MATRICES

1. Rectangular array of numbers

\[
\begin{bmatrix}
3 & 5 & 7 & 8 \\
1 & 2 & 3 & 7
\end{bmatrix} = A_{2\times4}
\]

Symbol - cap letter \( A, B, C \)

2. Vectors

3. Elements \( a_{ij} \)

4. Operations

(a) Addition or subtraction
   
   Element by element

(b) Multiplication

Vector \((1, 3, -5, 1)\)

\[
\begin{bmatrix}
2 \\
0 \\
3 \\
-2
\end{bmatrix} = 2 + 0 - 15 - 2 = -15
\]
Matrices

\[
\begin{bmatrix}
1 & 3 & 5 \\
1 & -2 & 2
\end{bmatrix}
\begin{bmatrix}
2 & 3 \\
1 & 2
\end{bmatrix}
= 
\begin{bmatrix}
7 & 8 \\
-2 & -15
\end{bmatrix}
\]

Note: In general, \(BA\) does not equal \(AB\) (may have different dimensions)

(c) Transpose A’

Note: \((AB)’ = (B’) (A’)\)

d) Scalar Multiplication

\[
3 \begin{bmatrix}
1 & 2 \\
-2 & 3
\end{bmatrix}
= 
\begin{bmatrix}
3 & 6 \\
-6 & 9
\end{bmatrix}
\]

5. Identity Matrix I

\[
I = 
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\]

\(IA = AI = A\) for \(A_{3 \times 3}\)

\(IB = B\) for \(B_{3 \times C}\)

\(CI = C\) for \(C_{r \times 3}\)
6. Rank and dependence

\[ A = \begin{bmatrix} 1 & 1 & 1 \\ 3 & 1 & 5 \\ 2 & 3 & 1 \end{bmatrix} \text{ columns } = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}, \text{ and } \begin{bmatrix} 1 \end{bmatrix} \]

Note: \( 2C_1 - C_2 - C_3 = \Phi \) where \( \Phi \) is a column of 0's.

We say that \( C_1, C_2, \) and \( C_3 \) are linearly dependent.

In general: \( k \) columns \( C_1, C_2, \ldots, C_k \) are dependent if there exist scalars \( \lambda_1, \lambda_2, \ldots, \lambda_k \) such that

\[
\begin{align*}
(1) \quad & \lambda_1 C_1 + \lambda_2 C_2 + \ldots + \lambda_k C_k = \Phi \\
\text{and} \quad & \text{(2) At least one of the } \lambda \text{'s is not 0.}
\end{align*}
\]

If the \( k \) columns are not dependent we call them linearly independent. The combination of columns in (1) is called a linear combination, a phrase we will often use. Thus, \( k \) columns are linearly independent if the only linear combination of them which will produce the zero vector is the linear combination with all \( \lambda \)'s 0. Often we will collect the \( \lambda \)'s together in a vector \( \Lambda \).

The rank of a matrix is the maximum number of linearly independent columns which can be selected from the columns of the matrix. Thus the rank of \( A \) is two. Notice that if the rank
of a matrix is 1, then there is one column such that all other columns are direct multiples.

For any matrix \( X \), the rank of \( X \) is the same as the rank of \( X'X \). The row rank of any matrix is always equal to the column rank.

7. Inverse of a matrix.

Symbol: \( A^{-1} \)

The inverse of an \( n \times n \) matrix \( A \) is an \( n \times n \) matrix \( B \) such that \( AB = I \). Such a matrix \( B \) will exist only if \( A \) is of rank \( n \). In this case it is also true that \( BA = I \).

Example:

\[
A = \begin{bmatrix}
1 & -1 & 1 \\
2 & 1 & 0 \\
3 & 1 & 1
\end{bmatrix}
\quad A^{-1} = \begin{bmatrix}
0.5 & 1 & -0.5 \\
-1.0 & -1 & 1.0 \\
-0.5 & -2 & 1.5
\end{bmatrix}
\]

(Check by multiplication.)
Solving equations

\[
\begin{align*}
    b_0 - b_1 + b_2 &= 2 \\
    2b_0 + b_1 &= 7 \\
    3b_0 + b_1 + b_2 &= -5
\end{align*}
\]

\[
\begin{bmatrix}
    1 & -1 & 1 \\
    2 & 1 & 0 \\
    3 & 1 & 1
\end{bmatrix}
\begin{bmatrix}
    b_0 \\
    b_1 \\
    b_2
\end{bmatrix}
=
\begin{bmatrix}
    2 \\
    7 \\
    -5
\end{bmatrix}
\]

\[
\begin{bmatrix}
    b_0 \\
    b_1 \\
    b_2
\end{bmatrix}
=
\begin{bmatrix}
    0.5 & 1 & -0.5 \\
    -1.0 & -1 & 1.0 \\
    -0.5 & -2 & 1.5
\end{bmatrix}
\begin{bmatrix}
    2 \\
    7 \\
    -5
\end{bmatrix}
=
\begin{bmatrix}
    10.5 \\
    -14.0 \\
    -22.5
\end{bmatrix}
\]

\[Ab = c \Rightarrow b = A^{-1} c\]

For a 2 × 2 matrix only we have the formula:

\[
\begin{bmatrix}
    a & b \\
    c & d
\end{bmatrix}^{-1}
=
\frac{1}{ad - bc}
\begin{bmatrix}
    d & -b \\
    -c & a
\end{bmatrix}
\]

Note: if A and B are square matrices then \( A^{-1} B^{-1} = (BA)^{-1} \)
RANDOM VECTORS

We have described a univariate random variable like weight $W$ by writing

$$W \sim N(150, 100)$$

We might also measure height in inches $H$ and have

$$H \sim N(68, 16)$$

Now the above tells us nothing about how the two variables height and weight covary. The covariance between $H$ and $W$ tells us how weight depends on height. If we know an individual is taller than the mean 68, would we predict that his weight will exceed the mean 150? If so we are claiming a positive covariance between height and weight. Formally, recall that the variance of weights is defined as an expected value, namely

$$\text{variance } (W) = E\left\{ (W - 150)^2 \right\}.$$ 

The covariance between $W$ and $H$ is defined as

$$\text{cov } (W, H) = E\left\{ (W - 150) (H - 68) \right\}.$$ 

Suppose the covariance is $\text{cov}(W, H) = 30$. We put this all together as
\[
\begin{bmatrix}
W \\
H
\end{bmatrix} \sim \text{MVN} \left[
\begin{bmatrix}
150 \\
68
\end{bmatrix},
\begin{bmatrix}
100 & 30 \\
30 & 16
\end{bmatrix}
\right]
\]

In general we write
\[Y \sim \text{MVN} (\mu, V)\]
where \(Y\) is a vector with \(i^{th}\) element \(Y_i\), \(\mu\) is a vector with \(i^{th}\) element \(\mu_i\) and \(V\) is a matrix whose \(ij^{th}\) element is the covariance between \(Y_i\) and \(Y_j\).

******

Fact: If \(A\) and \(B\) are matrices of constants and \(Y\) is a random vector with

\[Y \sim \text{MVN} (\mu, V)\]

then
\[AY + B \sim \text{MVN} (A\mu + B, AVA')\]

Example:
\[
Y \sim \text{MVN} \left[
\begin{bmatrix}
4 \\
6
\end{bmatrix},
\begin{bmatrix}
8 & 5 & 0 \\
5 & 12 & 4 \\
10 & 0 & 9
\end{bmatrix}
\right]
\]

(1) Find the distribution of \(Z = Y_1 - Y_2 + Y_3\).

(2) Let \(W = Y_1 - 3Y_2 + 2Y_3\). Find the joint distribution of \(Z\) and \(W\).
REGRESSION - HAND CALCULATION

Review of calculations from first semester:

DATA
\[
\begin{array}{c|c|c|c|c}
X & 4 & 7 & 5 & 4 \\
Y & 5 & 7 & 6 & 6 \\
\end{array}
\]

\[
\begin{array}{ccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\end{array}
\]

\(\bar{X} = 5, \bar{Y} = 6\)

\[
\begin{array}{cccccccc}
X - \bar{X} & (X - \bar{X})^2 & Y - \bar{Y} & (Y - \bar{Y})^2 & (X - \bar{X})(Y - \bar{Y}) \\
-1 & 1 & -1 & 1 & 1 \\
2 & 4 & 1 & 1 & 2 \\
0 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 \\
0 & 6 & 0 & 2 & 3 \\
\end{array}
\]

slope = \(b = \frac{3}{6} = 0.5\)

intercept = \(a = 6 - (0.5)(5) = 3.5\)

True slope \(\beta.\) Recall that \(b \sim N(\beta, \sigma^2/\sum(X_i - \bar{X})^2).\)

True intercept \(\alpha.\)

Recall that \(a \sim N\left(\alpha, \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]\right)\)
Estimate $\sigma^2$ by $s^2 = (\text{total SSq} - \text{regn SSq})/\text{df}$

Regn. SSq = 3 * 3/6 = 1.5
Total SSq = 2

$df = n - 2 = 4 - 2 = 2$
so $s^2 = \text{MSE} = (2 - 1.5)/2 = 0.25$

Estimated variance of $a$ is $0.25(1/4 + 25/6) = 1.1042$
Estimated variance of $b$ is $0.25/6 = 0.041667$
Notice that 511 formulas gave no indication of covariance.

EXAMPLES:

Test for no relationship between $Y$ and $X$.
$H_0: \beta=0$, \quad $t = \frac{b-0}{\text{std. err. of } b} = .5 / \sqrt{.041667}$

Give 95% confidence interval for mean $Y$ at $X = 6$.
$3.5 + .5 (6) = 6.5 = \text{prediction}$
add and subtract $t$ times std. err. of prediction
std. err. of prediction = $\sqrt{\left[\frac{1}{n}+(6-\bar{X})^2/\sum(X_i - \bar{X})^2\right]} 0.25$

Give 95% prediction interval for individual $Y$ at $X = 6$.
same prediction, 6.5
std. err. of prediction = $\sqrt{\left[1+1/n+(6-\bar{X})^2/\sum(X_i - \bar{X})^2\right]} 0.25$
Example of a regression:

Wilson and Mather, JAMA 229 (1994)

<table>
<thead>
<tr>
<th>LINE</th>
<th>X</th>
<th>XSQ</th>
<th>XY</th>
<th>AGE</th>
<th>Y</th>
<th>YSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.75</td>
<td>0.552</td>
<td>0.3047</td>
<td>-26.3083</td>
<td>19</td>
<td>-47.66</td>
<td>2271.48</td>
</tr>
<tr>
<td>9.00</td>
<td>-0.198</td>
<td>0.0392</td>
<td>5.2787</td>
<td>40</td>
<td>-26.66</td>
<td>710.76</td>
</tr>
<tr>
<td>9.75</td>
<td>0.552</td>
<td>0.3047</td>
<td>11.780</td>
<td>21.34</td>
<td>455.40</td>
<td></td>
</tr>
<tr>
<td>9.00</td>
<td>-0.198</td>
<td>0.0392</td>
<td>-5.413</td>
<td>27.34</td>
<td>747.48</td>
<td></td>
</tr>
</tbody>
</table>

Wilson and Mather, JAMA 229 (1994)

Plot of AGE*LINE. Legend: A = 1 obs, B = 2 obs, etc.
Plot of YHAT*LINE. Symbol used is ' - '.
Discussion of correlation coefficients:

We define the (population) correlation between two variables as the covariance divided by the square root of the product of the variances. For height and weight as just given, 
\[ \rho = \frac{30}{\sqrt{100 \times 16}} = 0.75. \]
Now suppose we take \( n = 103 \) people and measure their height \( H_i \) and weight \( W_i \). As usual, we can estimate variances as

\[ S_{W}^2 = \frac{\sum (W_i - \bar{W})^2}{n - 1} \quad \text{and} \quad S_{H}^2 = \frac{\sum (H_i - \bar{H})^2}{n - 1}. \]

The covariance is estimated as

\[ S_{WH} = \frac{\sum (W_i - \bar{W})(H_i - \bar{H})}{n - 1}. \]

Now these are just estimates of the true values so let us assume we get, say,

\[ S_{W}^2 = 125, \quad S_{H}^2 = 20, \quad \text{and} \quad S_{WH} = 40. \]

Our estimated covariance matrix is then

\[ \hat{V} = \begin{bmatrix} 125 & 40 \\ 40 & 20 \end{bmatrix} \]

and we compute a sample estimate, \( r \), of \( \rho \) as

\[ r = \frac{40}{\sqrt{125 \times 20}} = 0.8. \]
Notice that

\[ r^2 = \frac{\left[ \sum(W_i - \bar{W})(H_i - \bar{H}) \right]^2}{\sum(W_i - \bar{W})^2 \sum(H_i - \bar{H})^2} = \text{regression SS/total SS} \]

from a regression of either H on W or of W on H. Thus we have explained 64% of the variability in W by regressing it on H.

A test that \( \rho = 0 \) is just the t-test on the coefficient in the regression of W on H (or equivalently H on W). To test any other hypothesis like \( H_0: \rho = 0.5 \) or to put a confidence interval around \( r \), Fisher's transformation to \( Z \) is used. We define

\[ Z_r = 0.5 \ln((1 + r)/(1 - r)) \]

and define \( Z_\rho \) similarly. Now approximately we have

\[ Z_r \sim N(Z_\rho, 1/(n - 3)) \]

Thus we get \( Z_{0.8} = 1.09861 \) so

\[ 1.09861 - 0.196 \]

and

\[ 1.09861 + 0.196 \]

are the lower and upper confidence bounds for \( Z_\rho \).

Converting from \( Z_\rho \) to \( \rho \) we get \( 0.71 \leq \rho \leq 0.86 \). Table A.12 can be used to do the conversions or simply do them on a hand calculator.
REGRESSION IN MATRIX FRAMEWORK

The equations: (Henceforth intercept is $\beta_0$, slope is $\beta_1$)

\[
\begin{align*}
5 &= \beta_0 + 4 \beta_1 + e_1 \\
7 &= \beta_0 + 7 \beta_1 + e_2 \\
6 &= \beta_0 + 5 \beta_1 + e_3 \\
6 &= \beta_0 + 4 \beta_1 + e_4
\end{align*}
\]

Matrix form:

\[
\begin{bmatrix}
5 \\
7 \\
6 \\
6
\end{bmatrix} =
\begin{bmatrix}
1 & 4 \\
1 & 7 \\
1 & 5 \\
1 & 4
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1
\end{bmatrix} +
\begin{bmatrix}
e_1 \\
e_2 \\
e_3 \\
e_4
\end{bmatrix}
\Rightarrow
Y = X \beta + e
\]

Estimate $\beta$ by $b$ such that $(Y - Xb)'(Y - Xb)$ is minimized.

Note: this is the sum of squares of residuals $Y - Xb$.

Using calculus we can show the sum of squares is minimized by solving the following “normal equations.”

********

\[(X'X)b = X'Y \quad \text{i.e.} \quad b = (X'X)^{-1} (X'Y)\]

********

THIS FORMULA IS IMPORTANT
Our little example:

\[ X'X = \begin{bmatrix} 4 & 20 \\ 20 & 106 \end{bmatrix} \Rightarrow (X'X)^{-1} = \frac{1}{424 - 400} \begin{bmatrix} 106 & -20 \\ -20 & 4 \end{bmatrix} \]

We find the solution

\[ \mathbf{b} = \begin{bmatrix} 4.4167 \\ -0.8333 \end{bmatrix} \begin{bmatrix} -0.8333 \\ 0.1667 \end{bmatrix} \begin{bmatrix} 24 \\ 123 \end{bmatrix} = \begin{bmatrix} 3.5 \\ 0.5 \end{bmatrix} \]

Now let's relate the vector of estimated parameters (\( \mathbf{b} \)) to the vector of actual parameters (\( \boldsymbol{\beta} \)). We have

\[ \mathbf{b} = (X'X)^{-1} (X'Y) = (X'X)^{-1} (X' (\boldsymbol{\beta} + \mathbf{e})) = \boldsymbol{\beta} + (X'X)^{-1} (X'\mathbf{e}). \]

The difference between \( \mathbf{b} \) and \( \boldsymbol{\beta} \) is \( \mathbf{b} - \boldsymbol{\beta} = (X'X)^{-1}(X'\mathbf{e}) \).

Results:

(1) \( \mathbf{b} \) is an unbiased estimate of \( \boldsymbol{\beta} \).
(2) The variance-covariance matrix of \( b \) (\( V_b \)) is related to that of \( e \) (\( V_e \)) by the formula

\[
V_b = ((X'X)^{-1} X'V_e X(X'X)^{-1})
\]

(3) Let us assume that all the \( e \)'s have the same variance, have mean 0, and are uncorrelated. That means

\[
V_e = \sigma^2 I
\]

and after all the smoke clears, we obtain the crucial formula

\[
V_b = \text{var}(b) = (X'X)^{-1} \sigma^2
\]

**ESTIMATION OF VARIANCE OF \( e \)**

(1) Sum of squared residuals

(sum of squared errors, SSE)

\[
Y - Xb \text{ vector of residuals}
\]

For any vector \( a \), \( a' a \) is sum of squared elements
\[ \text{SSE} = (Y - Xb)' (Y - Xb) \]

\[ = \frac{Y'Y}{\text{Uncorr. Total SSq}} - b'X'Y \]
\[ = \frac{(Y'Y - n\bar{y}^2)}{\text{Corrected total SSq}} - \frac{(b'X'Y - n\bar{y}^2)}{\text{Corrected Regn. SSq}} \]

(2) Error df = n - 2 (2 = one intercept + one slope)

\[ \text{MSE} = \frac{\text{SSE}}{\text{df}} \]

(3) For our little example, check this against ST511 formula.

(4) Error degrees of freedom will always give degrees of freedom for \( t \) statistics.

Example (Continued.)

Variance-covariance matrix of parameter estimates

\[ \mathbf{V}_b = \begin{bmatrix} 4.4167 & -0.8333 \\ -0.8333 & 0.1667 \end{bmatrix} (0.25) = \begin{bmatrix} 1.1042 & -0.2083 \\ -0.2083 & 0.0417 \end{bmatrix} \]

(1) Test that slope is 0:

\[ t = \frac{0.5}{\sqrt{0.0417}} \]

(2 degrees of freedom)

\[ = \frac{0.5}{0.2041} = 2.45 \]
(2) Test that intercept is 0: \[ t = \frac{3.5}{\sqrt{1.1042}} = 3.33 \]
(2 degrees of freedom)

(3) Estimate mean value of \( Y \) at \( X = 6 \) and give 95% confidence interval.

Estimate is \( 3.5 + 0.5 \times 6 = 6.5 \)

We are estimating \( \beta_0 + 6 \beta_1 = (1, 6) \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \)

Our estimate is \( (1, 6) b = b_0 + 6b_1 = 3.5 + (6)(0.5) = 6.5 \)

Letting \( \Lambda' = (1, 6) \) we now want the variance of \( \Lambda' b \) but we know how to get that:

\[ \text{Var} (\Lambda' b) = \Lambda' V b \Lambda = (1, 6) \begin{bmatrix} 1.1042 & -0.2083 \\ -0.2083 & 0.0417 \end{bmatrix} \begin{bmatrix} 1 \\ 6 \end{bmatrix} = 0.1042 \]

95% confidence interval: \[ 6.5 \pm t * \sqrt{0.1042} \]

(4) Predict future individual value at \( X = 6 \)

Individual will differ from mean by a deviation with variance estimated by MSE = 0.25. Any future individual value at \( X = 6 \) is equal to mean at \( X = 6 \) plus deviation. Thus the variance is the sum of the variances of these two parts, namely, \( 0.25 + 0.1042 = 0.3542 \). Notice that the prediction interval
for an individual future value is wider than the corresponding confidence interval for the mean. We get, for our little example,

\[ 6.5 \pm t \cdot \sqrt{0.3542} \]

Since \( t \) has 2 d.f. we obtain \( t = 4.30 \) and thus

Confidence interval (5.11, 7.89)

Prediction interval (3.94, 9.06)

Example: You are in charge of predicting wheat yields from rainfall through July for your country so that you can place import quotas in early August. You have historic data on rain and yields. Which of the above formulas do you use and why?

Example: 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?)
SUMMARY OF REGRESSION FORMULAS

Model: \( Y = X \beta + e \) where \( e \sim \text{MVN}(0, \sigma^2 I) \)

Normal Equations: \( X'X b = X'Y \)
Solution: \( b = (X'X)^{-1} (X'Y) \) provided \( X'X \) is full rank

Estimate of variance-covariance matrix of \( b \): \( (X'X)^{-1} \text{MSE} \)

Predictions for observed \( Y \)'s is: vector \( \hat{X}b \).
We write \( \hat{Y} = \hat{X}b \) (\(^\wedge\) denotes predictor).

Residuals: \( = Y - \hat{X}b \)
SSE = \( (Y - \hat{X}b)' (Y - \hat{X}b) = Y'Y - b'X'Y = \)
\( \text{total SSq} - \text{regn. SSq.} \)

\( df = n \text{ minus rank of } X \text{ matrix.} \)
(Usually, rank = number of columns in \( X \), i.e. full rank.)

Prediction of future \( Y \)'s at some configuration of \( X \)'s:
Prediction written as \( \hat{Y} = \Lambda' b \) where \( \Lambda' = (1, X \text{-values}) \). In our example \( \Lambda' = (1, 6) \).
Variances of predictions:

for mean at configuration of X's in $\Delta$:
$$(\Delta'(X'X)^{-1} \Delta) \text{ MSE}$$

for individual at configuration of X's in $\Delta$:
$$(\Delta'(X'X)^{-1} \Delta + 1) \text{ MSE}$$

ANOVA (Column of 1's and k other columns.)

