Module 9: Linear regression

Siyue Yang

06/22/2022

Part 1: Linear regression in R

```
library(tidyverse) #ggplot2, dplyr, etc.
library(reshape2) #need this for melt()
library(knitr) #need this for kable
library(MASS) #contains dataset
```

Load the birthwt data. This data contains 189 observations, 9 predictors, and an outcome, birthweight, available both as a continuous measure and a binary indicator for low birth weight.

```
data(birthwt)
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

1. Plot a scatterplot of birthweight (bwt) and mother's weight (lwt).
2. Use OLS to fit the regression of birthweight on mother's weight.
3. Extract the following: estimated coefficients, standard errors, variance-covariance matrix, and confidence intervals.
4. Plot the regression line and interpret the intercept and slope
5. Does the interpretation of the intercept make sense? How might we change this?
6. Now, we want to fit a model that includes race, mother's age, and smoking status in the model. Race takes on value 1 for white, 2 for black, and 3 for other. Mother's age is continuous. Smoking status is binary. Write out the regression function we may be interested in.
7. Use OLS to calculate the coefficient estimates in this model.
8. Interpret all the coefficient estimates.
9. Print the results in Rmarkdown using kable().

Part 2: Regression model with interaction terms

1. Take a random sample of size 10,000 from the NC dataset to work with for this problem. Make sure everyone in your group uses the same seed, so that you draw the same sample.
2. For this problem, you will be working with the following model where Y is birth weight, X_1 is weight gain during pregnancy and X_2 is smoking. What does β_3 represent? Why might this be of interest?

$$E[Y \mid X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

3. Create a scatter plot of maternal weight gain and birth weight. Color observations according to smoking status.
4. Use the expression $\hat{\beta}$ given in the slides to find the estimates of the coefficients. Note: for this question, you will need to create the "design" matrix, $\mathbf{X}$.
5. Fit this regression using the lm() function. How does this compare with the results from part 4 ?
6. Interpret the coefficients for weight gain and smoker. Be as precise as possible.
7. Plot the regression line for smokers and non-smokers on part 3. Hint: use stat_function() in gpplot and define your own function.
8. Do you see large differences in the slopes of these lines? Which p-value in the regression output formally tests this? Does this align with your expectations?