Statistical Techniques II

EXST7015

Regression with Matrix Algebra

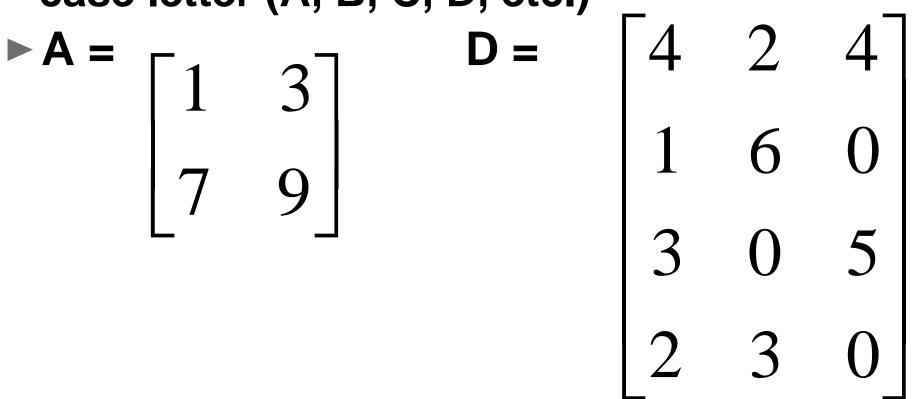
06a_Matrix SLR 1

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.

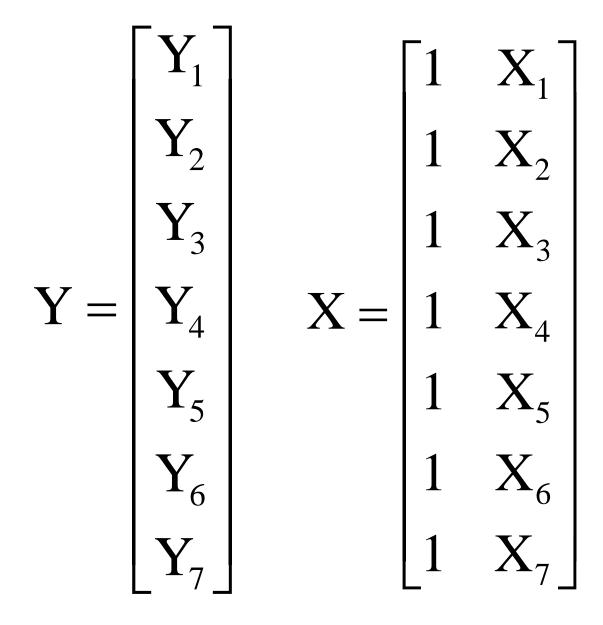
So, what is a matrix?

A matrix is a rectangular arrangement of numbers, usually represented by an upper case letter (A, B, C, D, etc.)



- 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

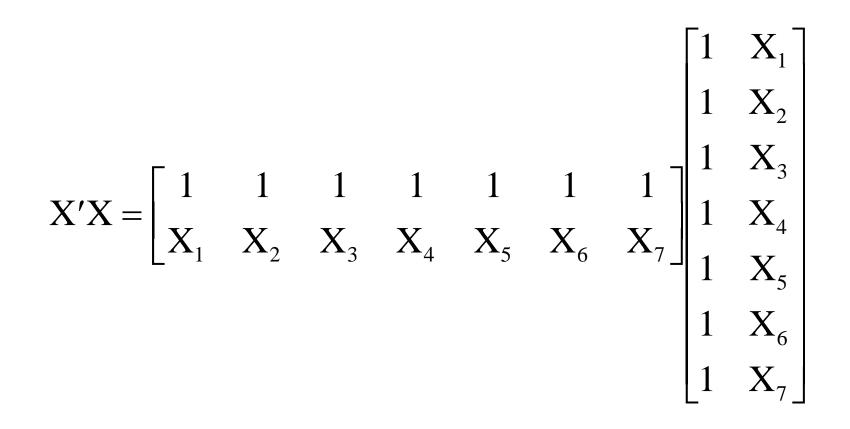
- For a simple linear regression the matrices of initial interest would be the data matrices, a Y matrix of values of the dependent variable and an X matrix of values of the independent variable.
- The X matrix also has a column of ones added to fit the intercept.



