Logs Transformation in a Regression Equation

Logs as the Predictor

The interpretation of the slope and intercept in a regression change when the predictor (X) is put on a log scale. In this case, the intercept is the expected value of the response when the predictor is 1, and the slope measures the expected change in the response when the predictor increases by a fixed percentage.

These properties of the regression equation are most clear in the context of an example, such as the display example from the casebook. In that example, the *estimated* least squares regression equation is

Sales = $84 + 139 \log(\text{Feet})$

To interpret the intercept 84 in this equation, we need to remove the term involving the slope. If the number of feet is 1, then the estimated equation becomes

Sales =
$$84 + 139 \log(1) = 84 + 139 (0) = 84$$

So, as promised, the intercept is the expected level of sales (here, \$84) when the number of feet used in the display is set to 1.

It's a little harder to figure out the meaning of the slope. One method (as in the casebook and repeated in class) is to consider the effect of small *percentage* changes in the predictor. For example, suppose we start with 2 feet of product on display, with an expected level of sales of

Sales (2 feet) =
$$84 + 139 \log(2)$$

If we now increase this by 1% (a quite small change), then the expected sales grows by

Sales(2 (1.01) feet) = $84 + 139 \log(2 (1.01))$ = $84 + 139 \log(2) + 139 \log (1.01)$ = Sales(2 feet) + 139 log 1.01 = Sales(2 feet) + 1.39

That is, we expect sales to increase by \$1.39 for every 1% increase in display footage. A 1% increase in the predictor leads to a change in the response of 1% of the value of the slope.

Said differently, when we use a log scale for the predictor, we are saying that a given percentage change in the predictor has the same impact on the response. Every time we increase the predictor by, say 20%, we expect the same change on average in the response.

For x feet on display:	Sales (x) = $84 + 139 \log x$
For 20% more on display:	Sales(1.2 x) = $84 + 139 \log (1.2 x)$ = $84 + 139 \log x + 139 \log 1.2$

Every time we increase the footage by 20%, we expect to see sales increase on average by 139 $\log 1.2 = 25.34 . In contrast, when we use a *linear* model, we are saying that a given fixed change in the value of the predictor has the same impact.

Logs as the Response

What happens when the response is on a log scale, but the predictor is expressed in the original units? Here is an example. The response is the value of a used car (expressed in thousands of dollars) and the predictor is the age of the car. The transformed model in this figure uses a log of the response and the age.



The fitted (or estimated) regression equation is

Log(Value) = 3.03 - 0.2 Age

The intercept is pretty easy to figure out. It gives the estimated value of the response (now on a log scale) when the age is zero. We would estimate the value of a "new" Accord (foolish using only data from used Accords) as

$$Log(Value for Age=0) = 3.03$$

so that the value itself would be about $e^{3.03} = 20.7 thousand, which you can check from the plot.

For the slope, again look for the percentage change, but now in the response. In this case, changes of the age in years will produce *percentage* changes in the value. For example, from this model, the value (on average) of a used Accord drop 20% for every additional year of age. The next paragraph explains this interpretation.

To arrive at this interpretation, recall that the slope tells us how changes in the predictor affect the response. Since the predictor is in the original units of the problem, we begin by seeing what happens to the estimated value for a one year change.

Value at age x years:	Log(Value at x) = 3.03 - 0.2 x
at age x+1:	Log(Value at $x+1$) = $3.03 - 0.2$ ($x+1$)

Subtracting the two equations, the change in the value for each year of aging is

Log(Value at age x+1) - Log(Value at age x) = -0.2

Combining the two log terms gives (using the property $\log a - \log b = \log a/b$)

Log((Value at x+1)/(Value at x)) = -0.2

If we think about the value at age x as the value at the older age plus a bit more, Value at age x+1 = V alue at age x + (change in value),

then

 $\begin{array}{ll} -0.2 &= Log((Value at x+1)/(Value at x)) \\ &= Log(((Value at x) + (change in value))/(Value at x)) \\ &= Log(1 + (change in value)/(Value at x)) \\ &\approx (change in value) /(Value at x) = percentage change \end{array}$

approximately for small changes. That's *about* a 20% drop for each year of aging. It's only approximate since this is a rather large change, and the approximation for $\log(1+x) \approx x$ only works for small x's.

If you check these calculations directly, you get a similar value. For example, at Age 1 year we find a value estimated to be

Value at age 1 = \$16.9 (thousand) at 2 = \$13.9at 3 = \$11.4

The drop from age 1 to age 2 is 3/16.9 or 18% and from age 2 to 3 is 2.5/13.9 which is also 18%. In general, it is easier (and quicker) to approximate this effect from the slope directly. (Note that if you do the percentage change using the smaller value in the bottom, you get changes of 22% per year – the slope gives a sort of "average" between these two ways to compute the percentage change.)

Logs as the Response: Another Example

Logs of the response are often used to model time trends as well. Here is an example. In the following plot, it appears the growth is not linear, but rather faster than linear.



From the following summary, the linear model implies constant growth of sales of 0.26 per time period (the slope from the fitted model).

Parameter	Estimates					
Term	Estimate	Std	Error	t Ratio	Prob> t	
Intercept	1.04		0.32	3.28	0.0019	
Time	0.26		0.01	23.95	<.0001	

In contrast, the alternative model using the log of sales as the response implies growth of about 3.6% per time period, a model with compound growth.

Parameter	Estimates					
Term	Estimate	Std Error	t Ratio	Prob> t		
Intercept	0.9909	0.0231	42.83	<.0001		
Time	0.0356	0.0008	45.05	<.0001		

To sort out this interpretation with a log transformed response, we can work through the details again as in the example with the decline in value of a car. The transformed model implies that

Ave $(\log(\text{Sales}_t) | \text{time}) = 1 + 0.034 \text{ time}$ = 1 + 0.034 t

To interpret slope value, compare the sales in adjacent periods as before.

$$log(Sales_t) = 1 + 0.034 t$$

log(Sales_{t+1}) = 1 + 0.034 (t+1)

so that

 $log(Sales_{t+1}) - log(Sales_t) = 0.034$

Now simplify the left side,

$$log(Sales_{t+1}) - log(Sales_t) = log(Sales_{t+1} / Sales_t)$$

= log (1 + (Sales_{t+1} - Sales_t) / Sales_t)
\$\approx (Sales_{t+1} - Sales_t) / Sales_t\$

as long as the changes are "small" relative to past values. Thus we have shown that on average, sales increase 3.4% per period.

Logs as the Predictor and the Response

In this case, the coefficient is known as an elasticity. Elasticities are described (albeit in a multiple regression) in the casebook (pages 148-151).