Modern Applied Statistical Models
Topic I: Regression Models for Independent Data

DT9209: MSc Applied Mathematics

Dr Joe Condon

School of Mathematical Sciences
Dublin Institute of technology

©J. Condon 2016

Regression Models

Regression models that look at the relationship between a response variable and one or more predictors.

They are used everywhere in science and economics.

Some examples:

- Life expectancy given medical and lifestyle factors
- Consumer sentiment and interest rates
- Yield from a chemical reaction and amount of catalysts used
- Annual salary and education
- Probability of defaulting on a bank loan based on socio-economic factors
- Response to a drug and dose used
- Many, many more...

A defining feature of the regression model is that the relationship between the response and the predictors has a stochastic (random) element - i.e. the relationship is not fully deterministic.

We initially look at a sub-class of regression models - linear models.

A regression model is applied to datasets consisting of a number of variables (response and predictors) measured against an experimental (or sample) unit. This collection of variables recorded on an experimental unit is called an ‘observation’.

Example: LDL Data

i	Weight (kg)	Cholestoral (mg/dL)
1	100	160
2	105	150
3	90	120
4	80	90
5	80	110
6	85	130
7	87	110
8	92	140
9	90	130
10	95	140
11	93	120
12	85	120
13	85	110
14	70	100
15	85	100

We use the following convention:
Each observation is given an index i,
i =1=first observation,
i =2=second observation, etc.
Each observation has a response variable y_i. In this case y_i is the LDL level for each child

Each observation has at least one predictor variable. These are denoted $x_{i1}, x_{i2}, x_{i3}, \ldots$

If there is is only one predictor variable, then we denote it as x_i.

Example: Dose-response data

An experiment was performed to test the effect of a steroid on the activity levels of rats. Eleven rats were used in the study, different doses of the steroid were administered and their activity levels were recorded over a 4 hour period.

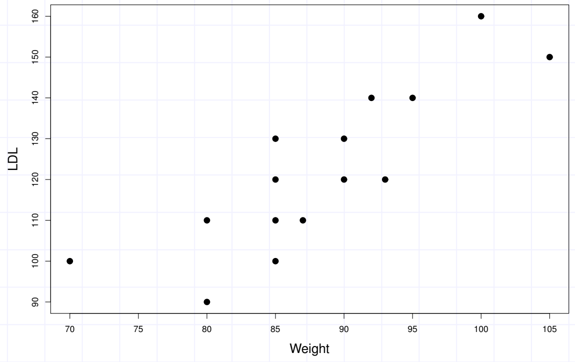
Dose	Activity
10.45	991
12.57	1233
15.62	1229
25.98	1684
30.52	1862
34.06	1919
41.17	2082
50.78	1776
61.01	1528
71.76	881
79.20	1101

Example: Bread-wrapper data

Seal Strength	Sealing Temp.	% polyethylene
6.6	225	0.5
6.9	285	0.5
7.9	225	0.5
6.1	285	0.5
9.2	225	1.7
6.8	285	1.7
10.4	225	1.7
7.3	285	1.7
9.8	204.5	1.1
5	305.5	1.1
6.9	255	1.1
6.3	255	1.1
4	255	0.09
8.6	255	2.11
10.1	255	1.1
9.9	255	1.1
12.2	255	1.1
9.7	255	1.1
9.7	255	1.1
9.6	255	1.1

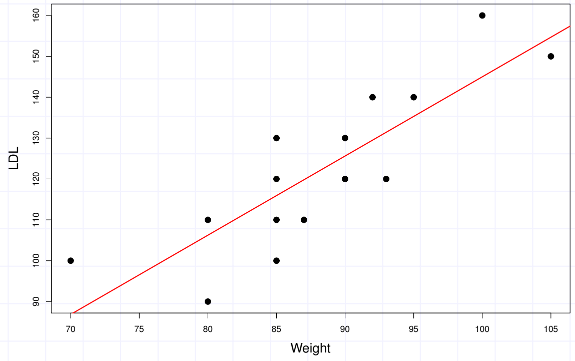
Motivating Examples

LDL Data: The relationship between the weigh of obese children and LDL ('bad') cholesterol.



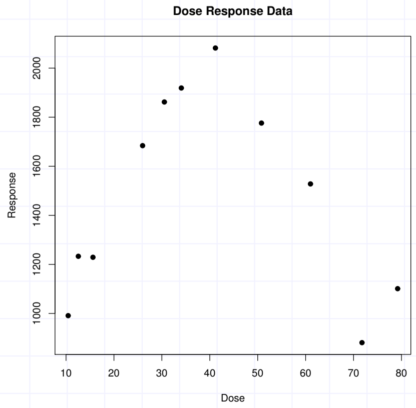
A straight line model seems appropriate here, i.e.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



This is called a 'simple linear regression model'.

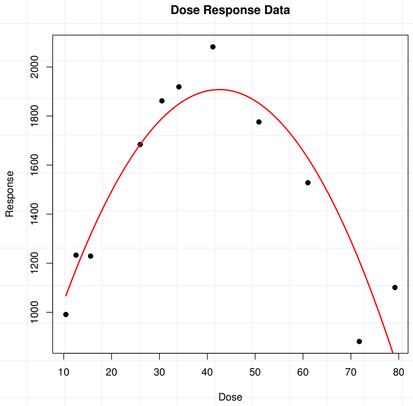
Dose-response Data: The relationship between the dose of a steroid given to rats and their activity levels.



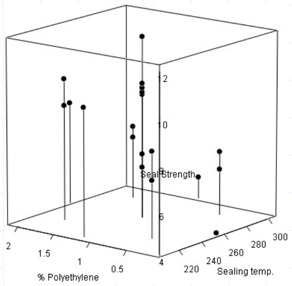
A quadratic model may be appropriate here, i.e.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

where y_i is the activity level for rat i and x_i is the dose of steroid they received.



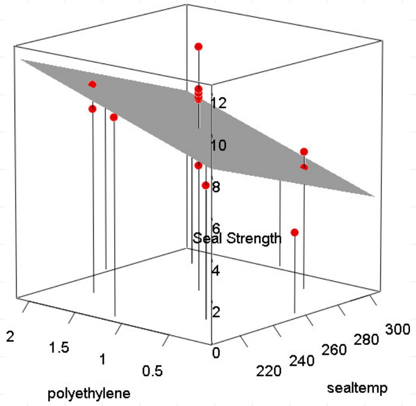
Example: The Bread-Wrapper Data: The relationship between the strength of the wrapping on a loaf of bread, and the sealing temperature t which the stock (i.e. glue) was applied and the % polyethylene in the stock.



A planar response surface model may be appropriate here, i.e.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where y_i is the strength for wrapper i, x_{i1} is the sealing temperature and x_{i2} is the % polyethylene used in the stock.



For the three models considered here so far, i.e.,

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \\ y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \end{aligned}$$

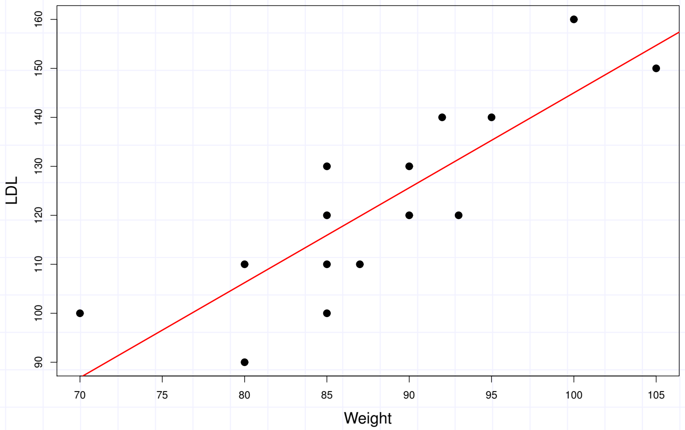
NB. The unknown regression parameters, the β's, enter each of the models as linear coefficients.

The Simple Linear Regression Model

LDL Data

- A medical researcher is investigating the relationship between weight in obese children and LDL Cholesterol in their blood stream.
- A central question: does increasing weight predict an increase in LDL Cholesterol levels in obese children?
- They randomly recruit 15 obese children and measure (i) their weight (kg) and (ii) their blood LDL level (mg/dL).
- They the draw a scatterplot of the results.

The data are shown in the plots below.



The line is called the 'regression line'.

- We are modelling the relationship as a straight line.
- What straight line do we choose from the infinity of lines available?
- There are numerous choice, but for linear statistical models we choose a particular mathematical criterion for finding the line, called the criterion of **least squares**.
- The criterion of least squares has some 'nice' statistical properties - it is also very mathematically convenient.

The model for the straight line is:

$$\text{LDL}= \beta_0 + \beta_1(\text{weight}) + \varepsilon.$$

Where
β_0 is the intercept
β_1 is the slope
ε is the error term - i.e. the distance of the observed data point to the line.

We say that we are *regressing* LDL on Weight.

More generally, letting y = LDL and x = Weight, we are regressing y on x.

So the general simple linear regression model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1)$$

NB. y_i is also called the response or dependent variable, and x_i the predictor or independent variable.

In all that follows we assume the following:

1. Both x_i and y_i variables are on the continuous level of measurement.
2. The x_i variables are not random variables.
3. The y_i variables are random variables - consisting of both a non-random (deterministic/structural) and random components.

Least Squares Criterion

The idea is to choose a line that is simultaneously close to all points by finding the line that minimises the sum of squared vertical distances from the observed data points to the line.

Mathematically this is: measure the squared vertical distances from the points to the line - so need to add up this distance over all points. For a given point y_i we get:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Squaring and writing in terms of the distance we get,

$$\varepsilon_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2$$

and finally summing over the n data points we get an overall measure of squared distance, i.e. the Sum of Squares

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (2)$$

The task now is to find the values of β_0, β_1 such that the objective function Q in equation (2) is minimised.

To do this solve the following equations simultaneously.

$$(1)\ \frac{\partial Q}{\partial \beta_0} = 0 \qquad (2)\ \frac{\partial Q}{\partial \beta_1} = 0$$

The solutions for the two parameters (i.e. β's) are,

$$\hat{\beta}_1 = S_{xy}/S_{xx} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xy} = \sum_{i=1}^{n} xy - \frac{\sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n} \qquad S_{xx} = \sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n}$$

The hats are used to indicate that these are estimates (in fact MLEs - more on this later).

We can see know why this is mathematically convenient - taking derivatives is easy - wouldn't be the case, if say, absolute values had been used.

Since we have minimised the squared distance, this method is called the Method of Least Squares or simply Least Squares (LS). It can be shown that under fairly general conditions these estimates are the unique minimisers of (2).

Solving for the LS estimates for the LDL data we get,
$\hat{\beta}_0 = -48.781$,
$\hat{\beta}_1 = 1.938$.

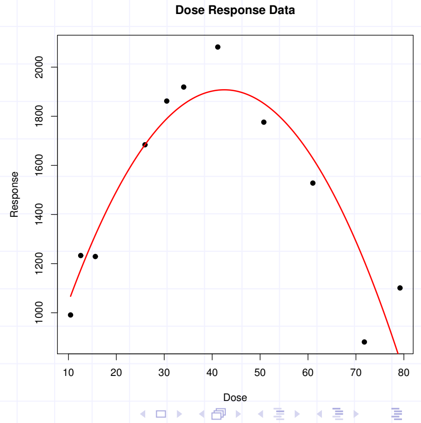
[You should verify these results as an exercise]

Multiple Regression

Dose-Response Data

An experiment was performed to test the effect of a steroid on the activity levels of rats. Eleven rats were used in the study, different doses of the steroid were administered and their activity levels were recorded over a 4 hour period.

Dose	Activity
10.45	991
12.57	1233
15.62	1229
25.98	1684
30.52	1862
34.06	1919
41.17	2082
50.78	1776
61.01	1528
71.76	881
79.20	1101



The quadratic model is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

where y_i is the activity level for rat i and x_i is the dose of steroid they received.

How do we find the estimates for β_0, β_1 and β_2?

We use the LS method again.

Define the objective function:

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

NB. This is a function of three unknowns.

Take the partial derivative WRT each unknown and set each equal to zero. Then solve the three equations simultaneously.

$$(1)\ \frac{\partial Q}{\partial \beta_0} = 0 \quad \Rightarrow \quad \beta_0 = \frac{1}{n}\left[\sum_i y_i - \beta_1 \sum_i x_i - \beta_2 \sum_i x_i^2\right]$$

$$(2)\ \frac{\partial Q}{\partial \beta_1} = 0 \quad \Rightarrow \quad \beta_1 = \frac{1}{\sum_i x_i^2}\left[\sum_i x_i y_i - \beta_0 \sum_i x_i - \beta_2 \sum_i x_i^3\right]$$

$$(3)\ \frac{\partial Q}{\partial \beta_2} = 0 \quad \Rightarrow \quad \beta_2 = \frac{1}{\sum_i x_i^4}\left[\sum_i x_i^2 y_i - \beta_0 \sum_i x_i^2 - \beta_1 \sum_i x_i^3\right]$$

This is entirely feasible - but we can be more mathematically efficient.

