1. Big Picture and Notation

Definitions

- $y = f(x_1, x_2, \cdots, x_k, u) = \beta_0 + \beta_1 x + u$: This is the population regression equation of y on x. We assume that this is the true data generating process. We often assume the relationship is linear.
- $\mu = y - \beta_0 - \beta_1 x$: The variable μ is called the error term, or disturbance in the relationship, and represents factors other than x that affect y. The disturbance arises for several reasons, primarily because we cannot hope to capture every influence on an economic variable in a model.
- $E(y|x) = \beta_0 + \beta_1 x$: This is the **population regression function** (PRF), $E(y|x)$ is a linear function of x. The linearity implies that a one-unit increase in x changes the *expected* value of y by the amount β_1. For any given value of x, the distribution of y is centered about $E(y|x)$.

  Note this definition relies on the assumption (which we will investiage later) that $E(u|x) = E(u)$, which is essentially saying that no observations on x convey any information about the expected value of the distrubance.[1] We can also write:

  $$y_i = E(y|x) + u_i$$

  This says that any variable y_i can be decomposed into a piece that is explained by x, $E(y|x)$, and some piece that is left over u, which we don't observe.
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$: This is the fitted regression line. It can be thought of as our best guess for y given a certain value of x. This equation is also called the **sample regression function** (SRF) because it is the estimated version of the PRF.
- $y = \hat{\beta}_0 + \hat{\beta} x + \hat{\mu} = \hat{y} + \hat{\mu}$: This is now our estimated model. The hat symbol above our beta's indicate that these are calculated estimates of the true beta value they represent. Again we see how we can decompose y_i into two parts: a fitted value (best guess) and a residual.
- $\hat{\mu} = y - \hat{y}$: The variable $\hat{\mu}$ is called the residual, it can be thought of as the deviations between the real y_i value and the predicted $\hat{y}_i$ value.

Graphs

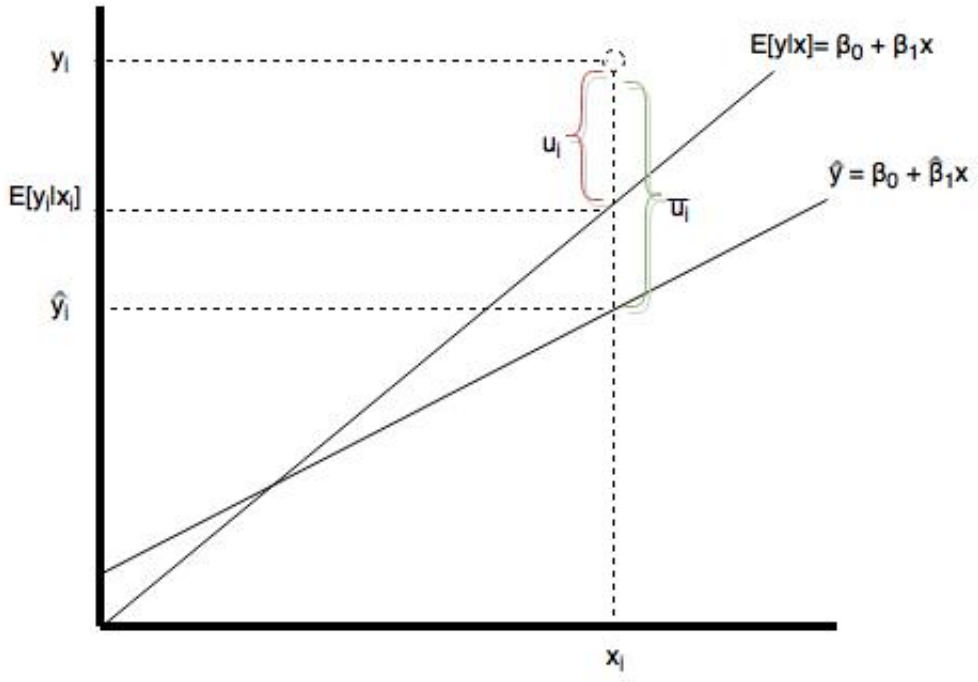
- Figure 1 illustrates each one of the concepts above in turn
- Figure 2 plots the population regression function of log weekly wages given schooling for men from the 1980 US census (assume for illustration purposes, we interviewed the entire population men in the country). The distribution of earnings is also plotted for a few key values: 4,8,12, and 16 years of schooling.
  - The PRF tells us how the average value of y changes with x: it does not say that y equals $\beta_0 + \beta_1 x$ for all units in the population. For example, suppose x=4, then on average this implies log weekly earnings of 5.9 dollars. This does not mean that everyone with 4 years of schooling makes 5.9 dollars.

