Lab 1: Theory vs. Simulation

Robin Mejia

Email: *rjmejia@berkeley.edu*

Office Hours: **Thursday, 9am-10:45am, 111A Haviland**

September 2, 2015

1. Theory
   - Distribution of Data
   - Prediction

2. Simulation
   - Estimates and Inference
   - Asymptotic Properties

Stata Code for Simulation 1, n=100

```
clear
**n=100
set obs 100
gen X1=runiform()
gen X2=1*X1+rnormal(-0.25,0.25)
scalar b0 = 0.5
scalar b1 = 1.0
scalar b2 = 0.0
gen Y = b0+b1*X1+b2*X2+ rnormal(0, 0.5)
```

Describe the True Distribution

Based on the code used to simulate the data, describe the true distribution of the data including the model of regression.

Example model of regression:

$$\mathbb{E}(Y|X_1 = x_1, X_2 = x_2) = b_0 + b_1 x_1 + b_2 x_2$$
$$\text{where } b_0 = 1.2, b_1 = 0.5, b_2 = 0.3$$

Solution

Here we are interested in the joint probability distribution of the data, that is how do you describe $Pr(Y = y, X_1 = x_1, X_2 = x_2)$? Remember, if Y, X_1, and X_2 were independent, we would have

$$Pr(Y = y, X_1 = x_1, X_2 = x_2) = Pr(Y = y) \times Pr(X_1 = x_1) \times Pr(X_2 = x_2)$$

However, a glance at the simulation code tells us that we definitely do not have independent random variables. We turn to conditional probability:

$$\begin{aligned} Pr(Y, X_1, X_2) &= Pr(Y|X_1, X_2) * Pr(X_1, X_2) \\ &= Pr(Y|X_1, X_2) * Pr(X_2|X_1) * Pr(X_1) \end{aligned}$$

Solution

Theoretical distribution of:

- X_1 - We used `gen X1=runiform()` to generate X_1, so we say

$$X_1 \sim Uniform(0,1)$$

- X_2 - Stata code: `gen X2 = X1 + rnormal(-0.25,0.25)`

$$X_2 \sim X_1 + Normal(\mu = -0.25, \sigma = 0.25)$$

- Y -
  Stata code: `gen Y = b0+b1*X1+b2*X2+ rnormal(0, 0.5)`

$$Y \sim b_0 + b_1 X_1 + b_2 X_2 + Normal(\mu = 0, \sigma = 0.5)$$

Model of Regression

Because we are given

```
scalar b0 = 0.5
scalar b1 = 1.0
scalar b2 = 0.0
gen Y = b0+b1*X1+b2*X2+ rnormal(0, 0.5)
```

We can say our model of regression is

$$
\begin{aligned}
\mathbb{E}(Y|X_1 = x_1, X_2 = x_2) &= \mathbb{E}(b_0 + b_1 X_1 + b_2 X_2 + e|X_1 = x_1, X_2 = x_2) \\
&= \mathbb{E}(0.5 + 1.0X_1 + 0X_2 + e|X_1 = x_1, X_2 = x_2) \\
&= 0.5 + 1.0x_1 + 0.0x_2 + \mathbb{E}(e|X_1 = x_1, X_2 = x_2) \\
&\text{Expectation of a normal random variable is ...} \\
&= 0.5 + x_1
\end{aligned}
$$

Next Question

Calculate the predicted value at $X_1 = 0, X_2 = 1$.

Solution

$$\mathbb{E}(Y|X_1 = 0, X_2 = 1) = 0.5 + 0 = 0.5$$

Note: remember that predicted values come from regressions.

Last but not Least

- What is the true change in the mean of Y when X_1 changes by 0.5?
- What is the true change in the mean of Y when X_2 changes at all?

Solution

Please remember that changes in the mean of Y are calculated using expectations.

$$
\begin{aligned}
&\mathbb{E}(Y|X_1 = x_1 + 0.5, X_2 = x_2) - \mathbb{E}(Y|X_1 = x_1, X_2 = x_2) \\
&= 0.5 + b_1(x_1 + 0.5) - [0.5 + b_1 x_1] \\
&= 0.5 * b_1 = 0.5
\end{aligned}
$$

There is no change in the mean of Y when X_2 changes.

