Homework2

Rachel Holman

2023-06-24

1. The data set mammals from the MASS package contains the average brain and body weights for 62 species of land mammals. We wish to see how body weight (x) could explain the brain weight (y) of land mammals.

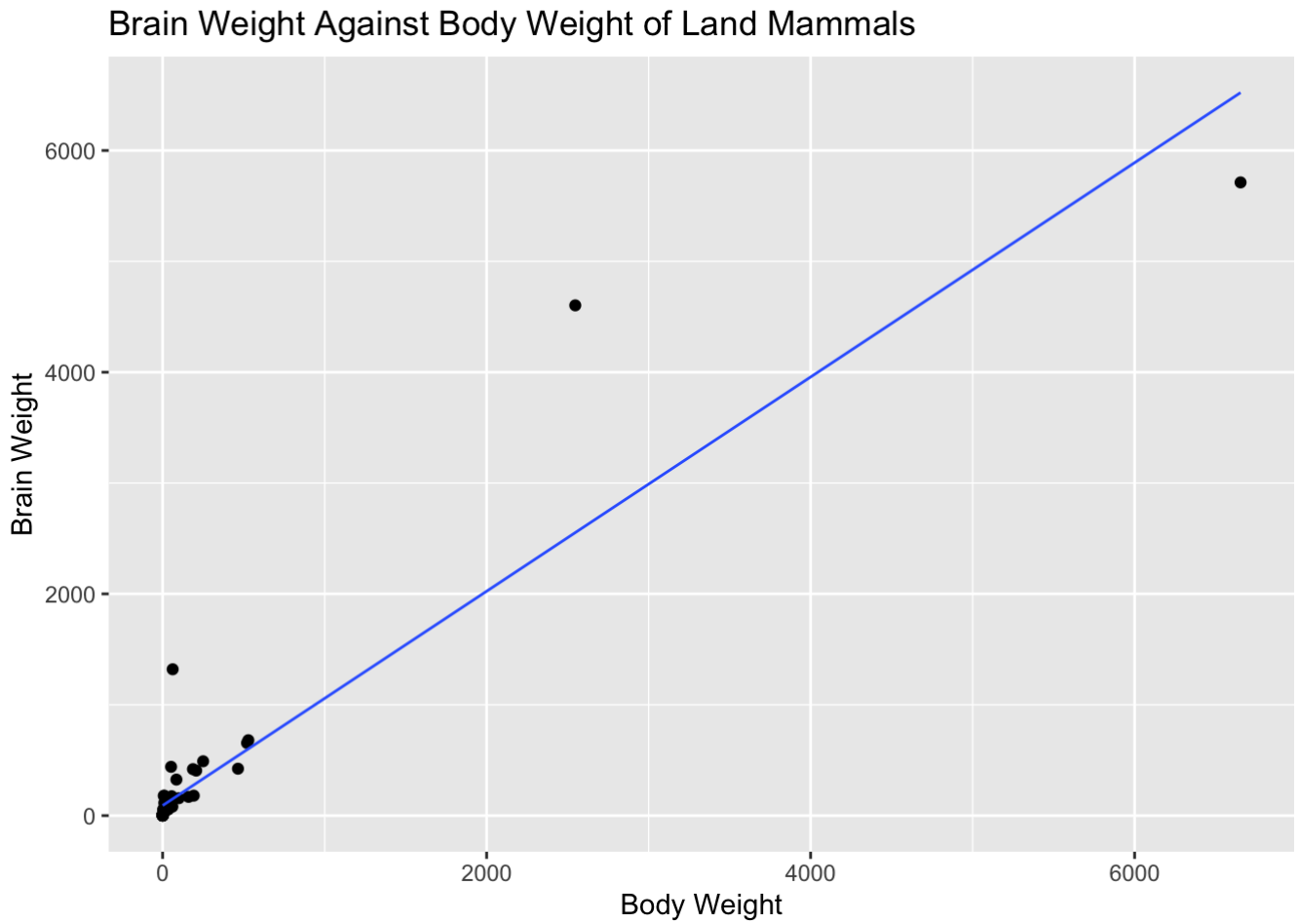
```
library(MASS)
head(mammals)
```

```
##                    body brain
## Arctic fox        3.385  44.5
## Owl monkey        0.480  15.5
## Mountain beaver   1.350   8.1
## Cow             465.000 423.0
## Grey wolf        36.330 119.5
## Goat             27.660 115.0
```

a. **Create a scatter plot of brain weight against body weight of land mammals. Comment on the appearance of the plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?**

```
ggplot(mammals, aes(x=body,y=brain))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE, linewidth=0.5)+
  labs(x="Body Weight", y="Brain Weight", title="Brain Weight Against Body Weight of Lan
d Mammals")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



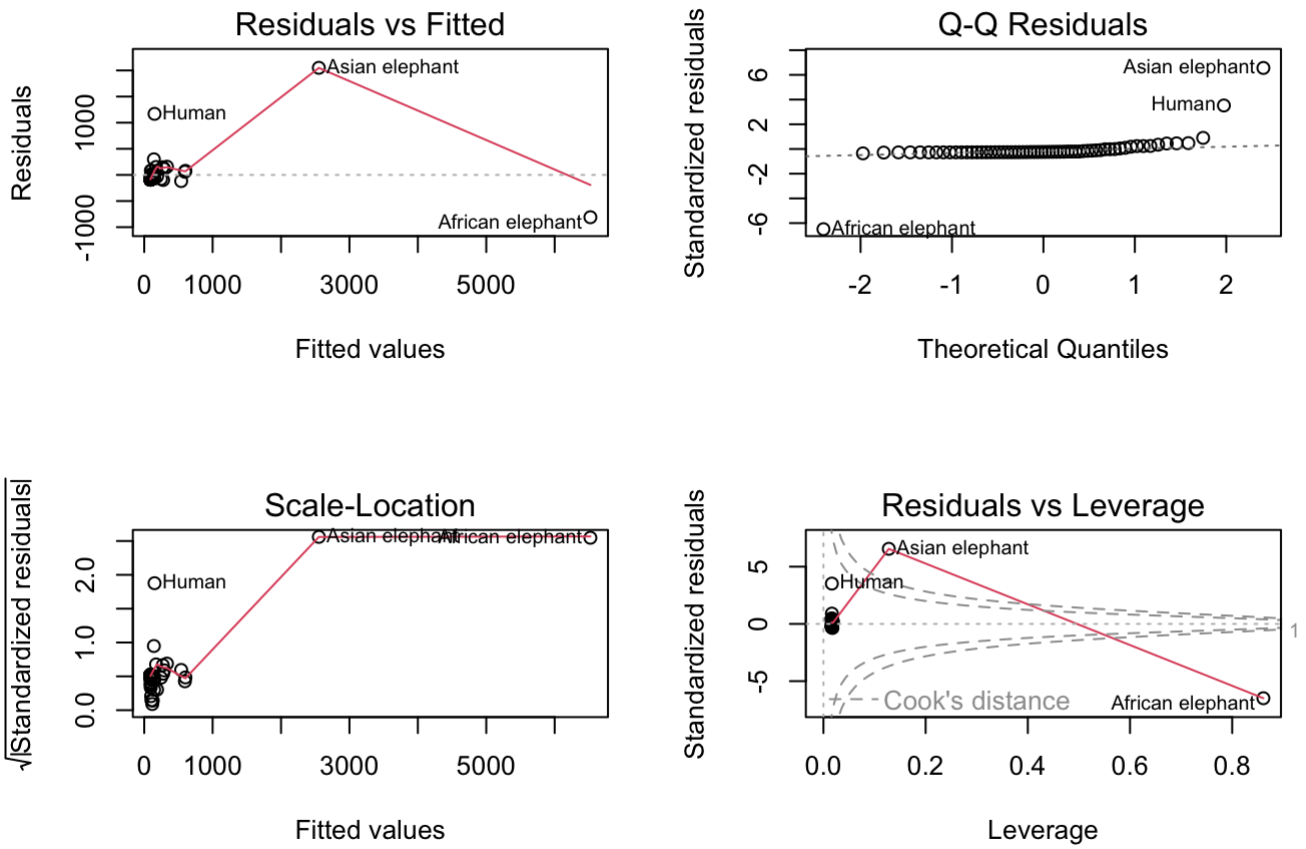
Generally speaking, there seems to be an increasing association between body weight and brain weight. The relationship appears to be more logarithmic than linear, but it is hard to tell because there are a few outliers that make the graph difficult to interpret.

To assess assumption 1, the data points should be evenly scattered on both sides of the regression line, as we move from left to right. We do not see this in the scatterplot, so assumption 1 is not met. When body weight is between 1000 and 6000 the data point(s) are above the line. When age is above 6000, the data point(s) are below the line.

To assess assumption 2, the vertical spread of the data points should be constant as we move from left to right. The spread seems to be increasing as we move from left to right (or in other words, the spread is increasing as the response increases), so assumption 2 is not met.

b. **Fit a simple linear regression to the data, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?**

