General Theory on Modeling and Data Fitting

Notes

In modeling our goal is typically to explain and/or predict the values of a target variable, given values of other variables and some modeling assumptions.

Terminology and General Methodology

Here are some key terms:

Target Variable One variable is the target of our modeling process. Our ultimate goal is to understand that variable in terms of others.

Predictor Variables Zero or more variables, which we expect to be related to the target variable. We want to be able to say something about the target variable *given* values for the predictor variables

Model Formula A specific formula / function of the predictor variables and some parameters, whose output is meant to model the target variable. We would write this in general form as:

$$Y \sim f(X_1, X_2, ...; \beta_1, \beta_2, ...)$$

where Y is the target variable, X_1, X_2, ... are the predictor variables, and β_1, β_2, ... are parameters, values to be determined.

Model Fit We say that we *fit* the model to the data, if using the data we determine values for the parameters β_1, ... that give in some sense the "optimal" fit. This results in a **specific model**, rather than the **general model** of the previous point.

Predicted Values Given a specific model, for any set of values for X_1, X_2, ..., we can use the formula to compute a specific value for Y. This is called the *predicted value*. These values are usually denoted by $\hat{y}$.

Actual Values These are the value that Y has in the actual data. More often denoted as y.

Residuals The differences between the Actual Values and the Predicted Values, $y - \hat{y}$. You can think of the residuals as telling you how far your specific model is from accurately predicting your actual values. In other words, they measure how much of an error you are making when predicting.

SSR/RSS The "Sum of Squared Residuals", sometimes called "Residual Sum of Squares". This is a measure of the overall error you are making at all your data points together.

Some times we adjust this by dividing by $n - 1$ or something similar.

Normal modeling techniques try to choose the parameters so as to minimize this sum.

It is important to identify some key steps in the process:

1. You have to decide what variables X_1, X_2, ... to include. We will largely dodge this question.
2. You have to decide how to combine them, i.e. what form the function f will take. We will restrict ourselves to linear functions, but there are other options out there.
3. You have to decide how to assess how good a fit you have. We will use SSR for this, but there are other options out there.
4. You have to find the parameter values that for your given choice of form for the function f achieve the best fit, in whatever way you have defined it.
5. You have to assess if that is a good fit, or whether you should look for other function forms f.

We will focus exclusively on linear functions of one predictor variable, so we will not be able to do many of these. But in this introduction I will talk about some of these concepts in more general terms.

Model Examples

Here are some basic examples of the above ideas. Our target variable is student GPA, we want to try to find a way to predict it.

Constant Model This is the simplest model we can try to fit. We have no predictor variables, so all we can do is predict a single number:

$$Y \sim \beta$$

Where β is the parameter. All we have to do to "fit" the model is to provide a constant value for the parameter. So we could for instance say $Y \sim 2.93$, meaning that we predict that the student's GPA is equal to 2.93.

The predicted values in this case are always equal to this constant value. The residuals are the differences $y - \beta$.

It turns out, that if we want to choose a value for β that makes the SSR as small as possible, then we must choose $\beta = \bar{y}$, the mean of the y values.

The best constant model fit is when that constant is equal to the mean of the y values, $\bar{y}$. The Adjusted SSR measure in this case equals the *variance* of y.