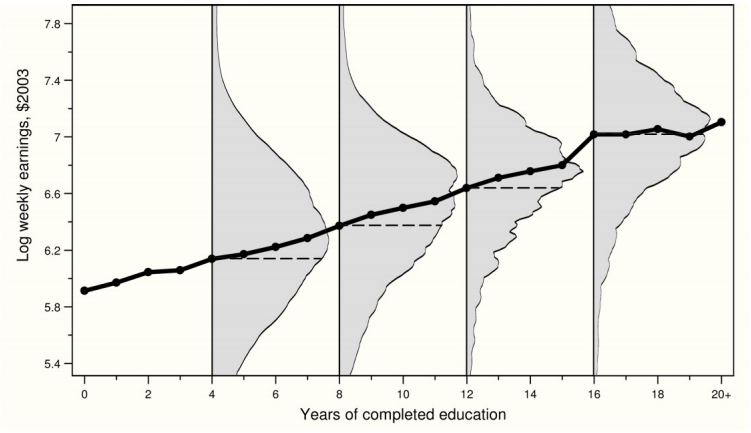
[1] This assumption will allow us to interpret the β coefficient (in the population) as the causal effect of an additional unit of x on the expected value of y. We can still fit a line to our data without this assumption, but we won't be able to interpret the estimate as causal. More on this later (but think of investigating the impact of education on income. If there is something unobservable like ability that varies with the level of education (higher educated people also have more ability) such that $E[u|x] \neq 0$, then we wont be able to say that the coefficient associated with education reveals the true effect of education on income because we are confounding the effect of ability and education

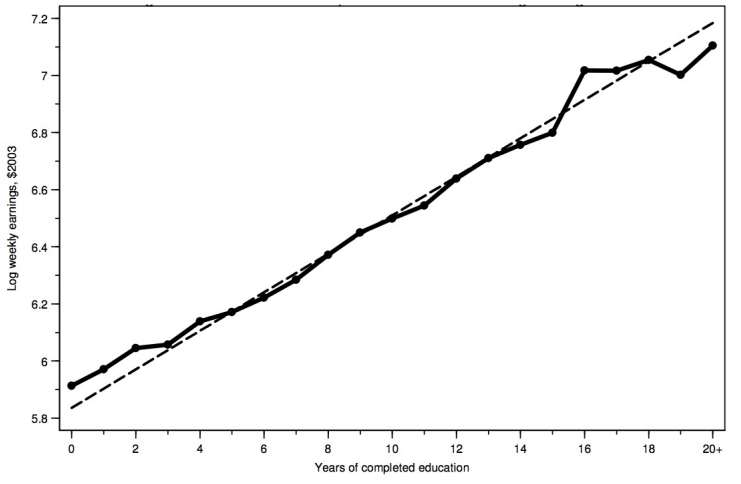
  - In this picture the PRF isn't actually linear, but for the purposes of this class we assume that it is.

- Figure 3 shows the fitted regression line for a sample of these men drawn from the census (sample regression function). The ideal case is for $\hat{u}_i = 0$, so that the line exactly predicts y_i. But in most cases every residual is not equal to 0, as can be seen on the figure.

  - This graph shows us superimposing the true population regression equation with the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$ we estimate.







2. Properties of $\hat{\beta}_0$, $\hat{\beta}_1$

i. Deriving the Estimators

In lecture, we considered the model $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, and we were given the following formulas for computing $\hat{\beta}_0$ and $\hat{\beta}_1$, which you will use on your problem set:

$$\hat{\beta}_1 = \frac{s_{xy}(x,y)}{s_{x^2}} = \frac{cov(x,y)}{var(x)} == \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x$$

Where did these formulas come from? One way to derive them is to recognize the fact that we want our regression line to minimize the distance between the observed y value and the predicted value $\hat{y}$. In other words we want to to make the set of residuals we obtain very small. There is some debate as to how to go about doing this (some advocate minimizing the absolute value of the residuals, while others argue for minimizing the sum of squared residuals). Minimizing the sum of squared residual gives more weight to large residuals, that is, outliers in which predicted values are far from actual observation (think of a line of best fit that is trying to "accomodate" the large outliers). More importantly for our puposes, this approach will produce an estimator with desirable properties (more on this later). This is what we refer to as **OLS**. Note we would never choose our estimates to minimize, say, the sum of residuals themselves , as residuals large in magitude but with opposite signs would tend to cancel out.

