

# Chapter 6 — The Formalities of Multiple Regression

February 12, 2005

## 1 Introduction

There can now be  $x_1, x_2, \dots, x_p$  predictors collected in a  $p \times 1$  vector  $\mathbf{x}$  and *terms* that can be constructed from one or more predictors.

## 2 Terms and Predictors

Now,  $y$  depends on  $\mathbf{x}$ , where the  $\mathbf{x}$  is the set of  $p$  predictors. Predictors may be:

- numerical (i.e., equal interval) even if with a limited range (e.g., only positive);
- categorical (e.g., ethnicity), with binary as a special case; or
- ordinal (e.g., military rank), but this can lead to troubles we will talk about later.

*Terms* are built from predictors. Thus, one can write

$$E(y|\mathbf{x}) = \eta_0 u_0 + \eta_1 u_1 + \dots + \eta_{k-1} u_{k-1}. \quad (1)$$

There are  $k$  regression coefficients (including the intercept) with each of the  $k - 1$  terms computed from known values of the  $p$  predictors.

Each of the  $k - 1$  terms is represented by  $u_j$  ( $j = 1, 2, \dots, k - 1$ ). The values for  $u_0$  are an  $n \times 1$  vector of 1's, with  $n$  equal to the number of observations.

The constraints typically imposed and the information often brought to bear from outside should now be familiar.

1. The concern is with the conditional means of  $y$ .
2. The goal is to characterize the path of the means with a hyperplane.
3. The least squares criterion is applied, which imposes symmetry on the errors and places special weight on the largest departures from the regression line.
4. For statistical inference, we assume that the data were produced by random sampling, a natural approximation thereof from a well-defined population, or through a well-understood data-generation process accurately characterized by a model.
5. We assume that the natural world constructed the data in the population so that a hyperplane passes through each of the conditional means. Alternatively, the process by which the data were generated produces errors that are uncorrelated with the terms in the model.
6. In either case, the errors are independent of one another.
7. We require that the natural world forces all of the conditional variances  $\text{Var}(y|\mathbf{x})$  to be the same.
8. We require no systematic measurement error for any of the variables and no random measurement error for  $\mathbf{x}$ .
9. In some instances, we require that  $y|\mathbf{x}$  was normally distributed.
10. We impose the thought experiment of a limitless number of independent samples from the given population or a limitless number of independent realizations of the data generation process.
11. For causal inference, we assume that one of the response schedule formulations apply in the sense that the data were generated as if the world actually operated as the response schedule requires.

### 3 Some Notation for Multiple Regression

At this point, we need to provide the matrix notation as follows:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

$$\mathbf{u} = \begin{pmatrix} u_0 = \mathbf{1} \\ u_1 \\ \vdots \\ u_{k-1} \end{pmatrix}$$

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_{k-1} \end{pmatrix}.$$

The linear model with more than a single predictor can now be written as

$$E(y|\mathbf{x}) = \boldsymbol{\eta}^T \mathbf{u}. \quad (2)$$

### 4 Estimation

1. Again we use least squares and minimize

$$RSS(\mathbf{h}) = \sum_{i=1}^n (y_i - \mathbf{h}^T \mathbf{u}_i)^2, \quad (3)$$

where  $\mathbf{h}$  is a set of  $k - 1$  “trial” coefficients. This leads to

$$\hat{\boldsymbol{\eta}} = (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \mathbf{y}, \quad (4)$$

where  $\mathbf{U}$  is an  $n \times k$  matrix of terms (including  $u_0$ ). The estimates of the regression coefficients  $\hat{\boldsymbol{\eta}}$  are unique, given the data. However, no variables can be an exact linear combination of any other variables.

