Discussion Section

Week 3

Bret Stevens

University of California, Davis

Logged Variables

What is a logged variable and why do we do this?

- Usually when econometricians say they "logged" a variable it means that they took the natural log of a particular variable
- This literally means that they took every observation of that variable and applied $\ln(x_i)$
- We do this for a couple reasons:
  - Coefficients are easier to interpret
  - Handles outliers better
- You can have four different types of regression:
  - Level-Level
  - Level-Log
  - Log-Level
  - Log-Log

Logged Variables

What exactly do I mean by 'handles outliers better'?

Logged Variables

What exactly do I mean by 'handles outliers better'?

- I'll show you

Logged Variables

What exactly do I mean when I say 'coefficients are easier to interpret'?

- Let's look at what happens when we log all the variables in a simple regression:

$$
\begin{aligned}
& ln(y_i) = b_0 + b_1 ln(x_i) + e_i \\
\Rightarrow & \frac{1}{y} dy = b_1 \frac{1}{x} dx \\
\Rightarrow & \frac{dy}{dx} \left(\frac{x}{y} \right) = b_1 \\
\Rightarrow & \frac{\%\Delta Y}{\%\Delta X} = b_1
\end{aligned}
$$

- Thus, we can interpret coefficients of log-log regressions as elasticities
- This is nice as there are no units to keep track of, everything is relative

Logged Variables

Sometimes we may want to only log one variable. Let's say we want to log only y, how would we interpret that variable?

$$
\begin{aligned}
& ln(y_i) = b_0 + b_1 x_i + e_i \\
\Rightarrow & \frac{1}{y} dy = b_1 dx \\
\Rightarrow & \frac{dy}{y dx} = b_1 \\
\Rightarrow & \frac{\%Y}{\text{unit } X} = 100 b_1
\end{aligned}
$$

This means, if we multiply b_1 by 100 in a log-level regression, we can interpret it as a percentage change in y given a 1 unit change in x.

Logged Variables

If we then only log x:

$$
\begin{aligned}
& y_i = b_0 + b_1 ln(x_i) + e_i \\
\Rightarrow & dy = b_1 \frac{1}{x} dx \\
\Rightarrow & \frac{dy}{dx} x = b_1 \\
\Rightarrow & \frac{\text{unit } Y}{\%X} = \frac{b_1}{100}
\end{aligned}
$$

This means, if we divide b_1 by 100 in a level-log regression, we can interpret it as a unit change in y given a percentage change in x.

Prediction

What is prediction?

Prediction

What is prediction? We just plug in real values to our model!
Let's say we have 3 data points:

y	x
10	4
5	2
15	5

Its a bad idea to run a regression with three data points, but lets do it anyway. We get:

Prediction

What is prediction? We just plug in real values to our model!
Let's say we have 3 data points:

y	x
10	4
5	2
15	5

Its a bad idea to run a regression with three data points, but lets do it anyway. We get:

$$b_0 = -1.7857$$

$$b_1 = 3.2143$$

Prediction

We then can predict a value for each observation:

$$\hat{y_1} : -1.7857 + 3.2143 \cdot (4) = 11.071$$
$$\hat{y_2} : -1.7857 + 3.2143 \cdot (2) = 4.6428$$
$$\hat{y_3} : -1.7857 + 3.2143 \cdot (5) = 14.285$$

Prediction

We then can predict a value for each observation:

$$\hat{y_1} : -1.7857 + 3.2143 \cdot (4) = 11.071$$
$$\hat{y_2} : -1.7857 + 3.2143 \cdot (2) = 4.6428$$
$$\hat{y_3} : -1.7857 + 3.2143 \cdot (5) = 14.285$$

Lets see what Python says

Sample vs. Population

What is the difference between a sample and population?

Sample vs. Population

What is the difference between a sample and population?
Notation:

$$\text{Sample: } y_i = b_0 + b_1 x_i + e_i$$

$$\text{Population: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Sample vs. Population

What is the difference between a sample and population?
Notation:

$$\text{Sample: } y_i = b_0 + b_1 x_i + e_i$$
$$\text{Population: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

But what is a population regression?

