

**NOTES ON REGRESSION**  
**With Applications to Estimating**  
**the Earnings Equation**

**Economics 150**  
**Fall 2001**

**William A. Sundstrom**  
**Department of Economics**  
**Santa Clara University**

**Please do not cite or reproduce**  
**without author's permission.**

## 1. Introductio

In class we have studied the economic model of individual labor supply, which explains how people allocate their time between paid labor and leisure, depending on their preferences, their nonlabor income, and the wage rate they can earn for paid work. This model has the following predictions:

- (a) An increase in the wage rate (all else constant) may lead to an increase, a decrease, or no change in working hours, depending on the individual's tastes (balance of substitution and income effects).
- (b) An increase in non-labor income (all else equal) should reduce working hours, assuming that leisure time is a normal good.
- (c) Labor force participation should increase with increases in the wage rate and decrease with increases in nonlabor income.

The theory by itself is of limited use. First, its prediction of the effect of the wage on hours of labor supplied is ambiguous. Second, it makes no prediction about the magnitude of any of the effects: are they big, or small and insignificant? What are the elasticities? Answering these questions about the size of the effects is crucial to using the model for forecasting or policy analysis. For example, one justification of the minimum wage and the earned income tax credit as anti-poverty instruments is that they encourage more work hours among the poor. But we would need to know the direction and size of the wage effect to know if this is really true.

Regression analysis offers economists a way of estimating the mathematical relationship between variables such as hours of work, wages, and income, using individual-level data. In

other words, regression allows us to “fit” a supply curve to the data. With such an estimate of the supply-curve formula, we could hope to answer two important questions about our model:

- (1) Are the predictions of the model in fact supported by the data?
- (2) What are the magnitudes of the effects in the real world?

Regression can help us estimate many other theoretical relationships. One very important example in labor economics is the *earnings equation*. An earnings equation is simply the relationship between a worker’s earnings (Y) and her or his individual characteristics (X). For example, several different economic theories predict that an individual’s earnings will depend positively on how much education she has had. But again, we would like to know if the theory holds true in the real world. Furthermore, we would like to know how big the effect of education is. For instance, does an additional year of schooling increase annual earnings by 1%, 5%, or 10%? Answering this question is clearly critical to evaluating the economic value of education to both individuals and society.

These notes are to help you refresh your memory regarding linear regression, and to show how regression can be used to estimate a mathematical relationship predicted by economic theory. For various reasons, it will be much easier to apply regression analysis to estimating an earnings equation than it will to estimating labor supply. Therefore, these notes start with a brief overview of the idea of an earnings equation, which will be developed in greater detail later in the course. We then proceed to a review of regression, with illustrations drawn from estimating the earnings equation.

## 2. The earnings equation

One of the most important issues in labor economics is the determination of people's wages or earnings. Why do some people earn more than others? There are obviously many factors involved, but economic theory predicts that certain factors should be very important.

In particular, economic theory suggests that both education and work experience should tend to have a positive impact on earnings. There are several reasons this might be true. Probably the most obvious reason is that both education and experience tend to increase a worker's skill, or the value of her marginal product. In the labor market they will then earn more. This is known in economics as the theory of *human capital*: education and on-the-job learning are "investments" individuals make in their own skills and earning power.

Other theories also predict a positive correlation between schooling and earnings, as we shall see in class. For instance, the theory of labor-market *signaling* suggests that some individuals will obtain more education as a "signal" of their basic abilities (intelligence, drive, reliability) to employers. In this view, the value of schooling is not what you learn but what you signal to employers when you take the time and trouble of receiving more education.

Many other factors may affect earnings as well. Some of these, such as "ambition" or "talents" cannot be directly observed in our data. Others may be observable: for example, race or gender. There are various reasons that pay might depend on one's race or gender: discrimination in the market is one of those.

To estimate the impact of measurable variables on earnings in the real world, we can use regression analysis to estimate an earnings equation from cross-section data. For now, let's just

think about a simple mathematical formula that could describe the relationship between education, work experience, and earnings. For example, let  $Y$  = annual earnings in dollars,  $S$  = years of school (e.g.,  $S = 12$  years is a high-school education), and  $E$  = years of work experience. Then one way  $S$  and  $E$  might affect earnings is the following *earnings equation*:

$$Y = b_0 + b_1S + b_2E$$

Our goal using regression analysis of data will be to obtain estimates for the coefficients of this equation:  $b_0$ ,  $b_1$ , and  $b_2$ . In interpreting the equation, note that the coefficient  $b_1$  is the change in earnings (in \$) for one more year of schooling, given whatever value  $E$  takes. So if  $b_1$  were 2000, it would imply that one more year of schooling would increase earnings by \$2,000. This could be thought of as the average market value of a year of schooling, holding work experience constant.

It turns out that in practice economists usually assume that the earnings equation has a slightly different formula, using the *natural log* ( $\ln$ ) of earnings as the dependent variable. In this case the estimated equation becomes:

$$\ln(Y) = b_0 + b_1S + b_2E$$

This functional form is sometimes referred to as the *semilog* form. In this case the coefficient  $b_1$  is the change in the log of earnings for one more year of schooling. The change in the log of a variable is the proportional or percentage change in that variable. This follows from calculus:

$$d\ln(Y) = \frac{dY}{Y} = \text{proportional change in } Y$$

Therefore, we can interpret  $b_1$  as the proportional change in earnings for one more year of schooling. For instance, suppose in the above equation the coefficients were as follows:

$$\ln(Y) = 5 + 0.05S + 0.01E$$

This means that one more year of schooling increases earnings by the proportion 0.05 or 5%.