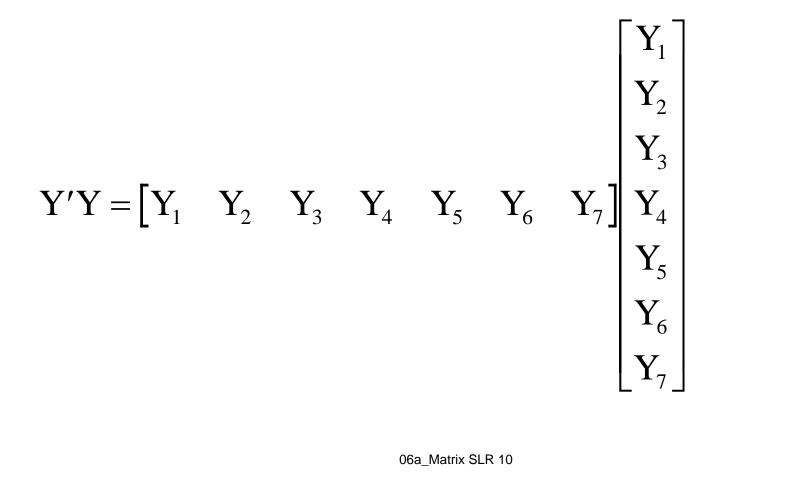
- As with our algebraic calculations we need some intermediate values; sums, sums of squares and cross-products. These are obtained by calculating
- First, a transpose matrix for both X and Y. This is simply the matrix turned on its side so the rows of the original matrix become the columns of the transpose.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

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

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



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



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

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_3 \\ \mathbf{Y}_4 \\ \mathbf{Y}_5 \\ \mathbf{Y}_6 \\ \mathbf{Y}_7 \end{bmatrix}$$

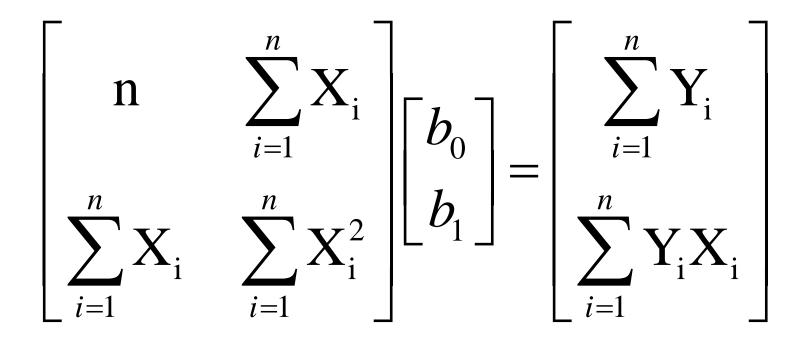
The results of these 3 calculations are; $\begin{bmatrix} n & \sum_{i=1}^{n} X_{i} \\ \sum_{i=1}^{n} X_{i} & \sum_{i=1}^{n} X_{i}^{2} \\ i = 1 \end{bmatrix} X'Y =$ X'X = $\sum_{i=1}^{n} \mathbf{Y}_{i}$ $\sum_{n=1}^{n} \mathbf{X} \cdot \mathbf{Y}_{i}$ • **Y'Y** = $\left[\sum_{i=1}^{n} Y_{i}^{2}\right]$

- 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.

■ The normal equations are ► $\mathbf{b}_0 \mathbf{n} + \mathbf{b}_1 \Sigma \mathbf{X}_i = \Sigma \mathbf{Y}_i$ ► $\mathbf{b}_0 \Sigma \mathbf{X}_i + \mathbf{b}_1 \Sigma \mathbf{X}_i^2 = \Sigma \mathbf{Y}_i \mathbf{X}_i$

If you solve these algebraically, you get the two equations we use to solve for b₀ and b₁.

When expressed as matrices this factors out to



- 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. b₀ and b₁). If we do this with algebra, we get the usual equations.
 Solving the matrix equations we get,
 - ► B=(X'X)⁻¹X'Y

- This equation is the matrix algebra solution for a simple linear regression.
 B=(X'X)⁻¹X'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)⁻¹X'Y = Yhat vector

- 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)⁻¹X' times the Y vector (MAIN DIAGONAL).