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>k</td>
<td>$b'X'Y - CT$</td>
</tr>
<tr>
<td>error</td>
<td>$n - k - 1$</td>
<td>$Y'Y - b'X'Y$</td>
</tr>
<tr>
<td>total</td>
<td>$n - 1$</td>
<td>$Y'Y - CT$</td>
</tr>
</tbody>
</table>

$R^2$ = coefficient of determination

$= \frac{\text{SSq(model)}}{\text{SSq(total)}}$

$= \frac{\text{corrected regn. SSq}}{\text{corrected total SSq}}$

$= 1 - \frac{\text{SSq(error)}}{\text{SSq(total)}}$

$$F^k_{n-k-1} = \frac{\text{MS(MODEL)}}{\text{MS(ERROR)}}$$

tests $H_0$: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
TYPE I (sequential) and TYPE III (partial) SUMS OF SQUARES

DATA

<table>
<thead>
<tr>
<th>Y</th>
<th>X0</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>1</td>
<td>-2</td>
<td>1</td>
</tr>
<tr>
<td>-3</td>
<td>1</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

\[ \mathbf{X}'\mathbf{X} \quad (\mathbf{X}'\mathbf{X})^{-1} \quad \mathbf{X}'\mathbf{Y} \quad \mathbf{b} \]

REGRESS Y ON X0 ONLY

\[
\begin{bmatrix} 4 & 1/4 \end{bmatrix} \begin{bmatrix} -3 \\ -3/4 \end{bmatrix}
\]

(SSR = b'X'Y = 9/4 = 2.25 = CT)

REGRESS Y ON X0, X1

\[
\begin{bmatrix} 4 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} 1/4 & 0 \\ 0 & 1/10 \end{bmatrix} \begin{bmatrix} -3 \\ 9 \end{bmatrix} \begin{bmatrix} -3/4 \\ 9/10 \end{bmatrix}
\]

(SSR = b'X'Y = 2.25 + 8.1 = 10.35)

REGRESS Y ON X0, X1, X2

\[
\begin{bmatrix} 4 & 0 & 3 \\ 0 & 10 & 0 \\ 3 & 0 & 5 \end{bmatrix} \begin{bmatrix} 10/22 & 0 & -3/11 \\ 0 & 1/10 & 0 \\ -3/11 & 0 & 4/11 \end{bmatrix} \begin{bmatrix} -3 \\ 9 \\ 2 \end{bmatrix} \begin{bmatrix} -21/11 \\ 9/10 \\ 17/11 \end{bmatrix}
\]

(SSR = 16.92)
Notes: No change in b0 from 1st to 2nd regression (orthogonality). Two b's change from 2nd to 3rd (not orthogonal).

Adding X2 to regression 2 increases the regression sum of squares from 10.35 to 16.92. We write

\[
\begin{align*}
R(X0, X1, X2) &= 16.92 \\
R(X0, X1) &= 10.35 \\
R(X2 | X0, X1) &= 6.57 = R(X0, X1, X2) - R(X0, X1)
\end{align*}
\]

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>TYPE I SSq (sequential)</th>
<th>TYPE III SSq (partial)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>R(X1</td>
<td>X0) = 8.1</td>
</tr>
<tr>
<td></td>
<td>(NOT usually equal)</td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>R(X2</td>
<td>X0, X1) = 6.57</td>
</tr>
<tr>
<td></td>
<td>R(X2</td>
<td>X0, X1) = 6.57</td>
</tr>
</tbody>
</table>

Note: The only reason type I and type III are equal for X1 is orthogonality. Generally they are not equal. Obviously type I = type III for the last X you have (X2 in our case).

Note: Partial SSq is "Type II" in PROC REG, " Type III " in GLM.
EXAMPLE 2:

DATA

<table>
<thead>
<tr>
<th>Y</th>
<th>X0</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>1</td>
<td>-2</td>
<td>1</td>
</tr>
<tr>
<td>-3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Y’Y = 17   SS(total) = 17 - 9/4 = 14.75

\[ (X’X)^{-1} \]

REGRESS Y ON X0 ONLY

\[ X’X \]

\[ (X’X)^{-1} \]

\[ X’Y \]

\[ b \]

REGRESS Y ON X0, X1

\[ \begin{bmatrix} 4 & 2 \\ 2 & 10 \end{bmatrix} \]

\[ \begin{bmatrix} 10/36 & -2/36 \\ -2/36 & 4/36 \end{bmatrix} \]

\[ \begin{bmatrix} -3 \\ 3 \end{bmatrix} \]

\[ \begin{bmatrix} -1.0 \\ 0.5 \end{bmatrix} \]

(\[ SSR = b’X’Y = 3 + 1.5 = 4.5 \])

R(X1 | X0) = 4.5 - 2.25 = 2.25
REGRESS Y ON X0, X1, X2

<table>
<thead>
<tr>
<th>$X'X$</th>
<th>$(X'X)^{-1}$</th>
<th>$X'Y$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\begin{bmatrix} 4 &amp; 2 &amp; 4 \ 2 &amp; 10 &amp; 1 \ 4 &amp; 1 &amp; 10 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 0.4670 &amp; -0.0755 &amp; -0.1792 \ -0.0755 &amp; 0.1132 &amp; 0.0189 \ -0.1792 &amp; 0.0189 &amp; 0.1698 \end{bmatrix}$</td>
<td>$\begin{bmatrix} -3 \ 3 \ 4 \end{bmatrix}$</td>
<td>$\begin{bmatrix} -2.3443 \ 0.6415 \ 1.2736 \end{bmatrix}$</td>
</tr>
</tbody>
</table>

$(SSR = b'X'Y = 14.052)$

$R(X0, X1, X2) = 14.052$

$R(X0, X1) = 4.500$

$R(X2 | X0, X1) = 9.552 = R(X0, X1, X2) - R(X0, X1)$

**EXERCISE:** Fill in the blank above by regressing Y on X0 and X2, etc. Are type I and type III equal for X1 in this example?

(ANS: 3.6353, no.)
GRADE – IQ EXAMPLE

<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>

Use the class computing account to enter the study time data. Regress GRADE on IQ. Regress GRADE on TIME, IQ. Finally, regress GRADE on TIME IQ TI where TI = TIME*IQ. The TI variable could, for example, be created in your data step. For the regression of GRADE on TIME and IQ, use the option /I in PROC REG. This will output the \((X'X)^{-1}\) matrix.

ANOVA (Grade on IQ)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ</td>
<td>1</td>
<td>15.9393</td>
<td>15.9393</td>
<td>0.153</td>
</tr>
<tr>
<td>Error</td>
<td>6</td>
<td>625.935</td>
<td>104.32</td>
<td></td>
</tr>
</tbody>
</table>

D. A. Dickey
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

\[
\text{ANOVA (Grade on IQ, Study Time)}
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>596.12</td>
<td>298.06</td>
<td>32.57</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>45.76</td>
<td>9.15</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{array}{cccc}
\text{SOURCE} & \text{df} & \text{SEQ} & \text{PART} \\
\hline
\text{IQ} & 1 & 15.94 & 121.24 \\
\text{STUDY} & 1 & 580.18 & 580.18 \\
\end{array}
\]

Parameter | Estimate | t  | Pr > |t|  | Std. Err. |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>0.74</td>
<td>0.05</td>
<td>0.9656</td>
<td>16.26</td>
<td>(0.9851)</td>
</tr>
<tr>
<td>IQ</td>
<td>0.47</td>
<td>3.64</td>
<td>0.0149</td>
<td>0.13</td>
<td>(0.0076)</td>
</tr>
<tr>
<td>STUDY</td>
<td>2.10</td>
<td>7.96</td>
<td>0.0005</td>
<td>0.26</td>
<td>(3.64)</td>
</tr>
</tbody>
</table>

From this regression we also can get

\[
(X'X)^{-1} = \begin{bmatrix}
28.8985 & -2261 & -.2242 \\
-2261 & .0018 & .0011 \\
-.2242 & .0011 & .0076
\end{bmatrix}
\]

1. To test H0: Coefficient on IQ is 0 (Note: calculations done with extra decimal accuracy.)

(a) Using t-test \( t = 0.47/\sqrt{0.0018*9.15} = 3.64 \)
(b) Using type III F-test, \( F = 121.24/9.15 = 13.25 = t^2. \)
Note: The type III sum of squares is defined by setting $t^2 = F$. This means that type III SSq = $b^*b/c$ where $b$ is the coefficient being tested and $c$ is the diagonal element of $(X'X)^{-1}$ which corresponds to $b$. We have $(0.47)(0.47)/0.0018 = 121.24$.

2. Estimate the mean grade for the population of all potential students with IQ = 113 and study time = 14 hours.

(a) Write this estimate as $\Lambda'b$ where $\Lambda' = (1, 113, 14)$.

(b) Variance of this is $\Lambda'(X'X)^{-1} \Lambda'MSE = 1.303$.

(c) Prediction is $\Lambda'b = 83.64$.

(d) To get confidence interval,
$$83.64 \pm 2.571 \sqrt{1.303}$$

(e) Interval (80.71, 86.57)

3. Estimate grade for individual with 113 IQ and 14 hours study time.
$$83.64 \pm 2.571 \sqrt{1.303} + 9.15$$
$$(75.33, 91.95)$$

4. What percent of grade variability is explained by IQ, STUDY?
R^2 = (corrected regn. SSQ)/(corrected total SSq) = 596.12/641.88 = 93%

5. Notes: When a new column is added to a regression, all the coefficients and their t-statistics can change. The t's could go from significance to insignificance or vice-versa.

The exception to the above case is when the added column of $X$ is orthogonal to the original columns. This means that the new $X'X$ has the old $X'X$ in the upper left corner, the sum of squares of the new column as the bottom right element, and all other elements 0.

Rerun this example adding a row 113 14 . at the end of the dataset. The dot implies a missing value. Use the statement MODEL GRADE = IQ STUDY / P CLM; . Compare to part 2 above. Rerun again with CLI instead of CLM. Compare to part 3 above. Was the extra data row used in computing the regression coefficients?

OPTIONS LS = 80 NODATE;
DATA GRADES; INPUT IQ STUDY GRADE @@;
  ST_IQ = STUDY*IQ; 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
PROC REG;
  MODEL GRADE = IQ STUDY ST_IQ/SS1 SS2;
TITLE "GRADE AND STUDY TIME EXAMPLE FROM ST 512 NOTES";
PROC PLOT;
PLOT STUDY*IQ = '/' / VPOS = 35;
DATA EXTRA;
   INPUT IQ STUDY GRADE;
CARDS;
   113 14 .
DATA BOTH;
   SET GRADES EXTRA;
PROC REG;
   MODEL GRADE = IQ STUDY/P CLM;
RUN;

GRADE AND STUDY TIME EXAMPLE FROM ST512 NOTES

DEP VARIABLE: GRADE

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARE</th>
<th>F VALUE</th>
<th>PROB&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>3</td>
<td>610.810</td>
<td>203.603</td>
<td>26.217</td>
<td>0.0043</td>
</tr>
<tr>
<td>ERROR</td>
<td>4</td>
<td>31.064674</td>
<td>7.766169</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C TOTAL</td>
<td>7</td>
<td>641.875</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ROOT MSE 2.786785  R-SQUARE 0.9516
DEP MEAN 81.375000  ADJ R-SQ 0.9153
C.V. 2.42462

| VARIABLE | DF | ESTIMATE | ERROR | T For H0: PARAMETER = 0 | PROB>|T| | TYPE I SS |
|----------|----|----------|-------|-------------------------|------|---------|
| INTERCEP | 1  | 72.206076| 54.072776| 1.335 0.2527 52975.125 |
| IQ       | 1  | -0.131170| 0.455300 | -0.288 0.7876 15.939299 |
| STUDY    | 1  | -4.111072| 4.524301 | -0.909 0.4149 580.176 |
| ST_IQ    | 1  | 0.053071 | 0.038581 | 1.376 0.2410 14.695210 |

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>DF</th>
<th>TYPE II SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEP</td>
<td>1</td>
<td>13.848316</td>
</tr>
<tr>
<td>IQ</td>
<td>1</td>
<td>0.644589</td>
</tr>
<tr>
<td>STUDY</td>
<td>1</td>
<td>6.412303</td>
</tr>
<tr>
<td>ST_IQ</td>
<td>1</td>
<td>14.695210</td>
</tr>
</tbody>
</table>
Discussion of the interaction model. We call the product \( I \times S = IQ \times STUDY \) an “interaction" term. Our model is

\[
\hat{G} = 72.21 - 0.13 I - 4.11 S + 0.0531 I \times S.
\]

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.

Suppose we measure IQ in deviations from 100 and STUDY in deviations from 8. What happens to the coefficients and t-tests in the interaction model? How about the main effects model?
GRADE AND STUDY TIME EXAMPLE FROM CLASS NOTES

Plot of STUDY*IQ. Symbol used is 'X'.
GRADE AND STUDY TIME EXAMPLE FROM NOTES

DEP VARIABLE: GRADE

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SUM OF SQUARES</th>
<th>MEAN SQUARE</th>
<th>F VALUE</th>
<th>PROB&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>2</td>
<td>596.115</td>
<td>298.058</td>
<td>32.568</td>
<td>0.0014</td>
</tr>
<tr>
<td>ERROR</td>
<td>5</td>
<td>45.759885</td>
<td>9.151977</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C TOTAL</td>
<td>7</td>
<td>641.875</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ROOT MSE 3.025223
R-SQUARE 0.9287
DEP MEAN 81.37500
ADJ R-SQ 0.9002
C.V. 3.717633

PARAMETER ESTIMATE STANDARD ERROR T FOR H0: PARAMETER = 0 PROB>|T|

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>DF</th>
<th>ESTIMATE</th>
<th>ERROR</th>
<th>PARAMETER = 0</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEP</td>
<td>1</td>
<td>0.736555</td>
<td>16.262800</td>
<td>0.045</td>
<td>0.9656</td>
</tr>
<tr>
<td>IQ</td>
<td>1</td>
<td>0.473084</td>
<td>0.129980</td>
<td>3.640</td>
<td>0.0149</td>
</tr>
<tr>
<td>STUDY</td>
<td>1</td>
<td>2.103436</td>
<td>0.264184</td>
<td>7.962</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

OBS ACTUAL PREDICT STD ERR LOWER 95% UPPER 95% MEAN MEAN RESIDUAL

<table>
<thead>
<tr>
<th>OBS</th>
<th>ACTUAL VALUE</th>
<th>PREDICT</th>
<th>STD ERR PREDICT</th>
<th>LOWER 95% MEAN</th>
<th>UPPER 95% MEAN</th>
<th>MEAN RESIDUAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75.000</td>
<td>71.445</td>
<td>1.933</td>
<td>66.477</td>
<td>76.412</td>
<td>3.555</td>
</tr>
<tr>
<td>2</td>
<td>79.000</td>
<td>78.017</td>
<td>1.270</td>
<td>74.752</td>
<td>81.282</td>
<td>0.983001</td>
</tr>
<tr>
<td>3</td>
<td>68.000</td>
<td>70.127</td>
<td>1.963</td>
<td>65.082</td>
<td>75.173</td>
<td>2.127</td>
</tr>
<tr>
<td>4</td>
<td>85.000</td>
<td>82.959</td>
<td>1.093</td>
<td>80.150</td>
<td>85.768</td>
<td>2.041</td>
</tr>
<tr>
<td>5</td>
<td>91.000</td>
<td>92.108</td>
<td>1.835</td>
<td>87.390</td>
<td>96.826</td>
<td>1.108</td>
</tr>
<tr>
<td>6</td>
<td>79.000</td>
<td>79.065</td>
<td>2.242</td>
<td>73.303</td>
<td>84.827</td>
<td>0.064928</td>
</tr>
<tr>
<td>7</td>
<td>98.000</td>
<td>96.737</td>
<td>2.224</td>
<td>91.019</td>
<td>102.455</td>
<td>1.263</td>
</tr>
<tr>
<td>8</td>
<td>76.000</td>
<td>80.543</td>
<td>1.929</td>
<td>75.585</td>
<td>85.500</td>
<td>4.543</td>
</tr>
<tr>
<td>9</td>
<td>83.643</td>
<td>1.141</td>
<td>80.709</td>
<td>86.577</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SUM OF RESIDUALS 7.10543E−15
SUM OF SQUARED RESIDUALS 45.75988

D. A. Dickey
ANALYSIS OF VARIANCE
AS A SPECIAL CASE OF REGRESSION

First, let us review ANOVA from last semester. Suppose we record the weights of 20 plants, where each of 4 fertilizers is used on one group of 5. We get these yields:

<table>
<thead>
<tr>
<th>FERTILIZER</th>
<th>DATA</th>
<th>MEAN</th>
<th>SUM</th>
<th>SSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>60</td>
<td>61</td>
<td>59</td>
<td>60</td>
</tr>
<tr>
<td>B</td>
<td>62</td>
<td>61</td>
<td>60</td>
<td>62</td>
</tr>
<tr>
<td>C</td>
<td>63</td>
<td>61</td>
<td>61</td>
<td>64</td>
</tr>
<tr>
<td>D</td>
<td>62</td>
<td>61</td>
<td>63</td>
<td>60</td>
</tr>
</tbody>
</table>

You should remember how to compute this table:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>FERTILIZER</td>
<td>3</td>
<td>25</td>
<td>8.333</td>
<td>3.92</td>
</tr>
<tr>
<td>ERROR</td>
<td>16</td>
<td>34</td>
<td>2.125</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>19</td>
<td>59</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

D. A. Dickey
You may also remember how to compare treatment means.

**CONTRAST I**

comparing A to B

Estimated difference $60 - 61 = -1$

Variance $(2.125)(1/5 + 1/5) = 0.85$

t-statistic $= -1 / \sqrt{0.85} = -1.0847$

**CONTRAST II**

comparing mean of A and B to C (or $A + B$ compared to $2C$)

Estimate of $[A \text{ mean} + B \text{ mean} - 2(C \text{ mean}) = 60 + 61 - 2(63) = -5$

Variance $(2.125)(1/5 + 1/5 + 4/5) = 2.55$

t-statistic $= -5 / \sqrt{2.55} = -3.1311$

Finally you may remember the model

$$Y_{ij} = \mu + \alpha_i + e_{ij}.$$  

The model says that each yield is an overall mean ($\mu$) plus a fertilizer effect ($\alpha_i$) plus a random deviation, $e_{ij}$. Now we can write this model as a matrix regression using these matrices:

D. A. Dickey
TABLE 1:

\[
\begin{pmatrix}
Y \\
60 \\
61 \\
59 \\
60 \\
60 \\
62 \\
61 \\
60 \\
60 \\
62 \\
60 \\
63 \\
61 \\
61 \\
64 \\
66 \\
62 \\
61 \\
63 \\
60 \\
64
\end{pmatrix}
= \begin{pmatrix}
X \\
1 1 1 0 0 0 \\
1 1 1 0 0 0 \\
1 1 1 0 0 0 \\
1 1 1 0 0 0 \\
1 1 0 1 0 0 \\
1 1 0 1 0 0 \\
1 1 1 0 0 0 \\
1 1 1 0 0 0 \\
1 1 0 1 0 0 \\
1 1 0 1 0 0 \\
1 1 0 1 0 0 \\
1 1 0 1 0 0 \\
1 0 0 1 0 \\
1 0 0 1 0 \\
1 0 0 1 0 \\
1 0 0 1 0 \\
1 0 0 0 1 \\
1 0 0 0 1 \\
1 0 0 0 1 \\
1 0 0 0 1 \\
1 0 0 0 1
\end{pmatrix}
\begin{pmatrix}
\beta \\
\mu \\
\alpha_1 \\
\alpha_2 \\
\alpha_3 \\
\alpha_4
\end{pmatrix}
+ \begin{pmatrix}
e \\
e_{11} \\
e_{12} \\
e_{13} \\
e_{14} \\
e_{15} \\
e_{21} \\
e_{22} \\
e_{23} \\
e_{24} \\
e_{25} \\
e_{31} \\
e_{32} \\
e_{33} \\
e_{34} \\
e_{35} \\
e_{41} \\
e_{42} \\
e_{43} \\
e_{44} \\
e_{45}
\end{pmatrix}
\]

It is obvious by multiplying the $X$ matrix by the $\beta$ vector that every observation in the first group is $(1)\mu + (1)\alpha_1 + (0)\alpha_2 + (0)\alpha_3 + (0)\alpha_4 = \mu + \alpha_1$ plus an error $e$, every element in group 2 is $\mu + \alpha_2$ plus an error, etc. This is exactly the ANOVA model. Now we are tempted to run a regression of $Y$ on the columns of $X$.
X to get our ANOVA but there is a problem. Look at the X matrix. Is it full rank? Obviously the answer is no. (Try inverting $X'X$ in SAS and see what happens.) I will now write down a new X matrix and $\beta$ vector.

TABLE 2:

<table>
<thead>
<tr>
<th>X</th>
<th>$\beta$ (Betas)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 0 0</td>
<td>$\mu + \alpha_4$ is now $\beta_0$</td>
</tr>
<tr>
<td>1 1 0 0</td>
<td>$\alpha_1 - \alpha_4$ is now $\beta_1$</td>
</tr>
<tr>
<td>1 1 0 0</td>
<td>$\alpha_2 - \alpha_4$ is now $\beta_2$</td>
</tr>
<tr>
<td>1 1 0 0</td>
<td>$\alpha_3 - \alpha_4$ is now $\beta_3$</td>
</tr>
<tr>
<td>1 1 0 0</td>
<td></td>
</tr>
<tr>
<td>1 0 1 0</td>
<td>Note: For rows 1 - 5 of X, multiplying row of X times the $\beta$ vector (recall multiplication of row times column)</td>
</tr>
<tr>
<td>1 0 1 0</td>
<td>$(1)(\mu+\alpha_4) + (1)(\alpha_1-\alpha_4) + 0 + 0 = \mu+\alpha_1.$</td>
</tr>
<tr>
<td>1 0 1 0</td>
<td></td>
</tr>
<tr>
<td>1 0 0 1</td>
<td></td>
</tr>
<tr>
<td>1 0 0 1</td>
<td></td>
</tr>
<tr>
<td>1 0 0 1</td>
<td></td>
</tr>
<tr>
<td>1 0 0 1</td>
<td></td>
</tr>
<tr>
<td>1 0 0 1</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0</td>
<td>Note: The last 5 rows multiplied by $\beta$ give $\mu + \alpha_4$ as they should.</td>
</tr>
<tr>
<td>1 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0</td>
<td></td>
</tr>
</tbody>
</table>

D. A. Dickey
For the above $X$ matrix and our fertilizer data, put the Y's and the last three columns of the last $X$ matrix in a SAS data set. Call the columns X1, X2, and X3. Regress Y on X1, X2, X3 and notice that you have produced the ANOVA table and F test for treatments (PROC REG supplies the column of 1s).

