Assignment 1 - CSC/DSC 265/465 - Spring 2020 - SOLUTIONS

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.

SOLUTION:

(a) The least squares estimates of regression coefficients $\boldsymbol{\beta}$ are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

as the stationary point of the SSE function. This function is convex, so a unique solution exists if and only if a stationary point exists, which can happen if and only if the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ exists.

(b) The design matrix $\mathbf{X}$ has dimension $n \times 4$. Let the element of row i and column j be $x_{i,j}$. The first column consists entirely of ones, that is, $x_{i1} = 1$, $i = 1, \ldots, n$. Columns $j = 2, 3, 4$ represent the indicator variables, so that $x_{ij} = X_{i,j-1} = I\{x_i = j-1\}$, $i = 1, \ldots, n$.

Let a_{jk} be element j, k of matrix $\mathbf{X}^T\mathbf{X}$. Then

$$a_{jk} = \sum_{i=1}^{n} x_{ij} \times x_{ik}.$$

Note that $\mathbf{X}^T\mathbf{X}$ is symmetric, that is, $a_{jk} = a_{kj}$.

The sums of columns $j = 1, 2, 3, 4$ are n, n_1, n_2, n_3, respectively. Then

$$a_{1,1} = \sum_{i=1}^{n} x_{i,1} \times x_{i,1} = \sum_{i=1}^{n} 1 \times 1 = n.$$

For $j = 2, 3, 4$

$$a_{j,1} = a_{1,j} = \sum_{i=1}^{n} x_{i,1} \times x_{i,j} = \sum_{i=1}^{n} 1 \times x_{i,j} = n_{j-1},$$
$$a_{j,j} = \sum_{i=1}^{n} x_{i,j} \times x_{i,j} = \sum_{i=1}^{n} x_{i,j}^2 = \sum_{i=1}^{n} x_{i,j} = n_{j-1},$$

since each entry in the matrix is 0 or 1. Finally, note that columns 2,3,4 are orthogonal, so $a_{2,3} = a_{3,2} = a_{3,4} = a_{4,3} = a_{2,4} = a_{4,2} = 0$. This gives matrix

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & n_1 & n_2 & n_3 \\ n_1 & n_1 & 0 & 0 \\ n_2 & 0 & n_2 & 0 \\ n_3 & 0 & 0 & n_3 \end{bmatrix}.$$

The matrix is not invertible. One way to show this is to use Gaussian elimination, noting that the sum of the last three rows equals the first row, hence $\mathbf{X}^T\mathbf{X}$ cannot be invertible.

(c) If we remove one of the predictors, this is equivalent to removing the corresponding column j from $\mathbf{X}$, to get, say $\mathbf{X}'$. Then we can get $[\mathbf{X}']^T[\mathbf{X}']$ by deleting the jth row and column from $\mathbf{X}^T\mathbf{X}$.

If we delete the β_0 term from Equation (2) we get

$$[\mathbf{X}']^T[\mathbf{X}'] = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix},$$

which is clearly invertible.

If we delete the $\beta_1 X_{i1}$ term from Equation (2), we then delete row $j = 2$ from $\mathbf{X}$ and we get

$$[\mathbf{X}']^T[\mathbf{X}'] = \begin{bmatrix} n & n_2 & n_3 \\ n_2 & n_2 & 0 \\ n_3 & 0 & n_3 \end{bmatrix}.$$

Using Gaussian elimination, subtract rows 2 and 3 from row 1 to get:

$$\begin{bmatrix} n - n_2 - n_3 & 0 & 0 \\ n_2 & n_2 & 0 \\ n_3 & 0 & n_3 \end{bmatrix}.$$

This is a lower triangular matrix with nonzero diagonal entries ($n - n_2 - n_3 = n_1$), and so is invertible (we can reasonably assume $n_1 > 0$, or we would have no need to model this outcome). This implies that $[\mathbf{X}']^T[\mathbf{X}']$ is also invertible.

The argument for any remaining predictor term is identical.

