STAT 302: Homework 4

Multiple and Logistic Regression

Jaiden Atterbury

Due: 05-07-23 at 11:59 PM

1. **Using the Lake dataset on Canvas.** Answer the following questions:

- a) At 10% level of significance, do the data provide sufficient evidence to conclude that the independent variables (Alkalinity and pH) significantly predict the outcome variable Mercury? Justify your answer.

To check and see if the data provide sufficient evidence to conclude that the independent variables (Alkalinity and pH) significantly predict the outcome variable Mercury, we will create a linear model of Mercury on Alkalinity and pH, and see if the F statistic is significant for the overall model.

```r
# Create the linear model:
model_1 <- lm(Mercury ~ Alkalinity + pH, data = Lake)

# Get the summary statistics:
summary(model_1)
```

```
##
## Call:
## lm(formula = Mercury ~ Alkalinity + pH, data = Lake)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42194 -0.13978 -0.02044  0.09076  0.57123
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.081900   0.211998   5.103 5.66e-06 ***
## Alkalinity  -0.004014   0.001232  -3.257  0.00207 **
## pH          -0.065484   0.036294  -1.804  0.07747 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2327 on 48 degrees of freedom
## Multiple R-squared:  0.4839, Adjusted R-squared:  0.4624
## F-statistic:  22.5 on 2 and 48 DF,  p-value: 1.275e-07
```

As can be seen the F-statistic is 22.5 on 2 and 48 degrees of freedom respectively, with a very low p-value of 1.275e-07. Thus at the 10% level of significance, and any other reasonable level of significance for that matter, we have evidence that the independent variables (Alkalinity and pH) significantly predict the outcome variable Mercury.

- b) Find the least-squares estimates for the regression line in part (a).

```
# Create confidence intervals for the estimates:
confint(model_1, level = 0.90)
```

```
##                      5 %         95 %
## (Intercept)  0.726332628  1.437468089
## Alkalinity  -0.006080904 -0.001947016
## pH          -0.126357186 -0.004610573
```

As computed above, the y-intercept estimate was 1.081900 with a 90% confidence interval of [0.726332628, 1.437468089]. The Alkalinity slope estimate was -0.004014 with a 90% confidence interval of [-0.006080904, -0.001947016]. Lastly, the pH slope estimate was -0.065484 with a 90% confidence interval of [-0.126357186, -0.004610573].

Putting this all together we obtain the least squares regression equation of:

$$\text{Mercury} = -0.004 \cdot \text{Alkalinity} - 0.066 \cdot \text{pH} + 1.082$$

- c) Interpret the value of the slopes and intercept in the context of the problem. In addition, state which variables are significant predictors at 10% level of significance? Explain.

Below we will interpret the intercept and the slopes of the regression equation in the context of the problem.

As computed above, the regression slope for pH was -0.065484 with a 90% confidence interval of [-0.126357186, -0.004610573]. This means, on average, for every 1 unit increase in the pH level of the lake water the mean mercury level of that lake is estimated to decrease by about 0.065484 units over the sampled range of mercury levels, holding all other variables in the model fixed. Furthermore, the p-value for this estimate was 0.07747 which means this estimate is significant at the 10% significance level.

As computed above, the regression slope for Alkalinity was was -0.004014 with a 90% confidence interval of [-0.006080904, -0.001947016]. This means, on average, for every 1 unit increase in the Alkalinity level of the lake water the mean mercury level of that lake is estimated to decrease by about 0.04014 units over the sampled range of mercury levels, holding all other variables in the model fixed. Furthermore, the p-value for this estimate was 0.00207 which means this estimate is highly significant, especially at the 10% significance level.

As computed above, the regression y-intercept was 1.081900 with a 90% confidence interval of [0.726332628, 1.437468089]. This means that the estimated mean mercury level in the lake water is equal to about 1.081900 when the pH level and Alkalinity of the lake water is zero. Furthermore, the p-value for this estimate was 5.66e-06 which means this estimate is highly significant, especially at the 10% significance level.

- d) Perform a residual analysis to decide whether considering the assumptions for regression inferences met by the variables in the dataset appears reasonable.

The conditions we will be checking are: normality of the response variable, no autocorrelation, no or little multucollinearity, linear relationship between the response and the predictor variables, normality of the residuals, equal variance of the residuals, and no outliers. As stated in the problem description we will be running these tests at the 10% significance level.

Checking for Normality of the Response:

```
# Run a Shapiro test to check the normality of the response:
shapiro.test(Lake$Mercury)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Lake$Mercury
## W = 0.94551, p-value = 0.02055
```

As can be seen from the above Shapiro-Wilk test, since the p-value is 0.02055 which is less than 0.1, we reject the null hypothesis and we assume that the dependent variable is not approximately normally distributed.

Checking for Autocorrelation:

```
dwtest(model_1)
```

```
##
##  Durbin-Watson test
##
## data:  model_1
## DW = 1.6278, p-value = 0.09029
## alternative hypothesis: true autocorrelation is greater than 0
```

As can be seen by the above Durbin-Watson test, since the p-value is less than 0.1, we reject the null hypothesis and we see that there is evidence that the residuals are auto-correlated.

Checking for Multicollinearity:

```
vif(model_1)
```

```
## Alkalinity         pH
##   2.024882   2.024882
```

As can be seen by the above vif tests, the multicollinearity for each variable is below 5, hence we can assume that there is little to no multicollinearity between the variables, and hence we don't violate the assumption.

Checking for Linearity:

```
# Run a Rainbow test to check the linearity of the variables:
raintest(model_1)
```

```
##
##  Rainbow test
##
## data:  model_1
## Rain = 1.0163, df1 = 26, df2 = 22, p-value = 0.489
```

As can be seen from the above Rainbow test, since the p-value is 0.489 which is greater than 0.1, we fail to reject the null hypothesis and we assume that the relationship between the independent and dependent variables is linear.

Checking for Normality of the Residuals:

```
# Calculate the residuals:
resid_1 <- residuals(model_1)

# Run a Shapiro test to check the normality of the residuals:
shapiro.test(resid_1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_1
## W = 0.94961, p-value = 0.03035
```

As can be seen from the above Shapiro-Wilk test, since the p-value is 0.03035 which is less than 0.1, we reject the null hypothesis and we assume that the residuals are not normally distributed.

Checking for equal variance:

