Assignment 1 - CSC/DSC 265/465 - Spring 2020 - Due February 27, 2020

Unless otherwise specified, statistical significance can be taken to hold when the relevant P-value is no larger than $\alpha = 0.05$. Note that problem **Q4** is reserved for graduate students. All questions have equal marks.

Q1: Consider the matrix representation of the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ response vector, $\mathbf{X}$ is a $n \times q$ matrix, $\boldsymbol{\beta}$ is a $q \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of error terms.

(a) Why is it the case that a unique least squares estimate of $\boldsymbol{\beta}$ can only exist if the matrix $\mathbf{X}^T\mathbf{X}$ is invertible?

(b) Suppose we are given paired observations of the form $(x_1, y_1), \ldots, (x_n, y_n)$, where each $x_i \in \{1, 2, 3\}$ is one of three values, and $y_i \sim N(\mu_k, \sigma^2)$ if $x_i = k$. Assume that the responses y_i are independent, and that the variance σ^2 is the same for all responses.

We decide to express this model as a linear regression model by defining three predictors X_1, X_2, X_3, associated with the three outcomes of x_i, using indicator variables, that is,

$$\begin{aligned} X_{i1} &= I\{x_i = 1\}, \\ X_{i2} &= I\{x_i = 2\}, \\ X_{i3} &= I\{x_i = 3\}, \end{aligned}$$

for $i = 1, \ldots, n$. Then suppose we attempt to fit the model

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \;\; i = 1, \ldots, n, \tag{2}$$

where $\epsilon_i \sim N(0, \sigma^2)$. We may express this model in the matrix form of Equation (1). Derive the matrix $\mathbf{X}^T\mathbf{X}$. Is this matrix invertible? **HINT:** Let n_k be the number of times $x_i = k$, for each $k = 1, 2, 3$.

(c) Show that if any of the four terms associated with coefficients $\beta_0, \ldots, \beta_3$ is deleted from Equation (2), then the resulting matrix $\mathbf{X}^T\mathbf{X}$ will be invertible.

(d) In Part (c), four linear regression models are obtained by deleting one of the four terms associated with the coefficients. Show that the least squares fit of each of these will give the same fitted values, and are therefore equivalent.

Q2: For this question, use the `cats` data set from the `MASS` package. This data includes the following observations for each of $n = 144$ cats:

```
Sex
sex: Factor with levels "F" and "M".

Bwt
body weight in kg.

Hwt
heart weight in g.
```

(a) Suppose we have linear relationship $y = \beta_0 + \beta_1 x$ between two variables x, y. If $\beta_1 \neq 0$, this can always be written as $x = \beta_0' + \beta_1' y$. Express β_0' and β_1' as functions of β_0 and β_1.

(b) Fit the following linear models using the `lm()` function:

`Hwt` $\sim$ `Bwt`

and

$$\texttt{Bwt} \sim \texttt{Hwt}.$$

Do the least squares coefficients of the two models conform to the equivalence relationship given in Part (a)? Construct a scatter plot of the `Hwt` and `Bwt` paired observations (place `Hwt` on the vertical axis). For both models superimpose on this plot the estimated linear relationship between `Hwt` and `Bwt`. In each case, ensure that `Hwt` is represented on the vertical axis. Provide a brief explanation for your results.

(c) Fit the following three models (expressed using `R`'s model formula notation):

$$\begin{aligned} \texttt{Hwt} &\sim \texttt{Bwt} \text{ [Model 1]} \\ \texttt{Hwt} &\sim \texttt{Bwt} + \texttt{Sex} \text{ [Model 2]} \\ \texttt{Hwt} &\sim \texttt{Bwt} * \texttt{Sex} \text{ [Model 3]} \end{aligned}$$

For each model construct a scatter plot of `Hwt` and `Bwt` (placee `Hwt` on the vertical axis) and superimpose the estimated regression line (plot separate lines for the two `Sex` classes, and use a legend to identify line associated with each class). Is there statistical evidence at an $\alpha = 0.05$ significance level that either Model 2 or 3 improves Model 1?

Q3: For this question, use the `Insurance` data set from the `MASS` package. This data includes the following observations for each of $n = 64$ insurance companies:

```
District
factor: district of residence of policyholder (1 to 4): 4 is major cities.

Group
an ordered factor: group of car with levels <1 litre, 1{1.5 litre, 1.5{2 litre, >2 litre.

Age
an ordered factor: the age of the insured in 4 groups labelled <25, 25{29, 30{35, >35.

Holders
numbers of policyholders.

Claims
numbers of claims
```

(a) Fit a linear model with response `Claims`, and the remaining variables as predictors. Create a residual plot (residuals against fitted values). Also create a normal quantile plot for the residuals. Do the usual assumptions for linear regression seem reasonable in this case? Comment briefly.

(b) We will try to transform `Claims` using the function $h(x) = \log(x + a)$ (use the natural logarithm). For the standard log-transformation we would set $a = 0$. Why can't we do that here? Repeat Part (a) after replacing response `Claims` with the transformed response $h(\texttt{Claims})$. Use $a = 1$, then $a = 10$. Which succeeds better in normalizing the residuals?

(c) We can, in principal, consider all models using some subset of the original four predictors, including the original four predictors, and no predictors. We can assume all models include an intercept term. How many such models are there.

(d) Create a list in `R` of model formulae representing the collection of models defined in Part (c). Note that we can obtain the full model formula, then remove a predictor from the model with the following code:

```
> fit1 = lm(log(Claims+10) ~ .,data=Insurance)
> full.formula = formula(terms(fit1))
> next.formula = update(full.formula, ~ . -District)
> full.formula
```

```
log(Claims + 10) ~ District + Group + Age + Holders
> next.formula
log(Claims + 10) ~ Group + Age + Holders
>
```

Use this list to calculate R^2_{adj} for each model. Identify the model with the largest R^2_{adj}.

Q4: [**For Graduate Students**] Consider question **Q2**.

(a) Using Model 3, show how to construct a two-sided hypothesis test against null hypothesis

$$H_o : \mu_x^M - \mu_x^F = 0,$$

where μ_x^M, μ_x^F are the mean heart weights of male and female cats of body weights x kg. Construct a plot of the observed t-statistic used in this hypothesis test as a function x, where x ranges from the 0 to 5 in increments of 0.1. Does it appear that the t-statistic is bounded over all x?

(b) What is the P-value for testing the null hypothesis that Model 3 does not improve Model 1? Is there a significant improvement at a $\alpha = 0.1$ significance level? What is the two-sided P-value against $H_o : \mu_{3.5}^M - \mu_{3.5}^F = 0$? If $\mu_{3.5}^M \neq \mu_{3.5}^F$, does this imply Model 1 is incorrect?

(c) For large samples, we may reject simultaneously at a level of significance α (two-sided) all hypotheses

$$H_0 : \boldsymbol{a}^T \boldsymbol{\beta} = 0$$

for which the absolute value of the t-statistic exceeds $(\chi^2_{p;\alpha})^{1/2}$, where $\chi^2_{p;\alpha}$ is the α critical value of a χ^2 distribution with p degrees of freedom, and p is the model degrees of freedom (for example, Cox & Ma (1995) *Biometrics*). What implication does this have for the issue raised in Part (b)?