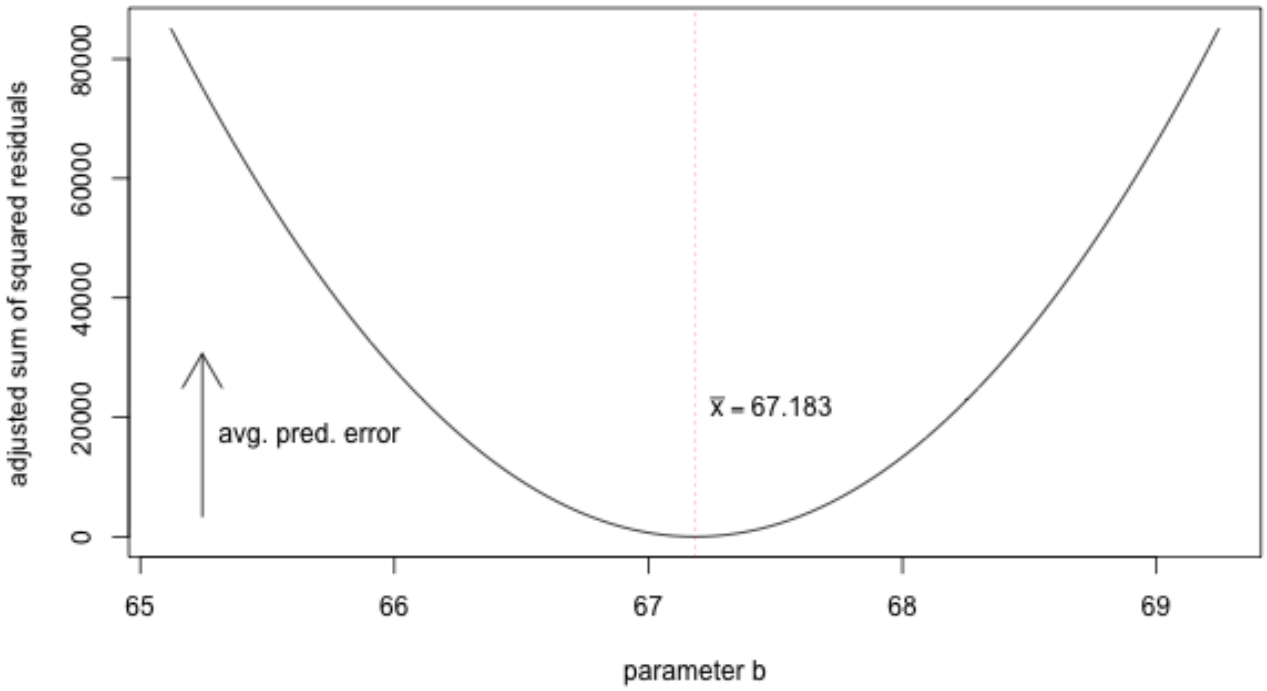

Figure 1: Prediction error. Constant model

As one example of this, consider the behavioral survey data we have been looking at. If we were asked to fit a constant model to the height variable in that dataset, then we would make the model $Y \sim 67.183$ as that value is the mean. We can see that this value is the one that minimizes the "sum of squared residuals":

$$\frac{\sum(y - \hat{y})^2}{n - 1}$$

In the case of a constant model, that ends up being what we called the variance, because $\hat{y}$ is always equal to the mean $\bar{y}$.

If we try to compute the sum

$$\frac{\sum(y - b)^2}{n - 1}$$

for different values of b, we can see that the mean achieves the smallest value. In the following graph, the x axis represents possible values for the constant b, and the y axis represents the corresponding average squared prediction error for that constant. Our goal is to minimize that error.

Factors Another common case is when we try to have a predictor variable that is categorical. These are often called "factors". For instance we could say that we will try to predict the student's GPA based on their gender.