2. The residual variance is estimated by

$$\hat{\sigma}^2 = \frac{RSS}{n - k}. \quad (5)$$

Now the estimate of the variance-covariance matrix of the regression coefficients (including the intercept) is

$$\text{Var}(\hat{\boldsymbol{\eta}}) = \hat{\sigma}^2 \mathbf{M}, \quad (6)$$

where  $\mathbf{M}$  is a  $k \times k$  matrix through which  $\sigma^2$  is transformed in the variance-covariance matrix of the regression coefficients (including the intercept). Formally,  $\mathbf{M} = (\mathbf{U}^T \mathbf{U})^{-1}$ . The diagonal elements in  $\hat{\sigma}^2 \mathbf{M}$  are the estimated squared standard errors, and the off-diagonal elements are the estimated covariances.

3. Now define the “hat matrix” as follows:

$$\mathbf{H} = \mathbf{U} (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T. \quad (7)$$

For each observation, there is a value  $h_i$  that is a function solely of the “input” matrix  $\mathbf{U}$ . The  $h_i$  is the  $ii$  (diagonal) element of  $\mathbf{H}$  (which is  $n$  by  $n$ ) and is called the leverage of observation  $i$ . It indicates the potential “pull” from regression terms an observation has on the regression fit.

The value of  $h_i$  is always between zero and one. If there are  $k - 1$  terms in the regression equation, the leverages sum to  $k - 1$ . Thus, the mean leverage is  $(k - 1)/n$ . If  $n$  is large relative to  $k - 1$ , individual leverages are unlikely to stand out (and matter).

4. Finally, we again consider a measure of goodness of fit:

$$R^2 = \frac{SYY - RSS}{SYY} = \frac{SS_{\text{reg}}}{SYY}. \quad (8)$$

The coefficient of determination, which denotes that percentage of the variance of  $y$  “explained” by the set of terms, indicates how closely the observations cluster around the fitted values. Although on intuitive grounds a “good fit” certainly seems desirable, on closer consideration, it is often unclear why a good fit is so good.

## 5 How Multiple Regression “Holds Constant”

Consider for now the case of two terms  $(x, z)$ :

$$E(y|x, z) = \eta_0 + \eta_1 x + \eta_2 z. \quad (9)$$

We will focus for now on the regression coefficient for  $x$  ( $\eta_1$ ). We estimate the parameters of the following two equations:

$$E(y|z) = \alpha_0 + \alpha_1 z, \quad (10)$$

$$E(x|z) = \beta_0 + \beta_1 z. \quad (11)$$

For each, we compute the residuals,  $e_{y|z}$  and  $e_{x|z}$ . Now consider

$$E(e_{y|z}|e_{x|z}) = \gamma_0 + \gamma_1 e_{x|z}. \quad (12)$$

Figure 1 is a scatter plot of two sets of such residuals. Overlaid is a least squares line based on Equation 12. The estimate of  $\gamma_1$  will be the same as the estimate of  $\eta_1$  coming out of the usual least squares formula applied to Equation 9. A parallel logic applies to what  $\eta_2$  means. In both cases, these are just manipulations of the data. Any correspondence to what goes in the real world needs to be argued.

To help underscore the impact the holding constant operation can have, consider the following example with  $x$  and  $z$  as predictors. Although  $y$  is a linear function of  $x$  and  $z$ , the relationship between  $x$  and  $z$  is nonlinear. Figure 2 shows a scatter plot of the relationship between  $y$  and  $x$  ignoring  $z$ . The “marginal response plot” that follows (“marginal” because only the two variables  $x$  and  $y$  are involved) indicates a strong nonlinear relationship. In contrast, the same plot, Figure 3, with  $y$  and  $x$  residualized for  $z$  shows a linear relationship.

## 6 Added Variables Plots

These are essentially the plots of the residualized response and each residualized predictor in turn. Think of them as regular (marginal) scatter plots, but with the relationship in question adjusted for all other confounders in the equation and the appropriate regression line overlaid. Added variable plots have the following properties.

