17.871 - Notes on PS2

Mike Sances

MIT

April 2, 2012

Interpreting Regression: Coefficient

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the coefficient estimate for distance.

Interpreting Regression: Coefficient

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the coefficient estimate for distance.
  - "A one-foot increase in distance is associated with a 4.1 percentage point decrease in the success rate."

Interpreting Regression: Coefficient

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the coefficient estimate for distance.
  - "A one-foot increase in distance is associated with a 4.1 percentage point decrease in the success rate."
  - "A one-foot increase in distance is associated with a 4.1% decrease in the success rate."

Interpreting Regression: Coefficient

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the coefficient estimate for distance.
  - "A one-foot increase in distance is associated with a 4.1 percentage point decrease in the success rate."
  - "A one-foot increase in distance is associated with a 4.1% decrease in the success rate."
    - No! 4.1% decrease in success rate means $.041 * .836 = 0.034$

Interpreting Regression: Coefficient

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the coefficient estimate for distance.
  - "A one-foot increase in distance is associated with a 4.1 percentage point decrease in the success rate."
  - "A one-foot increase in distance is associated with a 4.1% decrease in the success rate."
    - No! 4.1% decrease in success rate means $.041 * .836 = 0.034$
  - "A one-unit increase in X_k is associated with a $\hat{\beta}_k$ change in Y."

Interpreting Regression: Confidence Interval

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the confidence interval.

Interpreting Regression: Confidence Interval

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the confidence interval.
  - "If we were to repeatedly sample from the population and run this regression in each sample, then our confidence intervals will contain the true value of β in 95% of these samples."

Interpreting Regression: Confidence Interval

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the confidence interval.
  - "If we were to repeatedly sample from the population and run this regression in each sample, then our confidence intervals will contain the true value of β in 95% of these samples."
  - This gives us a measure of how uncertain we are about our estimate.

Interpreting Regression: Standard Error of Regression

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the Standard Error of Regression (SER, Stata calls it Root MSE).

Interpreting Regression: Standard Error of Regression

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the Standard Error of Regression (SER, Stata calls it Root MSE).
  - "On average, in sample predictions will be off the average in-sample mark by about 0.092."

Interpreting Regression: Standard Error of Regression

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the Standard Error of Regression (SER, Stata calls it Root MSE).
  - "On average, in sample predictions will be off the average in-sample mark by about 0.092."
  - "On average, in sample predictions will be off the average in-sample mark by about 9.2 percentage points."

Interpreting Regression: Standard Error of Regression

```
regress success_rate dist
      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  1,    17) =  113.35
       Model |  .952934346     1  .952934346           Prob > F      =  0.0000
    Residual |  .142915138    17  .008406773           R-squared     =  0.8696
-------------+------------------------------           Adj R-squared =  0.8619
       Total |  1.09584948    18  .060880527           Root MSE      =  .09169
------------------------------------------------------------------------------
success_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0408878   .0038404   -10.65   0.000    -.0489904   -.0327853
       _cons |   .8360873   .0471917    17.72   0.000     .7365215    .9356531
------------------------------------------------------------------------------
```

- Interpret the Standard Error of Regression (SER, Stata calls it Root MSE).
  - "On average, in sample predictions will be off the average in-sample mark by about 0.092."
  - "On average, in sample predictions will be off the average in-sample mark by about 9.2 percentage points."
- How "good" is the SER (9.2 percentage points) here?

Interpreting Regression: Standard Error of Regression

- Imagine you were playing golf and you found yourself 5 feet from the hole.
- Then the model tells us that your probability of success is

$$\begin{aligned} \hat{\alpha} + \hat{\beta} * 5 &= 0.84 + -0.04 * 5 \\ &= 0.84 - 0.20 \\ &= 0.64 \end{aligned}$$

and the SER tells us that you can expect this prediction to be off the mark by 0.092.

Observational Studies vs Randomization

- In observational studies, what are the two main problems researchers face with internal validity?

Observational Studies vs Randomization

- In observational studies, what are the two main problems researchers face with internal validity?
  - Problem 1: Confounding

Observational Studies vs Randomization

- In observational studies, what are the two main problems researchers face with internal validity?
  - Problem 1: Confounding
  - Problem 2: Reverse Causation

Observational Studies vs Randomization

- In observational studies, what are the two main problems researchers face with internal validity?
  - Problem 1: Confounding
  - Problem 2: Reverse Causation
- Why do experiments overcome these two problems?

Observational Studies vs Randomization

- In observational studies, what are the two main problems researchers face with internal validity?
  - Problem 1: Confounding
  - Problem 2: Reverse Causation
- Why do experiments overcome these two problems?
  - Random *assignment* to the treatment

Observational Studies vs Randomization

- In observational studies, what are the two main problems researchers face with internal validity?
  - Problem 1: Confounding
  - Problem 2: Reverse Causation
- Why do experiments overcome these two problems?
  - Random *assignment* to the treatment
  - Random *sampling* is not sufficient. Why not?

Random Sampling vs Random Assignment

- Say we have a model like this:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

And we're interested in the relationship between X_1 and Y. However, we aren't able to observe X_2, which is itself correlated with both X_1 and Y.

Random Sampling vs Random Assignment

- Say we have a model like this:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

  And we're interested in the relationship between X_1 and Y. However, we aren't able to observe X_2, which is itself correlated with both X_1 and Y.
- What type of variable is X_2?

Random Sampling vs Random Assignment

- Say we have a model like this:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

  And we're interested in the relationship between X_1 and Y. However, we aren't able to observe X_2, which is itself correlated with both X_1 and Y.
- What type of variable is X_2?
  - A "confound," "confounder," or "omitted variable."

Random Sampling vs Random Assignment

- Say we have a model like this:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

  And we're interested in the relationship between X_1 and Y. However, we aren't able to observe X_2, which is itself correlated with both X_1 and Y.
- What type of variable is X_2?
  - A "confound," "confounder," or "omitted variable."
- Confounding means we will not be able to estimate β_1 without bias, *even with an infinite and random sample.*

Random Sampling vs Random Assignment

- In the next slide I show regression estimates for β_1 when randomly sampling from a model like this:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- I set $\beta_2 = 2$ and $\beta_1 = 0$ and allow for some small correlation between X_1 and X_2.
- The black line is the estimate of β_1 for each *random* sample. I increase the size of the sample by 1 for each sample, starting with a sample of 10 and going to a sample of 1000. The red line is the true value of β_1.

Random Sampling vs Random Assignment