[Again, these results should be verified as as exercise.]

Matrix Formulation of LS Model

We could re-formulate the LS model using matrix algebra.

There are 11 rats in the experiment, so a matrix formulation of the model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

is:

$$\overbrace{\begin{pmatrix} y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\\ y_8\\ y_9\\ y_{10}\\ y_{11}\end{pmatrix}}^{Y} = \overbrace{\begin{pmatrix} 1 & x_1 & x_1^2\\ 1 & x_2 & x_2^2\\ 1 & x_3 & x_3^2\\ 1 & x_4 & x_4^2\\ 1 & x_5 & x_5^2\\ 1 & x_6 & x_6^2\\ 1 & x_7 & x_7^2\\ 1 & x_8 & x_8^2\\ 1 & x_9 & x_9^2\\ 1 & x_{10} & x_{10}^2\\ 1 & x_{11} & x_{11}^2\end{pmatrix}}^{X} \overbrace{\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\\ \varepsilon_5\\ \varepsilon_6\\ \varepsilon_7\\ \varepsilon_8\\ \varepsilon_9\\ \varepsilon_{10}\\ \varepsilon_{11}\end{pmatrix}}^{\varepsilon}$$

Using the actual data we get:

$$\overbrace{\begin{pmatrix} 991\\ 1233\\ 1229\\ 1684\\ 1862\\ 1919\\ 2082\\ 1776\\ 1528\\ 881\\ 1101\end{pmatrix}}^{Y} = \overbrace{\begin{pmatrix} 1 & 10.45 & 109.20\\ 1 & 12.57 & 158.00\\ 1 & 15.62 & 243.98\\ 1 & 25.98 & 674.96\\ 1 & 30.52 & 931.47\\ 1 & 34.06 & 1160.08\\ 1 & 41.17 & 1694.97\\ 1 & 50.78 & 2578.61\\ 1 & 61.01 & 3722.22\\ 1 & 71.76 & 5149.50\\ 1 & 79.20 & 6272.64\end{pmatrix}}^{X} \overbrace{\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\\ \varepsilon_5\\ \varepsilon_6\\ \varepsilon_7\\ \varepsilon_8\\ \varepsilon_9\\ \varepsilon_{10}\\ \varepsilon_{11}\end{pmatrix}}^{\varepsilon}$$

and with the vectors $Y,\ \beta,\ \varepsilon$ and the matrix X so defined, we can write very succinctly:

$$Y = X\beta + \varepsilon \tag{3}$$

Using these, and being careful about our matrix/vector dimensions and multiplication we get:

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) \tag{4}$$

NB: using A' to indicate the transpose of A. Also, I'm using uppercase for both vectors and matrices.

Now, we need to minimise a scalar Q with respect to the elements of the vector β.

Derivative of Q WRT to (elements of) a vector

- In reality what we are doing is taking the partial derivatives WRT to each element in the vector and stacking them in a column vector.
- The specialised notation used here consists of the following: (a) treat the vector as though it were a scalar and apply the usual rules of taking derivatives, but (b) keep a careful account of the dimensions of all the vectors and matrices in the resulting expressions to ensure they are correct/comfortable.

$$\begin{aligned} Q &= (Y - X\beta)'(Y - X\beta)\\ &\Rightarrow \frac{\partial Q}{\partial \beta} = -2X'Y + 2X'X\beta = 0\\ &\Rightarrow \hat{\beta} = (X'X)^{-1}X'Y \end{aligned}$$

[proof?]

Applying the formula to the dose-response data we get the following:

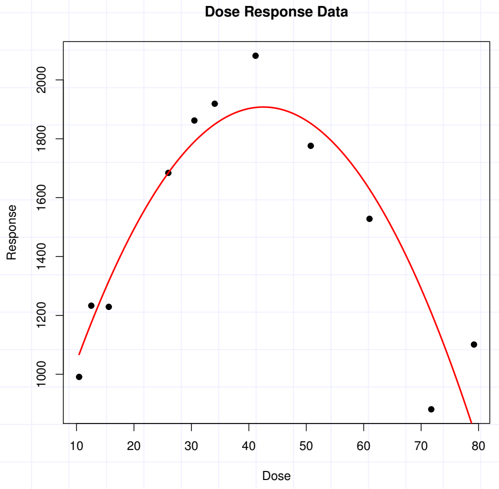
$$\overbrace{\begin{pmatrix}1.2632141 & -0.0639283 & 0.0006518\\ -0.0639284 & 0.0037908 & -0.0000414\\ 0.0006518 & -0.0000414 & 0.0000005\end{pmatrix}}^{(X'X)^{-1}}\overbrace{\begin{pmatrix}16286\\ 630536\\ 30939083\end{pmatrix}}^{X'Y} = \overbrace{\begin{pmatrix}430.0754\\ 69.5021\\ -0.8172\end{pmatrix}}^{\hat{\beta}}$$

So, the LS equation for the quadratic model to these data is:

$$y = 430.0754 + 69.5021x - 0.8172x^2$$

[You should verify these results as an exercise.]

If you plot this function over the scatterplot of the data - then you get the following curved line.



What happens if you decide, that a cubic model might be better than quadratic model for the dose-response data?

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

Now we have four parameters to estimate. Define the following:

$$\overbrace{\begin{pmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7\\ y_8\\ y_9\\ y_{10}\\ y_{11}\end{pmatrix}}^{Y} = \overbrace{\begin{pmatrix}1 & x_1 & x_1^2 & x_1^3\\ 1 & x_2 & x_2^2 & x_2^3\\ 1 & x_3 & x_3^2 & x_3^3\\ 1 & x_4 & x_4^2 & x_4^3\\ 1 & x_5 & x_5^2 & x_5^3\\ 1 & x_6 & x_6^2 & x_6^3\\ 1 & x_7 & x_7^2 & x_7^3\\ 1 & x_8 & x_8^2 & x_8^3\\ 1 & x_9 & x_9^2 & x_9^3\\ 1 & x_{10} & x_{10}^2 & x_{10}^3\\ 1 & x_{11} & x_{11}^2 & x_{11}^3\end{pmatrix}}^{X}\overbrace{\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\\ \varepsilon_5\\ \varepsilon_6\\ \varepsilon_7\\ \varepsilon_8\\ \varepsilon_9\\ \varepsilon_{10}\\ \varepsilon_{11}\end{pmatrix}}^{\varepsilon}$$

Notice that the equations can be written exactly the same as before, i.e.:

$$Y = X\beta + \varepsilon$$

$$\begin{aligned}\Rightarrow Q &= \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)\\ \Rightarrow \frac{\partial Q}{\partial \beta} &= -2X'Y + 2X'X\beta = 0\\ \Rightarrow \hat{\beta} &= (X'X)^{-1}X'Y\end{aligned}$$

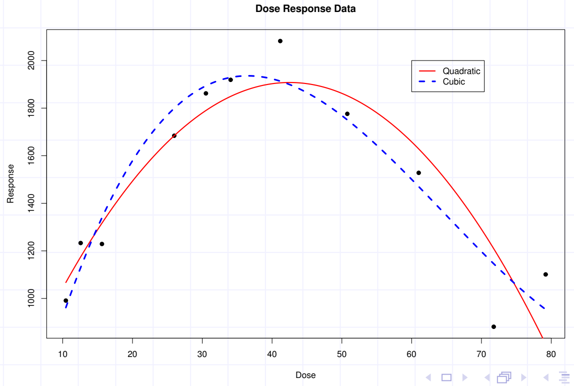
The only thing that has changed are the dimensions of the X matrix and the β vector (one extra column and row respectively).

The model from these specifications for the X matrix and β vector is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

Performing the matrix algebra we get:

$$\hat{\beta} = \begin{pmatrix} -157.87 \\ 132.06 \\ -2.52 \\ 0.013 \end{pmatrix}$$



Bread-wrapper Data

We can use exactly the same matrix algebra in fitting a model to the Bread-wrapper data.

Seal Strength	Sealing Temp.	% polyethylene
6.6	225	0.5
6.9	285	0.5
7.9	225	0.5
6.1	285	0.5
9.2	225	1.7
6.8	285	1.7
10.4	225	1.7
7.3	285	1.7
9.8	204.5	1.1
5	305.5	1.1
6.9	255	1.1
6.3	255	1.1
4	255	0.09
8.6	255	2.11
10.1	255	1.1
9.9	255	1.1
12.2	255	1.1
9.7	255	1.1
9.7	255	1.1
9.6	255	1.1

The model we are fitting is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

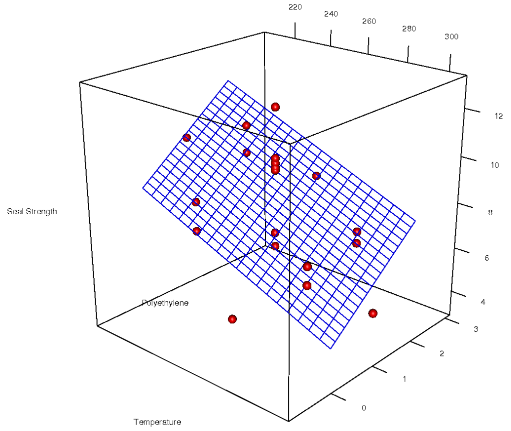
where x_{i1} is the sealing temperature for observation i and x_{i2} is the corresponding % of polyethylene used.

This is a plane in 3 dimensions.

$$\overbrace{\begin{pmatrix} 6.6 \\ 6.9 \\ 7.9 \\ 6.1 \\ 9.2 \\ 6.8 \\ 10.4 \\ 7.3 \\ 9.8 \\ 5 \\ 6.9 \\ 6.3 \\ 4 \\ 8.6 \\ 10.1 \\ 9.9 \\ 12.2 \\ 9.7 \\ 9.7 \\ 9.6 \end{pmatrix}}^{Y} = \overbrace{\begin{pmatrix} 1 & 225 & 0.5 \\ 1 & 285 & 0.5 \\ 1 & 225 & 0.5 \\ 1 & 285 & 0.5 \\ 1 & 225 & 1.7 \\ 1 & 285 & 1.7 \\ 1 & 225 & 1.7 \\ 1 & 285 & 1.7 \\ 1 & 204.5 & 1.1 \\ 1 & 305.5 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 0.09 \\ 1 & 255 & 2.11 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \\ 1 & 255 & 1.1 \end{pmatrix}}^{X} \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{17} \\ \varepsilon_{18} \\ \varepsilon_{19} \\ \varepsilon_{20} \end{pmatrix}}^{\varepsilon}$$

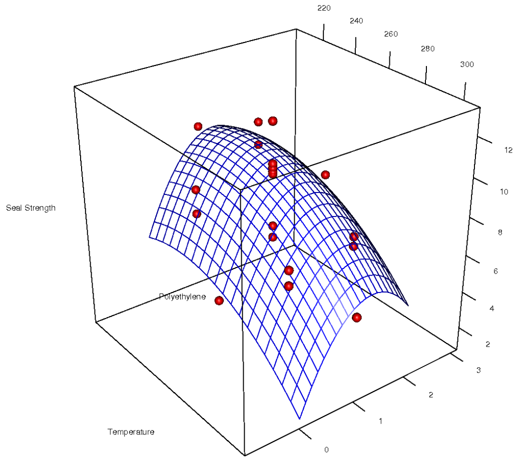
Once again we get the solution using the LS criterion:

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 15.658 \\ -0.037 \\ 1.700 \end{pmatrix}$$



We could also fit a quadratic surface in 3 dimensions:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \varepsilon_i$$



Regression & Statistical Inference

- Fitting Regression Models is typically only the first step in data analysis using regression.
- Statistical Hypothesis testing, Confidence Interval and Prediction Interval estimation is the second step.
- To do this second step we need to collect some statistical results.

Statistical Assumptions

We start with the following:

$$\begin{aligned} y_i &\sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \ldots, \ \sigma^2) \\ &\sim N(x_i'\beta, \ \sigma^2) \end{aligned}$$

NB. That x_i' is a row vector - it is the row of the X matrix for observation i.

