Multivariate Multiple Linear Regression

November 5, 2017

Multivariate Regression

Multivariate multiple linear regressions uses a linear model to relate two or more response variables to several covariates.

Example

Suppose the Air Force wishes to predict several measures of pilot efficiency. These response variables could be regressed against independent variables such as math and science skills, reaction time, eyesight acuity, and manual dexterity.

Each of the Y's are predicted by each of the X's.

model

The n observed values of the response vectors can be listed as rows:

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}$$

So each row contains the values of the p dependent variables measured on a subject and each column contains all n observations on one of the p variables. Each columns then corresponds to the response vector in a multiple regression model.

model

The design matrix is the same as for multiple regression.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix}$$

model

Each of the columns in the Y matrix will depend on the x's in their own way. So our parameters will be in a matrix, one column for each column in the Y matrix. The multivariate model then is

$$\mathbf{Y} = \mathbf{XB} + \Xi$$

where

- $\mathbf{Y}$ is an $n \times p$ matrix of response variables.
- $\mathbf{X}$ is an $n \times (q+1)$ matrix of covariates.
- $\mathbf{B}$ is a $(q+1) \times p$ matrix of parameters
- Ξ is an $n \times p$ matrix of errors

Example with $p = 2$ and $q = 3$

$$\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} & \epsilon_{12} \\ \epsilon_{21} & \epsilon_{22} \\ \vdots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} \end{pmatrix}$$

Model for first column of **Y**

$$\begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{21} \\ \beta_{31} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{n1} \end{pmatrix}$$

A similar model can be written for the second Y measure.

assumptions

Additional assumptions that lead to good parameter estimates are:

1. $E(\mathbf{Y}) = \mathbf{XB}$ or $E(\Xi) = \mathbf{0}$. This assumption is that the linear model is correct and no additional x's are needed to predict the y's.
2. $Cov(Y_i) = \Sigma$ for $i = 1, 2, \ldots, n$ where Y_i is the ith row of $\mathbf{Y}$ So each observed response vector has the same covariance.
3. $Cov(Y_i, Y_j) =)$ sample responses are independent of each other

Parameter Estimates

By analogy with the univariate case, we estimate $\mathbf{B}$ with

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$\hat{\mathbf{B}}$ is called the least squares estimator for $\mathbf{B}$ because it "minimizes"

$$\mathbf{E} = \hat{\Xi}'\hat{\Xi} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$$

example

A planned experiment involving a chemical reaction involved explanatory variables

- X_1 = temperature
- X_2 = concentration
- X_3 = time

The yield (dependent) variables are

- Y_1 = percentage of unchanged starting material
- Y_2 = percentage converted to the desired product
- Y_3 = percentage of unwanted by-product.

example

The least squares estimator for the regression of (Y_1, Y_2, Y_3) on (X_1, X_2, X_3) is

$$\hat{\mathbf{B}} = \begin{pmatrix} 332.11 & -26.04 & -164.08 \\ -1.55 & 0.40 & 0.91 \\ -1.42 & 0.29 & 0.90 \\ -2.24 & 1.03 & 1.15 \end{pmatrix}$$

The estimated regression lines are

- $Y_1 = 332.11 - 1.55X_1 - 1.42X_2 - 2.24X_3$
- $Y_2 = -26.04 + 0.4X_1 + 0.29X_2 + 1.03X_3$
- $Y_3 = -164.08 + 0.91X_1 + 0.9X_2 + 1.15X_3$