Lab 2: Multiple linear regression – Part 1

March 14, 2017

1 Review

This two weeks we began discussing multiple linear regression. The **model** that we fit is an extension of that fit in simple linear regression, and is given by:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i,$$

We **assume** that the observations Y_is are independent from each other, $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ and that $p < n$.

The β_j's are **interpreted** as the the change in the *expected response* (i.e., $E(Y)$) per unit change in X_j , holding the other X_i ($i \neq j$) constant.

1. What is the multiple linear regression model in matrix form?

2. What are each of the pieces of the model representing?

3. What is the least squares estimate for $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ in matrix form?

In addition to including multiple covariates, there are several reasons for using a multiple linear regression model. These include:

1. Creating a model with a predictor that is described by several dummy variables

$$E(Y_i) = \beta_0 + \beta_2 I_{i2} + ... + \beta_5 I_{ip}$$

2. Incorporating nonlinear effects by including polynomial terms of a predictor.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + ...\beta_p X_i^p + \varepsilon_i$$

3. Adjusting for confounding.

4. Incorporating interactions.

2 Example

The data set contains information from a study of 25 patients with cystic fibrosis. The investigators were interested in assessing predictors of *PEmax*, a measure of malnutrition. The data set contains a new categorical variable labeled $\texttt{FEV}_2$ that we will examine more closely this week. The categorical variable $\texttt{FEV}_2$ has three ordinal levels: 1,2 and 3. The data set named `cf2.sas7bdat` is posted on course web page in Moodle in the folder "Lab 2" under topic 3.

2.1 multiple linear regression with categorical predictors

We will being by considering the impact of the new variable in the data set, $\texttt{FEV}_2$ on `PEmax`.

- Create binary indicator variables to represent $\texttt{FEV}_2$, using level 1 of $\texttt{FEV}_2$ as the reference level. How many binary indicator variables do you need?

- Write the multiple linear regression model for prediction `PEmax` from $\texttt{FEV}_2$, using level 1 of $\texttt{FEV}_2$ as the reference level.

- We will now fit this model in SAS.

- Interpret the regression coefficients in this model?

2.2 confounding

We are interested in examining the impact of `Age` and $\texttt{FEV}_2$ on `PEmax`. In this example, our primary interest is with Age, but we also want to investigate if $\texttt{FEV}_2$ is a confounder.

First we will investigate confounding. There are two ways to do it.

One way by looking at the association between these three variables directly.

- Calculate Pearson correlation coefficient for continuous variables `Age` and `PEmax`. Is r significantly different from 0? Is there association between `Age` and `PEmax`?

- Investigate the association between $\texttt{FEV}_2$ and `PEmax`. Notice that $\texttt{FEV}_2$ is categorical variable with 3 levels and `PEmax` is continuous. What test should we use?

- Investigate the association between $\texttt{FEV}_2$ and `Age`. What test should we use?

- Assuming that there is no causal relationship between `Age` and $\texttt{FEV}_2$, do we think that $\texttt{FEV}_2$ is a confounder of the relationship between `Age` and `PEmax`? Why?

We can also compare the unadjusted β for `Age` with the adjusted β for `Age` after controlling for $\texttt{FEV}_2$ to see if $\texttt{FEV}_2$ confounds the association between `Age` and `PEmax`. Usually, we conclude that $\texttt{FEV}_2$ is a confounder when we see a change in β of 10% or more.

- To begin, we fit simple linear regression model with `Age` alone.

- Then we fit the multiple linear regression model with both `Age` and $\texttt{FEV}_2$ included.

