Midterm Exam 2 - Open Book Section (R) - Part 2

Instructions

The R Markdown file / Jupyter Notebook file includes the questions, the empty code chunk sections for your code, and the text blocks for your responses. Answer the questions below by completing the R Markdown file / Jupyter Notebook file. You must answer the questions using one of these files. You will submit a pdf, Word or html file using one of the two files. You may make other slight adjustments to get the file to knit/convert but otherwise keep the formatting the same. Once you've finished answering the questions, submit your responses in a single knitted/converted file (just like the homework peer assessments).

There are 10 questions total, each worth between 2-7 points. Partial credit may be given if your code is correct, but your conclusion is incorrect or vice versa.

Next Steps:

1. Save the .Rmd/.ipynb file in your R working directory - the same directory where you will download the abalone.csv data file into. Having both files in the same directory will help in reading the abalone.csv file.
2. Read the question and create the R code necessary within the code chunk section immediately below each question. Knitting this file will generate the output and insert it into the section below the code chunk.
3. Type your answer to the questions in the text block provided immediately after the response prompt.
4. Once you've finished answering all questions, knit this file and submit the knitted file on Canvas.

Example Question 10 - 4pts

This will be the exam question - each question is already copied from Canvas and inserted into individual text blocks below, *you do not need to copy/paste the questions from the online Canvas exam.*

```
# Example code chunk area. Enter your code below the comment‘
```

Response to Question 10: This is the section where you type your written answers to the question. Depending on the question asked, your typed response may be a number, a list of variables, a few sentences, or a combination of these elements.

Ready? Let's begin.

Background

For this exam, you will be predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task! Other measurements, which are easier to obtain, are used to predict the age.

The data consists of a data frame with 4177 observations on the following 8 variables:

1. *Sex*: M, F, and I (infant) (categorical)
2. *Length*: Longest shell measurement in mm (continuous)
3. *Diameter*: Perpendicular to length in mm (continuous)
4. *Height*: Height with meat in shell in mm (continuous)
5. *Whole*: Weight of whole abalone in grams (continuous)
6. *Viscera*: Gut weight (after bleeding) in grams (continuous)
7. *Shell*: Shell weight after being dried in grams (continuous)
8. *Rings*: Number of rings of the abalone – corresponds with the age (discrete)

Read the data

Read the data and answer the questions below using the supplied R Markdown / Jupyter notebook file.

```
# Load relevant libraries (add here if needed)
library(car)
```

```
## Loading required package: carData
```

```
library(aod)
```

```
## Warning: package 'aod' was built under R version 4.0.3
```

```
# Read the data set
abaloneFull = read.csv("C:/Users/mjpearl/Desktop/omsa/ISYE-6414-OAN/midterm2/abalone.csv",head=T)
row.cnt = nrow(abaloneFull)

# Split the data into training and testing sets
abaloneTest = abaloneFull[(row.cnt-9):row.cnt,]
abalone = abaloneFull[1:(row.cnt-10),]
```

