STAT 331: Assignment 4

Due "Officially": Tuesday July 30, 2019 by 11:59pm

Due "Unofficially": Wednesday July 31, 2019 by 11:59pm

Regarding the submission deadline: I cannot technically assign something and have it due after classes have finished. Thus, the "official" deadline for this assignment is Tuesday July 30th at 11:59pm. However, you may, if you wish, submit your assignment by 11:59pm on Wednesday July 31st with no penalty. You may also, if you wish, submit your assignment by 11:59pm on Thursday August 1st with a 50% penalty.

Submission

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically via *Crowdmark*. This means that your responses for different questions should be on separate pages.

For non-computational questions, you may upload solutions produced by LaTeX (or some other equation editing software), or handwritten solutions that have been scanned or photographed. Please note that if you scan or photograph your solutions it is your duty to ensure they are readable and of high quality. If the marking team cannot read your solutions, they simply will not be marked.

For computational questions, I highly recommend you produce your solutions as a nicely formatted .pdf file with R Markdown. R Markdown facilitates the seamless combination of written text, R code and output, and LaTeX equations. Your submission for these questions should include the code, the corresponding output, and the interpretations where appropriate.

Question 1 [5 points]

Consider the linear regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where X is the usual $n \times (p+1)$ matrix. Define $H = X(X^TX)^{-1}X^T$ to be the symmetric ($H = H^T$) and idempotent ($H = HH$) "hat" matrix.

(a) [2] Show that the trace of H (the sum of its diagonal elements) is equal to $p+1$.
[*Hint:* $\text{tr}(AB) = \text{tr}(BA)$]

(b) [1] Using the result from part (a), show that the average leverage is $(p+1)/n$.

(c) [2] Show that the diagonal elements of H (i.e., leverages) must lie between zero and one. In other words, show that $0 \leq h_{ii} \leq 1$.
[*Hint: Consider the ith diagonal element of HH and recognize that it is equivalent to the ith diagonal element of H due to idempotency.*]

Question 2 [5 points]
Consider the linear regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and the least squares estimate $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$. Let $X_{(i)}$ denote the X matrix with the i^{th} row $\boldsymbol{x}_i^T$ deleted, and let $\boldsymbol{y}_{(i)}$ and $\boldsymbol{\epsilon}_{(i)}$ denote respectively the vectors $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ without the i^{th} element. Note that without loss of generality we can write

$$X = \begin{bmatrix} X_{(i)} \\ \boldsymbol{x}_i^T \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_{(i)} \\ y_i \end{bmatrix} \qquad (1)$$

Without the i^{th} observation the model and the estimate of $\boldsymbol{\beta}$ become

$$\boldsymbol{y}_{(i)} = X_{(i)}\boldsymbol{\beta}_{(i)} + \boldsymbol{\epsilon}_{(i)} \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_{(i)} = \left(X_{(i)}{}^T X_{(i)}\right)^{-1} X_{(i)}{}^T \boldsymbol{y}_{(i)} \qquad (2)$$

From equation (1) we can write

$$X^T X = X_{(i)}{}^T X_{(i)} + \boldsymbol{x}_i \boldsymbol{x}_i^T \quad \text{and} \quad X^T \boldsymbol{y} = X_{(i)}{}^T \boldsymbol{y}_{(i)} + \boldsymbol{x}_i y_i \qquad (3)$$

Using the relationships given in equations (2) and (3) show

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \frac{(X^T X)^{-1} \boldsymbol{x}_i e_i}{(1 - h_{ii})}$$

Hint: Suppose that $\boldsymbol{u}$ and $\boldsymbol{v}$ are $n \times 1$ vectors, and A is an $n \times n$ invertible matrix. Then

$$(A - \boldsymbol{u}\boldsymbol{v}^T)^{-1} = A^{-1} + \frac{A^{-1}\boldsymbol{u}\boldsymbol{v}^T A^{-1}}{(1 - \boldsymbol{v}^T A^{-1} \boldsymbol{u})}$$

Question 3 [30 points]
In this question you will consider the forestry dataset found in the file `forestry.csv`. This dataset records the needle area and two other variables for 35 coniferous trees. Specifically, the dataset contains observations from the following variables:

- `area`: the total needle area of the tree
- `height`: the height of the tree
- `caliper`: a measure of the tree's trunk size

Interest lies developing a model that relates a tree's total needle area to its caliper and its height.

