Multivariate Regression (Chapter 10)

This week we'll cover multivariate regression and maybe a bit of canonical correlation. Today we'll mostly review univariate multivariate regression.

With multivariate regression, there are typically multiple dependent variables as well as multiple independent or explanatory variables. A special case of this is when the explanatory variables are categorical and the dependent variables are continuous (particularly multivariate normal), in which case we have MANOVA. For multivariate regression, we allow the explanatory variables to be continuous. This approach generalizes multiple regression much as MANOVA generalizes ANOVA.

Typically in regression, we think of the y variables as random and the x variables as fixed. For multivariate regression, we'll consider x variables as either fixed or random. We'll start with them being treated as fixed.

Multivariate regression

First, we'll review multiple (univariate) regression with fixed x variables. For this model, we have

$$
\begin{aligned}
y_1 &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{1j} + \varepsilon_1 \\
y_2 &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{2j} + \varepsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{nj} + \varepsilon_n
\end{aligned}
$$

Multivariate regression

The standard assumptions for multiple regression are

$$E(\varepsilon_i) = 0$$

$$Var(\varepsilon_i) = \sigma^2$$

$$cov(\varepsilon_i, \varepsilon_j) = 0$$

Equivalently, you can write

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

$$Cov(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$$

Multivariate regression

Under the assumption that the xs are fixed, we have

$$E(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{1j}$$

$$Var(y_i) = \sigma^2$$

$$Cov(y_i, y_j) = Cov(\varepsilon_i, \varepsilon_j) = 0$$

Equivalently,

$$E(\mathbf{y}) = \mathbf{X}\beta$$

$$Cov(\mathbf{y}) = \sigma^2 \mathbf{I}$$

Multivariate regression

The regression model using matrix notation is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

When I was an undergrad, my Calc III professor suggested that we get tattoos of

$$\mathbf{f} = \mathbf{ma},$$

but if you are a statistics,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

would be better....

Multivariate regression

Multivariate regression

Written out, the matrix form looks like this

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Multivariate regression

The $\mathbf{X}$ is called the design matrix and recall that it has a column of 1s which is necessary for the β_0 term.

For estimation and hypothesis testing (for which variances are needed), you need $n > q + 1$

Multivariate regression

The least squares approach for estimating $\boldsymbol{\beta}$ is to minimize the following

$$\begin{aligned} SSE &= \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} - \cdots - \widehat{\beta}_q x_{iq})^2 \end{aligned}$$

This problem can be solved with calculus, or with less effort, using matrix algebra:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

If you set $\widehat{\mathbf{y}}$ equal to its expectation and to solve for $\boldsymbol{\beta}$, then get

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

Multivariate regression

The previous solution for estimating β is the least squares solution regardless of the distribution of the error term. If the error terms are independent and identically distributed (i.i.d.) as $N(0, \sigma^2)$, then the solution is also the maximum likelihood solution.

Multivariate regression

An unbiased estimator for σ^2 is

$$s^2 = \frac{SSE}{n-q-1} = \frac{1}{n-q-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y}$$

Multivariate regression

Another way of writing the model is to center the xs, so you have

$$\overline{x}_1 = \sum_{i=1}^{n} x_{i1}, \quad \cdots, \overline{x}_q = \sum_{i=1}^{n} x_{iq}$$

Then we write (next slide)

$$y_i = \alpha + \beta_1(x_{i1} - \overline{x}_1) + \beta_2(x_{i2} - \overline{x}_2) + \cdots + \beta_q(x_{iq} - \overline{x}_q) + \varepsilon_i, \qquad (10.11)$$

where

$$\alpha = \beta_0 + \beta_1\overline{x}_1 + \beta_2\overline{x}_2 + \cdots + \beta_q\overline{x}_q. \qquad (10.12)$$

