Regression Models (Part XII: Multivariate Regression Analysis)

A. Multivariable regression analysis

Here we extend linear regression so that our models can contain more variables. A natural first approach is to assume additive effects, basically extending our linear model to a plane or hyperplane. This technique represents one of the most widely used and successful methods in statistics.

B. Multivariable regression analyses: adjustment

If I were to present evidence of a relationship between breath mint useage (mints per day, X) and pulmonary function (measured in FEV), you would be skeptical. Likely, you would say, 'smokers tend to use more breath mints than non smokers, smoking is related to a loss in pulmonary function. That's probably the culprit.' If asked what would convince you, you would likely say, 'If non-smoking breath mint users had lower lung function than non-smoking non-breath mint users and, similarly, if smoking breath mint users had lower lung function than smoking non-breath mint users, I'd be more inclined to believe you'. In other words, to even consider my results, I would have to demonstrate that they hold while holding smoking status fixed.

This is one of the main uses of multivariate regression, to consider a relationship between a predictor and response, while accounting for other variables.

C. Multivariable regression analyses: prediction

An insurance company is interested in how last year's claims can predict a person's time in the hospital this year. They want to use an enormous amount of data contained in claims to predict a single number. Simple linear regression is not equipped to handle more than one predictor. How can one generalize SLR to incorporate lots of regressors for the purpose of prediction? What are the consequences of adding lots of regressors? Surely there must be consequences to throwing variables in that aren't related to Y? Surely there must also be consequences to omitting variables that are?

D. The linear model

The general linear model extends simple linear regression (SLR) by adding terms linearly into the model.

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^{p} X_{ik} \beta_j + \epsilon_i$$

Here $X_{1i} = 1$ typically, so that an intercept is included. Least squares (and hence ML estimates under iid Gaussianity of the errors) minimizes:

$$\sum_{i=1}^{n} (Y_i - \sum_{k=1}^{p} X_{ki} \beta_j)^2$$

Note, the important linearity is linearity in the coefficients. Thus

$$Y_i = \beta_1 X_{1i}^2 + \beta_2 X_{2i}^2 + \cdots + \beta_p X_{pi}^2 + \epsilon_i$$

is still a linear model. We've just squared the elements of the predictor variables.

E. Estimation

Recall, the LS estimate for regression through the origin is,

$$E[Y_i] = X_{1i}\beta_1, was \sum X_i Y_i / \sum X_i^2.$$

Let's consider two regressors, $E[Y_i] = X_{1i}\beta_1 + X_{2i}\beta_2 = \mu_i$. Least squares tries to minimize:

$$\sum_{i=1}^{n} (Y_i - X_{1i}\beta_1 - X_{2i}\beta_2)^2$$

We describe fitting with two regressors using residuals, since it will help us to understand how multivariable regression adjusts an effect for another variable. The result is that the estimate for β_1 is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} e_{i,Y|X_2} \, e_{i,X_1|X_2}}{\sum_{i=1}^{n} e_{i,X_1|X_2}^2}$$

,

where, $e_{i,Y|X_2}$ is the residual having $fitX_2$ on Y and $e_{i,X_1|X_2}$ is the residual having $fitX_2$ on Y. That is, the regression estimate for β_1 is the regression through the origin estimate having regressed X_2 out of both the response and the predictor. Similarly, the regression estimate for β_2 is the regression through the origin estimate having regressed X_1 out of both the response and the predictor.

More generally, multivariate regression estimates are exactly those having removed the linear relationship of the other variables from both the regressor and response. This demonstrates the sense in which multivariate regression variables adjust for the effect of the other variables.

F. Example with two variables, simple linear regression

We already have one of the most important examples of two variables down, linear regression. The linear regression model is:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i}$$

where $X_{2i} = 1$ is an intercept term. Let's double check our rule, since we already know what the least squares estimates are in this case.

Notice the fitted coefficient of $X_{2i} on Y_i is \overline{Y}$, the mean of the Ys. Then the residuals are $e_{i,Y|X_2} = Y_i - \overline{Y}$.

Similarly, the fitted coefficient of $X_{2i} on X_{1i}$ is $\overline{X}_1$. Then, the residuals are $e_{i,X_1|X_2} = X_{1i} - \overline{X}_1$.

Thus let's work out the estimate for β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n e_{i,Y|X_2}\, e_{i,X_1|X_2}}{\sum_{i=1}^n e_{i,X_1|X_2}^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = Cor(X,Y)\frac{Sd(Y)}{Sd(X)}$$

This agrees with the estimate that we came up with before.