# EEP/IAS 118 - Introductory Applied Econometrics, Lecture 8

Gregory Lane

July 2017

# This Lecture

Topics

- Scaling & Standardized Effects
- Confidence Intervals for Predictions
- Choice between non-nested models

# Scaling Variables

Often times the units that variables come in are not the most useful for interpretation or analysis.

- Rescaling monetary units - $ thousands, $ billions, etc.
- Distance per second into distance per hour

Example:

$$\widehat{sleep} = 3315.574 - 12.189 educ + 2.7454 age$$

Where sleep is measured in minutes per night. Here, $\hat{\beta}_{educ}$ is interpreted:

- One more year of education is estimated to decrease predicted sleep by 12.189 minutes per week, holding age constant

# Scaling Variables

$$\widehat{sleep} = 3315.574 - 12.189educ + 2.7454age$$

Lets say we instead want to change the dependent variable to be measured in hours rather than minutes.

- Do this simply by changing our $y$ variable into $\tilde{y} = \frac{y}{60}$

How would this change our $\hat{\beta}$?

- The new $\beta_{educ}$ estimate would be $\frac{12.189}{60} = 0.2$ hours per night

The entire regression result changes to this:

$$\widehat{sleep} = 55.260 - .2032educ + .0458age$$

# Scaling Variables

In general, when we re-scale the outcome variable by $\alpha$

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + ... + \tilde{\beta}_k x_k + u$$
$$\alpha y = \alpha\beta_0 + \alpha\beta_1 x_1 + ... + \alpha\beta_k x_k + u$$

In the above example, $\alpha = \frac{1}{60}$, so the new $\hat{\beta}$s will be divided by 60 too.

- **Note:** nothing else about the regression will change ($R^2$, t-stats, p-values, etc.)

# Scaling Variables

Let's say instead we rescale an independent $x$ variable:

- Rescale education to be in units of half-years (6 months) - i.e. we multiply $educ$ by 2
- The new regression would give us:

$$\widehat{sleep} = 3315.574 - 6.095 educ + 2.7454 age$$

- Only the coefficient on the independent variable we modified has changed

# Scaling Variables

In general, if we scale $x$ by $\alpha$, the equation becomes:

$$\begin{aligned} y =& \beta_0 + \tilde{\beta}_1 \tilde{x}_1 + ... + \beta_k x_k + u \\ =& \beta_0 + \frac{\beta_1}{\alpha}(\alpha x_1) + ... + \beta_k x_k + u \end{aligned}$$

- In the above example, we had $\alpha = 2$, which meant we had to scale our estimate of $\hat{\beta}_{educ}$ by $\frac{1}{2}$.

# Standardizing Variables

Up until now we've been considering cases where we want to change the units of a variable into units that are more useful

- But what if we don't want units at all? Why would we want this?
  - Want to compare the relative effects of two variables that don't have the same unit - e.g. education and SAT score on income
- This is useful for many economic models: Hedonic Price Model

# Hedonic Price Model

Idea behind Hedonic Price Model:

- We want to measure "Willingness to Pay" (WTP) for certain amenities:
  - Environmental amenities (clean water, clean air, parks, ect.)
  - House Characteristics ( school district, local pollution, etc.)
- These can be difficult to measure, as most people are never asked to explicitly "buy" these goods
- How do we measure their value:
  1. Directly ask: "What is your WTP?" via survey
     *Problem:* Question framing important, people will inflate / deflate values because choice is not real
  2. Revealed Preference: you reveal your preference for amenities via the value you paid to obtain them

# Hedonic Price Model

Two common ways to do this:

1. **Travel cost method:** Used for value of fishing / beaches - you can infer the value of these amenities by how much people pay to travel to access them (especially over closer locations without these amenities)
2. **Hedonic Price:** When you choose a place to live, your WTP for the house reveals your preference for the value of all the amenities the house has access to

Hedonic Price:

$$price = f(\#rooms,\ size\, yard,\ \ldots,\ pollution,\ crime,\ school\, quality)$$

# Hedonic Price Model

$$price = f(\#rooms,\ size\, yard,\ \ldots, pollution,\ crime,\ school\, quality)$$

