

# Logarithms in Regression Analysis with Asiaphoria

ECO220Y1Y: 2016/17; Written by Jennifer Murdock

## 1 Required supplement to the textbook

Many applications of regression analysis involve variables that have had a logarithmic transformation. This supplement supports your study of this important topic and is required reading for all sections of ECO220Y1Y. It supplements the textbook, which *does* address non-linearity in regression analysis.<sup>1</sup>

So, why do you need extra required reading? This supplement addresses how to *interpret* the coefficients when logarithms are used. The textbook touches on this on page 705 but otherwise does not address the *interpretation* of these extremely common regression results.

Study this supplement right after Chapters 6 - 7. Review it, along with Chapters 6 - 7, when you begin Chapters 18 - 21. Work through all exercises in Section 10 and use Section 11 to check your work. Last, but not least, work through the Asiaphoria case study in Section 12.

## 2 Logarithms in simple regression, descriptive statistics

This supplement works within simple regression. For example, how happiness (y variable) is related to wealth (x variable). Simple regression is sometimes called bivariate regression because there are two variables. Chapters 20 and 21 extend to multiple regression. For example, how happiness (y variable) relates to wealth, health, age, gender, ... (multiple x variables). However, *simple regression* comes up two times: once with Chapters 6 - 7 and again with Chapters 18 - 19. The later chapters apply formal methods of statistical inference (hypothesis testing and interval estimation) to regression analysis. However, the key concepts specific to the use of logarithms can be explained within simple regression and descriptive statistics (i.e. Chapter 6 - 7), which is what this supplement does.

## 3 Motivating example: Human Development Index (HDI)

The United Nations (UN) reports the the Human Development Index (HDI) across countries and over time. It is a “way of measuring development by combining indicators of life expectancy, educational attainment and income into a composite human development index, the HDI.”<sup>2</sup> On page 203 of Section 7.8, the textbook studies cell phone penetration versus the HDI. Let’s use the more recent HDI data (2012) downloaded from the UN website for a cross-section of 187 countries and the more recent cellular telephone data downloaded from the ITU website<sup>3</sup> for a cross-section of 157 countries.<sup>4</sup>

---

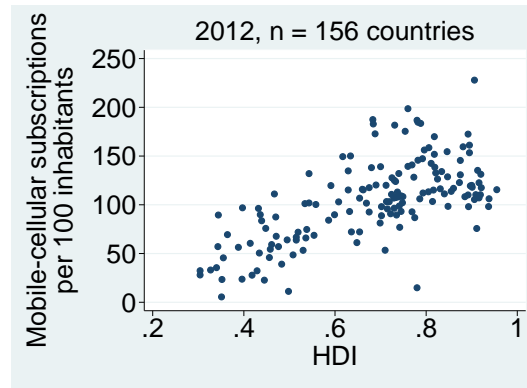
<sup>1</sup>See Section 6.4 “Straightening Scatterplots,” Section 7.8 “Nonlinear Relationships,” “Watch out for changing spread” on page 667, Section 19.6 “Linearity,” Section 19.7 “Transforming (Re-expressing) Data,” Section 19.8 “The Ladder of Powers,” “Don’t fit a linear regression to data that aren’t straight” on page 762, and “Watch out for the plot thickening” on page 762.

<sup>2</sup>See page entitled “Human Development Index (HDI)” at <http://hdr.undp.org/en/statistics/hdi>.

<sup>3</sup>See pp. 228 - 229 of “Measuring the Information Society: 2013” by the International Telecommunication Union (ITU): <http://www.itu.int/en/ITU-D/Statistics/Pages/publications/mis2013.aspx>.

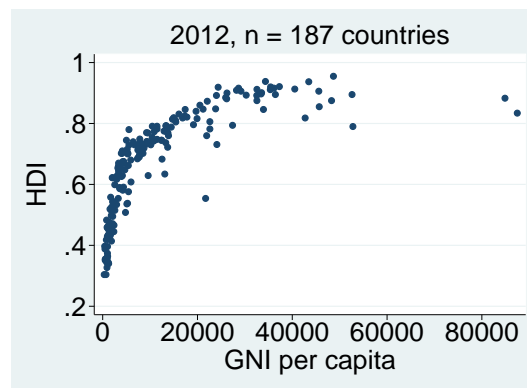
<sup>4</sup>When these two data sets are merged there are 156 observations. The ITU data include Macao, China as an observation but the UN data do not. All of the other ITU observations are also in the UN data.

Unlike the scatter diagram in the book (Figure 7.7), **Figure 1** shows a fairly linear association.<sup>5</sup>

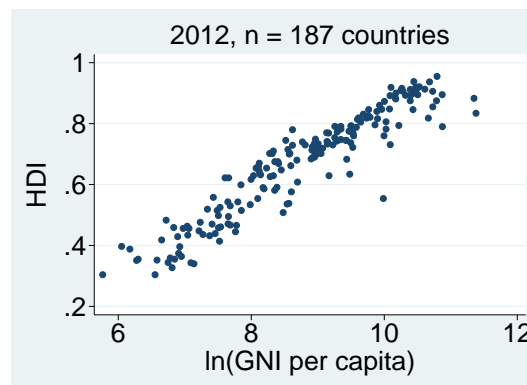


**Figure 1:** A scatter plot of mobile-cellular subscriptions per 100 inhabitants versus the HDI for a cross-section of 156 countries is fairly linear.

Income affects the HDI. The UN uses GNI (Gross National Income). **Figure 2** shows a nonlinear association between the HDI and GNI. Any statistic that assumes linearity –  $R^2$ , correlation, OLS slope – will describe it poorly. Logarithmically transforming either the x variable, y variable, or both often straightens relationships. Which combination, if any, will straighten this scatter plot?



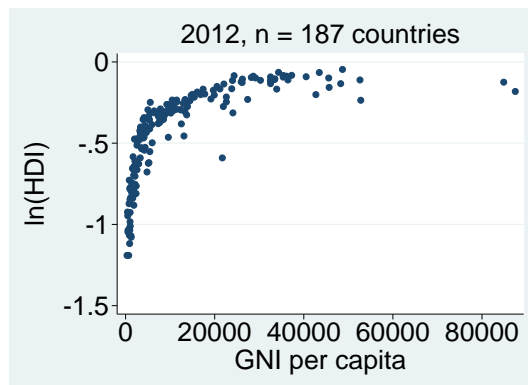
**Figure 2:** A scatter plot of the HDI in 2012 versus GNI per capita in 2012 for a cross-section of 187 countries shows a highly non-linear association.



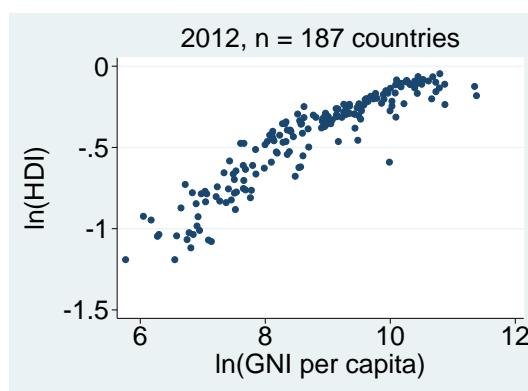
**Figure 3:** HDI versus the natural log of GNI per capita is fairly linear.

<sup>5</sup>Aside from using older data, the textbook uses a different measure of mobile-cellular penetration but there is insufficient information provided to determine exactly which measure is used.

**Figure 3** shows a fairly linear association between HDI and the natural logarithm of GNI per capita. Hence logarithmically transforming the x variable straightened this scatter plot. The other two possible combinations of logarithmic transformations – given in **Figure 4** and **Figure 5** – do not work or do not work as well.



**Figure 4:** The natural log of HDI versus GNI per capita shows that logarithmically transforming the y variable does not help straighten the scatter plot and seems to make it even more non-linear.



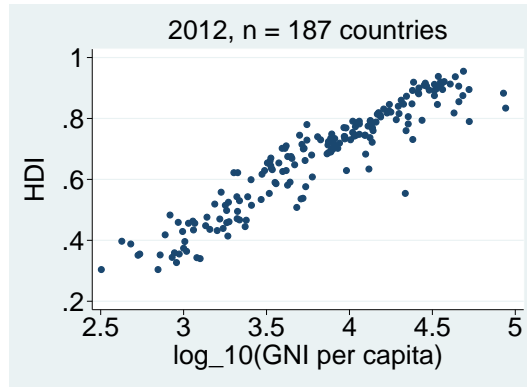
**Figure 5:** The natural logarithm of HDI versus the natural log of GNI per capita shows that logarithmically transforming both the x and y variables helps straighten the scatter plot compared to **Figure 2** but does not do as good a job as in **Figure 3**.

### 3.1 Natural logarithm versus base 10 logarithm

Before getting to the main point – how to interpret the OLS coefficient when one or more of the variables have been logarithmically transformed – consider natural logarithms versus base 10 logarithms. In applied work, and certainly amongst economists, when someone says “logarithm” they mean “natural logarithm.” It is common to write  $\log()$  even when talking about a natural logarithm.<sup>6</sup>

What if a base 10 logarithmic transformation is applied to GNI per capita? **Figure 6** looks very similar to **Figure 3** but for the scale of the horizontal axis. In fact, the  $R^2$  is 0.8854 in both cases. However, you will see why the natural logarithm is convenient in the next sections.

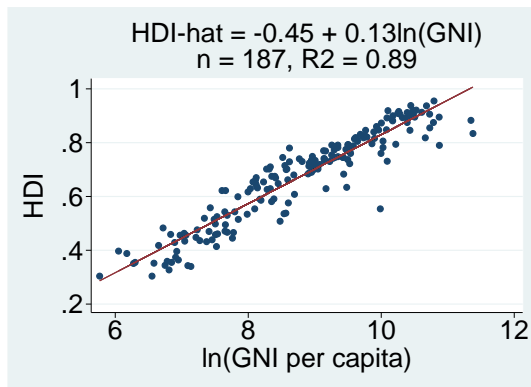
<sup>6</sup>Popular software reflect this convention. In STATA both the function  $\log$  and  $\ln$  return the natural logarithm: you must type  $\log10$  to get the base 10 logarithm. However, Excel is an exception:  $\log$  means a base 10 logarithm.