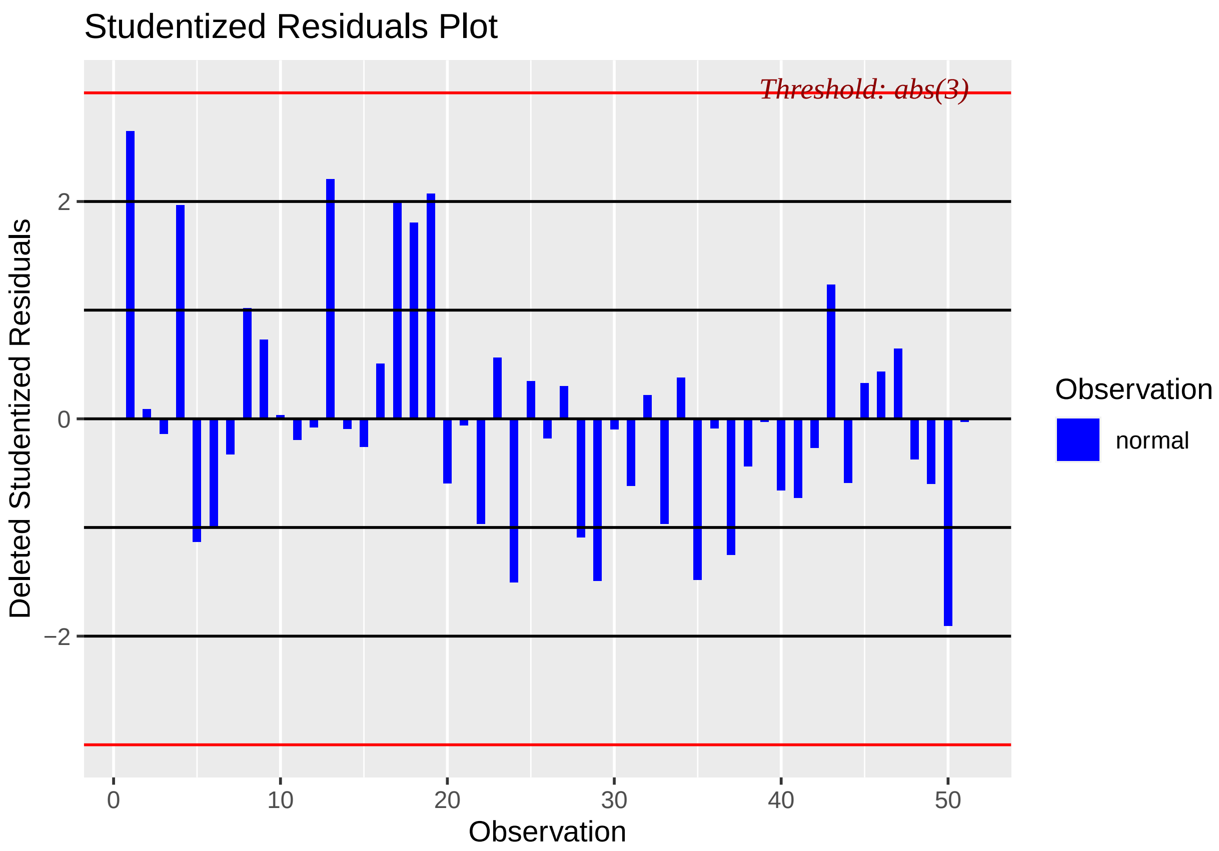
```
# Run the Breusch Pagan Test for Heteroskedasticity to test for equal variance:
ols_test_breusch_pagan(model_1)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##               Data
##  ------------------------------------
##  Response : Mercury
##  Variables: fitted values of Mercury
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    5.5134
##  Prob > Chi2   =    0.01887133
```

As can be seen from the above Breusch Pagan Test for Heteroskedasticity, since the p-value is 0.01887133 which is less than 0.1, we reject the null hypothesis, thus we have evidence that the equal/constant variance assumption is violated.

Checking for no outliers:

```
# Create a Studentized Residual Plot to check for outliers:
ols_plot_resid_stud(model_1)
```



As can be seen from the above studentized residual plot, there are no outliers present in the data.

Due to the fact that we violated the normality of the response, autocorrelation, normality of the residuals, and the equal variance assumption, it is safe to assume that the assumptions for regression inference are violated and we would have to do some extra work to fit this model.

2. **Using the Focus dataset on Canvas.** Find the regression equation (or least squares equation) for predicting GPA using Credits and Class (using Junior as your reference group). Carefully interpret your results.

Since the problem doesn't ask to check the residual assumptions we will skip those for this model, however, we will check and see if the dependent variable is approximately normal before fitting the model.

Checking for the normality of the response variable:

```
shapiro.test(Focus$GPA)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Focus$GPA
## W = 0.98968, p-value = 0.1601
```

As can be seen from the above Shapiro-Wilk normality tests, since the p-value is 0.1601 which is greater than 0.05, we fail to reject the null hypothesis and assume that the response variable is approximately normally distributed.

Below we will create the model for predicting GPA using Credits and Class (using Junior as the reference group).

```
# Changing reference category to Juniors:
Focus$CLASS <- relevel(factor(Focus$CLASS), ref = "Junior")

# Create the model for predicting GPA using Credits and Class (using Junior as
# your reference group):
model_2 <- lm(GPA ~ CREDITS + factor(CLASS), data = Focus)

# Get the summary statistics:
summary(model_2)
```

```
##
## Call:
## lm(formula = GPA ~ CREDITS + factor(CLASS), data = Focus)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28823 -0.40598  0.01608  0.35991  1.21478
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.0154608  0.2150784  14.020   <2e-16 ***
## CREDITS                -0.0007517  0.0028222  -0.266    0.790
## factor(CLASS)Freshman  -0.2167055  0.1919765  -1.129    0.260
## factor(CLASS)Senior     0.1426732  0.1499860   0.951    0.343
## factor(CLASS)Sophomore -0.0874455  0.1274669  -0.686    0.494
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.512 on 195 degrees of freedom
## Multiple R-squared:  0.03475,    Adjusted R-squared:  0.01495
## F-statistic: 1.755 on 4 and 195 DF,  p-value: 0.1395
```

As computed above, the F-statistic is 1.755 on 4 and 195 degrees of freedom, with a p-value of 0.1395. Hence the model overall is insignificant at predicting GPA. Usually we would start over and choose different variables/build another model, but for the sake of this problem we will continue.

Below we will calculate the corresponding 95% confidence intervals of the above estimates.

```
# Calculate the 95% confidence intervals of the above estimates:
confint(model_2)
```

```
##                               2.5 %      97.5 %
## (Intercept)             2.591282180 3.439639354
## CREDITS                -0.006317736 0.004814241
## factor(CLASS)Freshman  -0.595322438 0.161911367
## factor(CLASS)Senior    -0.153129807 0.438476306
## factor(CLASS)Sophomore -0.338836179 0.163945170
```

As computed above, the y-intercept estimate was 3.0154608 with a 95% confidence interval of [2.591282180, 3.439639354]. The Credits slope estimate was -0.0007517 with a 95% confidence interval of [-0.006317736

0.004814241]. The Freshman class slope estimate was -0.2167055 with a 95% confidence interval of [-0.595322438 0.161911367]. The Sophomore class slope estimate was -0.0874455 with a 95% confidence interval of [-0.338836179 0.163945170]. Lastly, the Senior class slope estimate was 0.1426732 with a 95% confidence interval of [-0.153129807 0.438476306].

Putting this all together we obtain the least squares regression equation of:

$$\text{GPA} = -0.0008 \cdot \text{Credits} - 0.217 \cdot \text{Freshamn} - 0.0087 \cdot \text{Sophomore} + 0.143 \cdot \text{Senior} + 3.015$$

As computed above, the regression slope for Credits was -0.0007517 with a 95% confidence interval of [-0.006317736 0.004814241]. This means, on average, for every 1 unit increase in the amount of credits taken, the mean GPA of that student is estimated to decrease by about 0.0008 units over the sampled range of GPA scores, holding all other variables in the model fixed. Furthermore, the p-value for this estimate was 0.790 which means this estimate is not significant at the 5% significance level.

As computed above, the regression slope for Freshman class was -0.2167055 with a 95% confidence interval of [-0.595322438 0.161911367]. This means that the GPA of Juniors is higher than the GPA of Freshman by 0.2167055 on average, holding all other variables in the regression model constant. Furthermore, the p-value for this estimate was 0.260 which means this estimate is not significant at the 5% significance level.

As computed above, the regression slope for Sophomore class was -0.0874455 with a 95% confidence interval of [-0.338836179 0.163945170]. This means that the GPA of Juniors is higher than the GPA of Sophomores by 0.0874455 on average, holding all other variables in the regression model constant. Furthermore, the p-value for this estimate was 0.494 which means this estimate is not significant at the 5% significance level.

As computed above, the regression slope for Senior class was 0.1426732 with a 95% confidence interval of [-0.153129807 0.438476306]. This means that the GPA of Juniors is less than the GPA of Seniors by 0.1426732 on average, holding all other variables in the regression model constant. Furthermore, the p-value for this estimate was 0.343 which means this estimate is not significant at the 5% significance level.

As computed above, the regression y-intercept 0.0154608 with a 95% confidence interval of [2.591282180, 3.439639354]. This means that the estimated mean GPA is equal to about 0.0154608 when all of the predictor variables are zero meaning the student is a Junior taking zero credits. Furthermore, the p-value for this estimate was <2e-16 which means this estimate is practically zero and thus highly significant, especially at the 5% significance level. It is important to note however, that this intercept doesn't give us much value, because in reality if a student is taking zero credits, their GPA is zero/not able to be calculated.