(a) [4] Fit a multiple linear regression model relating `area` to the two explanatory variables listed above and construct the following residual plots:
  i. Studentized Residuals vs. Index
  ii. Studentized Residuals vs. Fitted Values
  iii. Histogram of Studentized Residuals
  iv. QQ-plot of Studentized Residuals

(b) [4] Based on the plots in (a), answer “Yes” or “No” to the following questions and give a one sentence justification.

   i. Do the residuals appear to be independent?
   ii. Do the residuals appear to have constant variance?
   iii. Do the residuals appear to be normally distributed?
   iv. Do the residuals suggest the existence of an outlier?

(c) [1] Which observation has the largest Studentized residual?

(d) [3] Calculate the leverage for each observation and construct a plot of them vs. their index. Which observations have ‘high’ leverage (i.e., leverage larger than twice the average leverage)?

(e) [3] Calculate Cook’s D-statistic for each observation and construct a plot of them vs. their index. List the top three most influential points.

(f) [2] Fit a multiple linear regression model relating `area` to the two explanatory variables but with observations 10 and 29 deleted and construct a plot of the Studentized residuals vs. the fitted values.

(g) [2] Consider the plot from (f). Do these residuals appear to have constant variance? Answer “Yes” or “No” with a one sentence justification.

(h) Fit the three models listed below (to the data with observations 10 and 29 deleted), and for each model construct a plot of the Studentized residuals vs the fitted values and state “Yes” or “No”, indicating whether the transformation has stabilized the variability of the residuals relative to what was observed in (g).

   i. [3] The full regression model where the response variables has been log-transformed.
   ii. [3] The full regression model where the response variables has been square root-transformed.
   iii. [3] The full regression model where the response variable has been Box-Cox transformed. Be sure to state the optimal value of λ.

(i) [2] For interpretability purposes the model with the log-transformed response is the optimal model from (h). Interpret e^{β_1} and e^{β_2}.

Question 4 [15 points]

In this question you will return to the baseball dataset found in the file `hitters.csv`. Recall, this dataset records the salary of $n = 263$ Major League Baseball players during the 1987 season as well as $q = 19$ statistics associated with the performance of each player during the previous season. Specifically, the dataset contains observations from the following variables:

- `AtBat`: Number of times at bat in 1986
- `Hits`: Number of hits in 1986
- `HmRun`: Number of home runs in 1986
- `Runs`: Number of runs in 1986
- `RBI`: Number of runs batted in in 1986
- `Walks`: Number of walks in 1986
- `Years`: Number of years in the major leagues
- `CAtBat`: Number of times at bat during his career
- `CHits`: Number of hits during his career
- `CHmRun`: Number of home runs during his career
- `CRuns`: Number of runs during his career
- `CRBI`: Number of runs batted in during his career
- `CWalks`: Number of walks during his career
- `League`: A categorical variable with levels `A` (for American) and `N` (for National) indicating the player's league at the end of 1986
- `Division`: A factor with levels `E` (for East) and `W` (for West) indicating the player's division at the end of 1986
- `PutOuts`: Number of put outs in 1986
- `Assists`: Number of assists in 1986
- `Errors`: Number of errors in 1986
- `Salary`: 1987 annual salary on opening day in thousands of dollars
- `NewLeague`: A factor with levels `A` and `N` indicating the player's league at the beginning of 1987

As before, interest lies in developing a model that relates a player's annual salary to their previous performance.

(a) [3] Calculate the variance inflation factor (VIF) for each of the explanatory variables. Comment on whether multicollinearity appears to be an issue. If it is, identify the three explanatory variables that are most seriously affected by the issue.

(b) [3] Using the *all-possible-subsets* approach, find the model that best fits the observed data. This procedure may be automated using the `regsubsets()` function (using BIC as the decision criteria), but you must explain in your own words how this algorithm identifies the 'best' model.

(c) [3] Find the optimal model according to the *forward-stepwise-selection* approach using BIC as the decision criteria. This procedure may be automated using the `stepAIC()` function, but you must record in a table which explanatory variable is added at each step and also the associated decrease in BIC that results.

(d) [3] Find the optimal model according to the *backward-stepwise-elimination* approach using BIC as the decision criteria. This procedure may be automated using the `stepAIC()` function, but you must record in a table which explanatory variable is removed at each step and also the associated decrease in BIC that results.

(e) [3] Find the optimal model according to the *hybrid-stepwise-selection* approach using BIC as the decision criteria. This procedure may be automated using the `stepAIC()` function, but you must record in a table which explanatory variable is added or eliminated at each step and also the associated decrease in BIC that results.