17.874
Multivariate Statistics
Spring 2004
Problem Set 2

1. Using Simulations to learn about Omitted Variables.

In STATA create a simulated data set as follows.

```
set obs 250
gen x1 = invnorm(uniform())
gen x2 = x1 + invnorm(uniform())
gen y = 1 + 3*x1 - 3*x2 + 2*invnorm(uniform())
```

(a) What are the means and variances of the variables? What are the covariances and correlations among the variables?

(b) Regress y on x1 and x2 and y on x1 and y on x2. Compare your estimates.

(c) Generate a new dataset in which the coefficient on x2 is +3. Redo parts (a) and (b).

(d) Generate a new dataset in which x1 and x2 are uncorrelated. I.e., gen x2 = invnorm(uniform()). Redo parts (a) and (b)

(e) Compare your estimated coefficients in parts (b), (c), and (d). How do the estimates differ? What do you learn about the nature of bias created by omitted variables?

(f) Compare the estimated variance of e and the standard errors. How did omitting variables affect the precision and efficiency of the estimates in each of these cases?

2. Matrix Multiplication

Consider two matrices:

$$A = \begin{pmatrix} 1 & 3 & 3 \\ 2 & 4 & 1 \end{pmatrix}$$

$$B = \begin{pmatrix} 2 & 4 \\ 1 & 5 \\ 6 & 2 \end{pmatrix}.$$

Compute AB, A'B', and BA.

3. Matrix Multiplication with Statistical Applications

Consider the matrices:

$$X = \begin{pmatrix} 1 & 3 \\ 1 & 1 \\ 1 & 2 \\ 1 & 5 \\ 1 & 4 \\ 1 & 0 \\ 1 & 0 \\ 1 & 7 \\ 1 & 4 \end{pmatrix}$$

$$y = \begin{pmatrix} -2 \\ -5 \\ -3 \\ 4 \\ 0 \\ 2 \\ -3 \\ 6 \\ -1 \end{pmatrix}.$$

(a) Compute X'X and X'y.

(b) Consider the equations $(X'X)b = X'y$, where b is a column vector with values b_0 and b_1. Find values of b_0 and b_1 that satisfy this equation.

(c) Create a small dataset in the STATA editor in which the values of second column of X are Var1 and the values of the vector y are Var2. Regress Var2 on Var1 (i.e., reg Var2 Var1). What are the slope and intercept?

4. Using the matrix commands in STATA.

(a) Create an X matrix for the first simulation above. Be sure to create a column of 1's as the first column of X. Produce X'X and X'y. What do the values of the matrices mean in statistical terms?

(b) Subtract the mean values from each of the variables to create mean-deviated forms of the variables. Create a new X matrix of the mean deviated values: X^*. Produce X^*'X^* and X^*'y^*. What do these values mean in statistical terms?