Recall CONTRAST I and CONTRAST II on the previous page. We can even make the computer estimate those contrasts and give us the t-statistics. Here are the $X$ matrix columns to do the job:
TABLE 3:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 0 & -2 & 1 \\ 1 & 0 & -2 & 1 \\ 1 & 0 & -2 & 1 \\ 1 & 0 & -2 & 1 \\ 1 & 0 & -2 & 1 \\ 1 & 0 & 0 & -3 \\ 1 & 0 & 0 & -3 \\ 1 & 0 & 0 & -3 \\ 1 & 0 & 0 & -3 \\ 1 & 0 & 0 & -3 \end{pmatrix}$$

$$X'X = \begin{bmatrix} 20 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 30 & 0 \\ 0 & 0 & 0 & 60 \end{bmatrix}$$

Consider the second element of $X'Y$. It will be

$$(\text{treatment 1 sum}) - (\text{treatment 2 sum}) = 300 - 305 = Q$$
Since $X'X$ is diagonal the second $b$ estimated will be $b_1 = (300 - 305)/10$ which is simply 1/2 of CONTRAST I. Similarly $b_2$ is some multiple of CONTRAST II and finally $b_3$ will be some multiple of the contrast comparing the mean of A, B, and C to D. The t-statistics will be exactly those computed above for the contrasts. Notice that we have used three orthogonal contrasts, that is, contrasts which give columns in $X$ so that $X'X$ is a diagonal matrix. We could easily compute the regression coefficients and their variances by hand since $X'X$ is so nice (or just run it on the computer).

You can easily compute the following:

$$X'Y = \begin{bmatrix} 1230 \\ -5 \\ -25 \\ -10 \end{bmatrix} \quad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 1230/20 \\ -5/10 \\ -25/30 \\ -10/60 \end{bmatrix} \quad \text{MSE} = \frac{34/16}{\sqrt{3}} = \frac{2.125}{\sqrt{3}}$$

$$\text{SS(regression)} = b'X'Y = (1230)^2/20 + (-5)^2/10 + (-25)^2/30 + (-10)^2/60 \quad = 75645 + 2.5 + 20.8333 + 1.6667 = 75645 + 25 \quad = \text{CT} + 25$$
From PROC REG or PROC GLM (or easily by hand) you will get:

**ANOVA**

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>TYPE I SS</th>
<th>TYPE III SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td>2.5000</td>
<td>2.5000</td>
</tr>
<tr>
<td>X2</td>
<td>1</td>
<td>20.8333</td>
<td>20.8333</td>
</tr>
<tr>
<td>X3</td>
<td>1</td>
<td>1.6667</td>
<td>1.6667</td>
</tr>
<tr>
<td>ERROR</td>
<td>16</td>
<td>34.0000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COEFFICIENT</th>
<th>T-STAT</th>
<th>STD. ERR.</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>-.5000</td>
<td>-1.0847</td>
</tr>
<tr>
<td>X2</td>
<td>-.8333</td>
<td>-3.1311</td>
</tr>
<tr>
<td>X3</td>
<td>-.1667</td>
<td>-0.8856</td>
</tr>
</tbody>
</table>

**SUMMARY:**

1. An ANOVA is just a special case of regression with dummy variables. You have several choices as to how to enter the dummy variables.

2. PROC GLM will automatically create \( k - 1 \) columns from a single column of \( k \) numbers if we use a class statement. For example if we had a column FERT in our dataset with each entry 1, 2, 3, or 4 we could call PROC GLM; CLASS FERT; MODEL YIELD = FERT; then the procedure will automatically create the columns of table 2 a couple of pages back. It will
lump all the 1 degree of freedom sums of squares into the usual 3 df ANOVA sum of squares (25).

3. Because we had orthogonal contrasts (X’X diagonal) the hand computations were easy and the TYPE I and TYPE III sums of squares were exactly the same. Also note that you could have produced exactly the same coefficients and regression sums of squares by running simple linear regressions on each individual column of X. This computational simplicity is due solely to orthogonality.

4. You could also throw in 4 columns of block indicator variables if the experiment had been blocked.

Example with 4 treatments in 3 blocks

<table>
<thead>
<tr>
<th>Y_{ij}</th>
<th>int</th>
<th>blk</th>
<th>blk</th>
<th>trt</th>
<th>trt</th>
<th>trt</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(Of course these could be replaced with columns of CONTRASTS as well).

D. A. Dickey
SAS code for FERTILIZER example:

DATA FERT;
INPUT FERT $ YIELD @@;
  X1=(FERT='A');
  X2=(FERT='B');
  X3=(FERT='C');
  X4=(FERT='D');
CARDS;
A 60  A 61  A 59  A 60  A 60
B 62  B 61  B 60  B 62  B 60
C 63  C 61  C 61  C 64  C 66
D 62  D 61  D 63  D 60  D 64

PROC PRINT; TITLE "FERTILIZER DATA SET";

PROC GLM;
  CLASS FERT; MODEL YIELD=FERT/SOLUTION;
  TITLE "CLASS STATEMENT IN GLM";

PROC REG; MODEL YIELD=X1 X2 X3;
  TITLE "CREATING YOUR OWN DUMMY VARIABLES";
RUN;
### FERTILIZER DATA SET

<table>
<thead>
<tr>
<th>OBS</th>
<th>FERT</th>
<th>YIELD</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>60</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>61</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>59</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>60</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>60</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>62</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>61</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>60</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td>62</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
<td>60</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>63</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>C</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>C</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>C</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>D</td>
<td>62</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>D</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>D</td>
<td>63</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>D</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>D</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

### CLASS STATEMENT IN GLM

**General Linear Models Procedure**

**Class Level Information**

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>FERT</td>
<td>4</td>
<td>A B C D</td>
</tr>
</tbody>
</table>

Number of observations in data set = 20

D. A. Dickey
Dependent Variable: YIELD

<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>25.000000</td>
<td>8.33333</td>
<td>3.92</td>
<td>0.0283</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>34.000000</td>
<td>2.12500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>59.000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square | C.V.  | Root MSE | YIELD Mean
---------|-------|----------|-------------
0.423729 | 2.370306 | 1.4577    | 61.500

Source | DF | Type I SS | Mean Square | F Value | Pr > F |
-------|----|-----------|-------------|---------|--------|
FERT   | 3  | 25.000000 | 8.333333    | 3.92    | 0.0283 |

Source | DF | Type III SS | Mean Square | F Value | Pr > F |
-------|----|-------------|-------------|---------|--------|
FERT   | 3  | 25.000000   | 8.333333    | 3.92    | 0.0283 |

Parameter | Estimate | Parameter=0 Estimate
----------|----------|----------------------
INTERCEPT | 62.0000 B | 95.10 0.0001 0.65192024
FERT A   | -2.0000 B | -2.17 0.0455 0.92195445
| B -1.0000 B | -1.08 0.2941 0.92195445
| C 1.0000 B | 1.08 0.2941 0.92195445
| D 0.0000 B | . . . .

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.
CREATING YOUR OWN DUMMY VARIABLES

Model: MODEL1
Dependent Variable: YIELD

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>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>25.00000</td>
<td>8.33333</td>
<td>3.922</td>
<td>0.0283</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>34.00000</td>
<td>2.12500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>19</td>
<td>59.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 1.45774, R-square 0.4237, Dep Mean 61.50000, Adj R-sq 0.3157, C.V. 2.37031

Parameter Estimates

| Parameter | Standard Error | T for H0: Parameter=0 | Prob > |T| |
|-----------|---------------|------------------------|--------|
| INTERCEP  | 0.65192024    | 95.104                | 0.0001 |
| X1        | 0.92195445    | -2.169                | 0.0455 |
| X2        | 0.92195445    | -1.085                | 0.2941 |
| X3        | 0.92195445    | 1.085                 | 0.2941 |
POLYNOMIAL REGRESSION

Polynomial means form

\[ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k \]

EXAMPLE: We try work crews of various sizes and record the number of parts produced in an hour for each work crew. We see that up to a point, increasing the number of workers increases per worker production but beyond that point, production begins to decrease. Our goal is to model per worker production as a function of crew size.

DATA (Y=PARTS PER WORKER, X=CREW SIZE)

<table>
<thead>
<tr>
<th>Y</th>
<th>3 5 7 12 8 10 10 5 6 4</th>
<th>mean = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>2 2 3 4 4 4 5 5 5 6</td>
<td>mean = 4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>SSq</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 3, 5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3 7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4 12, 8, 10</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>5 10, 5, 6</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>6 4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ \text{sum} = 24 \]

ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
<th>Pr&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>4</td>
<td>54</td>
<td>13.5</td>
<td>2.81</td>
<td>0.1436</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>24</td>
<td>4.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Note: The ANOVA essentially fits means to each crew size. Below we plot the production against crew size and indicate the fitted means from the ANOVA model. Note the ANOVA error sum of squares, 24, is just the pooled (summed) SSq from within each treatment group. The predictions (group means) are not constrained in any way.

Data plot:

PARTS

12  X
11  
10  -X-  X  Line segments are ANOVA treatment means.
  9  X
  8  
  7  -X-  ---
  6  X
  5  X  X
  4  ---  -X- 
  3  X
2  -----------+-----+-------+----------+-----------+ CREW SIZE
  2  3  4  5  6

Next we will fit a quadratic to the data (forcing predicted values to lie on a parabola) and observe how much the fit deteriorates.
DATA A;  INPUT Y X XSQ;
CARDS;
  3  2  4
  5  2  4
  7  3  9
 10  4 16
  8  4 16
 12  4 16
  5  5 25
 10  5 25
  6  5 25
  4  6 36
PROC REG;  MODEL Y = X XSQ;

The output contains the following information:

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
<th>Pr&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>2</td>
<td>49.42</td>
<td>24.71</td>
<td>6.05</td>
<td>0.0298</td>
</tr>
<tr>
<td>ERROR</td>
<td>7</td>
<td>28.58</td>
<td>4.08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| PARAMETER | ESTIMATE | T   | PR>||T| |
|-----------|----------|-----|------|
| INTERCEPT | -12.5280 | -2.19| .0644|
| X         | 11.0311  | 3.47| .0104|
| XSQ       | -1.3975  | -3.40| .0115|

The job can be done more easily in PROC GLM.
In fact, we will do several steps:

1) Do an ANOVA of production with crew size as treatment.
2) Fit a quadratic to the data.
3) Fit a degree 4 polynomial to the data. (For only 5 distinct $X$ values, a degree 4 polynomial is the highest degree you can fit.)
PROC GLM; CLASS X; MODEL Y = X;

Try this and you will get a printout like this:
(this is the computerized way of getting the ANOVA table).

GENERAL LINEAR MODELS PROCEDURE
CLASS LEVEL INFORMATION

<table>
<thead>
<tr>
<th>CLASS</th>
<th>LEVELS</th>
<th>VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREW</td>
<td>5</td>
<td>2 3 4 5 6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>4</td>
<td>54</td>
<td>13.5</td>
<td>2.81</td>
</tr>
<tr>
<td>ERROR</td>
<td>5</td>
<td>24</td>
<td>4.8</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>TYPE I SS</th>
<th>TYPE III SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>4</td>
<td>54</td>
<td>54</td>
</tr>
</tbody>
</table>

Notice that SAS has created an \( X \) matrix like table 2 a few pages back, because we used a CLASS statement. Also, it will not give individual coefficients etc. because of the lack of a unique parameterization (they can be recovered through the \( /SOLUTION \) option.)

Notice also that the ANOVA, which basically fits group means to the 5 treatment groups, has increased the regression sum of squares over that of the quadratic, from 49.4161 (2df) to 54 (4df). Later, we will show that the ANOVA (means) model is like a “full model” and the quadratic like a “reduced model.”

The test which asks if the regression does as well as the ANOVA in fitting the data is called a “lack of fit F test.” We compute
\[ F = \frac{[(54 - 49.4161)/2]}{(4.8)} = 0.4775 \]

Since \( F \) is insignificant (2 and 5 df) we say there is no significant lack of fit. We conclude that a model forcing production to be a quadratic in crew size seems to explain production as well as a model in which unconstrained means are fit to the data.

To show that the F test is a full versus reduced model F test, I will show that the ANOVA approach is the same as fitting the highest degree polynomial possible to the data. Since there are \( m = 5 \) values of \( X \), the highest possible degree is \( m - 1 = 4 \). Thus we issue the commands:

```plaintext
PROC GLM; MODEL Y = X X*X X*X*X X*X*X*X;
```

Notice that no CLASS statement was used and that PROC GLM will actually compute the powers of \( X \) for you. We get:

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>4</td>
<td>54</td>
<td>13.5</td>
<td>2.81</td>
</tr>
<tr>
<td>ERROR</td>
<td>5</td>
<td>24</td>
<td>4.8</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>TYPE I SS</th>
<th>TYPE III SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>2.25</td>
<td>2.93</td>
</tr>
<tr>
<td>X*X</td>
<td>47.17</td>
<td>3.57</td>
</tr>
<tr>
<td>X<em>X</em>X</td>
<td>0.45</td>
<td>3.98</td>
</tr>
<tr>
<td>X<em>X</em>X*X</td>
<td>4.13</td>
<td>4.13</td>
</tr>
</tbody>
</table>

D. A. Dickey
Using the TYPE I SSq we compute the lack of fit F as:

\[ F = \left( \frac{(0.45 + 4.13)/2}{4.8} \right) = 0.4771 \]

the same as before and we thus see that the lack of fit statistic is testing for the powers of X up to the highest possible power you can fit to the data. The only reason that there is an error term left to test this against is the fact that some X's had repeated Y's with them and so the highest possible degree, \( m - 1 = 4 \), is less than the total degrees of freedom \( n - 1 = 9 \) leaving 5 degrees of freedom (with sum of squares 24) for “pure error.”

As before, we see that it is incorrect and dangerous to make a conclusion about the joint significance of all the coefficients taken together if we look only at the t statistics.

Try fitting the data yourself. Try fitting a fifth degree polynomial. What do you get on your printout to indicate that you can not fit a degree larger than 4? Also: Note that just fitting means to the data gave no evidence of crew size effect
but imposing the reasonable assumption of a smooth response (quadratic) gave us more power and we found an effect.

In a *response surface* model, a response is some function (usually quadratic) of one or more control variables. Try this example (of yields in a chemical reaction) on the computer:

```
DATA REACT;  INPUT YIELD PH TEMP@@;
   PSQ = PH**2;  TSQ = TEMP**2;  PT = PH*TEMP;
CARDS;
   90 5 60 100 5 80 95 5 100 105 5.5 80
   100 6 60 130 6 80 125 6 100 140 6.5 80
   135 7 60 142 7 80 126 7 100
;
PROC PRINT;
PROC REG;  MODEL YIELD = PH TEMP PSQ TSQ PT/P;
PROC RSREG;  MODEL YIELD = PH TEMP;
```

Note the use of @@ to keep SAS from going to a new line for each observation read. If we omitted @@ we would get only 2 observations in our data set. Use the “missing Y trick” to get SAS to compute a 95% prediction interval for the yield of a future reaction at PH 6.3 and temperature 92 degrees.
### CHEMICAL PROCESS YIELDS

<table>
<thead>
<tr>
<th>OBS</th>
<th>YIELD</th>
<th>PH</th>
<th>TEMP</th>
<th>PSQ</th>
<th>TSQ</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90</td>
<td>5.0</td>
<td>60</td>
<td>25.00</td>
<td>3600</td>
<td>300</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>5.0</td>
<td>80</td>
<td>25.00</td>
<td>6400</td>
<td>400</td>
</tr>
<tr>
<td>3</td>
<td>95</td>
<td>5.0</td>
<td>100</td>
<td>25.00</td>
<td>10000</td>
<td>500</td>
</tr>
<tr>
<td>4</td>
<td>105</td>
<td>5.5</td>
<td>80</td>
<td>30.25</td>
<td>6400</td>
<td>440</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>6.0</td>
<td>60</td>
<td>36.00</td>
<td>3600</td>
<td>360</td>
</tr>
<tr>
<td>6</td>
<td>130</td>
<td>6.0</td>
<td>80</td>
<td>36.00</td>
<td>6400</td>
<td>480</td>
</tr>
<tr>
<td>7</td>
<td>125</td>
<td>6.0</td>
<td>100</td>
<td>36.00</td>
<td>10000</td>
<td>600</td>
</tr>
<tr>
<td>8</td>
<td>140</td>
<td>6.5</td>
<td>80</td>
<td>42.25</td>
<td>6400</td>
<td>520</td>
</tr>
<tr>
<td>9</td>
<td>135</td>
<td>7.0</td>
<td>60</td>
<td>49.00</td>
<td>3600</td>
<td>420</td>
</tr>
<tr>
<td>10</td>
<td>142</td>
<td>7.0</td>
<td>80</td>
<td>49.00</td>
<td>6400</td>
<td>560</td>
</tr>
<tr>
<td>11</td>
<td>126</td>
<td>7.0</td>
<td>100</td>
<td>49.00</td>
<td>10000</td>
<td>700</td>
</tr>
</tbody>
</table>

### CHEMICAL PROCESS YIELDS

Model: MODEL1  
Dependent Variable: YIELD

#### 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>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>3331.65539</td>
<td>666.33108</td>
<td>8.429</td>
<td>0.0177</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>395.25370</td>
<td>79.05074</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>10</td>
<td>3726.90909</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE  8.89105  
R-square  0.8939  
Dep Mean  117.09091  
Adj R-sq  0.7879  
C.V. 7.59329

D. A. Dickey
Parameter Estimates

| Parameter | DF | Estimate | Standard Error | T for HO: Parameter=0 | Prob>|T| |
|-----------|----|----------|----------------|-----------------------|-------|
| INTERCEPT | 1  | -382.624093 | 239.89941787 | -1.595 | 0.1716 |
| PH        | 1  | 70.619739   | 74.04775138  | 0.954 | 0.3840 |
| TEMP      | 1  | 5.652925    | 2.57066503   | 2.199 | 0.0792 |
| PSQ       | 1  | -2.981132   | 5.98302278   | -0.498 | 0.6394 |
| TSQ       | 1  | -0.027675   | 0.01368841   | -2.022 | 0.0991 |
| PT        | 1  | -0.175000   | 0.22227621   | -0.787 | 0.4668 |

Dep Var | Predict
--- | ---
Obs | YIELD | Value | Residual
--- | --- | --- | ---
1  | 90.0  | 83.0  | 7.0065
2  | 100.0 | 101.1 | -1.0633
3  | 95.0  | 97.0  | -1.9935
4  | 105.0 | 113.7 | -8.7222
5  | 100.0 | 110.3 | -10.3208
6  | 130.0 | 124.9 | 5.1094
7  | 125.0 | 117.3 | 7.6792
8  | 140.0 | 134.6 | 5.4316
9  | 135.0 | 131.7 | 3.3142
10 | 142.0 | 142.8 | -0.7556
11 | 126.0 | 131.7 | -5.6858

Sum of Residuals 0
Sum of Squared Residuals 395.2537
Predicted Resid SS (Press) 3155.9162
Coding Coefficients for the Independent Variables

<table>
<thead>
<tr>
<th>Factor</th>
<th>Subtracted off</th>
<th>Divided by</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH</td>
<td>6.000000</td>
<td>1.000000</td>
</tr>
<tr>
<td>TEMP</td>
<td>80.000000</td>
<td>20.000000</td>
</tr>
</tbody>
</table>

Response Surface for Variable YIELD

Response Mean: 117.090909
Root MSE: 8.891048
R-Square: 0.8939
Coef. of Variation: 7.5933

<table>
<thead>
<tr>
<th>Degrees of Type I Regression Freedom</th>
<th>SSq</th>
<th>R-Square</th>
<th>F-Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>2</td>
<td>2898.15</td>
<td>0.7776</td>
<td>18.331</td>
</tr>
<tr>
<td>Quadratic</td>
<td>2</td>
<td>384.50</td>
<td>0.1032</td>
<td>2.432</td>
</tr>
<tr>
<td>Crossproduct</td>
<td>1</td>
<td>49.00</td>
<td>0.0131</td>
<td>0.620</td>
</tr>
<tr>
<td>Total Regress</td>
<td>5</td>
<td>3331.66</td>
<td>0.8939</td>
<td>8.429</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Degrees of Sum of Residual Freedom</th>
<th>Total Error</th>
<th>395.253701</th>
<th>79.050740</th>
</tr>
</thead>
</table>
### Degrees of Freedom Parameter Standard Error Parameter=0
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Freedom</th>
<th>Estimate</th>
<th>Error</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>1</td>
<td>-382.6241</td>
<td>239.8994</td>
<td>-1.595</td>
</tr>
<tr>
<td>PH</td>
<td>1</td>
<td>70.6197</td>
<td>74.0478</td>
<td>0.954</td>
</tr>
<tr>
<td>TEMP</td>
<td>1</td>
<td>5.6529</td>
<td>2.5707</td>
<td>2.199</td>
</tr>
<tr>
<td>PH*PH</td>
<td>1</td>
<td>-2.9811</td>
<td>5.9830</td>
<td>-0.498</td>
</tr>
<tr>
<td>TEMP*PH</td>
<td>1</td>
<td>-0.1750</td>
<td>0.2223</td>
<td>-0.787</td>
</tr>
<tr>
<td>TEMP*TEMP</td>
<td>1</td>
<td>-0.0277</td>
<td>0.0137</td>
<td>-2.022</td>
</tr>
</tbody>
</table>

### Parameter Estimate from Coded Data
| Parameter     | Prob > |T| | Data         |
|---------------|--------|----------------|-------------|
| INTERCEPT     | 0.1716 | 124.890566 |
| PH            | 0.3840 | 20.846154 |
| TEMP          | 0.0792 | 3.500000 |
| PH*PH         | 0.6394 | -2.981132 |
| TEMP*PH       | 0.4668 | -3.500000 |
| TEMP*TEMP     | 0.0991 | -11.069811 |

### Degrees of Freedom Sum of Squares Mean Square F-Ratio Prob > F
<table>
<thead>
<tr>
<th>Factor</th>
<th>Freedom</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F-Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH</td>
<td>3</td>
<td>2893.2796</td>
<td>964.426544</td>
<td>12.200</td>
<td>0.0098</td>
</tr>
<tr>
<td>TEMP</td>
<td>3</td>
<td>445.6173</td>
<td>148.539109</td>
<td>1.879</td>
<td>0.2508</td>
</tr>
</tbody>
</table>

D. A. Dickey
Canonical Analysis of Response Surface
(based on coded data)

<table>
<thead>
<tr>
<th>Factor</th>
<th>Coded</th>
<th>Uncoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH</td>
<td>3.751711</td>
<td>9.751711</td>
</tr>
<tr>
<td>TEMP</td>
<td>-0.435011</td>
<td>71.299771</td>
</tr>
</tbody>
</table>

Predicted value at stationary point 163.233672

<table>
<thead>
<tr>
<th>Eigenvectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalues</td>
</tr>
<tr>
<td>-2.618751</td>
</tr>
<tr>
<td>-11.432192</td>
</tr>
</tbody>
</table>

Stationary point is a maximum.

