Homework 4

Due on Tuesday, Nov. 7th by 11:00 am

You are expected to solve all the four problems by yourselves. Discussions within your group are encouraged ***but you must write out your own answers in your own words****. Duplicate homework will not receive credit. Make sure you show all of your work and attach your R-script for full credit. The datasets for the problems can be downloaded from Canvas. Please turn in your homework report right before the lecture time.*

1. (20 points) True or False. The following statements are based on a multiple linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

Let $(\hat{\beta}_0, ..., \hat{\beta}_p)$ be the least square estimators for the model.

(a) The least square estimator of the slope coefficient in the added-variable plot for X_k is equal to $\hat{\beta}_k$.

(b) Before applying inverse response plots, we only need to check whether each of $X_1, ...X_p$ is symmetric.

(c) If the two fitted curves in the marginal model plot do not agree with each other, our model is not a valid model.

(d) If a model has multicollinearity issue, we should drop all insignificant predictors that have large variance inflation factors in a single step to obtain a reduced model.

2. (10 points) Describe how to construct an added-variable plot for a multiple regression model.

3. (10 points) Researchers are interested in studying how to **maintain weight loss**. Based on a survey of almost 3000 adults, researchers Wyatt et al. (Obesity Research, 2002) reported that those who ate breakfast regularly tended to be more successful at maintaining their weight loss. Based on this study, could we conclude “eating breakfast regularly” and “maintaining weight loss” are in the “cause-and-effect" relationship? Give reasons to support your answer.

4. (60 points) Sec. 6.7, Problem 5 in textbook, part (a), (b) and

(c) The model developed in part (b) shows many predictors are not significant given all others constant. We are considering dropping some of them to get a reduced model. First, please check whether there is a “multicolinearity” issue in the model (list all the VIF to answer this question).

(d) Consider to drop the nonsignificnat predictor with highest VIF from the full model. Write down the reduced model and conclude which model you prefer using a partial F-test.

(e) You may find the reduced model above still has many insignificant predictors. Here I propose another reduced model, which is

$$\log(Y) = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + \beta_6 X_6 + \epsilon.(e)$$

Which reduced model you preferred? The one above or the one in part (d)? (Applying the partial F test.)

(f) Draw the added variable plots for model (e). If there are any points that influentially affect the slope coefficient estimators $\hat{\beta}_2$, $\hat{\beta}_4$ and $\hat{\beta}_6$, please circle them.

(g) Draw the marginal model plots for model (e). Do they show that model (e) adequately model the mean of $\log(Y)$?