Math 218 Mathematical Statistics

Prof. D. Joyce, Clark University

1 Apr 2008

Monday. Presentations from exercises beginning on page 390: 13, 16, 17, 18, 19, 20, 21, 22, 23.

Last time. . Regression diagnostics.

Multiple linear regression. This is a very common method used to understand numerical relations among numerical factors. For this method, we assume that we have k independent variables $x_1, \ldots, x_k$ that we can set, then they probabilistically determine an outcome Y. Furthermore, we assume that Y is linearly dependent on the factors according to

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where ϵ is a normal error. This is just like we had for simple linear regression except k doesn't have to be 1.

In an experiment, we have n observations, n typically being much more than k. For the i^{th} observation we set the independent variables to the values

$$x_{i1}, x_{i2} \ldots, x_{ik}$$

and measure a value y_i for the random variable Y_i. Thus, the model can be described by the equations

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

for $i = 1, 2, \ldots, n$, where the errors ϵ_i are independent normal variables each with mean 0 and the same unknown variance σ^2.

Altogether this model for multiple linear regression has $k+2$ unknown parameters: $\beta_0, \beta_1, \ldots, \beta_k$, and σ^2.

When k was equal to 1, we found the least squares line $y = \hat\beta_0 + \hat\beta_1 x$. It was a line in the plane $\mathbf{R}^2$. Now with $k \geq 1$, we'll have a least squares hyperplane

$$y = \hat\beta_0 + \hat\beta_1 x_1 + \hat\beta_2 x_2 + \cdots + \hat\beta_k x_k$$

in $\mathbf{R}^{k+1}$. The way to find the estimators $\hat\beta_0, \hat\beta_1, \ldots$, and $\hat\beta_k$ is exactly the same, namely, take the partial derivatives of the squared error

$$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}))^2$$

with respect to the $k+1$ variables $\beta_0, \ldots, \beta_k$, set them all to 0, and solve to find the critical point. There will only be one critical point except in exceptional situations, like when $n < k$, and it will give the minimum least squared error Q. To find that value, $k+1$ linear equations need to be solved simultaneously for $k+1$ unknowns, so methods of linear algebra are needed, but not very advanced methods since it's just solving a simultaneous system of linear equations.

When that system is solved we have fitted values

$$\hat y_i = \hat\beta_0 + \hat\beta_1 x_{i1} + \hat\beta_2 x_{i2} + \cdots + \hat\beta_k x_{ik}$$

for $i = 1, \ldots, n$ that should be close to the actual values y_i. The differences are the residuals

$$e_i = y_i - \hat y_i.$$

The error sum of squares, SSE, is defined as before, namely as

$$\text{SSE} = \sum e_i^2$$

and it is used along with the total sum of squares,

$$\text{SST} = \sum (y_i - \overline{y})^2$$

and the regression sum of squares

$$\text{SSR} = \sum (\hat{y}_i - \overline{y})^2$$

exactly as in the case when $k = 1$, so that

$$\text{SST} = \text{SSE} + \text{SSR}$$

and

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

The only difference is that we don't have a correlation coefficient r. Instead, we define the positive square root of r^2 to be r and we call it the *multiple correlation coefficient.*

We'll look at example 11.1 in the text next.

Vector and matrix notation. There's a lot more of linear algebra that we could use than just the notation for vectors and matrices, but that alone helps considerably to simplify the notation and concepts of multiple linear regression. We'll also need a some of the algebra of matrices, namely, addition, subtraction, multiplications, and inverses of square matrices. A complete study of multiple linear regression requires more, but we'll be satisfied with this much for our survey of multiple linear regression.

Matrices and vectors. A matrix is a rectangular array of numbers. For example,

$$\begin{bmatrix} 5 & 2 \\ 3 & -1 \\ 1 & 3 \end{bmatrix}$$

This matrix has 3 rows and 2 columns and is called a 3×2 matrix.

Some matrices have only one row or column. A matrix with only one column is called a *column matrix* or a *column vector*, while one with only one row is called a *row matrix* or *row vector*. The way things are set up here, all our vectors are columns.

We'll use boldface for vectors and matrices. We can collect several variables together into one vector. For instance, if we have n data-values $y_1, y_2, \ldots, y_n$, we can put them in a column vector

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

We'll use column vectors for other groupings of n values, too, such as the random variables $Y_1, Y_2, \ldots, Y_n$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

and the error random variables $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

and so forth.

When symbols are used for the elements (that is, entries) of a matrix, they are often doubly indexed (that is, subscripted), and the indices indicate where the entry is located. For instance, a_{34} indicates the element in the 3rd row and 4th column. Note that the first index gives the row number and the second index gives the column number. When a generic row is needed, usually i is used, and when a generic column is needed, usually j is used. So a_{ij} is the element in the ith row and jth column.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

The multiple regression model has doubly indexed variables x_{ij} where i refers to the i^{th} observation (varying from 1 to n) and j refers to the j^{th} random variable (varying from 1 to k), so we'll have use for matrices.