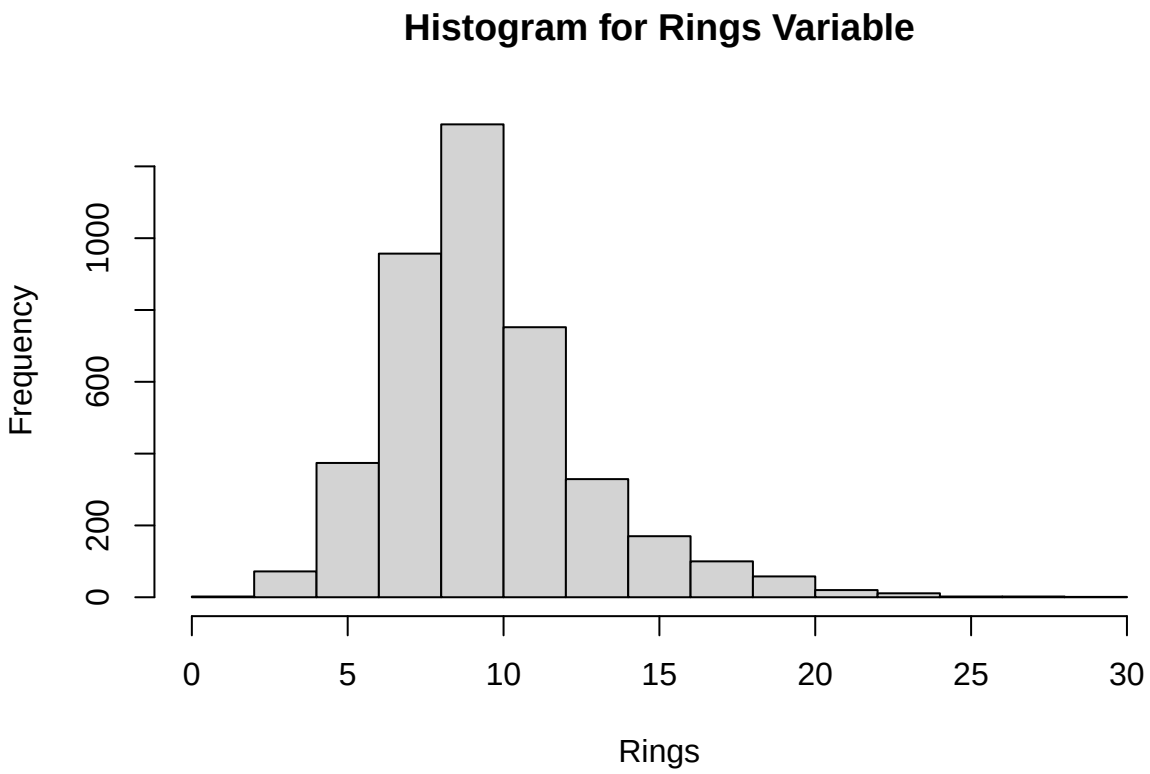
Note: Use *abalone* as your data set for the following questions unless otherwise stated.

Note: Treat all variables as quantitative variables, except for *Sex*.

Question 1 - 4 points

Create a histogram of the variable *Rings*. What generalized linear regression model(s) discussed in this course could be used to model this response variable? Explain.

```
# Code to build histogram
hist(abalone$Rings,xlab='Rings',main="Histogram for Rings Variable")
```



Response to Question 1: Based on our histogram we see that the response variable Rings tends to follow a normal distribution and doesn't exhibit any significant skewing when looking at our data. Based on our data being a continuous response, we can use a Standard or Multiple Linear Regression model in order to model this response variable. If we wanted to use the Sex variable in this case we would obviously need to convert this from a categorical variable to a dummy variable so we can convert it to two numeric fields.

Question 2 - 4 points

A) Build a multiple linear regression model named **model1** with *Rings* as the response variable and all other variables as predicting variables. Include an intercept. Display the summary table of the model.

B) Is the overall regression significant at the 0.01 alpha level? Explain.

```
# Code to create model/summary
model1 <- lm(Rings ~ Sex+Length+Diameter+Height+Whole+Viscera+Shell,data=abalone)
summary(model1)
```

```
##
## Call:
## lm(formula = Rings ~ Sex + Length + Diameter + Height + Whole +
##     Viscera + Shell, data = abalone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4620  -1.4164  -0.4250   0.8664  16.5084
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.79432    0.30912  15.510  < 2e-16 ***
## SexI        -1.02483    0.10917  -9.387  < 2e-16 ***
## SexM        -0.06053    0.08907  -0.680   0.4968
## Length      -3.77737    1.92922  -1.958   0.0503 .
## Diameter    11.58716    2.38165   4.865 1.19e-06 ***
## Height      11.88517    1.64283   7.235 5.53e-13 ***
## Whole       -5.57497    0.43472 -12.824  < 2e-16 ***
## Viscera     -2.38965    1.33586  -1.789   0.0737 .
## Shell       25.70361    0.94033  27.335  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.345 on 4158 degrees of freedom
## Multiple R-squared:  0.473,  Adjusted R-squared:  0.472
## F-statistic: 466.5 on 8 and 4158 DF,  p-value: < 2.2e-16
```

Response to Question 2 B): Since $\alpha = 0.01$ exceeds the observed significance level, p = <2.2e-16, we reject the null hypothesis. The data provide strong evidence that at least one of the slope coefficients is nonzero. The overall model appears to be statistically useful in predicting the Rings variable.

