Andrew Noymer, UC Irvine
Handout for 221B: Logged dependent variables (illustration by example)

Start with the **hire771** data set.

Create a logged salary variable:

```
gen logsal = ln(salary)
```

Note that **log** now works in Stata just like **ln**... the natural logarithm.

(1) Imagine a model, for instance:

```
regress logsal age age2 sex educ
```

(2) Interpretation of estimates.

This is a little bit different than what we have been covering up till now.

Whether or not a variable is significant has not changed — this is still determined by whether or not the value in the **P>|t|** column is greater than or less than our critical value (which is typically 0.05, but does not have to be; this is sometimes called α [just to confuse you] but it is not the same as the intercept which is also α).

The meaning of the β coefficients has changed, however.

Take a regression equation like we have been using all quarter:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Now, suppose we want to know the effect on Y of a change in X_2, holding X_1 constant. By now you should know that this value is represented by β_2.

We can show this more formally using calculus. *I know this is foreign to many of you, but please bear with me, because:* (1) *relax — you will not be required to derive this yourself, and* (2) *even if you don't really understand it, you can still learn something by taking the derivations for granted and looking at the results.*

As we have seen before, in calculus, the symbolic way to represent a "change in something given a unit change in X_2" is $\frac{\partial}{\partial X_2}$.

So if we want to know the change in Y for a unit change in X_2 we do this:

$$\frac{\partial}{\partial X_2} Y = \frac{\partial}{\partial X_2} (\alpha + \beta_1 X_1 + \beta_2 X_2)$$

And, take my word for it, this works out to:

$$\frac{\partial}{\partial X_2} Y = \beta_2$$

So, β_2 is the change in Y for a unit change in X_2, holding all else constant. You knew this already.

Now, today we have a different kind of regression equation. We have:

$$\log(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2$$

We can still differentiate both sides just as we did a moment ago:

$$\frac{\partial}{\partial X_2} \log(Y) = \frac{\partial}{\partial X_2} (\alpha + \beta_1 X_1 + \beta_2 X_2)$$

Notice that right hand side is no different than the previous case. So we have:

$$\frac{\partial}{\partial X_2} \log(Y) = \beta_2$$

But, because of the $\log(\cdot)$, the left hand side obeys different rules of calculus than in the case above. It works out to:

$$\frac{\partial}{\partial X_2} \log(Y) = \frac{\frac{\partial}{\partial X_2}(Y)}{Y}$$

Thus:

$$\frac{\frac{\partial}{\partial X_2}(Y)}{Y} = \beta_2$$

So you can see that in this case, the interpretation of β_2 is that of a *proportional* change in Y given a unit change in X_2, holding all else constant. Proportional change because it is the change in Y divided by the value of Y. An example: if the price of bananas goes up by \$0.04/lb starting from \$0.40/lb, that is a (.04/.4)=.1 (or 10%) proportional change ((Δ_{price})/price). Logged dependent variable; proportional change. In the old case we had instead: Δ_{price}=\$0.04; absolute change. *Both formulations have their uses, as you can see from the banana example. Which one is better depends on what you are trying to accomplish.* Do you want to know that bananas cost four cents more per pound? Or do you want to know that they cost ten percent more? Either one is potentially good information — it just depends on how you want to think about the problem. Yes, we have no bananas.

Recall that regressions deal with predictions of the mean value. Without going through the mathematics, note that the logged case refers to the proportional change in the geometric mean, while the regular case refers to the absolute change in the arithmetic mean. The geometric mean is a cousin of the arithmetic mean; don't worry too much about this right now.[1]

[1]The geometric mean is: $\exp[\sum \log(w)/N]$ whereas the arithmetic mean is: $\sum(w)/N$.

There is one more "wrinkle" that needs to be mentioned. In order to properly interpret the β coefficient as proportional change, it must be exponentiated. Proportional *rates* of change need to be exponentiated to yield actual *proportional changes* (viz., β_2 is a proportional *rate* of change — and because of this mathematical oddity, it needs to be exponentiated[2]). In other words, $\exp(\beta_2)$ is the proportional change in the geometric mean of Y for a unit change in X_2 holding all else constant. Here is a concrete example:

In our dataset, men and women have different geometric mean salary. We can see this by using the STATA commands **`means salary if sex==0`** for men and **`means salary if sex==1`** for women. It looks like this:

```
. means salary if sex==0

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+------------------------------------------------------------
      salary | Arithmetic     483    218.3892        206.9816   229.7969
             |  Geometric     483    193.6175        185.8568   201.7022
             |   Harmonic     483    177.0213        171.2967   183.1416
-------------------------------------------------------------------------

. means salary if sex==1

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
-------------+------------------------------------------------------------
      salary | Arithmetic    2648    145.5604        143.9073   147.2136
             |  Geometric    2648    141.3544        140.1525   142.5666
             |   Harmonic    2648    138.2386        137.2279   139.2644
-------------------------------------------------------------------------
```

So, we see that men have a geometric mean salary of 193.6175 and women have a geometric mean salary of 141.3544. The proportional change between the sexes is $141.3544/193.6175 = .73007037 \doteq 73\%$.

Now let's run a simple regression in STATA:

```
. reg logsal sex
------------------------------------------------------------------------------
      logsal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.3146141   .0135258   -23.26   0.000    -.3411344   -.2880937
       _cons |   5.265884   .0124389   423.34   0.000     5.241495    5.290274
------------------------------------------------------------------------------
```

[2] Another example to illustrate why you need to exponentiate. Suppose you put \$100.00 into a savings account with continuous compounding and expect to earn \$10.00 of interest at the end of a year. The annual interest *rate* [APR, annual percentage rate] is not 10% [that's the APY, annual percentage yield]. The formula is $P_1 = P_0 \exp(rt)$. In our example, t is 1 (year) and P_0 is \$100, and P_1 is \$110, so: $\frac{110}{100} = \exp(r)$, or $r = \log(1.1) \doteq 0.0953$. The APR is about 9.53%. Our coefficients are like these interest rates in a continuously-compounded account... you need to exponentiate.

The sex coefficient is $-.31$, which tells us that women have *approximately* 31% lower geometric mean salary than men. And, indeed, since we already know that female geometric mean salary in these data is 73% of men's, which corresponds to a 27% drop, we know that this is approximately correct. But why is it not *exactly* correct? It will be if we exponentiate.

In STATA:

```
. disp exp(-.3146141)
.73007055
```

Here we have agreement out to the sixth decimal place, which is exact for all intents and purposes.

To summarize:

Exact agreement:

$$\frac{\text{geometric mean, females}}{\text{geometric mean, males}} = \frac{141.3544}{193.6175} = .73007037$$

$$\exp(\beta_{\text{sex}}) = \exp(-.3146141) = .73007055$$

Approximate agreement:
Female geometric mean salary is 73% that of males. Thus, females have 27% lower average salary. $\beta_{\text{sex}} = -0.31 \approx -0.27$.

The closer β_{sex} is to zero, the better this approximation works.