Example:
\[ \hat{Y} = -382 + 70.62 P + 5.65 T - 2.98 P^2 - 0.028 T^2 - 0.175 PT \]

- Critical point: \( P = 9.7517, T = 71.2998 \)

- \( \hat{Y} = 163.23 + \)
  \[ (P-9.7517 \ T-71.2998) \begin{pmatrix} -2.9800 & -0.0875 \\ -0.0875 & -0.0280 \end{pmatrix} \begin{pmatrix} P - 9.7517 \\ T - 71.2998 \end{pmatrix} \]
  \[ = 163.23 + \mathbf{X}'\mathbf{A}\mathbf{X} \]
\[ \left( \begin{array}{cc}
-2.98 & -0.0875 \\
-0.0875 & -0.028
\end{array} \right) = A = Z L Z' = \\
\left( \begin{array}{cc}
-0.0296 & .9996 \\
.9996 & 0.0296
\end{array} \right) \left( \begin{array}{cc}
-0.0251 & 0 \\
0 & -2.9837
\end{array} \right) \left( \begin{array}{cc}
-0.0296 & .9996 \\
.9996 & 0.0296
\end{array} \right) \]

\[ \hat{Y} = 163.23 + X'Z L Z'X = 163.23 + W' L W = \\
163.23 + (-.0251) w_1^2 + (-2.984) w_2^2 \]

\[ (w_1, w_2) = (0,0) = \text{critical point, response is 163.23. Any movement away from critical point reduces response.} \]

Additional Points: Critical point may be max, min, or saddle point. It may be nowhere near experimental region. Ridge Analysis (not ridge regression) takes spheres of ever increasing radius around some point in the experimental region. On each sphere, coordinates of response maximizer (minimizer) are computed resulting in a path of maximum increase (decrease). PROC RSREG has a RIDGE statement to do this. Again eigenvalues are involved.
FACTORIAL EXPERIMENTS

TERMINOLOGY

We will have a yield \( Y \) like, for example, the amount of a certain chemical in refrigerated meat where the amount of the chemical indicates the degree of spoilage which has occurred.

We will speak of factors \( A, B, C \) like, for example, \( A \) = storage temperature, \( B \) = type of wrap, \( C \) = spoilage retardant.

We will speak of the levels of the factor such as

<table>
<thead>
<tr>
<th>FACTOR</th>
<th>LEVELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>35, 28, 21, 14 degrees</td>
</tr>
<tr>
<td>B</td>
<td>foil, paper, plastic</td>
</tr>
<tr>
<td>C</td>
<td>absent, present</td>
</tr>
</tbody>
</table>

The factors may be quantitative (\( A \)) at equally or unequally spaced levels. They may be qualitative (\( B, C \)). (Factor \( C \) could be made quantitative.)

We will have all possible treatment combinations represented and, in fact, replicated. Thus we would store \( 4 \times 3 \times 2 = 24 \) packages of meat in each replicate of this experiment. If we have 120 packages of meat available, we could have 5
replications which could be done in a completely randomized, randomized complete block, or even an incomplete block design.

Technically speaking we refer to a factorial arrangement of treatments, not a factorial “design.” The design would be completely random, randomized complete block, or whatever.

We have described an experiment which would be called a 4*3*2 factorial experiment with 5 replications.

Suppose the replications were in blocks. Then we might produce an initial ANOVA as follows:

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOCKS</td>
<td>4</td>
<td>300</td>
<td>75</td>
<td>7.5</td>
</tr>
<tr>
<td>TREATMENTS</td>
<td>23</td>
<td>943</td>
<td>41</td>
<td>4.1</td>
</tr>
<tr>
<td>ERROR</td>
<td>92</td>
<td>920</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Since there are treatment effects, we want to break down the 23 treatment degrees of freedom into some meaningful pieces (main effects and interactions as they will be called).

We begin by looking at a simple 2*2 factorial with 5 replications in a CRD (completely randomized design).
EXAMPLE 1: 2*2 FACTORIAL

Call the factors N and P. Let the levels of N be denoted \( n_1 \) and \( n_2 \). Similarly define \( p_1 \) and \( p_2 \).

<table>
<thead>
<tr>
<th>TREATMENT COMBINATION</th>
<th>DATA</th>
<th>MEAN</th>
<th>SSq within</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_1 ) and ( p_1 )</td>
<td>35, 26, 25, 33, 31</td>
<td>30</td>
<td>76</td>
</tr>
<tr>
<td>( n_1 ) and ( p_2 )</td>
<td>39, 33, 41, 31, 36</td>
<td>36</td>
<td>68</td>
</tr>
<tr>
<td>( n_2 ) and ( p_1 )</td>
<td>37, 27, 35, 27, 34</td>
<td>32</td>
<td>88</td>
</tr>
<tr>
<td>( n_2 ) and ( p_2 )</td>
<td>49, 39, 39, 47, 46</td>
<td>44</td>
<td>88</td>
</tr>
</tbody>
</table>

\[ \sum SSq = 320 \]

Let the following table denote mean yields, \( \bar{Y}_{ijk} \), of the 5 observations \( Y_{ijk} \) in each treatment combination:

<table>
<thead>
<tr>
<th>TABLE OF TREATMENT MEANS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_1 )</td>
</tr>
<tr>
<td>( p_1 )</td>
</tr>
<tr>
<td>( p_2 )</td>
</tr>
<tr>
<td>mean</td>
</tr>
</tbody>
</table>

\[ 5 = (8+2)/2 \]
\[ = 38-33 \]

D. A. Dickey
MAIN EFFECT OF P, MAIN EFFECT OF N

What we see is that the difference between low P and high P is a difference in mean yield of 40 - 31 = 9 in our sample. Is there a nonzero difference in the population? Similarly, main effect of N is 5.

INTERACTION OF P AND N

We see that this difference, 9, is not consistent over the levels of N. When N is low (n₁) a change in P produces an increase 6 in yield. When N is high, the increase in P (from p₁ to p₂) produces an increase in yield of 12. The effect of P depends on the level of N in our sample. Another way to say this is that N and P interact in the sample. Is this interaction large enough to declare a nonzero interaction in the population?

We approach the question of significance of “main effects" and “interactions" just as before, that is, we measure the inherent variability among items treated alike (MSE), use this to assign a variance to our computed numbers (9, 6, and 12) and then do a t-test (or an F). In the next couple of sections we show how it is done.
PRELIMINARY ANOVA

Table of treatment totals $Y_{ij}$.

<table>
<thead>
<tr>
<th>$n_1$</th>
<th>$n_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>160</td>
</tr>
<tr>
<td>180</td>
<td>220</td>
</tr>
<tr>
<td>330</td>
<td>380</td>
</tr>
</tbody>
</table>

$CT = (710^2)/20 = 25205$

$SS(trts) = (150*150)/5 + (160*160)/5 + (180*180)/5 + (220*220)/5 - CT = 25780 - 25205 = 575$ (remember?)

We have $SS(total) = 26100 - 25205 = 895$

by subtraction (or by summing $SSq(within)$ from data table)

$SSE = 320$

ANOVA

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>$Mn,Sq$</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREATMENTS</td>
<td>3</td>
<td>575</td>
<td>191.67</td>
<td>9.58</td>
</tr>
<tr>
<td>ERROR</td>
<td>16</td>
<td>320</td>
<td>20.00</td>
<td></td>
</tr>
</tbody>
</table>

Now to analyze the 3 df for treatments in detail. First, we make a little plot of yield against the levels of N as follows:
Draw a line connecting the $p_2$'s and another connecting the $p_1$'s. As a result of the interaction, these lines are not parallel. Above the $n_1$, the lines are 6 units apart while above the $n_2$ they are 12 units apart.

Obviously, if the distance between the plotted $p_1$ and $p_2$ were the same (say 9) above $n_1$ and above $n_2$ then your "$p_1$ line" and "$p_2$ line" would be parallel. The effect of changing from $p_1$ to $p_2$ would be the same (yield increase of 9) regardless of the level on N.

With the above two paragraphs in mind, let us measure the failure of the lines to be parallel and then ask if this could have happened by chance.

1. Distance from $p_1$ to $p_2$ at $n_1$ is $36 - 30 = 6$
2. Distance from $p_1$ to $p_2$ at $n_2$ is $44 - 32 = 12$

3. If lines were parallel, the difference of the above numbers would be 0 but instead we got $12 - 6 = 6$

4. Tracing our steps we took $(44 - 32) - (36 - 30) = 6$.
   Now $44 - 32 - 36 + 30$ is a linear combination of means. Each mean has estimated variance $= \text{MSE}/5$. 

5. Is the number 6 (from part 3 above) significantly different from 0? Estimated variance $= \text{MSE}/5 + \text{MSE}/5 + \text{MSE}/5 + \text{MSE}/5 = 16$
   
   $t = 6/\sqrt{16} = 6/4 = 1.5$ $\Rightarrow$ insignificant interaction.

6. If you prefer, compute $F = \frac{t^2}{n_1p_1 n_2p_1 n_1p_2 n_2p_2}$

   TOTAL $150$ $160$ $180$ $220$ $Q = 30$

   CONTRAST $1$ $-1$ $-1$ $1$ $\text{denom} = 5(4)$

   $Q =$ sum of cross products $= 150(1) + 160(-1) + 180(-1) + 220(1) = 30$
   
   $\text{denom} = (\text{reps})*(\text{sum of squared coefficients}) = 5(1^2 + (-1)^2 + (-1)^2 + 1^2)$
   
   $\text{sum of squares} = Q*Q/\text{denom} = 900/20 = 45$
   
   $F = 45/\text{MSE} = 45/20 = 2.25$

The step 6 above is a sort of “old fashioned" way of testing for interaction. You will not be surprised to find that $Q$ is really a regression $X'Y$ and that denom is really a regression $X'X$. In the
regression approach we will compute a coefficient equal to $(X'Y)/(X'X)$. We write this as a fraction because the matrices are scalars. Next the regression sum of squares

$$b'X'Y = (X'Y)/(X'X) = Q*Q/denom.$$  

*** We would also like to talk about the “main effect” of N. This means the change in average yield as we go from the low to high level of N. For main effects, the averaging is over the reps and all the other factors. The main effect of N in our example is

$$(44 + 32)/2 - (36 + 30)/2 = 0.5(44 + 32 - 36 - 30) = 5$$

Its estimated variance is

$$0.25(MSE/5 + MSE/5 + MSE/5 + MSE/5) = 0.25(16) = 4$$

We have $t = 5/2 = 2.5$ or $F = 6.25$.

Can we get $F$ directly as before? Yes. In fact let us get the sums of squares and $F$ statistics for the main effects of N and P and for the NP interaction using the treatment group totals:

<table>
<thead>
<tr>
<th></th>
<th>$n_1p_1$</th>
<th>$n_1p_2$</th>
<th>$n_2p_1$</th>
<th>$n_2p_2$</th>
<th>$Q$</th>
<th>den.</th>
<th>SSq= $Q*Q/den.$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTALS</td>
<td>150</td>
<td>180</td>
<td>160</td>
<td>220</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N MAIN</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>50</td>
<td>20</td>
<td>125</td>
</tr>
<tr>
<td>P MAIN</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>90</td>
<td>20</td>
<td>405</td>
</tr>
<tr>
<td>NP</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>30</td>
<td>20</td>
<td>45</td>
</tr>
</tbody>
</table>

We can now produce a more detailed ANOVA table as follows:
ANOVA

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>1</td>
<td>125</td>
<td>125</td>
<td>6.25</td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>405</td>
<td>405</td>
<td>20.25</td>
</tr>
<tr>
<td>NP</td>
<td>1</td>
<td>45</td>
<td>45</td>
<td>2.25</td>
</tr>
<tr>
<td>ERROR</td>
<td>16</td>
<td>320</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

SUMMARY AND GENERAL NOTATION (Example 1)

We will denote means as follows:

\[ n_1p_1 = 30, \text{ etc.} \]

We will define main effects and interactions in terms of means. The definitions are natural in terms of the main effects as we argued in paragraph *** on the preceding page. We compute the main effect of P as

\[ 0.5(-n_1p_1 + n_1p_2 - n_2p_1 + n_2p_2) = 9 \]

\[ = 0.5(-30 + 36 - 32 + 44) \]

The high levels of P get a + and the low a -. Similarly the main effect of N applies a 1 to high N and -1 to low N to get:

\[ 0.5(-n_1p_1- n_1p_2 + n_2p_1 + n_2p_2) = 5 \text{ (see *** above)} \]

To conform, the interaction is also defined with a 0.5 multiplier and applies −1 when N and P are at different levels, + 1 when N and P are at the same levels.
We can see that the main effects are just the differences between the mean of all observations at the high level of that factor and the mean of all observations at the low level. The computations can be summarized in a table almost exactly the same as the one with totals.

\[
0.5(n_1p_1 - n_1p_2 - n_2p_1 + n_2p_2) = 3
\]

Often the totals are denoted by enclosing the symbol for the corresponding mean in square brackets. Thus for example we would write \([n_1p_1] = 150\) since the square brackets denote the total in that first treatment group. Obviously we could write general formulas for the sums of squares in terms of these square bracket symbols. Exercise: write the formula for \(SS(N)\) using \([ ]\) notation.

Now for a little reward for having studied matrices and regression. This entire analysis can be done by a regression program on the computer. We set up a dataset using the appropriate contrast columns as follows:

```plaintext
DATA FACTOR; INPUT YIELD
```
N P NP REP;
CARDS;

35 -1 -1 1 1
26 -1 -1 1 2
25 -1 -1 1 3
33 -1 -1 1 4
31 -1 -1 1 5
39 -1 1 -1 1
33 -1 1 -1 2
41 -1 1 -1 3
31 -1 1 -1 4
36 -1 1 -1 5
37 1 -1 -1 1
27 1 -1 -1 2
35 1 -1 -1 3
27 1 -1 -1 4
34 1 -1 -1 5
49 1 1 1 1
39 1 1 1 2
39 1 1 1 3
47 1 1 1 4
46 1 1 1 5

\[ X'X = \begin{array}{cccc}
20 & 0 & 0 & 0 \\
0 & 20 & 0 & 0 \\
0 & 0 & 20 & 0 \\
0 & 0 & 0 & 20
\end{array} \]

(Remember: column of 1's inserted.)

\[ X'Y = \begin{array}{c}
150 + 180 + 160 + 220 = 710 \\
-150 - 180 + 160 + 220 = 50 \\
-150 + 180 - 160 + 220 = 90 \\
150 - 180 - 160 + 220 = 30
\end{array} \]

SSR(uncorrected) = \frac{710 \times 710}{20} + \frac{50 \times 50}{20} + \frac{90 \times 90}{20} + \frac{30 \times 30}{20} = CT + \frac{125}{N} + \frac{405}{P} + \frac{45}{NP}

Although the hand computations above are quite simple, our purpose is to use the computer so we submit this code:
PROC GLM DATA = FACTOR; MODEL YIELD = N P NP;

The resulting ANOVA is read off the computer output. Try submitting this job.

ANOVA

SOURCE | DF | TYPE I SS | TYPE III SS | F  
--- | --- | --- | --- | ---  
N | 1 | 125 | 125 | 6.25  
P | 1 | 405 | 405 | 20.25  
NP | 1 | 45 | 45 | 2.25  
ERROR | 16 | 320 | (MSE = 20)  

Benefits:

All computations done on computer (ease of computation).
F-tests and p-values automatically produced.
Degrees of freedom make sense (correspond to columns).
Easily extends to more levels, blocked designs.

Example: If the reps were blocks (say, greenhouse benches) we could run this code:

PROC GLM; CLASS REP; MODEL YIELD = REP N P NP;

Try this with the above dataset. Do the analysis by hand and check the numbers against the computer. Note that the initial hand analysis is a RCB just as last semester – we simply break down the treatment sum of squares further. Are the treatment
sums of squares affected by computing a block sum of squares? What is affected?

\[
( \text{SS(block)} = 124.5, \text{SS(trt)} = 575, \text{SS(error)} = 195.5 )
\]

Example 2: 2*2 factorial in blocked design.

RESPONSE: WEIGHT GAIN IN MICE

FACTORS:

F FEED SUPPLEMENT
- \( f_1 \): none, \( f_2 \): supplement given daily

L LIGHT
- \( t_1 \): 8 hrs light, \( t_2 \): 14 hrs light

DESIGN:

Four littermates from each of 3 litters are chosen. The 4 mice are assigned at random to the 4 treatment combinations. After 6 weeks the weight gains are recorded with these results:

<table>
<thead>
<tr>
<th>Litter 1</th>
<th>Litter 2</th>
<th>Litter 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_1 t_1 )</td>
<td>10</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>( f_1 t_2 )</td>
<td>12</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>( f_2 t_1 )</td>
<td>13</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>( f_2 t_2 )</td>
<td>18</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>53</td>
<td>69</td>
<td>64</td>
</tr>
</tbody>
</table>

COMPUTATIONS:

D. A. Dickey
\[
CT = 2883 \\
SS(\text{blocks}) = 53 \times 53/4 + 69 \times 69/4 + 64 \times 64/4 - CT = 33.5 \\
SS(\text{trt}) = 39 \times 39/3 + 42 \times 42/3 + 45 \times 45/3 + 60 \times 60/3 - CT = 87 \\
SS(\text{total}) = 10 \times 10 + \ldots + 20 \times 20 - CT = 125 \\
SS(\text{error}) = SS(\text{total}) - SS(\text{blocks}) - SS(\text{trt}) = 125 - 33.5 - 87 = 4.5
\]

<table>
<thead>
<tr>
<th>ANOVA</th>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOCKS</td>
<td>2</td>
<td>33.5</td>
<td>16.75</td>
<td>22.33</td>
<td></td>
</tr>
<tr>
<td>FEED</td>
<td>1</td>
<td>48.0</td>
<td>48.00</td>
<td>64.00</td>
<td></td>
</tr>
<tr>
<td>LIGHT</td>
<td>1</td>
<td>27.0</td>
<td>27.00</td>
<td>36.00</td>
<td></td>
</tr>
<tr>
<td>FD*LT</td>
<td>1</td>
<td>12.0</td>
<td>12.00</td>
<td>16.00</td>
<td></td>
</tr>
<tr>
<td>ERROR</td>
<td>6</td>
<td>4.5</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Can you verify the breakdown of treatment sums of squares?)

We see a significant interaction between light and the food supplement. What is the meaning of a main effect in the presence of an interaction? Increasing the hours of light seems to increase growth in these animals (from 14 to 17 in our sample) but this +3 main effect of light is averaged over the levels of the other factors. The sample increase without the supplement is from 13 to 14 (+1 is simple effect of light without the food supplement). In the presence of food supplement, the increased light exposure causes a sample average increase from 15 to 20! The average increase (main effect) of 3 may not have any meaning in the sense that there is no way to achieve it...
by manipulating the animals' food supply. It may be that any amount of supplement triggers the high response to light while without supplement, the animals do not respond to light. Then what does the average value 3 mean? Nothing. We are in fact interested in the two increases 1 and 5 within the two food regimes. Perhaps one is significant and the other is not.

What is our next step? We would like to reconsider our ANOVA table in light of the significant interaction. Let us contemplate the following ANOVA tables:

<table>
<thead>
<tr>
<th>ANOVA 1</th>
<th>ANOVA 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOURCE</td>
<td>DF</td>
</tr>
<tr>
<td>BLOCKS</td>
<td>2</td>
</tr>
<tr>
<td>FEED</td>
<td>1</td>
</tr>
<tr>
<td>L in f₁</td>
<td>1</td>
</tr>
<tr>
<td>L in f₂</td>
<td>1</td>
</tr>
<tr>
<td>ERROR</td>
<td>6</td>
</tr>
</tbody>
</table>

Here is the old-fashioned way to compute the sums of squares using treatment totals and contrasts:
ANCOVA 1 is nice. It very clearly shows what I think is going on. First, there is an overall effect of the food supplement producing a significant increase in weight gain. In the presence of this supplement, there seems to be a sensitivity of the animal to photoperiod, namely longer daylight hours seem conducive to weight gains. This may simply reflect a longer foraging time and hence higher intake of the supplement. The animal does not seem to respond to photoperiod in the absence of this food supplement. Each ANCOVA breaks down the treatment sum of squares in an additive way.

At this point it will come as no surprise to see this whole analysis run as a regression using PROC GLM or PROC REG.
First, let us create a dataset with all the variables we will need for our various analyses. I will put in contrasts in the blocks for no reason other than to make $X'X$ diagonal and thus allow me to do the computations by hand easily as well.

```plaintext
DATA GROWTH;
    INPUT GAIN LITTER FEED $ LIGHT 
        BLIVS3 BL2VS13 F L FL LINF1 LINF2;
CARDS;
10 1 NO 8 -1 -1 -1 1 -1 0
16 2 NO 8 0 2 -1 -1 1 -1 0
13 3 NO 8 1 -1 -1 1 -1 0
12 1 NO 14 -1 -1 1 -1 1 0
15 2 NO 14 0 2 -1 1 -1 1 0
15 3 NO 14 1 -1 -1 1 -1 1 0
13 1 YES 8 -1 -1 1 -1 -1 0 -1
16 2 YES 8 0 2 1 -1 -1 0 -1
16 3 YES 8 1 -1 1 -1 -1 0 -1
18 1 YES 14 -1 -1 1 1 1 0 1
22 2 YES 14 0 2 1 1 1 0 1
20 3 YES 14 1 -1 1 1 1 0 1
;
PROC GLM; CLASS LITTER FEED LIGHT;
    MODEL GAIN = LITTER FEED LIGHT FEED*LIGHT;
PROC GLM; MODEL GAIN = BLIVS3 BL2VS13 F L FL;
PROC GLM; CLASS LITTER FEED LIGHT;
    MODEL GAIN = LITTER FEED LIGHT(FEED) /SOLUTION;
PROC GLM;
    MODEL GAIN = BLIVS3 BL2VS13 F LINF1 LINF2;
```

The first and third regressions are the ones we would normally do in research. The second and fourth are just there
for pedagogical reasons. They illustrate that the block sum of squares can be removed by orthogonal contrasts and enable us to easily see what the computer is calculating (because the whole $X$ matrix is orthogonal). You should try running all of these regressions and study the interrelationships among the printouts.