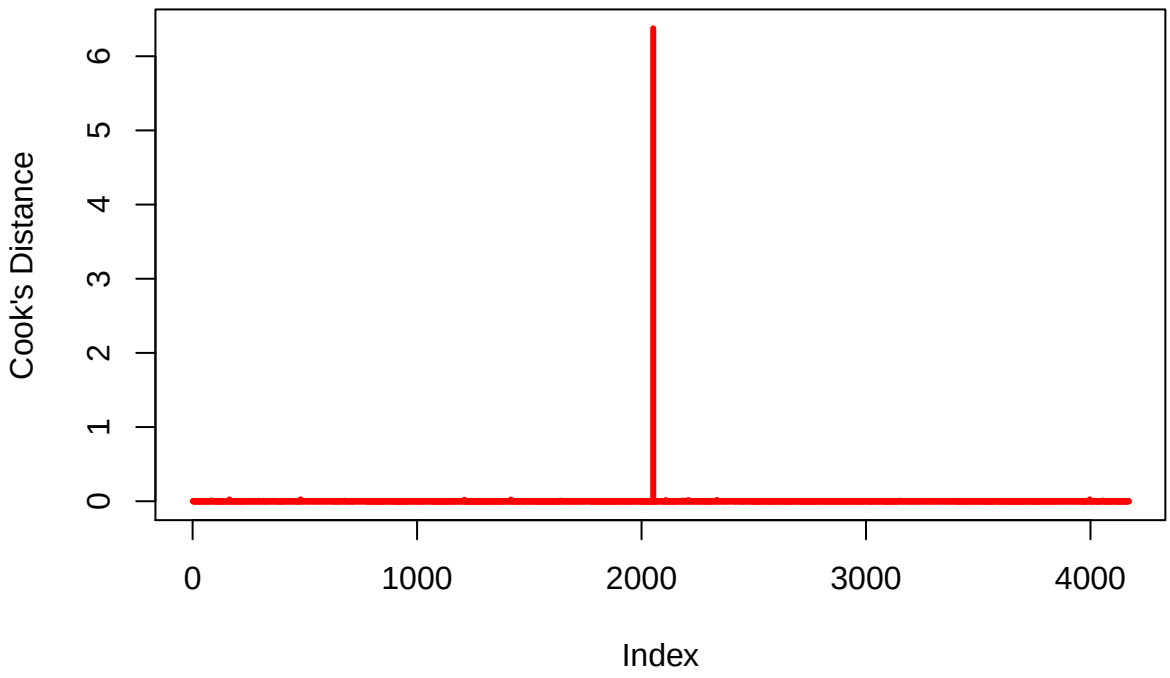
Question 3 - 6 points

A) Using **model1**, calculate the Cook's distance of the points in the dataset and create a plot for the Cook's Distances.

B) Identify the row number of the observation with the highest Cook's distance.

C) Remove this observation from the *abalone* dataset. Call this new dataset *abalone2* and create a new multiple linear regression model, called *model2*, using all predictors with *Rings* as the response. Display the summary table of this model. Discuss differences between the models with and without the outlier. Would you classify this observation as influential?

```
# Code to calculate cooks distance, create plot, identify row number, create new dataset and new model,
cook = cooks.distance(model1)
plot(cook,type="h",lwd=3,col='red',ylab = "Cook's Distance")
```




```
cook[cook>1]
```

```
##     2052
## 6.375861
```

Response to Question 3 B): Based on our data we can observe that the row number 2052 contains an influential point that's well above our typical threshold.

Response to Question 3 C):

```
removeRows <- function(rowNum, data) {
      newData <- data[-rowNum, , drop = FALSE]
      rownames(newData) <- NULL
      newData
}

abalone2 = removeRows(2052,abalone)
model2 <- lm(Rings ~ Sex+Length+Diameter+Height+Whole+Viscera+Shell, data=abalone2)
summary(model2)
```

```
##
## Call:
## lm(formula = Rings ~ Sex + Length + Diameter + Height + Whole +
##     Viscera + Shell, data = abalone2)
##
```