Overall, due to the fact that none of the estimates are significant, and the insignificance of the model overall, this model should definitely not be used for GPA prediction.

3. **Using the Fish Consumption dataset on Canvas.** At 10% level of significance, perform logistic regression analysis using switch fishing as the dependent variable. Predictor variables include fish consumption, stress level and IQ. Answer the following questions:

- a) Write out the form of the model and identify which of the variables are positively associated when controlling for other variables.

Before we perform our logistic regression we will create a frequency table, a contingency table, and run a chi-square test of our one categorical variable; fish consumption, to see if it is reasonable to use in our logistic regression.

```
table(FC_Edited$FC)
```

```
##
##       No    Often Sometime
##       38      125       36
```

As we can see there are no cells with a low amount of observations, so now we will check if this holds when creating a contingency table with the switch fishing variable.

```
table(FC_Edited$Switch, FC_Edited$FC)
```

```
##
##       No Often Sometime
##   no  24    62       21
##   yes 14    63       15
```

As we can see, none of these cells contain less than ten observations so we are good to check if there is a significant relationship between these variables.

```
chisq.test(FC_Edited$Switch, FC_Edited$FC)
```

```
##
##  Pearson's Chi-squared test
##
## data:  FC_Edited$Switch and FC_Edited$FC
## X-squared = 2.5233, df = 2, p-value = 0.2832
```

As seen above, due to the p-value of 0.2832, there is not a statistically significant association between switch fishing and fish consumption. In the real world, we would have to critically assess if we should use the fish consumption variable or not, in this case we will proceed with using fish consumption as a predictor variable for completeness.

Below we will create the logistic model to perform logistic regression analysis using switch fishing as the dependent variable and fish consumption, stress level and IQ as the independent variables.

```
# Re-coding our dependent variable:
FC_Edited$Switch_cat <- revalue(FC_Edited$Switch, c("no" = "0", "yes" = "1"))

# Create the model for predicting switch fishing using fish consumption,
# stress level and IQ as the independent variables:
model_3 <- glm(factor(Switch_cat) ~ factor(FC) + IQ + Stress, data = FC_Edited,
               family = "binomial")

# Get the summary statistics:
summary(model_3)
```

```
##
## Call:
## glm(formula = factor(Switch_cat) ~ factor(FC) + IQ + Stress,
##     family = "binomial", data = FC_Edited)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4397  -1.1119  -0.8665   1.2085   1.7082
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.58113    2.05855  -2.225   0.0261 *
## factor(FC)Often     0.49138    0.38898   1.263   0.2065
## factor(FC)Sometime  0.12528    0.48532   0.258   0.7963
## IQ                  0.03429    0.01902   1.803   0.0714 .
## Stress              0.02775    0.01739   1.596   0.1105
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 274.74  on 198  degrees of freedom
## Residual deviance: 266.43  on 194  degrees of freedom
## AIC: 276.43
##
## Number of Fisher Scoring iterations: 4
```

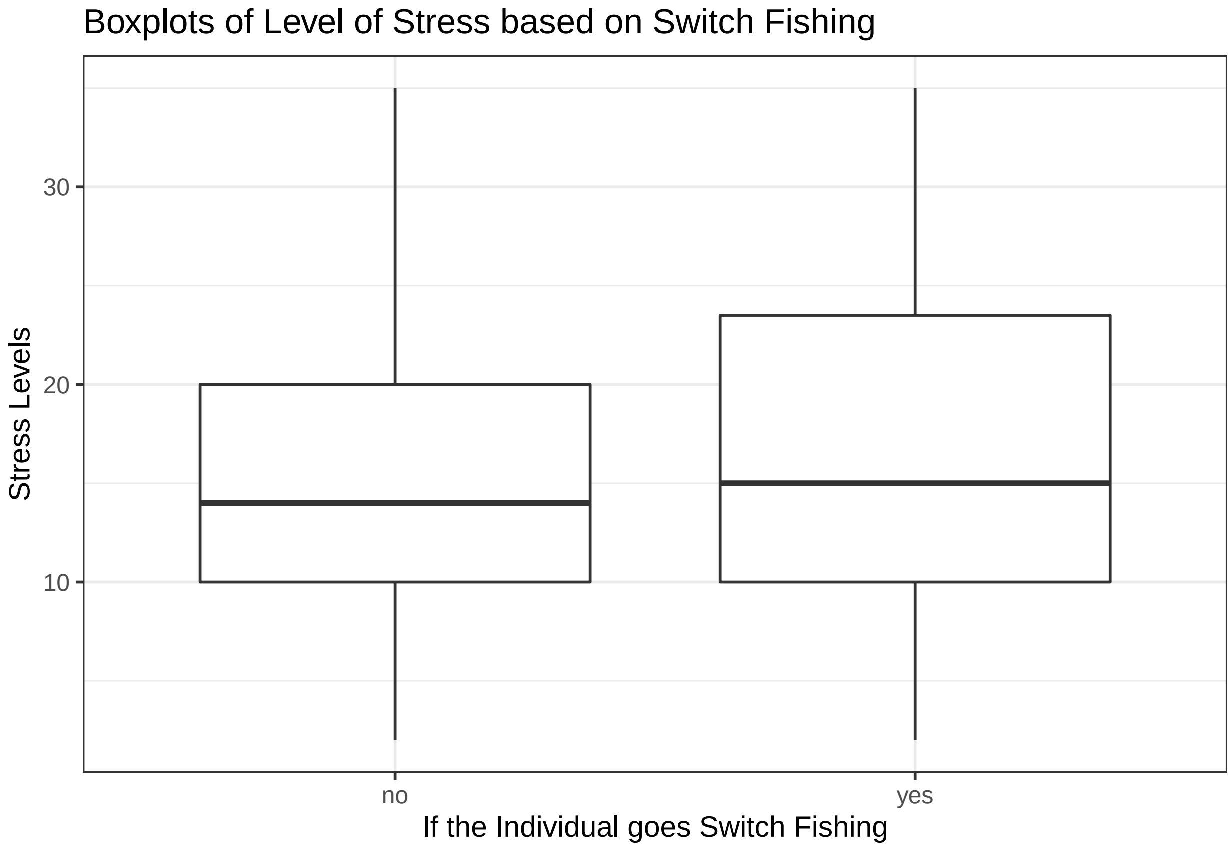
As we can see from above, the levels of the fish consumption variable Often and Sometimes, as well as the variables IQ and Stress all have positive slope coefficients. Hence all of these variables are positively associated when controlling for other variables. This means that for each unit increase of the above variables, the log odds ratio increases by the magnitude of the given slope.

- b) Examine each of the predictors. Are there any outliers that are likely to have a very large influence on the logistic regression model?

We will now examine each of the predictor variables for outliers to see if there are any data points that are likely to have a very large influence on our logistic regression model. Since the fish consumption variable is categorical, there is no sense of outliers and in the previous part we showed that the fish consumption has adequate cell values for both levels of switch fishing. Hence we will only analyze the Stress and IQ variables.

Looking for outliers in the Stress variable: Below we will make a boxplot of Stress versus the two levels of Switch Fishing to see if there are any outliers/influential points.