One more year of work experience adds 1% to earnings.

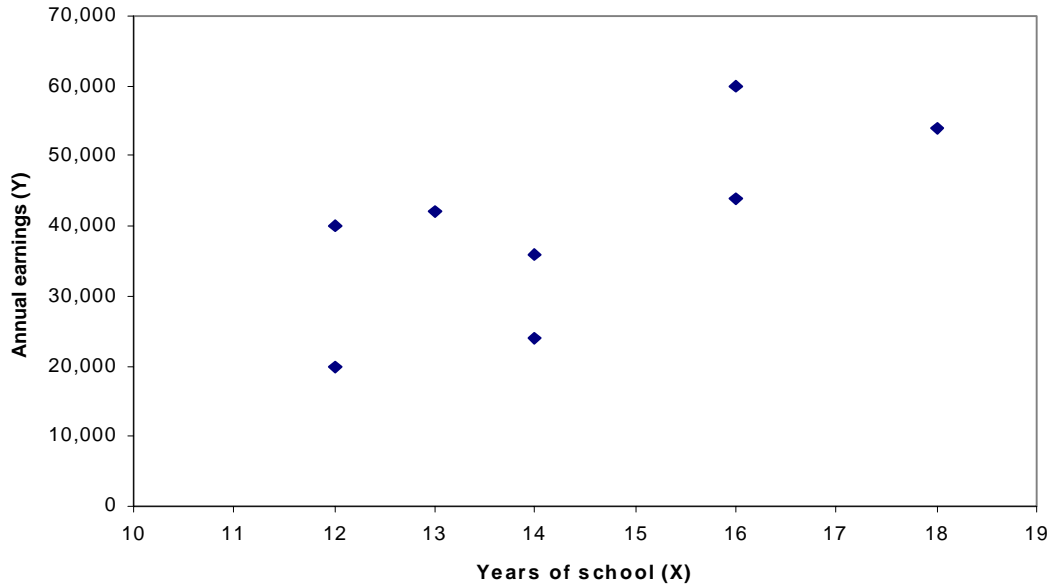
Aside from this nifty interpretation of the results, there are two major reasons for using the log of earnings. One is theoretical. According to the theory of human capital, individuals invest in education (which imposes direct and opportunity costs on them) in order to reap a return on their investment. The coefficient  $b_1$  turns out to be closely related to the rate of return on that investment. The second reason is statistical, and has to do with the statistical distribution of earnings, an issue we will return to. We will compare the linear and semilog earnings equations later.

### 3. The idea of simple regression: fitting a line

We begin with the data. If we are estimating the earnings equation, we are interested in the relationship between earnings, on the one hand, and schooling, experience, and other variables on the other. Our data set will be a cross section of individual workers. For each worker, we will have information on their earnings and other characteristics. For now, let me focus on the relationship between earnings and years of education. Here's a simple example of a made-up data set with 8 observations (people):

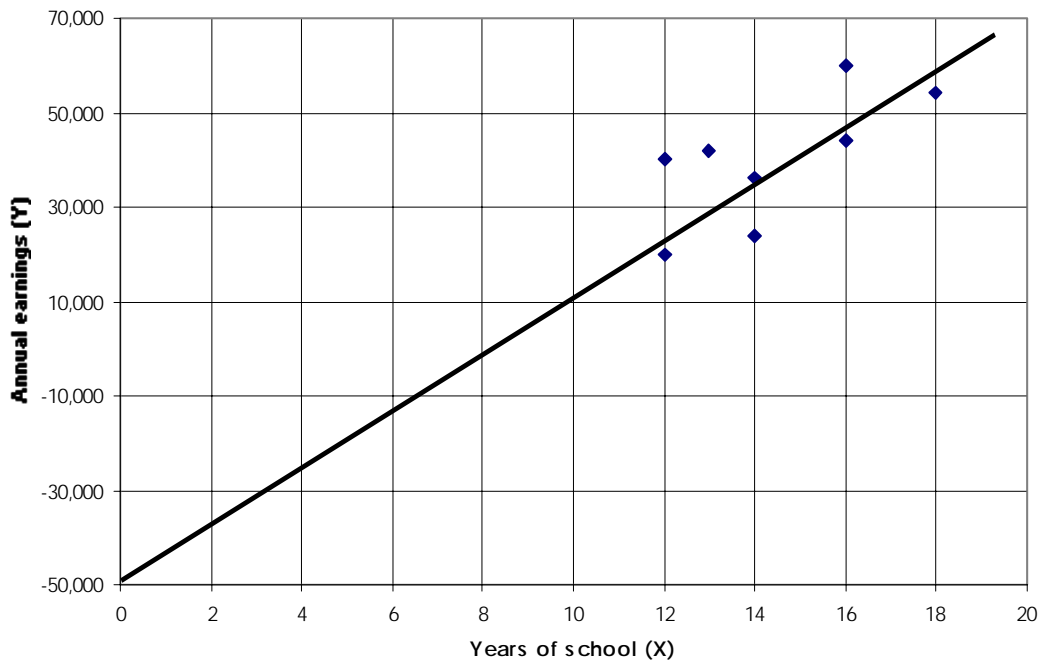
Person (i)	Years of school (X)	Annual earnings (Y)
1	12	20,000
2	14	24,000
3	14	36,000
4	18	54,000
5	16	44,000
6	16	60,000
7	12	40,000
8	13	42,000

Here is a scatter plot of these 8 points:



You can see from this scatter that the relationship between  $X$  and  $Y$  is generally a positive one, but not perfectly so. We expect schooling to be positively associated with earnings, but because other factors affect earnings as well, there's no reason to expect that each individual always makes more money than every other person with less education.