Sample vs. Population

What is the difference between a sample and population?
Notation:

$$\text{Sample: } y_i = b_0 + b_1 x_i + e_i$$
$$\text{Population: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

But what is a population regression? It is the best possible regression to explain the variation in some variable. If we were God, and could know every bit of information we ever wanted to, we could construct the population regression. However, this can essentially never happen. As mortals, we can only attempt to estimate this "perfect" regression. Since we are only estimating, we have to be careful of how we interpret the results of a statistical analysis.

Sample vs. Population

How do we deal with our less than perfect data?

- Be very careful about defining your population and research question. These things are very closely linked. Is your population just UC Davis students? All Californian university students? All university students in the world? A dataset containing information on every UC Davis student would be the ideal dataset for a question only about UC Davis students, but an awful dataset looking at every university student in the world. That is, you could extrapolate a lot about the population for the first case, but not be able to say anything at all in the second case.

Sample vs. Population

How do we deal with our less than perfect data?

- Be very careful about defining your population and research question. These things are very closely linked. Is your population just UC Davis students? All Californian university students? All university students in the world? A dataset containing information on every UC Davis student would be the ideal dataset for a question only about UC Davis students, but an awful dataset looking at every university student in the world. That is, you could extrapolate a lot about the population for the first case, but not be able to say anything at all in the second case.
  Can you think of some examples of questions with different population sizes?

Sample vs. Population

- We need to make (and understand the implications of) some assumptions about our sample. In order for OLS to be the best method to analyze our data, a few things must be true:
  1. Representative Sample
  2. Homoskedastic Errors - $Var(\epsilon_i) = \sigma^2$
  3. Uncorrelated Errors - $Cor(\epsilon_i, \epsilon_j) = 0$
  4. Normally Distributed Errors - ϵ_i $N(0, \sigma^2)$
  5. Exogeniety - $E[\epsilon_i | X_i] = 0$

  Its hard to know if these assumptions hold, although there are tests for some. Thus, sometimes you must just argue verbally if you think they will or not. Thus, it is important to understand these assumptions at some depth. Further, understanding what problems may come up given your situation and data type is vital. To recap the types of data are:
  1. Cross-sectional - Many individuals in one time
  2. Time-series - One individual over many time periods
  3. Panel - Many individuals over many time periods

Sample vs. Population

To really understand this difference, we must first understand that the sample estimates b are essentially random variables. We can run regressions on different samples and get different estimates of b. These estimates will have a normal distribution around the true population average. However, this range can be fairly large (draw picture on board).

Unbiasedness

What is unbiasedness?

Unbiasedness

What is unbiasedness?

- First of all, what is the thing that is unbiased? We say that an *estimator* is unbiased when the average sample estimate of some statistic is the population statistic. (Draw a picture)
  - The formula for a sample average is an unbiased estimator of the population mean. There are many different ways in which we could "estimate" the population mean, but the sample average is the best.
- In the context of regression, we want to know when OLS is an unbiased estimator.
- We derived what is necessary for OLS to be unbiased in lecture (Lecture 6 Slide 26). We can do the derivation if you'd like. However, the result is:

$$E[x_i e_i] = 0$$

This essentially means that that x_i and e_i are not correlated.

Unbiasedness

When should we worry about $Cov(x_i, e_i) \neq 0$?

- Non-random samples: When the sample you have was not drawn randomly, the value of x_i will be correlated with e_i. Think of e_i as being "all of the variation in y_i we can't explain with x_i". Thus, when you select your sample, you will be selecting for something that is not included in the regression, leading to a correlation. (Give basketball example)
- Endogeniety - In this case, y_i and x_i have a chicken and egg relationship. This is not accounted for in your regression and thus will lead to bias. (Give price & quantity example)
- Measurement error - If you variables are not measured precisely, this will also lead to bias. If the measurement error is random, it will have less of an effect than if the error is related to the size of x_i

Standard Errors

- Given that our estimates of b are not precise, its important to know how confident we are in the accuracy of our estimates. The way to quantify this is a *standard error*. A standard error lets us know a reasonable range of b given our data. Our *point estimate* is the b that OLS gives us. This is the "best" candidate for b given our data. However, just because it was the best does not mean its necessarily true. Given more data, we may find a wholly different "best" estimate for b.