```
mammalslm<-lm(brain~body, data=mammals)
par(mfrow = c(2, 2))
plot(mammalslm)
```



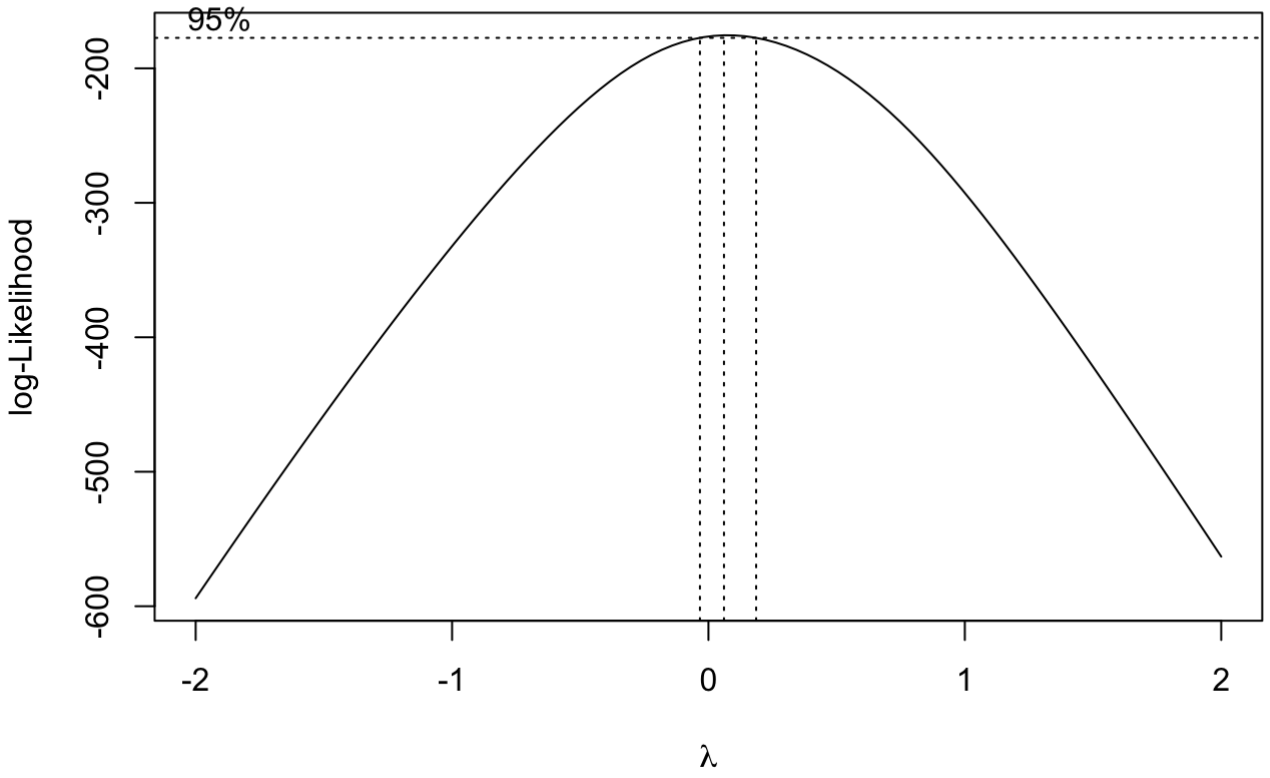
- The first plot (top left) is the **residual plot**, with residuals on the y-axis and fitted values on the x-axis. The residual plot can be used to address assumptions 1 and 2. A red line is overlayed to represent the average value of the residuals for differing values along the x-axis. This line should be along the x-axis without any apparent curvature to indicate the form of our model is reasonable. This is not what we see, as we see a clear curved pattern. So *assumption 1 is not met*. For assumption 2, we want to see the vertical spread of the residuals to be fairly constant as we move from left to right. We do not see this in the residual plot; the vertical spread increases as we move from left to right, so *assumption 2 is not met*.
- The second plot (top right) is the normal probability plot (also called a **QQ plot**), and addresses assumption 4. If the residuals are normal, the residuals should fall along the 45 degree line. The regression model is fairly robust to this assumption though; the normality assumption is the least crucial of the four. There are clearly very influential outliers that cause the QQplot to deviate at the tails so *assumption 4 may not be met*.
- The third plot (bottom left) is a plot of the square root of the absolute value of the standardized residuals against the fitted values (**scale-location**). This plot should be used to assess assumption 2, the constant variance assumption. A red line is overlayed to represent the average value on the vertical axis for differing values along the x-axis. If the variance is constant, the red line should be horizontal and the vertical spread of the plot should be constant. It is clear that assumption 2 is certainly not met, which tell a similar story to the first plot.
- The last plot (bottom right) is a plot to identify **influential outliers**. Data points that lie in the contour lines with large Cook's distance are influential. Two of our data points have Cook's distance greater than 1 and are therefore are flagged as influential.

c. **Based on your answers to parts 1a and 1b, do we need to transform at least one of the variables? Briefly explain.**

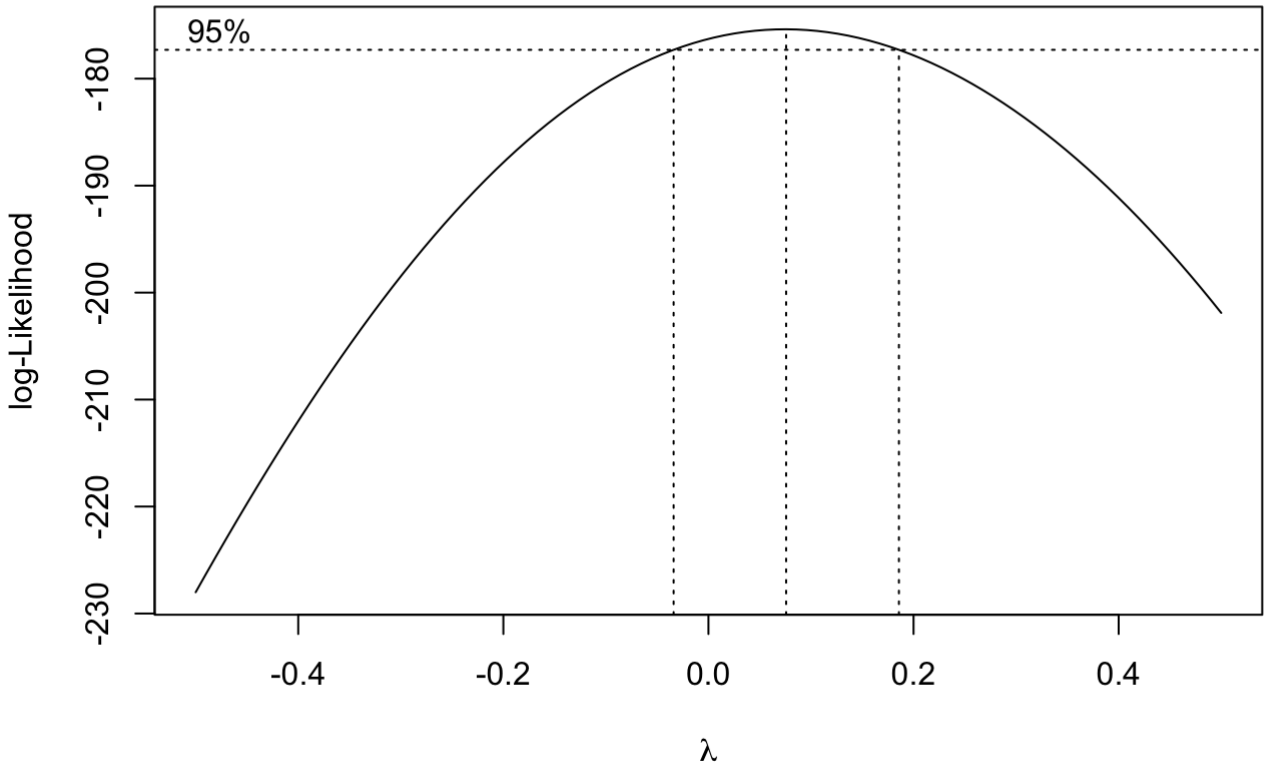
Because assumptions 1 and 2 (errors have mean 0 and errors constant variance) are both violated, we want to first transform the response variable to stabilize the variance, and once we fix the variance we can solve the non-linearity needed for assumption 1 by transforming the predictor variable.

d. **For the simple linear regression in part 1b, create a Box Cox plot. What transformation, if any, would you apply to the response variable? Briefly explain.**

```
boxcox(mammalslm)
```




```
boxcox(mammalslm, lambda = seq(-0.5, 0.5, 1/10))
```



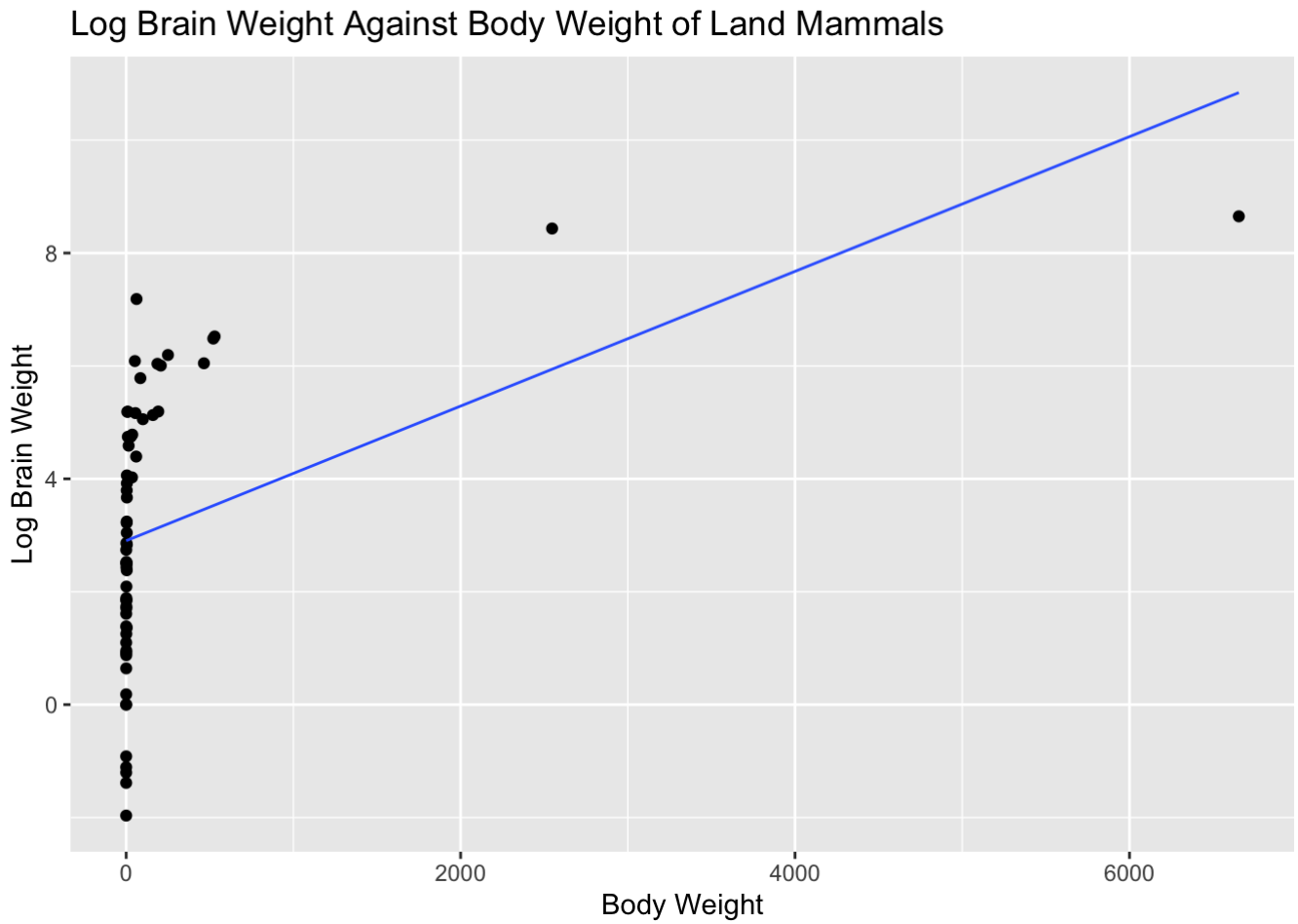
Because 0 lies in the confidence interval (CI), we can choose $\lambda = 0$ to log transform the response variable to get $y_* = \log(y)$. This is an ideal solution because log transformations are interpretable and the boxcox supports this solution.

e. **Apply the transformation you specified in part 1d, and let y_* denote the transformed response variable. Create a scatterplot of y_* against x. Comment on the appearance of the plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?**

