A multiple regression model relates a single *response* variable $y$ (dependent variable) to the values of $k$ *regressor* variables $x_1, x_2, \ldots, x_k$ (predictors, independent variables).

A multiple *linear* regression model does so using a linear function of the regressors, with a random error term $\epsilon$:

$$ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon. $$
The model is called *linear* because it is a linear function of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$. However, some $x$’s may be functions of others.

For instance,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$ 

and

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2 + \epsilon.$$ 

are both linear regression models.
Parameter Estimation

Inference

Suppose we have

- $n$ observations of the response, $y_1, y_2, \ldots, y_n$
- corresponding values of the regressors; $x_{i,j}$ is the value of the $j^{th}$ regressor associated with the $i^{th}$ observation.
- Assume that $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$.
- What can we say (infer) about $\beta_0, \beta_1, \ldots, \beta_k$?
Method of *least squares*: the best values of the parameters are the ones that minimize

\[ L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{i,j} \right)^2. \]

\( L \) is a quadratic function of \( \beta_0, \beta_1, \ldots, \beta_k \), so we can find the minimum by equating the gradient to 0.

We obtain \( p = k + 1 \) linear equations (the *normal* equations) in the \( p \) unknowns.
The equations may be written compactly in terms of vectors and matrices:

\[ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{1,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,1} & x_{n,2} & \cdots & x_{n,k} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \]

and

\[ \mathbf{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \]
In terms of these vectors and matrices, the model may be written

\[ y = X\beta + \epsilon, \]

and the normal equations are

\[ X'X\hat{\beta} = X'y. \]
If \( X'X \) is non-singular, and hence has an inverse, the normal equations may be solved to give

\[
\hat{\beta} = (X'X)^{-1} X'y.
\]

If not, the equations still have solutions, but they are not unique.

The *fitted values* and *residuals* are

\[
\hat{y} = X\hat{\beta} \quad \text{and} \quad e = y - \hat{y},
\]

and are unique even when \( \hat{\beta} \) is not.
Estimating $\sigma^2$

The residual sum of squares is

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e = y'y - \hat{\beta}'X'y.$$

We can show that $SS_E$ has $n - p$ degrees of freedom, and

$$E(SS_E) = (n - p)\sigma^2,$$

so that the corresponding mean square

$$\hat{\sigma}^2 = \frac{SS_E}{n - p}$$

is an unbiased estimator of $\sigma^2$. 
Properties of $\hat{\beta}$

Unbiasedness:

$$E(\hat{\beta}) = \beta.$$ 

Variances and covariances:

$$\text{Cov}(\hat{\beta}) = \begin{bmatrix} V(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \ldots & \text{Cov}(\hat{\beta}_0, \hat{\beta}_k) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_0) & V(\hat{\beta}_1) & \ldots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\hat{\beta}_k, \hat{\beta}_0) & \text{Cov}(\hat{\beta}_k, \hat{\beta}_1) & \ldots & V(\hat{\beta}_k) \end{bmatrix} = \sigma^2 (X'X)^{-1}.$$
Example: Viscosity of a polymer

viscosity.txt

<table>
<thead>
<tr>
<th>Temperature</th>
<th>CatalystFeedRate</th>
<th>Viscosity</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>8</td>
<td>2256</td>
</tr>
<tr>
<td>93</td>
<td>9</td>
<td>2340</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>2426</td>
</tr>
<tr>
<td>82</td>
<td>12</td>
<td>2293</td>
</tr>
<tr>
<td>90</td>
<td>11</td>
<td>2330</td>
</tr>
<tr>
<td>99</td>
<td>8</td>
<td>2368</td>
</tr>
<tr>
<td>81</td>
<td>8</td>
<td>2250</td>
</tr>
<tr>
<td>96</td>
<td>10</td>
<td>2409</td>
</tr>
<tr>
<td>94</td>
<td>12</td>
<td>2364</td>
</tr>
<tr>
<td>93</td>
<td>11</td>
<td>2379</td>
</tr>
<tr>
<td>97</td>
<td>13</td>
<td>2440</td>
</tr>
<tr>
<td>95</td>
<td>11</td>
<td>2364</td>
</tr>
<tr>
<td>100</td>
<td>8</td>
<td>2404</td>
</tr>
<tr>
<td>85</td>
<td>12</td>
<td>2317</td>
</tr>
<tr>
<td>86</td>
<td>9</td>
<td>2309</td>
</tr>
<tr>
<td>87</td>
<td>12</td>
<td>2328</td>
</tr>
</tbody>
</table>
R commands

viscosity <- read.table("data/viscosity.txt", header = TRUE)
viscosityLm <- lm(Viscosity ~ Temperature + CatalystFeedRate, viscosity)
summary(viscosityLm)