- 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)⁻¹X' matrix main diagonal is a diagnostic that we will use (hat diag).

- 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.

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
►XIX Inverse	Parameter Estima	tog and CCF	
A V THAT BE'	Falametel Estima	Les, and sse	
	INTERCEP	DBH	WEIGHT
		-	WEIGHT -729.3963003
•	INTERCEP	DBH	

- The first two rows and columns of numbers contain the X'X matrix, which has the values for n, ΣX_i and ΣX²_i
 - Model Cross-products X'X X'Y Y'Y
 X'X INTERCEP DBH
 INTERCEP 47 289.2
 - DBH 289.2 1981.98

- The last column has X'Y (values for ΣY_i and ΣX_iY_i) and the last value is Y'Y (ΣY²_i).
 - Model Cross-products X'X X'Y Y'Y
 - ►X'X WEIGHT
 - ► INTERCEP 17359
 - ►DBH 142968.3
 - ► WEIGHT

13537551

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 b₀ and b₁.

▶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

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.

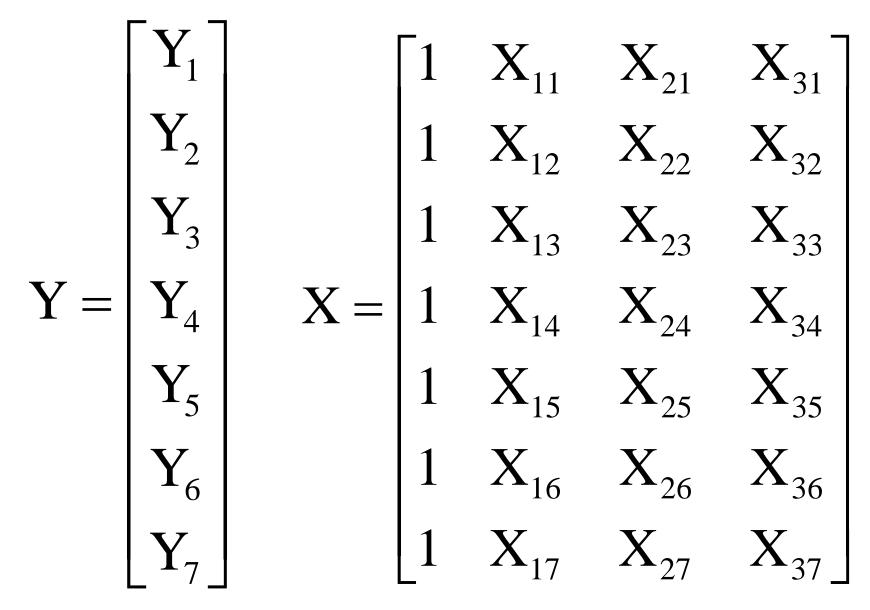
►Model Cross-products X'X X'Y Y'Y

►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 Estimate	es, and SSE	
		5511	

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

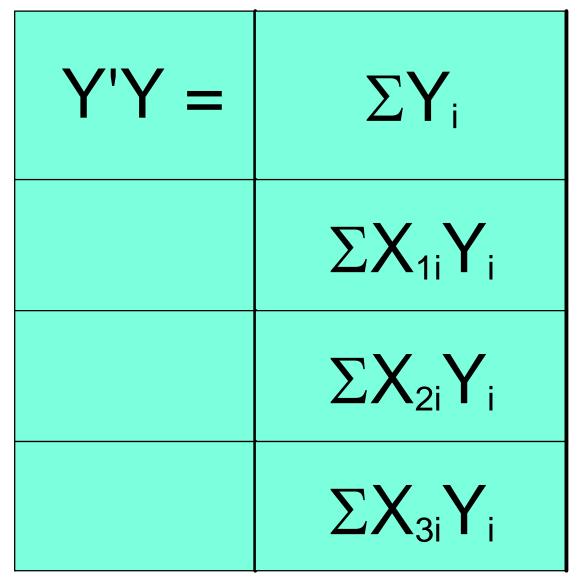
Multiple Regression

- 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 (X₁, X₂, X₃ and an intercept) the X'X matrix will be 4x4.



 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).

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



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

- 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

- 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.

- 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.

- 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 (b₀+b₁X_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)⁻¹X'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.