Statistics 104 - FALL, 2004 — Assignment 2

Due: Wednesday October 13, 2003

Readings (Moore and McCabe)

- Chapter 2

Against All Odds videotape

The relevant tapes for this week are numbers 7 (Models for Growth), 8 (Relationships), 9 (Correlation), 10 (Multidimensional Data), and 11 (The Question of Causation). This is completely optional supplementary viewing!

Written assignment (Moore and McCabe)

- **MM:** 2.4, 2.8, 2.30, 2.42, 2.46, 2.66, 2.78, 2.86.

Suggested problems - not to be handed in

- **MM:** 2.52, 2.72, 2.87, 2.128

Computer Problems

You will analyze the Companies dataset which can be found on the Stat 104 Assignments page. If you have problems with the Stata version of the file (Companies.dta), a tab delimitted version (Companies.txt) is also available on the same page.

This dataset holds information about 77 companies selected from the Forbes 500 list for 1986. This is a 1/10 systematic sample from the alphabetical list of companies. The Forbes 500 includes all companies in the top 500 of any of the criteria, and thus has almost 800 comapanies in the list. Companies of often interested in how to increase sales. We will examine this dataset to see if we can find variables that can help us understand the sales data better. Note that many of the original variables in the dataset have been transformed to the log scale. The reason for this is that many of the variables are skewed – a common occurance with financial data – which suggests that much of the data are better analyzed after taking logarithms. (Note you might want to verify this skewnes for yourself. Take a look at the variables `assets, sales, market_value` for example.)

Include computer printouts as appropriate in your answers. Edit the printouts so they do not contain too many extraneous lines.

1. Plot histograms and boxplots for the distributions of `LogAssets`, `LogSales`, and `LogMarket`. Describe the features of these three distributions, e.g. skewness/symmetry, single/multiple peaks,

etc. Use the command `summarize` to get means and medians of these three variables. Comment on the relationships between the means and medians and the appearance of the histograms and box plots for these three variables. Did transforming `assets`, `sales`, and `market_value` to `LogAssets`, `LogSales`, `LogMarket` solve the skewness problem?

2. Figure out which variables (out of `LogAssets`, `LogSales`, and `LogMarket`) it makes sense to include in a correlation analysis. Use the `correlate var1 var2 ... vark` command to calculate the correlation matrix for all of the variables `var1` through `vark`. You can find a list of variables in the variables window on the Mac or the PCs. You can look at all the pairs plots (known as a scatterplot matrix) using the command `graph matrix var1 var2 ... vark` or with the menu selection `Graphics > Scatterplot matrix`. What do the three correlation values suggest about the relationships between these three variables?

3. Plot scatterplots for `LogSales` vs `LogAssets`, `LogSales` vs `LogMarket`, and `LogAssets` vs `LogMarket` and say in ordinary words what they tell you about the relationships between these three variables. Use the command `scatter var1 var2` where `var1` and `var2` are two variables you want to correlate.

4. From the correlation analysis and scatterplot, do `LogSales` and `LogAssets` appear to be related?

5. Run the regression of `LogSales` on `LogAssets`. The command for regression is `regress vary varx`, where vary is the name of the response variable and varx is the name of the predictor variable. Write down or print out the regression equation. Useful follow-up commands to `regress` are as follows: `predict resid, residuals` and `predict yhat, xb`. Type these now. For your information, `yhat` contains predicted values of `LogSales` from the regression on `LogAssets` and `resid` contains the difference between the original variable `LogSales` and its predicted value `yhat`. The fitted and residuals can also be gotten from via the menus `Statistics > General Post Estimation > Obtain predictions, residuals, etc after estimation`.

6. What does this equation tell you about the relationship between the variables, i.e., on the average, when `LogAssets` increases, what happens to `LogSales`? Explain why this seems either expected or surprising to you.

7. Plot the scatterplot with the fitted line on it as follows:

   ```
   twoway (lfit LogSales LogAssets) (scatter LogSales LogAssets)
   ```

   This graph should have `LogAssets` on the horizontal- or x-axis and both `LogSales` and `yhat` on the vertical- or y-axis. Print out this plot. This plot can be gotten via the menus `Graphics > Easy Graphs > Regression Fit`.

8. Using the regression equation, calculate the prediction for `LogSales` for three companies with `LogAssets` of 2.5, 3.5, and 4.8 respectively. You can do this by hand with a calculator. Which of these predictions would you least want to rely upon, and why? (Consult the scatter plot!)

9. Plot the residuals against the predictor variable `LogAssets` as follows:

```
scatter resid LogAssets
```

   Describe any interesting features you see in the residual plot (i.e. features of the data, particularly of the relationship between these two variables, that were not summarized by the regression).

10. Replot the residuals against the preditor variable `LogAssets`, but this time include the information about which sector each company is from as follows:

```
twoway (scatter resid LogAssets, mlabel(sector))
```

    . If you are to do this plot via the menus, you need to go to `Graphics > Twoway graph`. What does this plot suggest about only using `LogAssets` to describe the `LogSales` data.

11. Plot the residuals against the explanatory variable `LogMarket`. What does this plot suggest about only using `LogAssets` to describe the `LogSales` data.

12. While the two variables `LogProfit` and `LogCash` were included in the data set, they probably shouldn't have been. What is it about the original variables `profit` and `cash_flow` that makes `LogProfit` and `LogCash` poor variables to study `LogSales`.

Challenge Problem: Least Squares Coefficients

"Challenge problems" are extra credit problems that go beyond the basic requirements of this course. Doing these problems is completely voluntary! They may involve a bit more mathematics or computer skills than the regular homework, but otherwise they are self-contained, i.e., they do not assume a high level of preparation in probability or statistics.

Challenge problems will be counted for a modest number of extra credit points, roughly equivalent to one or two regular problems. More important, they should give you a glimpse of more advanced concepts and techniques in probability and statistics.

The regression formulas you learned in class give the coefficients for the "least squares regression line." This is defined as the line that is close to the data points in the sense that the sum of the squared residuals is as small as possible. An alternative model that is sometimes used is regression through the origin. In this problem, you will derive formulas for the coefficients of this line, against based on the least squares criterion.

Suppose that you have a data set consisting of x and y values, $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$. The fitted line will have the equation $y = b_1 x$.

1. Write down an equation for the sum of the squared residuals, in terms of b_1 and the data values. (The equation will involve a summation.)

2. Calculate the derivative of the sum with respect to b_1.

3. Minimize the sum of squared residuals: Set the derivatives equal to 0 and solve for b_1 to find the least squares coefficient. You may use any other method for the minimization if you prefer. For example, this minimization can be done only with algebra and some knowledge of quadratic functions.