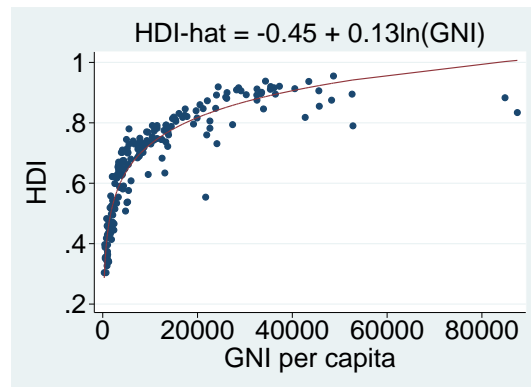
**Figure 6:** The horizontal scale differs from **Figure 3** but the straightening ability of a natural logarithm and a base 10 logarithm are indistinguishable.

#### 4 Logarithm of the x variable but not the y variable: $\hat{y}_i = b_0 + b_1 \ln(x_i)$

For HDI versus GNI taking the log of the x variable succeeded in straightening the scatter plot. **Figure 7** shows the OLS line fitted to the transformed data:  $\widehat{HDI}_i = -0.45 + 0.13 * \ln(GNI_i)$ . How to interpret 0.13? You *CANNOT* say “In countries where the natural logarithm of GNI per capita is 1 unit higher the HDI is 0.13 higher on average.” That is *NOT* an interpretation. Interpretation means explaining numbers so that people can understand them: it is hard to find a person who thinks in terms of the natural logarithm of dollars!



**Figure 7:** OLS: straightened scatter plot.



**Figure 8:** HDI versus GNI.

An interpretation requires understanding how y varies with respect to x. For example, how HDI varies with respect to GNI. There is a linear association between HDI and the natural logarithm of GNI – **Figure 7** – but a non-linear association between HDI and GNI – **Figure 8**. The interpretation is about HDI and GNI (not logged): i.e. the curve in **Figure 8**. To see how to get from the equation, which involves logarithms, to an interpretation that does not involve logarithms:

Start with the equation:  $y = b_0 + b_1 \ln(x)$

Take the derivative with respect to x (i.e. find the slope):  $\frac{dy}{dx} = b_1 \frac{1}{x}$

Rewrite to see how a change in x associates with a change in y:  $dy = \frac{dx}{x} b_1$

The slope of a curve is not constant: in this case  $\frac{dy}{dx} = \frac{b_1}{x}$ . However, if  $\frac{dx}{x}$  is fairly small – i.e. a change in  $x$  that is modest relative to the size of  $x$  – you can use a point estimate.

A percent change in  $x$  is  $100 * \frac{dx}{x}$ . For example, if GNI goes from \$10,000 to \$10,250 that is a 2.5% increase:  $2.5 = 100 * \frac{dx}{x} = 100 * \frac{250}{10000}$ . Hence,  $dy = \frac{dx}{x} b_1$  says that  $\frac{b_1}{100}$  can be interpreted as the change in  $y$  given a percent change in  $x$ :  $dy = (100 * \frac{dx}{x})(\frac{b_1}{100})$ .

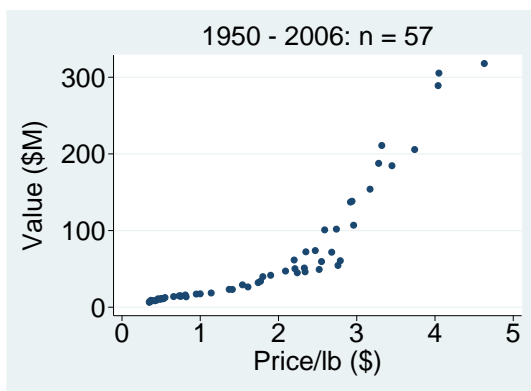
So, how to interpret 0.13? “In 2012, in countries where Gross National Income per capita is 1 percent higher the Human Development Index (an index between 0 and 1) is approximately 0.0013 units higher on average.” You may also write “In 2012, in countries where GNI per capita is 10 percent higher the HDI (an index between 0 and 1) is approximately 0.013 units higher on average.”

Can you see how the idea of percent changes in  $x$  captures the curve in **Figure 8**? A 10% increase in GNI if GNI is \$3,000 (poor country) is \$300 per capita. A 10% increase in GNI if GNI is \$60,000 (rich country) is \$6,000 per capita. According to the prediction of the regression line, both the \$300 and the \$6,000 change are associated with an HDI that is 0.013 higher on average. There are diminishing marginal returns: the HDI is constructed to increase with income but at a decreasing rate.

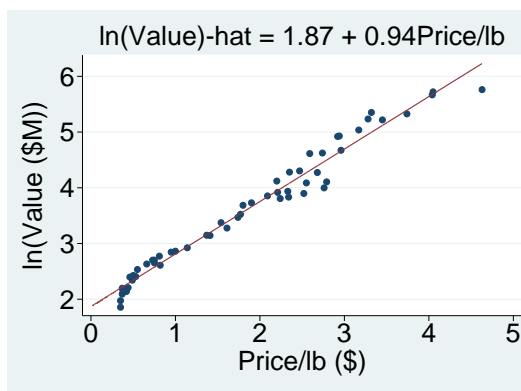
## 5 Logarithm of the $y$ variable but not the $x$ variable: $\ln(\widehat{y}_i) = b_0 + b_1 x_i$

In exercise 65 on page 220, the textbook studies data on Maine lobsters. Variables include total lobsters harvested (by weight), market value, price per pound, number of license holders, number of traps, and water temperature.<sup>7</sup> These data are time-series: the unit of observation is a year.

**Figure 9** shows the non-linear association between the dollar value of the Maine lobster harvest and the price per pound over the 1950 to 2006 period. **Figure 10** shows that a logarithmic transformation of the  $y$  variable somewhat straightens it, although there is still clear evidence of non-linearity at low prices. Ignoring the remaining concerns about non-linearity, the OLS line fitted to the transformed data is:  $\ln(\widehat{Value}_t) = 1.87 + 0.94 * Price_t$ . How to interpret 0.94?



**Figure 9:** Exercise 65: untransformed



**Figure 10:** Exercise 65:  $y$  transformed

<sup>7</sup>More recent data are available from the Maine Department of Natural Resources on a page entitled “Historical Maine Lobster Landings”: <http://www.maine.gov/dmr/commercial-fishing/landings/documents/lobster.table.pdf>.

The same approach as in Section 4 can illuminate the interpretation of the OLS coefficient:

Start with the equation:  $\ln(y) = b_0 + b_1x$

Rewrite:  $y = \exp(b_0 + b_1x)$

Take the derivative with respect to x (i.e. find the slope):  $\frac{dy}{dx} = b_1 \exp(b_0 + b_1x)$

Rewrite:  $\frac{dy}{y} = b_1 dx$

Rewrite to see how a change in x associates with a change in y:  $\frac{dy}{y} = b_1 dx$

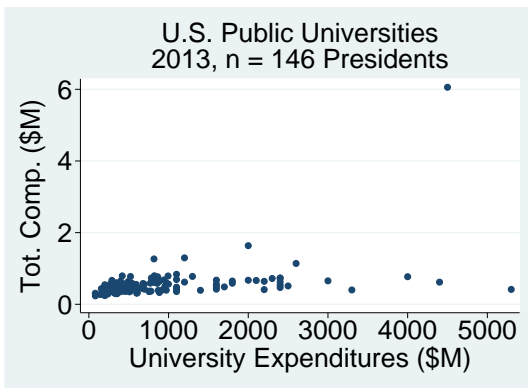
A percent change in y is  $100 * \frac{dy}{y}$ . For example, if the market value of the lobster harvest goes from \$50.0 million to \$50.7 million that is a 1.4% increase:  $1.4 = 100 * \frac{dy}{y} = 100 * \frac{0.7}{50}$ . Hence,  $\frac{dy}{y} = b_1 dx$  says that  $100 * b_1$  can be interpreted as the percent change in y given a unit change in x:  $100 * \frac{dy}{y} = 100 * b_1 dx$ .

So, how to interpret 0.94? “Over the period from 1950 through 2006, in years where the price per pound of lobster is \$0.10 higher the total dollar value of the annual Maine lobster harvest is approximately 9.4 percent higher on average.” Why use a price change of \$0.10 and not \$1? In this context a \$1 price change is very large and would strain the interpretation, which is an approximation.

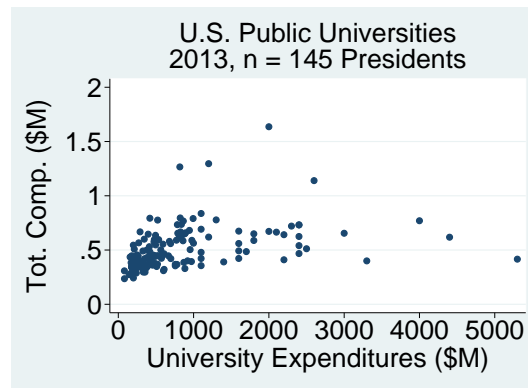
## 6 Logarithm of both the x and the y variables: $\ln(\widehat{y}_i) = b_0 + b_1 \ln(x_i)$

Consider data on compensation of presidents at public universities in the U.S.<sup>8</sup> Variables include the president’s total compensation, base pay, total university expenditures, and the date the president took office. These data are cross-sectional: the unit of observation is a university president.<sup>9</sup>

Do presidents at bigger schools make more money? Start with a scatter diagram of compensation versus total expenditures. **Figure 11** shows an outlier and non-linearity. **Figure 12** excludes the outlier but there is still non-linearity. Both cases clearly show heteroscedasticity (unequal spread).



**Figure 11:** E. Gordon Gee \$6.1 million

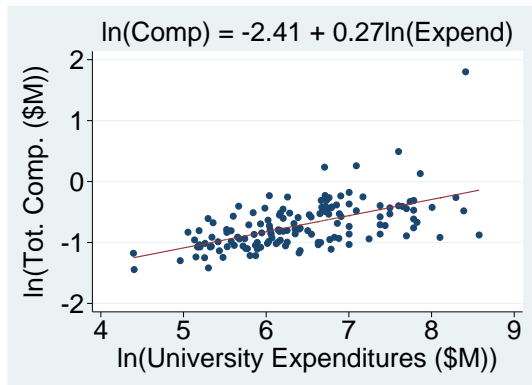


**Figure 12:** Non-linear even w/o outlier

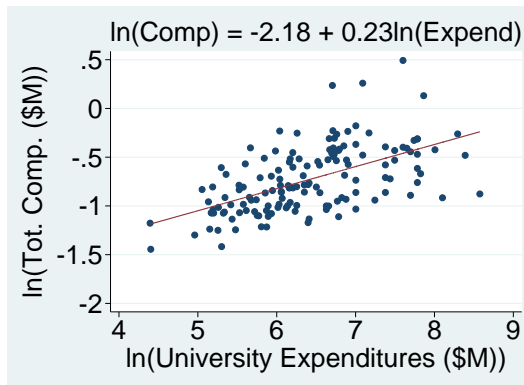
<sup>8</sup>On May 16, 2014 *The Chronicle of Higher Education* published “Executive Compensation at Public Colleges, 2013 Fiscal Year.” No link is given as these data are available only by subscription.