(d) Suppose we delete the β_0 term from Equation (2), to get model :

$$y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i.$$

However, note that the sum of columns $2, 3, 4$ of $\mathbf{X}$ equals column 1. Then we can write

$$\begin{aligned} y_i &= \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \\ &= (\beta_1 - \beta_3 + \beta_3) X_{i1} + (\beta_2 - \beta_3 + \beta_3) X_{i2} + \beta_3 X_{i3} + \epsilon_i \\ &= (\beta_1 - \beta_3) X_{i1} + (\beta_2 - \beta_3) X_{i2} + \beta_3 (X_{i1} + X_{i2} + X_{i3}) + \epsilon_i \\ &= (\beta_1 - \beta_3) X_{i1} + (\beta_2 - \beta_3) X_{i2} + \beta_3 + \epsilon_i. \end{aligned}$$

We can reparametrize $\beta_0^* = \beta_3$, $\beta_1^* = \beta_1 - \beta_3$, $\beta_2^* = \beta_2 - \beta_3$, so that the model can be equivalently expressed

$$y_i = \beta_0^* + \beta_1^* X_{i1} + \beta_2^* X_{i2} + \epsilon_i.$$

Thus the least squares solution to

$$y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

will be equivalent to to that of

$$y_i = \beta_0^* + \beta_1^* X_{i1} + \beta_2^* X_{i2} + \epsilon_i.$$

The same argument will apply to the deletion of any other column.

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:

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

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?

SOLUTION:

(a) If we have $y = \beta_0 + \beta_1 x$, this may be rewritten

$$x = \frac{y}{\beta_1} - \frac{\beta_0}{\beta_1} = \beta_0' + \beta_1' y,$$

where $\beta_0' = -\beta_0/\beta_1$ and $\beta_1' = 1/\beta_1$.

(b) The following `R` code can be used to answer Part (b). See Figure 1.

```
> library(MASS)
>
> ######
> ###### Q2 (b)
> ######
>
> ### Plot data Hwt vs Bwt
>
> plot(Hwt~Bwt,data=cats)
>
> ### Calculate both fits
>
> fit1 = lm(Hwt~Bwt,data=cats)
> fit2 = lm(Bwt~Hwt,data=cats)
>
> ### Copy coefficients
>
> cf1 = fit1$coefficients
> cf2 = fit2$coefficients
>
> ### Coefficients for the Hwt~Bwt model
>
> c(cf1)
(Intercept)         Bwt
 -0.3566624   4.0340627
>
> ### Invert Hwt~Bwt model
>
> c(-cf1[1]/cf1[2],1/cf1[2])
(Intercept)         Bwt
 0.08841271  0.24788906
>
> ### Coefficients for the Bwt~Hwt model
>
> c(cf2)
(Intercept)         Hwt
  1.0196367   0.1602902
>
```

```
> ### Plot regression line for Hwt~Bwt model
>
> abline(cf1)
>
> ### Plot regression line for Bwt~Wwt model,
> ### after inverting fuction.
> ### Use dashed lines.
>
> abline(-cf2[1]/cf2[2],1/cf2[2],lty=2)
>
> ### Create legend
>
> legend(’topleft’,legend=c(’Hwt~Bwt model’,’Bwt~Hwt model’),lty=c(1,2))
>
```

The fitted coefficients for the `Hwt` $\sim$ `Bwt` model are $(\hat{\beta}_0, \hat{\beta}_1) = (-0.3566624, 4.0340627)$. If we use the inversion formula of Part (a) we get

$$\hat{\beta}'_0 = -\hat{\beta}_0/\hat{\beta}_1 = 0.08841271$$
$$\hat{\beta}'_1 = 1/\hat{\beta}_1 = 0.24788906$$

(these values are calculated by the code). However, the fitted coefficients for the `Bwt` $\sim$ `Hwt` fit are $(\hat{\beta}^*_0, \hat{\beta}^*_1) = (1.0196367, 0.1602902)$, which are not equal to the transformed coefficients $(\hat{\beta}'_0, \hat{\beta}'_1)$. In addition, the two fitted models `Hwt` $\sim$ `Bwt` and `Bwt` $\sim$ `Hwt` shown in Figure 1, after suitable transformations, are clearly not equal.

Recall that least squares regression minimizes the total squared *vertical* distance over all pairs (x_i, y_i) from the *response* y_i to the regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$. If x and y are exchanged, then the new least squares fit is equivalent to minimizing the total squared *horizontal* distance over all pairs (x_i, y_i) from the *predictor variable* x_i to the regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$. These are two distinct optimization problems, and we should not expect them to infer the same relationship between x and y.