```
logbrain<-log(mammals$brain)
mammals2<-data.frame(mammals,logbrain)

ggplot(mammals2, aes(x=body,y=logbrain))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE, linewidth=0.5)+
  labs(x="Body Weight", y="Log Brain Weight", title="Log Brain Weight Against Body Weigh
t of Land Mammals")
```

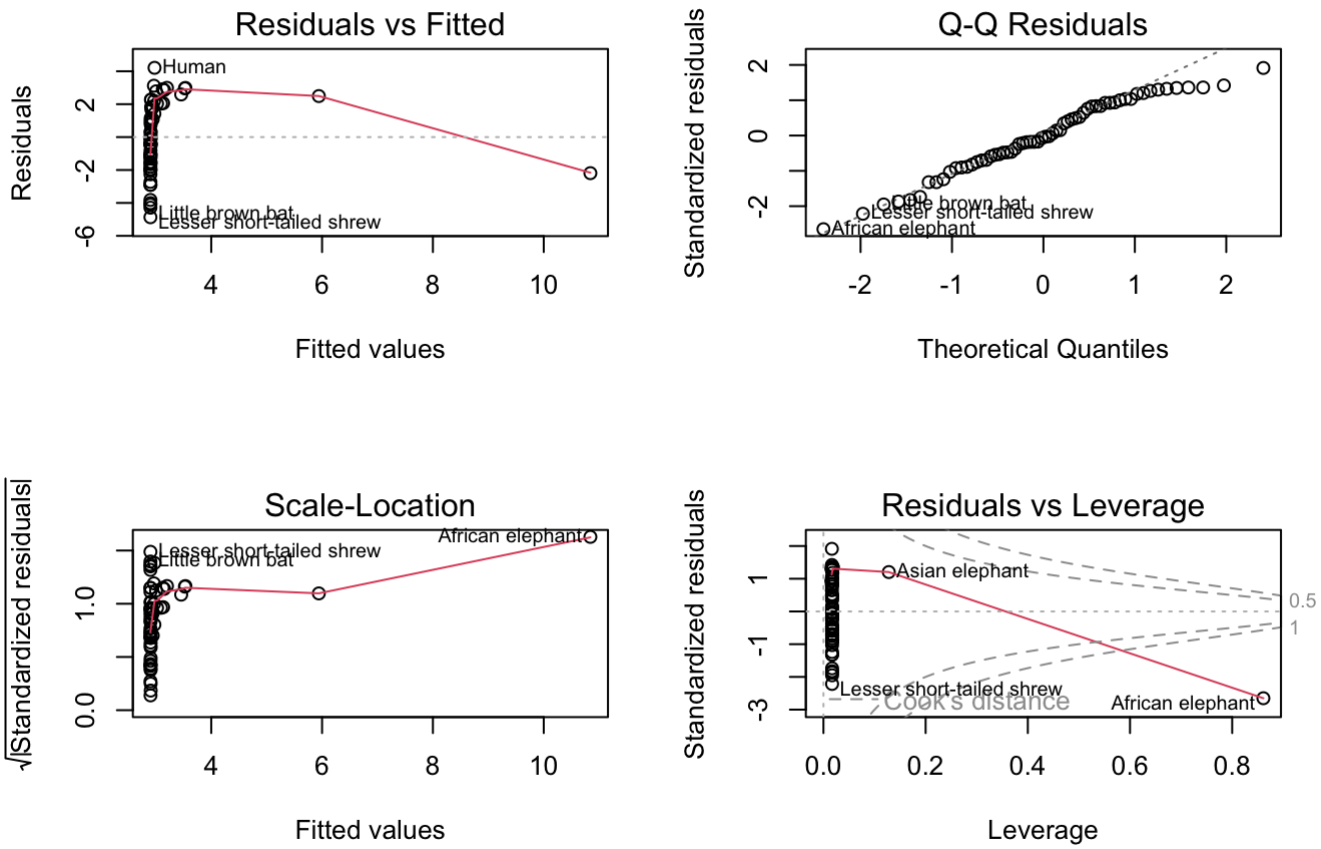
```
## `geom_smooth()` using formula = 'y ~ x'
```



The relationship between y*. and x does not appear to be linear. This plot looks very logarithmically distributed, so assumption 1 certainly appears to be violated. Assumption 2 is also not perfect, but I think transforming the x variable will help with that.

f. **Fit a simple linear regression to y_* against x, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?**

```
mammals2.ystar<-lm(logbrain~body, data=mammals2)
par(mfrow = c(2, 2))
plot(mammals2.ystar)
```



The variance assumption seems to be much improved, but the mean of errors does not appear to be 0 (they deviate from the red line in the residual vs. fitted plot). For this reason, assumption 1 seems to be violated.

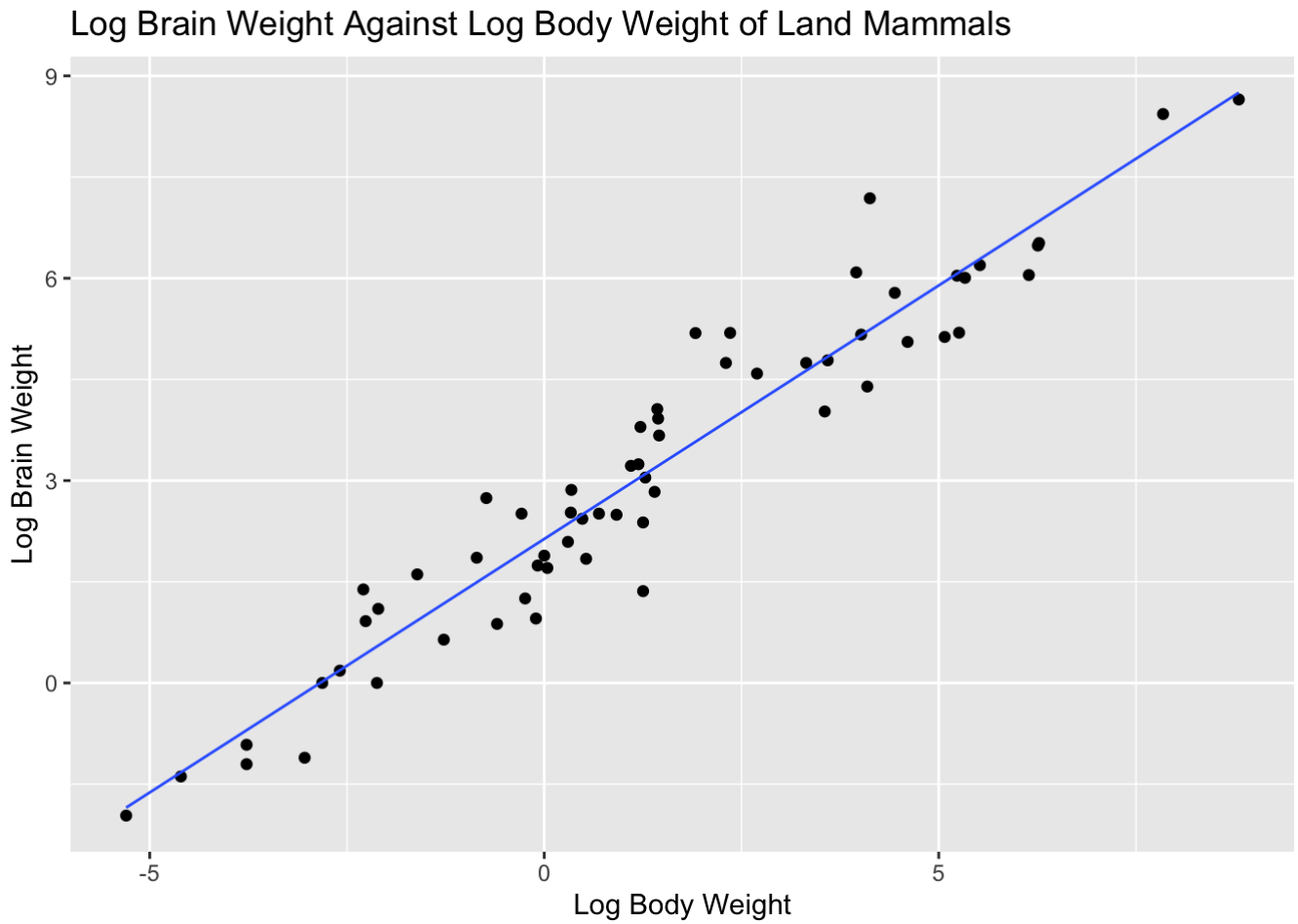
g. **Do we need to transform the x variable? If yes, what transformation(s) would you try? Briefly explain. Create a scatterplot of y_* against x_*. Do any assumptions for simple linear regression appear to be violated? If so, which ones?**

YES. Because assumption 1 is violated, we will need to transform the x variable. Looking at the scatterplot, it looks like a log() transformation would be best. I perform that transformation below:

```
logbody<-log(mammals2$body)
mammals2<-data.frame(mammals2,logbody)

ggplot(mammals2, aes(x=logbody,y=logbrain))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE, linewidth=0.5)+
  labs(x="Log Body Weight", y="Log Brain Weight", title="Log Brain Weight Against Log Bo
dy Weight of Land Mammals")
```

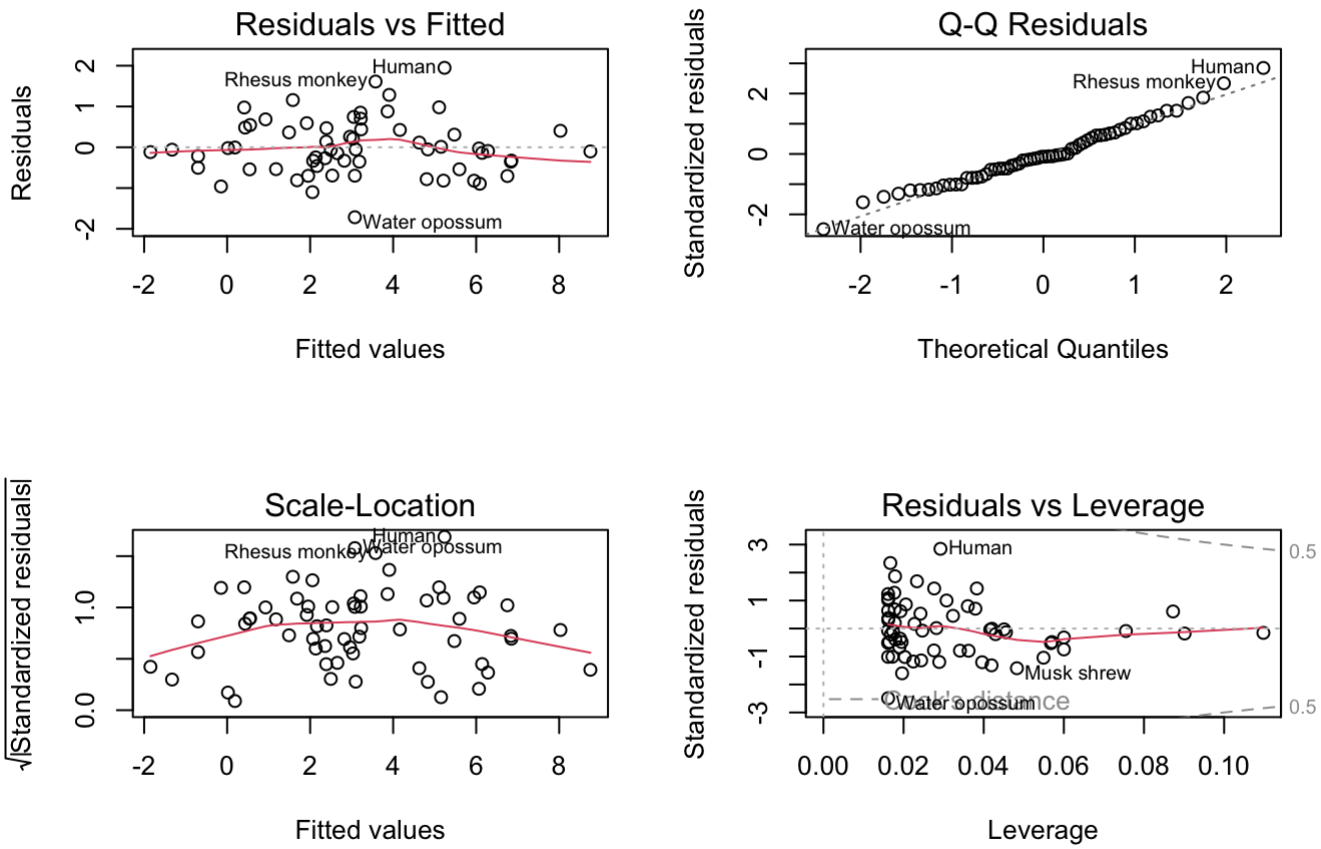
```
## `geom_smooth()` using formula = 'y ~ x'
```



Now all of the assumptions appear to be reasonably met! The points seem very positively linearly associated with relatively constant variance.

h. **Fit a simple linear regression to y_* against x_*, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones? If the assumptions are not met, repeat with a different transformation on the predictor until you are satisfied.**