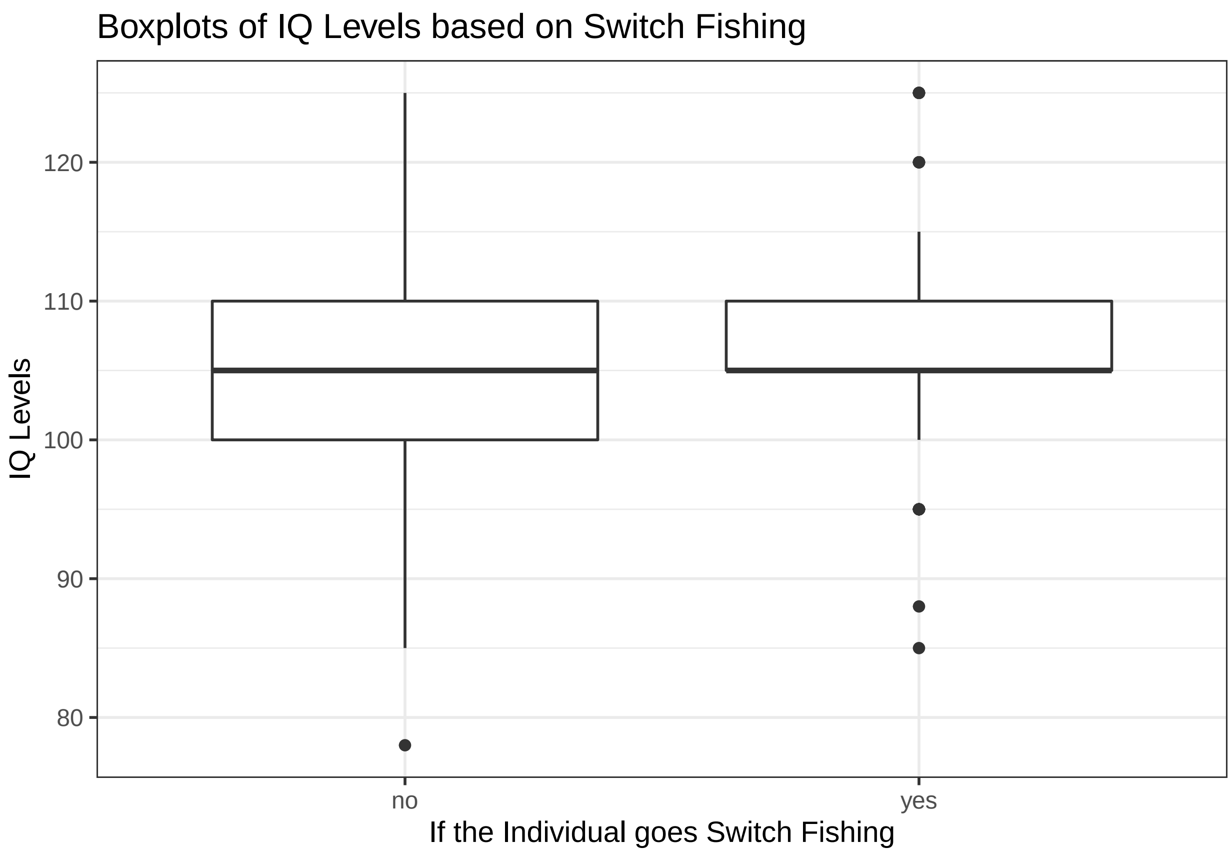
```
ggplot(data = FC_Edited, mapping = aes(x = Switch, y = Stress)) +
  geom_boxplot() +
  labs(title = "Boxplots of Level of Stress based on Switch Fishing",
       x = "If the Individual goes Switch Fishing",
       y = "Stress Levels") +
  theme_bw()
```



As can be seen by the above boxplots, there appears to be no significant outliers on either side of the lower or upper fence of either boxplots.

Looking for outliers in the IQ variable: Below we will make a boxplot of IQ versus the two levels of Switch Fishing to see if there are any outliers/influential points.

```
ggplot(data = FC_Edited, mapping = aes(x = Switch, y = IQ)) +
  geom_boxplot() +
  labs(title = "Boxplots of IQ Levels based on Switch Fishing",
       x = "If the Individual goes Switch Fishing",
       y = "IQ Levels") +
  theme_bw()
```



As can be seen by the above boxplots, unlike the Stress variable, there are outliers for both levels of Switch fishing. For individuals who don't go switch fishing there is one low outlier. Furthermore, for individuals who do go switch fishing, there are two high outliers, and three low outliers. All of these outliers could be influential. This means that including these points in the logistic regression could have huge impacts on the regression coefficients.

- c) Give a brief interpretation of the odds ratios and show how to compute them from the R output.

Below we will compute the estimated odds ratios of the given variables as well as there corresponding confidence intervals.

In order to compute the odds ratios from the logit model, we will exponentiate all of the numbers given from the logit model. This process is shown below.

```
# Using Standard Errors:
exp(cbind(OR = coef(model_3), confint.default(model_3, level = 0.90)))
```

```
##                             OR          5 %     95 %
## (Intercept)         0.01024331 0.000346666 0.302670
## factor(FC)Often     1.63457650 0.862060588 3.099365
## factor(FC)Sometime  1.13346766 0.510174744 2.518253
## IQ                  1.03488784 1.003006620 1.067782
## Stress              1.02814081 0.999146269 1.057977
```

Since the answer no was used as the reference group for fish consumption, people who often consume fish had higher odds of switch fishing compared to people who don't consume fish. In particular, the odds ratio was 1.63457650, and the corresponding 90% confidence interval was [0.862060588 3.099365]. Note that since 1 is contained in the interval, these results are insignificant at the 10% level.

Since the answer no was used as the reference group for fish consumption, people who sometimes consume fish had higher odds of switch fishing compared to people who don't consume fish. In particular, the odds ratio was 1.13346766, and the corresponding 90% confidence interval was [0.510174744 2.518253]. Note that since 1 is contained in the interval, these results are insignificant at the 10% level.

As was computed above, the IQ odds ratio estimate was 1.03488784 with a 90% confidence interval of [1.003006620, 1.067782]. Hence we can say that for a one unit increase in IQ, the odds of that person switch fishing (versus not switch fishing) increases by a factor of 1.03, holding other variables in the model fixed. Notice that since 1 isn't included in this interval, this result is significant at the 10% significance level.

As was computed above, the Stress odds ratio estimate was 1.02814081 with a 90% confidence interval of [0.999146269 1.057977]. Hence we can say that for a one unit increase in Stress, the odds of that person switch fishing (versus not switch fishing) increases by a factor of 1.02814081, holding other variables in the model fixed. Notice that since 1 is included in this interval, this result is insignificant at the 10% significance level.

4. **Using the Focus dataset on Canvas** - determine the regression equation for predicting GPA based on age, gender and credits.

Before answering the questions below, we will create a multiple regression model for predicting GPA based on age, gender and credits.

```r
# Setup the model for predicting GPA based on age, gender and credits:
model_4 <- lm(GPA ~ AGE + CREDITS + factor(SEX), data = Focus)

# Get the summary statistics:
summary(model_4)
```

```
##
## Call:
## lm(formula = GPA ~ AGE + CREDITS + factor(SEX), data = Focus)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21681 -0.37086  0.00697  0.32877  1.17585
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.356240   0.386381  11.274  < 2e-16 ***
## AGE          -0.084202   0.020821  -4.044 7.55e-05 ***
## CREDITS       0.005786   0.001261   4.589 7.95e-06 ***
## factor(SEX)M -0.120611   0.070774  -1.704   0.0899 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4893 on 196 degrees of freedom
## Multiple R-squared:  0.1139, Adjusted R-squared:  0.1003
## F-statistic: 8.398 on 3 and 196 DF,  p-value: 2.801e-05
```

As can be seen the F-statistic is 8.398 on 3 and 196 degrees of freedom respectively, with a very low p-value of 2.801e-05. Thus at the 5% level of significance, and any other reasonable level of significance for that matter, we have evidence that the independent variables Sex, Credits, and Age significantly predict GPA as a whole.

- a) Identify the dependent and independent variable(s)?

Since we are using age, gender and credit to predict GPA, it follows that the independent/predictor variables are age, gender, and credits, and the dependent/outcome variable is GPA.

