EEP/IAS 118
Andrew Dustan
Section Handout 9

1. Confidence Intervals for Predictions

We already know how to predict the value of *y* for given values of *x*. But the prediction is just a guess, so we also want to know how sure of this guess we are.

There are two kinds of predictions we might want to make:

(a) Predictions about the *average* value of *y* in the population, given *x*

These predictions basically use our regression model to answer the question:

"What is the average value of *y* in the population for everyone with this value of *x*?"[1]

(Note that here, *x* can refer to a list of variables: $x_1, x_2, \dots, x_k$.)

Given a value of $x = x^*$, how do we get this predicted average *y* and its standard error?

1. Remember, one objective of doing regressions is to estimate $\hat{\beta}$'s that let us take x^* and get an expected (average) value of *y*. That is, the regression gives us $\hat{E}(y|x = x^*)$. Problem is, they *don't* automatically give us standard errors for that expected *y*.
2. Realize that regressions *do* give us standard errors for each estimated coefficient. We want to use these standard errors to get the standard error for $\hat{E}(y|x = x^*)$.
3. Run this regression, modified from its simple form just to let us use the standard error of $\hat{\beta}_0$:
$$y = \beta_0 + \beta_1(x_1 - x_1^*) + \beta_2(x_2 - x_2^*) \dots + \beta_k(x_k - x_k^*) + u$$
Notice: when $x = x^*$, we simply have $y = \beta_0 + u$.

   Estimating the above regression will get you (among other things) $\hat{\beta}_0$ and its standard error.
4. Since $y = \beta_0 + u$ when $x = x^*$, we see that $\hat{E}(y|x = x^*) = \hat{\beta}_0$. Furthermore, we have an important result: $SE\left(\hat{E}(y|x = x^*)\right) = SE(\hat{\beta}_0)$.
5. Now you have both the things you need: $\hat{E}(y|x = x^*)$ and its standard error. Use this to construct confidence intervals for $E(y|x = x^*)$ or perform hypothesis tests, which you already know how to do.

Example (from Wooldridge):

How does the price of an airplane ticket depend on distance flown and the number of people aboard?

$$price = \beta_0 + \beta_1 distance + \beta_2 \log(passengers) + u$$

price: average one-way ticket price on route ($)
distance: distance of one-way flight (miles)
passengers: average number of passengers aboard flights on this route

I estimated this using 4596 domestic routes between 1997 and 2000. Here is the Stata output:

[1] Note to the curious: this statement isn't exactly right when some of the *x* variables are continuous: e.g., no two people in the population have *exactly* the same height or weight so we can't talk about the average of *y* over everyone with the exact same height and weight. We're actually talking about conditional expectations here, which can generalize our statement to continuous variables. We're really answering the question, "What is the conditional expectation of *y* given *x*?" But the wording I gave seems more intuitive.

```
      Source |       SS       df       MS              Number of obs =    4596
-------------+------------------------------           F(  2,  4593) = 1548.22
       Model |  10375345.9     2  5187672.95           Prob > F      =  0.0000
    Residual |  15389926.3  4593  3350.73509           R-squared     =  0.4027
-------------+------------------------------           Adj R-squared =  0.4024
       Total |  25765272.2  4595  5607.24096           Root MSE      =  57.886

------------------------------------------------------------------------------
        fare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |   .0749986   .0014016    53.51   0.000     .0722509    .0777463
     lpassen |  -10.01041   .9700543   -10.32   0.000    -11.91218   -8.108635
       _cons |   164.8002   6.180564    26.66   0.000     152.6833    176.9171
------------------------------------------------------------------------------
```

Question: On average, how much would we expect flights of **500 miles** with 200 passengers [$\mathbf{\log(passengers) = 5.30}$)] to cost?

We need to run this regression again, modifying the *x* values we put into the regression. How?

$$price = \beta_0 + \beta_1(distance - 500) + \beta_2(\log(\text{passengers}) - 5.30) + u$$

Let's do it:

```
      Source |       SS       df       MS              Number of obs =    4596
-------------+------------------------------           F(  2,  4593) = 1548.22
       Model |  10375345.9     2  5187672.95           Prob > F      =  0.0000
    Residual |  15389926.3  4593  3350.73509           R-squared     =  0.4027
-------------+------------------------------           Adj R-squared =  0.4024
       Total |  25765272.2  4595  5607.24096           Root MSE      =  57.886

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dist_500 |   .0749986   .0014016    53.51   0.000     .0722509    .0777463
 lpassen_200 |  -10.01041   .9700543   -10.32   0.000    -11.91218   -8.108636
       _cons |   149.2612   1.331763   112.08   0.000     146.6503    151.8721
------------------------------------------------------------------------------
```

Read off the answers from entries corresponding to $\hat{\beta}_0$:

$\hat{E}(price \mid distance = 500, passengers = 200)$	**$149.2612**
$SE(\hat{E}(price \mid distance = 500, passengers = 200))$	**$1.331763**
95% CI for $\hat{E}(price \mid distance = 500, passengers = 200)$	**[$146.6503, $151.8721]**

So we have a very precise estimate of the *average* ticket price for all flights of 500 miles and 200 passengers.

(b) Predictions about a *particular* value of *y* in the population, given *x*

These predictions basically use our regression model to answer the question:

"What is the value of *y* for a *specific person/house/flight/etc.* in the population, given that I know its *x*?"