This immediately implies the following:

$$\varepsilon_i \sim N(0, \sigma^2)$$

We also assume that the responses are uncorrelated (in fact we will assume they are statistically independent later on!), i.e.:

$$Cov[y_i, \ y_j] = 0 \text{ for } i \neq j \qquad Cov[y_i, \ y_i] = \sigma^2$$

$$Cov[\varepsilon_i, \ \varepsilon_j] = 0 \text{ for } i \neq j \qquad Cov[\varepsilon_i, \ \varepsilon_i] = \sigma^2$$

Properties of the LS estimator of β

Let Y be a random columnar vector with random element Y_i, $i = 1..n$.

$$E[Y] = E\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix}$$

$$Var[Y] = \begin{pmatrix} Var(Y_1) & Cov(Y_1, Y_2) & \ldots & Cov(Y_1, Y_n) \\ Cov(Y_2, Y_1) & Var(Y_2) & \ldots & Cov(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_n, Y_1) & Cov(Y_n, Y_2) & \ldots & Var(Y_n) \end{pmatrix}$$

$$E[AY] = AE[Y] \qquad Var[AY] = AVar[Y]A'$$

for a matrix/vector of constants A.

This means we can write our the linear regression assumptions in the following way:

$$Y \sim N(X\beta,\ I_n\sigma^2) \qquad \rightarrow \qquad \varepsilon \sim N(0, I_n\sigma^2)$$

Using these results we can get the following very easily:

$$\begin{aligned} E[\hat{\beta}] &= \beta \\ Var[\hat{\beta}] &= (X'X)^{-1}\sigma^2 \\ \rightarrow \hat{\beta} &\sim N\left(\beta,\ (X'X)^{-1}\sigma^2\right) \end{aligned}$$

[Proof?]

Estimating σ^2

We are going to need a 'good' estimate of σ^2 to proceed to hypothesis testing, CIs etc.

We have the following from the model:

$$\begin{aligned} \varepsilon &= Y - X\beta \\ Var[\varepsilon] &= \sigma^2 I_n \end{aligned}$$

We can 'estimate' ε using the fitted model:

$$e = Y - X\hat{\beta}$$

where e is called the vector of residuals with individual elements e_i given by:

$$e_i = y_i - x_i'\hat{\beta} = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots)$$

A 'sample' variance of these residuals can be used to get an estimate of σ^2.

First, square them and add them up, i.e.

$$e'e = \varepsilon(I - H)\varepsilon$$

where $H = X(X'X)^{-1}X'$ is called the Hat matrix.

The term $e'e$ is called the Sum of Squares (SS) for error (or residual).

NB. H is an $n \times\ n$ symmetric idempotent matrix of constants.[Proof?]

$\varepsilon'(I - H)\varepsilon$ is an example of a **random variable quadratic form**.

What we need to do is find the statistical properties of $\varepsilon'(I - H)\varepsilon$.

Quadratic forms: A quadratic form is a scalar expression of the form $z'Az$ where A is an $(n \times n)$ symmetric matrix and z is an $(n \times 1)$ vector. Expanding the algebra out we get,

$$z'Az = \sum_{i=1}^{n}\sum_{j=1}^{n} z_i a_{ij} z_j$$

For example the quadratic form $z_1^2 + 2z_1z_2 - 3z_2^2$ can be written as $z'Az$ where,

$$\begin{pmatrix} 1 & 1 \\ 1 & -3 \end{pmatrix} \text{ and } \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

Quadratic forms of independent random variables:

Let Z be an $n \times 1$ vector of **independent** random variables with mean vector μ, and variance-covariance matrix V.

If C is a symmetric $n \times n$ matrix of constants then we get:

$$\begin{aligned} Z'CZ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(z_j - \mu_j)(z_i - \mu_i) + \mu'CZ + Z'C\mu - \mu'C\mu \\ E[Z'CZ] &= trace(CV) + \mu'C\mu \end{aligned}$$

[Proof?]

NB: $trace(ABC) = trace(BCA) = trace(CAB)$

i.e. the trace of a matrix product is invariant to cyclical permutations of that product.

Now we get:

$$E[e'e] = E[\varepsilon'(I - H)\varepsilon] = (n-p)\sigma^2$$

where p is the number of parameters in the model, i.e. the number of columns in the X matrix. [Proof?]

Therefore, an unbiased estimate of σ^2 is given by:

$$s^2 = \frac{e'e}{n-p} = \frac{1}{n-p}Y'(I-H)Y = \frac{\sum_i (y_i - x_i'\hat{\beta})^2}{n-p}$$

Hypothesis Testing using the Matrix formulation

- We would like to be able to conduct hypothesis tests on the parameters in our regression model.
- These tests can be performed both on (a) individual parameters and (b) jointly on multiple parameters at the same time.
- We would also like to calculate confidence and prediction intervals as well.

Example: Recall the first model for the Bread Wrapper data:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where x_{i1} was the sealing temperature and x_{i2} was the % polyethylene used.

Is there any evidence from this model and data that temperature is a genuine predictor of the sealing strength?

We might consider the following null and alternative hypotheses:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_A : \beta_1 &\neq 0 \end{aligned}$$

The general framework for doing this that we will discuss is the general linear hypothesis.

General Linear Hypotheses

- General linear hypotheses are a very general way a testing many hypotheses from regression models.
- The tests we have looked at above can be approached this way.

The approach involves specifying an hypothesis of the form:

$$H_0 : \ L\beta = 0$$

where L is a $1 \times p$ vector or $r \times p$ matrix (full row rank) of coefficients defining an hypothesis of interest.

NB. The RHS need not be the zero vector, a more general $r \times 1$ vector $d \neq 0$ may be specified on the RHS instead.

For example: we could specify the following for the Bread Wrapper data:

$$H_0 : \quad \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = (0)$$

This is equivalent to testing

$$H_0 : \beta_1 = 0$$

which is testing if there is statistically significant evidence from the model that temperature is related to seal strength.

Now we specify the following:

$$\begin{aligned} \hat{\beta} &\sim N\left(\beta, \sigma^2 (X'X)^{-1}\right) \\ \Rightarrow L\hat{\beta} &\sim N\left(L\beta, \sigma^2 L (X'X)^{-1} L'\right) \end{aligned}$$

Define the following Sum of Squares (SS):

$$SS = (L\hat{\beta})'(L(X'X)^{-1}L')^{-1}(L\hat{\beta})$$

This is a random variable quadratic form with expected value $r\sigma^2$ under the null hypothesis.

[Proof?]

From our proof, this result holds for all hypotheses of the form:

$$H_0 : \ L\beta = 0$$

Use of Cochran's Theorem

The theorem proves the following:

1. Regardless of any null hypothesis specified:

$$e'e/\sigma^2 \sim \chi^2_{n-p}$$

2. That under the null hypothesis $L\beta = 0$:

$$(L\hat{\beta})'(L(X'X)^{-1}L')^{-1}(L\hat{\beta})/\sigma^2 \sim \chi^2_r$$

3. That these two random variables are independent of each other.

This implies that if the null hypothesis $L\beta = 0$ is true, the ratio:

$$\frac{(L\hat{\beta})'(L(X'X)^{-1}L')^{-1}(L\hat{\beta})}{rs^2} \sim F(r, n-p)$$

Example: we specify the following for the Bread Wrapper data:

$$H_0 : \quad \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = (0)$$

This is equivalent to testing

$$H_0 : \beta_1 = 0$$

1. Calculate the SS for this choice of null hypothesis using:

$$SS = (L\hat{\beta})'(L(X'X)^{-1}L')^{-1}(L\hat{\beta}) = 16.64$$

   where X is the regression matrix for the model given above, L is a specified in the null hypothesis and $\hat{\beta} = (X'X)^{-1}X'Y$.

2. Calculate the estimate $s^2 = e'e/(n-p) = 3.018$.

3. Calculate the F test statistic:

$$F = \frac{SS}{rs^2} = \frac{16.64}{3.018} = 5.51.$$

4. Calculate the p-value by integrating the $F(r, n-p)$ PDF over the domain $[F, \infty]$

$$\int_F^\infty f_{(r,n-p)}(x)dx = 0.03$$

   where $f_{(r,n-p)}(x)$ is the F density function with r and $n-p$ degrees of freedom - in this case 1 and 17.
   Alternatively, set a rejection region at a pre-specified α level - often the 0.05 (5%) level.

5. Reject or fail to reject the null hypothesis and state your conclusions.
   What are they in this particular example?

F Density function

The F probability density function with r and $n-p$ degrees of the freedom is:

$$f_{(r,n-p)}(x) = \frac{\Gamma((r+n-p)/2)}{\Gamma(r/2)\Gamma([n-p]/2)} \frac{(r/[n-p])^{r/2}x^{(r/2)-1}}{(1+(rx/[n-p]))^{(r+n-p)/2}}$$

where $\Gamma(z) = \int_0^\infty e^{-u}u^{z-1}\,du$

[there is no need to commit this density function to memory!]

Examples: Hypotheses for the Bread Wrapper data

What are the null hypotheses and conclusions in each case?

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \Rightarrow F = 150.13$$

$$L_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \Rightarrow F = 5.11$$

$$L_3 = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \Rightarrow F = 14.23$$

$$L_4 = \begin{pmatrix} 0 & 1 & -2 \end{pmatrix} \Rightarrow F = 4.82$$

Certain hypotheses are of particular relevance in nearly all regression situations - and therefore they typically form part of the default output from software packages such as SAS, R, SPSS, etc.

SS regression is the sum of squares associated with the following null hypothesis:

$$\begin{aligned} H_0: \quad & \beta_0 = \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ & \Rightarrow L = I_p \end{aligned}$$

SS model is the sum of squares associated with the following null hypothesis:

$$\begin{aligned} H_0: \quad & \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ & \Rightarrow L = \begin{pmatrix} 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 1 & \\ & \vdots & & & \ddots \end{pmatrix} \end{aligned}$$

It is relatively easy to show that SS regression and SS error represent a decomposition of the total uncorrected sum of squares for the response variables, i.e.:

$$Y'Y = \text{SS regression} + \text{SS error}$$

and that SS model and SS error represent a decomposition of the total corrected sum of squares for the response variables,i.e.:

$$(Y - \bar{Y})'(Y - \bar{Y}) = \text{SS model} + \text{SS error}$$

Also, that SS regression $-$ SS model $= n\bar{y}^2$.

This is sometimes called the correction factor and represents the SS regression that would be given by fitting the intercept on its own in a model. [Proof?]

ANOVA Table

The hypothesis test associated with SS model is always always of more interest than SS regression - and that is what is produced by R by default.

Results for the LS plane fitted to the Breadwrapper data.

```
summary(fit3)
Call:
lm(formula = y ~ x1 + x2)...

Residual standard error: 1.737 on 17 degrees of freedom
Multiple R-squared:  0.3756,  Adjusted R-squared:  0.3022
F-statistic: 5.113 on 2 and 17 DF,  p-value: 0.01825
```

How do we interpret this p value?

How do we interpret p-values generally?

How would the rejection region approach have been applied here?

The results for the LDL data and the Dose-Response data are given below. In each case state the null hypothesis, the alternative and use the NCST to state your conclusions concerning these hypotheses.

```
> summary(fit1)

Call:
lm(formula = y ~ x)...

Residual standard error: 11.13 on 13 degrees of freedom
Multiple R-squared:  0.7039,  Adjusted R-squared:  0.6811
F-statistic:  30.9 on 1 and 13 DF,  p-value: 9.248e-05
```

```
summary(fit2)...

Call:
lm(formula = y ~ x + x2)....

Residual standard error: 182.7 on 8 degrees of freedom
Multiple R-squared:  0.8428,  Adjusted R-squared:  0.8035
F-statistic: 21.45 on 2 and 8 DF,  p-value: 0.0006106
```

T-tests for a single regression parameter

Recall that a t-test statistic is applied to the mean(s) of normally distributed data.

The simplest form of this test statistic takes the form of:

$$\frac{\bar{x} - \mu}{\sqrt{s_{\bar{x}}^2}}$$

where $\bar{x}$ is the sample mean, μ is the hypothesised ‘true’ population mean and $s_{\bar{x}}^2$ is the estimate of variance of the mean from the same sample.

In the case of hypotheses regarding a single regression parameter, β_j we could use the following:

$$\hat{\beta}_j \sim N(\beta, \sigma_{\beta_j}^2)$$

Now, specify a null hypothesis, e.g.:

$$H_0 : \beta_j = 0$$

Calculate the following test statistic:

$$t = \frac{\hat{\beta}_j - 0}{\sqrt{s^2 (X'X)_{jj}^{-1}}} \sim t_{n-p}$$

Calculate the p-value as:

$$2 \int_{|t|}^{\infty} f(x)dx$$

where $f(x)$ is the t PDF with $n - p$ degrees of freedom.