In the examples from class, the daily assignment, and the problem set, you were asked to calculate each term $(x - \bar{x})$, $(y - \bar{y})$ and plug in to the formula to comute $\hat{\beta}_1$ and then $\hat{\beta}_0$.

Question: Why are we using $\bar{x}$ rather than $E(x)$. Answer: because we only have the sample of values we drew, and not the entire population.

Derivation

Let's define W as we did in class, plugging in our model for $\hat{y}_i$:

$$W = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We'd like to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that W is as small as possible. To do this we solve the following minimization problem with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} W = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Taking the first order conditions (partial derivatives):

$$\frac{\partial W}{\partial \hat{\beta}_0} = -\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (1)$$

$$\frac{\partial W}{\partial \hat{\beta}_1} = -\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \quad (2)$$

These equations can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$. Starting with equation (1)

$$2\left[-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\hat{\beta}_0 + \sum_{i=1}^{n}\hat{\beta}_1 x_i\right] = 0 \qquad \text{Distribute the Summation}$$

$$\left[-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\hat{\beta}_0 + \sum_{i=1}^{n}\hat{\beta}_1 x_i\right] = 0 \qquad \text{Get rid of the 2}$$

$$\sum_{i=1}^{n}\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{\beta}_1 x_i \qquad \text{Re-arranging terms}$$

Since β_0 and β_1 are same for all cases in the original linear equation, this further simplifies to:

$$n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

This is our final expression for β_0. Going back to equation (2) we will solve for β_1:

$$-\sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

$$\sum_{i=1}^{n} 2x_i(-y_i + \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i(-y_i + \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i(-y_i + (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i(-y_i + (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i(\bar{y} - y_i + \hat{\beta}_1(x_i - \bar{x})) = 0$$

$$\sum_{i=1}^{n} x_i(\hat{\beta}_1(x_i - \bar{x})) = \sum_{i=1}^{n} x_i(y_i - \bar{y})$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i(y_i - \bar{y})$$

From some properties of summation operation (see Appendix A.1 Woolridge for the full set of steps)

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1}$$

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{2}$$

Then plugging (1) and (2) into our previous expression:

$$\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

This is the expression we are familiar with.

ii. Method of Moments

There is another way to estimate the parameters β_0 and β_1 that was introduced in class: Method of Moments. The method of moments relies on two assumption:

1. $E(u) = 0$: this is simply an assumption about the distribution of the unobservables *in the population*. You can think of this as normalizing all the unobserved factors affecting y, so that their mean is zero
2. $E(u|x) = E(u) = 0$: this is the conditional mean assumption. Taking the example in class (which assumed that $u = ability$) this just says that the average ability of individuals in the population is the same regardless of the years of education.

Then from Assumption 2 we were able to get that $E(xu) = 0$:

$$Cov(x, u) = E[xu] - E[x]E[u] = E[xu] - E[x] \times 0 = E[xu]$$

Then using the Law of Iterated Expectations:

$$\begin{aligned} Cov(x, u) &= E[x, u] \\ &= E[E[xu|x]] \\ &= E[xE[u|x]] \\ &= E[xE[u]] \\ &= 0 \end{aligned}$$

And since $Cov(x, u) = E[xu]$, then $E[xu] = 0$.

Rewriting $E(u) = 0$ and $E[xu] = 0$, we have:

$$E(y - \beta_0 - \beta_1 x) = 0 \quad (1)$$
$$E[x(y - \beta_0 - \beta_1 x)] = 0 \quad (2)$$

Then give the sample of data we have, we will choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the *sample counterparts* of equations (1) and (2). Which gives:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (3)$$
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \quad (4)$$

Then, solving as we did for OLS, we get the same results for $\hat{\beta}_0$, and $\hat{\beta}_1$ (Note the fact that OLS=MOM is not generalizable. It is true when we assume a linear relationship between x and y).

iii. Interpreting estimates: Sign, Significance, Size

Whenever we ask you to "interpret" your estimated results, you need to address these three characteristics of the each coefficient you are examining:

1. **Sign:**
   - What sign did you expect the estimated parameter to have? Why?
   - Does your estimate have this sign (i.e. are you surprised or reassured by your results)?