To answer this question, recall the result of some work done in lecture:

$$var(\hat{y}) = var\left(\hat{E}(y|x = x^*)\right) + var(u)$$

This makes sense because there are two sources of error in our prediction, $\hat{y}$:

1. Error in estimating $\hat{E}(y|x = x^*)$, the average *y* given $x = x^*$. This is the same as in part (a).
2. Unobservable characteristics of the person/house/flight/whatever being different from zero.

Let's return to the example to show that this prediction error for $\hat{y}$ is higher than for $\hat{E}(y|x = x^*)$.

Example:

Question: I am about to book Flight 1154 from San Francisco to San Diego ($distance \approx 500$ miles). This flight usually has about 200 passengers. How much will my ticket cost?

Reproducing the table from the last page:

```
      Source |       SS       df       MS              Number of obs =    4596
-------------+------------------------------           F(  2,  4593) = 1548.22
       Model |  10375345.9     2  5187672.95           Prob > F      =  0.0000
    Residual |  15389926.3  4593  3350.73509           R-squared     =  0.4027
-------------+------------------------------           Adj R-squared =  0.4024
       Total |  25765272.2  4595  5607.24096           Root MSE      =  57.886

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dist_500 |   .0749986   .0014016    53.51   0.000     .0722509    .0777463
 lpassen_200 |  -10.01041   .9700543   -10.32   0.000    -11.91218   -8.108636
       _cons |   149.2612   1.331763   112.08   0.000     146.6503    151.8721
------------------------------------------------------------------------------
```

We want $SE(\widehat{price})$ for this particular flight. To get it, we need estimates for each of the following:

Population value	Estimate	Answer (a number)
$var\left(\hat{E}(y\|x = x^*)\right)$	$SE(\hat{\beta}_0)^2$	$1.33^2 = 1.77$
$var(u)$	$\hat{\sigma}^2 = \frac{SSR}{n-k-1}$	3350.74
$var(\hat{y}) = var\left(\hat{E}(y\|x = x^*)\right) + var(u)$	$SE(\hat{\beta}_0)^2 + \hat{\sigma}^2$	$1.77 + 3350.74 = 3352.51$

Take the square root of the last one to get the standard error, $SE(\widehat{price})$: _____$57.90_____

You can see that this is much bigger than the estimate for the average price of all flights with these observed characteristics. Even if you had an infinite number of observations, $SE(\widehat{price})$ would still be big, because almost all of the variance in the prediction is coming from the unobservables, not estimation error! Having a really great guess for the average price doesn't help you get rid of uncertainty due to unobservables.

Note about predicting *y* when the dependent variable is $log(y)$:
After you get a prediction for $\log(y)$, you still need to turn it into a prediction for *y*. To do this, don't just use $\hat{y} = e^{\widehat{\log(y)}}$. This is wrong. From lecture we know you have to use this estimate:

$$\hat{y} = e^{\widehat{\log(y)}} * e^{\frac{\hat{\sigma}^2}{2}}$$

Example:
If we do the above regression but with **log** ($price$) as the dependent variable instead of *price*, we get:

```
      Source |       SS       df       MS              Number of obs =    4596
-------------+------------------------------           F(  2,  4593) = 1544.89
       Model |  351.936774     2  175.968387           Prob > F      =  0.0000
    Residual |    523.1576  4593  .113903244           R-squared     =  0.4022
-------------+------------------------------           Adj R-squared =  0.4019
       Total |  875.094374  4595  .190444913           Root MSE      =   .3375

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dist_500 |   .0004222   8.17e-06    51.66   0.000     .0004062    .0004382
 lpassen_200 |  -.0888662   .0056558   -15.71   0.000    -.0999543   -.0777781
       _cons |   4.952713   .0077647   637.85   0.000      4.93749    4.967936
------------------------------------------------------------------------------
```

What's log $(\widehat{price})$ for a flight of 500 miles and 200 passengers? __ $\log(\widehat{price}) = \hat{\beta}_0 = 4.95$___

What's $\frac{\hat{\sigma}^2}{2}$? ____$0.1139/2 = .057$_____

Then what is $\widehat{price}$ for this flight? ____ $\widehat{price} = e^{4.95} * e^{.057} = \$149.46$_____

Note: Computing the standard error of these predictions is complicated, so we did not cover it.

2. Comparing Goodness of Fit between Linear and Logarithmic Regressions

You should **never** compare the R^2 between regressions where one has *y* as the dependent variable and the other has $\log(y)$. How do you choose between the two models, then?

1. Run the regression for the linear model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ and get its R^2
2. Run the regression for the log model $\log(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
3. Predict $\hat{y}$ using the formula from above: $\hat{y} = e^{\widehat{\log(y)}} * e^{\frac{\hat{\sigma}^2}{2}}$
4. Compute $cor(y, \hat{y})^2$ using your predictions from the log model in (3). This is like an R^2 for how well the log model explains *y* (as opposed to $\log(y)$)
5. Compare the R^2 from (1) with your $cor(y, \hat{y})^2$ from (4). If the linear model R^2 is bigger, then the linear model has a better fit. If the $cor(y, \hat{y})^2$ from the log model is bigger, then the log model has a better fit.

Why does this work? Because for a linear regression $y = \beta_0 + \cdots$, it is true that $R^2 = cor(y, \hat{y})^2$. So this comparison is just between the $cor(y, \hat{y})^2$ from the linear and log models. The model that has a higher correlation between its in-sample predictions and the actual sample values is the winner.

Which model for predicting airline ticket prices is better, given that for the log model, $cor(y, \hat{y})^2 = .388$?

The linear model, because its $R^2 = cor(y, \hat{y})^2 = .4027$ is higher than the log model's $cor(y, \hat{y})^2 = .388$.