Would use data from housing sales, with price, and all information about the house and location we can and then we would run a linear model:

$$price = \beta_0 + \beta_1 NO_2 + \beta_2 crime + \beta_3 rooms ...$$

# Hedonic Price Model

```
. reg price nox crime dist rooms lowstat stratio, beta

      Source |       SS       df       MS
-------------+------------------------------
       Model |  3.0150e+10     6  5.0250e+09
    Residual |  1.2675e+10   499  25401468.7
-------------+------------------------------
       Total |  4.2826e+10   505    84803032

------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------
         nox |  -1757.656   331.4642    -5.30   0.000
       crime |  -80.57672   30.47786    -2.64   0.008
        dist |  -1202.372   170.5011    -7.05   0.000
       rooms |   4412.584   415.8469    10.61   0.000
     lowstat |  -519.7665   48.41627   -10.74   0.000
     stratio |   -998.834    115.819    -8.62   0.000
       _cons |    34431.7   4732.075     7.28   0.000
------------------------------------------------------
```

# Hedonic Price Model

```
. reg price nox crime dist rooms lowstat stratio, beta

      Source |       SS       df       MS
-------------+------------------------------
       Model |  3.0150e+10     6  5.0250e+09
    Residual |  1.2675e+10   499  25401468.7
-------------+------------------------------
       Total |  4.2826e+10   505    84803032

------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------
         nox |  -1757.656   331.4642    -5.30   0.000
       crime |  -80.57672   30.47786    -2.64   0.008
        dist |  -1202.372   170.5011    -7.05   0.000
       rooms |   4412.584   415.8469    10.61   0.000
     lowstat |  -519.7665   48.41627   -10.74   0.000
     stratio |   -998.834    115.819    -8.62   0.000
       _cons |    34431.7   4732.075     7.28   0.000
------------------------------------------------------
```

But we want to *compare* these coefficients!

- What do consumers value more - crime or pollution?
- Problem is that crime and pollution have vastly different ranges

We can do this by *standardizing* the variables in the regression

# Standardize Variables

Standardizing means we will compare how a one standard deviation increase in $x_1$ affects $y$ to how a one Stdev increase in $x_2$ affects $y$

We do this by transforming all our variables by subtracting their mean and dividing by the standard deviation:

$$\tilde{y} = \left( \frac{y - \bar{y}}{\hat{\sigma}_y} \right)$$
$$\tilde{x} = \left( \frac{x_1 - \bar{x}_1}{\hat{\sigma}_{x_1}} \right)$$

This should look familiar - this is what we do with our t-stats! Idea is that we put all the variables on the same scale. Then we can compare relative effects.

# Standardize Variables

Once we standardize all the units, re-running the regression produces:

$$\left(\frac{y-\bar{y}}{\hat{\sigma}_y}\right) = \frac{\hat{\sigma}_{x_1}}{\hat{\sigma}_y}\hat{\beta}_1\left(\frac{x_1-\bar{x}}{\hat{\sigma}_{x_1}}\right) + \frac{\hat{\sigma}_{x_2}}{\sigma_y}\hat{\beta}_2\left(\frac{x_2-\bar{x}_2}{\hat{\sigma}_{x_2}}\right)$$

- The new parameters will be equal to the old parameters scaled by $\frac{\hat{\sigma}_{x_1}}{\hat{\sigma}_y}$
- This is called the "standardized coefficient" or the "beta coefficient"
- Note there is no $\beta_0$ because it will be zero (why?)
- In Stata we can produce these coefficients with the "beta" option (because transforming each variable is a pain)