(c) The following `R` code can be used to plot the required graphs for Part (c). See Figure 2.

```
par(mfrow=c(2,2))

### Plot model 1

plot(Hwt~Bwt,data=cats,col=2+(Sex=="M"))
legend(’topleft’,legend=c(’Female’,’Male’),pch=1,col=2:3)

cf = fit1$coefficients
abline(cf,col=’black’)

### Plot model 2

plot(Hwt~Bwt,data=cats,col=2+(Sex=="M"))
```

```
legend('topleft',legend=c('Female','Male'),lty=1,pch=1,col=2:3)

# Here, we only use the observed range of Bwt for each sex

cf = fit2$coefficients

x = seq(min(cats$Bwt[cats$Sex=="F"]),max(cats$Bwt[cats$Sex=="F"]),0.1)
lines(x, cf[1]+cf[2]*x,col='red')
x = seq(min(cats$Bwt[cats$Sex=="M"]),max(cats$Bwt[cats$Sex=="M"]),0.1)
lines(x, cf[1]+cf[3]+cf[2]*x,col='green')


### Plot model 3

plot(Hwt~Bwt,data=cats,col=2+(Sex=="M"))
legend('topleft',legend=c('Female','Male'),lty=1,pch=1,col=2:3)

# Here, we only use the observed range of Bwt for each sex

cf = fit3$coefficients

x = seq(min(cats$Bwt[cats$Sex=="F"]),max(cats$Bwt[cats$Sex=="F"]),0.1)
lines(x, cf[1]+cf[2]*x,col='red')
x = seq(min(cats$Bwt[cats$Sex=="M"]),max(cats$Bwt[cats$Sex=="M"]),0.1)
lines(x, cf[1]+cf[3]+(cf[2]+cf[4])*x,col='green')
```

The following `R` code can be used to perform the required tests for Part (c). Directly from the output, the F-test P-value for the Models 1 and 2 comparison is $P = 0.7875$ and the F-test P-value for the Models 1 and 3 comparison is $P = 0.1337$. We cannot conclude that either model 2 or 3 improves model 1 at an $\alpha = 0.05$ sigificance level.

```
>
> ### Use the anova() function to do a goodness of fit F-test
>
> anova(fit1,fit2)
Analysis of Variance Table


Model 1: Hwt ~ Bwt
Model 2: Hwt ~ Bwt + Sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    142 299.53
2    141 299.38  1    0.1548 0.0729 0.7875
> anova(fit1,fit3)
Analysis of Variance Table

Model 1: Hwt ~ Bwt
Model 2: Hwt ~ Bwt * Sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    142 299.53
2    140 291.05  2    8.4865 2.0411 0.1337
>
```

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 $h($`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}.

SOLUTION:

(a) The following R code can be used to plot the required graphs for Parts (a) and (b). See Figure 3. From the residual plot (first row of Figure 3) it is clear that the variance is larger for larger fitted values. From the normal quantile plot (first row of Figure 3) it is clear that the distribution of the residuals is not normal, especially values located in the left and right tails.

```
######
###### Q3
######

library(MASS)

### (a)

pdf("fig1A1Q3ab.pdf")
par(mfrow=c(3,2))

# Fit model

fit1 = lm(Claims ~ .,data=Insurance)

# Residual plot

plot(fit1$fitted.values,fit1$residuals,main="Residual plot Q3a")
abline(h=0)

# Normal quantile plot

qqnorm(fit1$residuals,main="Normal quantile plot Q3a")
qqline(fit1$residuals)


### (b)

# a = 1

fit2 = lm(log(Claims+1) ~ .,data=Insurance)
plot(log(fit2$fitted.values),fit1$residuals,main = "Residual plot Q3b, a = 1")
abline(h=0)
qqnorm(fit2$residuals,main="Normal quantile plot Q3b, a = 1")
qqline(fit2$residuals)


# a = 10

fit2 = lm(log(Claims+10) ~ .,data=Insurance)
```

```
plot(log(fit2$fitted.values),fit1$residuals,main = "Residual plot Q3b, a = 10")
abline(h=0)
qqnorm(fit2$residuals,main="Normal quantile plot Q3b, a = 10")
qqline(fit2$residuals)

dev.off()
```

(b) The minimum value of `Claims` is zero, for which the logarithm is not defined. Therefore, the $\log(x)$ transformation cannot be used. Figure 3 contains all required plots for this part. Neither transformation succeeds entirely in satisfying the constant variance assumption (see residuals plots). However, the $\log(x + 10)$ transformation yields residuals closer to the normal distribution (see normal quantile plots).

