Lecture 15
Regression II

Statistical properties:

Assumptions:
(1) E(Y_i) =$\alpha + \beta x_i$ - the relationship is linear.
(2) V(Y_i)=V(e_i)=σ^2 does not depend on i (X or Y).
(3) e_i independent.
(4) X_i measured without error.

r_i – residual , $\hat{Y}_i$ =a+bX_i - fitted value (a,b estimates)

$$r_i = Y_i - \hat{Y}_i$$

$$\sum r_i = 0, \quad \sum X_i r_i = 0.$$

Results:

E(b) = β E($\hat{Y}$) = $\alpha + \beta x_i$

V(b) =$\dfrac{\sigma^2}{SSX}$ V($\hat{Y}$)=$(\dfrac{1}{n} + \dfrac{(x_h - \bar{x})^2}{SSX})\sigma^2$

The expressions for E break down if assumptions(1) or (4) violated.
The failure of any of the assumptions makes the expressions for variance invalid.

Distributional assumption

$$e_i \sim N(0,\sigma^2) \quad (Y_i - \text{normally distributed})$$

As b and $\hat{Y}$ are linear combinations of Y they are also Normal with the means and variances above.

This allows us to construct confidence intervals and conduct tests.

The variance is estimated by:

$$\hat{\sigma}^2 = \frac{1}{n-2}\text{SSE}$$

Where SSE the residual sum of squares = $\sum r_i^2$
The t-distribution is used for these the df=n-2.

Example: Gibbs Sampler data

$$\sum xy = 61800, \sum x = 500, \sum y = 1100, \sum x^2 = 28400, n = 10$$

$$b = \frac{61800 - 10*50*110}{28400 - 10*2500} = 2$$

$$a = 110 - 2*50 = 10$$

SSX = 28400 – 10*2500 = 3400

To compute SSE you also need $\sum y^2$.

The formula is not worth remembering here SSE =60.
Confidence interval for the slope – the man-hours required for 1 unit.

b = 2

$\hat{\sigma}^2 = \frac{60}{8}$=7.5, V(b) =$\frac{7.5}{3400}$=0.0022, SD(b)=0.047

TINV(0.05,8)=2.31

b = 2 +/- 0.108

Test β = 0 (no relationship (linear) between job size and man-hours – clearly rejected.

Average time to complete a 60 parameter job.

$\hat{y}$(60)= 10 + 2*60 =130

Confidence interval:

$$130 \text{ +/- } 2.31 * \sqrt{7.5\left(\frac{1}{10} + \frac{(60-50)^2}{3400}\right)}$$

130 +/- 2.31 * 0.9852

Note there were 3 jobs run with 60 parameters 128, 132, 135 respectively.

$\bar{y} = 131.67$ not the same.

Using just these 3 observations :

CI 131.67 +/- 3.182 * 2.028 much wider.

As well as the t value being bigger the estimate of variance is bigger (accident) but also the multiplier in the regression estimate is:

$$\sqrt{\left(\frac{1}{10}\right) + \frac{100}{3400}} = 0.360 \text{ as compared to } \frac{1}{\sqrt{3}} = 0.577$$

for the simple mean of 3 observations.

The gain in precision comes from the fact that we use all 10 observations to get the estimate in the regression case.

But the regression estimate is only valid if the regression assumptions (linearity, const variance) are met. For the simple mean we only require data to be Normal.

Prediction Interval

The standard error that we associated with the prediction above comes from the uncertainty of the estimates a and b.
If am going to fulfil an order of 60 units I might want to know the confidence interval for that particular run.
The model we have assumed says this will take a time Y that is N(μ,σ^2). Where μ is estimated by 10 + 2*60 a particular run has a Normal distribution around this mean. Y = $\hat{Y}$ + N(0,σ^2). So if I want a reasonable range for the next run with 60 parameters I have an additional σ^2 to account for.

This is called a *prediction interval*:

The half width is : $t * \hat{\sigma} * \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSX}}$

The estimate itself being $\hat{Y}$ because my best guess what the next value of Y will be is its average.

Regression statistics and checking the assumptions.

The summaries that we get from the regression are three sums of squares:

Error

SSE – Sum of squares Error= residual sum of squares. The unexplained (remaining) variation in Y. The variation around the line.

SSTO – The total sum of squares $\sum (y_i - \bar{y})^2$. The variation in Y not allowing for X. What SSE would be if b=0.

SSR – Sum of Squares due to regression. The variation (in Y) explained by X.

SSR = SSTO – SSE.