Output

Call:
lm(formula = Viscosity ~ Temperature + CatalystFeedRate, data = viscosity)

Residuals:
     Min      1Q  Median       3Q      Max
Output, continued

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 1566.0778| 61.5918    | 25.43   | 1.80e-12 *** |
| Temperature    | 7.6213   | 0.6184     | 12.32   | 1.52e-08 *** |
| CatalystFeedRate | 8.5848  | 2.4387     | 3.52    | 0.00376 **  |

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 16.36 on 13 degrees of freedom
Multiple R-Squared: 0.927, Adjusted R-squared: 0.9157
F-statistic: 82.5 on 2 and 13 DF, p-value: 4.1e-08

Fitted model is

\[
\hat{y} = 1566.0778 + 7.6213 x_1 + 8.5848 x_2.
\]
Residual plots

Make four plots of the residuals:

```r
plot(viscosityLm)
```

The first three are the usual (Residuals vs. fitted, Q-Q, and Scale-Location), but the fourth now displays residuals vs. leverage.
Regression Models
Parameter Estimation

lm(Viscosity ~ Temperature + CatalystFeedRate)

Residuals vs Fitted

Fitted values

Im(Viscosity ~ Temperature + CatalystFeedRate)
Theoretical Quantiles
Standardized residuals
lm(Viscosity ~ Temperature + CatalystFeedRate)
lm(Viscosity ~ Temperature + CatalystFeedRate)
lm(Viscosity ~ Temperature + CatalystFeedRate)

Residuals vs Leverage

Leverage
Standardized residuals

Cook's distance

0.5

1

11

6

0

-1

0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35

Leverage

Im(Viscosity ~ Temperature + CatalystFeedRate)
Regression and Factorial Designs

We have used regression to find main effects and interactions in experiments with factorial (full and partial) designs, as an alternative to the hand calculation of effects and the ANOVA table.

If some observations are missing in a factorial design, unbiased estimates of effects can be calculated only using regression methods.
Example

A $2^3$ design with 4 center points (yield-10-2.txt):

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Pressure</th>
<th>Catalyst</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>32</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>46</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>57</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>65</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>68</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>56</td>
</tr>
</tbody>
</table>
R commands

```r
ex10p2 <- read.table("data/yield-10-2.txt", header = TRUE)
summary(lm(Yield ~ Temperature + Pressure + Catalyst, ex10p2))
```

Output

Call:
lm(formula = Yield ~ Temperature + Pressure + Catalyst, data = ex10p2)
Residuals:
  Min     1Q Median     3Q    Max
-7.000e+00 -1.031e+00 -3.483e-15 1.344e+00 5.000e+00

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            51.0000    0.9662  52.783  1.84e-11 ***
Temperature            5.6250    1.1834   4.753  0.00144 **
Pressure               10.6250    1.1834  8.979  1.89e-05 ***
Catalyst               1.1250    1.1834   0.951  0.36961

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
Residual standard error: 3.347 on 8 degrees of freedom
Multiple R-squared: 0.9286, Adjusted R-squared: 0.9019
F-statistic: 34.7 on 3 and 8 DF,  p-value: 6.196e-05
Same example, with run 8 missing

Call:
lm(formula = Yield ~ Temperature + Pressure + Catalyst, data = ex10p2[-8, ])

Residuals:
      Min       1Q     Median       3Q      Max
-7.0577 -1.1635   0.1538   1.5481   4.9423

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    51.058     1.108  46.083  5.92e-10 ***
Temperature     5.712     1.401   4.075  0.004717 **
Pressure       10.712     1.401   7.643  0.000122 ***
Catalyst       1.212     1.401   0.864   0.415959

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 3.573 on 7 degrees of freedom
Multiple R-squared: 0.905, Adjusted R-squared: 0.8643
F-statistic: 22.23 on 3 and 7 DF,  p-value: 0.000592
Regression coefficients, and hence estimated factor effects, changed, but not by much.