(c) If there are 4 predictors, then the number of predictor subsets is $2^4 = 16$, including the empty set. All models include the intercept, so the total number of models is 16.

(d) The following code can be used to answer Part (d). From the resulting table, the full model `log(Claims + 10)   1 + District + Group + Age + Holders` has the highest R^2_{adj} (= 0.9442990).

```
> ### (d)
>
> # This is a recursive subroutine that
> # returns all subsets of the elements of
> # a vector x of size k or less
>
> subset.enum = function(x,k) {
+
+   if (k == 0 | length(x) == 0) {
+     return(vector("list",1))
+   } else {
+     list1 = subset.enum(x[-1],k)
+     list2 = subset.enum(x[-1],k-1)
+     list3 = lapply(list2,function(y) {c(x[1],y)} )
+     return(append(list1,list3))
+   }
+ }
>
>
> # We can get the term labels this way
>
> fit1 = lm(log(Claims+10) ~ .,data=Insurance)
> full.formula = formula(terms(fit1))
> tm = attr(terms(full.formula),"term.labels")
> tm
[1] "District" "Group"    "Age"      "Holders"
>
> # for any subset of terms (say District and Group) we can construct a formula this wa
```

```
>
> as.formula(paste("log(Claims+10) ~ ", paste(tm[1:2], collapse="+")))
log(Claims + 10) ~ District + Group
>
> # We just need to enumerate all subsets
>
> subset.list = subset.enum(1:4,4)
>
> # This subroutine returns a formula using the predictor subset
> # defined by ondex subset subl.
>
> ff.sub = function(subl) {
+   if (is.null(subl)) {
+     fm = "log(Claims+10) ~ 1"
+   } else {
+     fm = as.formula(paste("log(Claims+10) ~ 1 +", paste(tm[subl], collapse="+")))
+   }
+   return(fm)
+ }
>
> # Create a list of formula for all predictor subsets
> # defined by the index subsets in subset.list.
>
> formula.list = lapply(subset.list, ff.sub)
>
> # For each formula fit the model, then extract the adjusted R-squared.
>
> radj.vector = sapply(formula.list, function(fm) {
+   fit1 = lm(fm,data=Insurance)
+   return(summary(fit1)$adj.r.squared)
+ })
>
> # Display the the adjusted R-squared for each formula.
>
> data.frame(as.character(formula.list),radj.vector)
                                as.character.formula.list. radj.vector
1                                       log(Claims+10) ~ 1   0.0000000
2                           log(Claims + 10) ~ 1 + Holders   0.6801578
3                               log(Claims + 10) ~ 1 + Age   0.5200682
4                     log(Claims + 10) ~ 1 + Age + Holders   0.7468752
5                             log(Claims + 10) ~ 1 + Group   0.1517469
6                   log(Claims + 10) ~ 1 + Group + Holders   0.7023331
7                       log(Claims + 10) ~ 1 + Group + Age   0.7071737
8             log(Claims + 10) ~ 1 + Group + Age + Holders   0.8100072
9                          log(Claims + 10) ~ 1 + District   0.1668537
10               log(Claims + 10) ~ 1 + District + Holders   0.7085192
11                   log(Claims + 10) ~ 1 + District + Age   0.7230757
```

```
12         log(Claims + 10) ~ 1 + District + Age + Holders   0.8248050
13                 log(Claims + 10) ~ 1 + District + Group   0.3353690
14       log(Claims + 10) ~ 1 + District + Group + Holders   0.7416876
15           log(Claims + 10) ~ 1 + District + Group + Age   0.9318542
16 log(Claims + 10) ~ 1 + District + Group + Age + Holders   0.9442990
>
```

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)?

