STATISTICS 452/652: Statistical Learning and Prediction

August 19, 2020

Regression Review Part 3: Multiple Linear Regression

(Reading: ISLR Sections 3.1–3.4)

1 Review of Multiple Linear Regression

- Simple linear regression is rarely used as a final analysis
- Most problems have more than one variable
  - Want to understand all of their effects
  - Want to build regression model
  - Want to know which variables are important
- Model structures is harder to visualize, except when there are only 2 variables
  - Generally referred to as "surfaces"
- Multiple linear regression model is just like simple linear model, but with more variables.
  - Main new element is p explanatory variables, $X_1, \ldots, X_p$
- Standard model is to assume that, for X value, Y originates from an expanded version of the simple linear regression model (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{1}$$

  - $\beta_0, \ldots, \beta_p$ are regression parameters or coefficients
    * β_0 is the intercept (mean value of Y when all $X_j = 0$)
    * β_j, $j = 1, \ldots, p$ are "(partial) regression coefficients"

      - · Change in mean Y for 1-unit change in X_j *holding all other variables constant.*

- Need to estimate the parameters
  - Minimize the LS criterion (2) again, expanded for full model

    $$\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}])^2,$$

    where $x_{i1}, x_{i2}, \ldots, x_{ip}$, $i = 1, \ldots, n$ are the n observed values of the variables $X_1, \ldots, X_p$
- Mean value of Y for given values of $X_1, \ldots, X_p$ is $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$
  - The shape of this function is a P-DIMENSIONAL HYPERPLANE.
  - This is a fancy term for something that reduces to a straight line in every direction
- When $p = 2$ we can visualize this
  - X_1, X_2, Y are points in 3 dimensions, so picture a cube
  - X_1,X_2 form the bottom surface of the cube
  - Y values are points floating above the surface in a cloud
  - Multiple linear regression model will fit a plane (a board) through the points (see example)
- Multiple regression has several challenges that are not faced when $p = 1$
  - When $p > 2$,we can no longer fully visualize the surface
  - We may not be sure that all variables are necessary
    * There is uncertainty about what is the "right" model
  - Multicollinearity—correlated explanatory variables—is an incredible nuisance
    * Can lead to variables having unexpected (and impossible) coefficients as variables "explain overlapping information" in Y
  - Harder to tell when there are problems with the fit of the model
    * Residual analysis is less reliable
- Despite these challenges, we *must* move forward
  - We can't control what the data give us; we just have to develop tools to extract what we can from it

Example: Prostate data (L2-ProstateData.R) Let's look at a 3D plot of the three variables we looked at in the previous example, `lpsa` vs. `lcavol` and `pgg45`. This is best done live with the program and code.

2 Exercises (Due Friday of Next Week)

Refer to the Air Quality data available in R as the data frame "airquality". Run `help(airquality)` to learn a little more about this data set. We will treat `Ozone` as the response variable and use `Temp`, `Wind`, and `Solar.R` as explanatory. We won't use `Month` or `Day`.

1. Create a separate data frame for these data containing only the variables we will need. You can use something like
   `AQ = airquality[,1:4]`.
   Then create a scatterplot matrix of these four variables. Comment on
   (a) Relationships of each X with Y
   (b) Relationships among the three explanatories.
2. Run separate simple linear regressions of `Ozone` against each explanatory variable.
   (a) Report the three slopes and t-values in a table.
   (b) Make three separate scatterplots and add the respective regression lines to each plot. Present the plots and comment on how well the lines seem to fit each variable.
3. Make a 3D plot of `Ozone` against temperature and wind speed. Rotate it around and notice to yourself what relationship the ozone might have jointly with temperature and wind. Take a screenshot from any angle you think helps you to see most of this relationship. No comments are needed.
4. Fit the multiple linear regression that corresponds to this 3D plot.
   (a) Report the slopes and t-values. Are they much different from when they were computed in simple linear regressions?
   (b) Add the plane surface to the 3D plot. Rotate it around and comment on the quality of the fit. Show a screenshot from some angle that helps to support your comment