Eryk Wdowiak
economics@doviak.net

Introductory Statistics

Homework #8

This is an assignment in basic linear regression. The goal of this assignment is to give you an opportunity to make practical use of the theory that we discussed in class. In particular, I hope this assignment helps you explore the relationships between different variables.

First, download the US state data from the course website. As you may recall, the file contains household and family median income, a measure of income inequality (the Gini coefficient), marginal tax rates, health insurance coverage rates, racial and ethnic composition and the percentage of the population that was enslaved in 1820 and 1860.

Please perform the following tasks, print your results and come to class prepared to discuss them.

♦ ♦ ♦

1. To perform a proper regression analysis, we need to work with the log of income. Why take the log? When working with the "raw" income series, we are examining dollar changes in income. When working with the log of income, we are examining percentage changes in income. So go to Gretl's "Add" menu and take the log of the income variables.

2. Create scatterplots of:
   a. the log of median income and the Gini coefficient
   b. the log of median income and percentage without health insurance
   c. the log of median income and percentage slave in 1860
   d. the Gini coefficient and percentage without health insurance
   e. the Gini coefficient and percentage slave in 1860
   f. percentage without health insurance and percentage slave in 1860

3. Create a correlation matrix to compute the coefficient of correlation between the log of median income, the Gini coefficient, percentage without health insurance and percentage slave in 1860.

4. How do the correlation coefficients summarize the information that you saw in the scatterplots? What do the scatterplots reveal that the correlation coefficients do not?

5. Using ordinary least squares, regress the log of median income on the Gini coefficient.
   a. Interpret the regression coefficients.
   b. Is the regression coefficient statistically significant from zero at the 10 percent significance level? Is it statistically significant from zero at the 5 percent significance level?
   c. Given the values of R-squared and the F-statistic, how well do you think the model explains the variance in log of median income?

6. Now regress the log of median income on the Gini coefficient, percentage without health insurance and percentage slave in 1860.
   a. Interpret the regression coefficients.
   b. Which of the regression coefficients are statistically significant from zero at the 10 percent significance level? Which are statistically significant from zero at the 5 percent significance level?
   c. Given the values of R-squared and the F-statistic, how well do you think the model explains the variance in log of median income?

7. Finally, let's perform an F-test on the contribution that percentage without health insurance makes to explaining the variance in log of median income.
   a. First, make note of the R-squared value from the regression that you just ran. This is the "unrestricted R-squared."
   b. Now regress the log of median income on the Gini coefficient and percentage slave in 1860. (Note that we're excluding percentage without health insurance).
   c. Now make note of the new R-squared value. This is the "restricted R-squared."
   d. How many restrictions have we imposed? This is the numerator degrees of freedom in the F-test.
   e. How many degrees of freedom did we have in the original model? This is the denominator degrees of freedom in the F-test.
   f. Calculate the F-statistic.
   g. Using Gretl's p-value finder, calculate the p-value of the null hypothesis that percentage without health insurance does not contribute to explaining the variance in log of median income.
   h. Given the p-value that you just computed, should we exclude percentage without health insurance from the regression model? Why or why not?