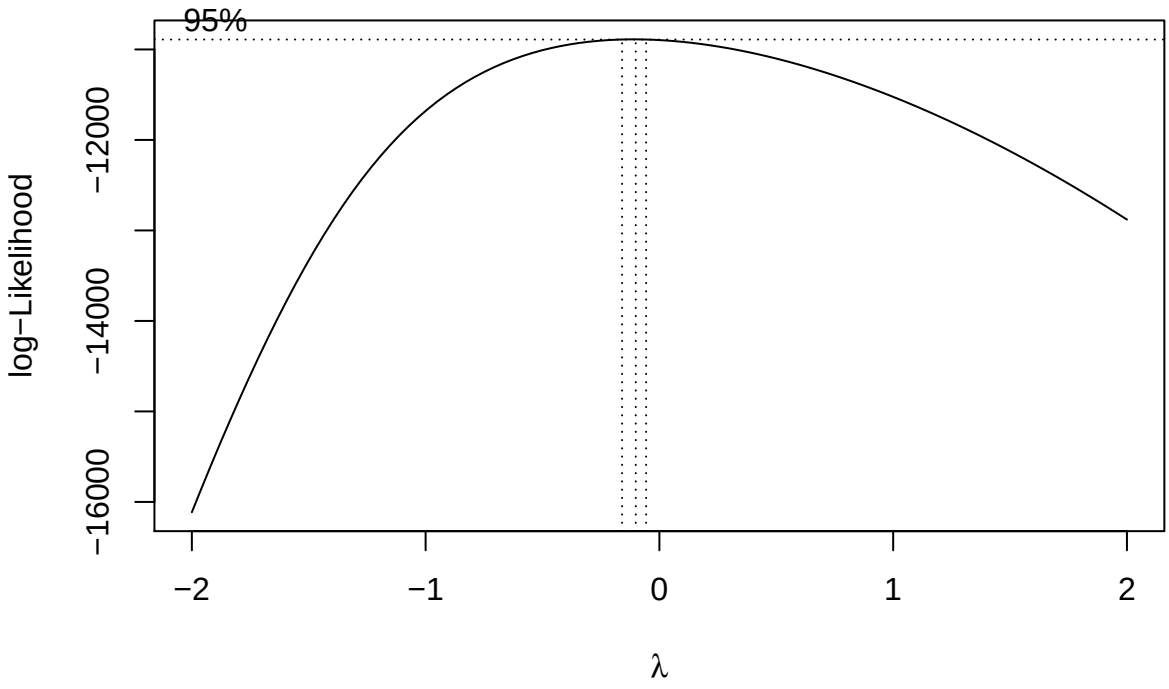
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4835 -1.4293 -0.4088  0.8632 16.6578
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.43440    0.31068  14.273  < 2e-16 ***
## SexI        -0.99300    0.10852  -9.151  < 2e-16 ***
## SexM        -0.06538    0.08847  -0.739   0.4599
## Length      -4.30330    1.91749  -2.244   0.0249 *
## Diameter     9.72743    2.37829   4.090 4.39e-05 ***
## Height      24.30973    2.31195  10.515  < 2e-16 ***
## Whole       -5.44717    0.43212 -12.606  < 2e-16 ***
## Viscera     -3.11269    1.33029  -2.340   0.0193 *
## Shell       24.37620    0.95025  25.652  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.329 on 4157 degrees of freedom
## Multiple R-squared:  0.4802, Adjusted R-squared:  0.4792
## F-statistic:   480 on 8 and 4157 DF,  p-value: < 2.2e-16
```

Based on our results we can see that removing the observation doesn't significantly improve our model's results. There was only a slight uptick to the Adjusted R^2 and R^2 values. We can therefore conclude that this observation is not an influential point as we don't exhibit a significant improvement in the model2's performance compared to model1.

Question 4 - 7 points

A) Find the optimal lambda, rounded to the nearest half integer, for a Box-cox transformation on **model2**.

```
# Code to find optimal lambda/create new model /calculate vif
library(MASS)
bc <- boxcox(model2,data=abalone2)
```




```
# Code to find optimal lambda/create new model /calculate vif
round(bc$x[which.max(bc$y)],0)
```

```
## [1] 0
```

B) Based on the results in (A), create a new model named **model3** with the appropriate Box-cox transformation. Display the summary table of the model. Note: Make sure to use the cleaned data set (*abalone2*).