The objective of simple regression is to find the straight line that best fits through this scatter of points. To see how the regression picks this line, let's try "eyeballing" it and drawing a straight line that appears to fit this scatter of points. Here's my best guess (I have changed the scales of the axes and added gridlines to make it easier to figure out the formula for my line).



**Sundstrom's best guess of the earnings equation for these data**

What is the formula for my line? Well, in general, a line has the formula  $Y = a + bX$ . So what are  $a$  (the  $Y$  intercept) and  $b$  (the slope)? Note that my line crosses the  $Y$  axis at about  $-50,000$  (that's  $a$ ) and then slopes up to around  $(X=18, Y=60,000)$ . This implies that the slope (rise over run) is  $b = \Delta Y / \Delta X = (60,000 - -50,000) / (18 - 0) = 110,000 / 18 = 6111$ . Thus the approximate formula for my line is  $Y = a + bX = -50,000 + 6,111X$ .

Now I want to distinguish between the actual values of  $Y$  (represented by the data points) and the values of  $Y$  predicted by my line. So instead of using  $Y$  in my formula, I am going to use the following notation to indicate the "fitted" or predicted value of  $Y$ , which I'll call "Y-hat":

$$\hat{Y} = a + bX$$

Does my line fit the data well? One way to judge is to look at the *deviation* (also known as the *residual*) between each actual value of  $Y$  in the data and the value predicted by my regression line ( $\hat{Y}$ ). For an observation  $i$ , this deviation is

$$d_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i)$$

For example, consider observation (person)  $i = 4$ . For that person,  $X = 18$  and  $Y = 54,000$ . My line predicts that their  $Y$  will be  $\hat{Y} = -50,000 + 6,111(18) = \$60,000$ . So my line is “off” by the deviation  $d = Y - \hat{Y} = 54,000 - 60,000 = -6,000$  for this person. The deviation is negative for a person who earns less than what the line predicts, and positive for a person who earns more.

Now imagine taking each individual in the sample, predicting their  $\hat{Y}$  from the line, and then calculating the deviation between their actual earnings ( $Y$ ) and the earnings predicted by the line ( $\hat{Y}$ ). Take each deviation, square it, and add them all up. What you have is the sum of squared deviations (SSD) between the actual values and the predicted values on the line:

$$SSD = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - a - bX_i)^2$$

where the  $\sum$  means the sum over all the observations (8 in our example). This is an overall measure of how far off the predicted values of  $Y$  are from the actual values.

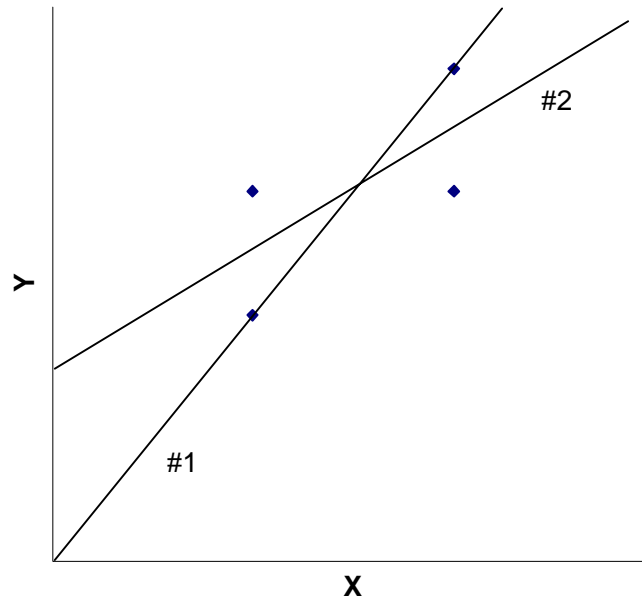
The technique of ordinary least-squares (OLS) regression chooses the best regression line to fit the data by choosing  $a$  and  $b$  to minimize the sum of squared deviations (SSD) between the actual values and the line.

There are formulas for the values of  $a$  and  $b$  that minimize SSD. But we will just let the computer find them for us. I’ll show you how using Excel in class. For now, let me just note that the OLS regression coefficients for my made-up data are  $a = -24,039$  and  $b = 4455$ . So my guess was not too close!

The OLS regression should be thought of as choosing the line that minimizes the average vertical distance between the data points and the line. This means that *the regression line offers the best prediction of  $Y$  given (conditional on) a particular value of  $X$* . The regression uses the information on  $X$  to help predict  $Y$ .

The following diagram helps illustrate this. For the data set shown, you might be inclined to guess that line #1 fits best. But the OLS regression would choose line #2. The reason is that

the regression line wants to slice through the middle of the data *given each value of X*. If you know X, then line #2 gives a more reliable prediction of Y than does line #1.