<sup>9</sup>Of the 254 observations, the analysis focuses on the 146 who were in office for the entire 2013 fiscal year.

For patterns like these taking the natural log of both variables often works.<sup>10</sup> **Figure 13** is much more in line with the underlying assumptions of OLS.<sup>11</sup> Even the outlier is less extreme. Often outliers disappear altogether but the former Ohio State president continues to stand out. To check how the results are affected by this single point: see **Figure 14**. The OLS coefficient does differ between the two figures – 0.27 versus 0.23 – but is that a large difference? To answer requires knowing how to interpret the OLS coefficient when both the x and y variables have been logged.



**Figure 13:** Both x and y logged



**Figure 14:** Both logged; w/o outlier

The same approach as the previous sections can illuminate the interpretation of the OLS coefficients:

Start with the equation:  $\ln(y) = b_0 + b_1 \ln(x)$

Rewrite:  $y = \exp(b_0 + b_1 \ln(x)) = \exp(b_0) * x^{b_1}$

Take the derivative with respect to x (i.e. find the slope):  $\frac{dy}{dx} = b_1 \exp(b_0) * x^{b_1-1}$

Rewrite:  $\frac{dy}{dx} = b_1 \frac{y}{x}$

Rewrite to see how a change in x associates with a change in y:  $\frac{dy}{y} = b_1 \frac{dx}{x}$

A percent change in x is  $100 * \frac{dx}{x}$  and percent change in y is  $100 * \frac{dy}{y}$ . Hence,  $\frac{dy}{y} = b_1 \frac{dx}{x}$  says that  $b_1$  can be interpreted as the percent change in y given a percent change in x:  $100 * \frac{dy}{y} = b_1 * 100 * \frac{dx}{x}$ . Economists would call this an elasticity. Hence, it is often called a constant elasticity functional form.

So, how to interpret the 0.27 and 0.23 OLS coefficients from **Figure 13** and **Figure 14**? “In 2013 in the U.S., public colleges and universities with total expenditures that are 1 percent higher typically pay the president approximately 0.27 percent more in total compensation. However, if the highest paid president (with total compensation over \$4 million more than the second-highest paid president) is removed, the estimate drops from 0.27 to 0.23. Either way, presidents at bigger schools (as measured by total expenditures) do receive somewhat more lucrative compensation packages.”

<sup>10</sup>For another such example see page 204 of the textbook.

<sup>11</sup>There is still some suggestion of heteroscedasticity but it is no longer a substantial violation.

## 7 Interpretation summary

To summarize, here is how to interpret OLS coefficients when logarithms are used. For comparison, the first case is the simple linear functional form with no logarithmic transformations.

**Linear specification** ( $\widehat{y}_i = b_0 + b_1x_i$ ) “A one unit change in  $x_i$  is associated with a  $b_1$  unit change in  $y_i$  on average.” The OLS coefficient  $b_1$  is a slope. **Note well:** A full interpretation *must* specify the context, give variable descriptions, and provide the units of measurement of both  $x$  and  $y$ .

**Partial-log specification** ( $\widehat{y}_i = b_0 + b_1\ln(x_i)$ ) “A one percent change in  $x_i$  is associated with approximately a  $b_1/100$  unit change in  $y_i$  on average.” **Note well:** A full interpretation *must* specify the context, give variable descriptions, and provide the units of measurement of  $y$ .

**Semi-log specification** ( $\widehat{\ln(y_i)} = b_0 + b_1x_i$ ) “A one unit change in  $x_i$  is associated with approximately a  $100 * b_1$  percent change in  $y_i$  on average.” **Note well:** A full interpretation *must* specify the context, give variable descriptions, and provide the units of measurement of  $x$ .

**Log-log specification** ( $\widehat{\ln(y_i)} = b_0 + b_1\ln(x_i)$ ) “A one percent change in  $x_i$  is associated with approximately a  $b_1$  percent change in  $y_i$  on average.” The OLS coefficient  $b_1$  is an elasticity. **Note well:** A full interpretation *must* specify the context and give variable descriptions.

### 7.1 Descriptive versus causal interpretations: you *must* take a stand

The interpretation templates above are descriptive. However, if you are able to infer causality then a full interpretation *must* make that clear. This is the difference between saying an extra year of education is associated with 7% higher annual earnings on average (descriptive) and saying that investing in an extra year of education yields 7% higher earnings on average (causal). The first describes patterns but is *not* a sufficient reason to advise people to get more education: it may be that those who choose to obtain extra education are systematically different from those that do not. In that case, it may be the systematic other differences (e.g. high motivation, successful parents, etc.) that are causing some (or even all) of the observed higher earnings. However, if a causal interpretation is warranted then it would be the basis for sound advice: people who were not planning on more education but get more on average will earn 7% higher annual earnings per additional year of education. Sometimes students think that always offering a descriptive interpretation (which is less strong) is safe and always correct. No. You must be clear on causality and imply it when justified. Otherwise you would fail to offer a full interpretation.

### 7.2 Percent changes versus percentage point changes

When the  $x$  or  $y$  variable (or both) have been logged, notice that the interpretation of OLS coefficients involves percent changes. Is there a difference between a *percent change* and a *percentage point change*? Yes, these are very different. Percent changes make sense regardless of the units of measurement of the original variable. For example, if a city had 26 smog-alert days in 2013 and 32 in 2014, then it is up by 6 days, which is a 23 percent increase ( $= 100 * (32 - 26)/26$ ). However, percentage point changes only make sense if a variable is measured as a percent: percentage points are the units



of measurement for a variable measured as a percent. For example, if a city had an unemployment rate of 7.2 percent in 2013 and 8.1 percent in 2014, then unemployment is up by 0.9 percentage points.

Consider this sentence: “Lottery winners are 14.1 percentage points more likely to enroll in college the fall after their senior year, a 49.0 percent increase from the control mean of 28.8 percent.”<sup>12</sup> What does this mean exactly? This means that 28.8 percent of lottery losers (the control group) enroll in college while 42.9 percent of lottery winners (the treatment group) enroll. There are two different ways to describe this (large) difference in these percents. You could say that the lottery winners are 14.1 *percentage points* more likely to enroll (= 42.9 – 28.8). Alternatively, you could say that the lottery winners are 49.0 *percent* more likely to enroll ( $100 * (42.9 - 28.8)/28.8$ ). Obviously 14.1 and 49.0 are very different numbers: percent and percentage point mean *different* things.

To illustrate this distinction when interpreting OLS coefficients, consider a situation where a variable measured as a percent either has or has not been logged. Consider a random sample of 40 publicly traded firms. The x variable is the percent of the CEO’s total annual compensation that is stock options (e.g. x could be 10, which means 10% of the CEO’s compensation is stock options). The y variable is the change in the firm’s stock price from 2013 to 2014 in dollars (e.g. y could be -0.53 if the stock price dropped by 53 cents). Suppose a scatter plot looks straight and the OLS regression results are:  $\hat{y}_i = 0.8 + 0.1x_i$ . How to interpret the OLS coefficient of 0.1? Start by recalling the units of measurement: x is measured as a percent and y in dollars. On average firms whose CEO’s have 1 *percentage point* more of their compensation in the form of stock options (e.g. 7% versus 6%) have annual growth in their stock price that is 10 cents (\$0.10) higher. Suppose a log transformation of the x variable is required to straighten the scatter plot and the OLS regression results are:  $\hat{y}_i = 0.6 + 0.4\ln(x_i)$ . How to interpret the OLS coefficient of 0.4? On average firms whose CEO’s have 1 *percent* more of their compensation in the form of stock options (e.g. 6.06% versus 6%) have annual growth in their stock price that is 0.4 cents (\$0.004) higher. In this example a 1 percent change is really tiny (6.06% versus 6%) so it makes more sense to use 10 percent in the interpretation: On average firms whose CEO’s have 10 percent more of their compensation in the form of stock options (e.g. 6.6% versus 6%) have annual growth in their stock price that is 4 cents (\$0.04) higher. In this example, specifications where the y variable is logged do not make sense as it can take negative values (stock prices often go down).

### 7.3 When logs are involved, coefficient interpretations are approximations

When reviewing the interpretation summaries at the start of Section 7, did you notice that the word *approximately* appears for each specification involving logarithms but not for the linear specification? These interpretations are approximations that are accurate for marginal changes. They can become nonsense for non-marginal changes. In contrast, for a linear specification it does not matter whether the change is marginal because the slope is constant.

For example, consider cross-sectional data for 20 randomly selected geographic areas. The x variable measures the unemployment rate: for example, x equal to 0.05 is 5% unemployment. The y variable

---

<sup>12</sup>This appears on page 16 of a 2013 NBER working paper entitled “The Medium-Term Impacts of High-Achieving Charter Schools on Non-Test Score Outcomes” by Will Dobbie and Roland Fryer <http://www.nber.org/papers/w19581>.

measures the fraction of voters that favor free trade policies (for example, 0.63 means that 63% favor free trade). OLS regression results are  $\widehat{\ln(y_i)} = -0.4 - 1.5x_i$ . How to interpret -1.5? A terrible answer is: A one unit increase in x is on average associated with a 150% decrease in y. This makes no mention of the context, fails to say what the variables are, forgets to specify units, *and* uses an approximation for a non-marginal scenario. The last mistake explains the crazy 150% decrease. (Something positive, like the fraction favoring free trade, could at most decrease by 100%, say from 0.60 to 0. A 150% decrease is not even mathematically possible.) What is a 1 unit change in x in this example? Given that x is the unemployment rate, it has a possible range from 0 to 1. Hence a 1 unit change would be comparing geographic areas with 0% unemployment to those with 100% unemployment. Because  $-1.5 \times 1$  is a huge (not marginal) predicted change in the natural logarithm of y, the approximate interpretation fails. We can still use the approximate interpretation for a smaller change in the x variable such as a 1 percentage point change in unemployment (e.g. going from 5 to 6%, which is a 0.01 change in the x variable). A correct interpretation is: Geographic areas with unemployment rates that are 1 percentage point higher (e.g. 8% versus 7%) on average have 1.5 percent less support for free trade (e.g. 59.1% versus 60%). Notice that this descriptive interpretation keeps percent and percentage points straight but forgot to say *approximately*. Technically it should have said approximately, but researchers often leave that word out (once they have made sure they are considering a marginal change when logarithms are involved).