```
# Model3 with log(Rings) used as the response variable
model3 <- lm(log(Rings) ~ Sex+Length+Diameter+Height+Whole+Viscera+Shell, data=abalone2)
summary(model3)
```

```
##
## Call:
## lm(formula = log(Rings) ~ Sex + Length + Diameter + Height +
##     Whole + Viscera + Shell, data = abalone2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38451 -0.13798 -0.02261  0.11246  0.87788
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.377904   0.028288  48.711  < 2e-16 ***
## SexI        -0.105868   0.009881 -10.715  < 2e-16 ***
## SexM        -0.001549   0.008056  -0.192   0.8476
## Length       0.198503   0.174588   1.137   0.2556
## Diameter     1.268198   0.216545   5.857 5.09e-09 ***
## Height       2.631469   0.210505  12.501  < 2e-16 ***
## Whole       -0.596839   0.039345 -15.169  < 2e-16 ***
## Viscera     -0.226678   0.121123  -1.871   0.0614 .
## Shell        1.885279   0.086521  21.790  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2121 on 4157 degrees of freedom
## Multiple R-squared:  0.5612, Adjusted R-squared:  0.5604
## F-statistic: 664.6 on 8 and 4157 DF,  p-value: < 2.2e-16
```

C) Compare and discuss the adjusted R^2 of **model2** and **model3**.

D) Conduct a multicollinearity test on the predictors of *model3*. Does multicollinearity seem to be a problem in this model? Explain.

```
# create new model /calculate vif
cat("VIF Threshold:", max(10, 1/(1-summary(model3)$r.squared)), "\n")
```

```
## VIF Threshold: 10
```

```
# create new model /calculate vif
vif(model3)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## Sex        1.535728  2        1.113214
## Length    40.767004  1        6.384904
## Diameter  42.830157  1        6.544475
## Height     6.220743  1        2.494142
## Whole     34.503384  1        5.873958
## Viscera   16.340894  1        4.042387
## Shell     13.449515  1        3.667358
```

Response to Question 4 A): Based on our results of the boxcox transformation, we can see that our lambda value when rounded to the nearest integer is = 0. Therefore, this suggests that we should apply a log transformation to the response variable for model3.

Response to Question 4 C): When looking at the Adjusted R^2 and R^2 values for model2 and model3 we can see that the log transformation to the Rings variable does provide improvement to the overall score as the R^2 increased from 0.4802 to 0.5612 and the Adjusted R^2 value increased from 0.4792 to 0.5604.

Response to Question 4 D): When looking at the output for the vif function. We can see that there are several variables that are exhibiting multicollinearity including the Length, Diameter, Whole, Viscera and Shell variables. Multicollinearity can cause issue for the convergence of the target variable, and may also provide inaccurate estimates of the coefficient values for our model. This can also potentially result in coefficient values for the model being classified as significant when they may actually be not.

Question 5 - 4 points

A) Using the cleaned data set (*abalone2*), build a Poisson regression model named **model4** with *Rings* as the response variable and all other variables as predicting variables. Include an intercept. Display the summary table of the model.

```
# Code to poisson regression model
model4 <- glm(Rings ~ Sex+Length+Diameter+Height+Whole+Viscera+Shell, data=abalone2, family='poisson')
summary(model4)
```

```
##
## Call:
## glm(formula = Rings ~ Sex + Length + Diameter + Height + Whole +
##     Viscera + Shell, family = "poisson", data = abalone2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9830  -0.4767  -0.1368   0.2800   4.0372
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.54892    0.04700  32.954  < 2e-16 ***
## SexI        -0.11520    0.01513  -7.615 2.64e-14 ***
## SexM        -0.00392    0.01149  -0.341    0.733
## Length      -0.11722    0.25754  -0.455    0.649
## Diameter     1.24976    0.31618   3.953 7.73e-05 ***
## Height       2.51223    0.28975   8.670  < 2e-16 ***
## Whole       -0.54380    0.05585  -9.737  < 2e-16 ***
## Viscera     -0.27319    0.16963  -1.610    0.107
## Shell        1.93476    0.11332  17.073  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4137.4  on 4165  degrees of freedom
## Residual deviance: 2056.5  on 4157  degrees of freedom
## AIC: 19164
##
## Number of Fisher Scoring iterations: 4
```

B) Perform a test for the overall regression, using $\alpha = 0.01$. Does the overall regression have explanatory power? Explain.

