Multiple Linear Regression

There is no difficulty in extending the idea to having more than one X in a regression

$Y = a + bX1 + c\ X2 + ...$

The least squares solution for a,b,c .. is a solution of a system of linear equations.
In practice use package.

The equation can still be used to predict Y, however standard errors of predictions involve complicated matrix formulae - need to use the package output for these : MINITAB is better at this than Data desk.

r^2 is still a valid criterion for deciding on the quality of the regression.

Residual plots should be used for model checking.

Equation development

A new issue arises in Multiple regression - which predictor X variables **really** influence Y?

A model should be parsimonious -using as few X variables as possible and yet reflect the data.

We want a simple model because:

1 -It will be easier to interpret not obscured by spurious relationships.

House prices in Dublin will be related to stocks of Cod in the North sea !

As the cod stocks have dwindled so the house prices have gone up.

2. The variance of the predictions improves if the model does not include spurious variables.

Thus we have the task of selecting an appropriate equation.

The basic tool is the **marginal t-test.**

All packages that do regression provide a test of each $\beta=0$.

This is of the form :

$$t = \frac{\text{Estimate of } \beta}{\text{s.e.}(\beta)}$$

A value of 2 (or -2) is used as the deciding criterion. (p-value 0.05)

Example:
Developing an equation for the price of a Car (Cars91 data).

In the initial model we include all possibly relevant variables.

Dependent variable is: Price
No Selector

R squared = 73.1% R squared (adjusted) = 71.5%
s = 4756 with 91 - 6 = 85 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	5.22351e9	5	1.0447e9	46.2
Residual	1.92267e9	85	2.26197e7	

Variable	Coefficient	s.e.	t-ratio	prob
Constant	-19395.	6878	-2.82	0.0060
Weight	6.89128	2.546	2.71	0.0082
Horsepower	117.728	23.46	5.02	< 0.0001
Eng. Displc	-36.4007	25.66	-1.42	0.1597
Cylinders	1006.49	949.8	1.06	0.2923
Drive ratio	346.258	1407	0.246	0.8062

r^2 is good at 73%, The last three variables are not significant (prob>0.05) so we drop them from the model.

We note though that dropping variables will in general alter the p-values of the others.

Drive ratio and Cylinders and Eng disp are a long way from 0.05 so we drop all these.

Dependent variable is: Price
No Selector

R squared = 72.4% R squared (adjusted) = 71.7%
s = 4738 with 91 - 3 = 88 degrees of freedom

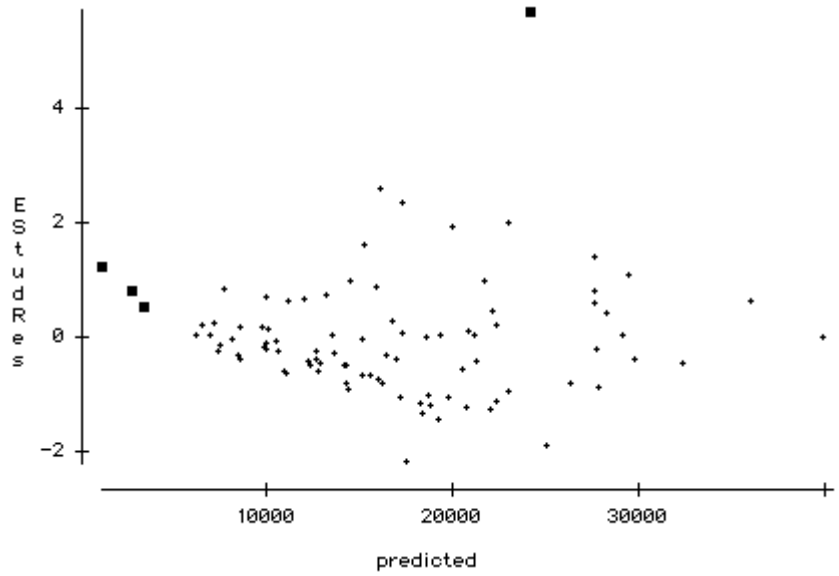
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	5.17093e9	2	2.58546e9	115
Residual	1.97526e9	88	2.24461e7	

Variable	Coef	s.e.	t-ratio	prob
Constant	-13811	3099	-4.46	<0.0001
Weight	4.70558	1.732	2.72	0.0079
Horsepower	125.699	21.6	5.82	< 0.0001

Of the variables available, Weight and Horse power are all that is needed to predict Price.
Note that the have been some changes in the coefs. when the non-significant variables were dropped.
r^2 is still very good at 72%
The equation is:

Price $ = -13811 +4.706*Weight +125.7*HP

Residual Plot



The highlighted points are troubling ...

The Merc (check) stands out as an extremely expensive car. This car does not fit the current model and should be excluded from the data.

The three cars with the lowest predicted prices all have positive residuals suggesting that the equation may not work for very cheap cars (less than $3500).

Excluding the merc gives:

Dependent variable is: Price
No Selector
91 total cases of which 1 is missing

R squared = 76.8% R squared (adjusted) = 76.3%
s = 4065 with 90 - 3 = 87 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	4.7716e9	2	2.3858e9	144
Residual	1.43755e9	87	1.65236e7	

Variable	Coefficient	s.e.	t-ratio	prob
Constant	-13266.9	2661	-4.99	<0.0001
Weight	4.72278	1.486	3.18	0.0021
Horsepower	119.336	18.57	6.43	< 0.0001