Multiple Regression

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Multiple Regression

The Multiple Regression Model

- The simple linear regression model states that

$$E(Y|X = x) = \beta_0 + \beta_1 x \quad (1)$$
$$\text{Var}(Y|X = x) = \sigma^2 \quad (2)$$

- In the multiple regression model, we simply add one or more predictors to the system. For example, if we add a single predictor X_2, we get

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (3)$$

An Example: The U.N. Fertility Data

- The `UN11` dataset contains U.N. Data on factors related to fertility for 199 countries.
- We are curious about the relationship between female life expectancy (`lifeExpF`) and two other variables.
- One potential predictor of life expectancy is `fertility`, measured by the average number of children per woman in the country.
- Another potential predictor is the general economic condition of the country. We measure that by log(ppgdp), the natural logarithm of ppgdp, the per capita gross domestic product in U.S. dollars.

An Example: The U.N. Fertility Data

- We start with a simple linear regression predicting female life expectancy from `fertility`, another simple linear regression predicting female life expectancy from from log(ppgdp, and a third relating the two potential predictors to each other.
- These three simple linear regressions are called marginal plots.
- Data for these 3 variables, plus the names of the localities for each data point, are in the `UN11` dataset.

An Example: The U.N. Fertility Data

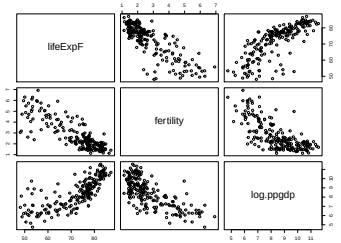
Preliminary Analyses with Simple Regression

We begin (after loading alr4) by predicting examining the correlations and scatterplots among the 3 variables.

```
> library(alr4)
> attach(UN11)
> log.ppgdp <- log(ppgdp)
> cor(cbind(lifeExpF,fertility,log.ppgdp))

             lifeExpF  fertility  log.ppgdp
lifeExpF    1.0000000 -0.8235881  0.7722587
fertility  -0.8235881  1.0000000 -0.7210800
log.ppgdp   0.7722587 -0.7210800  1.0000000

> pairs(~lifeExpF + fertility + log.ppgdp)
```



An Example: The U.N. Fertility Data

Preliminary Analyses with Simple Regression

- Both the scatterplots and the correlation matrix reveal that both `fertility` and log(*ttppgdp*) are excellent predictors of `lifeExpF`.
- However, they both are excellent predictors of each other, because, up to a point, `fertility` decreases sharply as *log*(`ppgdp`) increases.
- Since the two predictors are substantially redundant, the regression coefficients in a multiple regression equation using both predictors simultaneously may differ substantially from the slope of the two simple regression lines obtained by using only one predictor at a time.

An Example: The U.N. Fertility Data

Preliminary Analyses with Simple Regression

- Given what we have seen so far, what could we say about the proportion of variability of `lifeExpF` explained by `fertility` and `log.ppgdp`?
- A general rule in multiple regression is that adding a predictor can never decrease the proportion of variance accounted for.
- Moreover, if the predictors were uncorrelated, the proportions of variance would be additive, i.e. the pair of predictors would explain $67.8\% + 59.6\% = 117.4\%$ of the variance.
- In this case, the substantial redundancy of the two variables means that, together, they will predict substantially less than this sum.
- However, adding a predictor can never reduce the proportion of variance accounted for in a multiple regression, so we know that adding `log.ppgdp` to `fertility` will result in a system that accounts for somewhere between 67.8% and 100% of the variance in `lifeExpF`.

Effect of an Added Variable

The Added Variable Plot

- Suppose we have a simple regression in which `fertility` is used to predict `lifeExpF`, and we wish to evaluate the impact of adding `fertility` to the regression equation.
- It can be shown algebraically that the unique effect of adding `log.ppgdp` to a mean function that already includes `fertility` is determined by the relationship between the part of `lifeExpF` that is not explained by `fertility` and the part of `log.ppgdp` that is not explained by `fertility`.
- In a linear prediction system, the "unexplained parts" are just the residuals from these two simple regressions.

Effect of an Added Variable

The Added Variable Plot

- To examine these, we can look at the scatterplot (and the linear fit) of the residuals from the regression of `lifeExpF` on `fertility` versus the residuals from the regression of `log.ppgdp` on `fertility`.
- This scatterplot is called an added variable plot.
- The slope of the regression line in the added variable plot is identical to the OLS estimate of the regression coefficient β_2 for the added variable if the multiple regression model is fit with both predictors included.
- Recalling our earlier discussion of partial correlation, we realize that the added variable plot actually depicts the partial regression of `lifeExpF` on `log.ppgdp` with `fertility` partialled out.
- So a regression coefficient in multiple regression is a partial regression coefficient between a variable and the criterion, with all the other regressors partialled out of both variables.