An F test statistic with numerator degrees of freedom $= 1$, and denominator degrees of freedom $= n - p$, is the square of a t test statistic with the same null hypothesis and with degree of freedom $n - p$.

Therefore, both tests will give the same p-value and lead to the same conclusion - i.e. they are equivalent tests.

Example (bread-wrapper data):

$$L = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \Rightarrow F = 5.51$$

Try the following:

$$t = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} \sim t_{n-p}$$

The s.e. is the square root of the variance given by the appropriate element in:

$$s^2(X'X)^{-1} = \begin{pmatrix} 16.84742053 & -6.256595e-02 & -6.747308e-01 \\ -0.06256595 & 2.453567e-04 & -1.832334e-18 \\ -0.67473079 & -1.832334e-18 & 6.133916e-01 \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} 15.6583 \\ -0.03677 \\ 1.7003 \end{pmatrix}$$

So we get:

$$t = \frac{-0.03677}{\sqrt{0.0002455)}} = -2.3467$$

What is the conclusion?

We can do the same for % polyethylene using the same null and alternative hypotheses.

Do this now - what is your conclusion?

SAS produces this type of test for each of the parameters in a model.

Confidence and Prediction Intervals

The t-distribution is used in calculating Confidence Intervals and Prediction Intervals.

Confidence interval (CI): An interval estimate for the mean (average) response at a given value(s) of the predictor(s).

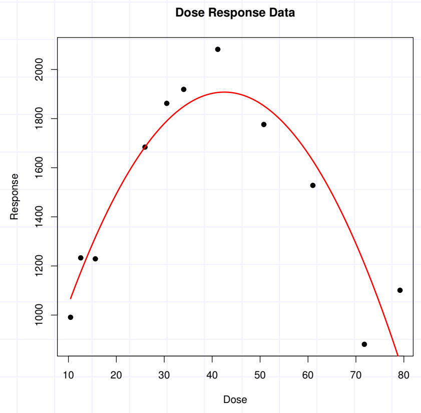
NB. The mean response at a given value of x_i' is also called the fitted value and can be denoted as $\hat{y}(x_i')$.

Prediction interval (PI): An interval estimate for an hypothetical new response at a given value(s) of the predictor(s).

This is the quadratic model for the Dose response data. The LS parameters were:

$$\hat{\beta} = \begin{pmatrix} 430.1 \\ 69.5 \\ -0.8 \end{pmatrix}$$

This suggests a maximum activity level in the vicinity of $x = 43$.
How can we find a 95% CI and prediction interval at $x = 43$?



The point estimate for the mean response at $x = 43$ is:

$$\hat{y}_i(43) = \hat{\beta}_0 + \hat{\beta}_1(43) + \hat{\beta}_2(1849) = L\hat{\beta} \qquad \text{where } L = \begin{pmatrix} 1 & 43 & 1849 \end{pmatrix}$$

Of course here the choice of L could also be written as x_i', i.e. a row of the X matrix that would correspond to $x = 43$.

Using our knowledge above concerning the t-test we get:

$$\text{CI}_{100(1-\alpha)\%} : \qquad \hat{y}_i(x_i') \pm t_{1-\alpha/2, df=n-p}\sqrt{s^2 x_i'(X'X)^{-1}x_i}$$

For the dose response example above we get the following 95% CI:

$$\text{CI}_{95\%}(x = 43) : \qquad 1907.6 \pm 2.306(86.98) = (1707.02,\ 2108.18)$$

A prediction interval is derived by considering the variation of a new y^* around the mean value:

$$\begin{aligned} Var[y^* - \hat{y}] &= Vay[y^*] + Var[\hat{y}] - 2Cov(y^*, \hat{y}) \\ &= Vay[y^*] + Var[\hat{y}] \\ &= s^2\left(1 + x_i'(X'X)^{-1}x_i\right) \end{aligned}$$

This gives the following prediction interval formula:

$$\text{PI}_{100(1-\alpha)\%} : \qquad \hat{y}_i(x_i') \pm t_{1-\alpha/2, df=n-p}\sqrt{s^2(1 + x_i'(X'X)^{-1}x_i)}$$

For the dose response example above we get the following for a 95% prediction interval:

$$\text{PI}_{95\%} : \qquad 1907.6 \pm 2.306(202.3) = (1441.1,\ 2374.1)$$

For the data used in the model, we can efficiently calculate the fitted values are their variances:

$$\hat{Y} = X\hat{\beta} \qquad Var[\hat{Y}] = \sigma^2 H \approx s^2 H$$

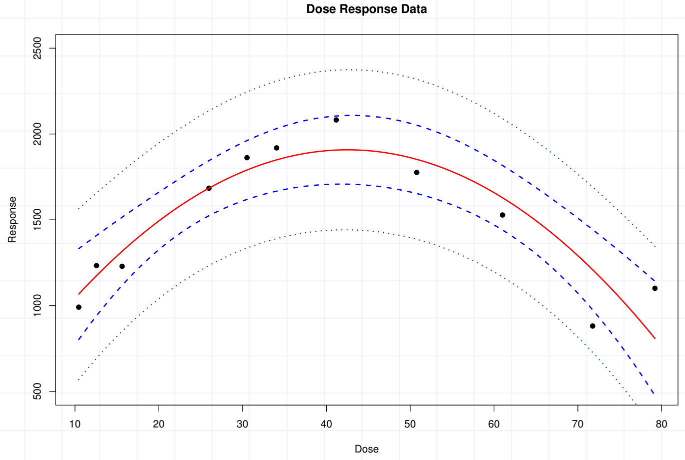
The variance of the fitted value for observation i is:

$$Var[\hat{y}_i] = \sigma^2 x_i'(X'X)^{-1}x_i = \sigma^2 H_{ii} \approx s^2 H_{ii}$$

The prediction variances are:

$$Var[Y^* - \hat{Y}] = \sigma^2(I + H) \approx s^2(I + H)_{ii}$$

In the case of polynomial models like the dose response data we can plot both the CI and PI across the range of the data:



Methods of Model Building

In many cases, we have to choose what predictors to use in our model.

Example: For the Bread Wrapper data should the model have both temperature and polyethylene as predictors, just one of them or none?

Example: For the Dose Response data is the quadratic the best model, or should we use the cubic or even a quartic model?

How do we select the best model? Define the best model as:

Best Model = The minimum, plausible model which adequately describes the data.

Underfitting

Imagine the true model is:

$$Y = X_A\beta_A + X_B\beta_B + \epsilon$$

where X_A is a $n \times p$ regression matrix, β_A is $p \times 1$, X_B is a $n \times q$ regression matrix and β_B is $q \times 1$. The underfitted model is:

$$Y = X_A\beta_A + \epsilon$$

What are the consequences of underfitting?

We can show that:

$$\begin{aligned} E[\hat{\beta_A}] &= \beta_A + A\beta_B \neq \beta_A \\ E[s^2] &= E\left[\frac{1}{n-p}Y'[I - H_A]Y\right] \\ &= \sigma^2 + \beta_B'[X_B - X_A A]'[X_B - X_A A]\beta_B > \sigma^2 \end{aligned}$$

where $A = (X_A'X_A)^{-1}X_A'X_B$ is called the Alias matrix.

[Proof?]

So, the effects of underfitting are:

- Bias the estimate of s^2 upwards and hence increase the Type II error rate.
- Bias the estimates of the parameters for the predictors that are fitted.

Overfitting

Imagine the true model is:

$$Y = X_A\beta_A + \epsilon$$

where X_A is a $n \times p$ regression matrix, β_A is $p \times 1$.

The overfitted model is:

$$Y = X_A\beta_A + X_B\beta_B + \epsilon$$

where X_B is a $n \times q$ regression matrix and β_B is $q \times 1$.

What are the consequences of overfitting? We can show that:

$$E[\hat{\beta_A}] = \beta_A \qquad E[s^2] = \sigma^2 \qquad \text{[Proof?]}$$

$$Var[\hat{\beta}_A|X_B] > Var[\hat{\beta}_A|X_B = 0]$$

So, the effects of underfitting are:

- Increases variance for the parameter estimates - increasing Type II error.
- In crease in the variance for fitted and predicted values.

The moral of all this is: We need to avoid underfitting and overfitting - but a small amount of overfitting is less of a problem.

In particular, a modest (but not excessive) amount of overfitting will still lead to an unbiased estimate $s^2 \approx \sigma^2$ and reasonable estimates of the parameters in β_A.

Model Selection for Polynomial Models

Question for the Dose-Response Data : Do we need the cubic term in the model? To answer these lets propose a few solutions.

Define the statistic R^2 as the proportion of variation in the data explained by the model. More exactly it is:

$$R^2 = \frac{SSmodel}{SStotal(corrected)}$$

So the explanatory power of the model can be assessed by reference to R^2.

Proposal (1): Just look at R^2 - the model (quadratic or cubic) with the highest R^2 is explaining most variation in the data and hence is to be preferred?

Problem: R^2 for quadratic fit 0.84, and for cubic 0.90. However, if I go on to fit a quartic (i.e. add (dose)^4 and another slope β_4, I get R^2 = 0.9656. And if I keep adding higher order terms R^2 keeps increasing.

Degree of Polynomial	R^2	Adjusted R^2
1	0.0120	-0.0978
2	0.8428	0.8035
3	0.9044	0.8634
4	0.9656	0.9427
5	0.9743	0.9485
6	0.9759	0.9398
7	0.9907	0.9691
8	0.9908	0.9542
9	0.9998	0.9978
10*	1.0000	1.0000

*Theoretical result - not from SAS

The adjusted R^2 above is an adjusted version with respect to the number of parameters in the model. It comes from the fact that R^2 can be written as,

$$R^2 = \frac{SS_{Model}}{SS_{Tot(corrected)}} = 1 - \frac{SS_e}{SS_{Tot(corrected)}}$$

$$Adjusted\ R^2 = 1 - \frac{SS_e/(n-p)}{SS_{Tot(corrected)}/(n-1)}$$

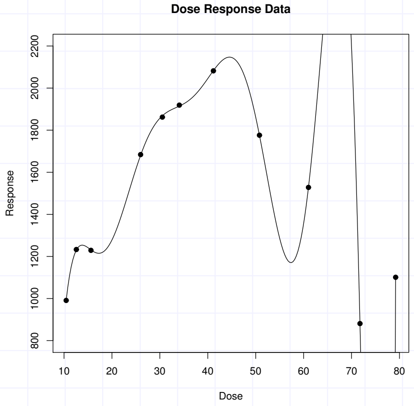
So, adjusted R^2 penalises models with more parameters.

As can be seen using R^2 as a guide will always favour a more complicated (bigger) model.

This is not good - we want to follow the law of parsimony, i.e. the simplest plausible model is best.

Adjusted R^2 is a bit better, but not much.

As an example, look at the polynomial of degree 10, here's a plot of the LS line it gives,



Which is an interpolating polynomial, which involves no simplification of the data.