**Line #2 is the OLS regression line**

### ***The regression as a model***

So far we have just considered regression as a matter of line fitting. But we'd really like it to be a way to do "model fitting." That is, we start with our economic theory, and come up with a model that predicts some relationship between observable variables. Then we use the regression to fit the actual parameters of the model.

Thinking about our regressions this way is very important, because it clues us into potential problems in interpreting the regression results. For instance, a regression line by itself is really just a fancy way to summarize the scatter diagram between two variables, such as education and earnings. But in our economic model of earnings, an individual's education is usually completed early in life, and then fixed thereafter. After that time, education becomes part of the individual's characteristics that help determine her or his earnings on the market at any point in time. Our model directs us toward a certain way of analyzing the data:

- (1) It directs us to focus in on education as one important variable determining earnings.

- (2) It directs us to think of education as the *exogenous* or *independent* (X) variable and the earnings as the *endogenous* or *dependent* (Y) variable. In other words, our model suggests that education causes earnings, rather than vice versa.
- (3) Therefore, the model directs us to estimate the relationship with education as the X variable and the earnings as the Y variable.

One problem most people will raise with economic models of individual behavior is that they are so mechanical. When an economist draws an earnings equation, suggesting a certain relationship between education and earnings, a non-economist might say, “Sure, but what about a poorly educated person who is a great athlete, or who just gets lucky and lands a great job? Your earnings equation can’t predict their pay. Or what about a well-educated but lazy person who earns very little? Your model makes it sound as though only schooling determines earnings, and that everyone with the same education will earn the same amount.”

In fact, economists are well aware that other variables determine earnings, and that individual “luck” or tastes for effort could play a role. But our non-economist raises an important challenge: How can our *statistical analysis* of the data take account of these additional factors? After all, the regression so far has just given us one line,  $\hat{Y} = a + bX$ , to describe everybody.

There are two ways we will address the problems raised here. First, note that we will want to see how other variables affect earnings, in addition to schooling, and we can do so statistically, using the technique of *multiple* regression. So one thing we can do is to add additional variables, such as age or years of work experience, and see how they affect earnings along with education.

Second, there is the issue of individual variability in terms of tastes, talents, luck, and other variables that we *cannot observe or measure*. To deal with these we must add to our model a random component. This is known as the “error term,” which will be represented by “u” in our formula. The idea of the error term is to include in our model a “random variable”: that stands for all the factors affecting Y that we have not controlled for. So now our simple regression model of earnings (Y) can be represented as follows:

$$Y_i = \alpha + \beta X_i + u_i$$

This model assumes that for each person  $i$ , their earnings has a component that depends systematically on their years of education (that's the  $\alpha + \beta X$  part), and a component that is particular to that person but cannot be observed (that's the  $u$  part). So you could think of the "u" as standing for "unobserved" or "unknown". In fact, even  $\alpha$  and  $\beta$  are unknown: they are hypothetical parameters of the earnings equation. But we can estimate these parameters using the regression coefficients  $a$  and  $b$ .

Actual parameters versus estimated parameters:

Our model of  $Y$  will suggest that  $Y$  depends on  $X$  and an unobserved "error",  $u$ :

$$Y = \alpha + \beta X + u$$

Our OLS regression will give us estimates of  $\alpha$  and  $\beta$  :

$$a = \text{OLS estimate of } \alpha \quad b = \text{OLS estimate of } \beta$$

It is very important to keep in mind that the coefficient estimates  $a$  and  $b$  from our regression are just that: estimates. As such, they are likely to be off one way or the other, just as a sample mean is likely to be different from the population mean.

The error term "u" is related to the idea of the deviation or residual in our earlier discussion. If our estimates  $a$  and  $b$  were exactly equal to the "true" values  $\alpha$  and  $\beta$ , then each residual would be exactly equal to that individual's "u". But because our estimates will generally not be exactly correct, we cannot consider the deviations to be exactly the same as the error term.

The reliability of the regression estimates as estimates of the model rests critically on the reliability of the model itself, as well as the nature of the "unknown" part, the error term. I now turn to take a look at some important regression statistics and diagnostics that can tell us something about the reliability of our results. I then turn to a list of some common things that may go wrong with a regression.

#### 4. Some useful regression statistics

In this section I review some of the most important regression statistics. The first set of statistics all have to do with how well the regression fits the data overall. The second set have to do with how well the regression has estimated particular coefficients— especially the slope,  $\beta$ . To make the discussion more concrete, I provide the following printout of the Excel regression results for a regression of annual earnings on years of schooling (SCHOOL), using a sample of 150 full-time, full-year workers from March, 1998:

##### SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.195573835
R Square	0.038249125
Adjusted R Square	0.031750808
Standard Error	30524.51552
Observations	150

##### ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	5484261932	5.48E+09	5.886005	0.01646454
Residual	148	1.37898E+11	9.32E+08		
Total	149	1.43383E+11			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	6640.986557	12069.84036	0.550213	0.583003	-17210.5	30492.473
SCHOOL	2171.128576	894.901485	2.426109	0.016465	402.693345	3939.5638

Note here that the estimated equation is  $\hat{Y} = 6641 + 2171 \cdot \text{SCHOOL}$ . The coefficient on SCHOOL of 2171 implies that a worker with an additional year of school earns \$2171 more per year.