To estimate

$$\boldsymbol{\beta}_1 = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix},$$

we use the centered x's in the matrix

$$\mathbf{X}_c = \begin{pmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \cdots & x_{1q} - \overline{x}_q \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & \cdots & x_{2q} - \overline{x}_q \\ \vdots & \vdots & & \vdots \\ x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 & \cdots & x_{nq} - \overline{x}_q \end{pmatrix} = \begin{pmatrix} (\mathbf{x}_1 - \overline{\mathbf{x}})' \\ (\mathbf{x}_2 - \overline{\mathbf{x}})' \\ \vdots \\ (\mathbf{x}_n - \overline{\mathbf{x}})' \end{pmatrix}, \qquad (10.13)$$

where $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{iq})$ and $\overline{\mathbf{x}}' = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_q)$. Then by analogy to (10.5), the least squares estimate of $\boldsymbol{\beta}_1$ is

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}. \qquad (10.14)$$

Multivariate regression

This approach is equivalent, and corresponds to the model

$$y_i = \alpha + \sum_{j=1}^{q} \beta_j (x_{ij} - \overline{x}_j)$$

so the xs are centered and the intercept term is changed and becomes

$$\widehat{\alpha} = \overline{y}$$

The term $\widehat{\boldsymbol{\beta}}_1$ is $(q-1) \times 1$ rather than $q \times 1$, so we have

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}_1)'$$

where

$$\widehat{\beta}_0 = \widehat{\alpha} - \sum_{j=1}^{q} \widehat{\beta}_j \overline{x}_j$$

Multivariate regression

To do hypothesis tests, the total sums of squares for y is partitioned into SSE and SSR. This is done as follows

$$\begin{aligned}
\mathbf{y}'\mathbf{y} &= \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}\mathbf{y} + \widehat{\boldsymbol{\beta}}'\mathbf{X}\mathbf{y} \\
&= SSE + \widehat{\boldsymbol{\beta}}'\mathbf{X}\mathbf{y} \\
&= SSE + \widehat{\boldsymbol{\beta}}'\mathbf{X}\mathbf{y} + n\overline{y}^2 - n\overline{y}^2 \\
&= SSE + SSR - n\overline{y}^2 \\
&\Rightarrow \mathbf{y}'\mathbf{y} + n\overline{y}^2 = SSE + SSR
\end{aligned}$$

Multivariate regression

A test for the nonintercept coefficients

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$$

is

$$F = \frac{SSR/q}{SSE/(n-q-1)}$$

which has an $F_{q,n-q-1}$ distribution under the null (and assuming normally distributed y values).

Multivariate regression

You can also test whether a subset of coefficients is 0. To do this, let $\boldsymbol{\beta}_d$ be the subset of interest so that the null is

$$H_0 : \boldsymbol{\beta}_d = \mathbf{0}$$

Have the betas arranged so that

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_r \\ \boldsymbol{\beta}_d \end{pmatrix}$$

The reduced model is

$$\mathbf{y} = \mathbf{X}_r \boldsymbol{\beta}_r + \varepsilon_r$$

The idea is that the reduced model has only the variables with nonzero coefficients.

Multivariate regression

The term $\boldsymbol{\beta}_r$ is estimated by

$$\widehat{\boldsymbol{\beta}_r} = (\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r\mathbf{y}$$

The reduced model is tested against the full model using

$$\begin{aligned} F &= \frac{(\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \widehat{\boldsymbol{\beta}}_r'\mathbf{X}_r'\mathbf{y})/h}{(\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y})/(n-q-1)} \\ &= \frac{SSR_f - SSR_r)/h}{SSE_f/(n-q-1)} = \frac{MSR}{MSE} \end{aligned}$$

where the subscript f refers to the full model and h is the number of parameters in $\boldsymbol{\beta}_d$. The test statistic is compared to a $F_{h,n-q-1}$ distirbution.

Multivariate regression

A special case is testing individual predictor variables, in which case $h = 1$, but the formulas hold for this case as well. In this particular case (with numerator degrees of freedom equal to 1), the F statistic is the square of a t statistic.

