Statistical Techniques II

EXST7015

Regression with Matrix Algebra

Matrix Algebra

- **We will not be doing our regressions with matrix algebra, except that the computer does employ matrices. In fact, there is really no other way to do the basic calculations.**
- **You will be responsible for knowing about matrices only to the extent that PROC REG or PROC GLM produces information. This is primarily the initial and final matrices.**

Matrix Algebra *(continued)*

- **So, what is a matrix?**
  - **A matrix is a rectangular arrangement of numbers, usually represented by an upper case letter (A, B, C, D, etc.)**
  - **A =** $\begin{bmatrix} 1 & 3 \\ 7 & 9 \end{bmatrix}$

D = $\begin{bmatrix} 4 & 2 & 4 \\ 1 & 6 & 0 \\ 3 & 0 & 5 \\ 2 & 3 & 0 \end{bmatrix}$

Matrix Algebra *(continued)*

- **The dimensions of a matrix are given by the number of rows and columns in the matrix (i.e. the dimensions are r by c). For the matrices above,**
  - **A is 2 by 2**
  - **D is 4 by 3**

Matrix Algebra *(continued)*

- **For a simple linear regression the matrices of initial interest would be the data matrices, a Y matrix of values of the dependent variable and an X matrix of values of the independent variable.**
- **The X matrix also has a column of ones added to fit the intercept.**

Matrix Algebra *(continued)*

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ 1 & X_4 \\ 1 & X_5 \\ 1 & X_6 \\ 1 & X_7 \end{bmatrix}$$

Matrix Algebra *(continued)*

- As with our algebraic calculations we need some intermediate values; sums, sums of squares and cross-products. These are obtained by calculating
- First, a transpose matrix for both X and Y. This is simply the matrix turned on its side so the rows of the original matrix become the columns of the transpose.
- These are denoted X' and Y'.

Matrix Algebra *(continued)*

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \end{bmatrix}$$

$$Y' = \begin{bmatrix} Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 \end{bmatrix}$$

Matrix Algebra *(continued)*

- **We now calculate 3 matrices, X'X, Y'Y and X'Y. This requires matrix multiplication.**

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ 1 & X_4 \\ 1 & X_5 \\ 1 & X_6 \\ 1 & X_7 \end{bmatrix}$$

Matrix Algebra *(continued)*

- **Calculate X'X, Y'Y and X'Y *(continued)*.**

$$\mathbf{Y'Y} = \begin{bmatrix} Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix}$$

Matrix Algebra *(continued)*

- **Calculate X'X, Y'Y and X'Y *(continued)*.**

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix}$$

Matrix Algebra *(continued)*

- **The results of these 3 calculations are;**
- **X'X =** $\begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix}$ **X'Y =** $\begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix}$
- **Y'Y =** $\begin{bmatrix} \sum_{i=1}^{n} Y_i^2 \end{bmatrix}$

Matrix Algebra *(continued)*

- **Notice that the contents of these 3 matrices are the same as the values we used for the algebraic solution.**
- **Normal equations - when the equations needed to solve a simple linear regression are derived, the result is two equations with two unknowns that must be resolved. These are called the normal equations.**
-

Matrix Algebra *(continued)*

- **The normal equations are**
  - $b_0 n + b_1 \Sigma X_i = \Sigma Y_i$
  - $b_0 \Sigma X_i + b_1 \Sigma X^2_i = \Sigma Y_i X_i$
-
- **If you solve these algebraically, you get the two equations we use to solve for b_0 and b_1.**

Matrix Algebra *(continued)*

- **When expressed as matrices this factors out to**

$$\begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i X_i \end{bmatrix}$$

Matrix Algebra *(continued)*

- In simple matrix notation, a B matrix (vector) times the X'X equals the X'Y.
  - $(X'X)B=X'Y$
- As with the algebraic equations we need to solve for B (i.e. b_0 and b_1). If we do this with algebra, we get the usual equations.
- Solving the matrix equations we get,
  - $B=(X'X)^{-1}X'Y$