   This is your opportunity to use your common sense, and state your prior hypothesis based on your real-world intuition. For example: I expect higher levels of education to be associated with higher income.

2. **Significance:**
   - Is the estimate statistically different from zero?
   - What is the t-statistic of this hypothesis?

   More on this later.

3. **Size:**
   - How do changes in this variable affect the dependent variable according to your estimation?
   - Is this an economically meaningful effect size?

   The answer to the first question will depend on the functiona form (see next section). The answer to the second is fairly open-ended. I suggest looking at summary statistics (mean, median, variance) of the variables in question for some perspective on size.

3. Functional Forms Review

Preliminaries: Proportions, Percentages, and Approximations

Suppose our variable of interest x has an initial value of x_0 and then increases to x_1.

- A *proportional change* in x: $\frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x_0}$
- A *percentage change* in x: $\frac{x_1 - x_0}{x_0} \times 100 = \frac{\Delta x}{x_0} \times 100$

Quick Example from World Bank Data

China's GDP per capita was 1853.45 in 1995 and in 2000 it had grown to 2673.66. Let's call the 1995 value x_0 and the 2000 value x_1.

- Then the proportional change is $\frac{\Delta x}{x_0} = \frac{820.21}{1853.45} = 0.4425$
- and the percentage change is $\frac{\Delta x}{x_0} \times 100 = 44.25\%$.

In lecture, we used the fact that $\Delta \log x = \frac{\Delta x}{x}$, i.e. that the change in the log of a variable is equal to the proportional change in the variable itself. Proof:

$$
\begin{aligned}
y &= y_0 + f'(x_0)(x - x_0) && \text{Equation of a Tangent Line at } x_0, y_0 \\
\Delta y &= f'(x_0)\Delta x \\
\Delta y &= \frac{1}{x_0}\Delta x && \text{Letting } \; y = f(x) = ln(x) \rightarrow f'(x_0) = \frac{1}{x_0} \\
\Delta y &= \frac{\Delta x}{x_0}
\end{aligned}
$$

Elasticities

Definition: the percent change in one variable in response to a given percent change in another variable, holding all other relevant variables constant. In other words it summarizes the responsiveness of one variable to a change in another variable.

$$E = \frac{\text{percent change in z}}{\text{percent change in x}} = \frac{\Delta z / z}{\Delta x / x} = \frac{\partial z}{\partial x}\frac{x}{z}$$

Note: As $\Delta x \rightarrow 0$, $\Delta z / \Delta x$ goes to the partial derivative $\frac{\partial z}{\partial x}$. Economists usually calculate elasticities only at this limit, i.e. for infinitesimal changes in x.

Example: At a point on a supply curve where the elasticity of supply $\eta = 0$, we say the supply curve is perfectly inelastic: The supply doesn't change as the price rises. If $0 < \eta < 1$, the supply curve is inelastic (but not perfectly inelastic): A 1% increase in the price causes a less than 1% rise in quantity supplied. If $\eta > 1$, the supply curve is elastic: A 1% increase in the price causes a more than 1% rise in quantity supplied.

Functional Forms and Marginal Effects

Choosing an appropriate functional form is a critical choice in econometric modeling. Your choice of model and selection of variables will greatly influence the fit of your model when mapping independent variables to your dependent variable.

1. Linear functions and Unit-Unit changes

If we assume a linear functional form, the model is: $y = \beta_0 + \beta_1 x$

Interpretation: First, take the derivative of the expression to get: $\frac{dy}{dx} = \beta_1$. Now, (though this is technically not very rigorous) for small enough changes in x and y, we can rewrite this as:

$$\frac{\Delta y}{\Delta x} \approx \frac{dy}{dx} = \beta_1$$

Then we can rearrange to see that

$$\Delta y = \beta_1 \Delta x$$

Suppose $\Delta x = 1$, so that x changes by 1 **unit**. Then we can plug this into the above expression to see that y will change by β_1 **units**.