```
# Code to build model/ perform test for overall regression ...
1-pchisq((4137.4-2056.5),(4165-4157))
```

```
## [1] 0
```

Response to Question 5 B) : Based on our observed p-value of 0, we reject the null hypothesis and conclude that at least one predicting variables is statistically significant.

Question 6 - 4 points

A) What is the estimated value of the *SexI* coefficient in **model4**?

B) Interpret the *SexI* coefficient in the context of the problem. *Note: Make sure that you are treating Female (F) as the baseline level.*

Response to Question 6 A): The estimated value of the SexI coefficient is -0.11520

Response to Question 6 B): We can interpret the SexI coefficient as the following: -The SexI coefficient is negative and statistically significant based on a p-value of 2.64e-14 and has a value of -0.11520 -The ratio of Rings to SexI to the baseline SexF = exp(SexI) = 0.8911. This suggest a lower rate of Rings for SexI compared to the baseline group SexF.

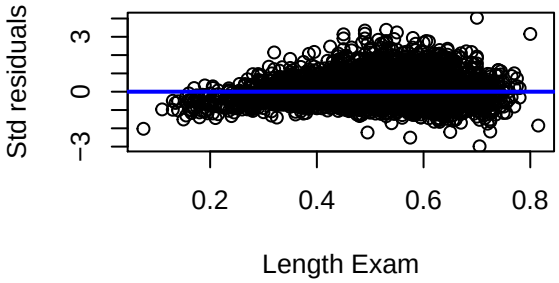
Question 7 - 5 points

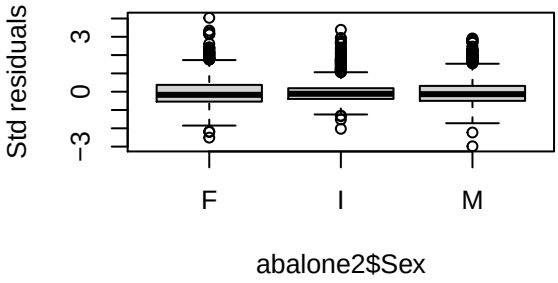
Perform a goodness-of-fit statistical test for **model4** using the deviance residuals and $\alpha = 0.01$. Provide the null and alternative hypotheses, test statistic, p-value, and conclusions in the context of the problem.

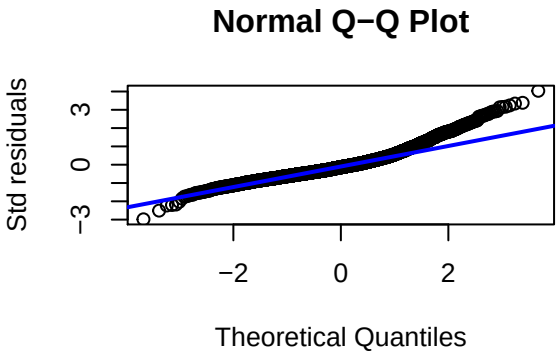
```
# Code to perform GOF test ...
with(model4, cbind(res.deviance =deviance,df=df.residual,p=1-pchisq(deviance,df.residual)))
```

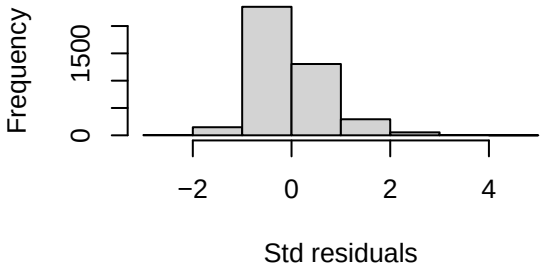
```
##      res.deviance   df p
## [1,]     2056.494 4157 1
```

```
# Code to perform GOF test ...
res = resid(model4,type="deviance")
par(mfrow=c(2,2))
plot(abalone2$Length,res,ylab="Std residuals",xlab="Length Exam")
abline(0,0,col="blue",lwd=2)
boxplot(res~abalone2$Sex,ylab = "Std residuals")
qqnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
hist(res,10,xlab="Std residuals", main="")
```









Based on our results of the residual analysis and selecting an arbitrary predicting variable Length, it does not seem that the normality assumption holds in this case as we seem to have a skewed distribution in our standard residuals plot, and our qq-plot is exhibiting a skew as well.

However, since the Poisson regression is based on a sample size, we can't use these plots to fully conclude that our model is not a good fit, and we may need to perform additional transformations to our predicting variable in order for these assumptions to hold.

Response to Question 7:

H_0 : The model provides a good fit to the data

H_a : The model does not provide a good fit to the data

Deviance Test statistic: 2056.494

p-value: 1

Conclusion: Based on our p-value of 1, which signifies a large value. Therefore, since we have a large p-value we do not reject the null hypothesis and can conclude that our model does provide a good fit to the data.

Question 8 - 3 points

Estimate the dispersion parameter for *model4*. Does overdispersion seem to be a problem in this model?