### 7.3.1 But what about a non-marginal change (i.e. when the approximation fails)?

What is the interpretation of a coefficient for a *non-marginal* change in a regression with logarithms? Even if the approximation fails, an exact interpretation is still possible. Continuing with the free trade and unemployment example, what does the regression coefficient say about about geographic areas with 100% unemployment versus others with 0% unemployment? (Note: In this example these are both likely outside the range of the data: this comparison would require extrapolation well beyond the evidence. However, put aside concerns about extrapolation to see what the coefficient mathematically implies about that comparison. In general, just because changes are non-marginal does *not* imply that they are necessarily outside the range of the data.) Define  $\ln(y1_i)$  as the predicted natural logarithm of free trade support in geographic areas with 100% unemployment ( $x = 1$ ) and define  $\ln(y0_i)$  as the predicted natural logarithm of free trade support in geographic areas with 0% unemployment ( $x = 0$ ). The difference of interest is  $\ln(y1_i) - \ln(y0_i)$ , which OLS estimates as -1.5 ( $-1.5 = (-0.4 - 1.5 * 1) - (-0.4 - 1.5 * 0)$ ).

$$\begin{aligned} \ln(y1_i) - \ln(y0_i) &= -1.5 \\ \ln\left(\frac{y1_i}{y0_i}\right) &= -1.5 \\ \ln\left(1 + \frac{y1_i - y0_i}{y0_i}\right) &= -1.5 \\ 1 + \frac{y1_i - y0_i}{y0_i} &= e^{-1.5} \\ \frac{y1_i - y0_i}{y0_i} &= e^{-1.5} - 1 = -0.78 \end{aligned}$$

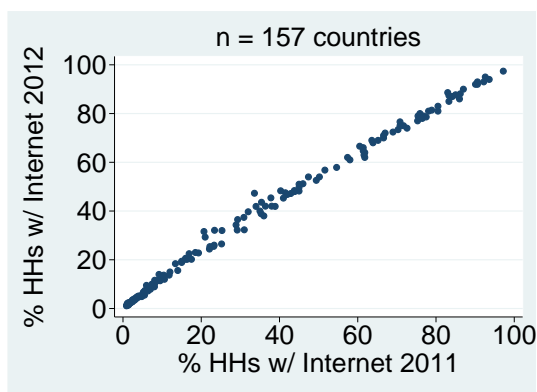
What does -0.78 mean? Geographic areas with unemployment rates that are 100 percentage point

higher on average have 78 percent less support for free trade (e.g. the difference between 67% and 15% of voters supporting free trade). Notice the big difference between saying 150 percent less (the failed approximation) and 78 percent less (the exact interpretation).

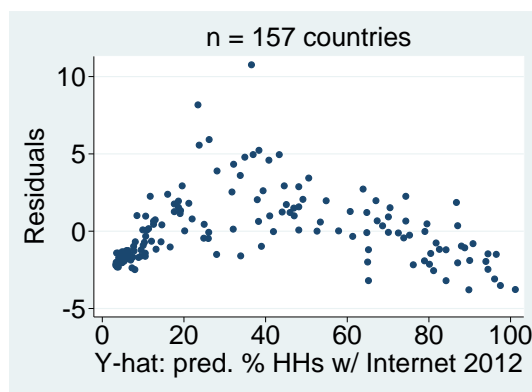
Given that the approximate interpretations should not be used for non-marginal changes, where is the boundary between a marginal and a non-marginal change? The answer is subjective. On page 94 of *Mastering ‘Metrics: The Path from Cause to Effect* (2015), the authors (prominent professors of economics at MIT and LSE) suggest that if the predicted difference in the logged values of  $y$  is between  $-0.2$  and  $0.2$  then this is sufficiently marginal to use the standard approximations given at the start of this section. To illustrate, suppose that the OLS results for the free trade and unemployment example had been  $\widehat{\ln(y_i)} = -0.4 - 0.1x_i$  (instead of  $\widehat{\ln(y_i)} = -0.4 - 1.5x_i$ ). Then  $\ln(y1_i) - \ln(y0_i) = -0.1$ , and  $-0.1$  is between  $-0.2$  and  $0.2$ . Geographic areas with unemployment rates that are 100 percentage point higher on average have *approximately* 10 percent less support for free trade, whereas the exact value is 9.5 percent less ( $e^{-0.1} - 1 = -0.095$ ). The approximation is excellent.

## 8 Sneaky non-linearities

Sometimes it is hard to see non-linearity. The ITU data in Section 3 include variables on fixed telephone subscriptions per 100 inhabitants, mobile-cellular telephone subscriptions per 100 inhabitants, percent of households with a computer, and percent of households with internet access. These are measured for 157 countries for each of two years (2011 and 2012), which means these are panel (longitudinal) data. How well does the 2011 percent of households with internet access predict the percent in 2012? **Figure 15** shows a strong association (as would be expected) between 2011 and 2012 internet penetration. It is hard to see but it is *NOT* straight. The best way to check for non-linearity is to try to run a linear regression and then graph the residuals ( $e_i = y_i - \widehat{y}_i$ ) versus the predicted values of  $y$  ( $\widehat{y}_i$ ). This makes violations of the linearity assumption more visually obvious. Your textbook makes this important point repeatedly (e.g. pages 197-198 and pages 702-704). Always check a graph of the residuals versus  $y$ -hat when checking for violations of the linearity assumption.



**Figure 15:** Seemingly straight



**Figure 16:** But actually curved

## 9 Limitations of logarithmic transformations

Logarithms cannot address all non-linearities: for example, a U-shaped association or any non-monotonic relationship (i.e. sometimes goes up and sometimes goes down). Taking the logarithm of a variable that sometimes takes negative values is not possible. The logarithm of zero also does not exist. If a variable contains some zeros, researchers sometimes add a small constant as a workaround. When logarithms can straighten a scatter plot they are often used because the coefficients have an easy (approximate) interpretation and using logarithms does not require adding extra variables.<sup>13</sup>

## 10 Exercises

- Q1.** The OLS line for **Figure 1** is  $\hat{y}_i = -2.32 + 156.1x_i$ . Give a *full interpretation* of each of the coefficients (i.e.  $b_0$  and  $b_1$ ).
- Q2.** The OECD (Organization for Economic Cooperation and Development) makes data about its 34 member nations publicly available (<http://stats.oecd.org/>). Consider data for 2013 with two variables: percent of females aged 15 - 64 who are employed and percent of males aged 15 - 64 who are employed. The mean female employment rate is 60.4% with a s.d. of 10.6%. For males the mean is 72.3% with a s.d. of 6.5%. How predictive is male employment of female employment? A simple regression of the female employment rate (%) on the male employment rate (%) yields  $\hat{y}_i = -16.03 + 1.06x_i$  with  $n = 34$  and  $R^2 = 0.42$ .
- Are these data cross-sectional, time series or panel?
  - Give a *full interpretation* of each of the coefficients (i.e.  $b_0$  and  $b_1$ ).
  - Does the fact that these two variables are measured as percentages mean that the coefficient is an elasticity ( $\% \Delta y$  given  $\% \Delta x$ ) like the log-log specification? Explain.
  - How predictive is the male employment rate of the female employment rate?
- Q3.** How do Olympic medals relate to a country's wealth? Let's refine the question to summer Olympics: a country's climate and topography heavily influence performance in the winter Olympics. Further, restrict attention to the 1988 and later Olympics where there are almost no boycotts<sup>14</sup> by major competitor nations: this includes 7 summer Olympics. Further, restrict attention to major competitor nations: those averaging 5 or more medals per summer Olympic games since it began competing as an independent nation or 1988 (whichever is later).<sup>15</sup> Use data on the number of gold, silver, and bronze medals won by a country in each of the summer Olympics<sup>16</sup> merged with data on the total (not per capita) real GDP of each country in each year.<sup>17</sup> Given that it takes time to invest in athletes, the GDP variable measures the GDP three

---

<sup>13</sup>An alternative is to add a squared term and higher order polynomials: these are more flexible (e.g. can accommodate non-monotonic relationships) but are a less parsimonious solution (i.e. require adding more variables). While this is possible with multiple regression analysis, more variables are only desirable if necessary.

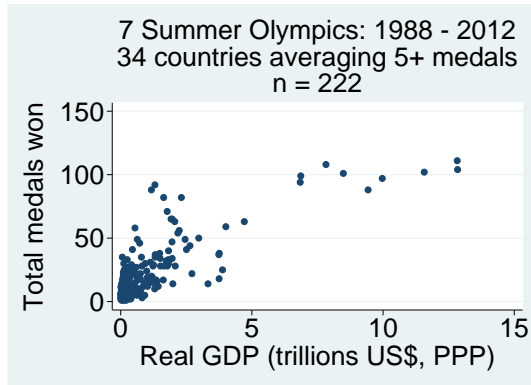
<sup>14</sup>The exception is Ethiopia in 1988.

<sup>15</sup>This includes: Australia, Azerbaijan, Belarus, Brazil, Bulgaria, Canada, China, Czech Republic, Denmark, Ethiopia, France, Georgia, Germany, Great Britain, Greece, Hungary, Italy, Jamaica, Japan, Kazakhstan, Kenya, New Zealand, Norway, Poland, Romania, Russia, South Korea, Spain, Sweden, Switzerland, The Netherlands, Turkey, Ukraine, and United States of America.

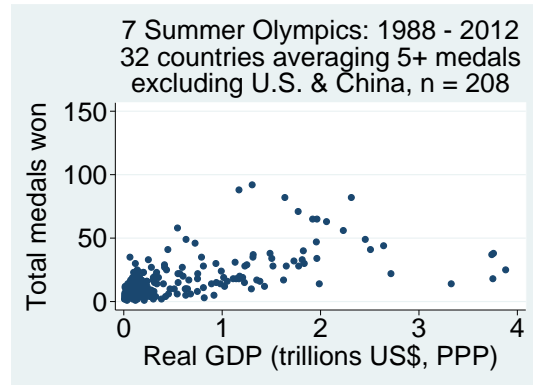
<sup>16</sup>Source is "Olympic Games" <http://www.olympic.it/english/game>.

<sup>17</sup>Source is "The Penn World Tables" <http://www.rug.nl/research/ggdc/data/penn-world-table>. In particular, "Output-side real GDP at chained PPPs (in millions 2005US\$)."

