

## 5 To Invest or not to Invest? That is the Question.

Before starting this lab, you should be familiar with these terms:

- response “y” (or dependent) and explanatory “x” (or independent) variables;
- slope and intercept in a linear regression equation;
- positive and negative correlations.

### In-class Activity

Suppose you are a Broadway producer. You would want your show to make as much money as possible, and one way of deciding whether or not to invest your time and money into a particular show would be to examine past shows to see how they did. We’ll examine a simple question: How does the size of the theater affect box office receipts? Does the size of the theater “predict” the average box office receipts?

To begin, download the data:

```
. use http://www.stat.ucla.edu/labs/datasets/broadway.dta
```

Make a scatterplot of the receipts against the capacity

```
. graph receipts capacity, xlab ylab
```

If you want to reveal the name of the show for any unusual observations, issue this command:

```
. graph receipts capacity, xlab ylab symbol([show])
```

*Question 1: Which show had the highest box office receipts? Which show appeared in the theater with the most seats?*

*Question 2: Describe the trend: How are receipts and capacity related? Would you say this is a linear relationship?*

The correlation coefficient helps us quantify the linear relationship of two variables.

*Question 3: Based on the scatterplot of receipts and capacity, what would you guess the correlation between these variables will be?*

Check your guess by typing

```
. corr receipts capacity
```

*Stata* allows you to calculate the correlations of multiple pairs of variables at once by including more variable names in the command.

*Question 4: Before you do, play a guessing game with the variables receipts, capacity, attendnc, and ratio. Which pair of variables will have the highest positive correlation? Which pairs, if any, will have a negative correlation?*

```
. corr receipts capacity attendnc ratio
```

We can further quantify the linear relationship with a least squares regression. (This is possible whether or not the relationship is really linear. If the relationship is not linear, then our least squares regression will be a very poor description - but we can still compute it.) Note that *Stata* gives us a lot more information than we are ready for right now. You'll return to this later in your studies. Type:

```
. regress receipts capacity
```

*Question 5: Look in the column headed by "Coef." (Coefficient) to find the estimated intercept and slope. Write the equation of the line here:*

*Question 6: Interpret the slope.*

To graph the line on top of the scatterplot, type:

```
. predict preceipts  
. graph receipts preceipts capacity, s(oi) c(.l) xlab ylab
```

The command “predict preceipts” calculates the predicted receipts for each value of capacity. It is based on the previous regression command. The predicted values all fall on the regression line. The last command “graph receipts preceipts capacity, s(oi) c(.l) xlab ylab” does the actual graphing. Be careful, the l for line is not the number one. The command plots receipts vs. capacity and then superimposes a plot of preceipts vs. capacity. The s(oi) sets the symbols for the plot, so that the first plot is done with circles (the o option) and the second plot is done with no symbols (the i for invisible option). The c(.l) option controls how the points are connected; in the first plot the points are not connected (the . option) but in the second plot the points are connected with a line (the l or line option).

*Question 7: What is the interpretation of this regression line? Is capacity a good predictor of average receipts? Explain.*

An examination of the residuals of a regression can help us discover errors. A residual is the difference between an observed value and the predicted value. In this context, it’s the difference between what a show actually made and what the linear regression says it was expected to make. Put slightly differently, it’s the difference between the actual average receipts for a particular show and the average receipts of all shows in theaters with the same capacity. To examine the residuals, type

```
. quietly regress receipts capacity  
. predict resids, residuals  
. graph resids capacity, symbol([show]) xlab ylab yline(0)
```

*Note: The quietly regress command performs the same regress as the regress command. The difference is that it does not output anything to the screen.*

*Question 8: Name two shows that made more than expected. Name two shows that made less. Which show did the best, in terms of beating expectations?*

*Question 9: As an investor, what have you learned? What type of theater would you be more inclined to invest in?*

*Question 10: This dataset contains many more variables, such as whether the production was a play or a musical, the average attendance, and what percentage of seats are filled during the average performance. What other analysis would you do to this data set before investing?*

## **Commands for To Invest or not to Invest**

Use the space next to each command to make notes on what that command does.

use filename

edit

label variable

sort x

graph y x

graph y x, xlabel ylabel

regress y x

```
quietly regress y x
```

```
predict newvar
```

```
predict newvar, residuals
```

## Assignment

You've probably been told, since the first day you complained about school, that education will help you get a better job. Certainly many jobs require a certain level of education, but does all that schoolwork pay off? It is difficult to claim solely on data of income and education level that the higher the level of education, the more money one makes. This is because there are many other factors that determine where we ultimately obtain a job, such as socio-economic status, or simply who you know. One way to try to bypass all these confounding factors and focus solely on the question "Will I make more money if I am more educated?" is to look at twins. Twins should have identical background factors. We expect them to have the same opportunities. Look at pairs of twins and analyze their differences in education level and differences in income.

Load this data set into *Stata*:

```
. use http://www.stat.ucla.edu/labs/datasets/twins.dat
```

Two variables of interest are *hrwageh* and *hrwagel*. These are the hourly wages of twins. You might want to focus your investigation on the difference in their hourly wage. To create this variable, type

```
. gen diffwage = hrwageh - hrwagel
```

Two other interesting variables are *educh*, the self-reported education level (in years) of the twin who reported earning *hrwageh*, and *educ1*, the self-reported education level of the twin who reported earning *hrwagel*. Create another variable to describe the difference in education level.

```
. gen diffeduc = educh - educ1
```

Are education and income related? Investigate this question with these data. Report on your findings. Your report should include answers to the following questions.

*Question 11: Do you expect the correlation between the twins' incomes to be positive or negative? High (close to positive or negative 1) or low (close to 0)? Check.*

*Question 12: Find the correlation matrix for these variables: hrwageh, hrwage, educl, educh, diffwage, diffeduc. What's the correlation between hrwageh and hrwage? Interpret. Why does the correlation between hrwageh and diffeduc have a different sign than the correlation between hrwageh and diffeduc?*

*Question 13: What's the typical difference in hourly wage between twins? Is it what you expected?*

*Question 14: Describe the distribution of the difference in hourly wage. Are there any unusual features?*

*Question 15: Make a scatterplot of the difference in income against the education level of either one of the twins. Interpret. Does it matter which twin's education level you chose?*

*Question 16: Perform a regression using the diffeduc variable to answer this question: is there evidence that the twin with more education makes more money?*

*Question 17: Examine the residuals from this last regression. For what types of twins did the model have the largest error (that is, the greatest difference between the predicted value and the observed value)? Do you see any possible outliers?*