# Standardize Variables

```
. reg price nox crime dist rooms lowstat stratio, beta

      Source |       SS       df       MS              Number of obs =     506
-------------+------------------------------           F(  6,   499) =  197.82
       Model |  3.0150e+10     6  5.0250e+09           Prob > F      =  0.0000
    Residual |  1.2675e+10   499  25401468.7           R-squared     =  0.7040
-------------+------------------------------           Adj R-squared =  0.7005
       Total |  4.2826e+10   505    84803032           Root MSE      =    5040

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
         nox |  -1757.656   331.4642    -5.30   0.000                -.2210981
       crime |  -80.57672   30.47786    -2.64   0.008                -.0751639
        dist |  -1202.372   170.5011    -7.05   0.000                -.2749917
       rooms |   4412.584   415.8469    10.61   0.000                 .3366601
     lowstat |  -519.7665   48.41627   -10.74   0.000                -.4085311
     stratio |   -998.834    115.819    -8.62   0.000                -.2349146
       _cons |    34431.7   4732.075     7.28   0.000                        .
------------------------------------------------------------------------------
```

- One SD increase in pollution leads to 0.22 SD decrease in prices
- One SD increase in crime leads to a 0.07 SD decrease in prices

# Confidence Intervals for $y$

There are some instances where we may care about the predicted value of the dependent variable $y$

We know that the estimated regression give us $\hat{y}$ which is our best guess for $y$ for and given $x$. However, $\hat{y}$ is a random variable (just like $\hat{\beta}$) and therefore has uncertainty.

- We can quantify this uncertainty and create a confidence interval for $\hat{y}$ for any specific combination of $x_j$

# Confidence Intervals for $y$

However, there are two types of CI that we may want to calculate:

1. A confidence interval for the **average** y given $x_1, ..., x_k$
2. A confidence interval for a **particular** y given $x_1, ..., x_k$

You can think of the difference as being the answer to these two questions:

1. How uncertain are we about the average income for this type of person?
2. If we asked a person of this type their income, what range would cover 95% of responses

# Confidence Intervals for average $y$

Recall that regression gives us an estimate of $y$ given $x$:

$$\widehat{\mathbb{E}}[y|x_1, x_2, x_2] = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

- If we want the best estimate for a particular value of $x_j$, we just plug those values into the equation
- To get a CI, then we only need to find the stand error for this prediction
- Recall that $\beta_0$ takes on the predicted value of $y$ when all the $x_j$ are zero

$$\hat{\beta}_0 = \hat{E}(y|x_1 = 0, x_2 = 0, x_3 = 0)$$

# Confidence Intervals for average $y$

$$\hat{\beta}_0 = \hat{E}(y|x_1 = 0, x_2 = 0, x_3 = 0)$$

- Therefore, if we transform our $x_j$ by subtracting the values ($\alpha_j$) for which we want a prediction:

  $$y = \beta_0 + \beta_1(x_1 - \alpha_1) + \beta_2(x_2 - \alpha_2) + \beta_3(x_3 - \alpha_3)$$

  Then

  $$\hat{\beta}_0 = \hat{E}(y|x_1 = \alpha_1, x_2 = \alpha_2, x_3 = \alpha_3)$$

  When we run the regression with these transformed variables, $\hat{\beta}_0$ will then be best prediction and Stata will produce the correct SE

# Confidence Intervals for average $y$

Process Summary for CI on average $y$:

1. Generate new variables: $\tilde{x}_j = x_j - \alpha_j$.
2. Run the regression of: $y = \tilde{\beta}_0 + \tilde{\beta}_1\tilde{x}_1 + ... + \tilde{\beta}_k\tilde{x}_k + \tilde{u}$
3. Then $\hat{\mathbb{E}}[y|x_1 = \alpha_1, ..., x_k = \alpha_k] = \tilde{\beta}_0$ and the standard error for this estimate is $SE(\tilde{\beta}_0)$.
4. Plug these values into the formula for confidence intervals and interpret.

$$[\tilde{\beta}_0 - c \cdot SE(\tilde{\beta}_0)\ ,\ \tilde{\beta}_0 + c \cdot SE(\tilde{\beta}_0)]$$

# Confidence Intervals for average $y$: Example

For an example, let's use Woolridge's birthweight data. Let's say we want to find a prediction for average birthweight for babies with family income of \$14,500 ($ln(14.5) = 2.674$), mothers with 12 years of education, and with 2 older siblings ($parity = 3$)