```
mammals2.stars<-lm(logbrain~logbody, data=mammals2)
par(mfrow = c(2, 2))
plot(mammals2.stars)
```



Now all the assumptions seem to be satisfied! No further transformations are needed. It is interesting that the influential data points from the original model are no longer influential.

i. **Write out the regression equation, and if possible, interpret the slope of the regression.**

```
summary(mammals2.stars)
```

```
##
## Call:
## lm(formula = logbrain ~ logbody, data = mammals2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71550 -0.49228 -0.06162  0.43597  1.94829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.13479    0.09604   22.23   <2e-16 ***
## logbody      0.75169    0.02846   26.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6943 on 60 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

The regression equation is $\log(y) = 2.13479 + \log(x) * 0.75169$ where y = land mammal brain weight and x = land mammal body weight.

β_1 = 0.75169. Since both variables were log transformed, this slope shows that for a 1% increase in body weight, the weight of the brain increases by approximately 0.75169%.

---

2. For this question, we will use the cornnit data set from the faraway package. Be sure to install and load the faraway package first, and then load the data set. The data explore the relationship between corn yield (bushels per acre) and nitrogen (pounds per acre) fertilizer application in a study carried out in Wisconsin.

```
library(faraway)
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:survival':
##
##     rats, solder
```

```
## The following object is masked from 'package:GGally':
##
##     happy
```

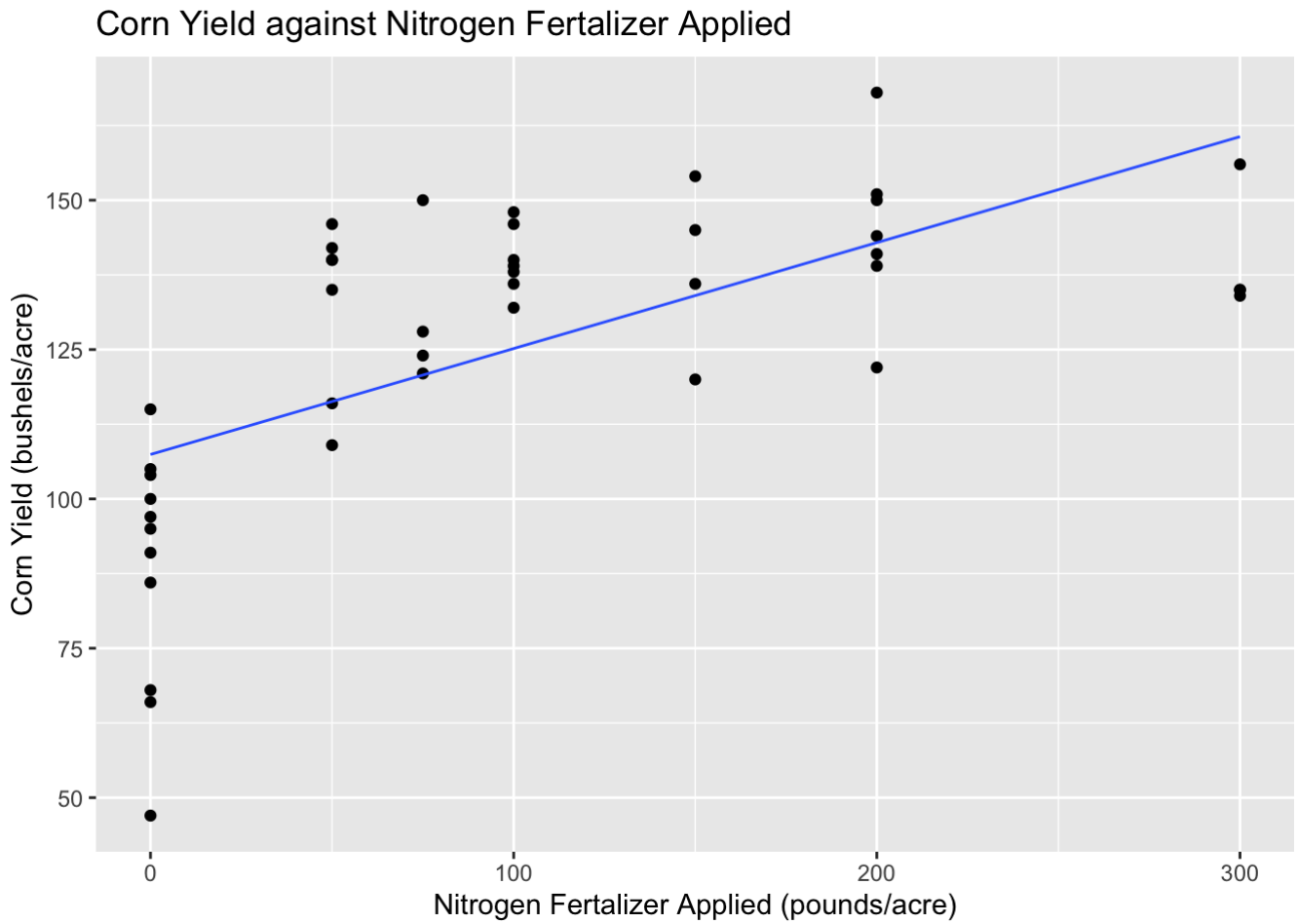
```
head(cornnit)
```

```
##   yield nitrogen
## 1   115        0
## 2   128       75
## 3   136      150
## 4   135      300
## 5    97        0
## 6   150       75
```

a. **What is the response variable and predictor for this study? Create a scatterplot of the data, and interpret the scatterplot.**

```
ggplot(cornnit, aes(x=nitrogen,y=yield))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE, linewidth=0.5)+
  labs(x="Nitrogen Fertalizer Applied (pounds/acre)", y="Corn Yield (bushels/acre)", tit
le="Corn Yield against Nitrogen Fertalizer Applied")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



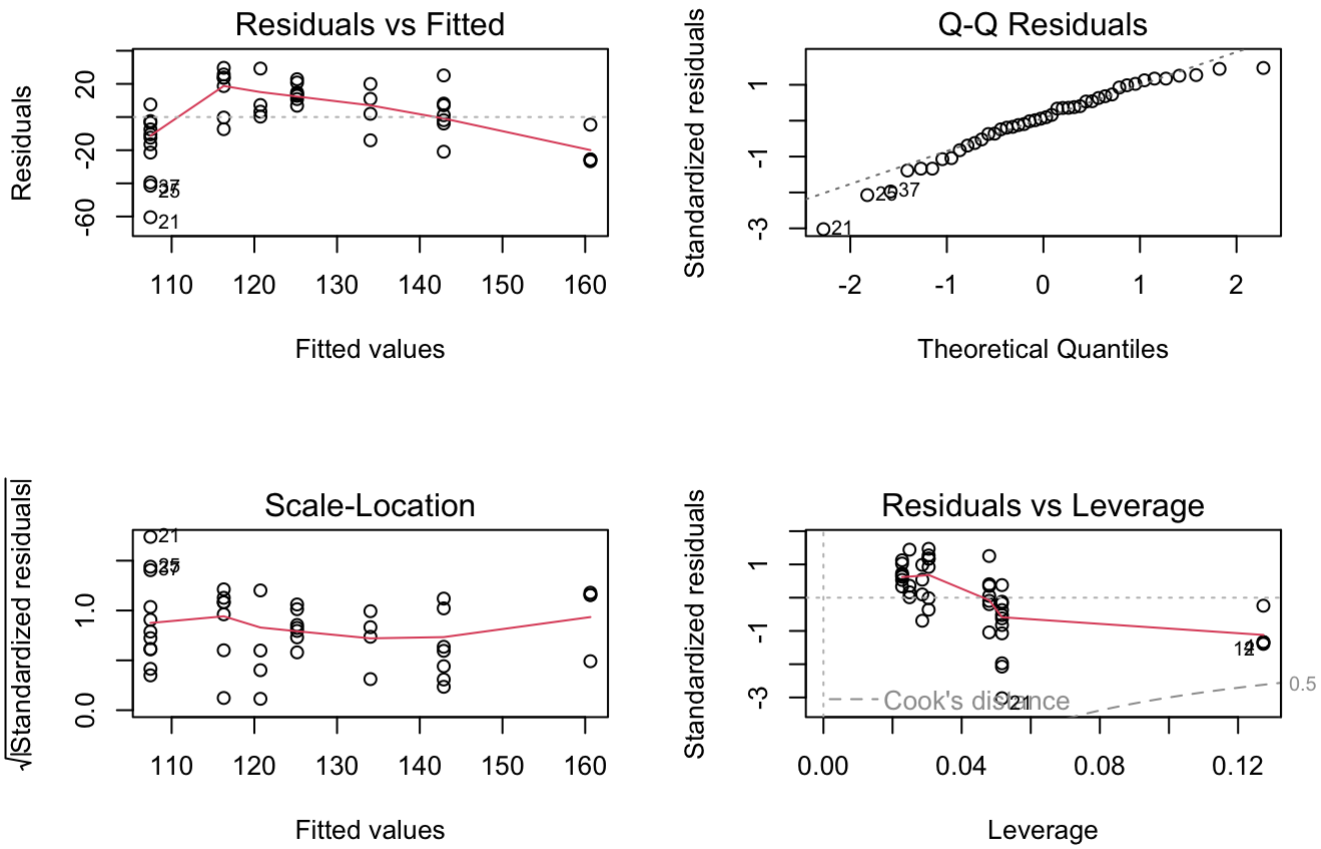
Generally speaking, there seems to be an increasing association between the number of bushels per acre of corn yielded and the pounds per acre of nitrogen fertilizer applied. The relationship appears to be more logarithmic than linear.

To assess assumption 1, the data points should be evenly scattered on both sides of the regression line and follow the linear shape of the line as we move from left to right. We do not see this in the scatterplot, so assumption 1 is not met. When nitrogen fertilizer applied is below 50 and above 250, the data point(s) are below the line. When nitrogen fertilizer applied is between 50 and 200, the majority of the data point(s) are above the line. Assumption 1 is not met.

To assess assumption 2, the vertical spread of the data points should be constant as we move from left to right. The spread seems to be inconsistent, so assumption 2 is not met either.

b. **Fit a linear regression without any transformations. Create the corresponding residual plot. Based only on the residual plot, what transformation will you consider first? Be sure to explain your reason.**

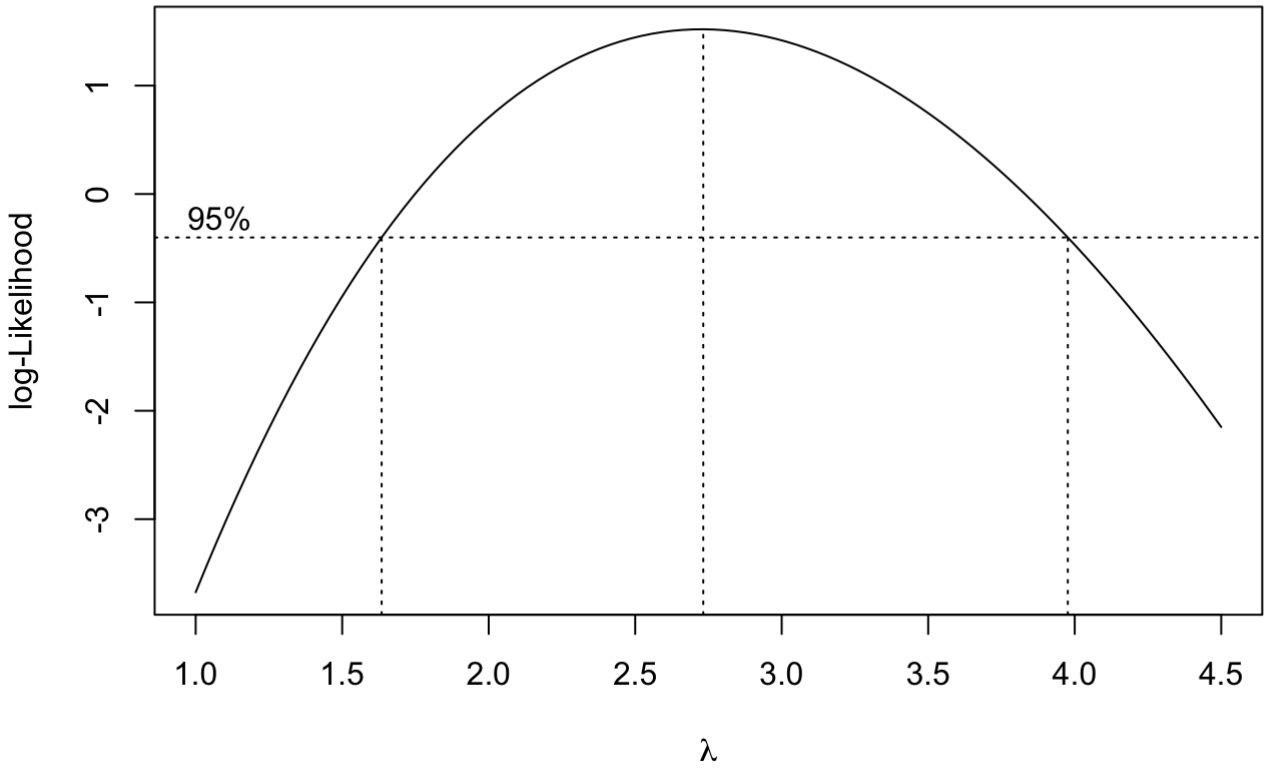
```
cornlm<-lm(yield~nitrogen, data=cornnit)
par(mfrow = c(2, 2))
plot(cornlm)
```



The residuals vs. fitted and scale-location plots seem to show variation with a bit of fanning toward the right-hand side, so assumption 2 is not met and that assumption should be fixed before we adjust to fix assumption 1. In other words, we should transform the y-variable to create constant variation before adjusting the x-variable to ensure the errors have mean 0.

c. **Create a Box Cox plot for the profile loglikelihoods. How does this plot aid in your data transformation?**