2. Logarithmic functions and Percent-Unit changes

If we assume a logarithmic functional form, the model is: $y = \beta_0 + \beta_1 \log(x)$

Interpretation: First, take the derivative of our model, $\frac{dy}{dx} = \frac{\beta_1}{x}$ and again notice that we can rewrite this:

$$\frac{\Delta y}{\Delta x} \approx \frac{dy}{dx} = \frac{\beta_1}{x}$$

Then we can rearrange to see that

$$\Delta y = \beta_1 \frac{\Delta x}{x}$$

Suppose we know that x changes by 10 **percent**, so that the proportional change in x is 0.1: $\frac{\Delta x}{x} = 0.1$. Plug this value into the expression we derived, and we see that y will change by $\beta_1 * 0.1$ **units**.

3. Exponential functions and Unit-Percent changes

If we assume an exponential functional form, the model is: $y = e^{\beta_0+\beta_1 x}$ or $\log(y) = \beta_0 + \beta_1 x$

Interpretation: Once again, we take a derivative of our model with respect to x to find $\frac{d\log(y)}{dx} = \beta_1$, and we rewrite it in terms of small changes in $\log(y)$ and x:

$$\frac{\Delta \log(y)}{\Delta x} \approx \frac{d\log(y)}{dx} = \beta_1$$

Use the fact that $\Delta \log(y) = \frac{\Delta y}{y}$:

$$\frac{\left(\frac{\Delta y}{y}\right)}{\Delta x} \approx \frac{d\log(y)}{dx} = \beta_1$$

Then we can rearrange to see that

$$\frac{\Delta y}{y} = \beta_1 \Delta x$$

Suppose x changes by 5 **units** and plug this into the expression we just derived. We see that the proportional change in y is $5\beta_1$, so that y will change by $100 * 5\beta_1$ **percent**.

4. Log-Log functions and Percent-Percent changes

If we assume a log-log functional form, the model is: $\log(y) = \beta_0 + \beta_1 \log(x)$

Interpretation: As usual, start by taking a derivative of our model, $\frac{d\log(y)}{dx} = \beta_1 \frac{1}{x}$ and re-writing it in terms of small changes:

$$\begin{aligned}\frac{\Delta \log(y)}{\Delta x} \approx \frac{d\log(y)}{dx} = \beta_1\left(\frac{1}{x}\right) &\Rightarrow \beta_1\left(\frac{\Delta x}{x}\right) = \Delta \log(y) = \frac{\Delta y}{y} \\ &\Rightarrow \frac{\Delta y}{y} = \beta_1\left(\frac{\Delta x}{x}\right) \\ &\Rightarrow \frac{\Delta y}{y} \times 100 = \beta_1\left(\frac{\Delta x}{x}\right) \times 100\end{aligned}$$

Suppose we know that x changes by 10 **percent**. Plug this value into the expression we derived, and we see that y will change by $\beta_1 * 10$ **percent**.

Practice

This Table (Table 2.3 in Wooldridge) is meant to practice and continue familiarizing ourselves with these functional forms.

Model	DepVar	IndepVar	How does Δy relate to Δx?	Interpretation
Linear	y	x	$\Delta y = \beta_1 \Delta x$	$\Delta y = \beta_1 \Delta x$
Logarithmic	y	$\log(x)$	$\Delta y = \beta_1 \frac{\Delta x}{x}$	$\Delta y = (\beta_1/100)\%\Delta x$
Exponential	$\log(y)$	x	$\frac{\Delta y}{y} = \beta_1 \Delta x$	$\%\Delta y = (100\beta_1)\Delta x$
Log-Log	$\log(y)$	$\log(x)$	$\frac{\Delta y}{y} = \beta_1 \frac{\Delta x}{x}$	$\%\Delta y = \beta_1 \%\Delta x$

Examples

Example 1. Suppose you've collected data on household gasoline consumption (gallons) in the Bay Area and gas prices ($ per gallon), and you estimate the following model:

$$\log(gasoline) = 12 - 0.21 price$$

According to the model, how does gas consumption change when *price* increases by $1?

Example 2. Professor Villas-Boas in the ARE/EEP department used scanner data from a national grocery store to investigate how chicken consumption was affected by gas prices. Specifically, she looked at the share of chicken purchases that were made while the chicken was on sale. The following model was estimated:

$$\log(chickenshare) = 0.83 + 0.491 \log(gasprice)$$

How does *chickenshare* change if gas prices rise by 2%? Does this relationship make sense?