Let us trace through the computations for the last regression above. It is the one we would like to present since we have interaction and thus want to look at effects of $L$ within the various levels of $F$.

$$X = \begin{bmatrix}
1 & -1 & -1 & -1 & 0 \\
1 & 0 & 2 & -1 & -1 & 0 \\
1 & 1 & -1 & -1 & 1 & 0 \\
1 & -1 & -1 & 1 & 0 & -1 \\
1 & 0 & 2 & 1 & 0 & -1 \\
1 & 1 & -1 & 1 & 0 & -1 \\
1 & -1 & -1 & 1 & 0 & 1 \\
1 & 0 & 2 & 1 & 0 & 1 \\
1 & 1 & -1 & 1 & 0 & 1 \\
\end{bmatrix}$$

$$X'X = \begin{bmatrix}
12 & 0 & 0 & 0 & 0 & 0 \\
0 & 8 & 0 & 0 & 0 & 0 \\
0 & 0 & 24 & 0 & 0 & 0 \\
0 & 0 & 0 & 12 & 0 & 0 \\
0 & 0 & 0 & 0 & 6 & 0 \\
0 & 0 & 0 & 0 & 0 & 6 \\
\end{bmatrix}$$

$$(X'Y)' = (186 \ 11 \ 21 \ 24 \ 3 \ 15)$$
Now using our usual regression formulas, we see that the computer will calculate:

\[ SS(\text{regn}) = b'X'Y = \]

\[
\begin{align*}
\text{(blocks)} & \quad \text{(blocks)} & \quad \text{(feeds)} & \quad \text{L(F1)} & \quad \text{L(F2)} \\
11/8 & \quad 21/24 & \quad 24/12 & \quad 3/6 & \quad 15/6
\end{align*}
\]

This shows exactly how the computer can calculate ANOVA sums of squares using regression computations. The Q's from the "old fashioned" computational method are just entries of \(X'Y\). The "denom" values are just the entries in \(XX\). Now, since \(XX\) is a diagonal matrix the SS(regn) is just the sum of terms of the form Q*Q/denom, justifying the old fashioned computations. We finish the example by writing

\[ SS(\text{regn}) = CT + (15.125 + 18.375) + 48 + 1.5 + 37.5 \]

and thus

\[
\begin{array}{cccc}
\text{SOURCE} & \text{DF} & \text{SSq} & \text{Mn Sq} & \text{F} \\
\text{BLOCKS} & 2 & 33.5 & 16.75 & \\
F & 1 & 48.0 & 48.00 & 64 \\
\text{L}(f_1) & 1 & 1.5 & 1.50 & 2 \\
\end{array}
\]
One final computational note is this: The feed and light sums of squares could have been computed from totals. For example, the feed totals are 81 and 105, each one a total of 6 observations. Thus

$$SS(\text{FEED}) = 81\times81/6 + 105\times105/6 - CT = 48$$ (Check it out!)

To get the $SS(\text{FD}\times\text{LT})$ we take

$$SS(\text{trts}) - SS(\text{FEED}) - SS(\text{LIGHT}).$$

Similarly, the sums of squares for light within the feed levels can be computed from totals. Within $f_1$, the totals are 39 and 42, each a total of 3 observations. Using the correction term $CT^* = 81\times81/6$, we compute

$$SS(\text{L in } f_1) = 39\times39/3 + 42\times42/3 - 81\times81/6 = 1.5$$

See if you can compute $SS(\text{L in } f_2) = 37.5$ in this same way. What is $CT^*$ now?

At this point you are well advised to run all the regressions above and to check out the sums of squares by hand using totals and the $Q^*Q/denom$ method.

In addition, here is an example which you might want to pursue with your study group: Five 8 lb. packages of ground meat are available, each from a different animal. Split each package into four – 2 lb. packages and apply these treatments:
R spoilage retardant \( (r_1 = \text{none}, \ r_2 = \text{retardant added}) \)

T temperature \( (t_1 = 5 \text{ degrees}, \ t_2 = 25 \text{ degrees}) \).

From each animal, assign the four packages at random to treatments and store them that way. After 3 weeks, measure \( Y \) = amount of a chemical in parts per million where the chemical is formed by decomposition of the meat.

**DATA**

\[
\begin{array}{c|ccccc}
\text{TEMP} & 5 & 5 & 25 & 25 \\
\text{RETARDANT} & \text{NO} & \text{YES} & \text{NO} & \text{YES} \\
\text{ANIMAL} & \text{TOTAL} \\
1 & 30 & 28 & 50 & 30 & 138 \\
2 & 20 & 17 & 35 & 22 & 94 \\
3 & 25 & 26 & 40 & 25 & 116 \\
4 & 50 & 48 & 70 & 55 & 223 \\
5 & 42 & 45 & 68 & 52 & 207 \\
\end{array}
\]

**ANOVA**

\[
\begin{array}{l|l}
\text{BLOCKS} & 4 & 3214.3 \\
\text{TEMP} & 1 & 672.8 \\
\text{RETARDANT} & 1 & 336.2 \\
\text{T*R} & 1 & 288.8 \\
\text{ERROR} & 12 & 77.7 & \text{MSE} = 6.475 \\
\end{array}
\]

For practice, compute all sums of squares by hand and by computer. Discuss the interpretation of the data.
Now we extend the ideas above in two ways. First, we allow more than 2 factors and second, we allow the factors to appear at an arbitrary number of levels. One of the factors will be quantitative at equally spaced levels, allowing us to use orthogonal polynomials in our analysis. As usual we approach the topic through a contrived, but hopefully realistic, example.

EXAMPLE:

FACTOR
D: DOSE OF ANESTHETIC AGENT (5, 10, 15)
(units are 0.0001 gm per gm body wt.)

FACTOR S: SEX (M, F)
FACTOR A: AGE OF MOUSE AT TIME OF TEST
(6 MONTHS, 12 MONTHS)
(different mice at different times)

RESPONSE:
Time mouse remains in “surgical plane of anesthesia."

DESIGN:
RCB using 12 mice from each of 2 colonies.
Colonies are blocks here.

DATA:

<table>
<thead>
<tr>
<th>Age</th>
<th>Dose=5</th>
<th>Dose=10</th>
<th>Dose=15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
<td>M</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>12</td>
<td>17</td>
<td>23</td>
<td>24</td>
</tr>
</tbody>
</table>

Left number in each pair is from colony 1.
Colony totals: 332 and 388. (verify)

\[ CT = 720 \times 720 / 24 = 21600 \]
SS(total) = 18*18 + ... + 39*39 - CT = 1116
SS(blocks) = 332*332/12 + 388*388/12 - CT = 130.67
SS(trts) = 40*40/2 + ... + 72*72/2 - CT = 968
Thus SSE = 1116 - 130.67 - 968 = 17.33 with 11 d.f.

PRELIMINARY ANOVA

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOCKS</td>
<td>1</td>
<td>130.67</td>
<td></td>
</tr>
<tr>
<td>TRTS</td>
<td>11</td>
<td>968.00</td>
<td>88.0</td>
</tr>
<tr>
<td>ERROR</td>
<td>11</td>
<td>17.33</td>
<td>1.5754</td>
</tr>
</tbody>
</table>

There obviously are treatment effects (F = 55.86) so let us break the sums of squares down further to investigate them. We do this by creating “two way tables” and analyzing those as if there were only those 2 factors in the experiment (as you will see).

The S*A table

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>F</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 6</td>
<td>178</td>
<td>180</td>
<td>358</td>
</tr>
<tr>
<td>Age 12</td>
<td>178</td>
<td>184</td>
<td>362</td>
</tr>
<tr>
<td></td>
<td>356</td>
<td>364</td>
<td>720</td>
</tr>
</tbody>
</table>
Try using orthogonal contrasts with the totals to get these sums of squares by the Q*Q/denom method

<table>
<thead>
<tr>
<th>TOTAL</th>
<th>178</th>
<th>180</th>
<th>178</th>
<th>184</th>
<th>Q</th>
<th>denom</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>−1</td>
<td>+1</td>
<td>−1</td>
<td>+1</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>A</td>
<td>−1</td>
<td>−1</td>
<td>+1</td>
<td>+1</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>S*A</td>
<td>+1</td>
<td>−1</td>
<td>−1</td>
<td>+1</td>
<td></td>
<td>24</td>
</tr>
</tbody>
</table>

The S*A table yields sums of squares for S, A, and S*A. We construct the other two possible tables (S*D and A*D) as follows.

The S*D Table

<table>
<thead>
<tr>
<th></th>
<th>Dose=5</th>
<th>Dose=10</th>
<th>Dose=15</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>80</td>
<td>118</td>
<td>158</td>
<td>356</td>
</tr>
<tr>
<td>Female</td>
<td>102</td>
<td>120</td>
<td>142</td>
<td>364</td>
</tr>
<tr>
<td>Totals</td>
<td>182</td>
<td>238</td>
<td>300</td>
<td>720</td>
</tr>
</tbody>
</table>

Check these out:

SS(DOSE) = 871, SS(SEX) = 2.67, SS(table) = 964
SS(D*S) = 964 − 871 − 2.67 = 90.33
From a table of "orthogonal polynomial coefficients" we will get some contrasts whose sum of squares turn out to be exactly the same as the type I SS we would have obtained if we had regressed our Y variable on X, X*X, X*X*X, etc. In our case X is DOSE and since it is a 3 levels (2df) we will go only up to a quadratic term. From the table on page 387 of Steel & et al we find the contrasts for a factor (DOSE) at 3 levels. Here are the results:

<table>
<thead>
<tr>
<th>TOTAL</th>
<th>182</th>
<th>238</th>
<th>300</th>
<th>Q</th>
<th>den</th>
<th>SSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LINEAR</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>118</td>
<td>16</td>
<td>870.25</td>
</tr>
<tr>
<td>QUADRATIC</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>6</td>
<td>48</td>
<td>0.75</td>
</tr>
</tbody>
</table>

We can also decompose the interaction into a (SEX)*(DOSE LINEAR) and a (SEX)*(DOSE QUADRATIC) part by applying the weights in the following tables to the corresponding totals in our S*D table:

<table>
<thead>
<tr>
<th>Weights for S*D&lt;sub&gt;L&lt;/sub&gt;</th>
<th>Weights for S*D&lt;sub&gt;Q&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>D&lt;sub&gt;L&lt;/sub&gt;:</td>
<td>-1</td>
</tr>
<tr>
<td>S</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
</tr>
</tbody>
</table>

\[
SS(S*D<sub>L</sub>) = \frac{(80 - 158 - 102 + 142)^2}{16}
\]

\[
SS(S*D<sub>Q</sub>) = \frac{(-80 + 236 - 158 + 102 - 240 + 142)^2}{48}
\]

\[
SS(S*D<sub>L</sub>) = 90.25 \quad SS(S*D<sub>Q</sub>) = 0.0833
\]
Notes:
The 2 df for main effects of DOSE are decomposed into SS(D_L) and SS(D_Q). The interaction of dose with sex is similarly decomposed into linear and quadratic interactions.

The A*D Table

<table>
<thead>
<tr>
<th></th>
<th>Dose=5</th>
<th>Dose=10</th>
<th>Dose=15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age = 6</td>
<td>90</td>
<td>120</td>
<td>148</td>
</tr>
<tr>
<td>Age = 12</td>
<td>92</td>
<td>118</td>
<td>152</td>
</tr>
</tbody>
</table>

SS(table) = 874 SS(A*D) = 874 - 871 - 0.67 = 2.33

Recall that the treatment sum of squares was 968. Now the treatment sum of squares breaks down into SS(A) + SS(S) + SS(D) + SS(S*A) + SS(S*D) + SS(A*D) + SS(S*A*D). Since we have computed all the main effects and two way interactions we obtain the three way interaction by subtraction.

(This method can be extended to more factors. If there were another factor B we would compute the treatment sum of squares then sum over the levels of B to get a "three way table" like that at the beginning of this example. This would give us several main effects two and three way interactions. By analyzing all possible three way tables, we would get everything but the four way interaction which then would be obtained by subtraction.)
ANOVA

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOCKS</td>
<td>1</td>
<td>130.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>871.00</td>
<td>435.5</td>
<td>276.40</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>2.67</td>
<td></td>
<td>1.69 (careful !)</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>0.67</td>
<td></td>
<td>0.43</td>
</tr>
<tr>
<td>D*S</td>
<td>2</td>
<td>90.33</td>
<td>45.17</td>
<td>28.66 &lt;----*</td>
</tr>
<tr>
<td>D*A</td>
<td>2</td>
<td>2.33</td>
<td>1.17</td>
<td>0.74</td>
</tr>
<tr>
<td>S*A</td>
<td>1</td>
<td>0.67</td>
<td></td>
<td>0.43</td>
</tr>
<tr>
<td>D<em>S</em>A</td>
<td>2</td>
<td>0.33</td>
<td>0.17</td>
<td>0.11</td>
</tr>
<tr>
<td>ERROR</td>
<td>11</td>
<td>17.33</td>
<td>1.576 = MSE</td>
<td></td>
</tr>
</tbody>
</table>

The "careful" warning means that it would be rather misleading to interpret this as saying that factor S is not important. Factor S is part of a significant interaction. There is thus no reason to consider dropping S from the model.

SUMMARY:

1. Significant dose effect, but it seems linear in this dosage range.

2. Sex is important but it enters as an interaction, not a main effect.

3. No effect of age either as a main effect or interaction. What now? Analyze as sex and dose within sex. We now get
## ANOVA

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOCKS</td>
<td>1</td>
<td>130.67</td>
</tr>
<tr>
<td>SEX</td>
<td>1</td>
<td>2.67</td>
</tr>
<tr>
<td>$D_L(M)$</td>
<td>1</td>
<td>$760.50 = (-80 + 158)(-80 + 158)/8$</td>
</tr>
<tr>
<td>$D_L(F)$</td>
<td>1</td>
<td>$200.00 = (-102 + 142)(-102 + 142)/8$</td>
</tr>
<tr>
<td>OTHER</td>
<td>8</td>
<td>4.83</td>
</tr>
<tr>
<td>ERROR</td>
<td>11</td>
<td>17.33</td>
</tr>
</tbody>
</table>

Design note: We have used AGE as a variable here and have used colonies. I assume we can get enough mice of each sex and age from each colony.

CONCLUSION: We have traced the analysis down to a fairly simple conclusion. The male and female animals are reacting differently to the dosage of anesthetic. The response seems to be linear but with different slope and intercept in the two sexes. No dependence on age was found. It might now be nice to estimate the regression as follows:
DATA MOUSE;
INPUT Y SEX DLINM DLINF BLOCK;
CARDS;
18 1 5 0 1
22 1 5 0 0
22 0 0 5 1
28 0 0 5 0
SOURCE DF SS(TYPE I) SS(TYPE III)
26 1 10 0 1 BLOCK 1 130.67 130.67
34 1 10 0 0 SEX 1 2.67 88.59
27 0 0 10 1 DLINM 1 760.50 760.50
33 0 0 10 0 DLINF 1 200.00 200.00
38 1 15 0 1 ERROR 19 22.17
40 1 15 0 0 MSE = 1.167
33 0 0 15 1
37 0 0 15 0 PARAMETER ESTIMATE T STDERR
17 1 5 0 1 INTERCEPT 22.67 26.54 0.8539
23 1 5 0 0 BLOCK -4.67 -10.58 0.4410
24 0 0 5 1 SEX -10.17 -8.71 1.167
28 0 0 5 0 DLINM 1.95 25.53 0.076
27 1 10 0 1 DLINF 1.00 13.09 0.076
31 1 10 0 0
28 0 0 10 1 Note that we have lost some of our
32 0 0 10 0 orthogonality.
39 1 15 0 1
41 1 15 0 0
33 0 0 15 1
39 0 0 15 0

What do we make of this? First, there are 4 lines plotted above the “dose” axis. There is one for each (sex, colony) combination. The equations are:
Colony 1 Male
\[ \hat{Y} = (22.66 - 4.66 - 10.16) + 1.95*DOSE \]
Colony 2 Male
\[ \hat{Y} = (22.66 - 10.16) + 1.95*DOSE \]
Colony 1 Female
\[ \hat{Y} = (22.66 - 4.66) + 1.00*DOSE \]
Colony 2 Female
\[ \hat{Y} = (22.66) + 1.00*DOSE \]

Next, we notice that the male lines are not parallel to the female lines. It appears that males start out lower than females (male intercepts being 10.16 below the corresponding female ones) but that the males respond more to increases in dosage (1.95 per unit increase in dose as opposed to 1.00). Thus the males are out longer than females under high dosages. The reason that there are two male and two female lines is, of course, that we have plotted a line for each colony and the model allows the effect of colony to be only a shifting up or down of the level. The true response could not show all these effects and stay linear all the way down to dose 0 because 0 dose would give 0 response for males and females in both colonies.

A programming note: If we had more than the two blocks, (say k) we could have put BLOCKS in a class statement to let SAS create the k -1 block columns analogous to the 1 column we had. In that case, you would probably use the /SOLUTION option in your model statement to produce the parameter estimates.
THE UNDERLYING MODEL IN FACTORIAL EXPERIMENTS

We assume an experiment with no blocks and with fixed (as opposed to random) treatment effects. We use subscripts to denote the treatment combination and replication being referenced. We assume two factors A and B at a levels and b levels respectively. Assume r replications so \( n = abr \).

\[ Y_{ijk} = \text{Observation in } k^{\text{th}} \text{ replicate of } A \text{ at level } i \text{ and } B \text{ at level } j. \]
\[ \alpha_i = \text{Population effect of level } i \text{ of treatment } A \text{ (averaged over all other factors)} \]
\[ \beta_j = \text{Population effect of level } j \text{ of factor } B. \]
\[ (\alpha \beta)_{ij} = \text{Interaction effect in population.} \]

We also assume that the sums of these effects over \( i \) or \( j \) are 0. That is
\[ \sum_i \alpha_i = 0, \sum_j \beta_j = 0, \sum_i (\alpha \beta)_{ij} = 0, \sum_j (\alpha \beta)_{ij} = 0 \]

As a concrete example, let \( A \) be at 3 levels with
\[ \alpha_1 = 10, \alpha_2 = -7 \text{ and } \alpha_3 = -3. \]
Let \( B \) be at 2 levels with
\[ \beta_1 = 5 \text{ and thus } \beta_2 = -5. \]

Finally, assume \( \mu = 100 \). If we had \( Y_{ij} = \mu + \alpha_i + \beta_j \) then our \( Y \) observations would be as in this table:
TABLE WITH NO INTERACTION OR ERROR TERM

<table>
<thead>
<tr>
<th>$\mu = 100$</th>
<th>$\alpha_1 = 10$</th>
<th>$\alpha_2 = -7$</th>
<th>$\alpha_3 = -3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 = 5$</td>
<td>.</td>
<td>115</td>
<td>98</td>
</tr>
<tr>
<td>$\beta_2 = -5$</td>
<td></td>
<td>105</td>
<td>88</td>
</tr>
</tbody>
</table>

$Y_{ij} = \mu + \alpha_i + \beta_j$

Now let us throw in some interactions (they have to sum to 0 across rows and columns).

TABLE OF INTERACTIONS

<table>
<thead>
<tr>
<th>$(\alpha\beta)_{11} = -8$</th>
<th>$(\alpha\beta)_{21} = 3$</th>
<th>$(\alpha\beta)_{31} = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\alpha\beta)_{12} = 8$</td>
<td>$(\alpha\beta)_{22} = -3$</td>
<td>$(\alpha\beta)_{32} = -5$</td>
</tr>
</tbody>
</table>

$(\alpha\beta)_{ij}$

Now we add the numbers in the above two tables to illustrate what the data would look like if there were no sampling error.

FACTORIAL WITH INTERACTION, NO ERROR

| 107 | 101 | 107 |
| 113 | 85  | 87  |

$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$

Under this model, each rep would produce the numbers in the above table and our analysis would give 0 error sum of squares.
This is, of course unrealistic and so we add random error \( e \) to each observation in each replications and this brings us to the model

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

What I am trying to get across here is that the model we write down is supposed to draw to our mind the sequence of tables shown above. It illustrates how we are conceptualizing the creation of our data.

If we have 5 reps created from the model above then the observed table of sample means will, of course, not coincide with the last table above. We might get, for example

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>109</td>
<td>105</td>
<td>100</td>
</tr>
<tr>
<td>111</td>
<td>87</td>
<td>90</td>
</tr>
</tbody>
</table>

but we also get an MSE with which to judge the significance of main effects and interactions. Hopefully we could use the data to discover the nature of the population with which we are dealing. Suppose 109 is the average of 108, 110, 110, 107, 110. The error sum of squares is \((-1)^2 + 1^2 + 1^2 + (-2)^2 + 1^2\) plus similar contributions from the other 5 cells of the above table.

For example, if MSE = 80, the contrast of level 1 of A to level 2 of A would be 110 - 96 = 14 in the sample. This has variance \(2(MSE/10) = 16\). If you look at the model and remember that summing over \( j \) produces 0 for some of the model terms, you can see that
\[ \bar{Y}_{1..} = 110 = \mu + \alpha_1 + 0 + 0 + \bar{e}_{1..} \]
and
\[ \bar{Y}_{2..} = 96 = \mu + \alpha_2 + 0 + 0 + \bar{e}_{2..} \]
so that
\[ \bar{Y}_{1..} - \bar{Y}_{2..} = \alpha_1 - \alpha_2 + \bar{e}_{1..} - \bar{e}_{2..} \]

In other words, we are estimating \( \alpha_1 - \alpha_2 \) with an error that has mean 0 and variance 16. Since \( \alpha_1 = 10 \) and \( \alpha_2 = -7 \) we see that the difference is +17 which we have estimated as 14, easily within the sampling error of our experiment. In an experimental situation, of course, we do not know the parameter values and the above example is simply a pedagogical tool to illustrate our (statisticians') thinking.

RANDOM EFFECTS AND VARIANCE COMPONENTS

Let us start with an example of a two-factor factorial. Suppose we have

Factor A: two methods for counting somatic cells in milk.

Factor B: four technicians who will do the counting.

Now the two methods are the only two methods of interest. We are not trying to make inferences about counting methods other than those used in this experiment. This defines the effects of A as \textbf{fixed}.

On the other hand, we certainly do not want to restrict our inferences only to the four technicians in our experiment. Nevertheless, counting must be done by someone and we
cannot do the experiment without introducing some technician effects. If, for the moment, we assume no interaction, we write our model as

$$Y_{ij} = \mu + \tau_i + \rho_j + e_{ij}$$

The model expresses $Y$ as a sum of these parts:

An overall mean $\mu$

```
0  \mu
```

plus

A fixed treatment effect $\tau_i$ where

$$\tau_1 + \tau_2 = 0$$

```
\tau_1<-------0------->\tau_2
```

plus

A random effect $\rho_j$ from a $N(0, \sigma^2_\rho)$ distribution

```
X X _________ X
```

plus
A random error $e_{ij}$ from a $N(0, \sigma^2)$ population

Notice that, by setting $\mu$ properly, the “restriction” that the fixed effects sum to 0 is really no restriction at all on the possible values of the treatment group means in the two populations. There is also no real restriction in the assumption that the population mean of all technician effects is 0. However, there is also certainly no reason to expect that the average effect of the particular four technicians we chose will be exactly equal to the population mean 0. We summarize our assumptions on the various factors as:

$\mu$: a fixed constant,

$\tau_i$: two fixed constants which sum to 0, $\Sigma \tau_i = 0$
(Our goal: to contrast the two constants.)