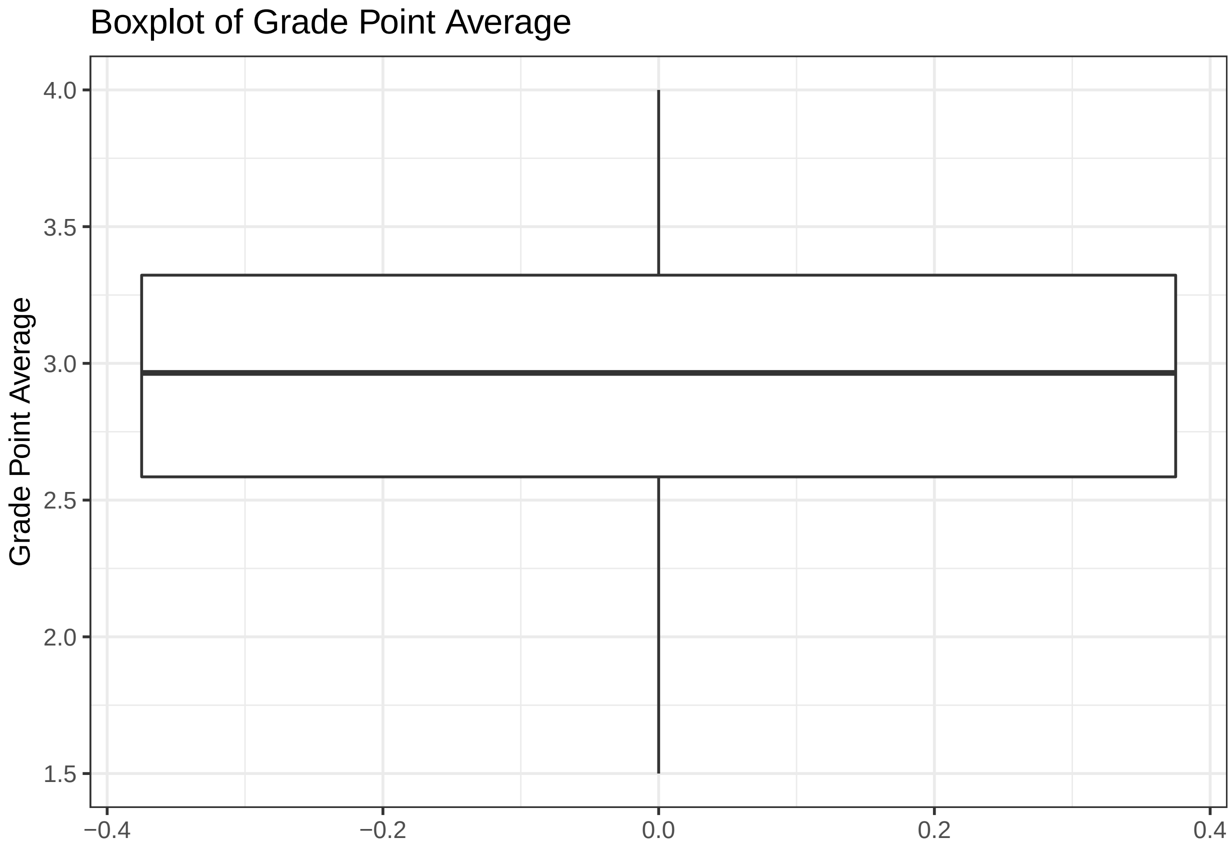
- b) Is the dependent variable normally distributed? Justify your answer? If not, make the necessary transformation and use the transformed variable for the rest of the analysis.

```
shapiro.test(Focus$GPA)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Focus$GPA
## W = 0.98968, p-value = 0.1601
```

As can be seen from the above Shapiro-Wilk test, since the p-value is 0.1601 which is greater than 0.05, we fail to reject the null hypothesis and we assume that the dependent variable is approximately normally distributed.

```
ggplot(data = Focus, mapping = aes(y = GPA)) +
  geom_boxplot() +
  labs(title = "Boxplot of Grade Point Average",
       y= "Grade Point Average") +
  theme_bw()
```



Furthermore, as can be seen by the above boxplot, the dependent variable GPA seems to be approximately normal, although a small left skew is present.

- c) What is the coefficient of determination and what does it mean?

Since R squared increases every time you add an independent variable to the model, the R squared value always increases. Thus, since we are running a multiple regression model our coefficient of determination is the adjusted R squared. In our model, the adjusted R squared was 0.1003. Our adjusted R squared value means, adjusting for the number of independent variables and their corresponding significance, around 10% of the variability in GPA can be explained by Sex, Credits, and Age.

- d) Find the least-squares estimates for the regression line.

Below we compute the corresponding 95% confidence intervals for the estimates.

```
confint(model_4)
```

```
##                    2.5 %       97.5 %
## (Intercept)   3.594242482  5.118237857
## AGE          -0.125264445 -0.043139145
## CREDITS       0.003299627  0.008273339
## factor(SEX)M -0.260186897  0.018965819
```

As computed above, the y-intercept estimate was 4.356240170 with a 95% confidence interval of [3.594242482, 5.118237857]. The Credits slope estimate was 0.005786483 with a 95% confidence interval of [0.003299627, 0.008273339]. The Age slope estimate was -0.084201795 with a 95% confidence interval of [-0.125264445, -0.043139145]. Lastly, the Male class slope estimate was -0.120610539 with a 95% confidence interval of [-0.260186897, 0.018965819].

Hence the regression equation is:

$$\text{GPA} = 0.006 \cdot \text{Credits} - 0.084 \cdot \text{Age} - 0.121 \cdot \text{Male} + 4.356$$

- e) Interpret the value of the slopes and intercept in the context of the problem. To get full credits, provide the estimate, standard error, p-value and 95% confidence interval for the intercept and slopes. Also, which variables are significant predictors? Explain.

As computed above, the regression slope for Credits was 0.005786483 with standard error 0.001261 and with a 95% confidence interval of [0.003299627, 0.008273339]. This means, on average, for every 1 unit increase in the amount of credits taken, the mean GPA of that student is estimated to increase by about 0.005786483 units over the sampled range of GPA scores, holding all other variables in the model fixed. Furthermore, the p-value for this estimate was 7.95e-06 which means this estimate is highly significant at the 5% significance level.

As computed above, the regression slope for Age class was -0.084202 with a standard error of 0.020821 and with a 95% confidence interval of [-0.125264445, -0.043139145]. This means that on average, for every 1 unit increase in the age of a student, the mean GPA of that student is estimated to decrease by about 0.084202 units over the sampled range of GPA scores, holding all other variables in the model fixed. Furthermore, the p-value for this estimate was 7.55e-05 which means this estimate is highly significant at the 5% significance level.

As computed above, the regression slope for Male sex was was -0.120610539 with a standard error of 0.070774 and with a 95% confidence interval of [-0.260186897, 0.018965819]. This means that the GPA of Females is higher than the GPA of Males by 0.120610539 on average, holding all other variables in the regression model constant. Furthermore, the p-value for this estimate was 0.0899 which means this estimate is not significant at the 5% significance level.

As computed above, the regression y-intercept was 4.356240170 with a standard error of 0.386381 with a 95% confidence interval of [3.594242482, 5.118237857]. This means that the estimated mean GPA is equal to about 4.356240170 when all of the predictor variables are zero meaning the student is a Female taking zero credits and aged zero. Furthermore, the p-value for this estimate was <2e-16 which means this estimate is practically zero and thus highly significant, especially at the 5% significance level. It is important to note however, that this intercept doesn't give us much value, because in reality if a student is taking zero credits and is of age zero, their GPA is zero/not able to be calculated.

- f) Predict the GPA for a female aged 20 student who took 80 credits.

Below is code to predict the GPA for a female aged 20 student who took 80 credits.

```
student <- data.frame(AGE = c(20), CREDITS = 80, SEX = "F")
predict(model_4, student)
```

```
##        1
## 3.135123
```

The equation for this prediction is $\text{GPA} = 0.006 \cdot \text{Credits} - 0.084 \cdot \text{Age} - 0.121 \cdot \text{Male} + 4.356$. With our specific values plugged in we get $\text{GPA} = 0.006 \cdot 80 - 0.084 \cdot 20 + 4.356 = 3.135123$. Hence our predicted GPA for a female aged 20 student who took 80 credits is 3.135123.

