Variable Transformations

- Replacing some x_j with a function $f(x_j)$
- Replacing y with a function $f(y)$
- Most common transformation: logarithm
  - Simple
  - Makes results easy to interpret
  - Can occasionally correct problems with OLS assumptions

Semilogarithmic Form

- Log of outcome variable
- Common for monetary measures: income, GDP
- Wage equation from labor economics, modeling log of wage instead of nominal wage:
  - $\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + u$
- Two common choices for base: base 10 and natural log

Base 10 Log

- In R: log10(y)
- Helpful when thinking about y in terms of powers of 10.
  - If right side of equation is close to 3, *y* will be about 1,000.
  - If slope coefficient is close to 1, every unit increase in *x* will multiply wage *y* by about 10.

Natural Log

- Log base e; in R: log(y)
- Gives elegant interpretation for slope coefficients
- Partial derivative of population model:

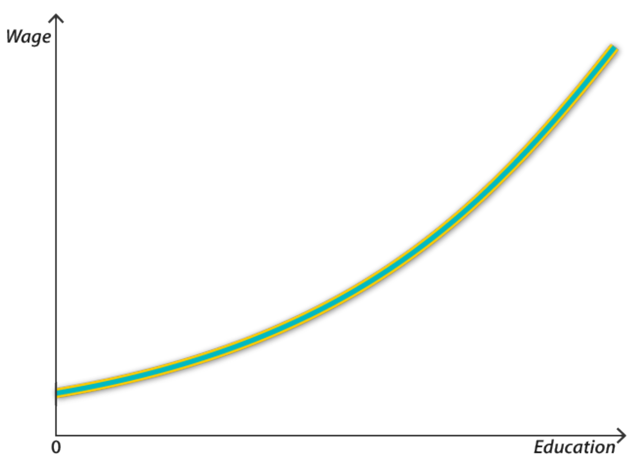
$$\beta_1 = \frac{\delta \log(wage)}{\delta educ} = \frac{1}{wage} \cdot \frac{\delta wage}{\delta educ} = \frac{\frac{\delta wage}{wage}}{\delta educ}$$

- β_1 : proportional increase in wage, as a result of an extra unit of education.
  - If we multiply *y* by a constant, β_1 doesn't change.

Natural Log (cont.)

- Say $\beta_1 = 0.15$; expect extra year of education to result in 15% higher wage.
    - Changes in equation are differential changes; interpretation is only exact in the limit as the changes become small.
    - For small percentage changes, increase in the log is close to the proportional increase in the variable.

Graphing Semilogarithmic Form



- Take exponent of both sides of population model:

$$wage = e^{\beta_0 + \beta_1 educ} = e^{\beta_0} e^{\beta_1 educ}$$

- Wage is exponential function of education.
- If $\beta_1 > 0$, wage increases to right; otherwise, it decreases to right.

Log-Log Form

Log of both *y* and *x* variable

- Example: log of a CEO's salary as linear function of log of firm's sales
  - $\log(\text{salary}) = \beta_0 + \beta_1 \log(\text{sales}) + u$
- Changes interpretation of regression coefficient
- Take the partial of the population model:

$$\beta_1 = \frac{\delta \log(salary)}{\delta \log(sales)} = \frac{\frac{\delta salary}{salary}}{\frac{\delta sales}{sales}}$$

Log-Log Form (cont.)

- Measuring percentage increase in salary, per 1% increase in sales
  - Only strictly true in the limit for small changes, but reasonably close for 10% or 20% changes
- Fitted regression:
  - $\widehat{\log}\left(\text{salary}\right) = 4.822 + .0257\log\left(\text{sales}\right)$
  - Each percentage increase in sales associated with a .257% increase in salary
- In economics, coefficient in log-log model called the **elasticity**
  - Slope of log-log plot of *y* against *x*
  - By choosing log-log, assumption of constant-elasticity relationship

Logarithm Rules of Thumb

- Look for variables that are naturally always positive.
    - Never add constant to make variable positive.
- Look for variables that have meaningful zero-point but no obvious maximum.
- Look for variables where percent change is meaningful.
- Taking logs mitigates influence of outliers in positive direction.
    - Useful for variables with large outliers
- Taking logs can help secure normality and homoskedasticity for OLS.
    - Only a concern for small samples when you can't rely on asymptotics.
    - For large sample, decision to be guided by what's more intuitive or gives better model fit.

Other Transformations

- Quadratic/higher-order polynomials (Y on X^2)
- Occasionally, you may see powers less than 1.
    - Corrects negative skew if you need a normal variable distribution
- Indicator functions convert metric variables to binary ones.
    - Assign a value of 1 whenever a variable is greater than its mean value
- Logit function takes variables bounded by [0,1] and maps them to the entire real line.
    - Idea behind logistic regression