**Measures of fit and residual variation:**

- *Sum of squared residuals (SSR)*
- *R-squared ( $R^2$ )*
- *Standard error (S.E.) of the regression*
- *F-statistic*

The goal of the regression is to estimate how Y depends on X. One way to get at how well the regression is doing that is to measure the total variation in Y in our data, and then see how much of that variation is accounted for by the regression.

The total variation in Y around its mean can be measured as the total sum of squares (TSS). This can be broken down into the sum of two parts: the part explained by the regression, and the residual or unexplained part. These components are given in the ANOVA table under the heading SS. The residual part is the sum of squared residuals, or SSR, which again is the same as the SSD (what the regression minimizes). In the above table, the SSR =  $1.37898 \times 10^{11} = 1.37898 \cdot 10^{11}$ , a very large number. The total sum of squares (TSS) =  $1.43383 \times 10^{11} = 1.43383 \cdot 10^{11}$ .

Using the SSR, Excel also calculates a very widely used statistic, called R-squared ( $R^2$ ). R-squared is often used as a measure of "goodness of fit." In words, R-squared tells us the proportion of the total variation in Y that is "explained" by variation in X along the regression line. Mathematically,

$$R^2 = \frac{\text{variation in } Y \text{ explained by regression}}{\text{total variation in } Y} = 1 - \frac{\text{SSR}}{\text{TSS}}$$

In our example, the R-squared of 0.038 means that the regression accounts for about 3.8% of the total variation in earnings in the sample. In this sample, then, variation in years of schooling accounts for rather little of the variation in earnings across individuals. Other, unaccounted for factors must explain the remaining 96.2%.

$R^2$  will always be between 0 and 1, with a higher number suggesting a better "fit."  $R^2$  can be a useful diagnostic, and other things equal we prefer a higher value to a lower value. But

there is no magic level of  $R^2$  that one should look for. In part the importance of  $R^2$  depends on why you are running a regression.

For instance, a regression with a high  $R^2$  may not necessarily be very informative. For example, if you regress nominal consumption on nominal GNP, you may get a high  $R^2$ . But this is partly due to inflation affecting both variables, and partly due to both happening to be on an upward trend.

It is equally mistaken to assume that a low  $R^2$  means your results are of no interest. Especially in cross-section studies, low  $R^2$  values are very common: even well below 10 percent (the example above shows this). Is that bad? Not necessarily. Suppose you are trying to get a good measure of the effect of education on earnings, for instance. This does not necessarily require that you do a good job explaining all the variation in earnings. It only requires that you accurately capture the relationship between X and Y. This does not necessarily require getting a high  $R^2$ .

Another statistic related to R-squared is the *standard error of the regression*. This is a measure of the variability of the data points above and below the regression line, and is in the same units as the dependent (Y) variable. In Excel, this number is found in the top part of the table, and is equal to 30524.51552 in the above example. The standard error of the regression can be used as an approximate prediction or forecast error. I will discuss this application in class.

The *F-statistic* in the ANOVA table provides a formal statistical test of the following hypothesis: *All the coefficients on the X variables are equal to zero*. For example, in our case above we have only one X variable, and the F-stat is 5.886. Directly to the right is the p-value for the F test, 0.0165. The conventional rule is that we can reject the null hypothesis if the p-value is less than 0.05. Therefore, we would reject the hypothesis that the coefficient is zero. In other words, we have confidence that there is an effect of education on earnings in this sample.

### ***The standard error of the slope coefficient and confidence intervals***

The other part of the regression output that I want to explore is the standard error of the coefficient. Our estimate  $b$  of the slope parameter is just that: an estimate. How good an estimate is it? In other words, how close is it likely to be to the "true value,"  $\beta$ ?

The standard error of  $b$  ( or  $s_b$  ) is an estimate of the accuracy ( or lack thereof ) of ou

parameter estimate. The larger the standard error of the coefficient, the less precise is our estimate. The SE of  $b$  appears in the regression output table next to the relevant slope coefficient. In our example, the standard error of  $b$  is thus approximately 894.9.

The standard error of  $b$  typically depends on the following factors:

- (1) It tends to be larger the smaller  $R$ -squared is.
- (2) It tends to be smaller the more observations ( $N$ ) we have in the sample.
- (3) It tends to be smaller the more variation there is in  $X$ .

The standard error of  $b$  is used to test hypotheses about  $\beta$ . One of the easiest ways to do this is to use the standard error to calculate a *confidence interval* around the estimate  $b$ . This will give us the range of values within which we are confident that the true slope lies. The size of that interval depends on two things: how much confidence we want to have (the more confidence, the wider the interval), and the SE of  $b$ .

It is conventional among economists and many other social scientists to use a 95% confidence interval. What this means in words is that we have “95% confidence that the true  $\beta$  lies within the interval.” This is conservative in the sense that 95% is a pretty high level of confidence.

For samples of a reasonable size (say,  $\geq 20$  or so), the 95% confidence interval for  $b$  is *approximately*  $b \pm 2s_b$ . Using Excel it is unnecessary to make such a calculation, because Excel automatically gives you the 95% confidence interval. For the slope coefficient, the 95% interval runs from 402.7 (lower 95%) to 3939.6 (upper 95%), and is centered on the point estimate of 2171.1.