SOLUTION:

(a) From Section 6.2, we may define a linear combination of regresion coefficients:

$$\eta = a_1\beta_1 + \ldots a_q\beta_q = \boldsymbol{a}^T\boldsymbol{\beta},$$

where $\boldsymbol{a} = [a_1 \cdots a_q]^T$ is the appropriate column vector. The estimator for η is

$$\hat{\eta} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}}. \tag{3}$$

The standard error $S_{\hat{\eta}}$ of this estimate is given by

$$S_{\hat{\eta}}^2 = MSE \times \boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{a}, \tag{4}$$

leading to a t-statistic $T = \hat{\eta}/S_{\hat{\eta}}$ for testing null hypothesis $H_0 : \boldsymbol{a}^T\boldsymbol{\beta} = 0$.

The following R code can be used to plot the required graphs for Part (a). See Figure 4. From the plot it is clear that the t-statistic is increasing with respect to $\widehat{\mathrm{B}}$wt, but is also bounded above and below by numbers close to ±2.

```
######
###### Q4
######

library(MASS)

### (a)

# Fit model

fit3 = lm(Hwt~Bwt*Sex,data=cats)

# Get the design matrix X

Xmatrix = model.matrix(Hwt~Bwt*Sex,data=cats)

# Calculate MSE

n = dim(cats)[1]
mse = sum(fit3$residuals^2)/(n-4)

# Get the XtX matrix, then the inverse

XtX = t(Xmatrix)%*%Xmatrix
XtX.inv = solve(XtX)

# Put the regression coefficients in matrix form

cf = fit3$coefficients
cfm = matrix(cf,4,1)
```

```
# loop through body weight variable bwt

t.vec=NULL
for (bwt in seq(0,5,0.1)) {

  # put linear combination in matrix form

  lin.coef = matrix(c(0,0,1,bwt),4,1)

  # calculate estimate of linear combination

  eta = t(lin.coef)%*%cf

  # calculate SE of estimate

  SE.eta = sqrt(t(lin.coef)%*%XtX.inv%*%lin.coef)*sqrt(mse)

  # capture the t-statistic

  t.vec = c(t.vec,eta/SE.eta)
}

pdf("fig1A1Q4a.pdf")
plot(seq(0,5,0.1),t.vec,xlab=’bwt’,ylab=’t-statistic’,type=’l’)
abline(h=0,lty=2)
dev.off()
```

(b) The P-value for comparing Models 1 and 3 using a goodness of fit F-test from **Q2** (c) was $P = 0.1337$, so Model 3 does not significantly improve Model 1 at a $\alpha = 0.1$ significant level (that is, the regression lines for the male and female groups are the same). However, we can directly test $H_o : \mu^M_{3.5} - \mu^F_{3.5} = 0$ against $H_a : \mu^M_{3.5} \neq \mu^F_{3.5}$, using a t-statistic of the type constructed in Part (a). This can be calculated using the following `R` code, which gives $T \approx 1.809$, with $n - 4 = 140$ degrees of freedom, and a two-sided P-value of $P \approx 0.0725$. This seems to lead to a contradiction. Clearly, if the regression lines for the male and female groups differ at any point, then they cannot be the same line. And if we accept the t-test, then by this logic we can conclude that the two lines differ with a significance level of $\alpha = 0.1$. The question we consider next is whether or not the procedure we did was actually a t-test.

```
> ### (b)
>
> # We need to single out bwt = 3.5 in the precding code
>
> bwt = 3.5
>
> # put linear combination in matrix form
```

```
>
> lin.coef = matrix(c(0,0,1,bwt),4,1)
>
> # calculate estimate of linear combination
>
> eta = t(lin.coef)%*%cf
>
> # calculate SE of estimate
>
> SE.eta = sqrt(t(lin.coef)%*%XtX.inv%*%lin.coef)*sqrt(mse)
>
> # capture estimate, SE, t-statistic and P-value
>
> c(eta,SE.eta,eta/SE.eta,2*pt(-abs(eta/SE.eta),df=n-4))
[1] 1.70152610 0.94037825 1.80940605 0.07253329
>
```

(c) Suppose we label the t-statistics of Part (b) as T_x, for body weight x. Suppose we fix a certain body weight range $x \in [a, b]$, and select $x^* \in [a, b]$ with maximizes $|T_x|$. Using this procedure, T_{x^*} can not be interpreted as a t-statistic for a specific hypothesis. Rather, it is the t-statistic selected from a set of t-statistics with the largest magnitude. More precisely

$$|T_{x^*}| = \max_{x \in [a,b]} \{|T_x|\}.$$

This will not have the same distribution as the absolute value of a random variable with a t-distribution. In Part (b), $T_{3.5}$ was selected using a similar procedure (that is, it was singled out because it was large).

