Stat 13, Lab 11-12, Correlation and Regression Analysis

Part I: Before Class

Objective: This lab will give you practice exploring the relationship between two variables by using correlation, linear regression and graphical techniques.

Before starting this lab, you should…

1) 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.

Part II: In-class Activity

Suppose you were 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?

To begin, download the data:

use http://www.stat.ucla.edu/~dinov/courses_students.dir/STAT13_Fall01/STAT13_Fall01/data.dir/broadway

Some potentially useful Stata Commands:

use filename **->** loads a Stata-format file dataset from the Web. If filename is specified without the Stata extension, ".dta" is assumed.

edit -> opens Stata's spreadsheet.

label variable *varname* **"**label**" ->** enables you to attach an extra piece of information (up to 80 characters) about a variable.

sort x **->** arranges the observations of the current data in ascending order of the values of the variables. There is no limit to the number of the variables in the data. Missing values are interpreted as being larger than any other number and are thus placed last.

graph y x **->** for a scatterplot of y vs.x.

graph y x, **xlabel ylabel ->** for scatterplots with improved axes (numerical lists).

regress y x **->** estimates a model from the list of variables using least-squares regression.

quietly regress y x **->** suppresses the regression output for the duration of the command.
predict *newvar***->** calculates the predicted values of a variable in a linear regression for each observation. The new values are stored under the name *newvar*. This command must follow a **regress** command.
predict ***newvar*, residuals** -> calculates the residuals from a regression and places them in the variable named *newvar*. This command must follow a **regress** command.
graph y *newvar* x**, connect (.s) symbol (oi) ->** displays a linear regression graph between two variables with fitted values (*newvar*) connected by a line.
corr x y z w**->** displays the correlation or covariance matrix for two or more continuous variables, or if they are not specified, for all variables in the data. Observations are excluded from the calculation when values are missing.

Does the size of a theater "predict" the average box-office receipts?

Your TA might ask the class the following questions, so you should jot down your observations and thoughts for class discussion.

1. Make a scatterplot of the receipts against the capacity

```
graph receipts capacity
```

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

```
graph receipts capacity, symbol([show])
```

Which show had the highest box office receipts? Which show appeared in a theater with the most seats?

2. Describe the trend: how are receipts and capacity related? Would you say this is a linear relationship?

3. We can quantify the linear relationship with a least squares regression. (This works whether or not the relationship is really linear. If it 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. But you'll return to this later in your studies. Type:

```
regress receipt capacity
```

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

Interpret the slope.

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

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

The first command **quietly regress receipts capacity** performs a regression that computes the slope and intercept of the regression line. The next command, **predict preceipts**, calculates the predicted receipts for each value of capacity. The predicted values all fall on the regression line. The last command, **graph receipts preceipts capacity, s(oi) c(.l)** does the actual graphing. 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 option).

6. What is the interpretation of this regression line? Is capacity a good predictor of average receipts? Explain.

7. 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 should have made. 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])
```

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

Are we better at predicting for smaller theaters than large? Are there any outliers?

8. 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
```

You can check the correlations between all pairs of variables by typing

```
corr receipts capacity attendnc ratio
```

Before you do, play this guessing game: which pair will have the highest positive correlation? Which pairs, if any, will have a negative correlation?

Part III: Take-home Problem

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 level of education, but does all that schoolwork pay off? Load this data set into Stata:

use http://www.stat.ucla.edu/projects/datasets/twins

This is data from a study of twins. You can learn details about the data, including how and why they were collected, at http://www.stat.ucla.edu/projects/datasets/twins-explanation.html.

Two variables of interest are *hrwageh* and *hrwagel*. These are the hourly wage of twin "1" and "2". (The twins were arbitrarily numbered.) 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 *educl*: the self-reported education level of the twin who reported earning *hrwagel*. For explanations of the other variables, see http://www.stat.ucla.edu/projects/datasets/twins-explanation.html

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

1) 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.

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

3) What's the typical difference in hourly wage between twins? Is it what you expected?

4) Describe the distribution of the difference in hourly wage. Are there any unusual features?

5) 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?

6) Perform a regression of difference in income against a twin's education level.

What does the estimated slope say about the effect of education on income? (You might want to superimpose the regression line on the graph for a clearer picture.)

Does your conclusion depend on which twin's education you used to predict the difference in income?

7) Create a new variable, diffeduc, that is the difference in education levels between twins. Perform a regression and use it to answer this question: is there evidence that the twin with more education makes more money?

8) 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?

This Lab was originally created by Prof. R. Gould A and Prof. V. Lew