Econometrics

Chapter 8: Nonlinear Regression Functions

Maybe the effect of X_1 on Y depends on the value of X_1 or maybe depends on some other variable. How to deal with this.

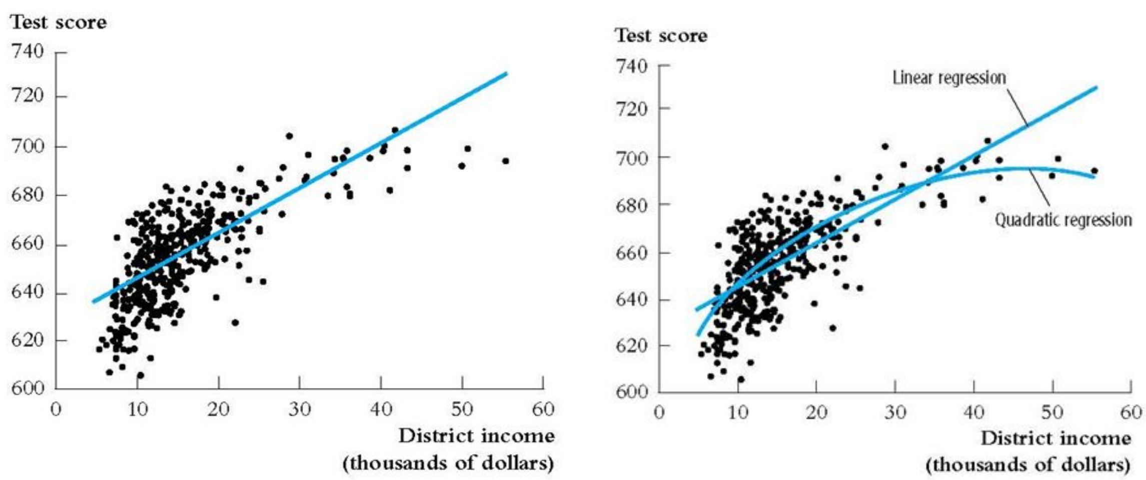
In this case the linear model is 'mis-specified' - the functional form is wrong.

Your estimates will not be correct (not even right on average likely).

Polynomial Terms

Example 1: quadratic effect - many things have large effects at low levels that then taper off.

Answer: add a X which is the square of the original $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ and just run OLS



(a) Maybe not Linear

(b) Maybe Quadratic

Figure 1: Polynomial Terms

$$\hat{TestScore} = \underset{(2.9)}{607} + \underset{(0.27)}{3.85} \cdot Income - \underset{(0.005)}{0.042} \cdot Income^2$$

From here on just like before, just treat the new variable as any other.

Effect on Y from change in X - no longer the simple coefficient.
So could take derivative or simple $f(X + \Delta X) - f(X)$

Ex: $\frac{dY}{dX} = 3.85 - 0.084 \cdot X$ - so we see the effect of X on Y changes as X changes - it decreases.

But! Don't take this too far. Do not extrapolate the effect outside the range of your data.

How many polynomial terms to use?

Any set of data points can be fit arbitrarily close with enough polynomial terms.

Could see by using the polynomials and then running f-test on the non-linear terms. Or judgement.

Logarithms

Could be of Y, of X, or both - and here we mean natural logarithm

Exponential function of x = e^x where is is the constant 2.718....., or: $exp(x)$

Natural log is the inverse of this function: $x = ln(e^x)$. Written $y = ln(x)$, Slope is 1/x.

Properties:
(1)$ln(1/x) = -ln(x)$; (2) $ln(ax) = ln(a) + ln(x)$; (3) $ln(x/a) = ln(x) - ln(a)$; (4) $ln(x^a) = aln(x)$

Link between logarithm and percentages: when Δx is small then $ln(x + \Delta x) - ln(x) \approx \frac{\Delta x}{x}$

$$\begin{aligned}
\%\Delta &= \frac{(x+\Delta x) - x}{x} \qquad \text{(ignoring the 100)} \\
&= \frac{x + \Delta x}{x} - 1 \\
\rightarrow & \frac{x+\Delta x}{x} = \%\Delta + 1 \\
\rightarrow & ln(x + \Delta x) - ln(x) = ln(\%\Delta + 1)
\end{aligned}$$

Note that derivative (slope) of ln(x) = 1/x.
And when %Δ is small this implies the slope (and so change in 'y') of ln(%Δ + 1) is close to 1.
And ln(1) is zero so ln(%Δ + 1) is close to %Δ.