The confidence interval is a convenient way to test hypotheses about the coefficient. We can reject a hypothesized value if it lies outside the confidence interval. Here are a couple of examples for the above estimate:

*Hypothesis 1: There is no effect of education on earnings.* This is equivalent to saying that  $\beta = 0$ . We can see that 0 does not lie within the 95% CI of [402.7, 3939.6]. Therefore, we can *reject* the hypothesis: in other words, we can conclude with 95% confidence that the effect is *not zero*. Another way we often put this is that *the slope is significantly different from zero*. Incidentally, the  $t$ -statistic in the table also tests this hypothesis. If the  $t$ -statistic has magnitude (positive or negative) bigger than about 2, we can reject the hypothesis that the slope is zero, with 95% confidence.

*Hypothesis II: An additional year of school will increase annual earnings by \$1000.* This would require that  $\beta = 1000$ . We can see that 1000 is inside the 95% CI, and therefore we *cannot reject* this hypothesis. Thus we could say that the hypothesis is plausible and cannot be ruled out with 95% confidence.

### ***Statistical and economic significance***

It is very important to check the statistical significance of your regression results. That is, check to see if the coefficients you are interested in are different from zero in the statistical sense. If not, you can have little confidence that there is something actually going on, rather than just a fluke of your particular data set. But it is also important to consider the *economic significance* of the results. For example, we have found that the slope  $b$  is significantly different from zero. But is the effect economically important? Does it suggest that education is a good investment? To answer this question, we might want to compare our estimate  $b$  with some estimate of the costs of education. The slope could be statistically significant without being economically significant.

In other words, always ask two questions about your coefficient estimate:

- (1) Is it statistically significant?
- (2) How big an effect is it?

### ***Looking at the residuals***

The regression results imply a residual or deviation for each observation. In Excel we can generate and plot these residuals in various ways. The residual series is a very useful diagnostic to check whether your regression results are valid. I will explore this in greater detail below. But for now, let me simply note that the regression results are most reliable if the residuals are normally distributed and uncorrelated both with each other and with the  $X$  variables. In class I will show you some ways to examine the residual series.

### ***Estimating the earnings equation using the natural log***

In some cases it may be desirable to estimate a relationship between two variables that is not a straight line. In the case of the earnings equation, I have noted that a common formula to use is the semilog specification. To implement this, we simply generate a new column of

numbers that is the log of earnings, and use this as the dependent variable in the regression. I will show how in class. Here are the results, using the same data as the preceding example.

## SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.286223
R Square	0.081924
Adjusted R Square	0.075721
Standard Error	0.664108
Observations	150

## ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	5.824667	5.824667	13.20668	0.000384
Residual	148	65.27385	0.44104		
Total	149	71.09851			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	9.307859	0.262598	35.44527	3.36E-74	8.788933	9.826785
SCHOOL	0.070756	0.01947	3.634099	0.000384	0.032281	0.109231

Note here that the estimated earnings equation is

$$\ln(EARNINGS) = 9.308 + 0.0708 \text{ SCHOOL}$$

The coefficient on SCHOOL is 0.071. Recall that when we use the semilog form, the slope coefficient is the proportional effect of the X variable. In other words, this estimate tells us that one more year of education is associated with 0.071 or 7.1% higher annual earnings.

## 5. Multiple regressio

Regression is at its most powerful when we use it to sort out the partial effects of a number of different variables on some variable. For instance, annual earnings might depend not just on education but also on the worker's age or experience, or race or gender.

In fact, if we do not control for various influences simultaneously, our results can be severely biased: the coefficient estimate may be systematically off. The bias associated with leaving out one or more important regressors is known as left-out variable bias or missing variable bias. I discuss this in the section on regression problems and pitfalls below.

Multiple regression assumes that the model now takes the following form for each individual  $i$ :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + u$$

For example, we could see how earnings depend on `SCHOOL` and potential years of work experience, which in these data is measured as `EXPER = worker's age - SCHOOL - 6`. (The idea is to measure how many years the person could possibly have been working after deducting years in school.) Here are the results of such a regression, using the linear form.

## SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.22288
R Square	0.049675
Adjusted R Square	0.036746
Standard Error	30445.68
Observations	150

## ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	7.12E+09	3.56E+09	3.841995	0.023637
Residual	147	1.36E+11	9.27E+08		
Total	149	1.43E+11			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-3146.87	14111.44	-0.223	0.823844	-31034.3	24740.61
SCHOOL	2432.603	914.0018	2.661486	0.008645	626.3236	4238.883
EXPER	300.8669	226.3077	1.329459	0.185756	-146.37	748.1035

The results now imply the following earnings equation:

$$EARNINGS = -3147 + 2433 \text{ SCHOOL} + 301 \text{ EXPER.}$$

With multiple regression, our coefficients are *partial effects*. For example, the coefficient on SCHOOL is now estimated to be 2433, which we can interpret as the change in earnings for an additional year of schooling, *holding the worker's years of work experience constant*. This is what makes multiple regression so powerful. For while we lack the ability to run controlled experiments much of the time, multiple regression is a way to control for other factors using available data. Note that our results also suggest that holding education constant, changes in years of work experience have the expected positive effect on earnings (each year of experience increases earnings by about \$300). As before, the standard errors of the coefficients can be used here to form confidence intervals and test hypotheses.