- g) Using information from part (f), find the residual if GPA is 3.5. Is this an overestimation or underestimation. Justify your answer.

If the actual/observed GPA of the student from part (f) is 3.5, and the formula for a residual is Residual = Observed value - Predicted value, then the residual of the prediction from part (f) is: 3.5 - 3.135123 = 0.364877. In our case, the prediction from part (f) is an underestimate of the true GPA of that student.

- h) Are the residuals as you would expect for a good model?

In order to check if the residuals are as we would expect for a good model we will run a residual analysis. The conditions we will be checking are: linear relationship between the response and the predictor variables, normality of the residuals, equal variance of the residuals, and no outliers. As stated in the problem description we will be running these tests at the 5% significance level.

Checking for Linearity:

```
# Run a Rainbow test to check the linearity of the variables:
raintest(model_4)
```

```
##
##  Rainbow test
##
## data:  model_4
## Rain = 1.0055, df1 = 100, df2 = 96, p-value = 0.4898
```

As can be seen from the above Rainbow test, since the p-value is 0.4898 which is greater than 0.05, hence we fail to reject the null hypothesis and we assume that the relationship between the independent and dependent variables is linear.

Checking for Normality of the Residuals:

```r
# Calculate the residuals:
resid_2 <- residuals(model_4)

# Run a Shapiro test to check the normality of the residuals:
shapiro.test(resid_2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_2
## W = 0.99119, p-value = 0.2649
```

As can be seen from the above Shapiro-Wilk test, since the p-value is 0.2649 which is greater than 0.2649, hence we fail to reject the null hypothesis and we assume that the residuals are normally distributed.

Checking for equal variance:

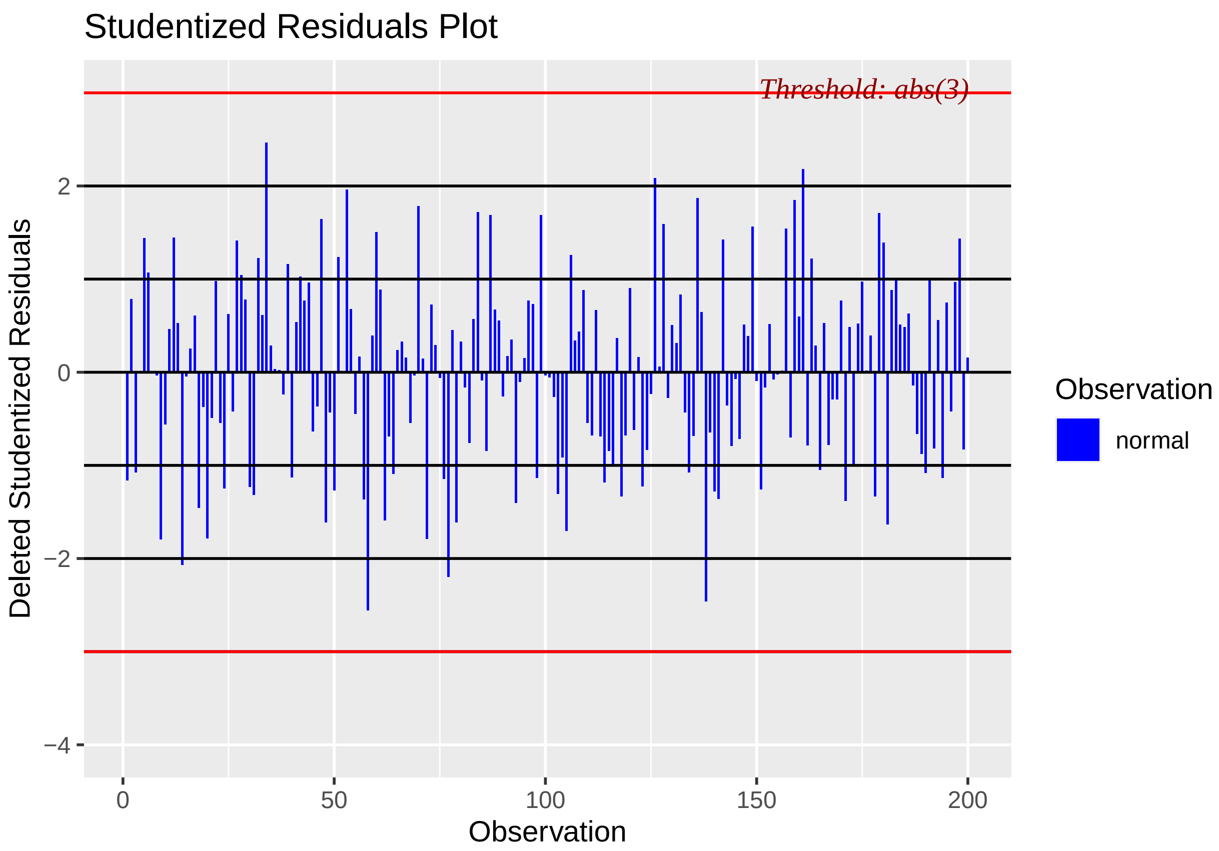
```r
# Run the Breusch Pagan Test for Heteroskedasticity to test for equal variance:
ols_test_breusch_pagan(model_4)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##              Data
##  -------------------------------
##  Response : GPA
##  Variables: fitted values of GPA
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    1.972315
##  Prob > Chi2   =    0.1602023
```

As can be seen from the above Breusch Pagan Test for Heteroskedasticity, since the p-value is 0.1602023 which is greater than 0.05, we fail to reject the null hypothesis, thus we don't have evidence that the equal/constant variance assumption is violated.

Checking for no outliers:

```r
# Create a Studentized Residual Plot to check for outliers:
ols_plot_resid_stud(model_4)
```



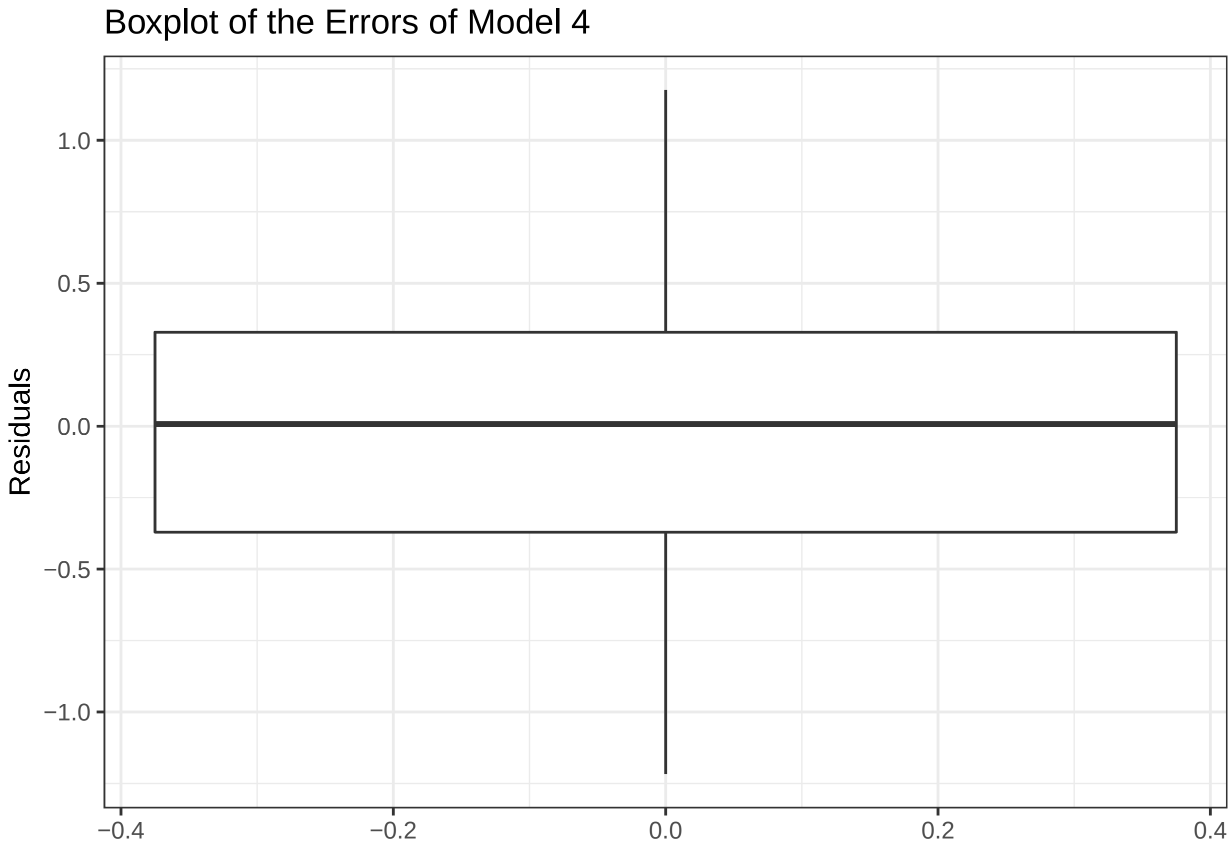
As can be seen from the above studentized residual plot, there are no outliers present in the data.

