STAT 331: Assignment 2

Due: Wednesday June 12, 2019 by 11:59pm

Derivation Component [30 points]

1. [5 points] Consider the regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where each ε_i is i.i.d. $N(0, \sigma^2)$, $i = 1, \ldots, n$. The least squares estimates of β_0 and β_1 are given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\beta}_1 = \frac{\hat{\rho} s_y}{s_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The model above can also be written in matrix notation as $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. By appropriately defining the matrix X, and the response vector $\boldsymbol{y}$, show that the least squares estimate $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ provides estimates equivalent to the ones shown above.

2. [10 points] Let $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$ with $\mathrm{E}[\boldsymbol{y}] = \begin{bmatrix} 2 \\ 6 \\ 4 \end{bmatrix}$ and $\mathrm{Var}[\boldsymbol{y}] = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & -1 \\ 1 & -1 & 3 \end{bmatrix}$.

Also define $\boldsymbol{a} = [1 \quad 2 \quad -1]$ and $A = \begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & -1 \\ 2 & 1 & 1 \end{bmatrix}$.

(a) [1] Find $\mathrm{E}[y_2]$.
(b) [1] Find $\mathrm{Var}[y_1]$.
(c) [1] Find $\mathrm{Cov}[y_1, y_2]$.
(d) [1] Find $\mathrm{Corr}[y_1, y_3]$.
(e) [1] Find $\mathrm{E}[y_2 - y_3]$.
(f) [1] Find $\mathrm{Var}[y_2 - y_3]$.
(g) [1] Find $\mathrm{E}[\boldsymbol{ay}]$.
(h) [1] Find $\mathrm{Var}[\boldsymbol{ay}]$.
(i) [1] Find $\mathrm{E}[A\boldsymbol{y}]$.
(j) [1] Find $\mathrm{Var}[A\boldsymbol{y}]$.

3. [10 points] In the context of the linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon} \sim \text{MVN}(\boldsymbol{0}, \sigma^2 I)$ we've seen that

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} \sim \text{MVN}(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

and that

$$\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\mu}, \sigma^2 H)$$

and

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{\mu}} \sim \text{MVN}\big(\boldsymbol{0}, \sigma^2 (I - H)\big)$$

where $H = X(X^T X)^{-1} X^T$ is the "hat" matrix. In class we have seen that $\boldsymbol{e}$ and $\widehat{\boldsymbol{\mu}}$ are orthogonal vectors. Your task here is to show that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\mu}}$ are also statistically independent random vectors. Specifically, perform the following tasks:

(a) [5] Find the distribution of the $(n + n) \times 1$ random vector $(\widehat{\boldsymbol{\mu}}, \boldsymbol{e})^T$.

(b) [5] Show that $\widehat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$ are independent random vectors.

Hint:
Let $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}} = H\boldsymbol{y}$ and $\boldsymbol{e} = (I - H)\boldsymbol{y}$ and consider $P\boldsymbol{y}$ where P is the $(n + n) \times n$ matrix formed by stacking H and $I - H$ on top of one another:

$$P = \begin{bmatrix} H \\ --- \\ I - H \end{bmatrix}$$

4. [5 points] Show that the least squares estimator $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1}$ is an *unbiased* estimator. Note that you may use, without proof or derivation, the fact that $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-p-1)}$.

Computation Component [58 points]

5. [26 points] The 2010 FIFA World Cup held in South Africa from June 11 – July 11, 2010 involved 595 players from 32 teams. The dataset found in the file `worldcup.csv` contains information on each of these 595 players. In particular, for each player, the dataset records their:
   - `Name`: the player's name
   - `Team`: the player's team
   - `Position`: the player's position (1 = Defender, 2 = Forward, 3 = Goalkeeper, 4 = Midfielder)
   - `Time`: the number of minutes played during the World Cup tournament
   - `Shots`: the number of shots taken during the World Cup tournament
   - `Passes`: the number of passes made during the World Cup tournament
   - `Tackles`: the number of tackles made during the World Cup tournament
   - `Saves`: the number of goals saved during the World Cup tournament

   Interest lies in quantifying the relationship between some of these quantitative measures and a player's position.

   (a) [2] Load the data into R and coerce the variable `Position` to be a factor variable with levels `Defender`, `Forward`, `Goalkeeper`, and `Midfielder` (instead of the of the default levels 1, 2, 3, 4).