When it comes to pure polynomial models, we can approach model fitting in the following way:

- The question is: what is the correct degree for the polynomial?
- If polynomial of degree r is fitted, then all coefficients in that polynomial should be left unconstrained - i.e. all coefficients are fitted even if some of them are statistically not significantly different from zero.
- The one exception to this rule, is the last coefficient of the polynomial, i.e. the coefficient for the x^r predictor.
- If the coefficient for x^r is not statistically significant (use a general linear hypothesis to test this) then it is removed and the simpler polynomial of degree $r-1$ is considered.
- You could proceed as follows:
  1. Start with the lowest degree polynomial that is plausible given a plot of the data.
  2. increase the degree by one, one step at a time. At each step check that that parameter for x^r is significantly different from zero - if it is, it stays in the model, if not use the polynomial with degree $r-1$.

```
> summary(lm(activity~dose+I(dose^2),data=dr))

Call:
lm(formula = activity ~ dose + I(dose^2), data = dr)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 430.0759   205.3052   2.095 0.069496 .
dose         69.5021    11.2468   6.180 0.000265 ***
I(dose^2)    -0.8172     0.1257  -6.502 0.000188 ***
---

Residual standard error: 182.7 on 8 degrees of freedom
Multiple R-squared:  0.8428,  Adjusted R-squared:  0.8035
F-statistic: 21.45 on 2 and 8 DF,  p-value: 0.0006106
```

Clearly the quadratic term is required. So, try adding the cubic term.

```
> summary(lm(activity~dose+I(dose^2)+I(dose^3),data=dr))

Call:
lm(formula = activity ~ dose + I(dose^2) + I(dose^3), data =
     dr)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.579e+02  3.255e+02  -0.485  0.64244
dose         1.321e+02  3.091e+01   4.272  0.00369 **
I(dose^2)   -2.522e+00  8.095e-01  -3.116  0.01695 *
I(dose^3)    1.303e-02  6.134e-03   2.124  0.07131 .
---

Residual standard error: 152.3 on 7 degrees of freedom
Multiple R-squared:  0.9044,  Adjusted R-squared:  0.8634
F-statistic: 22.08 on 3 and 7 DF,  p-value: 0.0006049
```

Looks like the cubic term is 'marginally' significant. Maybe we shouldn't discount it yet. Try adding a quartic term.

```
> summary(lm(activity~dose+I(dose^2)+I(dose^3)+I(dose^4),
    data=dr))

Call:
lm(formula = activity ~ dose + I(dose^2) + I(dose^3) + I(
    dose^4),
    data = dr)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.297e+03  4.925e+02   2.634   0.0389 *
dose        -7.376e+01  6.608e+01  -1.116   0.3070
I(dose^2)    6.338e+00  2.761e+00   2.296   0.0615 .
I(dose^3)   -1.335e-01  4.501e-02  -2.966   0.0251 *
I(dose^4)    8.178e-04  2.502e-04   3.269   0.0171 *
---

Residual standard error: 98.64 on 6 degrees of freedom
Multiple R-squared:  0.9656,  Adjusted R-squared:  0.9427
F-statistic: 42.13 on 4 and 6 DF,  p-value: 0.0001583
```

Looks like the cubic term is ‘marginally’ significant. Maybe we shouldn’t discount it yet. Try adding a quartic term.

```
> summary(lm(activity~dose+I(dose^2)+I(dose^3)+I(dose^4),
    data=dr))

Call:
lm(formula = activity ~ dose + I(dose^2) + I(dose^3) + I(
    dose^4),
    data = dr)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.297e+03  4.925e+02   2.634   0.0389 *
dose        -7.376e+01  6.608e+01  -1.116   0.3070
I(dose^2)    6.338e+00  2.761e+00   2.296   0.0615 .
I(dose^3)   -1.335e-01  4.501e-02  -2.966   0.0251 *
I(dose^4)    8.178e-04  2.502e-04   3.269   0.0171 *
---

Residual standard error: 98.64 on 6 degrees of freedom
Multiple R-squared:  0.9656,  Adjusted R-squared:  0.9427
F-statistic: 42.13 on 4 and 6 DF,  p-value: 0.0001583
```

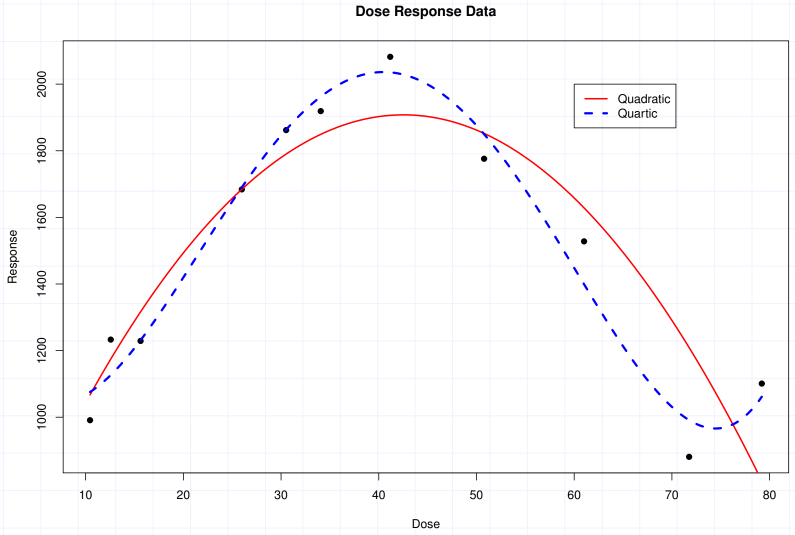
```
>summary(lm(activity~dose+I(dose^2)+I(dose^3)+I(dose^4)+I(
    dose^5),data=dr))

Call:
lm(formula = activity ~ dose + I(dose^2) + I(dose^3) + I(
    dose^4) +
    I(dose^5), data = dr)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.516e+02  9.992e+02   0.152    0.885
dose         1.240e+02  1.649e+02   0.752    0.486
I(dose^2)   -5.361e+00  9.394e+00  -0.571    0.593
I(dose^3)    1.742e-01  2.411e-01   0.722    0.502
I(dose^4)   -2.870e-03  2.854e-03  -1.005    0.361
I(dose^5)    1.640e-05  1.265e-05   1.297    0.251

Residual standard error: 93.48 on 5 degrees of freedom
Multiple R-squared:  0.9743,  Adjusted R-squared:  0.9485
F-statistic: 37.87 on 5 and 5 DF,  p-value: 0.000561
```

Stop there - it looks like a quartic model might be required.



Some points about all this:

- At early stages we might err on the side of including higher degree terms, by accepting a higher than normal p-value for inclusion. This would somewhat offset the tendency of underfitted models to have a large Type II error rate.
- We need to refit the model at each stage - as the estimate of s^2 changes as do the parameters estimates.
- We might be a bit concerned that this method favours more complicated models - as it those not penalise models for the number of parameters included - see later for a suggested fix for this problem.
- Type I SS depend on the order that predictors come in the X matrix - so whenever that is meaningful then they are of relevance - this obviously includes polynomial models.

Model Selection for General Models

Example: Fitness Data;

These measurements were made on men involved in a physical fitness course. The variables are Oxygen intake rate (ml per kg body weight per minute) - the response variable, Age (years), Weight (kg), time to run 1.5 miles (minutes), heart rate while resting, heart rate while running (same time Oxygen rate measured), and maximum heart rate recorded while running.

				Rest	Run	max
Oxygen	Age	Weight	Time	Pulse	Pulse	Pulse
44.609	44	89.47	11.37	62	178	182
45.313	40	75.07	10.07	62	185	185
54.297	44	85.84	8.65	45	156	168
59.571	42	68.15	8.17	40	166	172
49.874	38	89.02	9.22	55	178	180
44.811	47	77.45	11.63	58	176	176
45.681	40	75.98	11.95	70	176	180
49.091	43	81.19	10.85	64	162	170
39.442	44	81.42	13.08	63	174	176
60.055	38	81.87	8.63	48	170	186
50.541	44	73.03	10.13	45	168	168
37.388	45	87.66	14.03	56	186	192
44.754	45	66.45	11.12	51	176	176
47.273	47	79.15	10.60	47	162	164
51.855	54	83.12	10.33	50	166	170
49.156	49	81.42	8.95	44	180	185
40.836	51	69.63	10.95	57	168	172
46.672	51	77.91	10.00	48	162	168
46.774	48	91.63	10.25	48	162	164
50.388	49	73.37	10.08	67	168	168
39.407	57	73.37	12.63	58	174	176
46.080	54	79.38	11.17	62	156	165
45.441	52	76.32	9.63	48	164	166
54.625	50	70.87	8.92	48	146	155
45.118	51	67.25	11.08	48	172	172
39.203	54	91.63	12.88	44	168	172
45.790	51	73.71	10.47	59	186	188
50.545	57	59.08	9.93	49	148	155
48.673	49	76.32	9.40	56	186	188
47.920	48	61.24	11.50	52	170	176
47.467	52	82.78	10.50	53	170	172

Method 1: All possible Regressions

In this method we fit all possible regressions, i.e.,

- The intercept only model
- Set of all possible 1 predictor models
- Set of all possible 2 predictor models

⋮

- Set of all possible p-1 predictor models

Then identify the best model in each parameter set and choose between those models.

The best model is the model in each set with the largest adjusted R^2. Note: if there are $p-1$ predictor variables under consideration, then there are $2^{(p-1)}$ possible regressions - or 2^p if you are considering constrained regressions as well.

The models found with the best adjusted R^2 in each set were;

No. Predictors	Model Terms	Adjusted R^2
1	Run Time	0.7345
2	Age, Run Time	0.7474
3	Age, Run Time, Run Pulse	0.7901
4	Age, Run Time, Run Pulse, Max Pulse	0.8117
5	Age, Weight, Run Time, Run Pulse, Max Pulse	0.8176
6	Age, Weight, Run Time, Run Pulse, Rest Pulse, Max Pulse	0.8108

These are the best from the set of 63 regressions [See printout].

So, which model do we choose?

Well look at the 6 candidate models above and choose using sequential F tests.

Or, just take the smallest model with an acceptable level of adjusted R^2. There are two problems with all this;

1. suppose you had 25 predictors to start with, then you have some 33,554,432 regressions to fit. Even with fast computing this will take some time.
2. Since you are choosing models based on selecting the best model (i.e. the model with the lowest s^2 - this is equivalent to selecting the highest adjusted R^2) from a large number of models, then the p-values you calculate are being biased (this is sometimes called interrogating the data). Rejection of the H_0 : may be virtually guaranteed in many instances.

Information Criteria and Model building

There are a number of Information Criteria that are routinely used in model building. The most important is the Akaike Information Criterion (AIC).

AIC is based on fairly complex theory from information entropy - so we will omit any deep discussion. It takes a remarkably simple form however:

AIC $= -2$ log likelihood (evaluated at MLE) $+2p$

In this version the model with smaller AIC is preferred, therefore models with more parameters are penalised.

This is helpful, as the log likelihood is a non-decreasing function of the number of parameters.

The main advantage of the AIC is that is can be used to compare non-nested models - with the model with the smaller AIC preferred.

Drawback to AIC are:

(1) we don't know its distribution

(2) it can be misleading in small samples.

In the case of drawback (2) we can use the corrected AIC or AICc:

AICc $= -2$ log likelihood (evaluated at MLE) $+2p\frac{n}{n-p-1}$

Method 2: Forward Selection using AIC

Forward Selection: This is one of three main **stage-wise procedures**. (The others are backward selection and stepwise which are discussed below).

The forward selection algorithm proceeds as follows:

1. Consider all predictors not already included in the model one at a time for entry to the model. The single predictor selected as a candidate for entry at a stage is that predictor not already in the model that results in the largest **reduction** in AIC among all predictors under consideration for entry.
2. Once a predictor has been included in the model it is not removed.
3. Stop when the candidate predictor at a stage fails to reduce (i.e. improve) the AIC.

See output for Fitness DATA.

Method 3: Backward Selection using AIC

The Backward Selection algorithm is as follows:

1. Begin by calculating AIC for a model which includes all possible predictors.
2. Predictors are considered for removal from the model one at a time.
3. The predictor chosen for removal at any stage in that predictor whose removal results in the largest decrease in AIC.
4. Once a predictor is removed it is not considered for re-entry to the model.
5. Stop when the removal of any of the remaining predictors fails to reduce AIC

See output for Fitness DATA.

Method 4: Stepwise Selection using AIC

The stepwise algorithm proceeds as follows:

1. The stepwise algorithm is a compromise between the forward and backward algorithms. It differs from the forward selection algorithm by allowing predictors already included in the model may be removed at later stages.
2. As in forward selection, predictors are considered for entry to the model one at a time. The candidate predictor for inclusion at any stage is that predictor not already included that results in the largest decrease (i.e. improvement) in AIC.
3. At any stage where a predictor has been added to the model, the AIC values are computed for all predictors in the model at that stage. A predictor will be considered for removal from the model if its removal will result in a decrease of AIC. If several predictors achieve a reduction of AIC upon their removal - the predictor with the biggest reduction is removed first.

4. If a predictor is removed from the model, then the ‘amended’ model is refitted, and a check is made on those predictors to see if any others can be removed.
5. The stepwise process ends when the addition or removal of a predictor fails to improve the AIC.

See output for Fitness DATA.

NB: AIC can also be used as a measure in comparing polynomial models.

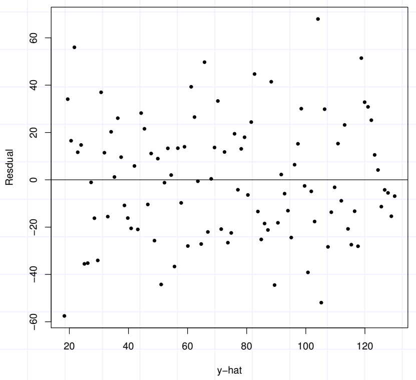
Doseresponse data - Quadratic model: AIC=117.1
Doseresponse data - Quartic model: AIC=104.3

Assessing Model fit - Residual Diagnostics, Influence and Leverage

- In fitting the above models and deciding what terms are significantly difference from zero etc., we have made some assumptions.
- In particular we have assumed that the residuals are iid normal with mean 0 and equal variance σ^2.
- Now we want to check that these assumptions are reasonable. We do this is validate what we are doing WRT model fitting. hypothesis testing and our conclusions.