However, we can simultaneously reject all null hypotheses of the form

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

in favor of the two-sided alternatives for all corresponding t-statistics $T_{\boldsymbol{a}}$ which satisfy $|T_{\boldsymbol{a}}| > (\chi^2_{p;\alpha})^{1/2}$, with a single level of signifigance α, where p is the model degrees of freedom. In this example, $p = 4$ and $(\chi^2_{p;\alpha})^{1/2} \approx 7.779^{1/2} \approx 2.79$. So, in Part (b), if $|T_{3.5}| > 2.79$ we would be able to conclude on that basis that the male and female regression lines were different. However, we observed $T_{3.5} \approx 1.809$ instead, so there is no contradiction.

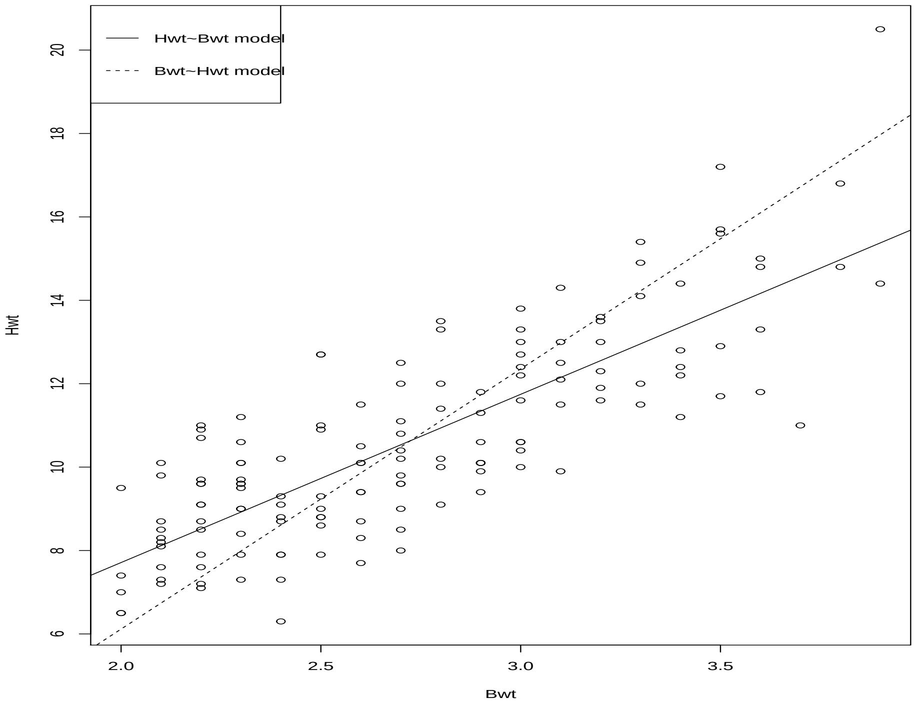

Figure 1: Plot for Q2b.