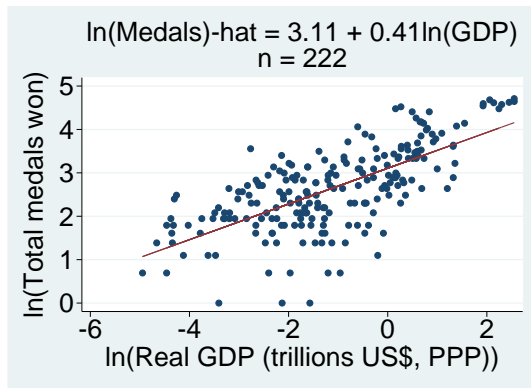
years before each Olympic games. **Figure 17** shows a highly non-linear association and some points stand out: the U.S. and China. However, **Figure 18** shows the non-linearity persists even without these two countries. **Figure 19** and **Figure 20** show logarithmic transformations applied to both variables. **Figure 21** and **Figure 22** drop the U.S. and China.



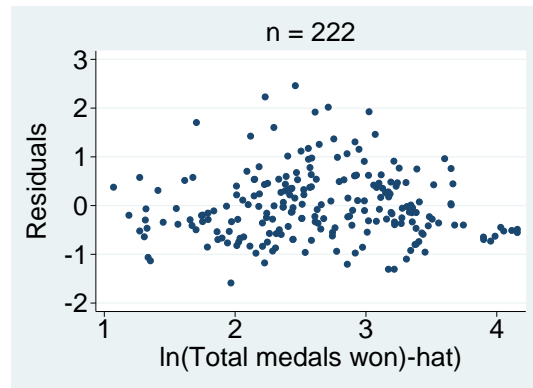
**Figure 17:** Olympic medals and wealth



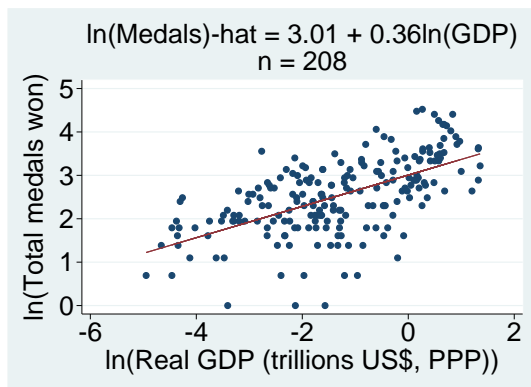
**Figure 18:** Excluding the U.S. and China



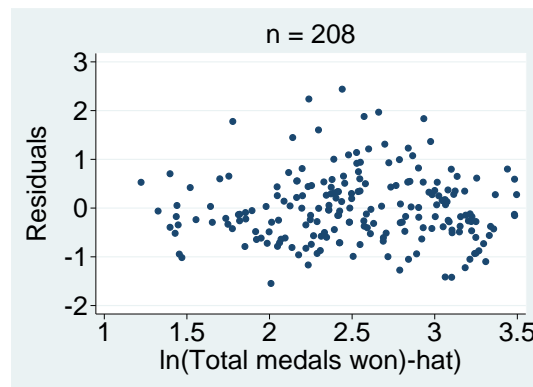
**Figure 19:** Log-log specification



**Figure 20:** Diagnostic scatter plot



**Figure 21:** Excluding the U.S. and China

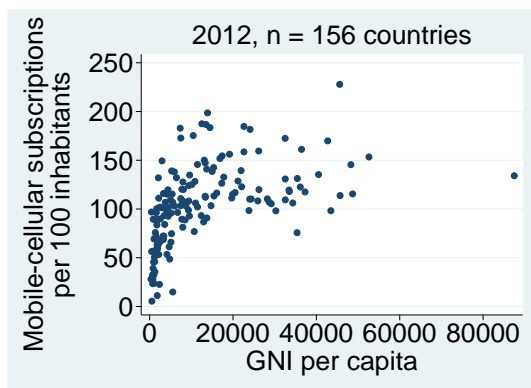


**Figure 22:** Diagnostic scatter plot

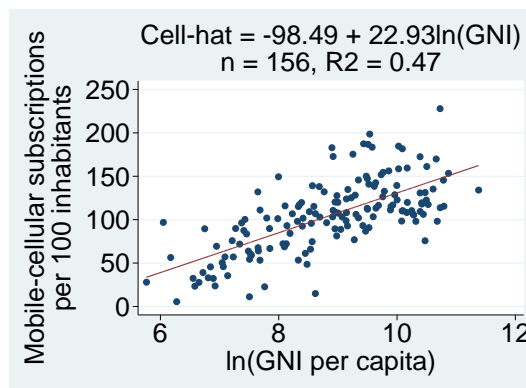
- Are these data cross-sectional, time series or panel?
- Why isn't the number of observations in **Figure 17** equal to 238 (= 7 \* 34)?
- What is the fundamental reason why it is wrong to find the OLS regression line associated with **Figure 17**?

- (d) What explains the difference in the number of observations in **Figure 17** versus **Figure 18**? Show the calculation.
- (e) The logarithm of zero does not exist. Given this, how is it possible that the number of observations in **Figure 19** is the same as in **Figure 17**?
- (f) Putting aside any concerns about violations of underlying assumptions, give a *full interpretation* of the number 0.41 that appears in **Figure 19**.
- (g) Is there a concern about a violation of linearity in **Figure 19**? How about heteroscedasticity? What about in **Figure 20**?
- (h) Putting aside any concerns about violations of underlying assumptions, give a *full interpretation* of the number 0.36 that appears in **Figure 21**.
- (i) Is there a concern about a violation of linearity in **Figure 21**? How about heteroscedasticity? What about in **Figure 22**?
- (j) If the OLS regression is  $\ln(\widehat{Medals}_{it}) = 3.01 + 0.36\ln(GDP_{it})$  when GDP is measured in trillions of US dollars, what would the OLS regression be if GDP were measured in millions of US dollars (which is actually how the data were originally reported)? Write your answer in the same format as the equation in the previous sentence but with the correct numbers.
- (k) How does your answer to the last question help explain why logarithmic transformations are popular with researchers?

**Q4.** Recall the data discussed in Section 3. **Figure 23** shows the number of mobile-cellular subscriptions per 100 inhabitants versus Gross National Income (GNI) per capita for 156 countries in 2012. **Figure 24** shows a similar scatter diagram but with a logarithmic transformation of the GNI per capita.



**Figure 23**

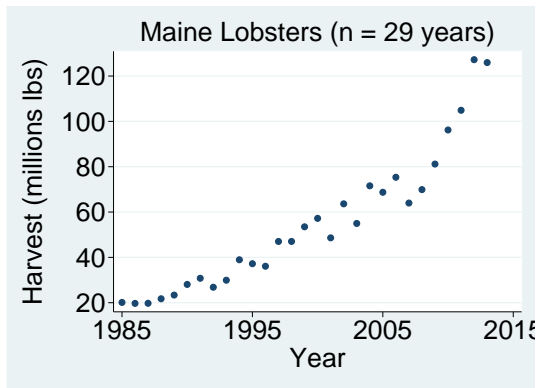


**Figure 24**

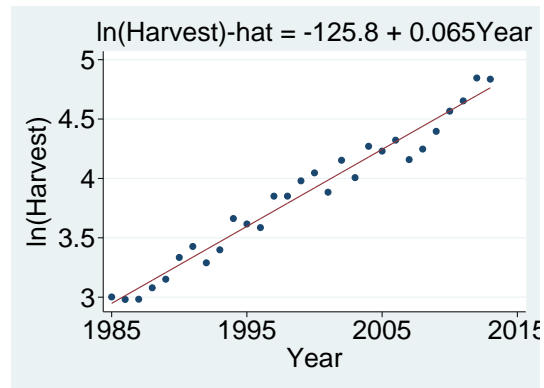
- (a) Referring to the figures, would it be appropriate to calculate the coefficient of correlation between mobile-cellular subscriptions per 100 inhabitants and GNI?
- (b) OLS regression should not be applied directly to the data shown in **Figure 23**. What is the fundamental reason why not?
- (c) What is the *full interpretation* of 22.93 in **Figure 24**?
- (d) What is the *full interpretation* of 0.47 in **Figure 24**?

- (e) Suppose you do not remember the units of measurement of GNI per capita (i.e. is it in thousands of dollars or tens of thousands of dollars or is it CAN\$ or US\$). Does that affect the interpretation of 22.93 in **Figure 24**?
- (f) The OLS regression in **Figure 24** can be written  $\widehat{Cell}_j = -98.49 + 22.93 \ln(GNI_j)$ . Supposing that mobile-cellular subscriptions are measured per 1000 inhabitants, what would that equation change to?

**Q5.** Recall the time-series lobster data discussed in Section 5 but now supplemented with the most recent publicly available data from Footnote 7. **Figure 25** shows the annual Maine lobster harvest (in millions of pounds) versus time (measured as years). **Figure 26** shows a similar scatter diagram but with the natural logarithm of harvest as the y variable. What is the *full interpretation* of 0.065 in **Figure 26**?

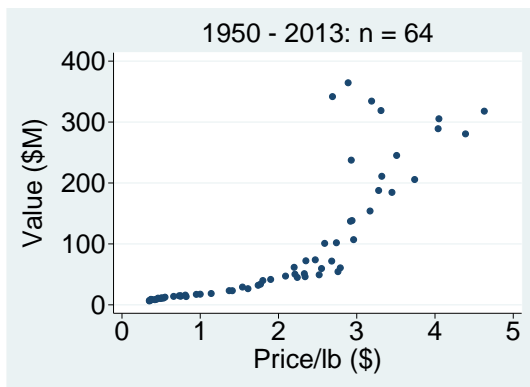


**Figure 25:** Linear specification

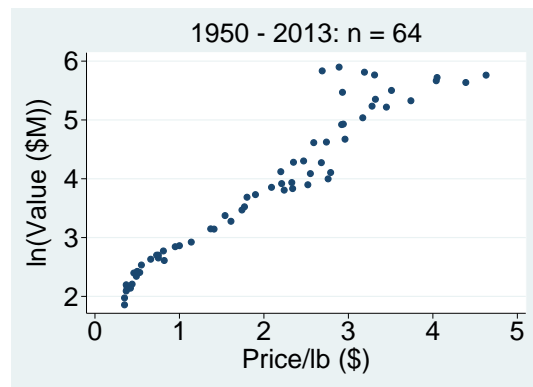


**Figure 26:** Semi-log specification

**Q6.** Review **Figure 9**: the dollar value of the Maine lobster harvest versus the price per pound over the 1950 to 2006 period. Also review **Figure 10**: the same data but with the natural logarithm of the dollar value. Since the textbook was first printed, seven more years of data have become publicly available: **Figure 27** and **Figure 28** add in seven more observations.