Running the standard regression:

$$\widehat{bwght} = 105.66 + 2.13ln(famine) + 0.317meduc + 1.53parity$$
$$\hat{y} = 105.66 + 2.13(2.674) + .317(12) + 1.53(3)$$
$$= 119.75 \text{ ounces}$$

Which is our best guess for $\hat{y}_{faminc=14.5,meduc=12,parity=3}$

# Confidence Intervals for average $y$: Example

To get the SE of this prediction, we run:

$$bwght = \beta_0 + \beta_1(lfaminc - 2.674) +$$

$$\beta_2(meduc - 12) + \beta_3(parity = 3) + u$$

| bwght | Coef. | Std. Err. | t | P>\|t\| |
|---|---|---|---|---|
| lfaminc_0 | 2.131266 | .6505986 | 3.28 | 0.001 |
| meduc_0 | .3171976 | .2519682 | 1.26 | 0.208 |
| parity_0 | 1.526144 | .6119145 | 2.49 | 0.013 |
| _cons | 119.6405 | 1.006928 | 118.82 | 0.000 |

Note, how now the "cons" takes on the predicted value and has a standard error!

# Confidence Intervals for average $y$: Example

Using this output, the 95% confidence interval for the average birthweight for babies given family income of \$14,500 ($ln(14.5) = 2.674$), mothers with 12 years of education, and with 2 older siblings ($parity = 3$) is:

$$\left[119.64 - 1.96(1.007), 119.64 + 1.96(1.007)\right] = \left[117.6653, 121.6158\right]$$

# Confidence Intervals for a particular $y$

Now let's turn to how we can create a confidence interval of $y$ for a *particular* individual with certain $x$.

- Again, this is different (larger) than our CI for the *average* $y$ in a sub-popultaion
- This is because we need to account for both the variance in our calculation of $\hat{y}$ as well as the variance unobserved error term $u$

# Confidence Intervals for a particular $y$

Let's see how to think about this using our example. Let $bwght^0$ denote the value for which we want to construct a confidence interval:

$$bwght^0 = \beta_0 + \beta_1 lfaminc^0 + \beta_2 meduc^0 + \beta_3 parity^0 + u^0$$

Our best prediction of $bwght^0$ is $\widehat{bwght^0}$, where

$$\widehat{bwght}^0 = \hat{\beta}_0 + \hat{\beta}_1 lfaminc^0 + \hat{\beta}_2 meduc^0 + \hat{\beta}_3 parity^0$$

Now there is some error associated with using $\widehat{bwght}^0$ to predict $bwght^0$:

$$\hat{u}^0 = bwght^0 - \widehat{bwght}^0 = \beta_0 + \beta_1 lfaminc^0 + \beta_2 educ^0 + \beta_3 par^0 + u^0 \\ - \hat{\beta}_0 + \hat{\beta}_1 lfaminc + \hat{\beta}_2 meduc + \hat{\beta}_3 parity$$

# Confidence Intervals for a particular $y$

To get a confidence interval, we need to quantify the variance of the error in this prediction:

$$
\begin{aligned}
Var(\hat{u}^0) &= Var(bwght^0 - \widehat{bwght}^0) \\
&= Var(\beta_0 + \beta_1 lfaminc^0 + \beta_2 educ^0 + \beta_3 parit^0 + u^0 - \widehat{bwght}^0) \\
&= Var(\widehat{bwght}^0) + Var(u^0) \\
&= Var(\widehat{bwght}^0) + \sigma^2 \\
\widehat{Var(\hat{u}^0)} &= Var(\widehat{bwght}^0) + \hat{\sigma}^2 \\
&= Var(\widehat{bwght}^0) + \frac{\sum \hat{u}_i^2}{n-k-1} = Var(\widehat{bwght}^0) + \frac{SSR}{n-k-1}
\end{aligned}
$$

# Confidence Intervals for a particular $y$

$$Var(\widehat{bwght}^0) + \frac{SSR}{n-k-1}$$

There are two sources of variation in $\hat{u}^0$

1. The sampling error in $\widehat{bwght}^0$ which arises because we have estimated the population parameters ($\beta$).
2. The variance of the error in the population ($u^0$).