Standard Errors

So then how do we calculate the standard error of b?

Standard Errors

So then how do we calculate the standard error of b?

- This depends on our assumptions from before. In class, we saw how to derive the standard error assuming homoskedasticity and uncorrelated errors. If those assumptions are broken, that formula will not work as it uses the assumptions to simplify the calculation. The "standard" formula for the standard error of b is:

$$\frac{1}{N-K-1}\sum_i (Y_i - \hat{Y}_i)^2 = \frac{ESS}{N-K-1}$$

  Where K is the number of x's on the right-hand side of your regression.
- Just remember that this formula only holds if CR2 (Homoskedasticity) and CR3(Uncorrelated Errors) are true.

Standard Errors

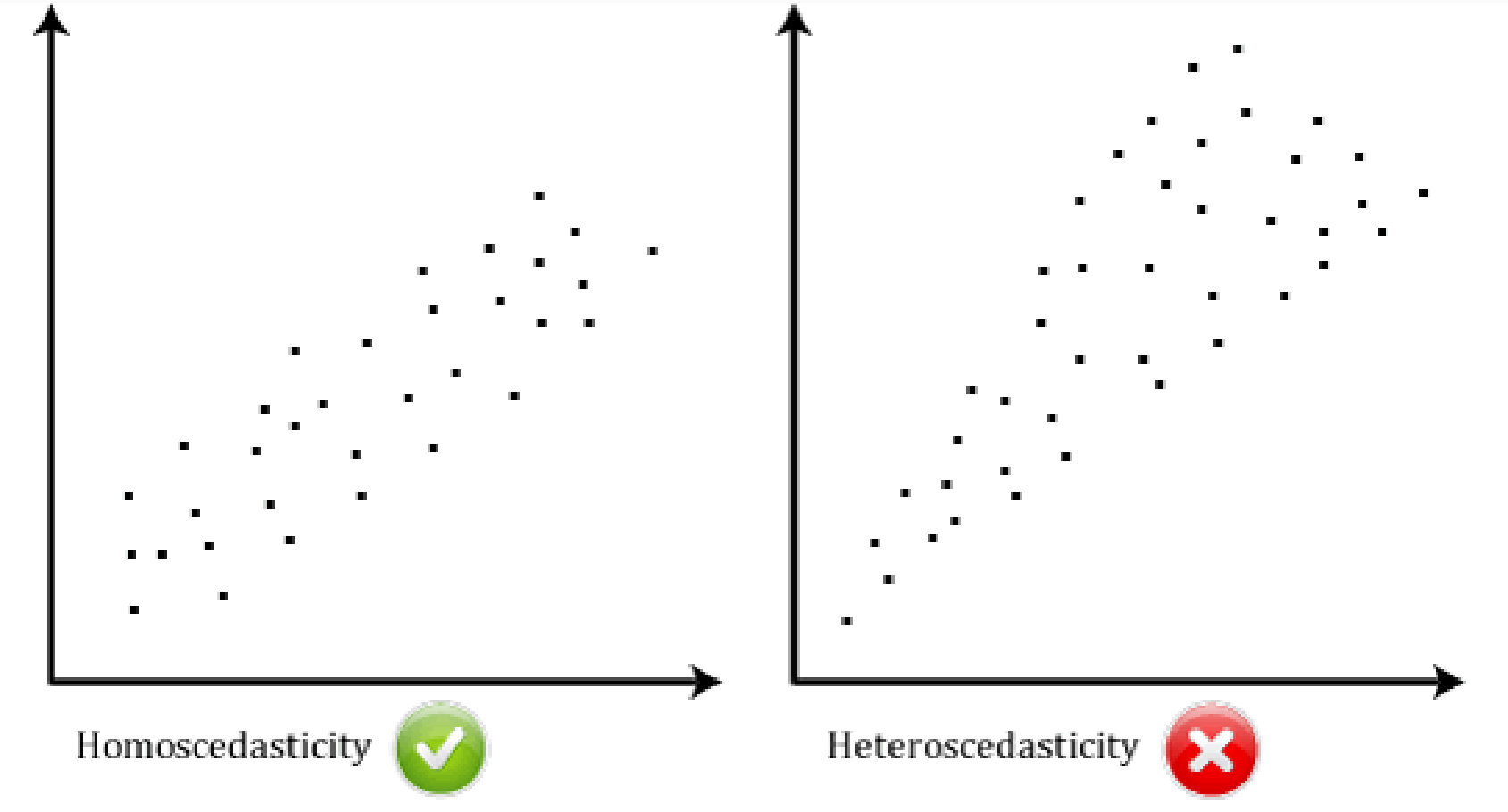
Great. So we know that these conditions have to hold, but what do they mean?