**Figure 27:** With 7 more years of data



**Figure 28:** With 7 more years of data

- (a) Did the functional form change after 2006?
- (b) The curved tail at low prices in **Figure 10** and **Figure 28** corresponds to the ten observations from 1950 to 1959. Does it also appear that the functional form changed after 1959?

- (c) How does this illustrate the perils of extrapolating beyond the data? (In this case trying to predict the volume of sales (measured in dollars) in the next year using price?)

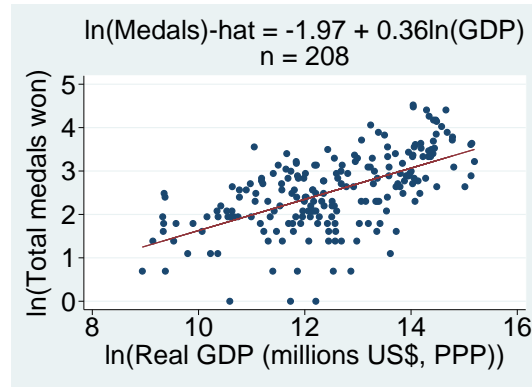
## 11 Answers to exercises

- A1.** In this linear specification the constant coefficient -2.32 is an intercept but it has no interpretation because no country has a Human Development Index (HDI) below around 0.3 (on a scale from 0 to 1) so a value of 0 is outside the range of the x variable (HDI). Because this is a linear specification the OLS coefficient 156.1 is a slope. In 2012, in countries where the Human Development Index (a number between 0 and 1) is 0.1 higher the number of mobile-cellular telephone subscriptions is 15.61 higher per 100 inhabitants on average.”
- A2.** (a) These data are cross-sectional as the unit of observation is a country. There are 34 member nations in the OECD (of which one is Canada) and in these data each of these countries in an observation ( $n = 34$ ).
- (b) The y variable is the female employment rate and the x variable is the male employment rate: y is “regressed on” x. In this linear specification the constant coefficient of -16.03 is an intercept but it has no interpretation. No OECD country has a male employment rate (the x variable) close to zero: that is far outside the range of the data. This example makes the perils of attempting to extrapolate beyond the range of the data very clear: obviously a female employment rate of -16.03 is not even theoretically possible. Because this is a linear specification the OLS coefficient 1.06 is a slope. “In 2013 amongst OECD members, in countries where the male employment rate (age 15 - 64) is 1 *percentage point* higher the female employment rate (age 15 - 64) is 1.06 *percentage points* higher on average.”
- (c) No: having a linear specification where both variables are measured as percentages does *not* correspond to a log-log specification. Remember the difference between a percent change and a percentage point change. For example, if in 2005 47 percent of consumers purchased an extended warranty with the purchase of a new television and that decreased 20 percent that means it is now 37.6 percent ( $= 0.8 * 47$ ). If it decreased 20 percentage points that means it is now 27 percent ( $= 47 - 20$ ). That is a big difference. With a linear specification of variables measured as percentages the interpretation is in percentage points. With a log-log specification involving variables measured in any units the interpretation is in terms of percents.
- (d) While there is a positive correlation between the male and female employment rates, it is fairly modest. The  $R^2$  is only 0.42, which means that only 42 percent of the variation in the 2013 female employment rate across OECD members can be explained by variation in the 2013 male employment rate across OECD members. The coefficient of correlation is 0.65 ( $= \sqrt{0.42}$ ). Some countries must have substantially higher or lower female employment rates than would be expected given the male employment rates: i.e. there is a fair amount of scatter.
- A3.** (a) These data are panel data as the unit of observation is a country in an Olympic year.
- (b) The number of observations is less than 238 (i.e. 222) because not all of the countries participated as independent nations for all 7 of the included summer Olympics. For example,



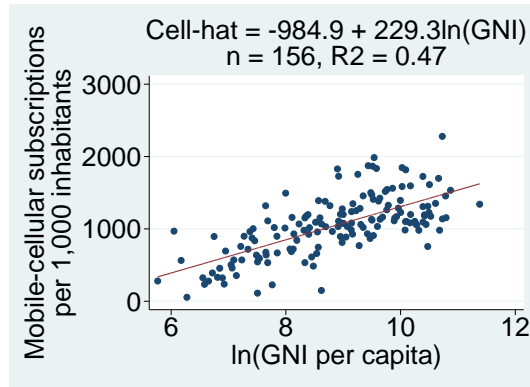
Belarus participated as an independent nation in the summer Olympics starting in 1996. For countries such as these there are fewer than 7 observations.

- (c) It does not make sense to fit a line to a non-linear relationship. This is why the linearity assumption is always listed FIRST. Often a violation of the linearity assumption will cause lots of problems (e.g. heteroscedasticity or seeming outliers) but the fundamental problem is a violation of linearity. Serious violations of linearity must be addressed before checking other underlying conditions such as homoscedasticity (equal spread).
- (d) The U.S. and China each participated in all 7 of the most recent summer Olympics. Hence,  $208 = 222 - 2 * 7$ .
- (e) None of the 34 major competitor nations participating in the most recent 7 summer Olympics earned zero medals. (Can you imagine the headlines if Canada, China, U.S., Italy, Russia, etc. came home from the summer Olympics with zero medals!) If they did then that would be an issue. This is one reason why the analysis focuses on only the 34 biggest competitor nations. (Of course no country has zero GDP.)
- (f) For the 34 biggest competitor nations for the 1988 through 2012 summer Olympics, a real GDP (three years earlier) that is 1 percent higher is on average associated with 0.41 percent more Olympic metals (gold, silver and bronze combined) won.
- (g) Yes, there is some concern about a violation of the linearity assumption in **Figure 19**. For very high levels of GDP (U.S. in all seven Olympics and China in the most recent two Olympics) the line systematically does not fit well. This is similarly apparent in the diagnostic scatter plot in **Figure 20**: that graph should show no pattern at all if the underlying conditions hold. Both figures also show some heteroscedasticity (violation of the equal spread condition). However, these violations are not severe.
- (h) For the 32 biggest competitor nations excluding the U.S. and China for the 1988 through 2012 summer Olympics, on average a real GDP that is 1 percent higher is associated with 0.36 percent more Olympic metals (gold, silver and bronze combined) won.
- (i) Both **Figure 21** and **Figure 22** show that there are no serious violations of either the linearity assumption or the homoscedasticity (equal spread) assumption.
- (j) The OLS regression would be  $\ln(\widehat{Medals}_{it}) = -1.97 + 0.36\ln(GDP_{it})$ . Why is 0.36 identical to the regression where real GDP is measured in trillions? Remember that for log-log you interpret the coefficient as the percent change in the y variable associated with a percent change in the x variable: for percent changes it does not make any difference whether units are millions or trillions of dollars. Another way to think about this is to recall that  $\ln(a * X) = \ln(a) + \ln(x)$  where  $a$  is a positive constant. Hence multiplying by a constant has no effect aside from a constant shift if a variable has been logged. Where does  $-1.97$  come from? Recall the formula for the constant term in a simple regression:  $b_0 = \bar{Y} - b_1\bar{X}$ . To go from GDP (trillions) to GDP (millions) means multiplying by 1,000,000.  $\ln(1,000,000 * GDP(trillions)) = \ln(1,000,000) + \ln(GDP(trillions))$  and  $\ln(1,000,000) = 13.82$ . Hence the mean real GDP in millions is 13.82 larger than the mean real GDP in trillions. Hence  $b_0$  will be  $-1.97 = 3.01 - 0.36 * 13.82$  (i.e. the old intercept minus the change in the mean of the x variable times  $b_1$ ).



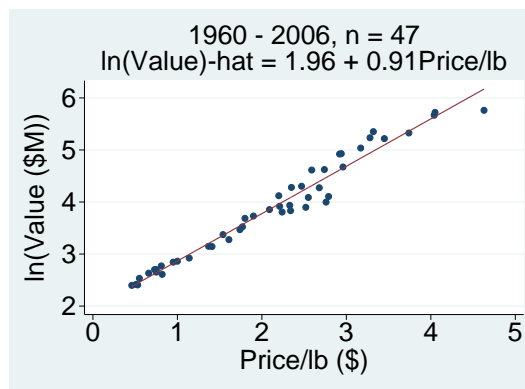
**Figure 29:** Changing the units of measurement of a logged variable by multiplying by a constant has no affect on  $b_1$ . It affects the constant term.

- (k) With logarithms you do not need to worry about multiplicative changes of units (e.g. from trillions to millions of dollars, from inches to cm, from an index on a 0 to 1 to 0 to 100 scale). In contrast, for a simple linear specification, changing the units will change the slope and its interpretation: you must know the units of measurement.
- A4.** (a) No. The coefficient of correlation (like the OLS line,  $R^2$ , and covariance) is based on the assumption that there is a linear relationship between two variables. Using the coefficient of correlation to summarize the strength of a non-linear relationship is as silly as measuring the circumference of someone's head to summarize her/his intelligence. **Figure 23** clearly shows a non-linear relationship with diminishing marginal returns.
- (b) The fundamental problem is a violation of the linearity assumption and NOT heteroscedasticity or outliers. Yes, there is heteroscedasticity and apparent outliers in **Figure 23** but these are by-products of the violation of linearity. **Figure 24** shows that once the non-linearity is addressed those other issues simply disappear.
- (c) In 2012 in countries with Gross National Income that is 10 percent higher on average 2.3 more mobile-cellular subscriptions per 100 inhabitants are observed. (Note: You could have used 1 percent and 0.23.)
- (d) The  $R^2$  of 0.47 must be interpreted in light of the logarithmic transformation: 47 percent of the variation in mobile-cellular subscriptions per 100 inhabitants across countries is explained by variation in the natural logarithm of GNI across those countries. Unfortunately it does get harder to understand  $R^2$  values when logarithms are involved because people do not naturally think in terms of logarithms. The  $R^2$  does still give a good sense of the amount of scatter in the scatter diagram of the transformed data.
- (e) No, it does not affect the interpretation because GNI is logged and the interpretation relates to a percent change, which is not affected by multiplying by a constant (i.e. changing dollar units from thousands to tens of thousands or converting US to CAN dollars or vice versus in a single year (i.e. at a fixed exchange rate)).
- (f) It would change to  $\widehat{Cell}_j = -984.9 + 229.3\ln(GNI_j)$ . **Figure 30** illustrates graphically.



