Assignment 4 — Due November 8, 2018

1. `online submission` Patient satisfaction: A hospital administrator wished to study the relation between patient satisfaction (Y, an index) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3, an index). The larger values of Y, X_2, and X_3 are associated with more satisfaction, increased severity of illness, and more anxiety, respectively. The administrator randomly selected 46 patients and collected data given in the data file `patient.txt`. Remember to interpret the results in the context of the study.

   (a) Obtain the scatter plot matrix and the correlation matrix. State your key findings.

   (b) State the model underlying a multiple linear regression of patient satisfaction on age, severity of illness, and anxiety level. We will consider this model to be the full model. What is the underlying population or the set of underlying populations?

   (c) Obtain the least squares estimates of the regression coefficients.

   (d) Test whether severity of illness has any effect on patient satisfaction, while controlling for patient's age and anxiety level.

   (e) Provide a 95% confidence intervals for the regression coefficient corresponding to severity of illness in the full model.

   (f) Provide an unbiased estimate for the error variance and a 95% confidence interval.

   (g) Obtain a point estimate and a 95% confidence interval for the mean satisfaction of patients when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$.

   (h) Obtain a point predictor and a 95% prediction interval for a new patient's satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$.

   (i) Obtain the coefficient of multiple determination R^2.

   (j) Perform an overall F-test for all predictors in the model. State the assumptions for the F-test and perform model diagnostics.

2. `online submission` Continue to work on the above dataset (patient satisfaction).

   (a) Fit a model with severity of illness (X_2) only. Test whether the regression coefficient for X_2 is zero or not and provide a 95% confidence interval for this regression coefficient. Compare the results with Problem 5(f) and 5(g) in Assignment 4 and explain the results.

   (b) Compute the VIF values for the explanatory variables in the full model with X_1, X_2, X_3.

   (c) Does SSR(X_1) equal SSR($X_1|X_3$)? Does SSR(X_2) equal SSR($X_2|X_3$)? Explain the results.

   (d) Use the best subsets methods with various model selection criteria (R^2, R_a^2, C_p) and 1 best model for each model size (i.e., set `nbest=1` in R function `my.regsub`).

   (e) Repeat (d) but with `nbest=4`. Briefly comment on the choice of `nbest`. How important is it to change `nbest` from 1 to 4?

   (f) Perform a backward elimination with AIC as the model selection criterion in R function `step`.

   (g) Repeat (f) but with BIC.

3. Consider the one-sample problem: $Y_i \sim N(\mu, \sigma^2), 1 \le i \le n$ with the Y_i's i.i.d. The MLE is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

   If we constrain $|\mu|^2 \le C$ and transform the problem to a penalize minimization problem we had to solve

$$\hat{\mu}_\lambda = \arg\min_{\mu} \sum_{i=1}^{n} (Y_i - \mu)^2 + \lambda\mu^2.$$

(a) Find a design matrix X such that
$$\hat{\mu} = (X^t X)^{-1} X^t Y.$$

(b) Show that
$$\hat{\mu}_\lambda = \frac{\hat{\mu}}{1 + \lambda/n}.$$

(c) Find a design matrix $X(\lambda)$ and a data vector $Y(\lambda)$ such that
$$\hat{\mu}_\lambda = (X(\lambda)^t X(\lambda))^{-1} X(\lambda)^t Y(\lambda).$$
(Hint: $Y(\lambda)$ will generally have to be of length $n+1$ or greater; i.e., you need to add an observation to the original Y, as well as an entry to the original X.)

(d) Generalize this to the constrained regression problem for a vector of non-negative constraints $\lambda = (\lambda_0, \ldots, \lambda_{p-1})$
$$\hat{\beta}_\lambda = \arg\min_\beta \left(\sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j X_{ij})^2 + \sum_{j=0}^{p-1} \lambda_j \beta_j^2 \right).$$

(e) Write a function in `R` that takes two arguments, one the output of `lm`, the other a vector λ of length p as above and return $\hat{\beta}_\lambda$. (Hint: The function `model.matrix` will likely be useful).

4. `online submission` A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. It is thought that the two most important variables affecting delivery time (Y) are the number of cases of product stocked (X_1) and the distance walked by the route driver (X_2). An industrial engineer collected 25 measurements on these three variables. The data are found in Montgomery, Peck, and Vining (2001) and are reproduced in the file `softdrink.txt`

(a) Obtain the scatterplot matrix and the correlation matrix for these three variables. State your key findings.

(b) Fit the linear regression model Y on the two explanatory variables. Report the R summary table from the fitted model.

(c) Test for an overall regression relation between Y and the two explanatory variables.

(d) Plot the standard (i.e., internally studentized) and the studentized (i.e., externally studentized) residuals versus the fitted values from the model in part (b). State your key findings.

(e) For the model in part (b) compute and investigate various regression diagnostics including the leverage values, DFFITS, Cook's distance, and DFBETAS. State your key findings.

(f) Based on your results in part (e), select a final model. Provide the R summary table and discuss the differences between the results here and the results in part (b).