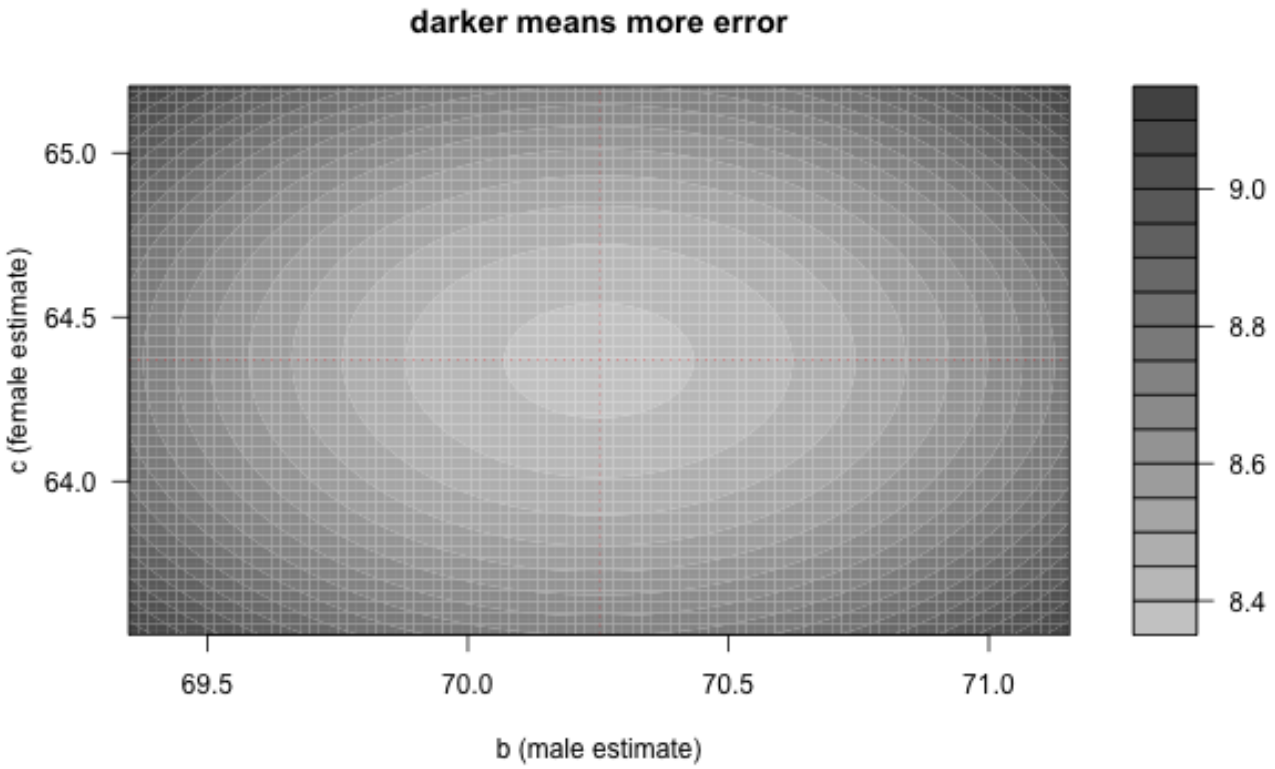

Figure 2: Prediction error. Factor model

In this case we would basically need to provide two parameter values: One parameter for our guess for the GPA if the student is male, and another for the GPA if the student is female.

It turns out that the best guesses in this case are again the means of the male and female students respectively:

> The best model for the case where Y is a scalar variable and X is a categorical variable is to assign for a particular value x_0 the average of the y values for those cases whose x value equals x_0.

As an example, in the behavioral survey, one of the variables is the gender. To build a model that based on the gender tries to predict the height, we simply need to provide a value for the male gender, and a value for the female gender. These values are respectively the height averages for males and females, 70.252 and 64.368 respectively.

In this case our model $Y \sim f(X)$ has two parameters: $f(\text{male}) = b$ and $f(\text{female}) = c$. These are chosen so at to make the corresponding sum of squared residuals as small as possible. The following two-dimensional graph gives us a sense of where this minimum overall error is achieved. The darker areas correspond to more error.

Linear Equation The most common model, and one we will spend more time with next week, is that of a linear equation. For example, perhaps we think that a student's high-school gpa should be a good predictor of their college gpa. In that case, if we denote a student's high-school GPA with x and their college gpa with y, we would be looking for an equation of the form:

$$y = \alpha + \beta x$$

Where α, β are the parameters, and we would like to choose their "best values" to fit the data in any given scenario. We will explore this more in the next section.