```
# Code to estimate dispersion parameter ...
dispersion = 2056.494/(length(model4$y)-length(coefficients(model4))-1)
dispersion
```

```
## [1] 0.4948253
```

Response to Question 8: Based on the value of 0.49 which is < 2, we can conclude that we are not exhibiting overdispersion for this model.

Question 9 - 6 points

A) Using the cleaned data set (*abalone2*), create a new Poisson regression model by adding interaction terms between *Sex* and *Height* to **model4**. Include an intercept. Call it **model5**. Display the summary table of the model.

Hint: Given that *Sex* has three levels, by adding the term Sex:Height to **model4**, **model5** will include two interaction terms: *SexI:Height* and *SexM:Height*. If needed, you can take a look at the help file by typing ?colon.

B) Perform a testing for subset of coefficients, comparing **model5** with **model4**, using $\alpha = 0.05$. Based on this test, is at least one of the two interaction terms significantly explanatory given all other predictors in **model5**? Explain.

```
# Code to create model5 and performing test subset of coeficcients...
model5 <- glm(Rings ~ Sex*Height+Sex+Length+Diameter+Height+Whole+Viscera+Shell, data=abalone2, family=
summary(model5)
```

```
##
## Call:
## glm(formula = Rings ~ Sex * Height + Sex + Length + Diameter +
##     Height + Whole + Viscera + Shell, family = "poisson", data = abalone2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4487  -0.4664  -0.1198   0.2758   4.1296
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.98346    0.07121  27.853  < 2e-16 ***
## SexI        -0.67487    0.06415 -10.521  < 2e-16 ***
## SexM        -0.12682    0.05829  -2.175  0.02959 *
## Height       1.10809    0.39045   2.838  0.00454 **
## Length      -0.54878    0.26032  -2.108  0.03502 *
## Diameter     1.00618    0.31584   3.186  0.00144 **
## Whole       -0.45028    0.05668  -7.945 1.95e-15 ***
## Viscera     -0.16113    0.17019  -0.947  0.34374
## Shell        2.02391    0.11415  17.731  < 2e-16 ***
## SexI:Height  4.31322    0.46025   9.371  < 2e-16 ***
## SexM:Height  0.73309    0.36088   2.031  0.04222 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4137.4  on 4165  degrees of freedom
## Residual deviance: 1959.8  on 4155  degrees of freedom
## AIC: 19071
##
## Number of Fisher Scoring iterations: 4
```

Response to Question 9 B):

```
# Code to calculate estimates/MSPE/PE...
library(aod)
wald.test(b=coef(model5), Sigma=vcov(model5), Terms=1:9)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 89399.7, df = 9, P(> X2) = 0.0
```

Based on our result and a corresponding p-value = 0, we reject the null hypothesis and can conclude that the interaction terms related to Sex:Height are stastistically significant based on an alpha level of 0.05.

Question 10 - 7 points

Estimate *Rings* for the last 10 rows of data (*abaloneTest*) using both **model3** and **model4**. Compare and discuss the mean squared prediction error (MSPE) and the precision error (PM) of both models. Which model performed the best?

```
# Code to calculate estimates/MSPE/PE...
pred3<-predict(model3, abaloneTest)
# Calculate MSPE
mse.model3<-mean((pred3-abaloneTest$Rings)^2)
# Calculate estimates in terms of the original data
pred4<-predict(model4, abaloneTest)^2
# Calculate MSPE
mse.model4<-mean((pred4-abaloneTest$Rings)^2)
cat("The MSPE of model3 is", mse.model3, "\n")
```

```
## The MSPE of model3 is 55.85572
```

```
cat("The MSPE of model4 is", mse.model4, "\n")
```

```
## The MSPE of model4 is 19.27076
```

Response to Question 10:

Since the MSPE of model4 is smaller than model3, we can conclude that model4 is preferred for predicting the response Rings in the abalone data.