Due to the fact that we violated none of the residual or initial assumptions, it is safe to assume that the assumptions for regression inferences are not violated and we would have some justification on using this model for prediction. However, we know that some of the predictor variables are not significant thus we would need to do more work.

- i) Is there evidence of normality of errors, auto-correlation and no multicollinearity?

As computed above we have evidence of the normality of errors, but we will show a boxplot to further emphasize this:

```
ggplot() +
  geom_boxplot(mapping = aes(y = resid_2)) +
  labs(title = "Boxplot of the Errors of Model 4",
       y= "Residuals") +
  theme_bw()
```



This looks very symmetric, thus as the Shapiro test stated, we are safe to assume the normality of the errors.

Checking for no autocorrelation:

```
# Run a Durbin Watson test to check autocorrelation:
dwtest(model_4)
```

```
##
##  Durbin-Watson test
##
## data:  model_4
## DW = 2.0225, p-value = 0.5668
## alternative hypothesis: true autocorrelation is greater than 0
```

As computed from the above Durbin-Watson test, the p-value is 0.5668 which is greater than 0.05, thus we fail to reject the null hypothesis and we can assume that there is no autocorrelation.

Checking for no multicollinaerity:

```
# Run a vif test to see if there is any multicollinearity:
vif(model_4)
```

```
##         AGE     CREDITS factor(SEX)
##    1.686409    1.698870    1.011989
```

Since the multicollinearity of each variable is below 5, we are safe to conclude that none of our variables are too closely related to each other

- j) Compare this particular model to another competing model with only two in- dependent variables: age and credits. What can you say by comparing this with actual model (with three independent variables)?

Below we will create a new model with only two independent variables and compare it to our model with three independent variables.

Since the only insignificant variable in model 4 was Sex, we will remove that variable when creating model 5.

```
# Create the new model:
model_5 <- lm(GPA ~ AGE + CREDITS, data = Focus)

compareLM(model_4, model_5)
```

```
## $Models
##   Formula
## 1 "GPA ~ AGE + CREDITS + factor(SEX)"
## 2 "GPA ~ AGE + CREDITS"
##
## $Fit.criteria
##   Rank Df.res   AIC  AICc   BIC R.squared Adj.R.sq   p.value Shapiro.W
## 1    4    196 287.7 288.0 304.2    0.1139  0.10030 2.801e-05    0.9912
## 2    3    197 288.6 288.8 301.8    0.1008  0.09165 2.858e-05    0.9902
##   Shapiro.p
## 1    0.2649
## 2    0.1922
```

As computed above since model 5 has a lower AIC and BIC than model 4, model 5, the simpler model, is the better one. It is important to note however that differences between these two models is little to none, and in fact the adjusted R squared is actually higher for model 4. In this case either model could be used for prediction, but the smaller model is preferred.

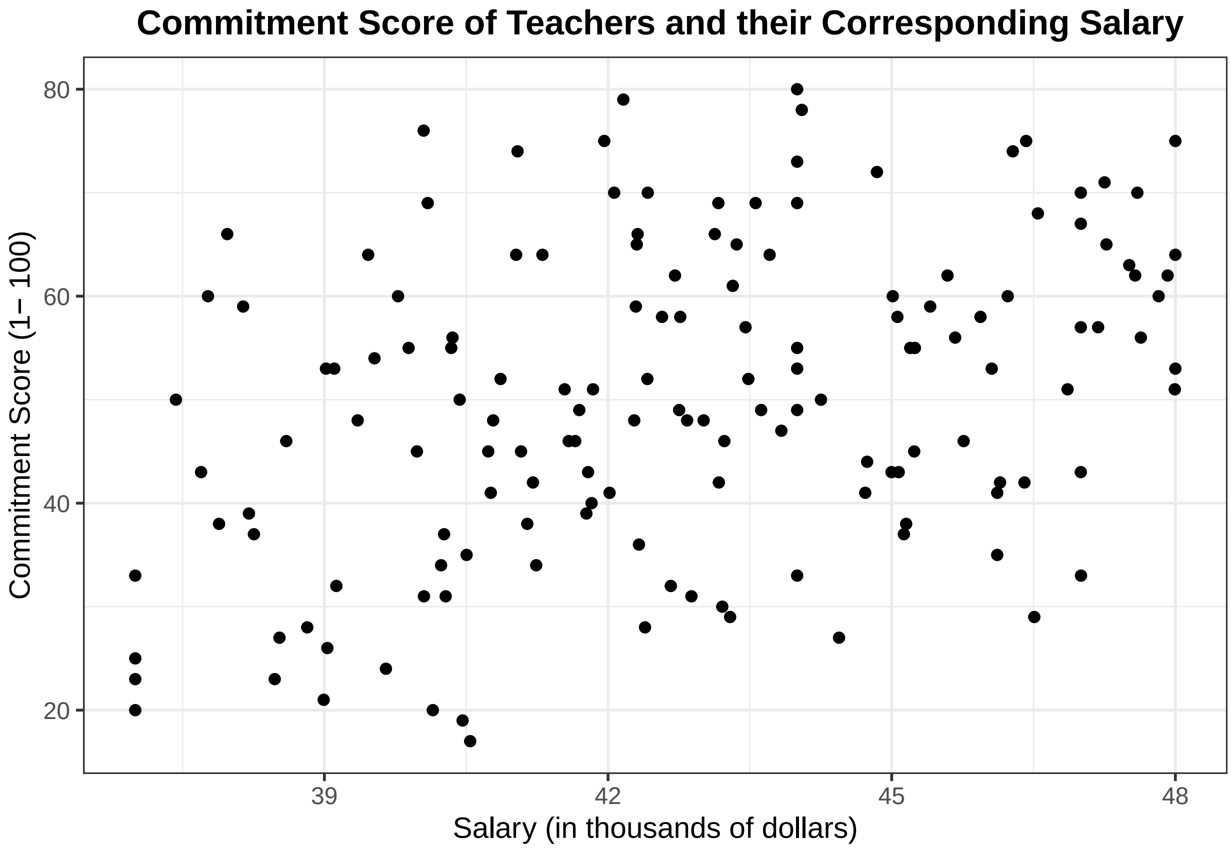
5. The dataset for this question is **Handout 1**. Select any four variables of your choice and use one visualization plot to tell a more complete and compelling story of the dataset masking use of all four variables selected. Consider using at least **color, faceting, theme** among many others in ggplot2. Write at least two paragraphs for this. One explaining why you choose those variables and the other what you see from the graph.