1. The estimated intercept will be zero (as long as you do not use a no-intercept model).
2. The slope will always be the same as the estimated regression obtained by the usual formula.
3. The residuals will also be identical to those of the full regression model.
4. If you calculate the Pearson correlated in the AVP data, you get the conventional partial correlation coefficient.
5. The AVP standard errors estimated will be a little off because the proper degrees of freedom will not be computed. The plot thinks there is only a single term.

The key use of AVPs is to visualize how the adjusted fit compares to the adjusted data. Does the functional form look right? Does it seem that a few observations are very influential on the fit?

Here is an example as part of regression analysis of how to construct a set of added variable plots for a data set called Freedman in the car library. I have had trouble with my Mac using some of the fancier options, such as cycling through each of the AVPs or printing all of the AVPs at once. It may work better on a Windows system. So, in the final step, the second argument is the variable whose AVP you want. Just use that command again for each variable of interest. This seems to be the most simple and robust way to proceed. On the other hand, feel free to try some of the fancier stuff. They may work on your system.

```
> library(car)
> help(avp)
> data(Freedman)
> attach(Freedman)
> summary(Freedman)
> out<-lm(crime~population+nonwhite+density)
> summary(out)
> avp(out,population)
```

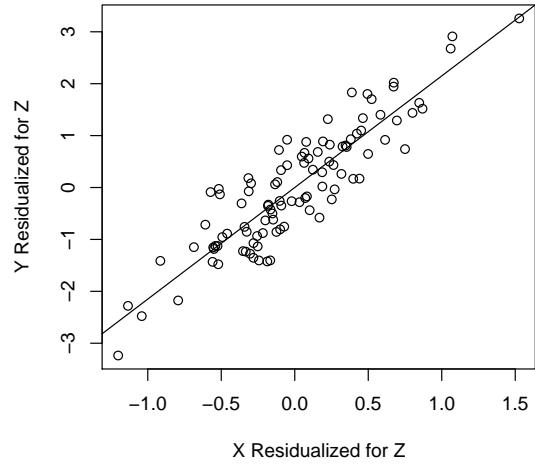


Figure 1: Scatter Plot of the Residualized Response and Residualized Term

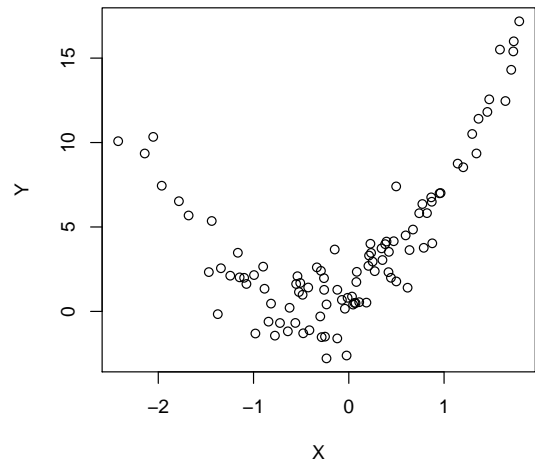


Figure 2: Scatter Plot of  $y$  on  $x$  with  $x$  a Nonlinear Function of  $z$

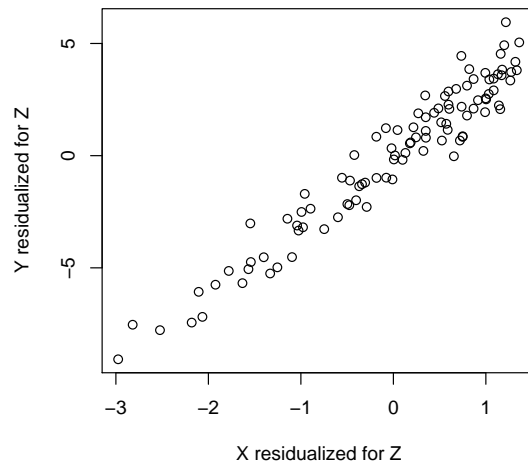


Figure 3: Residualized Scatter Plot of  $y$  on  $x$  with  $x$  a Nonlinear Function of  $z$