```
boxcox(cornlm, lambda = seq(1, 4.5, 1/10))
```

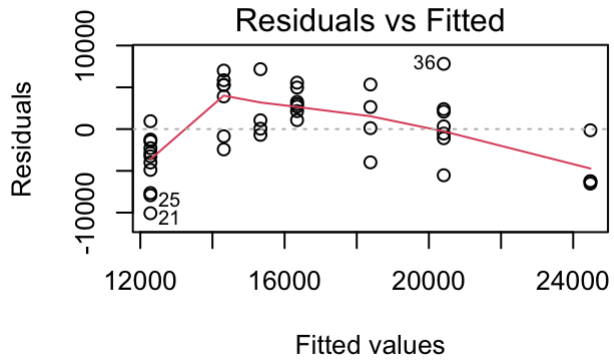


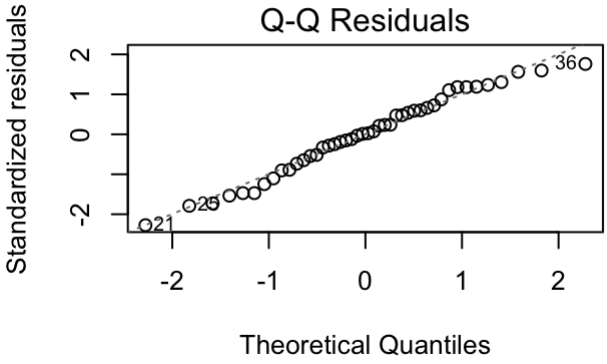
This boxcox plot shows that a transformation is certainly needed for the y-variable (corn yield) because 1 is not contained within the upper and lower bounds of the CI for λ. we can choose $\lambda = 2$ to square transform the response variable to get $y_* = y^2$.

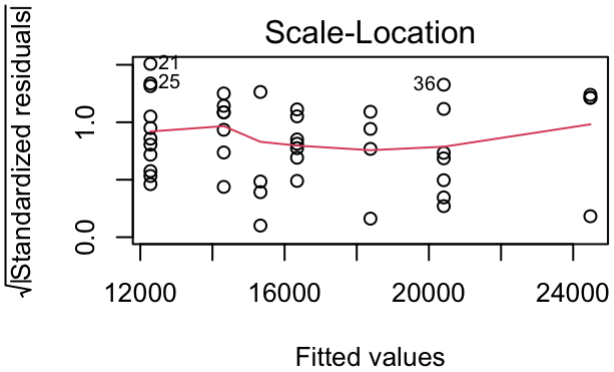
d. **Perform the necessary transformation to the data. Re fit the regression with the transformed variable(s) and assess the regression assumptions. You may have to apply transformations a number of times. Be sure to explain the reason behind each of your transformations. Perform the needed transformations until the regression assumptions are met. What is the regression equation that you will use?**

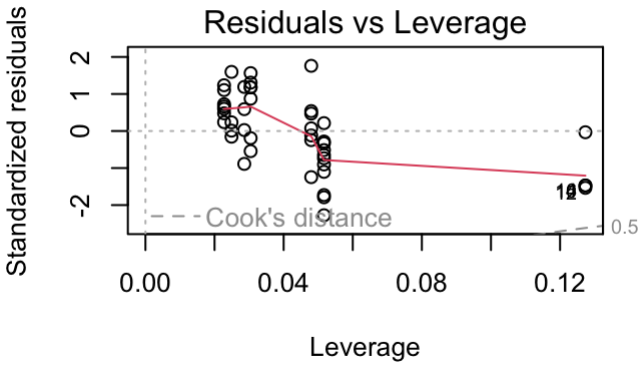
Note: in part 2d, there are a number of solutions that will work. You must clearly document your reasons for each of your transformations.

```
cornlm2<-lm((yield)^2~nitrogen, data=cornnit)
par(mfrow = c(2, 2))
plot(cornlm2)
```



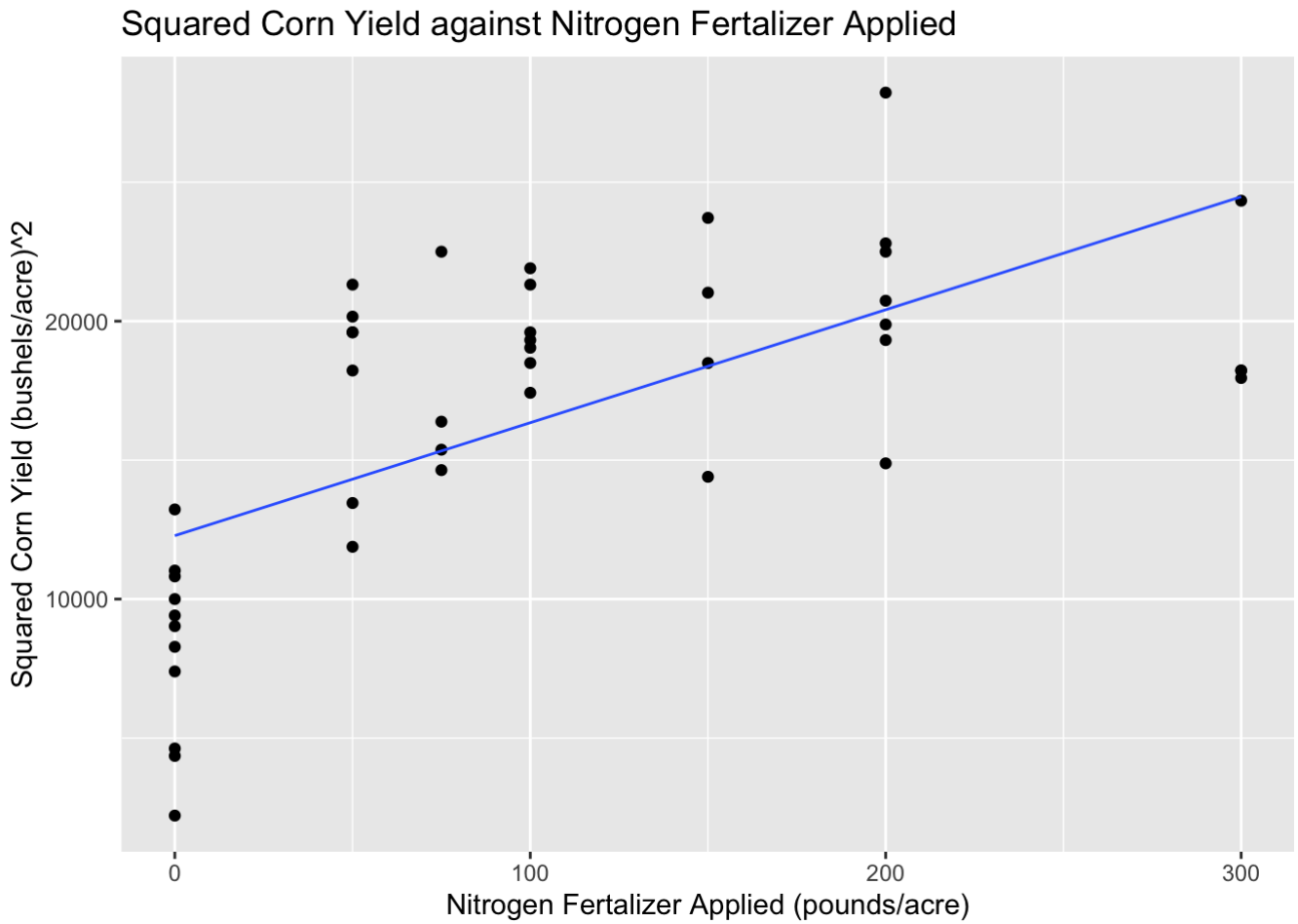






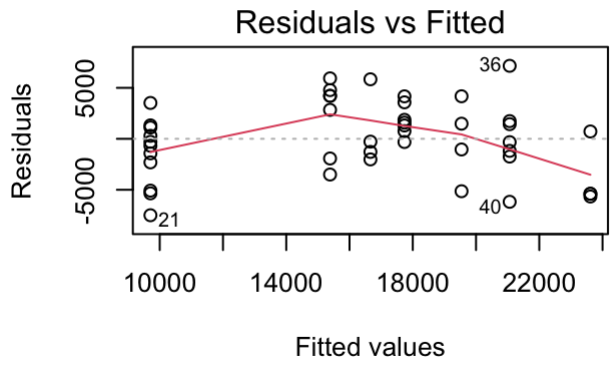