The following plots are constructed using the data in **Handout 1** which are from a "hypothetical study" of variables related to a teacher's intention to stay in the profession. The data were collected on 150 teachers who had been teaching for 5 years or less. These teachers responded to a questionnaire about their commitment to teaching and received a score (1 - 100) based on certain aspects of their commitment. In particular, the variables that show up in the below plots are `COMMIT`, `SALARY`, `SCHTYPE`, and lastly `SEX`.

Before I dive into this combined plot though, we will look at the normal scatter-plot with no faceting and color to see how much insight we can gain from our choices of faceting and coloring.

Plot before coloring and faceting:

```
# Plot the scatter-plot of commitment versus salary of teachers:
ggplot(data = Handout_1,
       mapping = aes(x = SALARY, y = COMMIT)) +
  geom_point() +
  labs(x = "Salary (in thousands of dollars)",
       y = "Commitment Score (1- 100)",
       title = "Commitment Score of Teachers and their Corresponding Salary") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.55, face = "bold", size = 13))
```



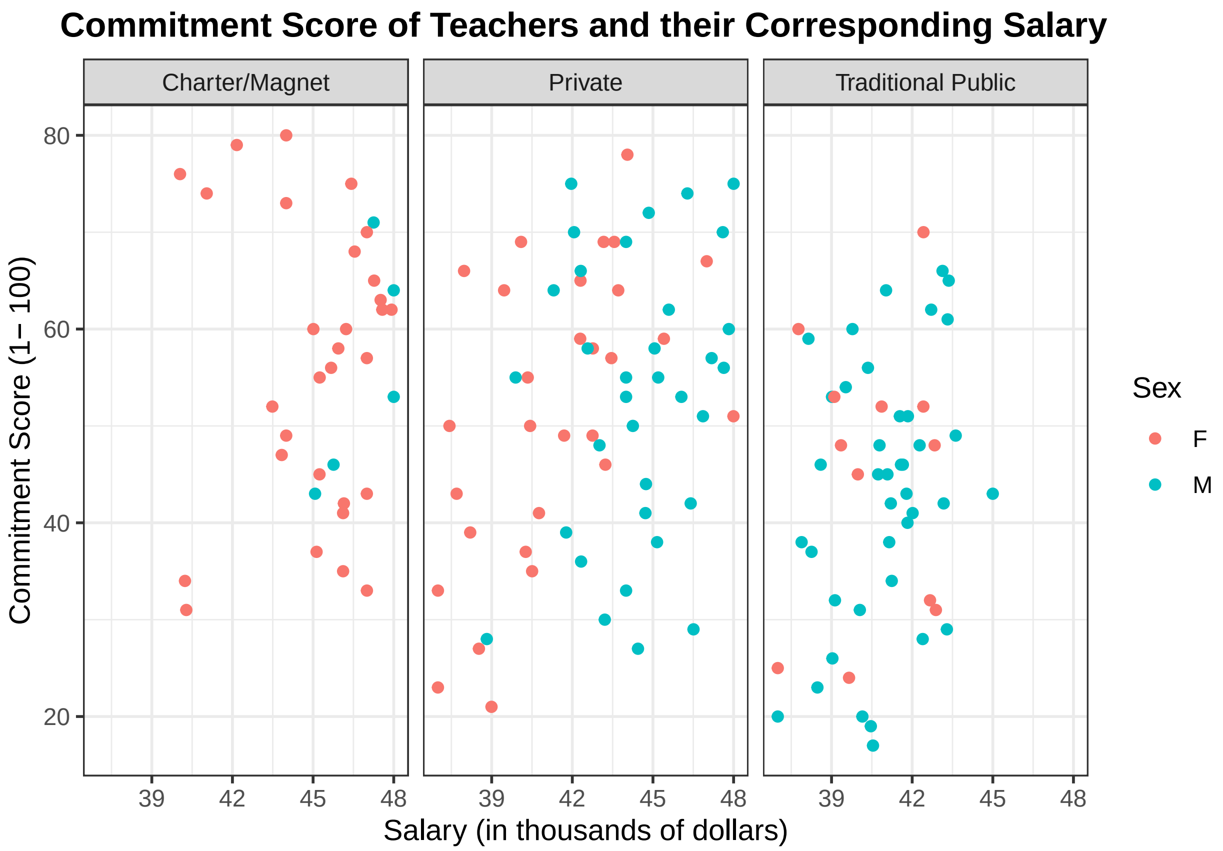
Findings:
As can be seen by the above scatter-plot comparing a teachers commitment score and their corresponding salary, we see that in general, as a teachers salary increases, their commitment score increases as well. For completeness, the range of the `SALARY` variable ranges from around 36,000 to 48,000, while the `COMMIT` variable ranges from around 15 to 65. This means, that despite the increase in commitment score with each increase in salary, this change isn't that much in terms of the magnitude of the score. Also, compared to the data in `Handout 2` it is obvious that this dataset is the same except for the missing outlier that was in the top left corner.

Plot with 4 variables:

```
# Recode categorical variables:
Handout_1 <- Handout_1 %>%
  mutate(Sex_cat = case_when(SEX == 1 ~ "M", SEX == 2 ~ "F"),
         Schtpye_cat = case_when(SCHTYPE == 1 ~ "Traditional Public",
                                 SCHTYPE == 2 ~ "Private",
                                 SCHTYPE == 3 ~ "Charter/Magnet"))

# Plot the scatter-plot of commitment versus salary of teachers faceted by
# school type and colored by sex:
```

```r
ggplot(data = Handout_1,
       mapping = aes(x = SALARY, y = COMMIT, color = Sex_cat)) +
  geom_point() +
  facet_wrap(~Schtpye_cat) +
  labs(x = "Salary (in thousands of dollars)",
       y = "Commitment Score (1- 100)",
       title = "Commitment Score of Teachers and their Corresponding Salary",
       color = "Sex") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.55, face = "bold", size = 13))
```



Why I choose the above plot:
As stated above, the variables I chose to analyze were `COMMIT` which corresponds to the commitment score of a given teacher in the sample, `SALARY` which corresponds to the annual salary of a given teacher in the sample, `SEX` which corresponds to the sex of a given teacher in the sample, and lastly `SCHTYPE` the school type taught at of a given teacher in the sample. In particular I decided to put `COMMIT` on the y-axis, `SALARY` on the x-axis, faceted on the `SCHTYPE` variable, with the coloring of the points dependent on `SEX`. First off, I decided to make my y-axis `COMMIT` and my x-axis `SALARY`, because I wanted to find out what variables are able to predict `COMMIT`. After looking at the data, one clear choice for my predictor variable was `SALARY`. Hence my main goal of this analysis was to answer "How does the annual salary of a teacher predict the commitment score of a teacher in this sample?" To dig deeper into this question and truly understand the relationship/nuances between the variables, I decided to facet on `SCHTYPE` in order to see how the relationship changes over the different school types, which in essence is how `COMMIT` changes based on different settings. Lastly, I decided to color each point in the scatter-plots based on `SEX` to see in each school type, how does the relationship differ between sexes.

Findings:
As can be seen in the initial scatter-plot, there is a moderate positive linear relationship between a teachers annual salary and their corresponding commitment score based on the data from the hypothetical teacher sample. Once we take a look at the faceted scatter-plots, we see that this trend somewhat holds up for the Public and Private class types. However, when we take a look at the Charter school type, there seems to be no apparent linear association between the two variables. One reason why this trend isn't as apparent as in the other faceted cases is because charter school are autonomous, meaning they aren't held federally accountable. Instead they are held publicly accountable by the families who choose to enroll their children into these schools. Thus they have more flexibility in their operations which leads to more input from the teachers and thus happier, better paid, and more committed teachers in general. On the basis of sex, the trends are opposite then was the case in the association explained above. On the basis of sex, we can see that, on average, females seem to be more committed than their male counterparts. Furthermore, we can see that, from this sample, majority of charter teachers are females.