The University of Georgia

Matrix Applications in Regression and Statistics

October 4, 2007

ERSH 8320

Today's Lecture

- Taking what we have covered to this point and expressing these values as a function of matrices.
  - General Linear Model for Regression/ANOVA.
  - All subsequent quantities that produce statistical estimates.
- Making sense of matrices.

Regression Analysis with Matrices

- Recall the "multiple regression" equation (for the i^{th} observation, prediction of Y_i by k variables X_{ik}):

$$Y_i = a + b_1 X_{i1} + b_2 X_2 + \ldots + b_k X_k + e$$

- The equation above can be expressed more compactly by a set of matrices:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

  - **y** is of size $(N \times 1)$.
  - **X** is of size $(N \times (k+1))$.
  - **b** is of size $(k \times 1)$.
  - **e** is of size $(N \times 1)$.

Regression with Matrices

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ 1 & X_{31} & X_{32} & \dots & X_{3k} \\ 1 & X_{41} & X_{42} & \dots & X_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \dots & X_{Nk} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_N \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X} \quad \mathbf{b} + \mathbf{e}$$

$$(N \times 1) \qquad (N \times (k+1)) \qquad (k \times 1) \qquad (N \times 1)$$

Regression with Matrices

- Working the matrix multiplication and addition for a single case gives:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ 1 & X_{31} & X_{32} & \dots & X_{3k} \\ 1 & X_{41} & X_{42} & \dots & X_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \dots & X_{Nk} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_N \end{bmatrix}$$

$$Y_1$$

Regression with Matrices

- Working the matrix multiplication and addition for a single case gives:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ 1 & X_{31} & X_{32} & \dots & X_{3k} \\ 1 & X_{41} & X_{42} & \dots & X_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \dots & X_{Nk} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_N \end{bmatrix}$$

$$Y_1 = a$$

Regression with Matrices

- Working the matrix multiplication and addition for a single case gives:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ 1 & X_{31} & X_{32} & \ldots & X_{3k} \\ 1 & X_{41} & X_{42} & \ldots & X_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \ldots & X_{Nk} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_N \end{bmatrix}$$

$$Y_1 = a + b_1 X_{11}$$

Regression with Matrices

- Working the matrix multiplication and addition for a single case gives:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \color{red}{X_{12}} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ 1 & X_{31} & X_{32} & \ldots & X_{3k} \\ 1 & X_{41} & X_{42} & \ldots & X_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \ldots & X_{Nk} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ \color{red}{b_2} \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_N \end{bmatrix}$$

$$Y_1 = a + b_1 X_{11} + \color{red}{b_2 X_{12}}$$

Regression with Matrices

- Working the matrix multiplication and addition for a single case gives:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ 1 & X_{31} & X_{32} & \ldots & X_{3k} \\ 1 & X_{41} & X_{42} & \ldots & X_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \ldots & X_{Nk} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_N \end{bmatrix}$$

$$Y_1 = a + b_1 X_{11} + b_2 X_{12} + \ldots$$

Regression with Matrices

- Working the matrix multiplication and addition for a single case gives:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ 1 & X_{31} & X_{32} & \dots & X_{3k} \\ 1 & X_{41} & X_{42} & \dots & X_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \dots & X_{Nk} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_N \end{bmatrix}$$

$$Y_1 = a + b_1 X_{11} + b_2 X_{12} + \ldots + b_k X_{1k}$$

Regression with Matrices

- Working the matrix multiplication and addition for a single case gives:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ 1 & X_{31} & X_{32} & \dots & X_{3k} \\ 1 & X_{41} & X_{42} & \dots & X_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \dots & X_{Nk} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_N \end{bmatrix}$$

$$Y_1 = a + b_1 X_{11} + b_2 X_{12} + \ldots + b_k X_{1k} + e_1$$

Regression with Matrices

- Note that most everything is really straightforward in terms of matrix algebra.
- The matrix of predictors, **X**, has the first column containing all ones.
    - This represents the intercept parameter a.
    - This is also an introduction to setting columns of the **X** matrix to represent design and or group controls (as in ANOVA).

Distribution of Errors

- Recall from previous classes that we often place distributional assumptions on our error terms, allowing for the development of tractable hypothesis tests.

- With matrices, the distributional assumptions are no different, except for things are approached in a multivariate fashion:

$$\mathbf{e} \sim N_N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N)$$

- Having a multivariate normal distribution with uncorrelated variables (from $\mathbf{I}_N$) is identical to saying:

$$e_i \sim N(0, \sigma_e^2)$$

for all i observations.

Regression Estimation with Matrices

- Regression estimates are typically found via least squares (called L_2 estimates).
- In least squares regression, the estimates are found by minimizing:

$$\sum_{i=1}^{N} e^2 = \sum_{i=1}^{N}(Y_i - Y_i')^2 = \sum_{i=1}^{N}(Y_i - a + b_1 X_{i1} + \ldots + b_k X_{ik})^2$$

- As you could guess, we could accomplish all of this via matrices.
- Equivalently:

$$\sum_{i=1}^{N} e^2 = \sum_{i=1}^{N}(y_i - \mathbf{x}_i'\mathbf{b})^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{e}'\mathbf{e}$$

Regression Estimation with Matrices

- Thankfully, there are people to figure out the equation for $\mathbf{b}$ that minimizes $\mathbf{e}'\mathbf{e}$.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\underset{((k+1)\times 1)}{\mathbf{b}} = (\underset{((k+1)\times N)}{\mathbf{X}'} \quad \underset{(N\times(k+1))}{\mathbf{X}})^{-1} \quad \underset{((k+1)\times N)}{\mathbf{X}'} \quad \underset{(N\times 1)}{\mathbf{y}}$$

- This equation is what I have been talking about for quite some time, the General Linear Model.