Simulation

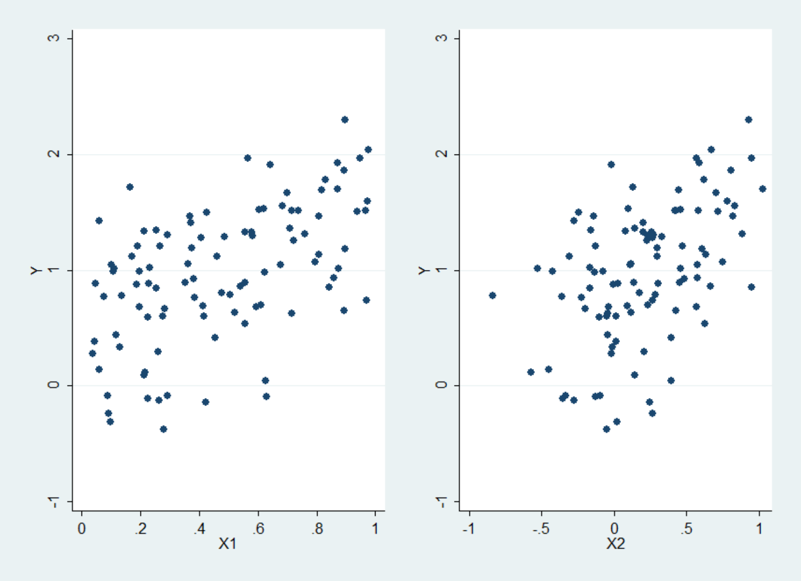
- Run the simulation with n=100.
- Try to graph (as a scatterplot) $Y \sim X1$ and $Y \sim X2$ *in the same window.*

  Your Stata code should look something like:

  ```
  scatter Y X1
  graph save "x1_scatter.gph"
  scatter Y X2
  graph save "x2_scatter.gph"
  gr combine "x1_scatter.gph" "x2_scatter.gph"
  ```

  Be sure you know what your working directory is before saving the graphs!

Graphical Representation of Simulation



Distribution

- X_1, X_2, and Y follow the same distributions as seen on slides 5 & 6.

- What is the model of regression? Run `regress Y X1 X2` to find the values of $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$.

Simulation 1, n=100

```
. regress Y X1 X2

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   20.39
       Model |  10.2006733     2  5.10033664           Prob > F      =  0.0000
    Residual |   24.268438    97  .250190083           R-squared     =  0.2959
-------------+------------------------------           Adj R-squared =  0.2814
       Total |  34.4691113    99  .348172842           Root MSE      =  .50019

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X1 |   .7669233   .2690555     2.85   0.005     .2329225    1.300924
          X2 |   .3210585    .197214     1.63   0.107    -.0703567    .7124738
       _cons |   .5381043   .1116183     4.82   0.000     .3165729    .7596357
------------------------------------------------------------------------------
```

Model of Regression

According to the previous slide, we have

$$\hat{b}_0 = 0.538 \qquad \hat{b}_1 = 0.767 \qquad \hat{b}_2 = 0.321$$

So our model of regression is:

$$\mathbb{E}[Y|X_1 = x_1, X_2 = x_2] = 0.538 + 0.767x_1 + 0.321x_2$$

Exercises

1. Calculate the predicted value at $X_1 = 0, X_2 = 1$ and provide a 95% confidence interval for your estimate.

2. What is the true change in the mean of Y when X_1 changes by 0.5? Provide a 95% confidence interval.

Solution #1

By hand,

$$\mathbb{E}[Y|X_1 = 0, X_2 = 1] = 0.538 + 0.767 * 0 + 0.321 * 1 = 0.859$$

Or use Stata's `lincom` command to do this!

```
. lincom _cons + X2

 ( 1)  X2 + _cons = 0

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .8591628   .2694559     3.19   0.002     .3243675    1.393958
------------------------------------------------------------------------------
```

Note: Calculating the confidence interval by hand is possible, but not covered in this course.