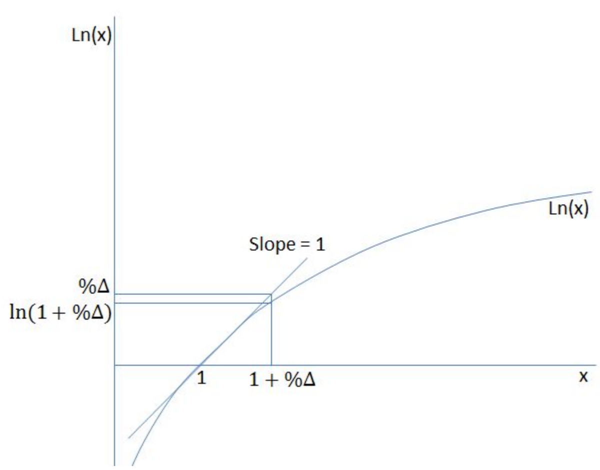


Figure 2: $ln(x + \Delta x) - ln(x) = ln(\%\Delta + 1) \approx \%\Delta$

So $ln(x + \Delta x) - ln(x) \approx \frac{(x+\Delta x)-x}{x} = \%\Delta$ (again this is with the x100 dropped).

Three Cases

1. X is in logs, Y is not

Call this the linear-log model

So what is the change in Y due to small change in X: $\beta_1[ln(X + \Delta x) - ln(X)] \approx \beta_1(\frac{\Delta X}{X})$

So a 1% change in X changes Y by about $\beta_1(0.01X/X) = 0.01\beta_1$

Nothing else different - just redefine your X before running the regression

Consider income effect on test scores. If in log's then it means at income of \$10,000 a \$1,000 increase has a much larger impact then at \$100,000. In the first the effect on Y is $0.1\beta_1$, in the second only $0.01\beta_1$. Similar in a sense to the quadratic.

2. Y is in logs X is not

Call this the log-linear model

In this case a one-unit change in X gives a $(100 \times \beta_1)$% change in Y.

Remember when X changes by one unit ln(Y) changes by β_1.

But changes is ln(Y) can be approximated by $\frac{\Delta Y}{Y}$ which can be written as: $\frac{\%change}{100}$.
So $\frac{\%changeinY}{100} \approx \beta_1$ or % change in Y $\approx 100 \times \beta_1$.

For example at many jobs get a (relatively) constant % increase in wages every year. So makes sense to estimate in log-linear (just your coefficient will be 100 x smaller).

3. Both are in logs

Call this the log-log

Here a 1 % change in X causes a β_1 % change in Y. So β_1 is the elasticity of Y with respect to X.

Which to use? Should base on economic knowledge.

Caution on interpreting the coefficients on a dummy variable in a log-linear model.

You cannot just do what we did above - because you can't take a derivative of ln(Y) with respect to a 0-1 variable.

To correctly do this note: $ln(y) = \beta_0 + \beta_1 D \rightarrow Y = exp(\beta_0)exp(\beta_1 D)$

So if D goes from 0 to 1 then $\frac{Y_1 - Y_0}{Y_0} = \frac{exp(\beta_0)exp(\beta_1 D) - exp(\beta_0)}{exp(\beta_0)} = exp(\beta_1) - 1$

and the % change in Y due to D going from 0 to 1 is $100(exp(\beta_1) - 1)$

and similarly can show the % change in Y do to D going from 1 to 0 is $100(exp(-\beta_1) - 1)$

Moreover these might be very different than $100\beta_1$

But there is one more problem. Since these conversions are non linear, though our estimate of β is unbiased this change is biased (due to Jensen's inequality). $E(\hat{\beta}) = \beta$ does not mean that $E[f(\hat{\beta})] = f(\beta)$

So we want to use a correction:

% change in Y due to D going from 0 to 1 $\approx 100[exp(\beta_1 - 0.5\sigma^2_{\beta_1}) - 1]$ where $\sigma^2_{\beta_1}$ is just the square of the SE for that coefficient that your program will give you.

Interaction Terms

1. Interacting two binary variables

Say looking at earnings and the effect of gender and having a college degree

This forces the effect of the college degree to be the same for both genders, maybe this is not so. Solution: add an interaction effect.

$Y = \beta_0 + \beta_1 Sex + \beta_2 College + \beta_3(Sex \times College) + u$

So if sex is 1 for male and college is 1, then the effect of college on women is β_2 and the effect on men is $\beta_2 + \beta_3$.

2. Interacting a continuous and a binary variable

Just as above but now the slop is different. Perhaps looking at effect of experience on wages and interact experience with gender.

3. Two continuous variables

No change, just be careful of interpretations.

In general always include main effects

Standard Error of Change

Though you rarely see this, still might be useful.

Say have some estimated change in Y from X increasing when X = 20 in a quadratic model: $\frac{dY}{dX} = \beta_1 + 2 \cdot \beta_2 \cdot 20$.

You can easily plug in the numbers and get an estimate of the expected change. But what about the SE of this change?

Well we could use an F-test. If we recall that an F-stat is a t-stat squared.

So run an F-test on the restriction $\beta_1 + 2 \cdot \beta_2 \cdot 20 = 0$. Now solve for the SE noting that :

$F = t^2 \rightarrow F = \frac{|\Delta \hat{Y}|}{SE(\Delta \hat{Y})} \rightarrow SE(\hat{Y}) = \frac{|\Delta \hat{Y}|}{\sqrt{F}}$