Lecture 17

Extending regression Transformations. Indicator variables.

The regression equation is a linear equation; we can however transform the variables to accommodate other relationship.

Suppose we want to fit the relationship:

$$Y = \alpha e^{\beta t}$$

This is achieved by transforming the Y variable:

$$\log(Y) = \alpha + \beta X .$$

The usual regression assumptions are assumed to hold for the transformed variables.

The full model is

$$\log(Y) = \alpha + \beta x + \varepsilon$$

Where ε are $N(0,\sigma^2)$.

Note: this means that for the original model we have a multiplicative error term (δ).

$$Y = \alpha e^{\beta x} \delta$$

δ has a log-normal distribution $(0,\infty)$ – the distribution so that its log() is Normal.

If we wanted to have the error to be additive

$$Y = \alpha e^{\beta x} + \varepsilon$$

This would be a non-linear least squares problem a much harder algorithm to minimise the sum of residuals . (iterative).

Fortunately it often makes sense in practice to have multiplicative error with such a relationship.

Example : Such a model is often used for growth of a biological entity.

Y= size (say weight)

x = time.

Having a multiplicative error means that the percentage variation is constant – that makes physical – large individuals vary more in weight than small ones in absolute units.

Other model sometimes used are:

$\log(Y) = \alpha + \beta \log(x)$ corresponding to a power relationship:

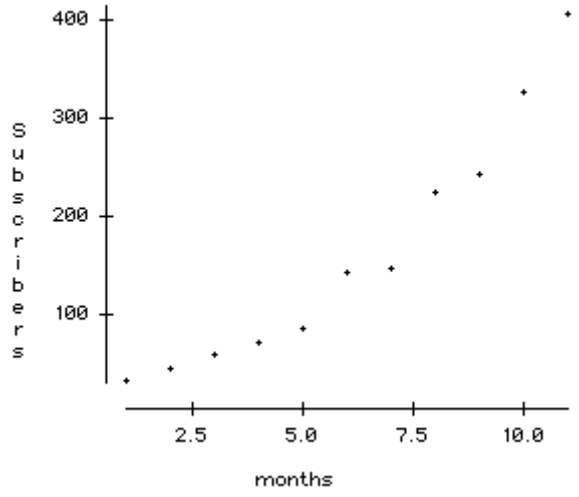
$$Y = \alpha X^{\beta}$$

$\frac{1}{Y} = \alpha + \beta x$ and $Y = \alpha + \beta \frac{1}{x}$ for reciprocals

The key issue about transformed models is that fitting and inference takes place in the transformed variable space.

Example
Broadband has hit the island of San Seriffe. The local telephone company has monthly records of the number of subscribers to broadband.

months	Subscribers
1	31
2	44
3	58
4	71
5	85
6	144
7	148
8	226
9	244
10	328
11	408



It is reasonably clear that the relationship is a curve. There is some evidence that the "adoption of new technologies" follows an exponential curve, at least in the early stages.

We want to fit the model :

$$\text{Subscribers} = \alpha e^{\beta months}$$

This is equivalent to:

$$\log(Subscribers) = a + b * months$$

Dependent variable is: ln(s)
No Selector

R squared = 99.0% R squared (adjusted) = 98.9%
s = 0.08973 with 11 - 2 = 9 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	7.17937	1	7.17937	892
Residual	0.0724632	9	0.00805147	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	3.25561	0.05803	56.1	< 0.0001
months	0.255474	0.008555	29.9	< 0.0001

Thus: a=3.25561 and b=0.255474

Transforming back to the original variables:

$$\alpha = e^{a} = 25.93$$

$$\beta = e^{\beta} = 1.291$$

Meaning : α size at time 0 i.e. about 26 subscribers initially

β multiplication factor for each unit of x (=month)

Thus the number of subscribers is increasing at 29.1% each month.

Confidence interval for the monthly increase:

Confidence interval for the b coefficient:

$$0.255474 \pm 2 * 0.008555$$

(0.238364, 0.272584)

We get the confidence interval for β coefficient by exponentiating the ends .

(1.269171, 1.313354)

Thus the monthly growth rate is estimated to be between 26.9% to 31.3%

12 month prediction:
We obtain the estimate for the ln(subscribers) for months=12

3.25561 + 12*0.255474 = 6.321298
corresponding to 556.29 subscribers.

To construct a confidence interval for this prediction we need $\bar{X}$ and SSX . The regression output does not provide these values.
The Data Desk Summary Reports provides:

Summary of months
No Selector

Count	11	
Sum	66	
Mean	6	
SSQ	506	(sum of squares)
Variance	11	

So

$$\bar{X} = 6$$

$$SSX = \sum X^2 - n\bar{X}^2$$

$$SSX = 506 - 11*36 = 110$$

The s.e. for a prediction at x is:

$$s\sqrt{\frac{1}{n}+\frac{(x-\bar{X})^2}{SSX}} = 0.08973*\sqrt{\frac{1}{11}+\frac{(12-6)^2}{110}} = 0.05826$$

Thus

$6.321298 \pm 2 * 0.05826$ for log subscribers.