   (b) [3] Construct boxplots (side-by-side) of `Time` versus `Position`. Be sure to include a title and label your axes. In one sentence describe the general relationship you see between these two variables.

   (c) [5] Fit a linear regression model in which `Time` is treated as the response variable and `Position` is treated as the explanatory variable. Interpret the values of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ in the context of the dataset.

   (d) [4] In the context of the model from part (c), provide a point estimate and a 95% confidence interval for the expected number of minutes played in each position.

   (e) [3] Construct boxplots (side-by-side) of `Shots` versus `Position`. Be sure to include a title and label your axes. In one sentence describe the general relationship you see between these two variables.

   (f) [5] Fit a linear regression model in which `Shots` is treated as the response variable and `Position` is treated as the explanatory variable. Interpret the values of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ in the context of the dataset.

   (g) [4] In the context of the model from part (e), provide a point estimate and a 95% confidence interval for the expected number of shots taken in each position.

6. [32 points] In this question you will consider the Boston housing dataset found in the file `boston.csv`. This dataset records the median house value and several other variables for 506 neighbourhoods around Boston in 1978. Specifically, the dataset contains observations from the following variables:
   - `crim`: per capita crime rate
   - `zn`: proportion of residential land zoned for lots over 25,000 sq. ft.
   - `indus`: proportion of non-retail business acreage
   - `chas`: dummy variable indicating whether the neighbourhood is adjacent to the Charles River (1) or not (0)
   - `nox`: nitrogen oxide pollution concentration (parts per 10 million)
   - `rm`: average number of rooms per house
   - `age`: proportion of owner-occupied houses built prior to 1940
   - `dis`: weighted mean of distance to five Boston employment centers
   - `rad`: index of accessibility to radial highways
   - `tax`: property tax rate per $10,000
   - `ptratio`: pupil-teacher ratio
   - `lstat`: percent of population that have a “low” socioeconomic status
   - `medv`: median value of owner-occupied houses in $1000s

Interest lies in understanding which factors influence the value of a house – and how they influence the value of a house. Interest also lies in predicting the value of a house in a neighbourhood with specific traits.

(a) [6] Construct two plots, each containing 6 scatterplots (3×2 layout), which visualize the relationship between `medv` and each of the 12 explanatory variables. Be sure to include titles and axis labels in each of these plots.

(b) [4] Calculate the correlation coefficient between `medv` and each of the 12 explanatory variables. On the basis of these correlation coefficients, identify the two variables most strongly associated with `medv`.

(c) [4] Fit a linear regression model relating `medv` to the twelve explanatory variables listed above. Calculate the test statistic and *p*-value associated with the hypothesis $\mathrm{H}_0\colon \beta_{\text{indus}} = 0$ vs. $\mathrm{H}_\mathrm{A}\colon \beta_{\text{indus}} \neq 0$. At a 5% significance level, determine whether median house value depends significantly on the proportion of non-retail business acreage.

(d) [4] Fit a reduced linear regression model relating `medv` to all of the explanatory variables listed above – except `indus`. Calculate the test statistic and *p*-value associated with the hypothesis $\mathrm{H}_0\colon \beta_{\text{age}} = 0$ vs. $\mathrm{H}_\mathrm{A}\colon \beta_{\text{age}} \neq 0$. At a 5% significance level, determine whether median house value depends significantly on the proportion of owner-occupied houses built prior to 1940.

(e) [11] Fit a reduced linear regression model relating `medv` to all of the explanatory variables listed above – except `indus` and `age`. In the context of this model, interpret each of the regression coefficients (i.e., all of the $\hat{\beta}$'s).

(f) [1] Based on results of the reduced model (without `indus` and `age`), rank the remaining 10 explanatory variables in terms of the strength of their relationship with `medv`, from most weakly associated to most strongly associated.

(g) [2] In the context of the reduced model (without `indus` and `age`), provide a point prediction and a 95% prediction interval for the median house value in a neighbourhood for which

- `crim` = 5
- `zn` = 25
- `chas` = 0
- `nox` = 0.6
- `rm` = 3
- `dis` = 4
- `rad` = 10
- `tax` = 500
- `ptratio` = 20
- `lstat` = 5

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 the derivation 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 the 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.