A simple indicator of the efficacy of the regression is:

coefficient of determination = $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$

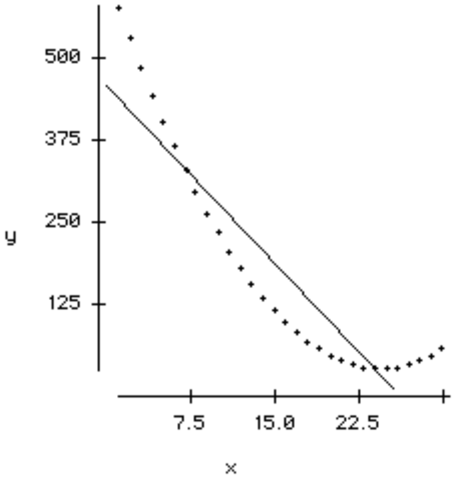
The proportion of explained variation.

Checking assumptions

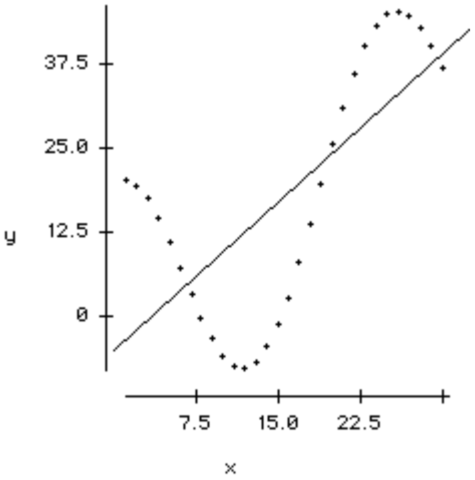
None of the formulas, tests etc are worth much if the regression model is inappropriate.
The algorithm doesn't check anything it just gives the best answers it can.

Here I have fitted regressions to some non-linear functions.

$y = x + (x-25)^2$

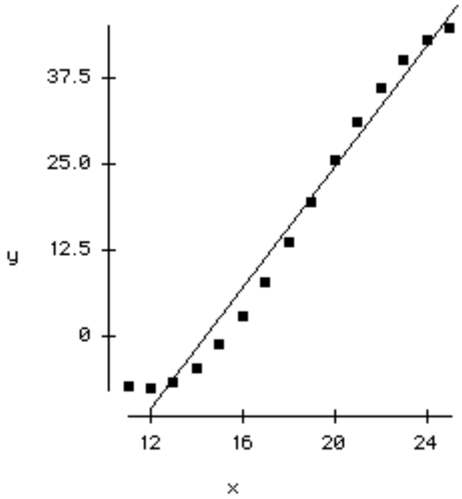


$y = x+20*\cos(x/4)$



Clearly there is more going on here than just a straight line.

Worse
In the second case suppose I only observed data X=11 to X=25



The regression is nearly a perfect fit. So confidently I use the equation to predict at x=5

Data desk output for this is:

```
Dependent variable is:    y
cases selected according to     Selected Data
30 total cases of which 15 are missing
R squared = 97.4%     R squared (adjusted) = 97.2%
s =  3.335  with  15 - 2 = 13  degrees of freedom

Source     Sum of Squares       df    Mean Square  F-ratio
Regression       5382.25        1     5382.25          484
Residual         144.564        13    11.1203

Variable  Coefficient      s.e. C       t-ratio     prob
Constant  -63.0688         3.689        -17.1   < 0.0001
x         4.38433          0.1993         22    < 0.0001
```

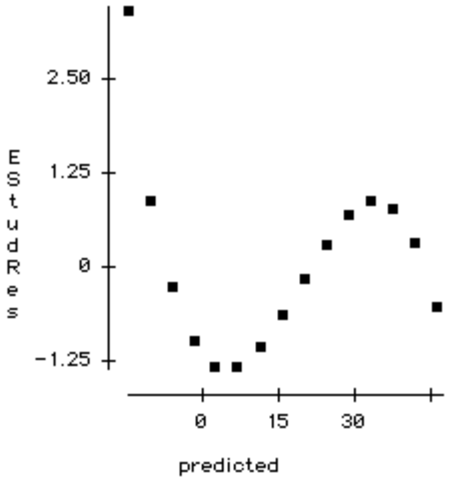
No indication anything is wrong !
So confidently Y(5) = -63.1 + 4.38 * 5 = -40.2

The actual value is +11.3

The main problem here is that the relationship is non-linear.

And I could have detected it (in the reduced data)

Residual Plot:



If the assumptions are met this plot is supposed to show no systematic pattern - a random scatter of points – it clearly does have a pattern.

The tools that we use are graphical.

The scatter Plot Y vs X

This will show up gross departures from linearity, and possibly non constant variance.

The residual plot: r vs $\hat{Y}$

Much more sensitive at detecting non-linearity,
observations that do not conform to the model,
and non-constant variance.
Often residuals are scaled (by their standard deviations) this gives us an absolute reference

|Estudres| > 2.5 is big (compared to N(0,1)).

Probability Plot of residuals

Checks the Normality assumption.

The Scatter plot and probability plot are straight forward to interpret.
The best of them is the residual plot.