$\rho_j$: a random value from a $N(0, \sigma^2)$ distribution,
(Our goal: to estimate $\sigma^2$ (or test that it is 0).)

Notes: It would not usually make sense to contrast technician means. Since the general readership for your results will not have these technicians available, they will only be interested in how much technician to technician variability to expect. On the other hand, the two methods of somatic cell counting should be contrasted and this is what your readers will want to see.
We should now understand several differences between random and fixed effects.

<table>
<thead>
<tr>
<th>RANDOM</th>
<th>FIXED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levels</td>
<td>selected at random from conceptually infinite collection of possibilities</td>
</tr>
<tr>
<td>Another expt.</td>
<td>would use different levels from same population</td>
</tr>
<tr>
<td>Goal</td>
<td>estimate variance components</td>
</tr>
<tr>
<td>Inference</td>
<td>for all levels of the factor (i.e., for population from which levels are selected)</td>
</tr>
<tr>
<td></td>
<td>only for levels actually used in the experiment</td>
</tr>
</tbody>
</table>

(* an exception is when Y has a polynomial relationship to some X - usually X would be considered fixed even though the polynomial can predict at any X)

The model currently under consideration is really a randomized complete block design with technicians as blocks. Note that the model does not allow for an interaction between technicians and methods of counting. We see that in general, a blocking factor is a random factor which does not interact with the treatments. Now there is nothing in the experiment that guarantees that there will be no interaction. If we are worried about this, we should check it. Our current experiment does not allow this
checking but we could fit an interaction term if we replicated each (method, technician) combination.

Let us suppose we do replicate each of our $2^*4 = 8$ treatment combinations, say $r$ times. We now have:

$$Y_{ijk} = \mu + \tau_i + \rho_j + (\tau \rho)_{ij} + e_{ijk}$$

The assumption on $(\tau \rho)_{ij}$ is that it is random with a distribution $\text{N}(0, \sigma_{\tau \rho}^2)$ but with the restriction that $\sum_i (\tau \rho)_{ij} = 0$. Using the letters $a$ and $b$ to stand in general for the number of levels of A and B, we can work out mathematically the expected values in repeated sampling of the mean squares in the ANOVA table.

First, recall that the sum of squares for A (methods) can be written in terms of means as follows:

$$\text{SS}(A) = rb\left[(\bar{Y}_{1..} - \bar{Y}_{...})^2 + (\bar{Y}_{2..} - \bar{Y}_{...})^2 + \ldots + (\bar{Y}_{a..} - \bar{Y}_{...})^2\right]$$

According to the model,

$$\bar{Y}_{i..} = \mu + \tau_i + \bar{\rho} + (\bar{\tau} \rho)_i + \bar{e}_{i..}$$

and

$$\bar{Y}_{...} = \mu + \bar{\tau} + \bar{\rho} + (\bar{\tau} \rho)_{..} + \bar{e}$$  \hspace{1cm} (Note: $\bar{\tau} = 0$ – Why?)

thus

$$\bar{Y}_{i..} - \bar{Y}_{...} = \tau_i + (\bar{\tau} \rho)_i - (\bar{\tau} \rho)_{..} + \bar{e}_{i..} - \bar{e}$$
Now if we assume that the random components of our model are independent, and if we remember how the variance of an average is computed, we can compute the expectation of SS(A).

We have

$$E\left\{ \sum (\bar{Y}_{i..} - \bar{Y}_{..})^2 \right\} =$$

$$E\left\{ \sum \tau_i^2 + \sum \left[ (\bar{\tau}_i)_{i..} - (\bar{\tau}_i)_{..} \right]^2 + \sum \left( \bar{e}_{i..} - \bar{e}_{..} \right)^2 \right\}$$

And we note that, for example, $\sum (\bar{e}_{i..} - \bar{e}_{..})^2$ is simply the numerator of a sample variance for a random sample of “$a$" values of $\bar{e}_{i..}$ and thus estimates

$$(a - 1) \text{ variance of } \bar{e}_{i..} = (a - 1) \sigma^2/(rb).$$

Similar reasoning shows that $\sum \left[ (\bar{\tau}_i)_{i..} - (\bar{\tau}_i)_{..} \right]^2$ estimates

$$(a - 1) \sigma_{\tau_i}^2/b.$$ 

Dividing by $a - 1$ to compute the mean square for $a$ and taking the expected value as above shows that the expected mean square for $A$ is

$$\sigma^2 + rb\sum \tau_i^2/(a - 1) + r\sigma_{\tau_i}^2$$

Similarly, we can compute expected mean squares for all the sources in our ANOVA table. Note that under $H_0$: no $A$ main effects, the mean square for $A$ is not estimating the same thing as MSE does. Therefore, MSE is not the appropriate denominator of $F$. 

---

D. A. Dickey
EXAMPLE: Two METHODS, four TECHNICIANS, 3 reps (also use T and M as suggestive subscripts).

ANOVA

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSQ</th>
<th>Mn Sq</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>METHOD</td>
<td>1</td>
<td>135</td>
<td>135</td>
<td>$\sigma^2 + 3\sigma_{MT}^2 + 12\left[\sum \tau_i^2 / (1)\right]$</td>
</tr>
<tr>
<td>TECHNICIAN</td>
<td>3</td>
<td>180</td>
<td>60</td>
<td>$\sigma^2 + 6\sigma_T^2$</td>
</tr>
<tr>
<td>METH*TECH</td>
<td>3</td>
<td>30</td>
<td>10</td>
<td>$\sigma^2 + 3\sigma_{MT}^2$</td>
</tr>
<tr>
<td>ERROR</td>
<td>16</td>
<td>64</td>
<td>4</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>

NOTE: Compute sums of squares just as in fixed effects case.

Test $H_0$: no METHOD effects:

$$F_{\frac{1}{3}} = 135/10 = 13.5$$

Test $H_0$: $\sigma_T^2 = 0$

$$F_{\frac{3}{16}} = 60/4 = 15$$

Estimate $\sigma_T^2 = \text{technician variance component}$:

$$(60 - 4)/6 = 9.33$$

Test $H_0$: $\sigma_{MT}^2 = 0$

$$F_{\frac{3}{16}} = 10/4 = 2.5$$
Contrast method 1 to method 2

\[ t = \frac{24.94 - 20.20}{\sqrt{2(10)/12}} = 3.67 \]

Note: for variance of a mean we have, as usual, divided an individual variance by 12 but the individual variance is not MSE, instead it is MS(MT) as used for testing the method effects before. This should seem sensible to you if you understood the F test.

Note: \( t \) has 3 degrees of freedom in this case.

Note: ANOVA sums of squares were just given in this problem. I did not show the computation from original data — it is the same as we have been doing all semester.

We do not want to do the expected mean square algebra on every problem. Fortunately there are some rules which can be used in the balanced case to give the expected mean squares. The rules are:

1. Write down all the variance components as though all factors are random.

2. To each variance component, assign a multiplier equal to the product of the number of levels of all the factors not represented in the variance component subscript.

3. For any effect (like ST in the example below) put \( X \) under each variance component whose subscript contains all the letters in the description of the effect (must contain both S and T) and put \( X \) under \( \sigma^2 \).
4. Look at the letters in each subscript which are not contained in the source. (For example in the ST row below, the source is ST and we see an X under $\sigma_{CST}^2$ so we consider the letter C.) If any of these remaining letters (those in the subscript but not in the source) represent fixed effects, erase the X. Thus the X is not erased.

5. If all letters in a subscript represent fixed effects, replace $\sigma^2$ by $\kappa^2$. On the previous page, the $12\sum T_i^2/(1)$ would become $12\kappa_T^2$ (optional).

EXAMPLE: Four cars of a certain type and 2 test tracks are selected by a regulatory agency. Each car is run on each test track three times at 70 MPH, three at 55 and three at 40 MPH. The amount of gas burned is measured by a very accurate device mounted on the carburetor of each car and miles per gallon, MPG, is found. The sources are symbolized

\begin{align*}
C: & \text{ cars random at } c = 4 \text{ levels} \\
T: & \text{ tracks random at } t = 2 \text{ levels} \\
S: & \text{ speeds fixed at } s = 3 \text{ levels} \\
r: & = 3 \text{ reps in a CRD.}
\end{align*}

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>$\sigma^2$</th>
<th>18 $\sigma_C^2$</th>
<th>36 $\sigma_T^2$</th>
<th>24 $\kappa_S^2$</th>
<th>9 $\sigma_{CT}^2$</th>
<th>12 $\kappa_T^2$</th>
<th>6 $\sigma_{CS}^2$</th>
<th>3 $\sigma_{CST}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>(X)</td>
<td>(X)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>(X)</td>
<td>(X)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>S</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>C*T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>C*S</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>S*T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>C<em>S</em>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>error</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

D. A. Dickey
The following SAS program uses some data which we will think of as coming from the mileage experiment above.

data cars; input speed car @;
  do track = 1 to 2; do rep=1 to 3; input MPG @;
  output; end; end;
*      <--Track 1--->     <--Track 2 -->   ;
cards;
70  1  19.3 18.3 20.3 20.8 21.2 20.2
70  2  19.0 21.7 20.2 18.4 19.7 19.4
70  3  16.5 16.2 15.2 14.7 16.4 17.0
70  4  16.6 16.6 16.8 17.6 18.0 18.9
55  1  18.9 18.1 19.2 20.4 21.7 21.0
55  2  19.4 18.7 20.7 21.9 23.0 21.0
55  3  20.5 19.4 18.9 20.1 20.0 20.5
55  4  17.1 16.5 17.2 18.0 19.4 18.3
40  1  22.6 24.8 22.2 25.3 26.1 27.1
40  2  23.2 20.9 20.6 22.7 24.0 21.9
40  3  18.3 17.8 19.3 20.4 19.0 20.0
40  4  21.8 21.7 19.5 22.7 20.7 22.6
;  
proc means; var mpg;

PROC  GLM; CLASS SPEED CAR TRACK;
  MODEL  MPG  = SPEED|CAR|TRACK;
  RANDOM CAR TRACK SPEED*CAR SPEED*TRACK CAR*TRACK
  SPEED*CAR*TRACK/TEST;
  CONTRAST 'SPEED_L' SPEED -1 0 1;
  CONTRAST 'SPEED_Q' SPEED -1 2 -1;

PROC MIXED covtest; CLASS SPEED CAR TRACK;
  MODEL MPG = SPEED / DDFM=SATTERTHWAITE;
  RANDOM CAR TRACK SPEED*CAR SPEED*TRACK
  CAR*TRACK SPEED*CAR*TRACK ;
  CONTRAST 'SPEED_L' SPEED -1 0 1;
  CONTRAST 'SPEED_Q' SPEED -1 2 -1;
run;
Analysis Variable : MPG

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>19.9180556</td>
<td>2.4854691</td>
<td>14.70000</td>
<td>27.10000</td>
</tr>
</tbody>
</table>

General Linear Models Procedure

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEED</td>
<td>3</td>
<td>40 55 70</td>
</tr>
<tr>
<td>CAR</td>
<td>4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>TRACK</td>
<td>2</td>
<td>1 2</td>
</tr>
</tbody>
</table>

Number of observations in data set = 72

Dependent Variable: MPG

<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>23</td>
<td>398.44653</td>
<td>17.32376</td>
<td>20.71</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>48</td>
<td>40.16000</td>
<td>0.83667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr Total</td>
<td>71</td>
<td>438.60653</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square          C.V.  Root MSE        MPG Mean
0.908437          4.592290      0.9147          19.918

<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>SPEED</td>
<td>2</td>
<td>158.93528</td>
<td>79.46764</td>
<td>94.98</td>
<td>0.0001</td>
</tr>
<tr>
<td>CAR</td>
<td>3</td>
<td>128.03042</td>
<td>42.67681</td>
<td>51.01</td>
<td>0.0001</td>
</tr>
<tr>
<td>SPEED*CAR</td>
<td>6</td>
<td>62.30917</td>
<td>10.38486</td>
<td>12.41</td>
<td>0.0001</td>
</tr>
<tr>
<td>TRACK</td>
<td>1</td>
<td>29.51681</td>
<td>29.51681</td>
<td>35.28</td>
<td>0.0001</td>
</tr>
<tr>
<td>SPEED*TRACK</td>
<td>2</td>
<td>5.97861</td>
<td>2.98931</td>
<td>3.57</td>
<td>0.0358</td>
</tr>
<tr>
<td>CAR*TRACK</td>
<td>3</td>
<td>6.67931</td>
<td>2.22644</td>
<td>2.66</td>
<td>0.0586</td>
</tr>
<tr>
<td>SPED<em>CR</em>TRK</td>
<td>6</td>
<td>6.99694</td>
<td>1.16616</td>
<td>1.39</td>
<td>0.2365</td>
</tr>
</tbody>
</table>

D. A. Dickey
### Summary 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>SPEED</td>
<td>2</td>
<td>158.93528</td>
<td>79.46764</td>
<td>94.98</td>
<td>0.0001</td>
</tr>
<tr>
<td>CAR</td>
<td>3</td>
<td>128.03042</td>
<td>42.67681</td>
<td>51.01</td>
<td>0.0001</td>
</tr>
<tr>
<td>SPEED*CAR</td>
<td>6</td>
<td>62.30917</td>
<td>10.38486</td>
<td>12.41</td>
<td>0.0001</td>
</tr>
<tr>
<td>TRACK</td>
<td>1</td>
<td>29.51681</td>
<td>29.51681</td>
<td>35.28</td>
<td>0.0001</td>
</tr>
<tr>
<td>SPEED*TRACK</td>
<td>2</td>
<td>5.97861</td>
<td>2.98931</td>
<td>3.57</td>
<td>0.0358</td>
</tr>
<tr>
<td>CAR*TRACK</td>
<td>3</td>
<td>6.67931</td>
<td>2.22644</td>
<td>2.66</td>
<td>0.0586</td>
</tr>
<tr>
<td>SPED<em>CR</em>TRK</td>
<td>6</td>
<td>6.99694</td>
<td>1.16616</td>
<td>1.39</td>
<td>0.2365</td>
</tr>
</tbody>
</table>

--- results of RANDOM statement ---

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEED</td>
<td>Var(Error) + 3 Var(SPEED<em>CAR</em>TRACK) + 12 Var(SPEED<em>TRACK) + 6 Var(SPEED</em>CAR) + Q(SPEED)</td>
</tr>
<tr>
<td>CAR</td>
<td>Var(Error) + 3 Var(SPEED<em>CAR</em>TRACK) + 9 Var(CAR<em>TRACK) + 6 Var(SPEED</em>CAR) + 18 Var(CAR)</td>
</tr>
<tr>
<td>SPEED*CAR</td>
<td>Var(Error) + 3 Var(SPEED<em>CAR</em>TRACK) + 6 Var(SPEED*CAR)</td>
</tr>
<tr>
<td>TRACK</td>
<td>Var(Error) + 3 Var(SPEED<em>CAR</em>TRACK) + 9 Var(CAR<em>TRACK) + 12 Var(SPEED</em>TRACK) + 36 Var(TRACK)</td>
</tr>
<tr>
<td>SPEED*TRACK</td>
<td>Var(Error) + 3 Var(SPEED<em>CAR</em>TRACK) + 12 Var(SPEED*TRACK)</td>
</tr>
<tr>
<td>CAR*TRACK</td>
<td>Var(Error) + 3 Var(SPEED<em>CAR</em>TRACK) + 9 Var(CAR*TRACK)</td>
</tr>
<tr>
<td>SPED<em>CR</em>TRK</td>
<td>Var(Error) + 3 Var(SPEED<em>CAR</em>TRACK)</td>
</tr>
</tbody>
</table>

{note that items in bold italic would not appear in the model with summing to 0 restrictions}
--- results of /TEST in RANDOM statement ------

General Linear Models Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: MPG

Source: SPEED
Error: MS(SPEED*CAR) + MS(SPEED*TRACK) - MS(SPEED*CAR*TRACK)

\[
\begin{array}{cccc}
\text{Denominator} & \text{DF} & \text{Type III MS} & \text{Denominator} \\
\hline
& 2 & 79.467 & 6.57 & 12.208 \\
\end{array}
\]

F Value  Pr > F  
6.5095  0.0276

Source: CAR
Error: MS(SPEED*CAR) + MS(CAR*TRACK) - MS(SPEED*CAR*TRACK)

\[
\begin{array}{cccc}
\text{Denominator} & \text{DF} & \text{Type III MS} & \text{Denominator} \\
\hline
& 3 & 42.677 & 6.60 & 11.445 \\
\end{array}
\]

F Value  Pr > F  
3.7288  0.0729

Source: SPEED*CAR
Error: MS(SPEED*CAR*TRACK)

\[
\begin{array}{cccc}
\text{Denominator} & \text{DF} & \text{Type III MS} & \text{Denominator} \\
\hline
& 6 & 10.385 & 2.58 & 4.049 \\
\end{array}
\]

F Value  Pr > F  
8.9052  0.0088

Source: TRACK
Error: MS(SPEED*TRACK) + MS(CAR*TRACK) - MS(SPEED*CAR*TRACK)

\[
\begin{array}{cccc}
\text{Denominator} & \text{DF} & \text{Type III MS} & \text{Denominator} \\
\hline
& 1 & 29.517 & 2.58 & 4.049 \\
\end{array}
\]

F Value  Pr > F  
7.2889  0.0868

D. A. Dickey
Source: SPEED*TRACK  
Error: MS(SPEED*CAR*TRACK)

<table>
<thead>
<tr>
<th>Denominator</th>
<th>Denominator</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF Type III MS</td>
<td>DF MS F Value Pr &gt; F</td>
</tr>
<tr>
<td>2 2.9893055556</td>
<td>6 1.1661574074 2.5634 0.1568</td>
</tr>
</tbody>
</table>

Source: CAR*TRACK  
Error: MS(SPEED*CAR*TRACK)

<table>
<thead>
<tr>
<th>Denominator</th>
<th>Denominator</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF Type III MS</td>
<td>DF MS F Value Pr &gt; F</td>
</tr>
<tr>
<td>3 2.2264351852</td>
<td>6 1.1661574074 1.9092 0.2293</td>
</tr>
</tbody>
</table>

Source: SPEED*CAR*TRACK  
Error: MS(Error)

<table>
<thead>
<tr>
<th>Denominator</th>
<th>Denominator</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF Type III MS</td>
<td>DF MS F Value Pr &gt; F</td>
</tr>
<tr>
<td>6 1.1661574074</td>
<td>48 0.8366666667 1.3938 0.2365</td>
</tr>
</tbody>
</table>

- results of CONTRAST statement in GLM: wrong denominators!! --

Contrast DF Contrast SS  Mean Square  F Value  Pr > F

<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>SPEED_L</td>
<td>1</td>
<td>154.80083</td>
<td>154.80083</td>
<td>185.02</td>
<td>0.0001</td>
</tr>
<tr>
<td>SPEED_Q</td>
<td>1</td>
<td>4.13444</td>
<td>4.13444</td>
<td>4.94</td>
<td>0.0310</td>
</tr>
</tbody>
</table>
The MIXED Procedure

Class Level Information

Class     Levels  Values
SPEED          3  40 55 70
CAR            4  1 2 3 4
TRACK          2  1 2

REML Estimation Iteration History

Iteration  Evaluations     Objective     Criterion
0            1  175.10023430
1            1  104.17593667    0.00000000
Convergence criteria met.

Covariance Parameter Estimates (REML)

Cov Parm       Estimate    Std Error      Z  Pr > |Z|
CAR              1.73509259   1.96725738   0.88    0.3778
TRACK            0.70742284   1.16374553   0.61    0.5433
SPEED*CAR        1.53645062   1.00556444   1.53    0.1265
SPEED*TRACK      0.15192901   0.25534910   0.59    0.5519
CAR*TRACK        0.11780864   0.21539466   0.55    0.5844
SPEED*CAR*TRACK  0.10983025   0.23153469   0.47    0.6352
Residual         0.83666667   0.17078387   4.90    0.0001

Model Fitting Information for MPG

Description                Value
Observations             72.0000
Res Log Likelihood      -115.495
Akaike's Information Criterion -122.495
Schwarz's Bayesian Criterion  -130.314
-2 Res Log Likelihood        230.9895
Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>NDF</th>
<th>DDF</th>
<th>Type III F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEED</td>
<td>2</td>
<td>6.57</td>
<td>6.51</td>
<td>0.0276</td>
</tr>
</tbody>
</table>

CONTRAST Statement Results

<table>
<thead>
<tr>
<th>Source</th>
<th>NDF</th>
<th>DDF</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEED_L</td>
<td>1</td>
<td>6.57</td>
<td>12.68</td>
<td>0.0102</td>
</tr>
<tr>
<td>SPEED_Q</td>
<td>1</td>
<td>6.57</td>
<td>0.34</td>
<td>0.5800</td>
</tr>
</tbody>
</table>

some notes:

MIXED gives Wald tests for variance components - unreliable even in moderate sized samples. We are unable to detect any random effects with this rough approximation.

We could do a full versus reduced model "likelihood ratio" test with 2 MIXED runs.

GLM gives exact F tests ot Satterthwaite approximations for random effects.

MIXED automatically gives correct standard errors for contrasts and corrected tests.

GLM cannot get correct denominators for contrasts when mixtures are involved.

MIXED never gives negative variance components estimates.

GLM can get negative variance component estimates. When this happens, MIXED will disagree with GLM in those places AND OTHERS.
HYPOTHESIS TESTS FOR MIXED MODELS:

Suppose we want to test for speed effects. The mean square for speed, 79.4676, is an estimate of

\[ \sigma^2 + 12 \sigma^2_{ST} + 6 \sigma^2_{CS} + 3 \sigma^2_{CST} + 24 \kappa_S^2 \]

where \( \kappa_S^2 = \sum_{i=1}^{3} \tau_i^2/(3-1) \) and \( \tau_i \) is the effect of driving at speed \( i \).