But the design is no longer orthogonal ($X'X \neq$ diagonal), and the coefficient estimates are now correlated:

```R
summary(ex10p2Lm, correlation = TRUE)$correlation
```

### Output

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>Temperature</th>
<th>Pressure</th>
<th>Catalyst</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.0000000</td>
<td>0.1581139</td>
<td>0.1581139</td>
<td>0.1581139</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.1581139</td>
<td>1.0000000</td>
<td>0.1875000</td>
<td>0.1875000</td>
</tr>
<tr>
<td>Pressure</td>
<td>0.1581139</td>
<td>0.1875000</td>
<td>1.0000000</td>
<td>0.1875000</td>
</tr>
<tr>
<td>Catalyst</td>
<td>0.1581139</td>
<td>0.1875000</td>
<td>0.1875000</td>
<td>1.0000000</td>
</tr>
</tbody>
</table>
Similarly, if factors cannot be controlled exactly at their nominal design levels, but can be measured, we can use regression to allow for the lack of control.

Suppose the actual levels in the example were:

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Pressure</th>
<th>Catalyst</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.75</td>
<td>-0.95</td>
<td>-1.133</td>
<td>32</td>
</tr>
<tr>
<td>0.90</td>
<td>-1</td>
<td>-1</td>
<td>46</td>
</tr>
<tr>
<td>-0.95</td>
<td>1.1</td>
<td>-1</td>
<td>57</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>65</td>
</tr>
<tr>
<td>-1.10</td>
<td>-1.05</td>
<td>1.4</td>
<td>36</td>
</tr>
<tr>
<td>1.15</td>
<td>-1</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>-0.90</td>
<td>1</td>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>1.25</td>
<td>1.15</td>
<td>1</td>
<td>68</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>44</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>56</td>
</tr>
</tbody>
</table>
R commands

```r
ex10p2modLm <- lm(Yield ~ Temperature + Pressure + Catalyst, ex10p2mod)
summary(ex10p2modLm)
```

Output

```r
lm(formula = Yield ~ Temperature + Pressure + Catalyst, data = ex10p2mod)

Residuals:
    Min     1Q Median     3Q    Max
-6.4939 -0.8745  0.2574  1.6221  5.5061

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)       50.49413   1.03384  48.812  3.43e-11 ***
Temperature      5.41056   1.25334   4.317  0.00256 **
Pressure         10.16315   1.22525   8.293  3.37e-05 ***
Catalyst          1.07207   1.17633   0.911   0.38890

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 3.575 on 8 degrees of freedom
Multiple R-squared: 0.9186, Adjusted R-squared: 0.8881
F-statistic: 30.1 on 3 and 8 DF, p-value: 0.0001044
```
**Correlations of coefficient estimates**

```
summary(ex10p2modLm, correlation = TRUE)$correlation
```

**Output**

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>Temperature</th>
<th>Pressure</th>
<th>Catalyst</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.00000000</td>
<td>-0.06011767</td>
<td>-0.02354036</td>
<td>-0.02726447</td>
</tr>
<tr>
<td>Temperature</td>
<td>-0.06011767</td>
<td>1.00000000</td>
<td>-0.03501220</td>
<td>0.01632200</td>
</tr>
<tr>
<td>Pressure</td>
<td>-0.02354036</td>
<td>-0.03501220</td>
<td>1.00000000</td>
<td>0.03873370</td>
</tr>
<tr>
<td>Catalyst</td>
<td>-0.02726447</td>
<td>0.01632200</td>
<td>0.03873370</td>
<td>1.00000000</td>
</tr>
</tbody>
</table>
Regression ideas can also be used to plan additional runs to de-alias interactions of interest in a fractional factorial design.

Since the new runs will usually need to be treated as a new block, care must be taken to control which interactions are confounded with blocks.