Standard Errors

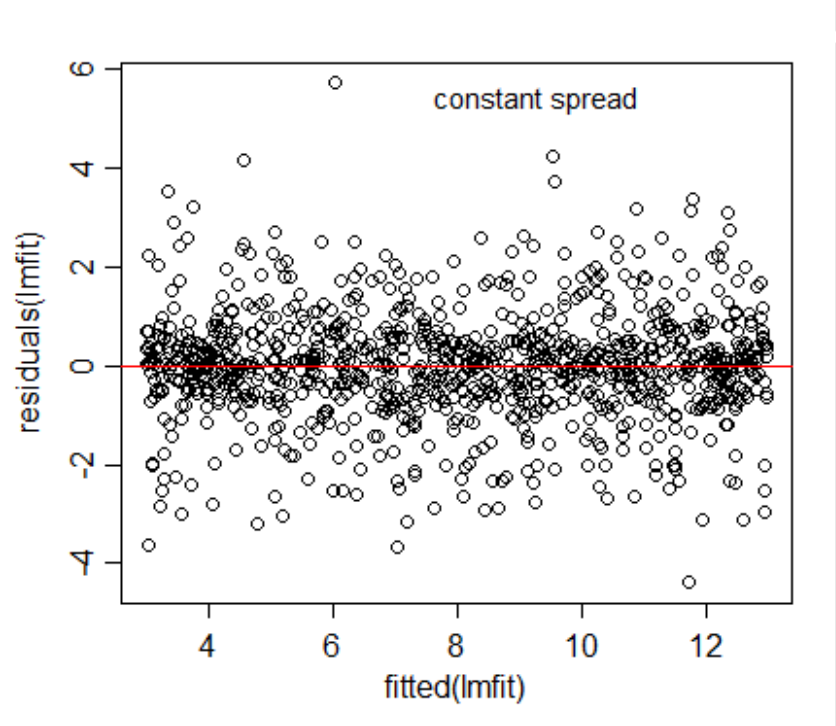
Great. So we know that these conditions have to hold, but what do they mean? Frankly, its easier just to look at it. Lets start with homoskedasticity:

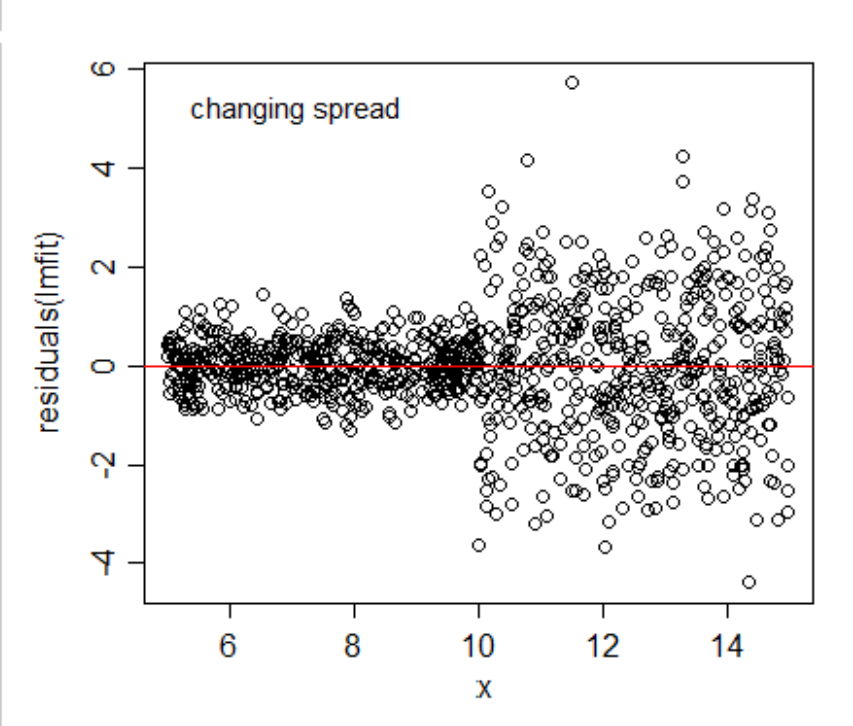
Standard Errors

Great. So we know that these conditions have to hold, but what do they mean? Frankly, its easier just to look at it. Lets start with homoskedasticity:

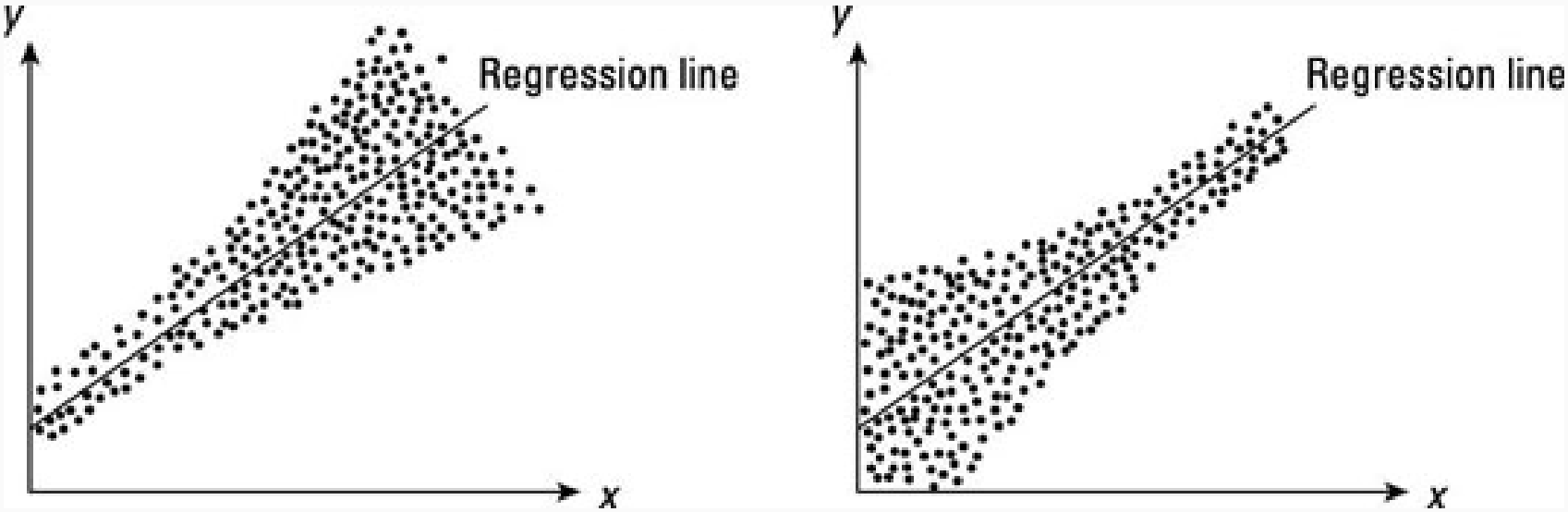


Standard Errors

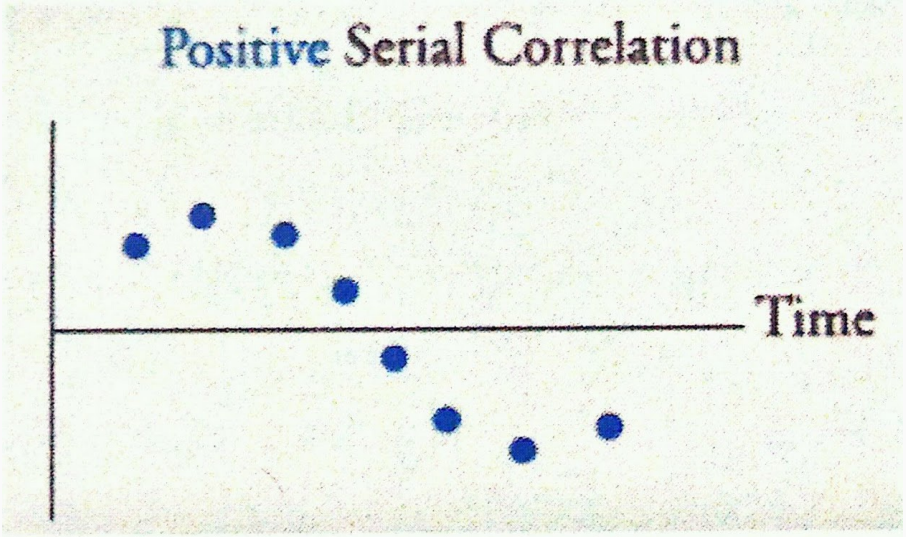


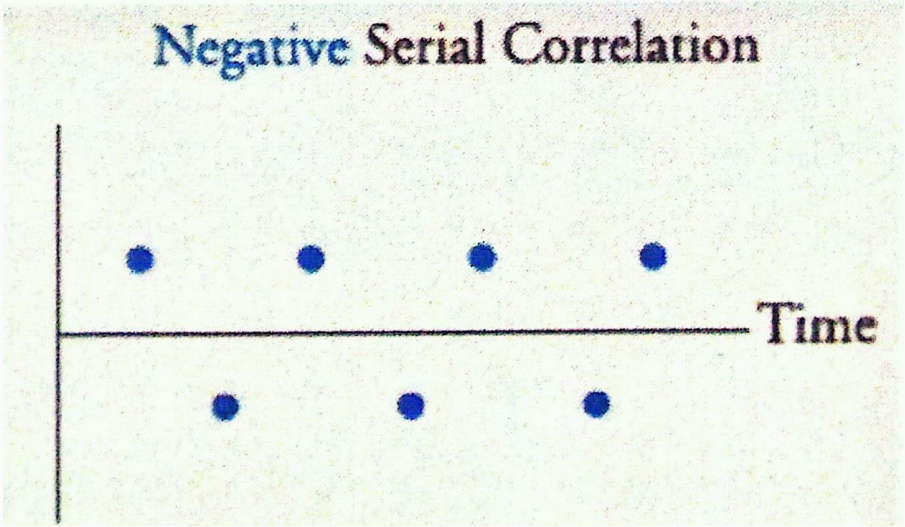


Standard Errors



Standard Errors





Standard Errors

So what do we want from standard errors?

Standard Errors

So what do we want from standard errors? For them to be small! The smaller they are the more confident we are in our point estimate of b. However, if CR2(Homoskedasticity) or CR3(Uncorrelated Errors) are not true, we will have to calculate our standard errors differently, which will make them larger. We always want to have homoskedastic errors, but rarely ever see them in practice.

- To get smaller standard errors we can use more data and have more variation in x.