```
ggplot(cornnit, aes(x=nitrogen,y=(yield^2)))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE, linewidth=0.5)+
  labs(x="Nitrogen Fertalizer Applied (pounds/acre)", y="Squared Corn Yield (bushels/acr
e)^2", title="Squared Corn Yield against Nitrogen Fertalizer Applied")
```

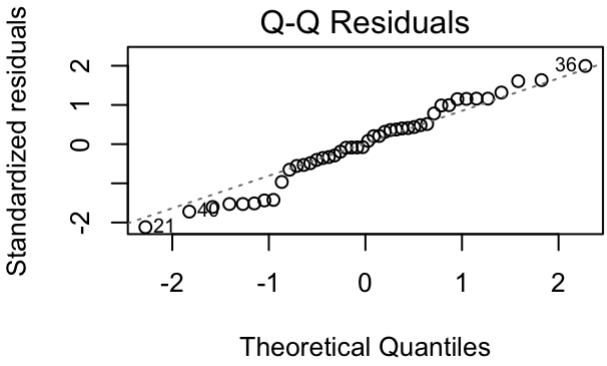
```
## `geom_smooth()` using formula = 'y ~ x'
```

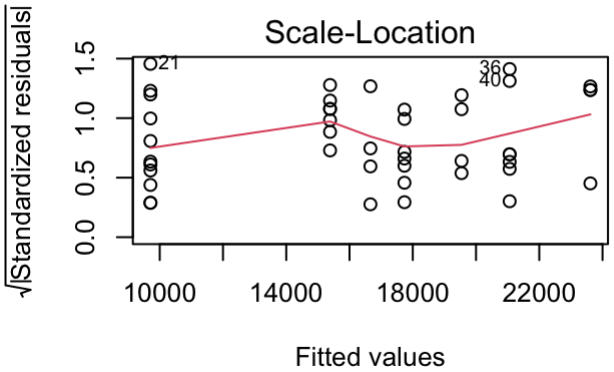


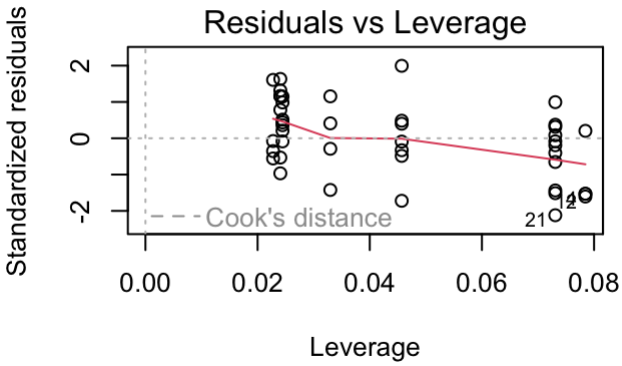
I first transformed the y variable (corn yield) by squaring it in order to create more constant variance as shown necessary by the boxcox plot. This seemed to cause improvement, but looking at the scatterplot it is clear that assumption 1 is still not met. Since the points seem to follow a log(x) or $\sqrt{x}$ shape, I decided to log transform the x value (nitrogen fertilizer applied). Because there are many x variables with a value of 0, and you cannot take the log of 0, I decided to take the square root of x (nitrogen fertilizer applied) instead.

```
cornlm3<-lm((yield)^2~sqrt(nitrogen), data=cornnit)
par(mfrow = c(2, 2))
plot(cornlm3)
```



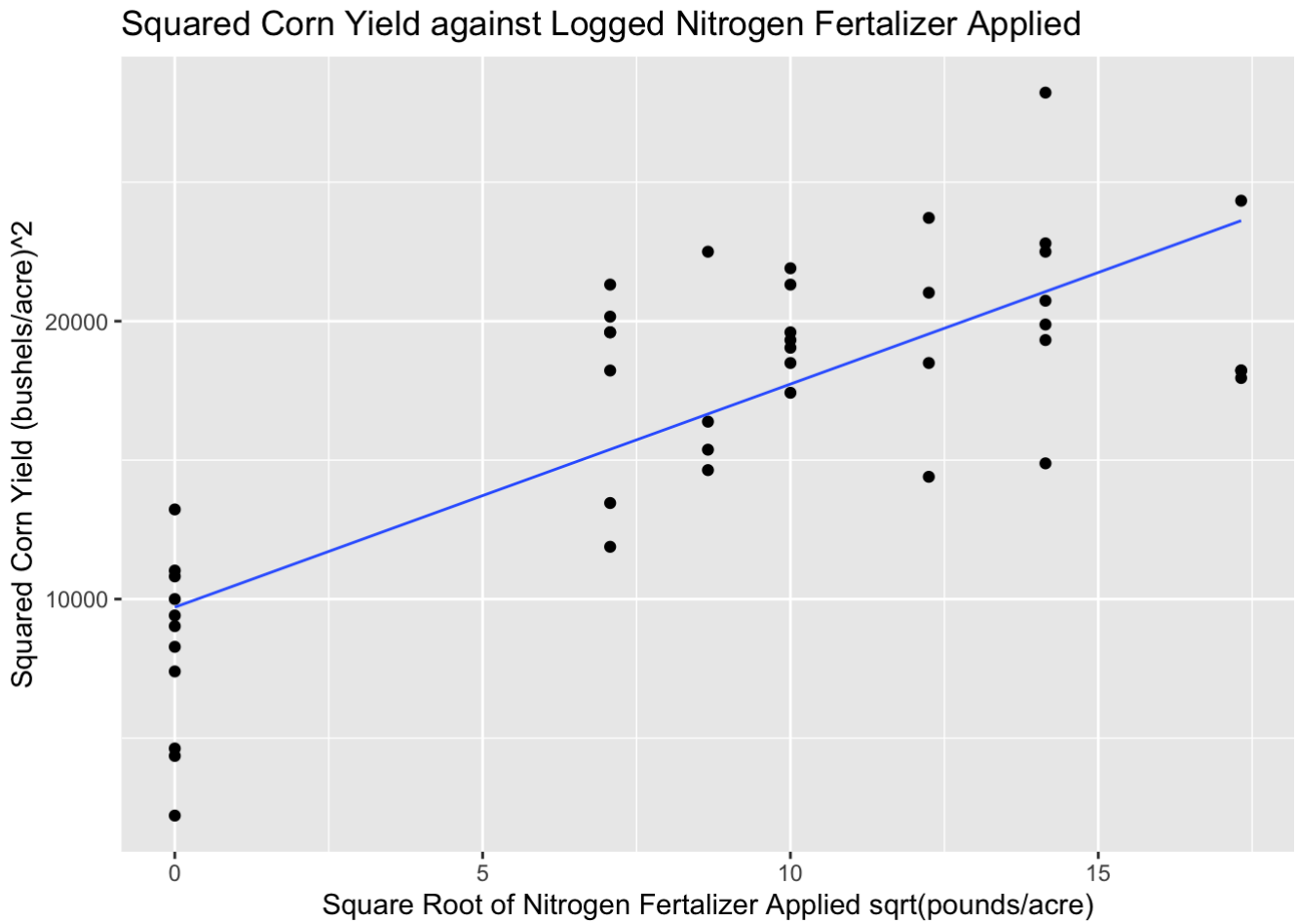







```
ggplot(cornnit, aes(x=sqrt(nitrogen),y=(yield^2)))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE, linewidth=0.5)+
  labs(x="Square Root of Nitrogen Fertalizer Applied sqrt(pounds/acre)", y="Squared Corn
Yield (bushels/acre)^2", title="Squared Corn Yield against Logged Nitrogen Fertalizer Ap
plied")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Although this scatterplot looks a bit odd with the large gap from 0 to 6 in sqrt(nitrogen fertilizer applied), this regression now fits all the assumptions relatively well.

```
summary(cornlm3)
```

```
##
## Call:
## lm(formula = (yield)^2 ~ sqrt(nitrogen), data = cornnit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7497.3 -1951.6     8.3  2107.3  7160.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9706.27     993.28   9.772 2.22e-12 ***
## sqrt(nitrogen)   803.07      97.68   8.222 2.75e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3674 on 42 degrees of freedom
## Multiple R-squared:  0.6168, Adjusted R-squared:  0.6076
## F-statistic:  67.6 on 1 and 42 DF,  p-value: 2.75e-10
```

The regression equation is $y^2 = 9706.27 + 803.07 * \sqrt{x}$ with y = corn yielded (bushel/acre) and x = nitrogen fertilizer applied (pounds/acre).

---

3. For this question, we will use the data set “nfl.txt”, which contains data on NFL team performance from the 1976 season. The variables are:

- `y` : Games won (14-game season)
- `x1` : Rushing yards (season)
- `x2` : Passing yards (season)
- `x3` : Punting average (yards/punt)
- `x4` : Field goal percentage (FGs made/FGs attempted)
- `x5` : Turnover differential (turnovers acquired - turnovers lost)
- `x6` : Penalty yards (season)
- `x7` : Percent rushing (rushing plays/total plays)
- `x8` : Opponents’ rushing yards (season)
- `x9` : Opponents’ passing yards (season)

```
nfl <- read.table("nfl.txt", header=TRUE)
head(nfl)
```

```
##    y   x1   x2   x3   x4 x5  x6   x7   x8   x9
## 1 10 2113 1985 38.9 64.7  4 868 59.7 2205 1917
## 2 11 2003 2855 38.8 61.3  3 615 55.0 2096 1575
## 3 11 2957 1737 40.1 60.0 14 914 65.6 1847 2175
## 4 13 2285 2905 41.6 45.3 -4 957 61.4 1903 2476
## 5 10 2971 1666 39.2 53.8 15 836 66.1 1457 1866
## 6 11 2309 2927 39.7 74.1  8 786 61.0 1848 2339
```

```
#GGally::ggpairs(nfl)
```

a. **Fit a multiple regression model for the number of games won against the following three predictors: the team’s passing yardage, the percentage of rushing plays, and the opponents’ yards rushing. Write the estimated regression equation.**

```
nfl.1 <- lm(y~x2+x7+x8, data=nfl)
summary(nfl.1)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = nfl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x2           0.003598   0.000695   5.177 2.66e-05 ***
## x7           0.193960   0.088233   2.198 0.037815 *
## x8          -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```

Our estimated regression equation is $Y = -1.808372 + 0.003598(X2) + 0.193960(X7) + -0.004816(X8)$ where Y= Games won, X2= Passing yards, X7= Percent rushing, and X8= Opponents' rushing yards.

b. **Interpret the estimated coefficient for the predictor x7 in context.**