Matrix Algebra *(continued)*

- **This equation is the matrix algebra solution for a simple linear regression.**
  - $B=(X'X)^{-1}X'Y$
  - **Note that there is not such thing as matrix "division".**
- **As with the algebraic values, if we multiply the B values (B matrix) by the X values we get the predicted values.**
  - $XB=X(X'X)^{-1}X'Y$ = Yhat vector

Matrix Algebra *(continued)*

- **What do we need to know about these matrix calculations?**
  - **We need to know that the solution to the problem using matrix algebra involves the same values as for the simple linear regression.**
  - **We need to know that the $(X'X)^{-1}$ is a key component to this solution.**
  - **We need to know that the predicted values require the matrix segment $X(X'X)^{-1}X'$ times the Y vector (MAIN DIAGONAL).**

Matrix Algebra *(continued)*

- **Why? We can get the matrices from SAS, but we want to understand what we have.**
  - **$(X'X)^{-1}$ is a key component not only of the solution for the regression coefficients, but also for the variance-covariance matrix.**
  - **The $X(X'X)^{-1}X'$ matrix main diagonal is a diagnostic that we will use (hat diag).**

Matrix Algebra *(continued)*

- **But the most important reason for using matrices is that the solution for simple linear and multiple regression are the same. Basically, matrix algebra is the ONLY way to solve multiple regressions.**
- **So, what do we get from SAS?**
- **If the options XPX and I are placed on the model statement, we can get the X'X matrix and the $(X'X)^{-1}$ matrix.**

Matrix Algebra *(continued)*

- **For the simple linear regression that we saw for the tree weights and diameters, these options produce the following output.**

```
Model Cross-products X'X X'Y Y'Y

X'X                INTERCEP                DBH             WEIGHT
INTERCEP                 47              289.2              17359
DBH                   289.2            1981.98             142968.3
WEIGHT                17359             142968.3            13537551


X'X Inverse, Parameter Estimates, and SSE

                   INTERCEP                DBH             WEIGHT
INTERCEP      0.2082694963      -0.030389579      -729.3963003
DBH           -0.030389579       0.004938832       178.56371409
WEIGHT        -729.3963003       178.56371409      670190.7322
```

Matrix Algebra *(continued)*

- **The first two rows and columns of numbers contain the X'X matrix, which has the values for n, ΣX_i and ΣX^2_i**
-

```
Model Cross-products X'X X'Y Y'Y
X'X          INTERCEP          DBH
INTERCEP           47         289.2
DBH             289.2       1981.98
```

Matrix Algebra *(continued)*

- **The last column has X'Y (values for ΣY_i and $\Sigma X_i Y_i$) and the last value is Y'Y (ΣY^2_i).**
  -
  - **Model Cross-products X'X X'Y Y'Y**
  - **X'X WEIGHT**
  - **INTERCEP 17359**
  - **DBH 142968.3**
  - **WEIGHT 13537551**

Matrix Algebra *(continued)*

- **In the X'X inverse matrix section, the first two rows and columns of numbers contain the $(X'X)^{-1}$ matrix and the value in the third row and third column is the SSE. The other values are b_0 and b_1.**

X'X Inverse, Parameter Estimates, and SSE

	INTERCEP	DBH	WEIGHT
INTERCEP	0.2082694963	-0.030389579	-729.3963003
DBH	-0.030389579	0.004938832	178.56371409
WEIGHT	-729.3963003	178.56371409	670190.7322

Matrix Algebra *(continued)*

- **You will be responsible only for knowing where the 6 intermediate values are for simple linear regression, and where to find the $(X'X)^{-1}$ matrix.**

Model Cross-products X'X X'Y Y'Y

X'X	INTERCEP	DBH	WEIGHT
INTERCEP	47	289.2	17359
DBH	289.2	1981.98	142968.3
WEIGHT	17359	142968.3	13537551

X'X Inverse, Parameter Estimates, and SSE

	INTERCEP	DBH	WEIGHT
INTERCEP	0.2082694963	-0.030389579	-729.3963003
DBH	-0.030389579	0.004938832	178.56371409
WEIGHT	-729.3963003	178.56371409	670190.7322