The R^2 value gives the proportion of variance "explained" by the model, which is

$$R^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{\widehat{\beta}' \mathbf{X}' \mathbf{y} - n\overline{y}^2}{\mathbf{y}' \mathbf{y} - n\overline{y}^2}$$

Multivariate regression

For multivariate regression, we have p variables for y, so that $\mathbf{Y} = (y_{ij})$ is an $n \times p$ matrix. The observation vectors are $\mathbf{y}_i'$, $i = 1, \ldots, n$. As usual, observation vectors are considered as column vectors even though they are written horizontally in the data file and even though they correspond to rows of $\mathbf{Y}$.

Multivariate regression

The design matrix $\mathbf{X}$ is as before with a column of 1s and q columns corresponding to x variables. However, there is now a column of q β coefficients for each of the p response variables. The model now has

$$\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) = (\beta_{ij}),$$

which is a $(q+1) \times p$ matrix. The model can be written as

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

The model for an individual column of $\mathbf{Y}$ is equivalent to a univariate multiple regression model. (It so happens that B is the capital of β in Greek. However Ξ is not the capital of ε, so this choice of notation seems a bit inconsistent. However $\mathbf{E}$ is used as $\widehat{\boldsymbol{\Xi}}'\widehat{\boldsymbol{\Xi}}$, which is the matrix analogue of SSE.

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

The model for the first column of $\mathbf{Y}$ is

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

Multivariate regression

The assumptions of the model are

1. $E(\mathbf{Y}) = \mathbf{XB}, E(\mathbf{\Xi}) = \mathbf{O}$
2. $Cov(\mathbf{y})_i) = \mathbf{\Sigma}$, for $i = 1, \ldots, n$, where $\mathbf{y}_i'$ is the ith row of $\mathbf{Y}$
3. $Cov(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{O}$ for $i \neq j$

Note that $Cov(\mathbf{y})_i$ is $p \times p$.

Multivariate regression

Similar to univrariate multiple regression,

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

so $\mathbf{y}$ was replaced with $\mathbf{Y}$ in the formula.

Multivariate regression

An estimator for the covariance matrix of $\mathbf{y}_i$ is

$$\mathbf{S}_e = \frac{\mathbf{E}}{n-q-1} = \frac{(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}})'(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}})}{n-q-1}$$

The $\mathbf{B}$ can be partitioned so that there is essentially a vector of intercept terms, one for each response variable, and a matrix of other non-intercept coefficients.

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_0' \\ \mathbf{B}_1 \end{pmatrix} = \left(\begin{array}{cccc} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \hline \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{array} \right).$$

By analogy with (10.14) and (10.15), the estimates are

$$\hat{\mathbf{B}}_1 = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{Y}, \tag{10.50}$$

$$\hat{\boldsymbol{\beta}}_0' = \overline{\mathbf{y}}' - \overline{\mathbf{x}}'\hat{\mathbf{B}}_1, \tag{10.51}$$

Multivariate regression

You can also express $\widehat{\mathbf{B}}$ as

$$\widehat{\mathbf{B}} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$$

where we use an estimated covariance matrix of all variables (whether or not they are really random):

$$y_1, \ldots, y_p, x_1, \ldots, x_q$$

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix}$$

Here $\mathbf{S}$ is $(p+q) \times (p+q)$.

Multivariate regression

We typically wish to test $H_0 : \mathbf{B}_1 = \mathbf{0}$ against $H_A : \mathbf{B}_1 \neq \mathbf{0}$. This only requires that one $\beta_{ij} \neq 0$ for some $i \geq 1$ and some $j \geq 1$.

