HANDOUT 8 – QUADRATIC AND LOGARITHMS

AGENDA

Non-linear regression functions

- Introduction
- Quadratics
- Logarithms
  - Log-level case
  - Log-log case
- Takeaways
- Vocabulary:
- Bonus:
  - Summary of logarithmic models
  - Logarithms review

BIBLIOGRAPHY FOR TODAY´S CLASS

- Ozimek (2017). Reducing immigration won't help places with the most Trump support. (*)
- Stock and Watson (2007), 8.1 (up to page 259), 8.2 (**)

INTRODUCTION

We continue our study of nonlinear relationships:

- Last class we considered binary dependent variables: We used the linear probability model (LPM) and the non-linear probit model
- Today we consider nonlinear relationships with a continuous dependent variable

QUADRATICS

Think back to our California class size, income and test score example

- Consider the multivariate regression we ran:

$$testscr = \beta_0 + \beta_1 str + \beta_2 income + \varepsilon$$

Interpretation:

- An additional $1,000 in income is associated with β2 change in test scores... regardless of whether that $1,000 is given to any family
- This linearity assumption may be unrealistic.
- Non-linear regression functions allow the predicted change in Y associated with a change in X to vary with X
- We will use quadratic and logarithmic regressions to fit curvature to data

A regression can fit any relationship between Y and X by adding higher order terms of X (like X^2, X^3, etc) to the right-hand side.

The quadratic (squared) regression function is written as:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

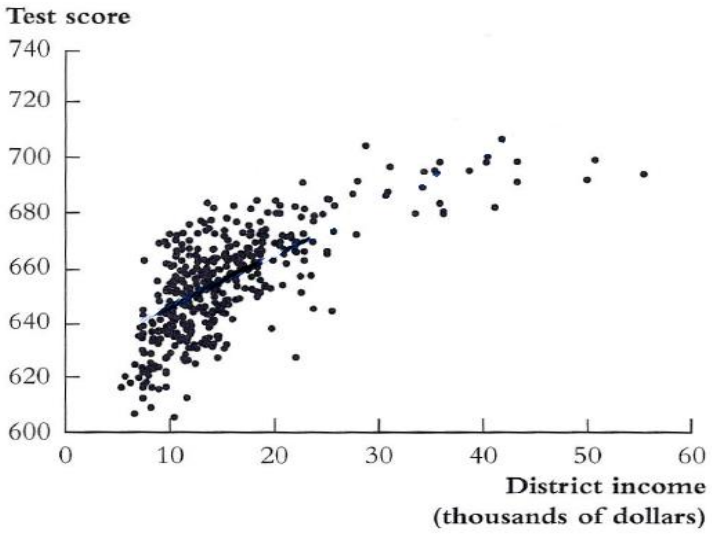
Quadratics functions fit U-shaped patterns: The coefficients determine:

- Where the U is centered.
- How flat or steep it is.
- Whether it faces up or down.

But β_1 no longer Represents the predicted change in Y associated with a one-unit change in X because now **the predicted change in Y associated with a one-unit change in X also depends on the level of X.**

Let's return to the California test score data.

We are interested in the relationship between school district average test scores and average income in that district (in thousands of dollars).



We could estimate this linear population regression function:

$$TestScore_i = \beta_0 + \beta_1 Income_i + u_i$$

…which would yield the sample regression function below:

Income	*TestScore* = 625.4 + 1.9*Income*	Change in $\widehat{testscore}$
10		
11		
40		
41		

The relationship between test scores and income does not appear to be linear.

Adding a quadratic term in the regression allows us to model the curvature observed in the data:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i$$

Just generate a squared version of the income variable and then use OLS:

```
. gen income2 = income*income

. reg testscore income income2, robust
```

testscore	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	3.850995	.2680941	14.36	0.000	3.32401	4.377979
income2	-.0423085	.0047803	-8.85	0.000	-.051705	-.0329119
_cons	607.3017	2.901754	209.29	0.000	601.5978	613.0056

The SRF is therefore:

$$\widehat{TestScore} = 607.3 + 3.9 Income - 0.04 Income^2$$

To see how predicted test scores change with a one unit change in income, we can no longer use only the coefficient on Income.