**Figure 30:** Changing the units of a non-logged y variable affects  $b_0$  &  $b_1$ .

- A5.** Over the 1985 to 2013 time period the Maine lobster harvest as measured by weight increased by about 6.5 percent per year on average. Note: You MUST specify it is measured by weight (and not dollars) because prices have not been constant over this period so that is not a simple change in units. However, it is not important to specify that it is in millions of pounds: even if it were measured in kilograms or thousands of pounds the 6.5 percent number would hold. This corresponds to an approximation of the geometric mean of the growth rate (compound growth rate): see page 103 - 104 of the textbook. (Note: This is one case where the textbook directly addresses interpretation: see page 705.)
- A6.** (a) Yes, the functional form has changed. With these new data, there is no way to straighten the scatter plot with any combination of logarithms. Sometimes logarithms do not work.
- (b) Yes. The semi-log functional form did not correct non-linearity for the first decade of data (in fact it made it worse). For the 1960 - 2006 period the functional form is fairly constant: the semi-log approach does fairly well at straightening the scatter plot. See **Figure 31**.



**Figure 31:** Fairly constant functional form.

- (c) Beyond the observed range of the data the functional form may differ. The lobster example illustrates this well. Had you used the regression line in **Figure 31** to predict past values (1959 or earlier) or future values (2007 or later), the predictions would have been off. Even when you learn about margins of error in prediction in the second half of the course, they are based on the assumption that the functional form is constant: the statistical margins of error do NOT take into account possible changes in the functional form.

## 12 Asiaphoria case study

The case study is “Asiaphoria Meets Regression to the Mean,” *NBER Working Paper 20573*, Oct. 2014, by Lant Pritchett and Larry Summers, abbreviated as “Asiaphoria” or Pritchett and Summers (2014). This is a working paper from a prominent series: the National Bureau of Economic Research (NBER)<sup>18</sup> Working papers have not yet been published in scholarly peer-reviewed journals. Because it can take *years* for a paper to be published in a journal, working papers are valuable sources at the cutting edge of research. The NBER working paper series includes many empirical papers (i.e. papers where analyses of data feature prominently). Papers from this series are frequently cited in the popular press (e.g. *The Economist* and *The New York Times*). In fact, there is a monthly NBER digest, which, for select working papers, includes an easy-to-read summary that captures the main message of the paper in one page of text and single figure (graphic).<sup>19</sup> Fortunately, Asiaphoira is one of those select papers featured in the digest, which provides you an easy entry point into this research.

You are not responsible for the entire working paper, but rather, select portions at the level of our course. Start with the summary of Asiaphoria in the Mar. 2015 *NBER Digest* (pp. 1-2, <http://www.nber.org/digest/mar15/mar15.pdf>). In the paper itself, read the Abstract and Sections 1, 2.1, and 6 (pp. 1-11, 56-59, <http://www.nber.org/papers/w20573.pdf>). See the “**Readings**” folder in portal if you have any trouble accessing these required readings or wish to access them off-campus.

### 12.1 Why this case study is important: goals and help for you

An important goal of ECO220Y is to prompt you to reach a sufficiently deep understanding of quantitative methods in economics (the title of our course) such that you can read, understand and critically evaluate real research that uses these methods. These are difficult skills to acquire.

Researchers usually presume that you understand foundational concepts/methods. Unlike your textbook and other course materials, research papers are not written with the goal of teaching core concepts. Instead, researchers apply concepts/tools to data to draw fresh insights about the world. A well-written research paper *will* explain these insights, how these are reached, and any limitations. However, researchers use a wide variety of terms and references, which presents a challenge to students looking for clear links between the research and what is learned in courses. Your ECO220Y materials include common synonyms and many examples of empirical research from a range of fields, with the goal of building a robust understanding. But there is always *at least* a little leap from your course work to a particular research paper. A goal of this case study is to help you practice making that leap.

As you may have guessed, given that this case appears at the end of “Logarithms in Regression Analysis,” Asiaphoria is a great example of current research that features logarithms and simple

---

<sup>18</sup>The NBER describes its working paper series as “NBER researchers initially report their findings in scientific papers aimed at other professional economists in academic institutions, business, government, and the business media around the world. Nearly 700 NBER Working Papers are published each year, and many subsequently appear in scholarly journals” (retrieved July 4, 2016 from <http://www.nber.org/pubs.html>).

<sup>19</sup>“The Digest is a monthly publication that summarizes at least four recent and newsworthy NBER Working Papers. Professional journalists write these summaries for a non-technical audience” (retrieved July 4, 2016 from <http://www.nber.org/pubs.html>).

regression used descriptively. In fact, it is rare for a key message of current research to be based on simple regression (most involve multiple regression or something even more complicated) and to be centered around a descriptive rather than causal question. Hence, aside from the timely and important topic of the research (economic growth across countries and over time), it is an ideal example for ECO220Y because of the prominent use of techniques and concepts we cover very early in the course and keep coming back to.

You are expected to carefully study the assigned portions of the research paper and to figure it out. While it is within your grasp and within our course, you will have to work on this case study. With study, you can appreciate the many links with various parts of our course material. Our goal is not to *teach* the case: that would rob you of the chance to make the leap from course material to real research, which is a skill you need in our course and beyond. That said, we do want to help enable you to productively do the necessary work yourself. First, Section 12.2 gives a reading guide, prepared with the goal of helping you making the leap. Think of this as getting hiking boots and a trail guide with tips about tricky portions of the trail, *not* as being carrying through the woods. Subsection 12.2.4 of this reading guide gets you started on replicating the working paper's key results with Excel using the original data that are publicly available. Replicating results is a great way to deepen your understanding. Also, because this case study appeared in 2015/16, you can see some old test questions in Subsection 12.2.5 to get a sense of the depth of understanding required (while remembering that there are *many* potential test questions). Additional help includes the usual course resources (e.g. TA office hours) for your *specific* questions. Work on the case study and figure out where you are stuck and ask for help on that sticking point. Do *not* expect help in response to questions like "Can you explain Asiaphoria to me?" or "I didn't have time to carefully do the readings, can you summarize what I need to know?"

## 12.2 Reading Guide for "Asiaphoria Meets Regression to the Mean"

This reading guide helps your study of Pritchett and Summers (2014). However, it does not walk you through the paper and does not reteach concepts from lectures, required readings, and homework. While the digest and the assigned sections of the paper are mostly straightforward, you may need help to fully understand how Tables 1 and 2 are constructed, which is *important*. The methods are all within our regular curriculum, except for the rank correlation, which Section 12.2.3 explains.

Read this guide, the digest and the assigned sections of the paper. You will likely need to read this guide at least twice (once before and once after reading the working paper). You may also need to brush up on regression topics using your regular course materials (e.g. if you are unsure about the  $R^2$ ). Also, make sure you can fully interpret the numbers in Tables 1 and 2.

For graded work, you are responsible for the digest, the assigned sections of the paper (not the entire paper) and having used this guide to assist your study. While some excerpts from the digest and/or the paper may appear on tests, you will not have copies of the readings or this guide. The tests may also explore the data in ways not presented in the original research. Also, since Pritchett and Summers (2014), updated data have been published. The tests may use the updated data, so the results may differ.

### 12.2.1 The data used in the analysis for Tables 1 and 2

Tables 1 and 2 of Pritchett and Summers (2014) use high-quality publicly available data: the Penn World Table (PWT) version 8.0. These are important data used by many researchers and an excellent opportunity for you to see how such data are presented and documented.

PWT 8.0: These data are available for download at <http://www.rug.nl/research/ggdc/data/pwt/pwt-8.0>. “PWT version 8.0 is a database with information on relative levels of income, output, inputs and productivity, covering 167 countries between 1950 and 2011. Released on: July 2, 2013 (DOI: 10.15141/S5159X)” (retrieved June 12, 2015).

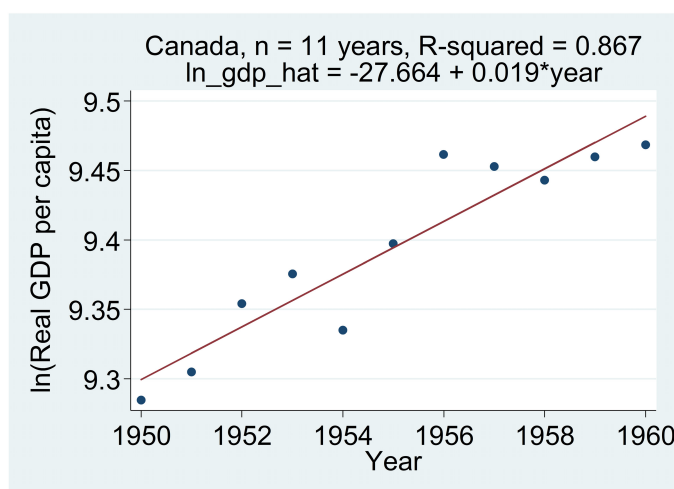
Visit this site and view the data: it is available in Excel format. The relevant measure of GDP (there are many) is “Real GDP at constant 2005 national prices (in mil. 2005US\$)” (`rgdpna`), which is described as “Real GDP using national-accounts growth rates, for studies comparing (output-based) growth rates across countries.” This number is divided by population (also in millions) to obtain real GDP per capita in 2005 US\$. Also, since version 8.0, there have been two updates:

PWT 8.1: These data are available for download at <http://www.rug.nl/research/ggdc/data/pwt/pwt-8.1>. “PWT 8.1 is an updated version of PWT 8.0, covering the same countries and period. Released on: April 13, 2015 (DOI: 10.15141/S5NP4S)” (retrieved June 8, 2015). Also, a published scholarly journal article describes these data: Feenstra, Robert C., Robert Inklaar and Marcel P. Timmer. 2015. “The Next Generation of the Penn World Table.” *American Economic Review*, 105(10): 3150-82 ([http://www.rug.nl/research/ggdc/data/pwt/v81/the\\_next\\_generation\\_of\\_the\\_penn\\_world\\_table.pdf](http://www.rug.nl/research/ggdc/data/pwt/v81/the_next_generation_of_the_penn_world_table.pdf))

PWT 9.0: These data are available for download at <http://www.rug.nl/research/ggdc/data/pwt/pwt-9.0>. “PWT version 9.0 is a database with information on relative levels of income, output, inputs and productivity, covering 182 countries between 1950 and 2014. Released on: June 9, 2016 (DOI: 10.15141/S5J01T)” (retrieved July 5, 2016).

