17.874, Spring 2004
Final Exam

1. In voting rights cases involving racial districting, the plaintiffs must show that the behavior of voters is such that the blacks and whites in a state vote as "blocs" and that they are "polarized." That is, whites vote together most of the time, blacks vote together most of the time, and blacks and whites vote against each other. Ecological regression is a common technique for measuring the behavioral parameters in a population when survey data are unavailable. The basic model holds that there are two variables, say vote for the black candidate Y and race X, for which we wish to measure the probability that $Y = 1$ given $X = 1$ and the probability that $Y = 1$ given $X = 0$.

The probability that an arbitrarily chosen person in the population votes for the black candidate in a two candidate race involving a black and a white can be written as follows:

$$Pr(Y = 1) = Pr(Y = 1|X = 1)Pr(X = 1) + Pr(Y = 1|X = 0)Pr(X = 0)$$

Let $\gamma_0 = Pr(Y = 1|X = 0)$ and $\gamma_1 = Pr(Y = 1|X = 1) - Pr(Y = 1|X = 0)$. Note that $Pr(X = 0) = 1 - Pr(X = 1)$. Then, we may write this as

$$Pr(Y = 1) = \gamma_0 + \gamma_1 Pr(X = 1) \tag{1}$$

In a given election we may observe the fraction black ($Pr(X = 1)$) and the fraction voting for the black candidate $Pr(Y = 1)$.

Ecological regression allows us to estimate the parameters, γ_0 and γ_1, as follows. The election district (say a state) is divided into many smaller jurisdictions (such as counties or precincts). We will use the index j to denote these smaller jurisdictions. For each j we can observe the fraction black (and thus the fraction white, assuming no other racial groups) and the fraction voting for the black candidate in stead of the white candidate. We denote the fraction black as x_j and the fraction voting for the black candidate as y_j. To estimate the parameters of interest, regress y on x across the jurisdictions:

$$y_j = \beta_0 + \beta_1 x_j + u_j$$

a. Write down the formula for least squares estimates of β_0 and β_1 in this model? What is the variance-covariance matrix for the estimates, (b_0, b_1)?

b. Assume that γ_0 and γ_1 vary across jurisdictions. Let γ_0 and γ_1 be the average values of the γ's in the entire area (e.g., state). Express the error term, u_j, in terms of γ_{0j}, γ_{1j}, and x_j. Write the parameters γ_0 and γ_1 in terms of β_0 and β_1.

c. If γ_0 and γ_1 are constant across jurisdictions, what form does the error in the regression equation take? [Note: Equation 1 is an identity.]

d. Consultants in court cases and scholars of race in elections frequently rely on ecological regressions to measure the behavior of the typical individual in each of the groups. Under what conditions on u_j is the least squares estimator of β_0 and β_1 unbiased?

e. The R^2 in ecological regressions involving racial voting are often quite high, sometimes nearly 1. In a court case, an expert testified that a high R^2 meant that the ecological regression assumptions were not violated (invoking part (c) above). Is this statement correct? Can you think of a simple example that has high R^2 but violates the assumptions?

f. Many plausible factors or behaviors could lead to violations of the necessary to ensure unbiasedness. In the race and racial voting example, γ might vary with another variable, such as income. One way to test for possible biases is to find auxiliary information that addresses these possible factors or behaviors. What other data might you bring to bear to validate the ecological regression assumption? How would you use the data to address the question of whether the regression assumption is violated?

2. I have emailed a data set to you containing the daily rate of return (return) and the market's daily rate of return (mktr) for two firms, ATT and IBM, over 2001 and 2002.

a. Estimate the β from the capital asset pricing model for each of these firms.

b. Look at the actual data for these firms (using appropriate plots). Are there outliers or other unusual cases that might affect your estimates? What are the dates of these unusual events? Estimate the results excluding these cases?

c. Conduct a single pooled regression including the returns for both firms. Are the intercepts and slopes of the two firms statistically different?

d. Present the means, standard deviations, and regression results for each of the stocks. Interpret these numbers in terms of the rates of return and risk associated with each of these stocks. Which would you have preferred to own over this two year period?