# Multiple Linear Regression Diagnostics

*Due:* Monday 10/04 (11.00PM)
*Submission:* On Gradescope

The Homework contains two parts:
Part I consists of practice problems that you can work on to practice; you do not need to submit these. Some of these will be discussed during Thursday's office hours. Part II consists of the problems that you have to submit. Use R and R Markdown as necessary and submit your solutions as a PDF or HTML file.

---

## Part I: Practice Questions

1. **Grocery Retailer**
   A large national grocery retailer tracks productivity and cost of its facilities closely. The data in the `grocery.txt` file were obtained from a single distribution center for a one year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped ($X_1$), the indirect costs of the total labor hours as a percentage ($X_2$), a qualitative predictor called holiday that is called in one of the week has a holy day and zero otherwise ($X_3$), and the total labor hours ($Y$).

   (a) Check the constant variance assumption.

   (b) Check the normality assumption.

   (c) Check for the structure of the relationship between the predictors and the response.

2. Use the `teengamb` data from the *faraway* library to fit a model with gamble as the response and the other variables as predictors.

   (a) Check the constant variance assumption.

   (b) Check the normality assumption.

   (c) Check for the structure of the relationship between the predictors and the response.

3. Consider the `prostate` data from the *faraway* library. Fit a model with `lpsa` as the response and the other variables as predictors.

   (a) Compute and comment on the condition numbers.

   (b) Compute and comment on the correlations between predictors.

   (c) Compute the variance inflation factors.

4. Consider the `cheddar` data from the *faraway* library. Fit a model with taste as the response and the other variables as predictors.

   (a) Check the model assumptions.

   (b) Is any transformation of the predictors suggested?

(c) Use the Box-Cox method to determine an optimal transformation of the response. Would it be reasonable to leave the response untransformed?

(d) Use the optimal transformation of the response and refit the additive model. Does this make any difference to the transformations suggested for the predictors?

# Part II: Homework Questions – to be submitted

1. If $n = p$ and the $\mathbf{X}$ matrix is invertible, show that the hat matrix $\mathbf{H}$ is given by the $p \times p$ identity matrix. In this case, what are $h_{ii}$ and $\hat{Y}_i$?

2. The `whitewines.csv` data set contains information related to white variants of the Portuguese "Vinho Verde" wine. Specifically, we have recorded the following information:

   (a) `fixed acidity`, (b) `volatile acidity`, (c) `citric acid` , (d) `residual sugar`, (e) `chlorides` , (f) `free sulfur dioxide`, (g) `total sulfur dioxide`, (h) `density`, (i) `pH`, (j) `sulphates`, (k) `alcohol`, (l) `quality` (score between 0 and 10)

   In this homework, our goal is to explain the relationship between `alcohol level` (dependent variable) and `residual sugar`, `pH`, `density` and `fixed acidity`.

   (a) Check the constant variance assumption.

   (b) Check the normality assumption.

   (c) Check for the structure of the relationship between the predictors and the response.

   (d) Is any transformation of the predictors suggested?

   (e) Use the Box-Cox method to determine an optimal transformation of the response. Would it be reasonable to leave the response untransformed?

   (f) Use the optimal transformation of the response and refit the additive model. Does this make any difference to the transformations suggested for the predictors?