Similar to MANOVA, we define matrices $\mathbf{E}$ and $\mathbf{H}$. To total sum of squares can be partitioned into these two matrices:

$$\begin{aligned} \mathbf{Y}'\mathbf{Y} - n\overline{y}'\overline{y} &= (\mathbf{Y}'\mathbf{Y} - \widehat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}) + (\widehat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - n\overline{y}'\overline{y}) \\ &= \mathbf{E} + \mathbf{H} \end{aligned}$$

Multivariate regression

Similar to MANOVA, the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ can be used to create test statistics for testing the null hypothesis.

$$\text{Wilk's Lambda: } |\Lambda = \prod_{i=1}^{\min(p,q)} \frac{1}{1+\lambda_i}$$

$$\text{Roy's greatest root: } \frac{\lambda_1}{1+\lambda_1}$$

$$\text{Pillai's test: } \sum_{i=1}^{\min(p,q)} \frac{1}{1+\lambda_i}$$

$$\text{Lawley-Hotelling test: } \sum_{i=1}^{\min(p,q)} \lambda_i$$

Multivariate regression

If you don't want to use specialized tables of critical values in the book for these statistics, you can use the same F approximations that we used for MANOVA for Wilk's Lambda, where $\Lambda = \Lambda_{q,p,n-p-1}$, so that the degrees of freedom for the F test are a function of q, p, and $n - p - 1$.

Multivariate regression

As in the univariate, multiple regression case, you can whether subsets of the x variables have coefficients of 0. In this case, there is a matrix in the null hypothesis, $H_0 : \mathbf{B}_d = \mathbf{0}$. The $\mathbf{E}$ and $\mathbf{H}$ matrices are given by

$$\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \widehat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}$$

$$\mathbf{H} = \widehat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - \widehat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y}$$

And the test statistics are given as before.

It is also possible to try to pick a subset of the y variables if some of the y variables are not well-explained by the x variables. This can also be done with stepwise procedures.

Canonical correlation analysis

Correlation between two variables measure the linear relationship between those two variables. In canonical correlation, we measure the linear relationship between two sets of variables. Typically, variables within each set will be related in some way, for example a set of student aptitudes or qualifications (high school GPA, SAT scores) and outcomes (college GPA, GRE scores), or variables on a child and similar variables on their parent.

Canonical correlation analysis

If you only have one variable in one set, y, and q variables in the other set, $x_1, \ldots, x_q$, then you can define

$$\mathbf{S} = \begin{pmatrix} s_y^2 & \mathbf{s}'_{yx} \\ \mathbf{s}_{xy} & \mathbf{S}_{xx} \end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix} 1 & \mathbf{r}'_{yx} \\ \mathbf{r}_{xy} & \mathbf{R}_{xx} \end{pmatrix}$$

where $\mathbf{r}'_{yx}$ is a vectorwith sample correlations between y and x_i, $i = 1, \ldots, q$.
The squared multiple correlation between y and $x_1, \ldots, x_q$ is

$$R^2 = \mathbf{r}'_{yx}\mathbf{R}_{xx}^{-1}\mathbf{r}_{xy}$$

Canonical correlation analysis

When there are multiple y variables, we use

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix}$$

A measure of association is

$$R_M^2 = \frac{|\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}|}{|\mathbf{S}_{yy}|} = |\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}| = \prod_{i=1}^{\min(p,q)} r_i^2$$

where the r_i^2 terms are the eigenvalues of $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$.

The values r_i, $i = 1, \ldots, \min(p, q)$ are called the **canonical correlations**.

Canonical correlation analysis

The largest canonical correlation r_1, is used as a measure of association of the two sets of variables. An interpretation of r_1^2 is that it is the maximum squared correlation between a linear combination of the y variables and a linear combination of the x variables.

With each canonical correlation, there is a set of associated linear combinations so that there exist $\mathbf{a}_i$ and $\mathbf{b}_i$ such that

$$r_i = cor(\mathbf{a}'\mathbf{y}, \mathbf{b}'\mathbf{x})$$

Canonical correlation analysis

There is some interesting discussion in the book about how the author thinks that canonical correlation is often misapplied in practice.

If you are ever asked to use canonical correlation, try looking this up!