$R^2$  has the same interpretation in multiple regression as in simple regression. We can no use  $R^2$  to see if our fit has improved by adding or changing some of the X variables. However, because  $R^2$  always goes up with the number of regressors, it's better to compare the *adjusted*  $R^2$ , which compensates for the effect of adding more regressors.

We can also put confidence intervals around either coefficient estimate, and Excel provides these in the table.

Multiple regression allows us to estimate very sophisticated models. First of all, we can control for many variables at once. Second, we can fit nonlinear relationships using a polynomial or other transformations of the data. For example, I could create a new variable that is the square of SCHOOL, and add it to my original regression. Then I would be estimating a quadratic function, which can have a curvature:

$$EARNINGS = b_0 + b_1 SCHOOL + b_2 SCHOOL^2.$$

### ***Dummy variables and interactions***

Multiple regression can also be used to capture the effects of qualitative characteristics b using “dummy variables” as regressors. For instance, suppose I wanted to see if married men earned more per year than unmarried men, controlling for their education. Marital status is not a continuous condition: it is either-or. But I can still generate a variable that captures the effect of marital status. I simply create a variable MARR, which is =1 for married men and =0 for unmarried, and then add it as a regressor in the following multiple regression:

$$EARNINGS = b_0 + b_1 SCHOOL + b_2 MARR$$

According to this formulation, when a man is unmarried, MARR = 0, and his earnings are described by

$$EARNINGS = b_0 + b_1 SCHOOL + b_2 * 0 = b_0 + b_1 SCHOOL$$

Whereas if a man is married, MARR = 1, and his earnings are described by

$$EARNINGS = b_0 + b_1 SCHOOL + b_2 * 1 = b_0 + b_2 + b_1 SCHOOL$$

So in this case the coefficient on the dummy variable ( $b_2$ ) has the effect of shifting the whole earnings equation up or down. The slope ( $b_1$ ) is not affected.

Of course, you might also wonder if the effect of education was similar for married versus unmarried men. This involves seeing whether the slope changes with marital status. To capture this possibility in the regression, we can add an *interaction term*, INTER = MARR\*SCHOOL. Now we estimate the following regression:

$$EARNINGS = b_0 + b_1 SCHOOL + b_2 MARR + b_3 INTER$$

This equation implies that when a man is unmarried, MARR = 0, and so INTER = MARR\*SCHOOL = 0, and his earnings are described b

$$EARNINGS = b_0 + b_1 SCHOOL + b_2 0 + b_3 0 = b_0 + b_1 SCHOOL$$

Whereas if a man is married,  $MARR = 1$  and  $INTER = MARR * SCHOOL = SCHOOL$ , so his earnings are described by

$$EARNINGS = b_0 + b_1 SCHOOL + b_2 1 + b_3 SCHOOL = b_0 + b_2 + (b_1 + b_3) SCHOOL$$

In this case the coefficient on the dummy variable ( $b_2$ ) has the effect of shifting the intercept for married men, while the coefficient on the interaction ( $b_3$ ) represents a difference in the slope between married and unmarried men:

	Intercept	Slope on wage
Unmarried	$b_0$	$b_1$
Married	$b_0 + b_2$	$b_1 + b_3$

## 6. Regression problems and pitfalls

Regression analysis is a very powerful tool. But the validity of regression results depends critically on the validity of certain assumptions. When these assumptions fail, the regression results can be invalid in various ways. Although some of these problems can be diagnosed and corrected, it must be stressed that some problems are hard to fix, and you cannot always tell when you have a problem. Being aware of the limitations of regression analysis and thinking about potential sources of error are crucial to good research.

The problems I want to consider come under two general headings: those that *bias* the coefficients, and those that invalidate our confidence intervals and hypothesis tests.

### *Problems that lead to bias in the coefficients*

A coefficient estimate is biased if it tends to be systematically off in one direction or the other. In other words, suppose the true effect of one more year of schooling on earnings is \$1000, but there is some problem that makes it most likely that we obtain an estimate around \$2000. Then our estimate is biased.

Bias occurs when one or more of the X variables (regressors) are correlated with the error term,  $u$ . There are several reasons this can occur.

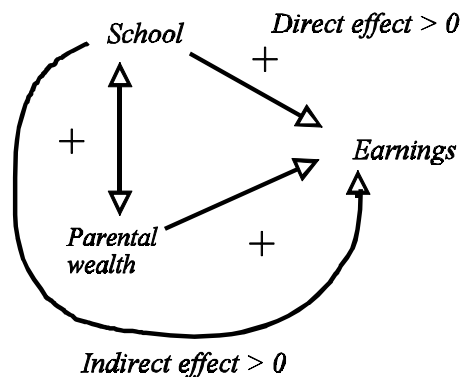
(1) *Misspecified functional form.* If the true relationship is a curve, but we fit a straight line, then our slope coefficient is misleading at best. *Diagnosis:* Plot the residuals against the X variables (I'll show this in class). The residuals should look like "white noise." *Treatment:* Try an alternative functional form, using a quadratic, or logs of the variables.

(2) *Left-out variable bias*. This occurs if you have failed to include adequate controls in your regression. Here's an example that illustrates the problem.