- For many types of data, in many differing analyses, this equation will provide estimates:
  - Multiple Regression
  - ANOVA
  - Analysis of Covariance (ANCOVA).
  - Multiple Regression with Curvilinear relationships in X.

Numerical Example

- For the past week we have strayed away from examples, but now that we have been introduced to the GLM, an example is warranted.
- From Chapter 5 of Pedhazur, p. 98, an example data set is given.
- The data set attempts to predict Reading Achievement (Y) as a function of Verbal Aptitude, (X_1), and Achievement Motivation, (X_2).
- To demonstrate the GLM estimates with matrices, I will show the results of the analysis from SPSS, and then replicate these results in MATLAB.
- This data set will be the focal point of subsequent demonstrations in this lecture.

SPSS Results

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.471	1.194		-.394	.698
	Verbal Aptitude (X1)	.705	.175	.602	4.021	.001
	Achievement Motivation (X2)	.592	.244	.363	2.428	.027

a. Dependent Variable: Reading Achievement (Y)

Other Regression Terms with Matrices

- Clearly, getting estimates for **b** isn't the only focus of this class.
- Most everything accomplished previously can be obtain via matrices:
  - Mean Vectors.
  - Variance-Covariance Matrices.
  - Correlation Matrices.
  - Sums of Squares.
  - Linear Combinations - Principal Components.

A Vector of Ones

- Helpful in many applications is a simple vector of ones:

$$\mathbf{1}_N = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

- Note that this vector is not the identity matrix (where ones are on the diagonal).
- You can probably see the use for such a vector by the following equation:

$$\sum_{i=1}^{N} X_i = \mathbf{x}'\mathbf{1}$$

The Mean Vector

- To compute a vector of means for a set of variables $X_1, \ldots, X_k$:

$$\bar{\mathbf{X}}_{(k \times 1)} = \mathbf{X}'_{k \times N} \mathbf{1}_{(N \times 1)} (\mathbf{1}'_{(1 \times N)} \mathbf{1}_{(N \times 1)})^{-1}$$

- Don't be dismayed by $(\mathbf{1}'_{(1 \times N)} \mathbf{1}_{(N \times 1)})^{-1}$, this is simply:

$$\frac{1}{N}$$

Sums of Squares and Cross Products

- The SSCP matrix contains:
  - On the diagonal, Sums of Squares for each variable Z_k:

$$\sum_{i=1}^{N}(X_{ik}-\bar{X}_{\cdot k})^2$$

  - On the off-diagonal the Sum of Cross Products for each pair of variables, X_k and $X_{k'}$:

$$\sum_{i=1}^{N}(X_{ik}-\bar{X}_{\cdot k})(X_{ik'}-\bar{X}_{\cdot k'})$$

  - The matrix expression for the SSCP matrix is:

$$\mathbf{SSCP}=(\mathbf{X}-\mathbf{1}\bar{\mathbf{X}}')'(\mathbf{X}-\mathbf{1}\bar{\mathbf{X}}')$$

Covariance Matrices

- Once the SSCP is obtained, the covariance matrix is easily obtained:
  - On the diagonal, the variance for each variable X_k:

$$\frac{\sum_{i=1}^{N}(X_{ik} - \bar{X}_{.k})^2}{N}$$

  - On the off-diagonal the covariance for each pair of variables, X_k and $X_{k'}$:

$$\frac{\sum_{i=1}^{N}(X_{ik} - \bar{X}_{.k})(X_{ik'} - \bar{X}_{.k'})}{N}$$

  - The matrix expression for the covariance matrix is:

$$\textbf{COV} = \frac{1}{N}(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})$$

Correlation Matrices

- Once the covariance matrix is obtained, the correlation matrix is easily obtained.
- Let D be a diagonal matrix consisting of the standard deviation of all variables (the square root of the diagonal elements of the covariance matrix).
- The matrix expression for the correlation matrix is:

$$\textbf{CORR} = \textbf{D}^{-1}\textbf{COV D}^{-1}$$

Regression Sums of Squares

- The Sum of Squares for the Regression can be obtained by matrix calculations:

$$SS_{reg} = \mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^{N} Y)^2}{N} = \mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{1}{N}(\mathbf{y}'\mathbf{1})^2$$

Residual Sums of Squares

- The Residual Sum of Squares is also obtained by matrix calculations:

$$SS_{res} = \sum_{i=1}^{N} {e_i}^2 = \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}$$

Squared Correlation

- Recall that the Squared Multiple Correlation Coefficient, or R^2, is found from:

$$R^2 = \frac{SS_{reg}}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

- This is also obtainable by matrix calculations:

$$R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{1}{N}(\mathbf{y}'\mathbf{1})^2}{\mathbf{y}'\mathbf{y} - \frac{1}{N}(\mathbf{y}'\mathbf{1})^2}$$

Variance of Regression Estimators

- The covariance matrix of the regression parameters contains useful information regarding the standard errors of the estimates (found along the diagonal).

- To find the covariance matrix of the estimates, the Residual Sum of Squares is needed (dividing this term by the residual degrees of freedom gives the error variance, or σ_e^2):

$$var(\mathbf{b}) = \sigma_e^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{SS_{res}}{N-k-1}(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{N-k-1}\mathbf{e}'\mathbf{e}(\mathbf{X}'\mathbf{X})^{-1}$$

Final Thought

- Matrices are prevalent in many statistical applications, but do not fear them.

- You will encounter matrix notation from time to time, but that should not be cause for concern.
- Leave the heavy matrix lifting to the experts: let a stat or math computer package do advanced matrix calculations.

Next Time

- Chapter 7: Partial/Semipartial Correlation.