Department of Statistics, The Chinese University of Hong Kong
STAT 5102 Regression in Practice (Term 1, 2018–19)

Assignment 2 · due on 29th October 2018 (Mo)

1. [Berkeley Guidance Study] The Berkeley Guidance Study enrolled children born in Berkeley, California, between January 1928 and June 1929, and then measured them periodically until age eighteen (Tuddenham and Snyder, 1954). The data we use is described in Table 1, and the data is given in the data files `BGSgirls.txt` for girls only, and `BGSall.txt` for boys and girls combined.

Table 1: Variable Definitions for the Berkeley Guidance Study in the Files

Variable	Description
Sex	0 for males, 1 for females
WT2	Age 2 weight, kg
HT2	Age 2 height, cm
WT9	Age 9 weight, kg
HT9	Age 9 height, cm
LG9	Age 9 leg circumference, cm
ST9	Age 9 strength, kg
WT18	Age 18 weight, kg
HT18	Age 18 height, cm
LG18	Age 18 leg circumference, cm
ST18	Age 18 strength, kg
Soma	Somatotype, a scale from 1, very thin, to 7, obese, of body type

   (a) For the girls only, draw the scatterplot matrix of all the age two variables, all the age nine variables and *Soma*. Write a summary of the information in this scatterplot matrix. Also obtain the matrix of sample correlations between the height variables.

   (b) Fit the multiple linear regression model with mean function

$$E(Soma \mid X) = \beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 HT9 + \beta_4 WT9 + \beta_5 ST9$$

   based on `BGSall.txt`. Find σ^2, R^2, the overall analysis of variance table and overall F-test. Compute the t-statistics to be used to test each of the β_j to be zero against two-sided alternatives. Explicitly state the hypotheses tested and the conclusions.

2. [A Genetic Study] Seeds sampled from trees in the eastern US and Canada were planted in a genetic study. The time of leafing out of these seedlings can be related to the latitude and mean July temperature of the place of origin of the seed. The variables are $X_1 =$ latitude, $X_2 =$ July mean temperature, and $Y =$ weighted mean index of leafing out time. (Y is a measure of the degree to which the leafing out process has occurred. A high value is indicative that the leafing out process is well advanced.) The data is below and in the file `maple.txt` on Blackboard.

   (a) Find the regression of `LeafIndex` on `Latitude`. Is latitude a useful predictor of leaf index?

   (b) Repeat part (a) for the regression of `LeafIndex` on `JulyTemp`.

(c) Find the regression of `LeafIndex` on `Latitude` and `JulyTemp`. Compare the results of this analysis with your results from (a) and (b). How different are the slope coefficients in each case. What best explains the differences in their values?

(d) Find ANOVA tables for the model in part (a) ($\texttt{LeafIndex} = \beta_0 + \beta_1 \texttt{Latitude} + \epsilon$) and the model in part (c) ($\texttt{LeafIndex} = \beta_0 + \beta_1 \texttt{ Latitude} + \beta_2 \texttt{ JulyTemp} + \epsilon$). What parts of the row of the ANOVA table corresponding to `Latitude` are the same and what parts are different? To what formal hypothesis test does the p-value in the `Latitude` row of each ANOVA table correspond? Why are the p-values different?