Example 3. Suppose you've collected data on CEO salaries (hundred thousand $) and annual firm sales (million $), and you estimate the following model:

$$salary = 2.23 + 1.1 \log(sales)$$

According to the model, how does *salary* change if annual firm sales increase by 10%?

Exercise Wooldridge exercise 3.4: Here is the result of a regression of median salary for new law school graduates on their LSAT score, median undergraduate GPA of the class, number of volumes in the law library, cost of attendance, and rank of the law school (1 being the best). The unit of observation is a law school:

$$\widehat{\log(salary)} = 8.34 + .0047 lsat + .248 gpa + .095 \log(libvol) + .038 \log(cost) - .0033 rank$$

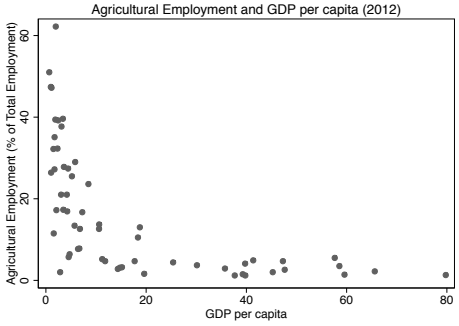
How does the median salary change when *libvol* (volumes in law library) change?

The coefficient on *rank* is pretty small. Does this mean the rank of a law school doesn't matter a lot for graduates' salaries?

4. Goodness of Fit, R^2

R^2 is a measure of the "goodness of fit," or how well our regression line fits the data, so that we are able to evaluate the quality of the model after estimating it. Specifically, R^2 is the proportion of variation in our dependent variable, y, that is explained by our model, $\beta_0 + \beta_1 x$.

Why is the R^2 important? Consider the following data from the World Bank that has this singular shape but we apply the linear model to it anyway. The line that minimizes the sum of squared errors is a flat line even though this clearly misses the underlying relationship between the variables. How can we tell that this is a poor model without looking at it? By looking at the R^2.



Before giving the explitic formula for the R^2, let's define a few additional terms. Consider these three variances, which we will call the Sum of Squares Total (SST), the Explained sum of squared (SSE), and the Sum of Squares Residual (SSR).

$$\begin{aligned} SST &= \sum_i^n (y_i - \bar{y})^2 \\ SSE &= \sum_i^n (\hat{y}_i - \bar{y})^2 \\ SSR &= \sum_i^n (y_i - \hat{y})^2 \end{aligned}$$

SST is a measure of the total sample variation in the y_i, that is it measures how spread out the y_i are in the sample. Similarly SSE measures the sample variation in the $\hat{y}_i$ (where we can use the fact $\bar{\hat{y}} = \bar{y}$). Finally the SSR measures the sample variation in the residuals $\hat{u}_i$. The total variation in y can always be expressed as the sum of the explained variation SSE, and the unexplained variation SSR. To see this, recall that $y_i = \hat{y}_i + \hat{u}_i$, i.e. the observed value of y_i is equal to the predicted value $\hat{y}_i$ and the difference between the two (the residual) $\hat{u}_i$. The formal proof is in Woolrdige p.39. Thus we can write,

$$SST = SSE + SSR$$

Next, let's define the R^2. We want the R^2 to express how well the regression line fits the data. One way to go about this is to express the fraction of the sample variation that is explained by x (i.e the proportion of variation that is explained by our model). In the sense that if we have a good model, then the sample variation in y should be mostly explained by x (and not by the residual that we don't explicity oberve)).

$$R^2 = \frac{SSE}{SST} = \frac{SSE}{SSE + SSR} = 1 - \frac{SSR}{SST}$$

The R^2 is always less than 1. Why? Well it's important to understand that if our variables x and y have *some* kind of relationship, knowing what x_i is should give us a little more information about what y is. Ex: if I know nothing about an individual but had to determine the probability that he/she had lung cancer, I would guess the average. However, if I find out that this particular individual is a smoker, I might think there's a slightly higher than average probability that he/she has cancer. We should think of $\hat{y}_i$ as a more knowledgeable guess for y_i than $\bar{y}$ as it uses x_i. Thus, the difference between what we observe and what we predict (SSR) should be smaller than the difference between what we observe and the average (SST). Thus $SSR/SST < 1$ and the $R^2 = 1 - SSR/SST$ will also be less than 1. If the model provides a perfect fit to the data, the $R^2 = 1$.