Variable tq
Covariance Structure Variance Components
Estimation Method Type 3
Residual Variance Method Factor
Fixed Effects SE Method Model-Based
Degrees of Freedom Method Containment

Class Levels Values

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>week</td>
<td>4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>variety</td>
<td>3</td>
<td>c g s</td>
</tr>
<tr>
<td>speed</td>
<td>3</td>
<td>Low Medium high</td>
</tr>
</tbody>
</table>

Dimensions

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariance Parameters</td>
<td>3</td>
</tr>
<tr>
<td>Columns in X</td>
<td>16</td>
</tr>
<tr>
<td>Columns in Z</td>
<td>16</td>
</tr>
<tr>
<td>Subjects</td>
<td>1</td>
</tr>
<tr>
<td>Max Obs Per Subject</td>
<td>36</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>variety</td>
<td>2</td>
<td>148.666667</td>
<td>74.333333</td>
</tr>
<tr>
<td>speed</td>
<td>2</td>
<td>112.166667</td>
<td>56.083333</td>
</tr>
<tr>
<td>variety*speed</td>
<td>4</td>
<td>2.166667</td>
<td>0.541667</td>
</tr>
<tr>
<td>week</td>
<td>3</td>
<td>19.222222</td>
<td>6.407407</td>
</tr>
<tr>
<td>week*variety</td>
<td>6</td>
<td>33.111111</td>
<td>5.518519</td>
</tr>
<tr>
<td>Residual</td>
<td>18</td>
<td>37.666667</td>
<td>2.092593</td>
</tr>
</tbody>
</table>

Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>Expected Mean Square</th>
<th>Error Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>variety</td>
<td>Var(Residual) + 3 Var(week<em>variety) + Q(variety,variety</em>speed)</td>
<td>MS(week*variety)</td>
</tr>
<tr>
<td>speed</td>
<td>Var(Residual) + Q(speed,variety*speed)</td>
<td>MS(Residual)</td>
</tr>
<tr>
<td>variety*speed</td>
<td>Var(Residual) + Q(variety*speed)</td>
<td>MS(Residual)</td>
</tr>
<tr>
<td>week</td>
<td>Var(Residual) + 9 Var(week)</td>
<td>MS(week*variety)</td>
</tr>
<tr>
<td>week*variety</td>
<td>Var(Residual) + 3 Var(week*variety)</td>
<td>MS(Residual)</td>
</tr>
<tr>
<td>Residual</td>
<td>Var(Residual)</td>
<td>.</td>
</tr>
</tbody>
</table>
### Type 3 Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>variety</td>
<td>6</td>
<td>13.47</td>
<td>0.0060</td>
</tr>
<tr>
<td>speed</td>
<td>18</td>
<td>26.80</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>variety*speed</td>
<td>18</td>
<td>0.26</td>
<td>0.9004</td>
</tr>
<tr>
<td>week</td>
<td>6</td>
<td>1.16</td>
<td>0.3990</td>
</tr>
<tr>
<td>week*variety</td>
<td>18</td>
<td>2.64</td>
<td>0.0516</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>week</td>
<td>0.09877</td>
</tr>
<tr>
<td>week*variety</td>
<td>1.1420</td>
</tr>
<tr>
<td>Residual</td>
<td>2.0926</td>
</tr>
</tbody>
</table>

### Fit Statistics

-2 Res Log Likelihood 118.2
AIC (smaller is better) 124.2
AICC (smaller is better) 125.3
BIC (smaller is better) 122.4

### Least Squares Means

| Effect          | variety | speed | Estimate | Error | DF | t Value | Pr > |t| |
|-----------------|---------|-------|----------|-------|----|---------|------|---|
| variety         | c       |       | 48.0000  | 0.6961| 6  | 68.95   | <.0001|
| variety         | g       |       | 52.6667  | 0.6961| 6  | 75.66   | <.0001|
| variety         | s       |       | 48.8333  | 0.6961| 6  | 70.15   | <.0001|
| speed           | Low     |       | 47.4167  | 0.5424| 18 | 87.41   | <.0001|
| speed           | Medium  |       | 50.5000  | 0.5424| 18 | 93.10   | <.0001|
| speed           | high    |       | 51.5833  | 0.5424| 18 | 95.10   | <.0001|
| variety*speed   | c Low   |       | 45.2500  | 0.9129| 18 | 49.57   | <.0001|
| variety*speed   | c Medium|       | 48.7500  | 0.9129| 18 | 53.40   | <.0001|
| variety*speed   | c high  |       | 50.0000  | 0.9129| 18 | 54.77   | <.0001|
| variety*speed   | g Low   |       | 50.5000  | 0.9129| 18 | 55.32   | <.0001|
| variety*speed   | g Medium|       | 53.0000  | 0.9129| 18 | 58.06   | <.0001|
| variety*speed   | g high  |       | 54.5000  | 0.9129| 18 | 59.70   | <.0001|
| variety*speed   | s Low   |       | 46.5000  | 0.9129| 18 | 50.94   | <.0001|
| variety*speed   | s Medium|       | 49.7500  | 0.9129| 18 | 54.50   | <.0001|
| variety*speed   | s high  |       | 50.2500  | 0.9129| 18 | 55.05   | <.0001|
### Differences of Least Squares Means

<table>
<thead>
<tr>
<th>Effect</th>
<th>variety</th>
<th>speed</th>
<th>_variety</th>
<th>_speed</th>
<th>Estimate</th>
<th>Error</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>variety</td>
<td>c</td>
<td>g</td>
<td></td>
<td></td>
<td>-4.6667</td>
<td>0.9590</td>
<td>6</td>
</tr>
<tr>
<td>variety</td>
<td>c</td>
<td>s</td>
<td></td>
<td></td>
<td>-0.8333</td>
<td>0.9590</td>
<td>6</td>
</tr>
<tr>
<td>variety</td>
<td>g</td>
<td>s</td>
<td></td>
<td></td>
<td>3.8333</td>
<td>0.9590</td>
<td>6</td>
</tr>
<tr>
<td>speed</td>
<td>Low</td>
<td>Medium</td>
<td></td>
<td></td>
<td>-3.0833</td>
<td>0.5906</td>
<td>18</td>
</tr>
<tr>
<td>speed</td>
<td>Low</td>
<td>high</td>
<td></td>
<td></td>
<td>-4.1667</td>
<td>0.5906</td>
<td>18</td>
</tr>
<tr>
<td>speed</td>
<td>Medium</td>
<td>high</td>
<td></td>
<td></td>
<td>-1.0833</td>
<td>0.5906</td>
<td>18</td>
</tr>
</tbody>
</table>

### Differences of Least Squares Means

| Effect  | variety | speed | _variety | _speed | t Value | Pr > |t|  |
|---------|---------|-------|----------|--------|---------|-------|---|
| variety | c       | g     |          |        | -4.87   | 0.0028|   |
| variety | c       | s     |          |        | -0.87   | 0.4183|   |
| variety | g       | s     |          |        | 4.00    | 0.0071|   |
| speed   | Low     | Medium|         |        | -5.22   | <.0001|   |
| speed   | Low     | high  |         |        | -7.06   | <.0001|   |
| speed   | Medium  | high  |         |        | -1.83   | 0.0832|   |