Multiple Regression

- **The only difference between simple linear regression and multiple regression is the fact that multiple regression has several independent variables (X_i variables).**
- **There for the matrix X'X will be larger. For a simple linear regression, X'X is 2x2. For a 3 factor multiple regression (X_1, X_2, X_3 and an intercept) the X'X matrix will be 4x4.**

Matrix Algebra *(continued)*

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{31} \\ 1 & X_{12} & X_{22} & X_{32} \\ 1 & X_{13} & X_{23} & X_{33} \\ 1 & X_{14} & X_{24} & X_{34} \\ 1 & X_{15} & X_{25} & X_{35} \\ 1 & X_{16} & X_{26} & X_{36} \\ 1 & X_{17} & X_{27} & X_{37} \end{bmatrix}$$

Multiple Regression *(continued)*

- **The values contained in the matrix include all of the sums, sums of squares and cross-products for all of the X_i variables (plus n in the upper left hand corner).**
-

Multiple Regression *(continued)*

$X'X =$	n	ΣX_{1i}	ΣX_{2i}	ΣX_{3i}
	ΣX_{1i}	ΣX^2_{1i}	$\Sigma X_{1i}X_{2i}$	$\Sigma X_{1i}X_{3i}$
	ΣX_{2i}	$\Sigma X_{1i}X_{2i}$	ΣX^2_{2i}	$\Sigma X_{2i}X_{3i}$
	ΣX_{3i}	$\Sigma X_{1i}X_{3i}$	$\Sigma X_{2i}X_{3i}$	ΣX^2_{3i}

Multiple Regression *(continued)*

$Y'Y =$	$\sum Y_i$
	$\sum X_{1i} Y_i$
	$\sum X_{2i} Y_i$
	$\sum X_{3i} Y_i$

Matrix Algebra *(continued)*

- For the multiple regression the solution is still given by;
  - $B=(X'X)^{-1}X'Y$
- The predicted values are still given by;
  - $XB=X(X'X)^{-1}X'Y$ = Yhat vector
- The residuals are given by;
  - $Yhat=Y_i - X(X'X)^{-1}X'Y$
- The $X(X'X)^{-1}X'$ matrix main diagonal is a diagnostic that we will use (hat diag).

Matrix Algebra *(continued)*

- **So basically everything works the same in multiple regression as in simple linear regression if we use matrix algebra.**
- **One last piece of the puzzle. We will need some estimates of variances and covariances. As usual these will all involve the MSE (Mean Square Error).**
- **We need all of the variances and covariances for the regression coefficients**

Matrix Algebra *(continued)*

- **These variances and covariances are obtained by multiplying the $(X'X)^{-1}$ matrix by the MSE. The resulting matrix contains all of the variances and covariances for the regression coefficients.**
- **The variances of regression coefficients are on the main diagonal. The square root of these values gives the standard error used for confidence intervals and testing of the b_i values.**

Multiple Regression *(continued)*

- **The $(X'X)^{-1}$ matrix can be obtained from SAS.**
- **When multiplied by the MSE value this gives the Variance-Covariance matrix. This can also be obtained from SAS.**
-

Multiple Regression *(continued)*

- **A note on the assumption of independence.**
  - **We assume that the e_i values are independent of each other.**
  - **We assume that the e_i are independent of the Yhat values ($b_0+b_1X_i$).**
  - **But WE DO NOT ASSUME THAT THE VARIOUS REGRESSION COEFFICIENTS ARE INDEPENDENT OF EACH OTHER.**
  - **Calculations do not assume zero covariances, they employ the Variance-Covariance matrix.**

Summary

- **Matrix solutions to regression begin with the matrices X'X, X'Y and Y'Y. The values in these matrices are the sums, sums of squares and cross-products, just as with simple linear regression.**
- **The normal equations are solved just as for SLR. The solution is $B=(X'X)^{-1}X'Y$.**
- **Using matrix algebra we can obtain predicted values and residuals, diagnostics, variances and covariances and all of the other values needed for testing and interpretation of the multiple regression.**