```
summary(nfl.1)$coefficients[3,1]
```

```
## [1] 0.1939602
```

The number of games won in a 14-game season (y) increases by 0.1939602 for a one unit increase in the percentage of rushing plays(x7) while holding the team's passing yardage(x2) and the opponents' yards rushing(x8) constant.

c. **A team with x2 = 2000 yards, x7 = 48 percent, and x8 = 2350 yards would like to estimate the number of games it would win. Also provide a relevant interval for this estimate with 95% confidence.**

```
newdata<-data.frame(x2=2000, x7=48, x8=2350)

predict(nfl.1, newdata, level=0.95, interval="prediction")
```

```
##        fit        lwr      upr
## 1 3.381448 -0.5163727 7.279268
```

The estimate for the number of games won in a 14-game season for a team with a passing yardage of 2000 yards, 48% of rushing plays, and the opponents' yards rushing equal to 2350 yards is 3.381448 games won.

We have 95% confidence that the number of games won in a 14-game season for a team with 2000 rushing yards, 48% rushing plays, and 2350 rushing yards by the opponent is between -0.5163727 games and 7.279268 games.

d. **Using the output for the multiple linear regression model from part 3a, answer the following question from a client: "Is this regression model useful in predicting the number of wins during the 1976 season?" Be sure to write the null and alternative hypotheses, state the value of the test statistic, state the p-value, and state a relevant conclusion.**

```
summary(nfl.1)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = nfl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x2           0.003598   0.000695   5.177 2.66e-05 ***
## x7           0.193960   0.088233   2.198 0.037815 *
## x8          -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```

$H_0 : \beta_{X2} = \beta_{X7} = \beta_{X8} = 0$ (We could drop one or more of the variables from our model in the presence of the other variables.)
H_A : At least one of the β coefficients is not 0 (and therefore should not be dropped from the model)

F-statistic = 29.44 and p-value= $3.273 * 10^{-8}$

Because this F-statistic is significantly large and the p-value of $3.273 * 10^{-8}$ is smaller than $\alpha = 0.05$, we H_0 in favor of H_A. In other words, we have enough evidence to conclude that we cannot drop any of the other predictors to simplify the model. The data supports the claim that the model with these three predictors is useful for predicting the number of wins during the 1976 season.

e. **Report the value of the t statistic for the predictor x7. What is the relevant conclusion from this t statistic?**

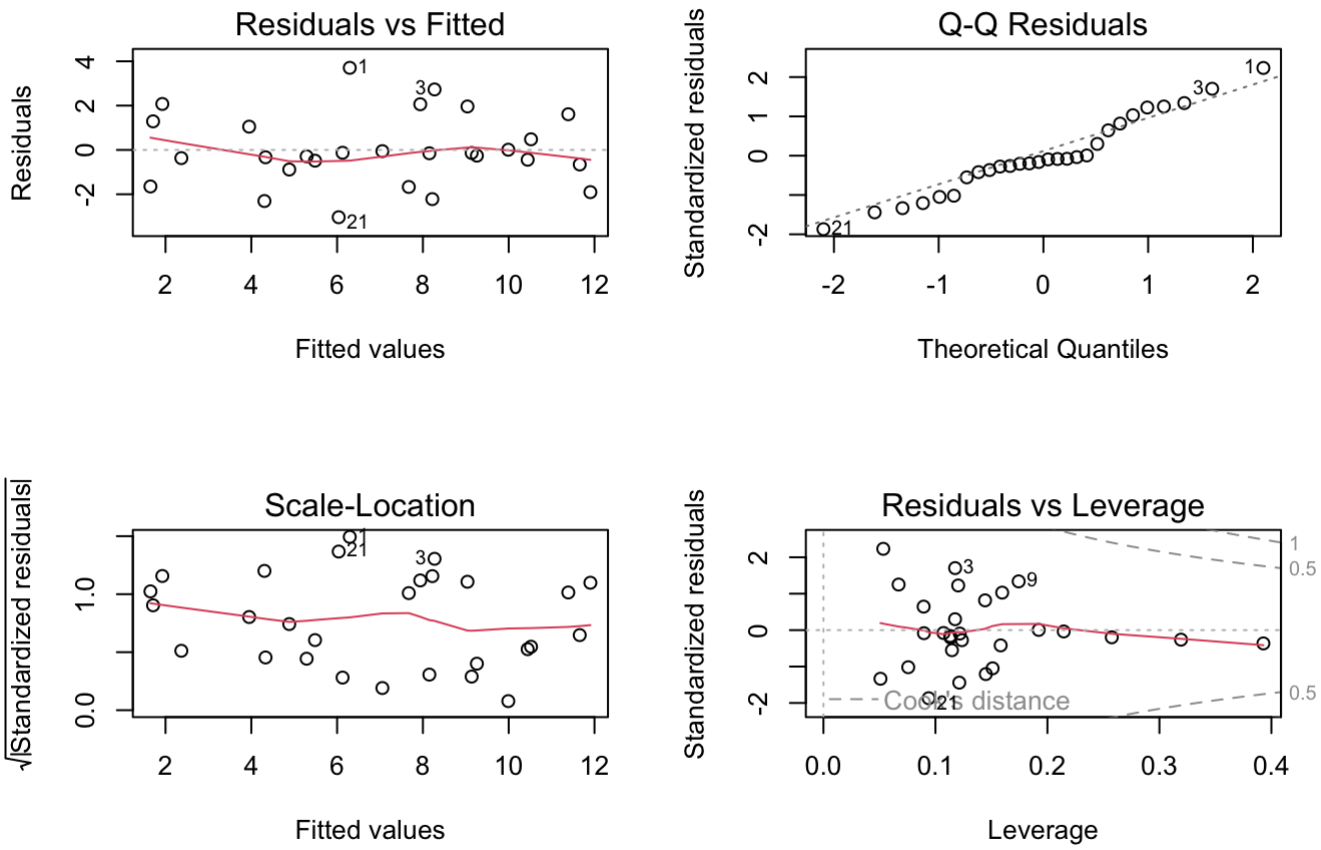
```
summary(nfl.1)$coefficients[3,3]
```

```
## [1] 2.198262
```

The t-statistic for the percentage of rushing plays(x7) is 2.198362. Because it is relatively large and the p-value for the percentage of rushing plays(x7) is smaller than $\alpha = 0.05$, we have enough evidence to conclude that this variable is significant for predicting the number of games won in a 14-game season(y) when in the presence of the other predictor variables.

f. **Check the regression assumptions by creating the diagnostic plots. Comment on these plots**

```
par(mfrow = c(2, 2))
plot(nfl.1)
```



All of the assumptions appear to be met. This is clear because the residuals are scattered randomly around 0 in the residual plot with constant variance.

g. **Consider adding another predictor, x1, the team's rushing yards for the season, to the model. Interpret the results of the t test for the coefficient of this predictor. A classmate says: "Since the result of the t test is insignificant, the team's rushing yards for the season is not linearly related to the number of wins." Do you agree with your classmate's statement?**

```
nfl.2 <- lm(y~x1+x2+x7+x8, data=nfl)
summary(nfl.2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x7 + x8, data = nfl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7456 -0.6801 -0.1941  1.1033  3.7580
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.8791718  8.1955007  -0.107  0.91550
## x1           0.0009045  0.0016489   0.549  0.58862
## x2           0.0035214  0.0007191   4.897 6.02e-05 ***
## x7           0.1437590  0.1280424   1.123  0.27313
## x8          -0.0046994  0.0013131  -3.579  0.00159 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.732 on 23 degrees of freedom
## Multiple R-squared:  0.7891, Adjusted R-squared:  0.7524
## F-statistic: 21.51 on 4 and 23 DF,  p-value: 1.702e-07
```

The t-statistic for x1, the team's rushing yards for the season, is 0.549 with a p-value of 0.58862. Because the result of the t-test is not significant, we can drop the predictor x1, the team's rushing yards for the season, while keeping the other predictors in the model.

I **disagree** with the classmate who said "Since the result of the t test is insignificant, the team's rushing yards for the season is not linearly related to the number of wins." An insignificant t test informs us we can drop that particular predictor, while leaving the other predictors in the model. It does not tell us anything about the linear relationship between the predictor and the response because there are other variables in the model whose interactions may be affecting its relationship with the response variable.

---

4. For this question, the data are from the faraway package in R. After installing the faraway package, load the seatpos dataset. Car drivers like to adjust the seat position for their own comfort. Car designers find it helpful to know where different drivers will position the seat. Researchers at the HuMoSim laboratory at the University of Michigan collected data on 38 drivers. The response variable is hipcenter, the horizontal distance of the midpoint of the hips from a fixed location in the car in mm. They measured the following eight predictors:

- `x1` : Age. Age in years
- `x2` : Weight. Weight in pounds
- `x3` : HtShoes. Height with shoes in cm
- `x4` : Ht. Height without shoes in cm
- `x5` : Seated. Seated height in cm
- `x6` : Arm. Arm length in cm
- `x7` : Thigh. Thigh length in cm
- `x8` : Leg. Lower leg length in cm

```
library(faraway)
head(seatpos)
```

```
##   Age Weight HtShoes    Ht Seated  Arm Thigh  Leg hipcenter
## 1  46    180   187.2 184.9   95.2 36.1  45.3 41.3  -206.300
## 2  31    175   167.5 165.5   83.8 32.9  36.5 35.9  -178.210
## 3  23    100   153.6 152.2   82.9 26.0  36.6 31.0   -71.673
## 4  19    185   190.3 187.4   97.3 37.4  44.1 41.0  -257.720
## 5  23    159   178.0 174.1   93.9 29.5  40.1 36.9  -173.230
## 6  47    170   178.7 177.0   92.4 36.0  43.2 37.4  -185.150
```

a. **Fit the full model with all the predictors. Using the summary() function, comment on the results of the t tests and ANOVA F test from the output.**

```
seatpos.full <- lm(hipcenter~., data=seatpos)
summary(seatpos.full)
```

```
##
## Call:
## lm(formula = hipcenter ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

The ANOVA F-statistic is 7.94 with a small p-value. So we reject the null hypothesis and state that our MLR model with all of these predictors is useful. However, notice the t tests are insignificant for the all of the predictor coefficients. Individually, each t test is informing us that we can drop that specific predictor, while leaving the other predictors in the model.

b. **Briefly explain why, based on your output from part 4a, you suspect the model shows signs of multicollinearity.**

Because almost all the t tests are insignificant, but the ANOVA F is highly significant, there is evidence of multicollinearity. Additionally, the standard errors for some are the estimated coefficients are very large which further suggests the presence of multicollinearity.

c. **Provide the output for all the pairwise correlations among the predictors. Comment briefly on the pairwise correlations.**

```
round(cor(seatpos[,-9]),3)
```

```
##             Age Weight HtShoes     Ht Seated   Arm Thigh    Leg
## Age       1.000  0.081  -0.079 -0.090 -0.170 0.360 0.091 -0.042
## Weight    0.081  1.000   0.828  0.829  0.776 0.698 0.573  0.784
## HtShoes  -0.079  0.828   1.000  0.998  0.930 0.752 0.725  0.908
## Ht       -0.090  0.829   0.998  1.000  0.928 0.752 0.735  0.910
## Seated   -0.170  0.776   0.930  0.928  1.000 0.625 0.607  0.812
## Arm       0.360  0.698   0.752  0.752  0.625 1.000 0.671  0.754
## Thigh     0.091  0.573   0.725  0.735  0.607 0.671 1.000  0.650
## Leg      -0.042  0.784   0.908  0.910  0.812 0.754 0.650  1.000
```

There are many variables that are highly correlated with one another! Nearly every correlation value shows moderate to high correlation. This is proof that there is certainly multicollinearity.

d. **Check the variance inflation factors (VIFs). What do these values indicate about multicollinearity?**