If speed has no effect then all \( \tau_i \)'s, and hence \( \kappa_S^2 \), are 0. The question, then, is whether 79.4676 is just an estimate of

\[ \sigma^2 + 12 \sigma^2_{ST} + 6 \sigma^2_{CS} + 3 \sigma^2_{CST}. \]

We want to take the ratio of 79.4676 to something that estimates \( \sigma^2 + 12 \sigma^2_{ST} + 6 \sigma^2_{CS} + 3 \sigma^2_{CST} \) but there is no single mean square that does the job. However, we see that if we take

\[ MS(CS) + MS(ST) - MS(CST) = 10.3849 + 2.9893 - 1.1662 \]

we have an estimate of \( \sigma^2 + 12 \sigma^2_{ST} + 6 \sigma^2_{CS} + 3 \sigma^2_{CST}. \) We see that \( 10.3849+2.9893-1.1662 = 12.2080 > 0 \) but there is nothing forcing the estimate or the resulting F to be positive. Thus F could be a meaningless negative number. In SAS PROC GLM, this is the method used by the RANDOM statement to construct F. PROC MIXED will not give negative variance component estimates.
\[ \sigma^2 + 18 \sigma_C^2 + 36 \sigma_T^2 + 24 \kappa_S^2 + 9 \sigma_{CT}^2 + 12 \sigma_{ST}^2 + 6 \sigma_{CS}^2 + 3 \sigma_{CST}^2 \]

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEED*CAR</td>
<td>6</td>
<td>62.30917</td>
<td>10.38486</td>
</tr>
<tr>
<td>SPEED*TRACK</td>
<td>2</td>
<td>5.97861</td>
<td>2.98931</td>
</tr>
<tr>
<td>SPED<em>CAR</em>TRK</td>
<td>6</td>
<td>6.99694</td>
<td>1.16616</td>
</tr>
</tbody>
</table>

Note that if we always add mean squares, for example we can take \( F \) as \( \frac{\text{MS(S) + MS(CST)}}{\text{MS(CS) + MS(ST)}} \) then we see that the expected value of the numerator is the same as that of the denominator plus 24 \( \kappa_S^2 \) so this will be testing the right thing and will never be negative.

Either way we choose to construct \( F \) (really \( F' \), an approximation to \( F \)) we need to get some approximate degrees of freedom. Satterthwaite is the researcher who showed that these ratios are distributed approximately as \( F \). His degrees of freedom for the linear combination

\[ \text{MS}_{\text{constructed}} = c_1 \text{MS}_1 + c_2 \text{MS}_2 + \ldots + c_k \text{MS}_k \]

is (Steel et al formula 7.17 or Ott & Longnecker pg. 987):
\[
\text{df} = \frac{(\text{MS}_{\text{constructed}})^2}{\sum_{i} \frac{(c_i \text{MS}_i)^2}{\text{df}_i}} = \frac{12.2080^2}{6} + \frac{2.9893^2}{2} + \frac{(-1.1662)^2}{6} = 6.57
\]

(for speed F denom.)

Now consider the random effect CAR and the F test for \( H_0 : \sigma_C^2 = 0 \). The SAS models give corrected F test 3.73 (P-value .073 and 6.60 denominator df) from GLM and a Wald test \( Z = 0.38 \) (P-value .3778 from the N(0,1) tables).

Our textbook model (summing to 0 restrictions assumed) produces the exact F test \( F^3_3 \) test \( \text{MS}(\text{CAR}) / \text{MS}(\text{CAR*TRACK}) = 42.6768 / 2.2264 = 19.17 \) (P-value 0.0185). We expect that there really is car to car variation in MPG and our summing to 0 assumption seems to have bought us a more powerful test - the only one that shows significance at .05. It is an exact test whereas GLM using Satterthwaite is approximate and MIXED using Wald gives a poor approximation to the normal distribution it uses for its P-values (it would be OK if we had hundreds of cars).

In summary, there are two types of models used for mixed effects based on whether one uses summing to 0 in the definition of the interaction of a fixed and random effect. It is at best difficult to extend the summing to 0 model to unbalanced data. Perhaps for this reason, no SAS procedure at present uses summing to 0. When an exact summing to 0 model is desired, we can use the TEST statement in GLM if Satterthwaite mixtures are not required. The TEST statement does not allow \( E = \) a mixture of mean squares. If the SAS model is used, GLM and MIXED often agree and are correct for the fixed effects tests assuming RANDOM / TEST is used in GLM and DDFM=SATTERTHWAITE is used in MIXED. I recommend these options every time a mixed model is analyzed. The fixed
effects are often those of most interest so this agreement is reassuring. Contrast statements are not corrected in GLM but are correct in MIXED. When negative variance component estimates are produced by GLM, the output will differ from that of MIXED for this and other effects. Finally, as to random effects, the Satterthwaite method of GLM or the likelihood ratio method (from two runs) in MIXED are reasonably accurate whereas the Wald tests should only be used when that random effect is observed at a very large number of levels, a rare situation in practice.

**Blocks versus random effects ?? (optional)**

We see that blocks are just the levels of a random (usually) factor in a factorial experiment but we *assume* they do not interact with treatments. If we attempt to fit a block*treatment interaction, we then have no degrees of freedom for error and hence no test. With 5 blocks and 4 treatments there are 12 degrees of freedom for error and block*treatment would use up all 12 of them. Tukey suggested taking a 1 degree of freedom sum of squares out of the error term. This sum of squares will be large for what he suggests is a common type of block*treatment interaction. An example of this situation occurs when blocks and treatments do not interact on the logarithmic scale but the experimenter fails to take logarithms. If Tukey's test is significant, try a logarithmic transformation.

The test is of the familiar form $Q^2/\text{denominator}$. Let the model be

$$
Y_{ij} = \mu + \rho_i + \tau_j + e_{ij}
$$
and compute sample block effects $\bar{Y}_{i.} - \bar{Y}_{..}$ and treatment effects $\bar{Y}_{.j} - \bar{Y}_{..}$ then compute

$$Q = \sum_i \sum_j (\bar{Y}_{i.} - \bar{Y}_{..}) Y_{ij}(\bar{Y}_{.j} - \bar{Y}_{..})$$

and

$$\text{denominator} = \left[ \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \right] \left[ \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2 \right]$$

For example, suppose we have 2 treatments and 3 blocks with this data:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>12</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>B</td>
<td>28</td>
<td>64</td>
<td>118</td>
</tr>
</tbody>
</table>

(20-45)=-25 \quad (70-45)=25

\[
\begin{align*}
20-45 & \quad 40-45 & \quad 75-45 \\
-25 & \quad -5 & \quad 30
\end{align*}
\]

$$Q = (-25)12(-25) + (-5)16(-25) + (30)32(-25) +
\quad (-25)28(25) + (-5)64(25) + (30)118(25) = 48500$$

and denominator = \[25^2+25^2][25^2+5^2+30^2] = 1937500

Thus the 1 df sum of squares is $Q^2/193750 = 1214.06$. 

**ANOVA**

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks</td>
<td>2</td>
<td>3100</td>
</tr>
<tr>
<td>Trt</td>
<td>1</td>
<td>3750</td>
</tr>
<tr>
<td>Tukey</td>
<td>1</td>
<td>1214</td>
</tr>
<tr>
<td>Error</td>
<td>1</td>
<td>14</td>
</tr>
</tbody>
</table>

D. A. Dickey
SPLIT PLOT EXAMPLE

To assess the effects of soils on growth in underwater plants, we take small dishes and in each dish, we put one of each of three soils. There are also two varieties of plants so we have a 3*2 factorial arrangement. Each replication thus requires 6 dishes.

The experiment will be performed in four aquaria. Each aquarium will hold six dishes and, of course, each aquarium will be filled with water.

The dependent variable will be the dry weight, in some units, of the harvested plants. (What kind of design do we have so far?)

Now suppose we are interested in some effect of the water surrounding the plants. For example, we may want to see what happens if we aerate the water. Note that the treatment units for aeration are the aquaria——that is, the blocks from the other experiment.

Now we combine the experiment on the aquaria with the experiment within the aquaria. Here are the dry weight yields:

Aquarium 1: aerated

<table>
<thead>
<tr>
<th>soil 1</th>
<th>soil 2</th>
<th>soil 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>variety 1</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>variety 2</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>total:</td>
<td>30</td>
<td>42</td>
</tr>
</tbody>
</table>
### Aquarium 2: not aerated

<table>
<thead>
<tr>
<th></th>
<th>soil 1</th>
<th>soil 2</th>
<th>soil 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>variety 1</td>
<td>8</td>
<td>12</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>variety 2</td>
<td>13</td>
<td>15</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>total:</td>
<td>21</td>
<td>27</td>
<td>22</td>
<td>70</td>
</tr>
</tbody>
</table>

### Aquarium 3: not aerated

<table>
<thead>
<tr>
<th></th>
<th>soil 1</th>
<th>soil 2</th>
<th>soil 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>variety 1</td>
<td>10</td>
<td>14</td>
<td>11</td>
<td>35</td>
</tr>
<tr>
<td>variety 2</td>
<td>13</td>
<td>17</td>
<td>15</td>
<td>45</td>
</tr>
<tr>
<td>total:</td>
<td>23</td>
<td>31</td>
<td>26</td>
<td>80</td>
</tr>
</tbody>
</table>

### Aquarium 4: aerated

<table>
<thead>
<tr>
<th></th>
<th>soil 1</th>
<th>soil 2</th>
<th>soil 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>variety 1</td>
<td>16</td>
<td>25</td>
<td>19</td>
<td>60</td>
</tr>
<tr>
<td>variety 2</td>
<td>19</td>
<td>28</td>
<td>23</td>
<td>70</td>
</tr>
<tr>
<td>total:</td>
<td>35</td>
<td>53</td>
<td>42</td>
<td>130</td>
</tr>
</tbody>
</table>
We can organize the data in a SAS dataset, compute sums of squares and tables of means using PROC MEANS and PROC GLM. In PROC GLM, error A would be specified as AQUARIUM(AIR). Here is a printout of such a dataset.

PROC PRINT; RUN;

```
OBS  YIELD  VAR AQUARIUM  SOIL  AIR
 1    10     1     1     1     1
 2    14     1     1     2     1
 3    12     1     1     3     1
 4    20     2     1     1     1
 5    28     2     1     2     1
 6    26     2     1     3     1
 7     8     1     2     1     0
 8    12     1     2     2     0
 9    10     1     2     3     0
10    13     2     2     1     0
11    15     2     2     2     0
12    12     2     2     3     0
13    10     1     3     1     0
14    14     1     3     2     0
15    11     1     3     3     0
16    13     2     3     1     0
17    17     2     3     2     0
18    15     2     3     3     0
19    16     1     4     1     1
20    25     1     4     2     1
21    19     1     4     3     1
22    19     2     4     1     1
23    28     2     4     2     1
24    23     2     4     3     1
```
PROC GLM; CLASS VAR AQUARIUM SOIL AIR;
MODEL YIELD = AIR AQUARIUM(AIR) SOIL VAR
SOIL*VAR AIR*SOIL AIR*VAR AIR*SOIL*VAR; RUN;

R-SQUARE C.V. ROOT MSE YIELD MEAN
0.906206 17.11 2.780887150 16.2500000

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>TYPE I</th>
<th>SS</th>
<th>F VALUE</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIR</td>
<td>1</td>
<td>337.5000</td>
<td>43.64</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>AQUARIUM(AIR)</td>
<td>2</td>
<td>41.6667</td>
<td>2.69</td>
<td>0.1159</td>
<td></td>
</tr>
<tr>
<td>SOIL</td>
<td>2</td>
<td>121.7500</td>
<td>7.87</td>
<td>0.0088</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>1</td>
<td>192.6667</td>
<td>24.91</td>
<td>0.0005</td>
<td></td>
</tr>
<tr>
<td>VAR*SOIL</td>
<td>2</td>
<td>0.5833</td>
<td>0.04</td>
<td>0.9631</td>
<td></td>
</tr>
<tr>
<td>SOIL*AIR</td>
<td>2</td>
<td>16.7500</td>
<td>1.08</td>
<td>0.3752</td>
<td></td>
</tr>
<tr>
<td>VAR*AIR</td>
<td>1</td>
<td>32.6667</td>
<td>4.22</td>
<td>0.0669</td>
<td></td>
</tr>
<tr>
<td>VAR<em>SOIL</em>AIR</td>
<td>2</td>
<td>3.5833</td>
<td>0.23</td>
<td>0.7973</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>TYPE III</th>
<th>SS</th>
<th>F VALUE</th>
<th>PR &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIR</td>
<td>1</td>
<td>337.5000</td>
<td>43.64</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>AQUARIUM(AIR)</td>
<td>2</td>
<td>41.6667</td>
<td>2.69</td>
<td>0.1159</td>
<td></td>
</tr>
<tr>
<td>SOIL</td>
<td>2</td>
<td>121.7500</td>
<td>7.87</td>
<td>0.0088</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>1</td>
<td>192.6667</td>
<td>24.91</td>
<td>0.0005</td>
<td></td>
</tr>
<tr>
<td>VAR*SOIL</td>
<td>2</td>
<td>0.5833</td>
<td>0.04</td>
<td>0.9631</td>
<td></td>
</tr>
<tr>
<td>SOIL*AIR</td>
<td>2</td>
<td>16.7500</td>
<td>1.08</td>
<td>0.3752</td>
<td></td>
</tr>
<tr>
<td>VAR*AIR</td>
<td>1</td>
<td>32.6667</td>
<td>4.22</td>
<td>0.0669</td>
<td></td>
</tr>
<tr>
<td>VAR<em>SOIL</em>AIR</td>
<td>2</td>
<td>3.5833</td>
<td>0.23</td>
<td>0.7973</td>
<td></td>
</tr>
</tbody>
</table>
Looking at the experiment as a randomized complete block design with factorial treatments VAR and SOIL (ignoring the AIR factor) we would obtain:

ANOVA

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks</td>
<td>3</td>
<td>379.167</td>
<td>126.4</td>
</tr>
<tr>
<td>Soils</td>
<td>2</td>
<td>121.750</td>
<td>60.9</td>
</tr>
<tr>
<td>Var</td>
<td>1</td>
<td>192.667</td>
<td>192.7</td>
</tr>
<tr>
<td>S*V</td>
<td>2</td>
<td>0.583</td>
<td>0.3</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>130.333</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Next consider the AIR experiment which was done on the aquaria. Now we envision an analysis which has 4 treatment units and 2 levels of air producing an ANOVA like this:

TRTS 1
ERROR 2
We will divide the BLOCK sum of squares, 379.167, into these 1 and 2 degree of freedom pieces. We will call the 2 df error term a “whole plot" error or ERROR A. The sums of squares are computed simply from a table of BLOCK (i.e., aquarium) totals:

\[
\begin{array}{cc}
\text{AIR} & \\
0 & 1 \\
70 & 110 \\
80 & 130 \\
150 & 240 \\
\end{array}
\]

\[
\text{Table SSq} = \frac{70 \times 70}{6} + \frac{80 \times 80}{6} + \frac{110 \times 110}{6} + \frac{130 \times 130}{6} - CT \\
= 379.167
\]

\[
\text{Air SSq} = \frac{150 \times 150}{12} + \frac{240 \times 240}{12} - CT = 337.500
\]

\[
\text{ERROR A SSq} = 379.167 - 337.500 = 41.667.
\]

Note that the air experiment was not blocked. Thus the rows of the table have no meaning. Another way to say this is that a table like the above except with 110 and 130 interchanged is just as valid a way to report the data as is the above table. If the rows had been blocks (like 2 benches on which the aquaria were set) then this interchange of 110 and 130 would destroy the meaning of the rows (which should be representing benches). In SAS, the 2 df sum of squares would be referred to as AQUARIUM(AIR), read as aquarium-to-aquarium variation within the levels of AIR. This instruction tells SAS to compute the table sum of squares and subtract off only the AIR SSq from it.
Now the ANOVA has two errors – ERROR A for the whole plot F tests and ERROR B for the split plot F tests. The table looks like this:

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIR</td>
<td>1</td>
<td>337.50</td>
<td>337.50</td>
<td>16.2</td>
</tr>
<tr>
<td>ERROR A</td>
<td>2</td>
<td>41.67</td>
<td>20.83</td>
<td>&lt;-----</td>
</tr>
<tr>
<td>SOIL</td>
<td>2</td>
<td>121.75</td>
<td>60.88</td>
<td>7.9</td>
</tr>
<tr>
<td>VAR</td>
<td>1</td>
<td>192.67</td>
<td>192.67</td>
<td>24.9</td>
</tr>
<tr>
<td>S*V</td>
<td>2</td>
<td>0.58</td>
<td>0.24</td>
<td>0.0</td>
</tr>
<tr>
<td>A*S</td>
<td>2</td>
<td>16.75</td>
<td>8.38</td>
<td>&gt; divide 1.1</td>
</tr>
<tr>
<td>A*V</td>
<td>1</td>
<td>32.67</td>
<td>32.67</td>
<td>&gt; by 4.2</td>
</tr>
<tr>
<td>A<em>S</em>V</td>
<td>2</td>
<td>3.58</td>
<td>1.79</td>
<td>/ 0.2</td>
</tr>
<tr>
<td>ERROR B</td>
<td>10</td>
<td>77.33</td>
<td>7.73</td>
<td>&lt; ------</td>
</tr>
</tbody>
</table>

Notice that the SAS printout does not give proper F statistics for the whole plot part of the table (it is possible to use a TEST statement and specify your own error term in SAS but it is just as easy to calculate the proper F by hand).

Now we might want to contrast various means from our experiment. In order to construct proper denominators for the t-tests, we will need to know what kind of model is being assumed for the data. For the moment, take a split plot in a completely randomized design with whole plot factor A at a levels, split plot factor B at b levels, ra whole plot units, and (thus) rab split plot units. The model is written
Assuming fixed effects $A_i$, $B_j$, and $(AB)_{ij}$ we can compute expected mean squares and also the variance of the difference of two means. It turns out that, for most of these differences, the variances are simple functions of the expected mean squares for ERROR A and ERROR B. For a fixed effects model we can compute the expected mean squares with "a" whole plot treatment levels, "b" split plot treatment levels and "r" replications in a CRD as follows:

\[ Y_{ijk} = \mu + A_i + D_{ik} + B_j + (AB)_{ij} + E_{ijk} \]

### ANOVA

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>df</th>
<th>SSq</th>
<th>Mn Sq</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHOLE TRT</td>
<td>a - 1</td>
<td>$\sigma^2 + b \sigma_D^2 + rb \kappa_A^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERROR A</td>
<td>(r-1)a</td>
<td>$E_a$</td>
<td>$\sigma^2 + b \sigma_D^2$</td>
<td></td>
</tr>
<tr>
<td>SPLIT TRT</td>
<td>b -1</td>
<td>$\sigma^2 + ra \kappa_B^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPL*WHL</td>
<td>(a -1)(b -1)</td>
<td>$\sigma^2 + r \kappa_{AB}^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERROR B</td>
<td>(r-1)(a)(b-1)</td>
<td>$E_b$</td>
<td>$\sigma^2$</td>
<td></td>
</tr>
</tbody>
</table>
Steel et al give a more complete listing of ANOVA EMS columns including blocked designs and mixed models. For our purposes, we will just consider the simple table above. Looking ahead we will need to use $E_a$ and $E_b$ (we know what they are estimating) and an estimate of the sum of the $D_{ik}$ variance and the $E_{ijk}$ variance. We see that

$$E\{E_a - E_b\} = b \sigma_D^2$$

so that

$$E\{(E_a - E_b + b E_b)/b\} = \sigma^2 + \sigma_D^2.$$

Calling this mixture of mean squares $E_m$ we have the formula

$$E_m = (E_a + (b - 1)E_b)/b.$$

How do we use $E_a$, $E_b$, and $E_m$? For one thing, they are useful in computing t-statistics to compare means. Remember that a t-statistic has an estimate of some quantity in the numerator and the square root of the variance of that estimate in the denominator. Steel lists the denominators for t-tests on page 381. If you look at these you will see under the square root sign, they all have the form $2^*(\text{variance})/(\text{number of things averaged})$ where “variance” changes depending on whether you are comparing split or whole plot means. In our notation we have these expressions (you will need to take the square root, of course).
Let us try some of these:

(1) Compare soil 1 to soil 3 (note ar = 4 so why have I used 8 as a divisor?)

\[
t = \frac{(109/8) - (128/8)}{\sqrt{2 \times 7.733 / 8}} = -1.71 \ (10 \text{ df})
\]

(2) Compare variety 1 to variety 2

\[
t = \frac{(161/12) - (229/12)}{\sqrt{2 \times 7.733 / 12}} = -4.99 \ (10 \text{ df})
\]
(3) Compare aeration to lack thereof

\[
t = \frac{(240/12) - (150/12)}{\sqrt{2} \times 20.833 / 12} = 4.02 \text{ (2 df)}
\]

(4) * Compute \( E_m \) and then compare aerated to not aerated for variety 1 in soil 1.

\[
E_m = \frac{(20.83 + (6 - 1)7.73)}{6} = 9.92
\]

\[
t = \frac{(10+16)/2 - (8+10)/2}{\sqrt{2} \times 9.92 / 2} = 1.27 \text{ (? df)}
\]

The obvious question is how many degrees of freedom to assign to our mixture \( E_m \) of error mean squares \( E_a \) with \( f_a \) degrees of freedom and \( E_b \) with \( f_b \) degrees of freedom. We can use the Satterthwaite approximation mentioned earlier or:

(a) Look up critical values \( t_a \) with \( f_a \) degrees of freedom and \( t_b \) with \( f_b \) degrees of freedom.

In our case \( f_a = 2 \) and \( f_b = 10 \), so on a two-sided 5% significance level we get \( t_a = 4.303 \) and \( t_b = 2.228 \). Now the appropriate \( t \) critical value (call it \( t' \)) lies between \( t_a \) and \( t_b \) so at this point we know we will not reject \( H_0 \), but clearly there will be instances where we can't stop here.

(b) Compute the approximate critical value \( t' \) as follows:

\[
t' = \frac{(b - 1)E_b t_b + E_a t_a}{(b - 1)E_b + E_a} = \frac{(5)(7.73)(2.228) + (20.83)(4.303)}{(5)(7.73) + (20.83)} = 2.95
\]

and we accept \( H_0 \) at the 5% level.

* The method is more complicated if you average over split plot cells as in variety 1 aerated versus variety 1 not aerated (averaged over soils). \( E_m \) will be different.