(425, 625) subscribers after 12 months.

We could use the model to produce predictions for

24 months = 11932
36 months = 255934 exceeds the total population of San Seriffe .

These are well beyond the observed data and rely on the model holding into the future – which it will not.

In Multiple regression we have the additional possibility that the individual Xs can be constructed variables.

Instead of fitting an exponential model to the data above we might use a polynomial in time:

$$Subs = \beta_0 + \beta_1 month + \beta_2 month^2 ...$$

We can create the powers of month using "Derived variables" in data desk.

Linear R squared = 91.3%

Variable	Coefficient	s.e.	t-ratio	prob
Constant	-53.1091	25.08	-2.12	0.0633
months	35.9273	3.698	9.72	<0.0001

Quadratic R squared = 99.0%

Variable	Coefficient	s.e. o	t-ratio	prob
Constant	43.7394	15.68	2.79	0.0236
months	-8.77203	6.006	-1.46	0.1823
m2	3.72494	0.4875	7.64	<0.0001

Cubic R squared = 99.1%

Variable	Coefficient	s.e.	t-ratio	prob
Constant	21.1061	24.74	0.853	0.4218
months	9.92327	17.07	0.581	0.5793
m2	-0.00582751	3.234	-0.0018	0.9986
m3	0.207265	0.1777	1.17	0.2817

The problem is that months, months^2, months^3 are similar variables and least squares cannot disentangle their individual contributions

Indicator variables
A very important class of variables are indicator variables and variables derived from them.

If condition 1 then Z1 =1 else Z1 =0
If condition 2 then Z2=1 else Z2 =0

Example
Two communication protocols XC23 and YD28 may be used for transferring files on a certain system. Test files of various sizes were transferred using the different protocols resulting in the following data:

Time to transfer, File Size, Protocol Used

Time	size kb	protocol
5.29	65	1
1.56	25	1
1.64	34	1
2.83	49	1
2.21	32	1
4.89	90	1
7.37	140	1
6.21	120	1
6.37	118	1
4.12	90	1
3.55	70	1
0.87	16	2
0.97	18	2
2.11	48	2
2.19	43	2
4.27	89	2
3.35	65	2
3.54	72	2
2.37	52	2
7.87	160	2

Code Z =1 if YD28 else 0.

We now consider 4 models for this data:

Model 0.

$time = \beta_0 + \beta_1 size$ - this model assumes no differences between protocols

β_0 -set up time, β_1 - secs/kb

Model 1.

$time = \beta_0 + \beta_1 size + \beta_2 Z$

For protocol XC23 this model is:

$time = \beta_0 + \beta_1 size$

For protocol YD28 it is:

$time = (\beta_0 + \beta_2) + \beta_1 size$

the set-up time for protocol 2 is $\beta_0 + \beta_2$ the time to transfer 1kb is *assumed* the same for both protocols.

We now create another artificial variable

W = Z*size

Thus W = size if YD28

= 0 if XC23

Model 2

$time = \beta_0 + \beta_1 size + \beta_3 W$

For XC23 $time = \beta_0 + \beta_1 size$

For YD28 $time = \beta_0 + (\beta_1 + \beta_3) size$

W and size have identical values if YD28.

β_3 measures the difference in slopes (secs/kb), set-up time is the same for both protocols.

Model 3

$time = \beta_0 + \beta_1 size + \beta_2 Z + \beta_3 W$

This allows for different lines – different set-up times and different transfer rates for the two protocols.

Model 0 is obtained by setting β_2 and β_3 to 0 in Model 3

Can compare using F-test

Similarly can compare 3 with 1 and 2, 2 with 0, 3 with 0

Cannot compare 2 and 3 using F-test as not nested.

Dependent variable is: Time
No Selector

R squared = 99.2% R squared (adjusted) = 99.0%
s = 0.2433 with 20 - 4 = 16 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	110.769	3	36.9229	624
Residual	0.947009	16	0.0591881	

Variable	Coefficient	s.e.	t-ratio	prob
Constant	2.20584	0.1657	13.3	<0.0001
size kb	0.050151	0.001962	25.6	<0.0001
Z	-0.218204	0.2221	-0.982	0.3405
W	0.0186435	0.002785	6.69	<0.0001

As the coefficient of Z is not significant we conclude:

Set-up time the same for both protocols 2.20 s
time to transfer 1kb using XC23 is estimated at 0.0502s
time to transfer 1kb using YD28 is 0.01864s longer i.e. 0.06866s