- Compute the $Var(\widehat{bwght}^0)$ exactly as before
- Second we can compute $\frac{SSR}{n-k-1}$ from our regression output
- Then the 95% confidence interval for $bwght^0$:

$$\hat{y} \pm 1.96 \cdot se(\hat{u}^0)$$

# Confidence Intervals for a particular $y$: Summary

1. Generate new variables: $\tilde{x}_j = x_j - \alpha_j$.
2. Run the regression of: $y = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_1 + ... + \tilde{\beta}_k \tilde{x}_k + \tilde{u}$
3. Then $\hat{\mathbb{E}}[y | x_1 = \alpha_1, ..., x_k = \alpha_k] = \tilde{\beta}_0$ and the standard error for this estimate is $SE(\tilde{\beta}_0)$.
4. Get an estimate for the variance of $\hat{u} = \hat{\sigma}^2$ from the Stata output.
5. Compute the standard error: $\sqrt{SE(\tilde{\beta}_0)^2 + \hat{\sigma}^2}$.
6. Plug these values into the formula for confidence intervals and interpret.

# Choice Between Non-Nested Models

You've been asked to do the following in past problems:

1. Deciding if one of your x variables is significant $\Rightarrow$ t-test
2. Deciding if multiple variables *together* are significant $\Rightarrow$ F-test.

These tests compare *nested* models

- Nested models are cases where one equation is just a special case of the other (e.g. fixing $\beta_3$ and $\beta_4 = 0$)

How do we compare *non-nested* models?

- Use **Adjusted** $R^2$

# Adjusted $R^2$: Comparing Non-nested Models

Regular $R^2$ is a measure of "goodness of fit", so why not just use that?

- $R^2$ will always (weakly) increase when you add more variables to the regression
- Not useful to choosing which model is better, more complex one will always win

Therefore, we use Adjusted $R^2$ which adds a penalty for each additional variable added to the model

# Adjusted $R^2$

The formula for adjusted $R^2$ is:

$$1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

Adding variables now has two effects:

1. The $SSR$ in the numerator will always (weakly) decrease with an additional variable
2. However, $k$ will also increase (making the numerator larger)

Therefore, the effect on the adjusted $R^2$ from adding an additional variable to the regression will depend on if the extra explanatory power is larger than the penalty

# Adj $R^2$: Two Models of Sleep

When does Adj $R^2$ come in handy:

- Choosing a functional form for the right hand side variables can be difficult
- A common example is a choice between $log(x)$ and a quadratic $x$ and $x^2$
- Both can be reasonable choices and it is difficult to eyeball which is better

# Adj $R^2$: Two Models of Sleep

```
reg sleep lnage

      Source |       SS       df       MS              Number of obs =     706
-------------+------------------------------           F(  1,   704) =    4.54
       Model |  891303.042     1  891303.042           Prob > F      =  0.0335
    Residual |   138348533   704  196517.802           R-squared     =  0.0064
-------------+------------------------------           Adj R-squared =  0.0050
       Total |   139239836   705  197503.313           Root MSE      =   443.3

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnage |   122.9174   57.71672     2.13   0.034     9.599897    236.2349
       _cons |   2821.777   209.4207    13.47   0.000     2410.613    3232.941
------------------------------------------------------------------------------


      Source |       SS       df       MS              Number of obs =     706
-------------+------------------------------           F(  2,   703) =    5.22
       Model |  2039007.98     2  1019503.99           Prob > F      =  0.0056
    Residual |   137200828   703  195164.762           R-squared     =  0.0146
-------------+------------------------------           Adj R-squared =  0.0118
       Total |   139239836   705  197503.313           Root MSE      =  441.77

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -21.4904   11.73674    -1.83   0.068    -44.53366    1.552851
  agesquared |   .3011932    .140117     2.15   0.032     .0260954     .576291
       _cons |    3608.03   230.6457    15.64   0.000     3155.193    4060.867
------------------------------------------------------------------------------
```

Which do we prefer? Look at $Adj - R^2$

# Choice Between $y$ and $ln(y)$

Rather than trying to choose between what $x$ to include in a model, what if we are trying to choose between different functional forms of $y$? A common example is the choice between $y$ and $ln(y)$