Residual Diagnostics

If the fitted model and the model assumptions are correct then our residuals should be iid normal. If this is the case then if we plot the residuals against say their fitted values we should get a random scatter of points - a featureless cloud of points. For example, an ideal situation might look like the following,



We know from the model that the ε_i's are $N(0, \sigma^2)$ but what about their estimates, i.e. the e_i's?

We need to check that the e_i's follow the assumptions concerning the ε_i's.

Some properties of the e_i's are:

$$\begin{aligned} E[e] &= 0 \\ Var[e] &= \sigma^2(I_n - H) \qquad (5) \\ & \qquad (6) \end{aligned}$$

where (5) is the variance-covariance matrix for the e's. Note that the diagonals of this matrix are the individual variances of the residuals and the off diagonals are their covariances.

[Proof?]

The diagonals of the Hat matrix play an important role in what follows. They have the following properties.

- Model with intercept

$$\frac{1}{n} \leq h_{ii} \leq 1$$

- Model with no intercept

$$0 \leq h_{ii} \leq 1$$

- Hat Diagonals

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

This is a normalised measure of the distance of the x_i vector from the vector of the means of the x's.
We say it is a measure of the distance of the vector x_i from the data centre in X space

This implies that $\sigma^2(I_n - H)$ are positive on the diagonals but are not necessarily zero on the off diagonals.

So the e_i's depart from the ε_i's insofar as they are not in general uncorrelated.

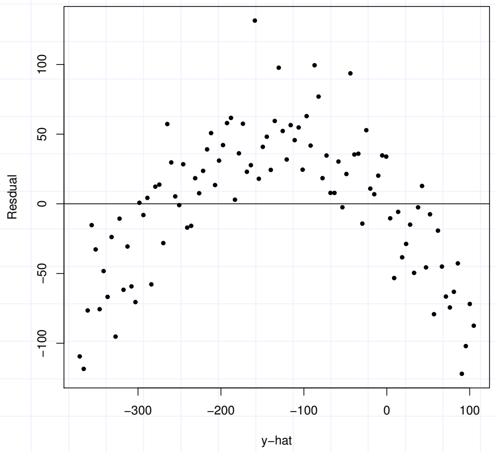
Despite this we can use the residuals in a simple way to show the following departures from the model assumptions;

- Model underspecification
- Departure from equal variation assumption (heterogeniety of variance)
- Existence of suspect data points (outliers)
- Identification of high influence points
- Departures from normality

Plotting of Raw Residuals

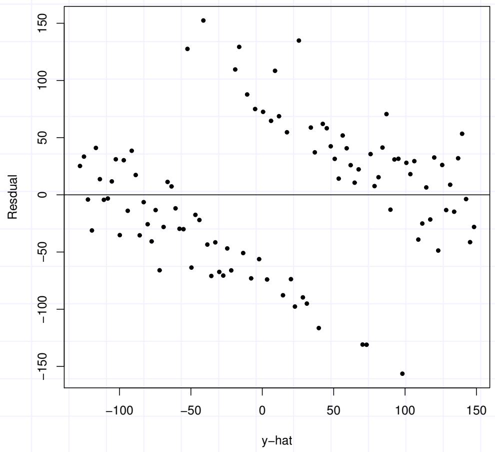
- The e's are the raw residuals. We can use them to detect model underspecification and non-constant variance.
- If the model is underspecified we will often see the structure of the underspecification in a plot of the residuals against the fitted values.
- So plot $(Y - \hat{Y})$ against $\hat{Y}$. Look for any structure in these plots which suggests problems.

The following plot suggests a missing higher degree polynomial term.



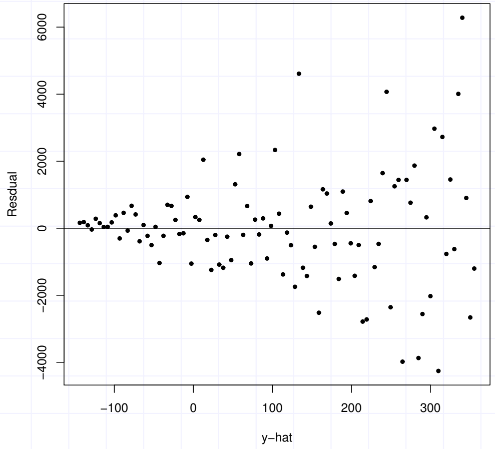
Remedy: Put the appropriate polynomial term in the model.

The following suggests a missing grouping term (categorical term).



Remedy: Use an appropriate categorical variable in the model.

The following plot indicates that the variances are not constant.



Remedy: Use weighted regression (if appropriate).

Outlier Detection

- To identify a potential outlier we need to find a point that yields a residual that is big in absolute value compared to the rest.
- This suggests that this point(s) are poorly fitted by the model and therefore corrective action may be needed. The question is how big is too big?
- It would be helpful if the e_i's were standardised - i.e. had zero mean and unit variance.
- It can be shown that the e_i's will have zero mean when an intercept is included. But we have seen that their individual variances are given by $\sigma^2(1 - h_{ii})$ where h_{ii} is the i^{th} diagonal of the hat matrix. These will not be the same in general.

It gets worse - look at the variance of the i^{th} residual i.e. $\sigma^2(1 - h_{ii})$.

This will be small when h_{ii} is close to 1.

It can be shown that h_{ii} is given by $x_i'(X'X)^{-1}x_i$, where x_i is the vector of predictor variables for the i^{th} observation, and this will be close to 1 when one or more of elements of x_i is far from the mean of the x's.

This implies that the variance of a residual for an observation which is far from the data centre (for one or more predictor variables) will be small.

So, it is helpful to standardise the residuals to give them unit variance - so they are compared on the same scale.

One way of doing this is by **studentised residuals** (also called standardised residuals).

$$\text{Studentised Residuals: } r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}} \tag{7}$$

This residual follows a t-like distribution (it is not exactly a t-distribution).

By removing the increase or decrease in the variance of the residuals due to the location of the observation in the X space we can more easily spot residuals that are bigger than general.

Example: Here are various residuals for the forestry data.

obs.	h_{ii}	residual	studentised	r-student	press
1	0.241	-0.005	-0.020	-0.019	-0.007
2	0.218	-0.290	-1.118	-1.127	-0.371
3	0.177	0.401	1.506	1.574	0.488
4	0.253	-0.351	-1.384	-1.428	-0.470
5	0.135	0.047	0.172	0.166	0.054
6	0.246	0.128	0.503	0.491	0.170
7	0.075	-0.149	-0.527	-0.515	-0.161
8	0.128	0.189	0.689	0.677	0.216
9	0.430	0.055	0.248	0.240	0.096
10	0.171	-0.571	-2.135	-2.444	-0.688
11	0.165	0.125	0.468	0.456	0.150
12	0.182	-0.099	-0.374	-0.364	-0.121
13	0.191	-0.230	-0.869	-0.862	-0.284
14	0.103	-0.125	-0.450	-0.438	-0.139
15	0.108	-0.194	-0.701	-0.689	-0.218
16	0.163	0.185	0.687	0.676	0.221
17	0.175	-0.120	-0.448	-0.437	-0.145
18	0.085	0.498	1.774	1.916	0.544
19	0.266	0.095	0.379	0.369	0.130
20	0.489	0.410	1.955	2.169	0.803

Compare observation (10) and (20). Looking at the raw residuals for (10) looks (in absolute value) much bigger than (20) - 139%.

But this is without taking into account their positions in X space which is known to have an effect on the raw residual (i.e. in a sense they are on a different scale).

So, comparing them with the studentised residual we find that (10) is ‘closer’ to (20) - 109%.

This is because (20) is more extreme in the X space than (10) [look at the data and the value of h_{ii}].

What do we do with large residuals?

There are three main causes of large residuals

1. true random variation - so do nothing
2. a mistake has been made in the data collection - if this can be established to have happened then the observation can be removed from the analysis or corrected
3. a model breakdown at some point in the data. In the case of (3) we have a particular problem - if the model breaks down at a given point then this is the same thing as an inadequate model for the data concerned.

There is a potential weakness of the studentised residual.

If an observation(s) is a true outlier, i.e. an observation that does not follow the model, then the s^2 will be inflated [recall what happens when we underfit].

If this is the case then is may be better to compute an s^2 estimate without the suspect observation included - denote this as s^2_{-i}.

This is an example of a ‘leave one out analysis’. In turns out that with some algebra we get;

$$s_{-i} = \sqrt{\frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}} \tag{8}$$

This is estimate is used instead of s^2 to give an **externally studentised** residual (hence the residual given by equation (7) is called internally studentised).

The crucial part of (8) is $e_i^2/(1-h_{ii})$.

The value $e_i/(1-h_{ii})$ is called the PRESS residual.

So the formula for the externally studentised (R-student) residual becomes,

$$\text{R-student:} \;\; t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} \tag{9}$$

The R-student residuals follow a t distribution when testing certain model breakdown hypotheses. The model breakdown could be,

1. The model breakdown is caused by a location shift, i.e. $E[\varepsilon_i] =\neq 0$. This is the mean shift outlier model. In this case we can use the R-student to test the null hypothesis $H_0 : \varepsilon_i = 0$.
2. The model breakdown is results in $Var[\varepsilon_i]$ being bigger than at the other data locations, i.e.$Var[\varepsilon_i] = \sigma^2 + \sigma_i^2$. In this case we can use the R-student to test the null hypothesis, $H_0 : \sigma_i^2 = 0$.

In both cases we use the critical values of the t distribution with $n-p-1$ degrees of freedom. For example for the Forestry data, (n=20, p=3) we can compare with the critical value of the t-distribution with 16 DF (critical value at $\alpha = 0.05$ is 2.12).

But, if we have no *a priori* suspicion of any points and so are investigating them all simultaneously, then we need to correct for the Type I error rate.

Therefore, use the Bonferroni correction, so if α is the type I error rate use α/n where n is the number of residuals being investigated.

This will result in a Type I error rate across all the residuals that is no larger than α - so it is conservative.

For the Forestry data therefore, the correct critical value is 3.58, (i.e. value for $t_{.05/20,df=16}$).

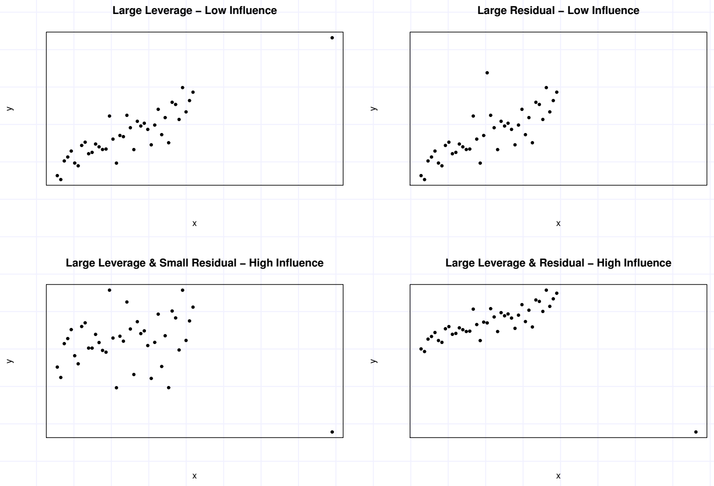
It is probably better to use the critical values only as an approximate yardstick which points at suspect data points which should get more attention.

In fact, any point for which $|t_i| > 2$ may be flagged as ‘suspect’ since it is more than 2 standard deviations away from the mean.

Influence

- Influence refers to the amount of influence a particular data point has on the regression statistics.
- These statistics can be the estimated parameters, s^2 and/or other performance related statistics (e.g. residuals). What we are looking for is the answer to the question; if the i^{th} observation was omitted from the model, how would things change?
- This is of interest as a large residual may indicate a failure of the model at a particular point, and if it also occurs with high leverage then that observation may influence the fitted model disproportionately.
- Leverage is a measure of the potential for an observation to exert influence. It is given as the diagonal of the HAT matrix (i.e. h_{ii}) which is a measure of standardised distance of the observation from the data center in terms of the predictor (X) space.

An observation will certainly be influential where both the leverage and the residual is large. But sometimes one will be large when the other small as seen in the following plots:



How do you identify influential observations?

Observations with high residuals potentially have high influence.

Points with large HAT diagonals (i.e. leverage) also have the potential to be influential.

How large is large for h_{ii}?

Recall that the trace of the HAT matrix = p, so a rule of thumb is $h_{ii} > 2p/n$ suggests the potential for exerting strong influence (this only works where $2p/n < 1$).