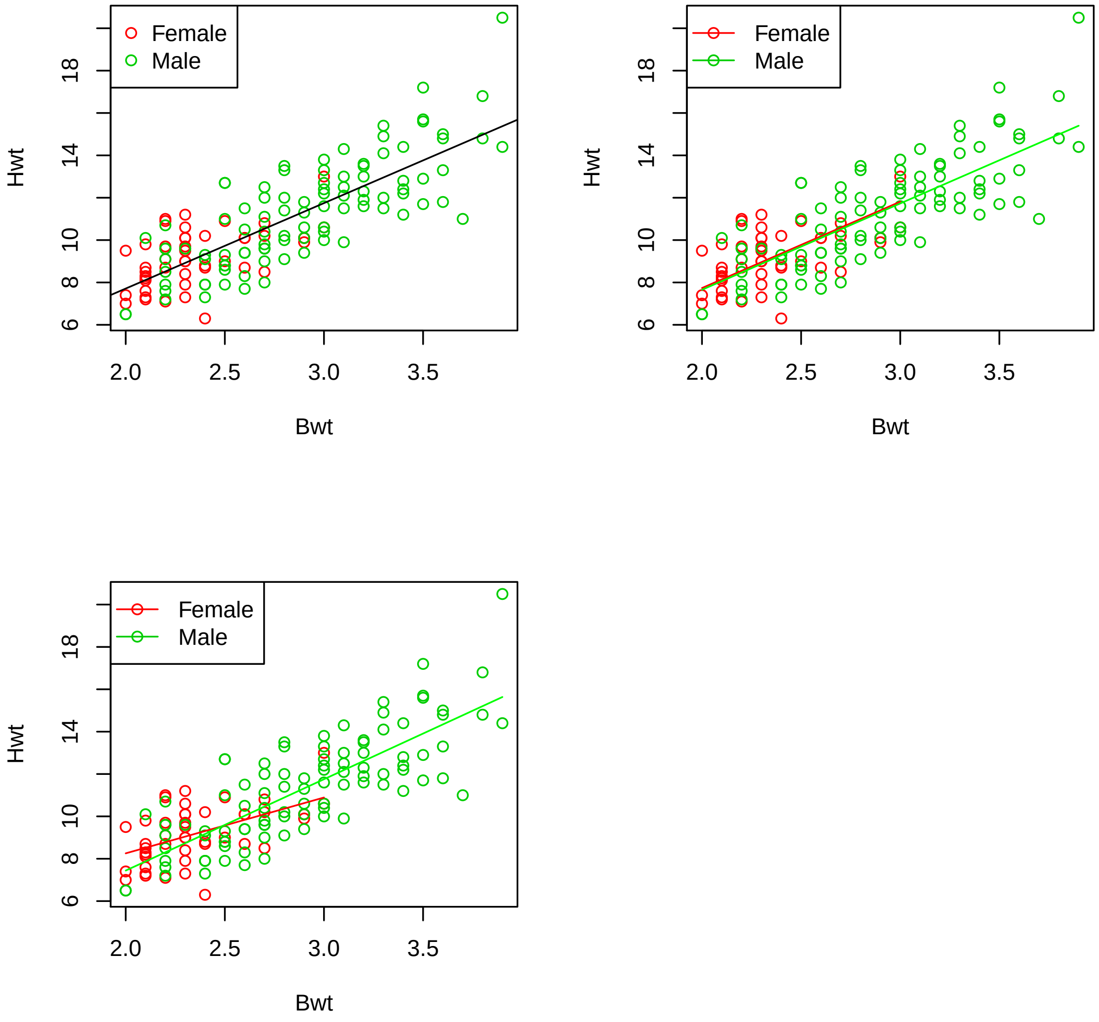

Figure 2: Plot for Q2c.