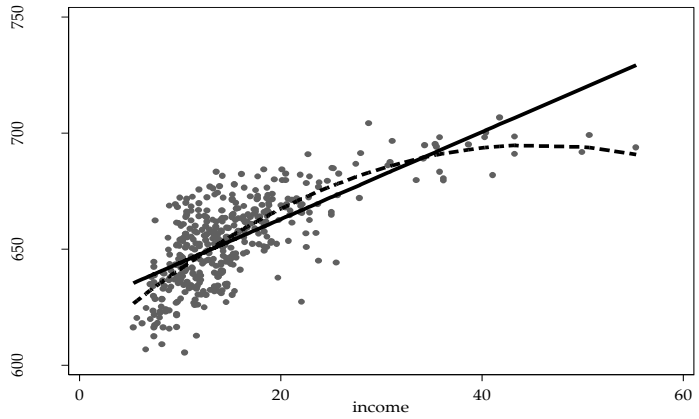
We need to specify a particular level of income because now the relationship between X and Y depends on the level of X

We must calculate this at each level of X we are interested in:

Income	$TestScore = 607.3 + 3.9Income - 0.04Income^2$	Change in $\hat{testscore}$
10		
11		
40		
41		

Is the marginal association of income with test scores increasing or decreasing?

The graph here shows the data, along with the fitted linear and quadratic curves estimated by the two regressions we ran above:



What are the big differences between these two models?

In which areas of the X spectrum are we most likely to err if we use linear?

LOGARITHMS

Logarithms are another way to allow curvature in a regression

Logarithms help describe relationships in terms of percentage changes, which may be helpful in certain contexts

This link between logarithms and percentages relies on a key fact derived from calculus:

$$When\ \Delta x\ is\ small, \ln(x+\Delta) - \ln(x) \cong \frac{\Delta}{x}$$

$$Example: \ln(101) - \ln(100) \cong \frac{1}{100} = 1\%$$

Note that all of the percentage change interpretations that follow are accurate when the change itself is relatively small (<10%).

The computation of OLS coefficients stays the same but **interpretation of those coefficients changes**.

We will discuss the two most common cases:

Case 1: Log-level (Y is in logs, X is not):

$$\ln(Y) = \beta_0 + \beta_1 X + u$$

A one unit increase in X is associated with a 100*β_1 percent change in Y.

Consider the following PRF:

$$\ln(income) = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 female + u$$

Here are estimates of PRF based on MA sample we've used before:

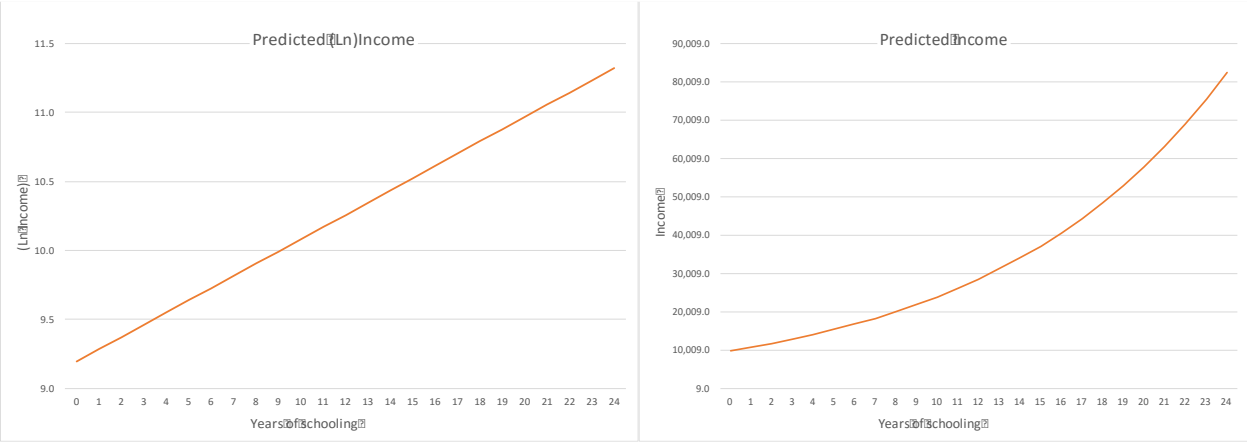
```
. gen ln_income = ln(income)

. reg ln_income educ age female, robust

Linear regression                               Number of obs =     709
                                                F(  3,   705) =   49.27
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1792
                                                Root MSE      =  .45204

------------------------------------------------------------------------------
             |               Robust
   ln_income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0884963   .0091553     9.67   0.000     .0705215    .1064712
         age |   .0372902   .0056259     6.63   0.000     .0262447    .0483358
      female |  -.1413939   .0346169    -4.08   0.000    -.2093585   -.0734294
       _cons |   8.333889   .2186836    38.11   0.000      7.90454    8.763238
------------------------------------------------------------------------------
```

Using average values for *female* and *age*, we can calculate the marginal association between years of schooling and *ln(income)* and *income*.