We really want to measure the actual influence on:

1. the fitted values and
2. the parameter estimates.

DFFITS
What would the fitted value $\hat{y}_i$ be for a set of predictors x_i' if observation i was not included in estimating the model regression parameters?

Denote the the fitted value with the ith observation included in the estimation $\hat{y}_i$ and denote $\hat{y}_{-i}$ the fitted value with it excluded from the estimation.

The logic is, if these two value are close, then observation i exerts little influence on the model fit. But, if they are far apart then observation i does exert strong influence on the model fit.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{-i}}{s_{-i}\sqrt{h_{ii}}} \qquad (10)$$

where h_{ii} and s_{-i} are used to standardise the difference.

DFBETAS
This looks at the influence an observation has on the parameter estimates directly.

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{s_{-i}\sqrt{c_{jj}}} \tag{11}$$

where $\hat{\beta}_{j,-i}$ is the j^{th} parameter estimate calculated without the i^{th} observation, and c_{jj} is the j^{th} diagonal element of the $(X'X)^{-1}$ matrix.

Note also that there will be p DFBETAS values for each observation.

Cook's D
An aggregate measure of influence on the parameters is given by Cook's D (Cook's Distance).

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})'(X'X)(\hat{\beta} - \hat{\beta}_{-i})}{ps^2} \tag{12}$$

Cook's D is a normalised measure of distance from the set of parameter estimates when the i^{th} observation included and excluded from the model fit. To identify what parameter values a large D_i is influencing, use DFBETAS.

Some authors advocate the use of yardstick for what are and are not large values for DFFITS, DFBETAS and Cook's D. for example;

Measure	Critical Yardstick
DFFITS	$2\sqrt{p/n}$
DFBETAS	$2/\sqrt{n}$
Cook's D	Use $\approx > 1$

Better to use comparisons among these measures - which is the biggest and why?

Is a suspect observation exerting high influence?

Is a particular sub-group of observation exerting high influence, etc.

These measures will identify any observation that are having a disproportionately large influence - the implications for the model may then be discovered.

Checking the normality assumption

- We assume that the errors are uncorrelated with the same variance.
- The residuals from the fitted model are centred and scaled to be uncorrelated with unit variance using the studentised residual.
- If we order these we should get an ordered group from a standard normal distribution.
- We can then use a result from order statistics that the expected value of the r^{th} order statistics from the standard normal is well approximated by:

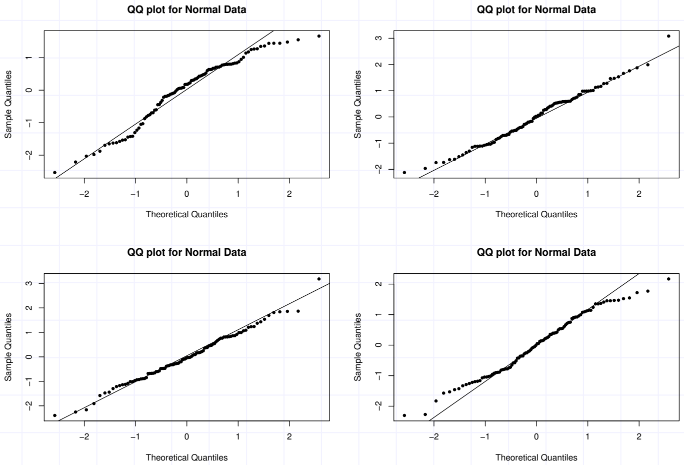
$$q_i = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right)$$

Where i is the order number of the q_i.

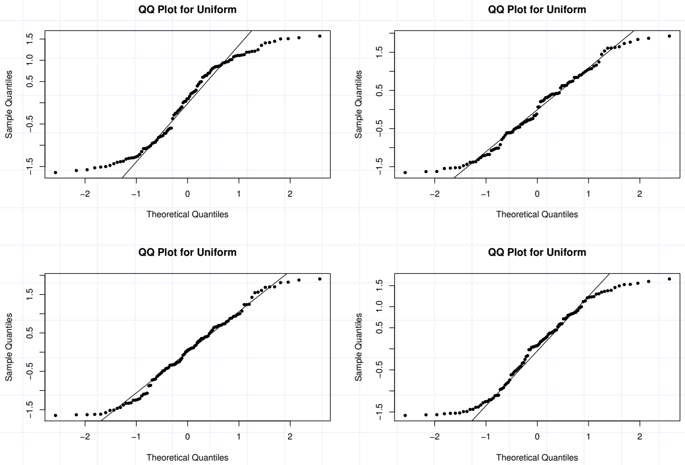
- An approach to check the normality assumption in linear models is to plot these expected **quantiles** against the observed quantiles of the studentised residuals.
- If the residuals are normally distributed then a straight line should be observed with an intercept at the origin. This is called *Normal probability Plot* or a *Quantile Quantile (QQ)* plot.
- In models involving real data we will never get a perfectly straight line, so it is important to learn to read a QQ plot to spot systematic departures from normality.

In the plots that follow the line is drawn though the 1^{st} and 3^{rd} quartiles of the data.

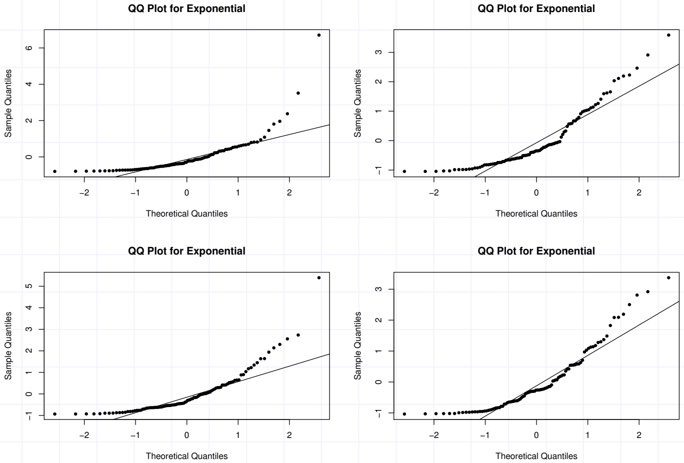
Normal Errors



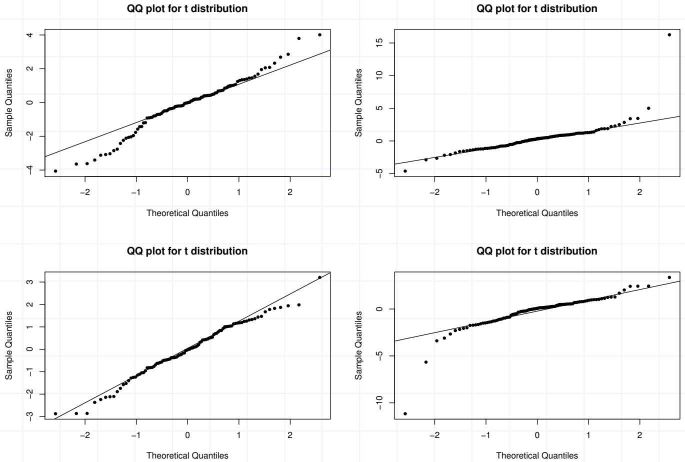
Light tailed Errors



Positive Skew, (i.e. light left tail, heavy right tail)



Heavy Tail Errors



Including Categorical Predictors

Clinical Trial Data: Medical experimenters wish to compare the effectiveness of a new type of drug versus the standard. They recruit 9 patients and randomly assign them to a placebo (treatment 1), the standard drug (treatment 2), and a test drug (treatment 3). They get the following results;

Response:	91	97	104	112	115	114	119	116	115
Treat:	1	1	1	2	2	2	3	3	3

We model the response as a linear function of an overall intercept and a treatment effect; i.e. $y_i = \mu + t_j + \varepsilon_i$ where t_j is the effect for treatment j. This is just another type of linear model which can be easily put in a regression framework.

$$Y = X\beta + \varepsilon$$

$$\begin{pmatrix} 91 \\ 97 \\ 104 \\ 112 \\ 115 \\ 114 \\ 119 \\ 116 \\ 115 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix}$$

where $\beta_0 = \mu$ and $\beta_j = t_j$ for $j = 1, \ldots, 3$. So we have a linear model again but with what are called dummy variables as predictors (also called categorical, indicator, factor variables etc.).

Why not just use the values of 1, 2 and 3 in one column of the regression matrix?

Nothing has changed from before, we just use least squares to find the solutions, i.e., $\hat{\beta} = (X'X)^{-1}X'Y$.

But there is one problem; from linear algebra we know that column rank(X)=rank($X'X$). Look at X above - it is not full column rank since the dummy variables will add to give the intercept column.

$$(X'X) = \begin{pmatrix} 9 & 3 & 3 & 3 \\ 3 & 3 & 0 & 0 \\ 3 & 0 & 3 & 0 \\ 3 & 0 & 0 & 3 \end{pmatrix}$$

So, $(X'X)$ is singular and therefore has no inverse. There are as more unknowns than linearly independent normal equations - we say that the model is over-parameterised. This will happen whenever a categorical variable is included in a linear model.

So, how is this type of model fitted? There are two main ways of dealing with singularity of $(X'X)$.

1. Impose a constraint on the model parameters in such a way as to make $(X'X)$ non-singular.
2. Use a generalised inverse (or pseudo inverse) of $(X'X)$. [This is a general method which includes the constraint method]

SAS uses a generalised inverse approx - however it can be shown to essentially equivalent to a particular choice of constraint for most practical purposes.

Constraints on parameters

This approach imposes a constraint that reduces the model to full-rank - we can also call this a reparameterisation of the model. A classic way of doing this (very popular in software) is to constrain one of the parameters to be zero. For example, in the clinical Trial data we arbitrarily set one of the treatment effects to be zero - lets say treatment 3. The matrix setup now becomes;

$$Y = X\beta^o + \varepsilon$$

$$\begin{pmatrix} 91 \\ 97 \\ 104 \\ 112 \\ 115 \\ 114 \\ 119 \\ 116 \\ 115 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_0^o \\ \beta_1^o \\ \beta_2^o \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix}$$

$$(X'X) = \begin{pmatrix} 9 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 1/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & 1/3 \\ -1/3 & 1/3 & 2/3 \end{pmatrix}$$

where β^0 is used to indicate that the parameters are dependent on the constraint chosen. What is ‘nice’ about all this is that nothing has changed from all that has gone before as regards Least Squares - so all that we have covered up to now applies to these models as well, i.e. hypothesis testing, model building, residual diagnostics etc.

But does this means we haven't estimated treatment 3?

Look at what has happened.

Response $=$ Unconstrained $\rightarrow$ Constrained $+$Error

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{pmatrix} = \begin{pmatrix} \mu + t_1 \\ \mu + t_1 \\ \mu + t_1 \\ \mu + t_2 \\ \mu + t_2 \\ \mu + t_2 \\ \mu + t_3 \\ \mu + t_3 \\ \mu + t_3 \end{pmatrix} \rightarrow \begin{pmatrix} \mu + t_1 \\ \mu + t_1 \\ \mu + t_1 \\ \mu + t_2 \\ \mu + t_2 \\ \mu + t_2 \\ \mu \\ \mu \\ \mu \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix}$$

So, the LS estimates ($\hat{\beta}$) can be written as;

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & -1/3 & -1/3 & -1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 & -1/3 & -1/3 & -1/3 \end{pmatrix} \begin{pmatrix} \mu + t_1 \\ \mu + t_1 \\ \mu + t_1 \\ \mu + t_2 \\ \mu + t_2 \\ \mu + t_2 \\ \mu + t_3 \\ \mu + t_3 \\ \mu + t_3 \end{pmatrix}$$

Therefore we get from the original model formulation;

$$E\begin{pmatrix} \hat{\beta}_0^o \\ \hat{\beta}_1^o \\ \hat{\beta}_2^o \end{pmatrix} = \begin{array}{lcr} \mu + t_3 & = & 116.67 \\ (\mu + t_1) - (\mu + t_3) = t_1 - t_3 & = & -19.33 \\ (\mu + t_2) - (\mu + t_3) = t_2 - t_3 & = & -3.00 \end{array}$$

- We cannot estimate $\mu,\ t_1,\ t_2,\ t_3$ uniquely (this is why we have an over-parameterised model).
- Also, this solution for $\hat{\beta}^o$ is just one of a infinite number of possible solutions - two very obvious alternative are to set t_1 or t_2 to zero. This type of constraint is sometimes called a **set to zero constraint**.
- If you think about the original problem - the questions of interest are the differences between the treatments and we do have access to this information - look at the definition of the parameters.

Another constraint that may be used is the **sum to zero constraint**.