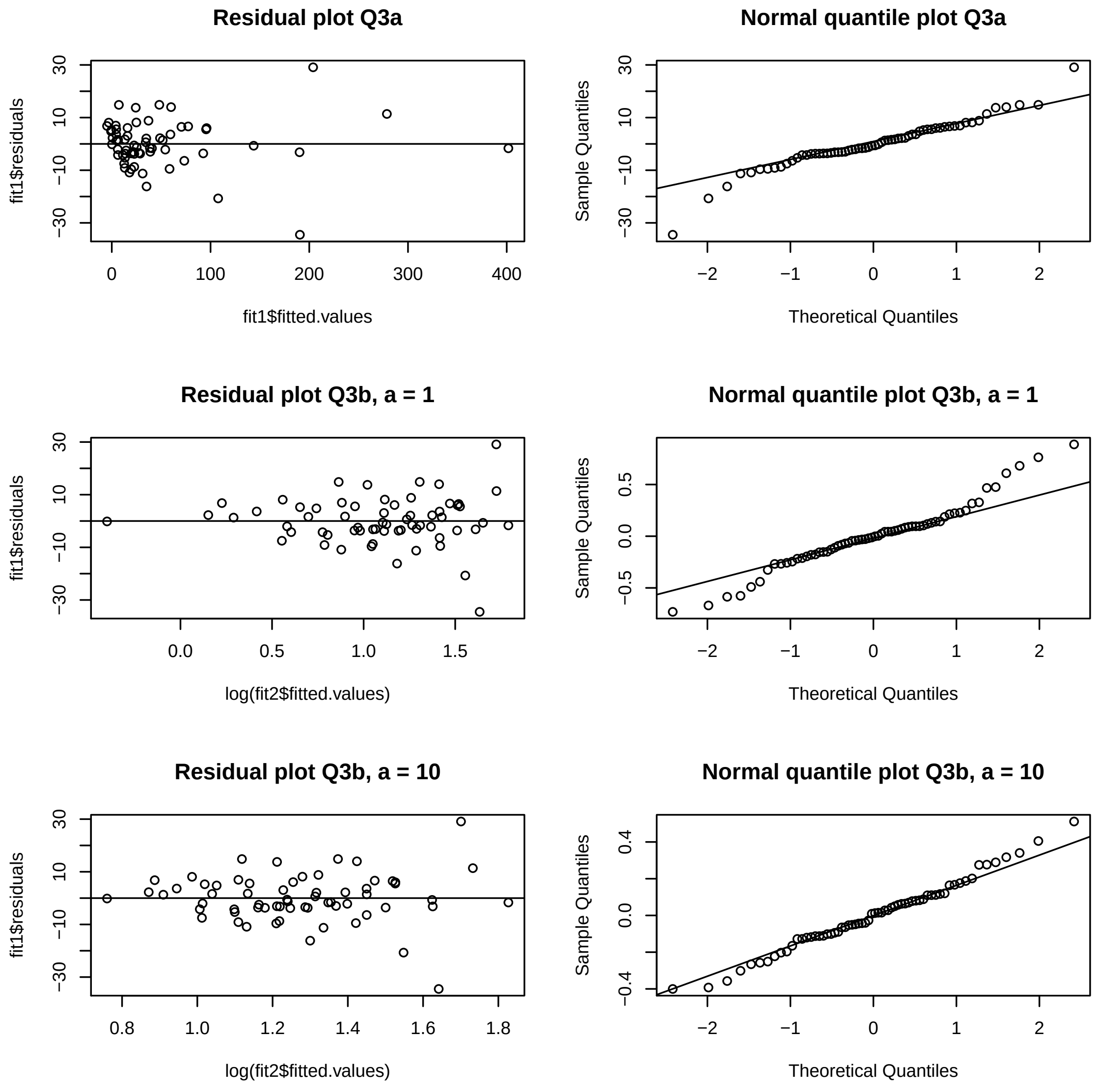

Figure 3: Plots for Q3a-b.

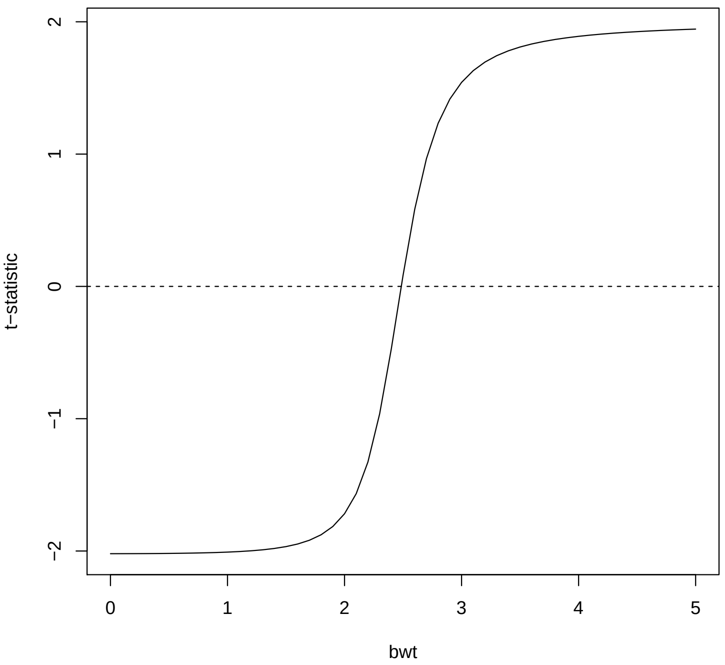

Figure 4: Plots for Q4a.