Solution #2

From slide 10, we know the true change in the mean of Y when X_1 changes by 0.5 is $0.5b_1$. We can therefore see

$$\mathbb{E}(Y|X_1 = x_1 + 0.5, X_2 = x_2) - \mathbb{E}(Y|X_1 = x_1, X_2 = x_2)$$
$$= 0.5\hat{b_1} = 0.5 * 0.767 = 0.3835$$

Using the `lincom` command to get our confidence interval:

```
. lincom 0.5*X1

 ( 1)  .5*X1 = 0

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3834616   .1345278     2.85   0.005     .1164613     .650462
------------------------------------------------------------------------------
```

Interpretation

Interpret to the best of your ability all the numbers in the row of the regression output corresponding to $X2$.

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
X2	.3210585	.197214	1.63	0.107	-.0703567	.7124738

Interpretation

- Coef $= \hat{b_2} = 0.321$: For a one unit increase in X_2, there is a 0.321 unit increase in the mean of Y holding X_1 constant.
- Std. Err. $= 0.197$: This is the estimated standard error of b_2, the coefficient of X_2 in the regression.
- $t = 1.63$: Test statistic, comes from $H_0 : b_2 = 0$. Is calculated by $t = \frac{\hat{b_2} - 0}{se(b_2)} = \frac{0.321 - 0}{0.197}$
- P$> |t| = 0.107$: Assuming the null is true, this is the probability of getting a t-statistic this extreme or more extreme.
- 95% Conf. Int. $= [-0.070, 0.712]$: If the experiment is repeated infinitely many times and 95% confidence intervals are calculated each time, 95% of those intervals would contain the true parameter, $b_2 = 0$.

Precision

What happens to the bias and standard error of $\hat{b}_1$ when we increase the sample size from $n = 100$ to $n = 500$?

Remember the following definition:

$$bias(\hat{b}_1) = \mathbb{E}[\hat{b}_1 - b_1]$$

Simulation 1, n=100

```
. regress Y X1 X2

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   20.39
       Model |  10.2006733     2  5.10033664           Prob > F      =  0.0000
    Residual |   24.268438    97  .250190083           R-squared     =  0.2959
-------------+------------------------------           Adj R-squared =  0.2814
       Total |  34.4691113    99  .348172842           Root MSE      =  .50019

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X1 |   .7669233   .2690555     2.85   0.005     .2329225    1.300924
          X2 |   .3210585    .197214     1.63   0.107    -.0703567    .7124738
       _cons |   .5381043   .1116183     4.82   0.000     .3165729    .7596357
------------------------------------------------------------------------------
```

Simulation 2, n=500

```
. regress Y X1 X2

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =   89.84
       Model |  44.4286117     2  22.2143059           Prob > F      =  0.0000
    Residual |  122.895974   497  .247275602           R-squared     =  0.2655
-------------+------------------------------           Adj R-squared =  0.2626
       Total |  167.324586   499  .335319812           Root MSE      =  .49727

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          X1 |    .917903   .1160903     7.91   0.000     .6898148    1.145991
          X2 |   .1205967   .0864875     1.39   0.164    -.0493294    .2905229
       _cons |    .498836   .0500369     9.97   0.000     .4005261    .5971459
------------------------------------------------------------------------------
```

Precision

Bias

- Sim 1: $bias(\hat{b}_1) = 0.767 - 1 = -0.233$
- Sim 2: $bias(\hat{b}_1) = 0.918 - 1 = -0.082$

Standard Error

- Sim 1: $se(\hat{b}_1) = 0.269$
- Sim 2: $se(\hat{b}_1) = 0.116$

Note: bias decreases (in absolute value) as n increases, and se decreases as n increases $\implies \hat{b}_1 \to b_1$ as $n \to \infty$.