Here we specify that the treatment effects sum to zero, $\sum_j t_j = 0$. This method may be applied by replacing any of the normal equation by this constraint, i.e.,

$$\overset{(X'X)}{\begin{pmatrix} 9 & 3 & 3 & 3 \\ 3 & 3 & 0 & 0 \\ 3 & 0 & 3 & 0 \\ 3 & 0 & 0 & 3 \end{pmatrix}} \overset{\beta}{\begin{pmatrix} \beta_0^o \\ \beta_1^o \\ \beta_2^o \\ \beta_3^o \end{pmatrix}} = \overset{X'Y}{\begin{pmatrix} 983 \\ 292 \\ 341 \\ 350 \end{pmatrix}}$$

Now replace one of the normal equations (say the 1st) with the sum to zero constraint;

$$\begin{array}{ccc} (X'X) & \beta^o & X'Y \\ \begin{pmatrix} 0 & 1 & 1 & 1 \\ 3 & 3 & 0 & 0 \\ 3 & 0 & 3 & 0 \\ 3 & 0 & 0 & 3 \end{pmatrix} & \begin{pmatrix} \beta_0^o \\ \beta_1^o \\ \beta_2^o \\ \beta_3^o \end{pmatrix} = & \begin{pmatrix} 0 \\ 292 \\ 341 \\ 350 \end{pmatrix} \end{array}$$

We can now invert this re-formed $(X'X)$ go get the following solution;

$$\hat{\beta}^{o'} = (109.23,\ -11.89,\ 4.44,\ 7.44)$$

So, the parameter values are different but the estimate of treatment difference is the same.

ANOVA table for models with singular X Matrix

To fit such a mode in R, we use the following code:

```
response=c(91,97,104,112,115,114,119,116,115)
treat=c(rep(c(1,2,3),rep(3,3)))
fit4=lm(response~factor(treat))
summary(fit4)
```

The Clinical Data: What is the overall F test for the model testing?

Notice the set-to-zero constraint used her - what are the interpretation of the parameter estimates?

```
> summary(fit4)

Call:
lm(formula = response ~ factor(treat))

Residuals:
    Min      1Q  Median      3Q     Max
-6.3333 -1.6667 -0.3333  1.3333  6.6667

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      97.333      2.333  41.714 1.27e-08 ***
factor(treat)2   16.333      3.300   4.950  0.00258 **
factor(treat)3   19.333      3.300   5.859  0.00109 **
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05
         .   0.1        1

Residual standard error: 4.041 on 6 degrees of freedom
Multiple R-squared:  0.8689,  Adjusted R-squared:  0.8252
F-statistic: 19.88 on 2 and 6 DF,  p-value: 0.002253
```

What are the H_0 : for each entry in this table and the conclusions?

We can also use customised hypothesis tests to test other interesting hypotheses.

Using a generalised linear hypothesis what are the following testing:

$$L_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$L_2 = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$$

```
# install.packages('multcomp')
library(multcomp)
L1=diag(3)[-1,]
glh1=glht(fit4,linfct=L1)
summary(glh1,test=Ftest())

L2=matrix(c(0,1,-1),nrow=1)
glh2=glht(fit4,linfct=L2)
summary(glh2,test=Ftest())
```

Consider again what the contrast L_2 is estimating:

$$L_2 = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$$

It would be of interest to get the actual estimate of this contrast and perhaps a 90% or 95% confidence interval for it.

We can do this by using the t-distribution version for such a contrast - which can use when L is a vector. This allows us to get both an estimate and a CI.

$$\text{CI}_{100(1-\alpha/2)\%} : \qquad L\hat{\beta} \pm t_{1-\alpha/2,df=n-p}\sqrt{s^2L(X'X)^{-1}L'}$$

Doing this we get the following:

```
L2=matrix(c(0,1,-1),nrow=1)
glh2=glht(fit4,linfct=L2)
summary(glh2,test=Ftest())
confint(glh2)
...

Linear Hypotheses:
       Estimate lwr       upr
1 == 0  -3.0000 -11.0744   5.0744
```

Least squares Means

The use of a vector L allows for other estimates of interest and their CI's to be computed.

In the data, there are the mean responses for the three treatments - again we will often want to estimate the means for each treatment and get a 95% CI for them.

Consider the following three L vectors:

$$L_a = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$$

$$L_b = \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}$$

$$L_c = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}$$

what are each estimating?

In the context of a regression model - these are often called the **least squares means**.

Models including continuous and categorical predictors

Quite complex models can be achieved using combinations of continuous and categorical predictors.

Turkey data:

Age (weeks)	Weight (pounds)	Feed Type
28	13.3	a
20	8.9	a
32	15.1	a
22	10.4	a
29	13.1	b
27	12.4	b
28	13.2	b
26	11.8	b
21	11.5	c
27	14.2	c
29	15.4	c
23	13.1	c
25	13.8	c

The question here: How is weight related to age and feed?

Fit model:

$$E(y_{ij}) = \beta_0 + b_j + \beta(age_i)$$

where b_j is the categorical effect of feed type. The regression matrix for this model is;

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 28 \\ 1 & 1 & 0 & 0 & 20 \\ 1 & 1 & 0 & 0 & 32 \\ 1 & 1 & 0 & 0 & 22 \\ 1 & 0 & 1 & 0 & 29 \\ 1 & 0 & 1 & 0 & 27 \\ 1 & 0 & 1 & 0 & 28 \\ 1 & 0 & 1 & 0 & 26 \\ 1 & 0 & 0 & 1 & 21 \\ 1 & 0 & 0 & 1 & 27 \\ 1 & 0 & 0 & 1 & 29 \\ 1 & 0 & 0 & 1 & 23 \\ 1 & 0 & 0 & 1 & 25 \end{pmatrix} \qquad X^* = \begin{pmatrix} 1 & 0 & 0 & 28 \\ 1 & 0 & 0 & 20 \\ 1 & 0 & 0 & 32 \\ 1 & 0 & 0 & 22 \\ 1 & 1 & 0 & 29 \\ 1 & 1 & 0 & 27 \\ 1 & 1 & 0 & 28 \\ 1 & 1 & 0 & 26 \\ 1 & 0 & 1 & 21 \\ 1 & 0 & 1 & 27 \\ 1 & 0 & 1 & 29 \\ 1 & 0 & 1 & 23 \\ 1 & 0 & 1 & 25 \end{pmatrix}$$

```
> fit5=lm(weight~factor(feed)+age,data=turkey)
> summary(fit5)

Call:
lm(formula = weight ~ factor(feed) + age, data = turkey)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37353 -0.15294  0.01103  0.17868  0.47353

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.48750    0.67340  -0.724    0.487
factor(feed)b   -0.27353    0.21844  -1.252    0.242
factor(feed)c    1.91838    0.20180   9.506 5.45e-06 ***
age              0.48676    0.02574  18.908 1.49e-08 ***
---

Residual standard error: 0.3002 on 9 degrees of freedom
Multiple R-squared:  0.9794,  Adjusted R-squared:  0.9726
F-statistic: 142.8 on 3 and 9 DF,  p-value: 6.6e-08
```

```
> anova(fit5)
Analysis of Variance Table

Response: weight
             Df Sum Sq Mean Sq F value    Pr(>F)
factor(feed)  2  6.382   3.191  35.404 5.431e-05 ***
age           1 32.224  32.224 357.523 1.489e-08 ***
Residuals     9  0.811   0.090
---
```

Under the set to zero constraint imposed above, this model is reparameterised as;

$$E(y_{ij}) = \beta_0 + \delta_b\beta_1 + \delta_c\beta_2 + \beta_3(age_i)$$

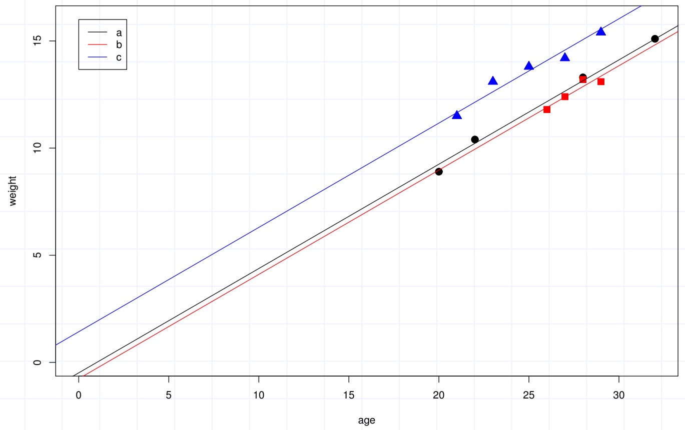
where δ_b is an indicator variable if the observation is for the turkey given feed b etc.

This is a regression model of weight on age, with a different intercept for each feed type but with a common slope β_3 for age across all feed types.

Feed	Model	age effect
a	$\beta_0 + \beta_3(age)$	$-0.49 + 0.49(age)$
b	$\beta_0 + \beta_1 + \beta_3(age)$	$-0.77 + 0.49(age)$
c	$\beta_0 + \beta_2 + \beta_3(age)$	$1.43 + 0.49(age)$

This type of model is sometimes called the **Common Slopes Model**

Common Slopes Model



Testing the Common Slopes model

- The Common Slopes model assumes that the relationship between age and weight is the same regardless of the feed used.
- It also assumes that this common relationship is a straight line - across the age range we are modelling.
- The alternative is that the either (a) the relationship is different depending on the feed used but still straight line, or (b) the relationship is not straight line, either for all feed types or for some feed types.
- Given the plot of these data, it seems likely that the relationship is approx. straight line for each feed type - but perhaps the slope may differ depending on feed type.
- One way of fitting such a model is to include an **interaction effect** of age and feed type.

Interaction of continuous and categorical predictors

Consider the following model:

$$E[y_{ij}] = \beta_0 + \delta_b\beta_1 + \delta_c\beta_2 + \beta_3(age) + \delta_b\beta_4(age) + \delta_c\beta_5(age)$$

i.e. different slopes (and intercepts) for each group.

This means that including an interaction allows for a different effect of age across the three different feed types.

```
fit5a=lm(weight~factor(feed)*age,data=turkey)
model.matrix(fit5a)
summary(fit5a)
anova(fit5a)
```

So, using this design matrix and the set to zero constraints, the model is effectively reparameterised as;

$$E[y_{ij}] = \begin{array}{l} \delta_a\left[(\beta_0) + (\beta_3)age\right] \\ \delta_b\left[(\beta_0+\beta_1) + (\beta_3+\beta_4)age\right] \\ \delta_c\left[(\beta_0+\beta_2) + (\beta_3+\beta_5)age\right] \end{array}$$

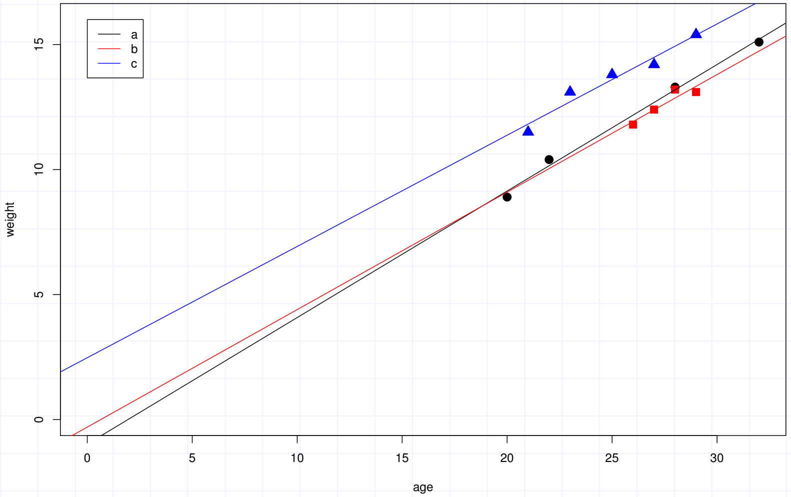
This is a model with a separate slope and intercept for the age effect for each feed type.

If we fail to reject the null hypothesis that $\beta_4 = \beta_5 = 0$ then we could conclude that the interaction is not significant and that the Common Slopes model was adequate.

If we reject the null hypothesis, then the we need to fit separate slope for each feed type.

```
> anova(fit5a)
Analysis of Variance Table

Response: weight
                 Df Sum Sq Mean Sq  F value    Pr(>F)
factor(feed)      2  6.382   3.191  31.6306 0.0003121 ***
age               1 32.224  32.224 319.4201 4.238e-07 ***
factor(feed):age  2  0.105   0.053   0.5204 0.6155896
Residuals         7  0.706   0.101
---
```


a
b
c
weight
0
5
10
15
age
0
5
10
15
20
25
30