Addition and subtraction of matrices. If two matrices are the same shape, that is, the same number of rows and the same number of columns, then you can add them together to produce another matrix by adding the corresponding elements together. Likewise you can subtract one matrix from another if they have the same shape. For example

$$\begin{bmatrix} 5 & 2 \\ 3 & -1 \\ 1 & 3 \end{bmatrix} + \begin{bmatrix} 4 & 0 \\ 0 & 3 \\ 6 & -5 \end{bmatrix} - \begin{bmatrix} 2 & 1 \\ -5 & 4 \\ 1 & -2 \end{bmatrix} = \begin{bmatrix} 7 & 1 \\ 8 & -2 \\ 6 & 0 \end{bmatrix}$$

The transpose of a matrix. A useful simple operation on matrices is *transposition*. It's where you exchange rows and columns in a matrix. For instance, if A is the matrix

$$A = \begin{bmatrix} 4 & 5 & 6 \\ 3 & -1 & 0 \end{bmatrix},$$

then the transpose of A is the matrix

$$A^T = \begin{bmatrix} 4 & 3 \\ 5 & -1 \\ 6 & 0 \end{bmatrix}.$$

The transpose of a matrix A is denoted in our text as A', but you'll find various different notations are in use such as A^T, A^t, A^*, and so forth. I'll use A' in these notes as the text does.

Note that if A is an $m \times n$ matrix, then its transpose A' is an $n \times m$ matrix. Also, note that the transpose of the transpose is the original matrix, that is $(A')' = A$. Furthermore, transposition turns row vectors into column vectors, and vice versa.

Matrix multiplication. Matrix multiplication is not at all like addition or subtraction. It actually corresponds to composition of transformations between various dimensional spaces, but that goes beyond what we have time to talk about here.

Take, for instance, the following two 3 by 3 matrices.

$$A = \begin{bmatrix} 4 & 5 & 6 \\ 3 & -1 & 0 \\ 2 & 0 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 4 & 5 \\ -2 & -3 & 0 \end{bmatrix}$$

Think of A as being made of three row vectors and B as being made of three column vectors.

$$A = \left[\begin{array}{ccc} 4 & 5 & 6 \\ \hline 3 & -1 & 0 \\ \hline 2 & 0 & -2 \end{array}\right], \quad B = \left[\begin{array}{c|c|c} 2 & 1 & 1 \\ 0 & 4 & 5 \\ -2 & -3 & 0 \end{array}\right]$$

The easiest way to see how it all works is to put the matrix A on the left the product matrix AB, and the matrix B above the product matrix AB as follows.

$$\begin{array}{cc} & \left[\begin{array}{c|c|c} 2 & 1 & 1 \\ 0 & 4 & 5 \\ -2 & -3 & 0 \end{array}\right] \\ \left[\begin{array}{ccc} 4 & 5 & 6 \\ \hline 3 & -1 & 0 \\ \hline 2 & 0 & -2 \end{array}\right] & \left[\begin{array}{c|c|c} -4 & 6 & 29 \\ \hline 6 & -1 & -2 \\ \hline 8 & 8 & 2 \end{array}\right] \end{array}$$

For each entry in the product matrix AB, look at the elements in the row of A to its left and the elements in the column of B above it. Then multiply the corresponding elements of the row to the left and the column above, and finally add the the products. Take, for example, the element in the second row and third column of the product. It turns out to be -2 since

$$3 \cdot 1 + (-1) \cdot 5 + 0 \cdot 0 = -2.$$

Note that matrix multiplication is seldom commutative, that is, AB usually doesn't equal BA. But it's always associative, that is, $(AB)C$ always equals $A(BC)$.

The size of matrices needed for multiplication. In order for there to be a product matrix AB, the length of a row in A has to be the same as the length of a column in B. If A is an m by n matrix, that means it has n entries in each row. So B will have to have n entries in each column, in other words, B has to have n rows. Then B is an n by k matrix (where k is the number of columns that B has, whatever that is). Thus, you can multiply an m by n matrix with an n by k matrix. The result will be an m by k matrix.

Inverse matrices. Matrix multiplication is unusual, and there is no operation of division. But for some square matrices, inversion sometimes does work.

First we need a matrix that acts as a multiplicative identity, and that's easy to find. The $n \times n$ identity matrix has 1's down the diagonal and 0's elsewhere. This I acts like an identity matrix since

$$AI = A \quad \text{and} \quad IB = B$$

for any matrix A that has n columns and any matrix B that has n rows.

We say that two square $n \times n$ matrices A and B are inverses of each other if

$$AB = BA = I,$$

and in that case we say that B is an inverse of A and that A is an inverse of B. A matrix A can have at most one inverse, and it's denoted A^{-1}. A fair amount of time in a linear algebra course is devoted to algorithms to find inverse matrices. We won't do that here, but we ought to have at least one example. Suppose that A is the matrix

$$A = \begin{bmatrix} 3 & -2 & 4 \\ 1 & 0 & 2 \\ 0 & 1 & 0 \end{bmatrix}$$

Then the inverse of A is

$$A^{-1} = \begin{bmatrix} 1 & -2 & 2 \\ 0 & 0 & 1 \\ -1/2 & 3/2 & -1 \end{bmatrix}$$

The multiple regression model in matrix notation. The model can be described by the n equations

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

for $i = 1, 2, \ldots, n$. We can put these in one matrix equation

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

that is, the equation

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

Note how by cleverly including a column of 1's in the matrix $\mathbf{X}$ we get to treat β_0 just like the rest of the β_i's.

As mentioned before, in order to minimize the squared error

$$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}))^2$$

you take its partial derivatives with respect to the $k+1$ variables $\beta_0, \ldots, \beta_k$, set them all to 0, and solve to find the critical point. The $k+1$ linear equations that need to be solved simultaneously can be written as the matrix equation

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}.$$

This will have a solution if the matrix $\mathbf{X}'\mathbf{X}$ has an inverse, and that solution is our estimator for β the least squares estimator:

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