Suppose we have estimated the relationship between EARNINGS and SCHOOL, and obtain a positive coefficient, but we have left out any measure of family background, such as parents' income or wealth. Furthermore, suppose that individuals from wealthy families tend to get more education and also tend to be better connected in the job market. If so, then our coefficient on SCHOOL is likely to be biased.

Why? Because the coefficient is picking up two effects. One is the *direct effect* of education on earnings, which is what we're trying to estimate: let's suppose it is positive. But the second *indirect effect* results in spurious correlation. When an individual has more education, she also is likely to have come from a wealthier family. Her family background is associated with higher earnings, due not to the schooling but to her social contacts. This is picked up by our SCHOOL variable. The upshot is that the coefficient on SCHOOL is positively biased: it shows too large an effect because it is also picking up the effect of parental wealth.

This can be illustrated using a "path diagram" of these two effects:



Our slope is a biased estimate of the effect of education (direct effect).

*Diagnosis:* There is no simple way to tell from your results if left-out variable bias is a problem. You should always think carefully about whether you have included adequate controls.

*Treatment:* Solving the left-out variable bias problem is easy if you have data on the left-out variable. For example, if our data included information on each person's social background or

parental wealth, we could add such a variable to SCHOOL in the regression. But you do not always have the necessary data, even under the best of circumstances. Parental variables are rare in most data sets, and other variables that might be correlated with both education and earnings, such as ambition, attitude toward risk, or willingness to wait, are almost never available. So one must take care in interpreting the coefficient on SCHOOL.

(3) *Errors in measuring the regressors (X)*. Very often the our data are measured with error. For example, our measure of education does not control for the quality of the school attended or the areas of study. If a regressor is measured with error, it biases the slope coefficient toward zero. That's because the measured relationship is noisy and weaker than the actual relationship. *Diagnosis*: This is another one that is difficult to diagnose from your results. It is again something one should be aware of. *Treatment*: There are fancy techniques for dealing with measurement error, but they are beyond the scope of this course. The best remedy is to get the best data you can and just be aware of the problem.

(4) *Endogeneity of the regressors*. Our basic assumption is that the X variables are exogenous, which for our purposes means they are not a function of Y. But very often the causality between X and Y runs in both directions. For example, we have assumed that education causes earnings, but the reverse causation is also possible. A person who earns a lot might be able to afford to go back to school, for example.

In such cases, the estimated slope  $b$  is a biased estimate of the impact of education on earnings. We are really picking up the mixed effects of two different causal relationships.

This problem is especially severe when we use market-level data on prices and quantities to estimate supply and demand curves. Because P and Q are always determined by the interaction of two curves (S and D), it is not usually possible to isolate or identify the separate curves using data on P and Q alone.

*Diagnosis*: Once again, there is not a simple diagnosis. Always think carefully about potential sources of endogeneity in your data. *Treatment*: There are good techniques for dealing with endogeneity, but again they are beyond the scope of this course.

### ***Problems that lead to invalid or very large standard errors***

When the error term ( $u$ ) is not normally distributed with the same mean and variance for each individual, the estimated standard errors will be incorrect. In that case, your confidence

intervals will be wrong, and any hypothesis tests may be invalid. I will discuss some sources of this kind of problem in class. Generally, it is a good idea to check the residuals for normality (a bell-shaped distribution). Excel has a nice graphical test for normality that we will look at in class. If the residuals do not look normal, there may be some simple ways to change the regression that help improve its validity.

In multiple regression, standard errors can be very large if there is *multicollinearity* in the data. Multicollinearity essentially occurs when two or more of the regressors are highly correlated with each other. When this happens, it is very hard for the regression to sort out the separate contributions of the two variables— it's almost as if they are really just one variable, because they tend to move together. An extreme example would be if you ran the earnings equation using both years of school and months of school as regressors. These variables are obviously highly correlated, and the regression could not estimate their separate effects accurately. The result is huge standard errors.

One “solution” to multicollinearity is to drop one of the correlated X variables out of the regression, but in doing so, you risk a serious left-out variable bias. The results must be interpreted carefully in such situations.

### ***Choosing a specification***

One of the most important decisions in regression analysis is what *specification* to use. The specification is simply the details of your regression: In particular, what X-variables (regressors) should you include? Which should you leave out? Should your regression equation be linear, or some other functional form?

There are no straightforward answers to these questions. I offer the following general guidelines:

(1) Use economic theory to decide which variables should be in the model before you run any regressions. For example, our model of earnings will suggest that both education and work experience could be expected to affect earnings. So both variables belong in the equation.

(2) Use variables that will help you test a hypothesis. For instance, suppose you are interested in whether men earn more than women, given the same schooling and experience. Then you could answer this by including a dummy variable for FEMALE, and testing whether it was significantly less than zero.

(3) Examine the regression residuals. If you are trying to decide between two alternative specifications, one factor to consider is which one results in more random, “normal”-looking residuals. Because many of our statistical tests rest on the assumption of random, normal distributed errors, there is some reason to favor a regression with residuals that look that way.

(4) Often when you run a multiple regression there will be some coefficients that are not significantly different from zero (large standard errors, small t-statistics). Should these regressors be left in the regression, or dropped out? The best general rule is to leave in an regressors that you really think ought to matter. We will consider this issue in more detail in class.