### 12.2.2 Getting from the PWT 8.0 data to Tables 1 and 2

Pritchett and Summers (2014) describe each of the steps necessary to replicate Tables 1 and 2. The first issue is that the PWT data describe the *levels* of GDP in each country in each year. However, their research question is about *growth rates*, not levels. To get the required data on growth rates they run simple regressions. In fact, the authors run many regressions – one for each country in each time period (either a decade or two decades) – to obtain a measure of the GDP growth rate in that country in that decade (or two decades). For example, Figure 32 shows the regression to obtain Canada’s growth rate from 1950 to 1960. The unit of observation is a year.



**Figure 32:** Simple regression using time-series data on annual real GDP per capita in Canada from 1950 - 1960.

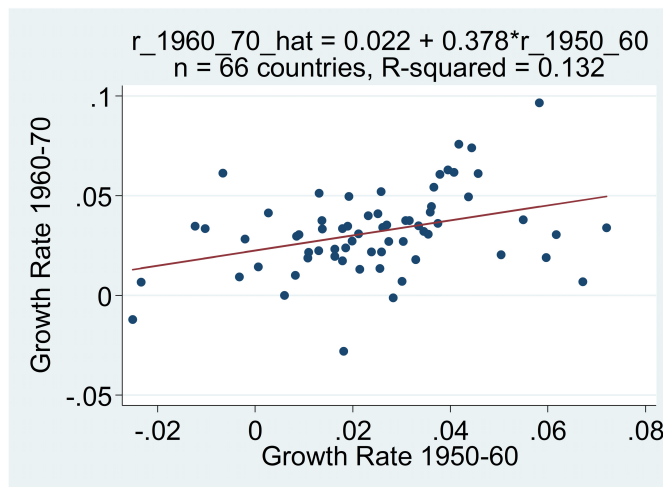
Here are all the OLS regression results for Canada for the time periods relevant for Tables 1 and 2.

Simple regressions for Canada:

1950-1960:  $\ln\_gdp\_hat = -27.664 + 0.019 \cdot year$ , n = 11  
 1960-1970:  $\ln\_gdp\_hat = -58.332 + 0.035 \cdot year$ , n = 11  
 1970-1980:  $\ln\_gdp\_hat = -45.625 + 0.028 \cdot year$ , n = 11  
 1980-1990:  $\ln\_gdp\_hat = -28.484 + 0.019 \cdot year$ , n = 11  
 1990-2000:  $\ln\_gdp\_hat = -30.747 + 0.021 \cdot year$ , n = 11  
 2000-2010:  $\ln\_gdp\_hat = -7.914 + 0.010 \cdot year$ , n = 11  
 1950-1970:  $\ln\_gdp\_hat = -38.907 + 0.025 \cdot year$ , n = 21  
 1960-1980:  $\ln\_gdp\_hat = -51.114 + 0.031 \cdot year$ , n = 21  
 1970-1990:  $\ln\_gdp\_hat = -31.441 + 0.021 \cdot year$ , n = 21  
 1980-2000:  $\ln\_gdp\_hat = -18.646 + 0.014 \cdot year$ , n = 21  
 1990-2010:  $\ln\_gdp\_hat = -26.730 + 0.019 \cdot year$ , n = 21

This is done not just for Canada, but each country. The OLS “slope” coefficients become the raw data for the OLS regressions reported in Tables 1 and 2. **Note well:** Make sure you can interpret the OLS slope coefficients in Figure 32 the all the Canada regressions above before reading on. (If you have trouble, review the earlier parts of this document including the exercises in Section 10.)

The regressions reported in Tables 1 and 2, look at how well growth rates from a previous time period predict subsequent growth across countries. For example, consider the first row of results in Table 1. That looks at how well growth rates during the period from 1950 - 1960 predict growth rates during 1960 - 1970. Below is the scatter plot and the OLS regression for that forecast. The unit of observation is a country. **Note well:** Make sure you can match up the results in Figure 33 with the first row of results in Table 1.



**Figure 33:** Simple regression using cross-sectional data on countries' GDP growth rates in the 1950s to predict growth rates in the 1960s.

Table A gives an excerpt of the data used in Figure 33. Look at the row for Canada and make sure you see the connection between each of the two numbers in that row and Figure 32 and the simple regressions for Canada reported just after Figure 32.

**Table A:** Excerpt of the data used to estimate the regression reported in the first row of Table 1 in Pritchett and Summers (2014)

country	countrycode	r_1950_60	r_1960_70	...
Argentina	ARG	0.013	0.022	
Australia	AUS	0.018	0.034	
Austria	AUT	0.055	0.038	
Belgium	BEL	0.023	0.040	
Bolivia	BOL	-0.023	0.007	
Brazil	BRA	0.037	0.036	
Canada	CAN	0.019	0.035	
Chile	CHL	0.011	0.022	
China	CHN	0.050	0.020	
...				
Venezuela	VEN	0.027	0.027	
Zimbabwe	ZWE	0.030	0.007	

Tables 1 and 2 show many regressions like Figure 33: 15 regressions in total. Each of these regressions use data created from the results of the (many) country-level regressions (like the illustrative set shown for Canada). That's right: regressions are run on data that have been created by running regressions. **Note well:** Make sure you can fully interpret all of the regression results in Tables 1 and 2 in Pritchett and Summers (2014).



### 12.2.3 Rank correlation

You should now have figured out exactly where all of the numbers in Tables 1 and 2 come from and what they mean, with one exception. There is a column labeled “Rank Correlation.” This is an interesting topic that is closely related to the correlation statistic you have already studied.

The rank correlation is simply the (regular) correlation between two rankings. The formal name for the rank correlation is the Spearman rank correlation. In contrast, the formal name for the (regular) correlation that you have already studied is the Pearson correlation. (If someone simply writes “correlation,” it nearly always means the Pearson correlation.) To find the rank correlation, instead of calculating the correlation between growth rates in 1950 - 60 with growth rates in 1960 - 70, rank countries from slowest to fastest growth and compute the correlation of their rankings. In other words, does a country with the slowest growth (rank 1) in one decade tend to have the slowest growth in the next decade (i.e. still stuck at rank 1)? A perfect rank correlation of 1 would mean that the rankings of countries are unchanged. (Note: This does not mean a perfect regular correlation.)

The rank correlation has two notable differences from the regular correlation: it can be used for non-linear associations (so long as they are monotonic) and it is not sensitive to outliers (because no matter how extremely poorly a country does, it cannot be ranked lower than 1). To illustrate, Table B gives an excerpt of data used in Pritchett and Summers (2014). The correlation between `r_1950_60` and `r_1960_70` is 0.363 whereas the rank correlation is the (regular) correlation between `rank_1950_60` and `rank_1960_70` and is 0.381. In Tables 1 and 2, the correlation and rank correlations are comparable. You may feel that that makes this extra column redundant. However, it is useful information for the reader. It makes it clear that the low (regular) correlations are not being caused by an outlier or a violation of the linearity assumption.

**Table B:** Excerpt of the data used in Pritchett and Summers (2014)

country	countrycode	r_1950_60	r_1960_70	...	rank_1950_60	rank_1960_70	...
Argentina	ARG	0.013	0.022		16	22	
Australia	AUS	0.018	0.034		22	37	
Austria	AUT	0.055	0.038		61	48	
Belgium	BEL	0.023	0.040		31	49	
Bolivia	BOL	-0.023	0.007		2	5	
Brazil	BRA	0.037	0.036		52	44	
Canada	CAN	0.019	0.035		26	40	
Chile	CHL	0.011	0.022		15	19	
China	CHN	0.050	0.020		60	18	
...							
Venezuela	VEN	0.027	0.027		39	25	
Zimbabwe	ZWE	0.030	0.007		41	7	

#### 12.2.4 Replicating results reported in Tables 1 and 2 in Pritchett and Summers (2014) using Excel (or other software)

As discussed in Section 12.2.1, you should download the PWT version 8.0 data and replicate some of the results in Tables 1 and 2 in Pritchett and Summers (2014). Replication means carefully reading the paper and reproducing the exact reported results. To help you along, here are some tips:

1. Make sure you use version 8.0 of the PWT data.
2. The relevant real GDP per capita measure is created as  $\text{real\_GDPPC} = \text{rgdpna}/\text{pop}$ , where  $\text{rgdpna}$  and  $\text{pop}$  are each variables in the PWT data. (The documentation of those data gives the exact variable definitions.)
3. In footnote 5 (p. 8) Pritchett and Summers (2014) write “we exclude countries with less than 25 years of data” and “we exclude Equatorial Guinea because it has a small population and is frequently a massive outlier.” Hence, your replication should drop those countries as well.
4. When finding the decade growth rate for a country, include the endpoints. Notice that Figure 32 includes both 1950 and 1960. Further, in footnote 5 (p. 8) Pritchett and Summers (2014) write “we calculate a growth rate if there is more than 7 years of data for the 10-year growth rates.” However, to be precise and to exactly replicate the results in Tables 1 and 2 this should say “we calculate a growth rate if there is 7 or more years of data for the 10-year growth rates.” (In other words, decades with exactly 7 years of data for a country *are* included.) (Similarly, for the 20-year periods, periods with exactly 14 years of *are* included.)
5. In replicating the Spearman rank correlation results, be careful with missing values in the early time periods. Specifically, you must compare the same group of countries in two different periods. If a particular country has missing data for either of the two periods it must be dropped completely before you rank the countries in each period.

#### 12.2.5 Sample old test questions on Asiaphoria

Some old test questions are publicly posted at [http://homes.chass.utoronto.ca/~murdockj/eco220/#old\\_tests\\_exams](http://homes.chass.utoronto.ca/~murdockj/eco220/#old_tests_exams). Here are the ones on Asiaphoria (arranged in a recommended order for your completion):

1. Test #4, Mar 2016: Question 3
2. Test #5, Apr 2016: Questions 20 - 28
3. Test #1, Oct 2015: Questions 32 - 35
4. Test #2, June 26, 2015: Question 1 and Test #2, Dec 2015: Question 2 (Note: These questions show much more similarity than usual between a test question and an old test question because the earlier one was a restricted exam.)

Make sure to completely solve the old test questions yourself and ONLY use the solutions/rubrics to grade your answer once you feel confident you have prepared an A+ answer. Otherwise these old test questions will not help you as much. There are many possible questions around this topic and old test questions are not a study guide. They are useful because they prompt you to think and apply your understanding and they give you an early signal about your depth of understanding.