A natural choice might be to run both regressions:

$$y = \hat{\beta}_0 + \hat{\beta}1x_1 + ... + \hat{\beta}_k x_k + \hat{u}$$
$$ln(y) = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + ... + \hat{\alpha}_k x_k + \hat{u}$$

And then look at the $R^2$ of each model to decide the best fit. But this is **wrong**

# Choice Between $y$ and $ln(y)$

How can we see the problem? Remember that $R^2 = corr(y, \hat{y})^2$ so what we're actually comparing is:

$$corr(y, \hat{y})^2 \qquad \text{to} \qquad corr(ln(y), \widehat{ln(y)})^2$$

- Comparing the $R^2$ for each model isn't an apples to apples comparison
- We need to do something else

# Choice Between $y$ and $ln(y)$

**Process to choose:**

1. Estimate the log model: $ln(y) = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + ... + \hat{\alpha}_k x_k + u$
2. Predict: $y$ from the log model: $\hat{y} = e^{\widehat{ln(y)}} e^{\frac{\hat{\sigma}^2}{2}}$
3. Find the correlation and square it (to get alternative $R^2_{log}$): $y$ and the $\hat{y}$ from the log model. This gives us an alternative $R^2$.
4. Estimate the linear model (to get $R^2_{lin}$): $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k + u$ and get it's $R^2$
5. Compare $R^2_{lin}$ to $R^2_{log}$ and choose the higher one

**Note:** that to predict $\hat{y}$ from $\widehat{ln(y)}$, you need to raise to the exponential *and* multiply by $e^{\frac{\hat{\sigma}^2}{2}}$ (where $\hat{\sigma}^2$ is found in the regression output under MS residual)

# Choice Between $y$ and $ln(y)$: Example

```
. reg lprice llotsize lsqrft bdrms
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   50.42
       Model |  5.15504028     3  1.71834676           Prob > F      =  0.0000
    Residual |  2.86256324    84  .034078134           R-squared     =  0.6430
-------------+------------------------------           Adj R-squared =  0.6302
       Total |  8.01760352    87  .092156362           Root MSE      =   .1846

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    llotsize |   .1679667   .0382812     4.39   0.000     .0918404     .244093
      lsqrft |   .7002324   .0928652     7.54   0.000     .5155597    .8849051
       bdrms |   .0369584   .0275313     1.34   0.183    -.0177906    .0917074
       _cons |  -1.297042   .6512836    -1.99   0.050    -2.592191   -.001893
------------------------------------------------------------------------------
```

# Choice Between $y$ and $ln(y)$: Example

```
. reg price lotsize sqrft bdrms
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   57.46
       Model |  617130.701     3  205710.234           Prob > F      =  0.0000
    Residual |  300723.805    84   3580.0453           R-squared     =  0.6724
-------------+------------------------------           Adj R-squared =  0.6607
       Total |  917854.506    87  10550.0518           Root MSE      =  59.833

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |   .0020677   .0006421     3.22   0.002     .0007908    .0033446
       sqrft |   .1227782   .0132374     9.28   0.000     .0964541    .1491022
       bdrms |   13.85252   9.010145     1.54   0.128    -4.065141    31.77018
       _cons |  -21.77031   29.47504    -0.74   0.462    -80.38466    36.84405
------------------------------------------------------------------------------
```

# Choice Between $y$ and $ln(y)$: Example

**Process:**

1. Estimate the log model:
   reg log(price) log(lotsize) log(sqrft) bdrms
2. Get your predictions from this regression: predict lpricehat
3. Predict: y from the log model: gen
   pricehat=exp(lpricehat)*exp(.034078134/2)
4. Find the correlation and square it: correl price pricehat
   $(= 0.7377)$
5. Find the $R^2$ from the linear regression $(R^2 = 0.6724)$
6. Compare: For predicting price, the log model is notably better.

# Review

1. One parameter: use t-stat and test significance
2. Multiple parameters: use F-test
3. Choosing between two non-nested models: Adj $R^2$
4. Choosing between different functional forms for $y$ ($y$ and $ln(y)$), use process above