Effect of an Added Variable

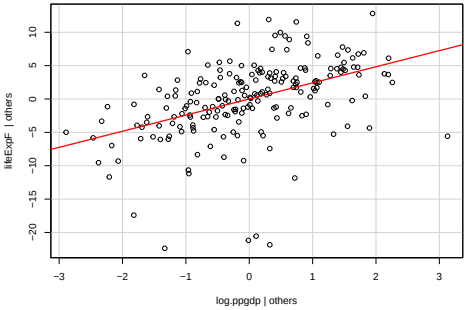
Automatic Generation of Added-Variable Plots

- The `avPlots` function generates added variable plots automatically.
- Here is an example. You first specify the full model.

```
> m2 <- lm(lifeExpF ~ fertility + log.ppgdp)
```

- Then you specify the terms you wish to have added variable plots for in the `terms` specifier.

```
> avPlots(m2,terms= ~ log.ppgdp)
```



Predictors and Regressors

Introduction

In Section 3.3 of ALR, Weisberg introduces a number of key ideas and nomenclature in connection with a regression model of the form

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{4}$$

with

$$\text{Var}(Y|X) = \sigma^2 \tag{5}$$

Predictors and Regressors

Introduction

- Regression problems start with a collection of potential predictors.
- Some of these may be continuous measurements, like the height or weight of an object.
- Some may be discrete but ordered, like a doctor's rating of overall health of a patient on a nine-point scale.
- Other potential predictors can be categorical, like eye color or an indicator of whether a particular unit received a treatment.
- All these types of potential predictors can be useful in multiple linear regression.
- A key notion is the distinction between *predictors* and regressors in the regression equation.
- In early discussions, these are often synonymous. However, we quickly learn that they need not be.

Predictors and Regressors

Types of Regressors

Many types of *regressors* can be created from a group of predictors. Here are some examples

- The intercept. We can rewrite the mean function on the previous slide as

$$E(Y|X) = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{6}$$

  where X_0 is a term that is always equal to one. Mean functions without an intercept would not have this term included.
- Predictors. The simplest type of regressor is simply one of the predictors.

Predictors and Regressors

Types of Regressors

- Transformations of predictors. Often we will transform one of the predictors to create a regressor. For example, X_1 in a previous example was the logarithm of one of the predictors.
- Polynomials. Sometimes, we fit curved functions by including polynomial regressors in the predictor variables. So, for example, X_1 might be a predictor, and X_2 might be its square.
- Interactions and other combinations of predictors. Combining several predictors is often useful. An example of this is using body mass index, given by weight in kg divided by the square of height in meters, in place of both height and weight, or using a total test score in place of the separate scores from each of several parts. Products of predictors called interactions are often included in a mean function along with the original predictors to allow for joint effects of two or more variables.

Predictors and Regressors

Types of Regressors

- Dummy variables and factors. A categorical predictor with two or more levels is called a factor. Factors are included in multiple linear regression using dummy variables, which are typically regressors that have only two values, often 0 and 1, indicating which category is present for a particular observation. A categorical predictor with two categories can be represented by one dummy variable, while a categorical predictor with many categories can require several dummy variables.
- Regression splines. Polynomials represent the effect of a predictor by using a sum of regressors, like $\beta_1 x + \beta_2 x^2 + \beta_3 x^3$. We can view this as a linear combination of basis functions, given in the polynomial case by the functions (x, x^2, x^3). Using splines is similar to fitting a polynomial, except we use different basis functions that can have useful properties under some circumstances.

Predictors and Regressors

Types of Regressors

- Principal components. In some cases, one has many predictors, and there are substantial redundancies between predictors.

  For reasons we shall explore a bit later in the course, this situation can lead to computational problems and lack of clarity in the analysis.

  Data reduction methods can help us in this kind of situation, by reducing the large number of predictors into a smaller set that captures the unique variability shared by those predictors.

  Principal components are linear combinations of the predictors that are themselves uncorrelated, yet capture the maximum possible variance in the predictor set.

Predictors and Regressors

Types of Regressors

- It is easy to see that a multiple regression with k predictors may have fewer than k regressors or more than k regressors.

Multiple Regression in Matrix Notation

- As we present additional results in multiple regression, it will prove useful to have at our disposal some basic matrix algebra, which we will present in the next two modules.