D. A. Dickey
EXAMPLE OF SPLIT PLOT IN BLOCKS

3 PAIRS OF IDENTICAL STEERS  3 BLOCKS
2 RATIONS (A, B)  2 WHOLE PLOT TRTS.
2 COOKING METHODS (1, 2)  2 SPLIT PLOT TRTS.

Whole plot units: steers
Split plot units: roasts

Within each pair of steers, one is assigned at random to feed A and one to feed B. After slaughter, two identical roasts are obtained from each steer and the two roasts are randomly assigned to the two cooking methods. Recorded data are weight losses due to cooking.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>RATION</th>
<th>PAIR1</th>
<th>PAIR2</th>
<th>PAIR3</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>11.0</td>
<td>17.0</td>
<td>11.0</td>
<td>39.0</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>2.5</td>
<td>9.0</td>
<td>6.5</td>
<td>18.0</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>5.0</td>
<td>8.0</td>
<td>8.0</td>
<td>21.0</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>3.5</td>
<td>4.0</td>
<td>4.5</td>
<td>12.0</td>
</tr>
</tbody>
</table>

BLOCK TOTALS = 22.0 38.0 30.0 90.0

Mean = 90/12 = 7.5.  CT = 90*90/12 = 675.

SUMS OF SQUARES

Total SS: 862 - CT = 187
Block SS: 22*22/4 + … + 30*30/4 - CT = 32
Rations SS: 57*57/6 + 33*33/6 - CT = 48

Error A SS(computed as Ration*Block) = 93.5 - 48 - 32 = 13.5
<table>
<thead>
<tr>
<th>Ration 1</th>
<th>Pair 2</th>
<th>Pair 3</th>
<th>Ration 2</th>
<th>Pair 2</th>
<th>Pair 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ration A</td>
<td>13.5</td>
<td>26</td>
<td>17.5</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>Ration B</td>
<td>8.5</td>
<td>12</td>
<td>12.5</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>22.0</td>
<td>38</td>
<td>30.0</td>
<td>90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(table SS = 93.5)

Treatment Sums of Squares:

<table>
<thead>
<tr>
<th>Method</th>
<th>SS</th>
<th>F</th>
<th>SM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ration 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>39</td>
<td>18</td>
<td>57</td>
</tr>
<tr>
<td>B</td>
<td>21</td>
<td>12</td>
<td>33</td>
</tr>
<tr>
<td>Totals</td>
<td>60</td>
<td>30</td>
<td>90</td>
</tr>
<tr>
<td>Table SS</td>
<td>135</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ration SS</td>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method SS</td>
<td>75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R*M SS</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ANOVA

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks</td>
<td>2</td>
<td>32</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Rations</td>
<td>1</td>
<td>48</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Error A</td>
<td>2</td>
<td>13.5</td>
<td>6.75</td>
<td></td>
</tr>
<tr>
<td>Methods</td>
<td>1</td>
<td>75</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>R*M</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Error B</td>
<td>4</td>
<td>6.5</td>
<td>1.625</td>
<td></td>
</tr>
</tbody>
</table>

Let us compute expected mean squares for the ANOVA table so that we can decide on appropriate F statistics. First, we write the model using i to index the blocks, j to index the rations, k to index the methods (this choice of subscripts is pretty arbitrary and does not make any real difference).
\[ Y_{ijk} = \mu + \rho_i + R_j + G_{ij} + M_k + (RM)_{jk} + \epsilon_{ijk}. \]

Notice that the whole plot error term \( G_{ij} \) has been indexed as though it were a block \((\rho)\) by whole plot treatment \((R)\) interaction. That is a good way to conceptualize this component of the model. The expected mean squares are given in the Steel & Torrie reference above and are quite like those of a CRD except for the first few rows which involve a blocking component. The expected mean squares are calculated basically using our rules for EMS but omitting (or perhaps I should say renaming) any block*treatment terms (why ?) we get:

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOCKS</td>
<td>( \sigma^2 + 2 \sigma_G^2 + 4 \sigma_P^2 )</td>
</tr>
<tr>
<td>RATIONS</td>
<td>( \sigma^2 + 2 \sigma_G^2 + 6 \kappa_R^2 )</td>
</tr>
<tr>
<td>ERROR A</td>
<td>( \sigma^2 + 2 \sigma_G^2 )</td>
</tr>
<tr>
<td>METHODS</td>
<td>( \sigma^2 + 6 \kappa_M^2 )</td>
</tr>
<tr>
<td>R*M</td>
<td>( \sigma^2 + 3 \kappa_{RM}^2 )</td>
</tr>
<tr>
<td>ERROR B</td>
<td>( \sigma^2 )</td>
</tr>
</tbody>
</table>

D. A. Dickey
SPLIT  –  SPLIT  PLOT (optional)

FOUR MICE – two fed diet A, two fed diet B

F: FEED

2 GLANDS PER MOUSE – come from two specific locations in body.

L: LOCATION

2 TISSUE SAMPLES PER GLAND – analyze 1 immediately, 1 next day

D: DELAY

YIELD–measure of toxin concentration in sample.

DATA:

Feed: Diet I (252)  Diet II (168)

<table>
<thead>
<tr>
<th></th>
<th>Mouse 1</th>
<th>Mouse 2</th>
<th>Mouse 3</th>
<th>Mouse 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(117)</td>
<td>(135)</td>
<td>(100)</td>
<td>(68)</td>
<td></td>
</tr>
</tbody>
</table>

Location of Gland:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 hrs.</td>
<td>35</td>
<td>28</td>
<td>40</td>
<td>35</td>
</tr>
<tr>
<td>24 hrs.</td>
<td>30</td>
<td>24</td>
<td>32</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>52</td>
<td>72</td>
<td>63</td>
</tr>
</tbody>
</table>

TREATMENT TOTALS:

<table>
<thead>
<tr>
<th>Location</th>
<th>Diet I</th>
<th>Diet II</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>75</td>
<td>52</td>
</tr>
<tr>
<td>B</td>
<td>63</td>
<td>41</td>
</tr>
</tbody>
</table>

Delay / 0  \24 hrs.

<table>
<thead>
<tr>
<th></th>
<th>62</th>
<th>52</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>43</td>
<td>32</td>
</tr>
</tbody>
</table>

D. A. Dickey
Total SS = 11878 - CT = 11878 - 11025 = 853
Trt. SS = 75*75/2 + ... + 32*32/2 - 420*420/16 = 675

TABLES:

### FEED*DELAY

<table>
<thead>
<tr>
<th></th>
<th>0 hrs</th>
<th>24 hrs</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet I</td>
<td>138</td>
<td>114</td>
<td>(252)</td>
</tr>
<tr>
<td>Diet II</td>
<td>93</td>
<td>75</td>
<td>(168)</td>
</tr>
<tr>
<td>sum</td>
<td>(231)</td>
<td>(189)</td>
<td></td>
</tr>
</tbody>
</table>

Table SS  553.50  
Delay SS   110.25  
Feed SS    441.00  
F*D SS     2.25

### FEED*LOCATION

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet I</td>
<td>137</td>
<td>115</td>
<td>(252)</td>
</tr>
<tr>
<td>Diet II</td>
<td>95</td>
<td>73</td>
<td>(168)</td>
</tr>
<tr>
<td>sum</td>
<td>(232)</td>
<td>(188)</td>
<td></td>
</tr>
</tbody>
</table>

Table SS  562   
Location SS 121  
Feed SS  441    
F*L SS    0

### DELAY*LOCATION

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 hrs.</td>
<td>127</td>
<td>104</td>
<td>(231)</td>
</tr>
<tr>
<td>24 hrs.</td>
<td>105</td>
<td>84</td>
<td>(189)</td>
</tr>
<tr>
<td>sum</td>
<td>(232)</td>
<td>(188)</td>
<td></td>
</tr>
</tbody>
</table>

Table SS  231.50   
Location SS 21.00
Delay SS 110.25  
L*D SS 0.25

Now compute SS for F*L*D using trt. SS = 675 minus all of the SS calculated above. Thus F*L*D SS = 0.25.

Mouse SS = 117*117/4 + … + 68*68/4 - CT = 609.5(3 df)

Error A SS = Mice(Feeds) = 609.5 - 441 = 168.5 (2 df)

Gland SS = 65*65/2 + … + 29*29/2 - CT = 733 (7 df)

Error B SS = 733 - 441 - 168.5 - 121 (see ANOVA)

ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SSq</th>
<th>Mn Sq</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Feed</td>
<td>1</td>
<td>441.00</td>
<td>441.00</td>
<td>n.s.</td>
</tr>
<tr>
<td>Error A</td>
<td>2</td>
<td>168.50</td>
<td>84.25</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>1</td>
<td>121.00</td>
<td>121.00</td>
<td>*</td>
</tr>
<tr>
<td>L*F</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>n.s.</td>
</tr>
<tr>
<td>Error B</td>
<td>2</td>
<td>2.5</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>Delay</td>
<td>1</td>
<td>110.25</td>
<td>110.25</td>
<td>**</td>
</tr>
<tr>
<td>D*F</td>
<td>1</td>
<td>2.25</td>
<td>2.25</td>
<td>n.s.</td>
</tr>
<tr>
<td>D*L</td>
<td>1</td>
<td>0.25</td>
<td>0.25</td>
<td>n.s.</td>
</tr>
<tr>
<td>D<em>L</em>F</td>
<td>1</td>
<td>0.25</td>
<td>0.25</td>
<td>n.s.</td>
</tr>
<tr>
<td>Error C</td>
<td>4</td>
<td>7.00</td>
<td>1.75</td>
<td></td>
</tr>
</tbody>
</table>

Results: It appears that both diets give about the same toxicity levels. The gland locations are significantly different in their levels of toxin and the delay in analysis seems to affect the measured toxin levels. The effect of delay seems to be consistent across gland locations since there is no significant
interaction. (The effect of a delay is the same for location A as for location B.)

COVARIANCE ANALYSIS

An experiment is run to assess the effects of two fertilizers on yield using a completely randomized design with 8 reps. The experiment is done in a greenhouse by randomly assigning fertilizer A to eight pots and fertilizer B to eight pots.

The yields were the weights of the roots of the plants after several weeks of growth. The yields for fertilizer A were 18, 15, 12, 11, 13, 17, 12, and 16. For fertilizer B the weights were 9, 10, 12, 13, 15, 15, 11, and 9.

It is easy to calculate the ANOVA table or equivalently the two sample t statistic for this data. The difference of the two means is 14.25 -11.75 = 2.5. The MSE = 6.357 and \( t = \frac{2.5}{\sqrt{2*6.357/8}} = 1.98 \) with 14 df. Equivalently, compute

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>Mn Sq</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fertilizer</td>
<td>1</td>
<td>25</td>
<td>25.00</td>
<td>3.93</td>
</tr>
<tr>
<td>Error</td>
<td>14</td>
<td>89</td>
<td>6.38</td>
<td>F(0.05) = 4.60</td>
</tr>
</tbody>
</table>

While harvesting the plants, it was noticed that the pots had become infested with insects. The degree of infestation was rated on a scale from 0 (no infestation) to 10 (high infestation) by estimating the number of insects in the pot.
The data are recorded in a SAS dataset and PROC PRINT is issued.

<table>
<thead>
<tr>
<th>YIELD</th>
<th>INFEST</th>
<th>X1</th>
<th>FERT</th>
<th>FD</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>-5</td>
<td>-5</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>4</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>3</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>2</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>-4</td>
<td>-4</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>5</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>-2</td>
<td>-2</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>-1</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>-5</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>-4</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>-5</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
</tbody>
</table>

The columns X1 and FD will be used in later analysis. The column INFEST is the infestation rating minus the sample mean rating which was 5. This subtraction of the mean is not really necessary. It is convenient since it tells at a glance how far any pot is above/below average infestation. We now run the following SAS step:

```
PROC PLOT; PLOT YIELD*INFEST = FERT;
```

The statements cause the plot symbol to be the value of FERT so that an A in the plot corresponds to fertilizer level 1 and a B corresponds to fertilizer level 2.
We see that the overall picture is two parallel lines. It is fairly obvious that there is a fertilizer effect but we did not pick it up in our ANOVA since the ANOVA model basically just fits two horizontal lines to the data. Any departure from this model is attributed to error variation and we see that this resulting variation is way too large. The ANOVA model is written:

\[ \text{YIELD}_{ij} = \mu + A_i + E_{ij} \]

where \( A_1 \) is fertilizer A effect and \( A_2 \) is fertilizer B effect.

The next step is to write a model which incorporates both the treatment effect and the linear effect of infestation as displayed in the graph. We write

\[ \text{Yield} = \mu + A_i + \beta^*(X_{ij} - \bar{X}) + E_{ij} \]
where $\bar{X}_\cdot$ is the sample mean of the "covariate" (infestation rating in our case). To fit this model, simply input the deviations of the covariate from its sample mean (the column INFEST in our case) and issue these commands:

```
PROC GLM; CLASS FERT;
MODEL YIELD = FERT INFEST/SOLUTION;
```

Notice that the CLASS statement will replace the column of $k$ fertilizer values with $k - 1$ columns of indicator variables. (In our case, $k = 2$ and it is really not necessary to use a class statement except for the fact that FERT is not numeric.)

Note that we do not put the variable INFEST in a class statement. We do not want SAS to replace that one column with a set of columns, we just want a regression coefficient to be computed. Our output contains these items:

| PARAMETER ESTIMATE | T-TEST | P>|T| | STD. ERR |
|--------------------|--------|------|---------|
| INTERCEPT          | 11.5139 B  42.78  0.0001  0.2691 |
| FERT A             | 2.9720 B  7.79   0.0001  0.3816 |
| FERT B             | 0.0000 B  0.00   0.0000  0.0000 |
| INFEST             | -0.6294 -11.89  0.0001  0.0529 |

Because we used a CLASS statement, SAS created a column which replaced the FERT column. This column would be the same as our FD, i.e. it has eight 1's followed by eight 0's. The coefficient on the new column is 2.9720 which means that the overall level for fertilizer A is $11.5139 + 2.9720 = 14.4859$. The level for B is then just $11.5139 + 0 = 11.5139$. Why do we see the B's? These
indicate that other combinations of numbers will fit the data equally well. For example suppose we choose

\[
\begin{align*}
\text{INTERCEPT} & \quad 11.0000 \\
\text{FERT A} & \quad 3.4859 \\
\text{B} & \quad 0.5139 \\
\text{INFEST} & \quad -0.6294
\end{align*}
\]

This gives exactly the same fit as the other parameter estimates. Now what is the effect of the covariate? For either variety, start with the overall level (11.5139 or 14.4859) as calculated above. Now subtract 0.6294 times the infestation rating. Recall that the infestation rating we are using is the original one with 5 subtracted off.

Another summary is to say that the fit consists of two parallel lines with slope $-0.6294$ and the vertical distance between them is 2.9720.

Finally, we see that if $X$ equals $X_{..}$ (i.e. if INFEST=0) then the slope gets multiplied by 0. This illustrates the fact that a covariance analysis simply adjusts all the observations to the levels they would have had if they had each been infested at rate $X_{..}$. Thus the overall levels referred to earlier, 11.5139 and 14.4859, are often called adjusted treatment means.

**TESTING FOR HOMOGENEOUS SLOPES**

It is fairly obvious that the nice interpretations in covariance analysis hinge critically on the assumption that the lines within each treatment level have the same slope. If the slope of the infestation line were different for fertilizer A than for fertilizer B then A and B would not differ by a constant amount (2.9720).
and in fact, B might be better than A for some X values and A better than B for others. Of course, our model forces the fitting of parallel lines but if that is inappropriate then our analysis is meaningless.

How can we check for parallelism? The answer is that we use the “full and reduced model F test” from our multiple regression theory. We will fit a model which allows different slopes at each level of the treatment variable (full model) and compare this to our covariance model with the parallel slopes.

One way to accomplish this fitting is to issue these commands:

```
PROC GLM; CLASS FERT;
MODEL YIELD = INFEST FERT FERT*INFEST;
```

The FERT*INFEST term will give the sum of squares for testing parallelism. Use the type III sum of squares. In our data, we see that FD is a dummy variable for fertilizer (which is what the CLASS statement will produce for us) and X1 is the product of FD times INFEST.

It is a bit easier to understand what is going on if we input our own interaction column X1 in the model. The column X1 serves to estimate interaction and is exactly what PROC GLM does behind the scenes when we call for FERT*INFEST with FERT in a CLASS statement.

We write our model as

\[ \text{Yield} = \mu + A_i + \beta_1(X_{ij} - \bar{X}_{..}) + \beta_2 X_1 + E_{ij} \]

and we see that in fertilizer group 2 (B) the slope is $\beta_1$ since X1 is 0 for this level. If we are in the fertilizer A group then X1 is the same as FERT so in that group we have $(\beta_1 + \beta_2)(X - \bar{X}_{..})$.
and thus the coefficient $\beta_2$ is the difference in slopes of the two regression lines. The t statistic for $\beta_2$ tests the parallelism hypothesis and if t is significant we conclude that the lines are not parallel and the idea of “adjusted treatment means" is meaningless. For more than 2 treatments, use the Type III F test for FERT*INFEST to check for parallel slopes.

It is easy to produce a list of adjusted treatment means along with either confidence intervals or prediction intervals for individuals. Simply append to your data as many lines as you have treatment groups (2 in our case). On these lines set YIELD = . and put in the appropriate levels of the treatments. For the covariate, just put in 0 (or $\bar{X}_.$) so that you are predicting at the mean level of the covariate. SAS will do the computations for you.

Here are the last two lines of our current dataset along with the additional lines for producing adjusted treatment means: we use 0 since we have centered our X's.

11 1 1 B
9 4 4 B
.0 0 0 A
.0 0 0 B

We now issue the following SAS code to perform our analysis:

```
PROC GLM; CLASS FERT;
MODEL YIELD = FERT INFEST/P CLM;
```

The output will contain predictions and confidence intervals for all the row configurations in your X matrix including the last two rows. The predictions in those last two rows will be the so-called “adjusted treatment means." An alternative is to use the
LSMEANS option in PROC GLM. (See the SAS manual for more details.)

data yields; input fert $ @;
  do rep = 1 to 9; input yield infest @; output; end;
cards;
  A  18 0 15 5 12 9 11 8 13 7 17 1 12 10 16 3 .5
  B  9 9 10 8 12 4 13 0 15 1 15 0 11 6 9 9 .5
;
proc means; var infest yield;
proc glm; class fert;
  model yield = fert infest /p;
  lsmeans fert / pdiff;
  means fert;

<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>INFEST</td>
<td>18</td>
<td>5.000</td>
<td>3.4979895</td>
<td>0</td>
<td>10.000</td>
</tr>
<tr>
<td>YIELD</td>
<td>16</td>
<td>13.000</td>
<td>2.7568098</td>
<td>9.000</td>
<td>18.000</td>
</tr>
</tbody>
</table>

General Linear Models Procedure

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>FERT</td>
<td>2</td>
<td>A B</td>
</tr>
</tbody>
</table>

Number of observations in data set = 18

NOTE: Due to missing values, only 16 observations can be used in this analysis.

Dependent Variable: YIELD

<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>
</table>

D. A. Dickey
### Model Summary

<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>FERT</td>
<td>1</td>
<td>25.000000</td>
<td>25.000000</td>
<td>43.38</td>
<td>0.0001</td>
</tr>
<tr>
<td>INFEST</td>
<td>1</td>
<td>81.507898</td>
<td>81.507898</td>
<td>141.43</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

### Source Summary

<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>FERT</td>
<td>1</td>
<td>34.950206</td>
<td>34.950206</td>
<td>60.64</td>
<td>0.0001</td>
</tr>
<tr>
<td>INFEST</td>
<td>1</td>
<td>81.507898</td>
<td>81.507898</td>
<td>141.43</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

### Observation Details

<table>
<thead>
<tr>
<th>Observation</th>
<th>Observed Value</th>
<th>Predicted Value</th>
<th>Residual Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.00000000</td>
<td>17.63304982</td>
<td>0.36695018</td>
</tr>
<tr>
<td>2</td>
<td>15.00000000</td>
<td>14.48602673</td>
<td>0.51397327</td>
</tr>
<tr>
<td>3</td>
<td>12.00000000</td>
<td>11.96840826</td>
<td>0.03159174</td>
</tr>
<tr>
<td>4</td>
<td>11.00000000</td>
<td>12.59781288</td>
<td>-1.59781288</td>
</tr>
<tr>
<td>5</td>
<td>13.00000000</td>
<td>13.22721750</td>
<td>-0.22721750</td>
</tr>
<tr>
<td>6</td>
<td>17.00000000</td>
<td>17.00364520</td>
<td>-0.00364520</td>
</tr>
<tr>
<td>7</td>
<td>12.00000000</td>
<td>11.33900365</td>
<td>0.66099635</td>
</tr>
<tr>
<td>8</td>
<td>16.00000000</td>
<td>15.74483597</td>
<td>0.25516403</td>
</tr>
<tr>
<td>9</td>
<td>*</td>
<td>14.48602673</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9.00000000</td>
<td>8.99635480</td>
<td>0.00364520</td>
</tr>
<tr>
<td>11</td>
<td>10.00000000</td>
<td>9.62575942</td>
<td>0.37424058</td>
</tr>
<tr>
<td>12</td>
<td>12.00000000</td>
<td>12.14337789</td>
<td>-0.14337789</td>
</tr>
<tr>
<td>13</td>
<td>13.00000000</td>
<td>14.66099635</td>
<td>-1.66099635</td>
</tr>
<tr>
<td>14</td>
<td>15.00000000</td>
<td>14.03159174</td>
<td>0.96840826</td>
</tr>
<tr>
<td>15</td>
<td>15.00000000</td>
<td>14.66099635</td>
<td>0.33900365</td>
</tr>
<tr>
<td>16</td>
<td>11.00000000</td>
<td>10.88456865</td>
<td>0.11543135</td>
</tr>
<tr>
<td>17</td>
<td>9.00000000</td>
<td>8.99635480</td>
<td>0.00364520</td>
</tr>
<tr>
<td>18</td>
<td>*</td>
<td>11.51397327</td>
<td></td>
</tr>
</tbody>
</table>

* Observation was not used in this analysis

Summary of Residuals

| Sum of Residuals | 0.00000000 |
| Sum of Squared Residuals | 7.49210207 |
| Sum of Squared Residuals - Error SS | -0.00000000 |

D. A. Dickey
First Order Autocorrelation: -0.04930460
Durbin-Watson D: 2.08063484

| FERT | YIELD   | Pr > |T| | H0:       |
|------|---------|------|----|----------|
|      | LSMEAN  | LSMEAN1=LSMEAN2 |
| A    | 14.4860267 | 0.0001 |
| B    | 11.5139733 |

<table>
<thead>
<tr>
<th>Level of FERT</th>
<th>N</th>
<th>Mean YIELD</th>
<th>SD</th>
<th>Mean INFEST</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>14.2500000</td>
<td>2.60494036</td>
<td>5.37500000</td>
<td>3.73927036</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>11.7500000</td>
<td>2.43486579</td>
<td>4.62500000</td>
<td>3.92564826</td>
</tr>
</tbody>
</table>