The slope is small for lower values of X and then increases rapidly.

Interpret the coefficient on *educ*. Is it statistically significant?

Interpret the coefficient on *female.* Is it statistically significant?

We could have used a linear PRF:

$$Income = \beta_0 + \beta_1 educ + \beta_2 age + \beta_2 female + u$$

Why might the log-level model be better than this linear function?

Case 2: Log-log (Y is in logs, X is in logs):

$$\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$$

The interpretation of the coefficient in a log-log relationship is the definition of an elasticity you surely learned in microeconomics…

Consider this estimated relationship between pollution and housing prices:

$$\ln(\widehat{price}) = 9.23 - 0.72 \ln(nox) + 0.31 \, rooms$$

where:

- *Price* represents median housing price in the community.
- *Nox* is the amount of nitrous oxide in the air (parts per million).
- *Rooms* is the mean number of rooms in houses in the community.

How would you interpret the coefficient on *log(nox)*?

How would you interpret the coefficient on *rooms*?

Difficulties in comparing logarithm specifications with non-logarithmic specifications:

Cannot compare output of regressions by any goodness of fit measure, because the outcome variables are different: Log (Y) is a different variable than Y

The best thing is to take the particular application and decide – based on economic theory or experts knowledge – whether it makes sense to specify Y in logarithms:

- In labor economics, studies on wage and wage comparisons are normally framed as percentage comparisons (elasticities)
- When studying test scores, it seems natural to discuss results in terms of points and not percentage increase in test scores

TAKEAWAYS

Quadratics (and higher-order polynomials) and logarithms provide ways of adding curvature to regressions

Allowing for such curvature can:

- Improve the predictions our models make.
- Highlight that the relationship between X and Y can change with X.
- Can yield coefficients with helpful interpretations (in the case of logarithms).

Choosing which non-linear model to use depends on your goals:

- Polynomials can be useful for fitting data
- Logarithms have coefficients with meaningful interpretations

Quadratics and logarithms can be used together in regressions, but be very careful to get the interpretation of the coefficients coming out of these more complex models right!

VOCABULARY

- Non-linear associations/effects of X and Y: non-linear associations is a type of relationship between two variables in which change in one variable does not correspond with a constant change in the other variable. Non-linear variables can also be related to each other in ways that are fairly predictable, but simply more complex than in a linear relationship.
- Quadratic functions (or higher order polynomials): polynomial functions of degree 2 include an *x*-variable squared.
- Logarithmic functions: Although there are logarithms in different basis, we will be using in class the **natural logarithm** of a number, which is its logarithm to the base of the mathematical constant *e*, where *e* is approximately equal to 2.718281828459…
    - Log-level: Regressions with left-hand side Y variable in logs, and all regressors in levels
    - Log-Log: Regressions with left-hand side Y variable in logs, and at least one regressor in logs. The coefficients of the Xs in logs can be interpreted as elasticities.

BONUS: SUMMARY OF LOGARITHMIC MODELS:

Model	Dep. Var.	Exp. Var.	Regression equation (PRF)	Interpretation of β_1 (in math)	Interpretation of β_1 (in words)
level-level	Y	X	$Y_i = \beta_0 + \beta_1 X_i + \mu_i$	$\Delta Y = \beta_1 \Delta X$	One-unit increase in X is associated with a β_1 change in Y
level-log	Y	ln(X)	$Y_i = \beta_0 + \beta_1 \ln(X_i) + \mu_i$	$\Delta Y = (\beta_1 / 100)\%\Delta X$	1% increase in X is associated with a 0.01 β_1 change in Y
log-level	ln(Y)	X	$\ln(Y_i) = \beta_0 + \beta_1 X_i + \mu_i$	$\%\Delta Y = (100\beta_1)\Delta X$	One-unit increase in X is associated with a 100 β_1 % change in Y
log-log	ln(Y)	ln(X)	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \mu_i$	$\%\Delta Y = \beta_1 (\%\Delta X)$	1% increase in X is associated with a β_1 % change in Y (β_1 is the elasticity of Y with respect to X)

BONUS: SUMMARY OF LOGARITHMS:

- Logarithms are the inverses of exponential functions.

  $2^3 = 8 \rightarrow \log2(8) = 3$

- The number e = 2.718281828459045… is the base of the natural logarithm (written "ln" or just "log").

  $e^1 = 2.71\ldots \rightarrow \ln(e) = 1$ [or just log(e) = 1]

- Because an exponential function like 2^X doubles with each unit increase in X, logarithms are useful for thinking about relative (not absolute) rates of change, such as percentage growth rates.