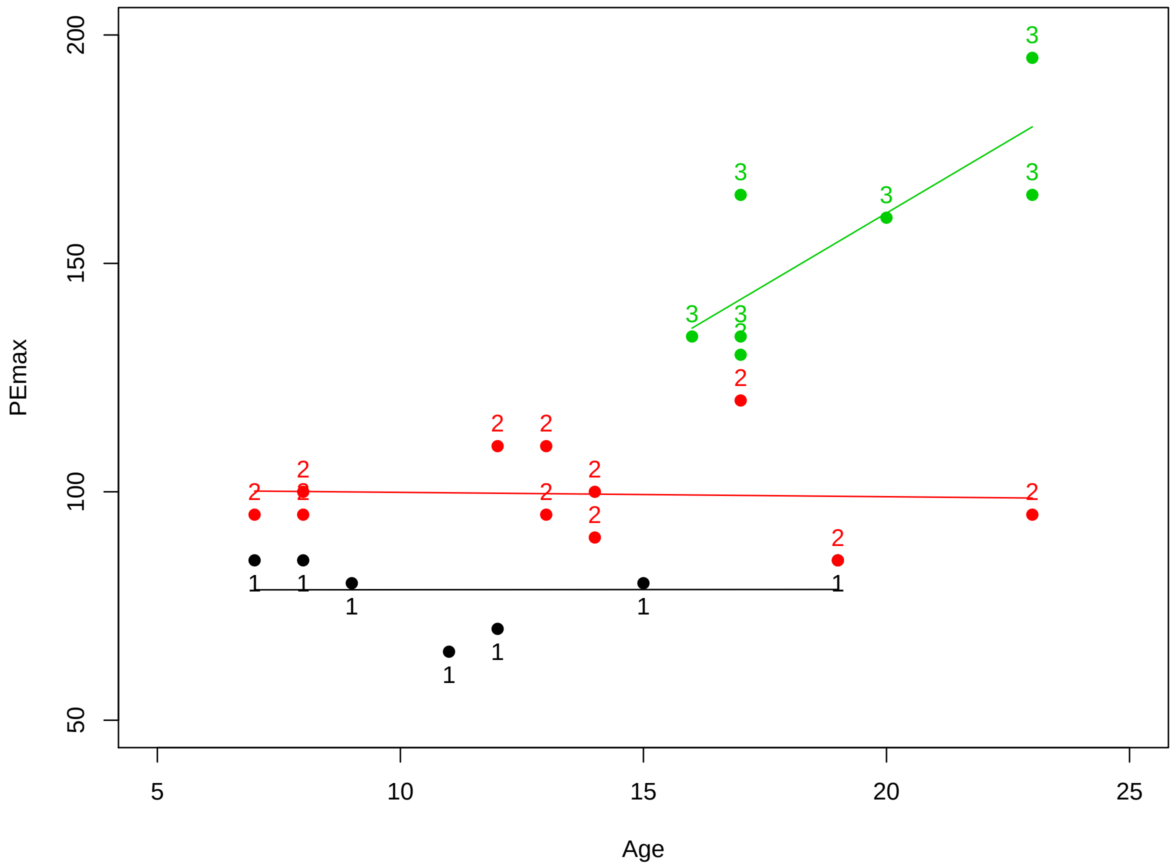
- Assuming that there is no causal relationship between `Age` and $\texttt{FEV}_2$, do we think that $\texttt{FEV}_2$ is a confounder of the relationship between `Age` and `PEmax`, after looking at the output from the two above models? Why?

- What is the expected (or average) `PEmax` score from someone who is `Age` 16 and has $\texttt{FEV}_2$ score of 1? $\texttt{FEV}_2$ score of 2? $\texttt{FEV}_2$ score of 3?

2.3 interactions

- Using `PEmax` as response variable, write out the full model for `Age` and each level of $\texttt{FEV}_2$, as well as interaction terms between `Age` and $\texttt{FEV}_2$.

- We will look at this relationship graphically. What do you notice from the plot?



- We will now fit the model with the interaction terms.

- What do you conclude from the model?

- What is the expected `PEmax` score from someone who is `Age` 16 and has $\texttt{FEV}_2$ score of 1? $\texttt{FEV}_2$ score of 2? $\texttt{FEV}_2$ score of 3?

- How does this compare to your previous estimate?