```
faraway::vif(seatpos.full)
```

```
##        Age     Weight    HtShoes         Ht     Seated        Arm      Thigh
##   1.997931   3.647030 307.429378 333.137832   8.951054   4.496368   2.762886
##        Leg
##   6.694291
```

Generally, VIFs greater than 5 indicate some degree of multicollinearity, and VIFs greater than 10 indicate a high level of multicollinearity. In the output aboce, we see VIF values as high as 333.137832 and 307.429378! These values, corresponding with Ht and HtShoes respectively, indicate very high levels of multicollinearity.

e. **Looking at the data, we may want to look at the correlations for the variables that describe length of body parts: HtShoes, Ht, Seated, Arm, Thigh, and Leg. Comment on the correlations of these six predictors.**

```
round(cor(seatpos[,-c(1,2,9)]),3)
```

```
##         HtShoes    Ht Seated   Arm Thigh   Leg
## HtShoes   1.000 0.998  0.930 0.752 0.725 0.908
## Ht        0.998 1.000  0.928 0.752 0.735 0.910
## Seated    0.930 0.928  1.000 0.625 0.607 0.812
## Arm       0.752 0.752  0.625 1.000 0.671 0.754
## Thigh     0.725 0.735  0.607 0.671 1.000 0.650
## Leg       0.908 0.910  0.812 0.754 0.650 1.000
```

All of the predictors that describe length of body parts (`HtShoes` , `Ht` , `Seated` , `Arm` , `Thigh` , and `Leg`) are moderately-highly positively linearly correlated. The largest correlation is between `Ht` and `HtShoes` .

f. **Since all the six predictors from the previous part are highly correlated, you may decide to just use one of the predictors and remove the other five from the model. Decide which predictor out of the six you want to keep, and briefly explain your choice.**

Because the largest correlation is between `Ht` and `HtShoes` , I want to use one of these two predictors. Looking at how each of these are correlated with the other predictors, I see that `Ht` has slightly higher correlation values with each of the other variables than `HtShoes` does, so I plan to remove `HtShoes` , `Seated` , `Arm` , `Thigh` , and `Leg` and keep only `Ht` . Additionally, height is a fairly reliable measurement that can be more consistantly measured than many of the other variables.

g. **Based on your choice in part 4f, fit a multiple regression with your choice of predictor to keep, along with the predictors x1 = Age and x2 =Weight. Check the VIFs for this model. Comment on whether we still have an issue with multicollinearity.**

```
seatpos.red <- lm(hipcenter~Age+Weight+Ht, data=seatpos)
faraway::vif(seatpos.red)
```

```
##      Age   Weight       Ht
## 1.093018 3.457681 3.463303
```

Because all the VIF values are less than 5 now, the multicollinearity issue has been resolved and there is no longer any strong multicollinearity.

h. **Conduct a partial F test to investigate if the predictors you dropped from the full model were jointly insignificant. Be sure to state a relevant conclusion.**

- Model 1: using `Age` , `Weight` , `HtShoes` , `Ht` , `Seated` , `Arm` , `Thigh` , and `Leg` as the predictors for `hipcenter`
- Model 2: using `Age` , `Weight` , and `Ht` as the predictors for `hipcenter`

Carry out the appropriate hypothesis test to decide which of models 1 or 2 should be used. Be sure to show all steps in your hypothesis test.

$H_0 = \beta_{HtShoes} = \beta_{Seated} = \beta_{Arm} = \beta_{Theigh} = \beta_{Leg} = 0$ (The parameters in the variables we wish to drop are 0, so the reduced model (model 2) is supported.)
H_0 = At least one coefficient in H_0 is not 0 (The full model (model 1) is supported.)

```
anova(seatpos.red, seatpos.full)
```

```
## Analysis of Variance Table
##
## Model 1: hipcenter ~ Age + Weight + Ht
## Model 2: hipcenter ~ Age + Weight + HtShoes + Ht + Seated + Arm + Thigh +
##     Leg
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     34 45262
## 2     29 41262  5    4000.3 0.5623 0.7279
```

```
qf(0.95, 5, 29)
```

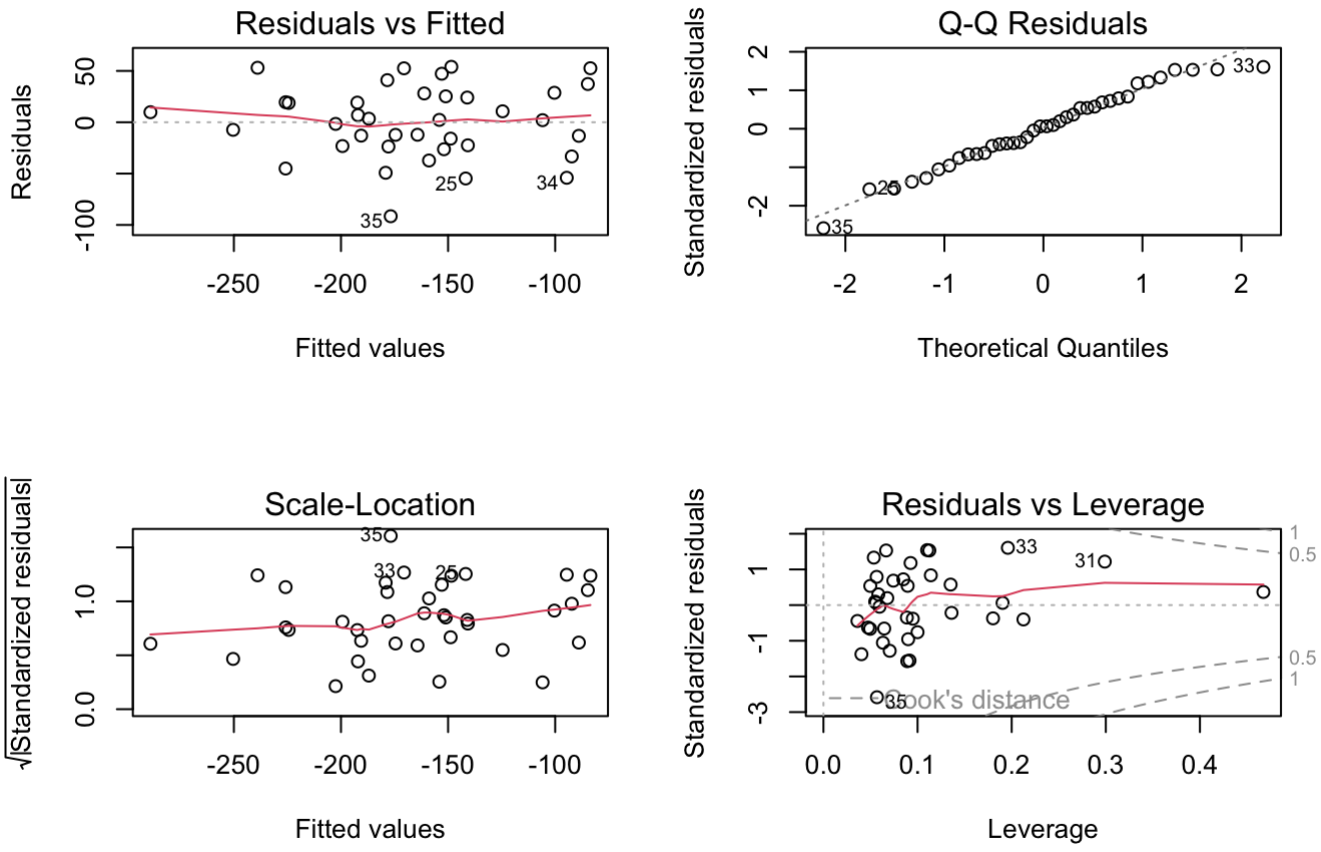
```
## [1] 2.545386
```

F-statistic = 0.5623
Critical value = 2.545386
P-Value = 0.7279

Because the F-statistic is smaller than the critical value, and the p-value is greater than $\alpha = 0.05$, we fail to reject H_0. In other words we do not have enough evidence to conclude that `HtShoes`, `Seated`, `Arm`, `Thigh`, and `Leg` are necessary to keep in the model in the presence of the other variables. We can drop `HtShoes`, `Seated`, `Arm`, `Thigh`, and `Leg` as they not necessary in this model, so a reduced model (Model 2) is prefered.

i. **Produce a residual plot for your model from part 4g. Based on the residual plot, comment on the assumptions for the multiple regression model.**

```
par(mfrow = c(2, 2))
plot(seatpos.red)
```



All of the assumptions appear to be met. This is clear because the residuals are scattered randomly around 0 in the residual plot with constant variance.

j. **Based on your results, write your estimated regression equation from part 4g. Also report the R2 of this model, and compare with the R2 you reported in part 4a, for the model with all predictors. Also comment on the adjusted R2 for both models.**

```
summary(seatpos.red)
```

```
##
## Call:
## lm(formula = hipcenter ~ Age + Weight + Ht, data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.526 -23.005   2.164  24.950  53.982
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 528.297729 135.312947   3.904 0.000426 ***
## Age           0.519504   0.408039   1.273 0.211593
## Weight        0.004271   0.311720   0.014 0.989149
## Ht           -4.211905   0.999056  -4.216 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.49 on 34 degrees of freedom
## Multiple R-squared:  0.6562, Adjusted R-squared:  0.6258
## F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08
```

```
summary(seatpos.red)$r.squared
```

```
## [1] 0.6561654
```

Our estimated regression equation is

$$\text{hipcenter} = 528.297729 + 0.519504(\text{Age}) + 0.004271(\text{Weight}) + -4.211905(\text{Ht})$$

The R^2 of this model is 0.6561654. The coefficient of determination informs us that about 65.62% of variation in the horizontal distance of the midpoint of the hips from a fixed location in the car in mm (`hipcenter`) can be explained by the driver's age, weight, and height. We notice that the R^2 value for this reduce model is lower than the R^2 for the full model which was 0.6866. However, when looking at the adjusted R^2 values for the models, the full model has an adjusted R^2 of 0.6001 while the reduced model has an adjusted R^2 of 0.6258. R^2 tends to increase with additional variables in the model, but after adjusting for multicollinearity, we notice the adjusted R^2 is lower in the full model than in the reduced model.

---

5. (You may only use R as a simple calculator or to find p-values or critical values) Data from n = 113 hospitals are used to evaluate factors related to the risk that patients get an infection while in the hospital. The response variable is InfctRsk, the percentage of patients who get an infection while hospitalized. The predictors are Stay, the average length of stay, Cultures, a ratio of the number of cultures performed per number of patients with no infection (times 100), Age, the average patient age, Census, the number of patients in the hospital, and Beds, the number of beds in the hospital. We consider the following multiple regression equation: E(InfctRsk) = β0+β1Stay+β2Cultures+β3Age+β4Census+β5Beds. SomeRoutputisshownbelow. You may assume the regression assumptions are met.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2051282  1.2075929   0.170   0.8654
Stay        0.2055252  0.0660885   3.110   0.0024 **
Cultures    0.0590369  0.0103096   5.726  9.5e-08 ***
Age         0.0173637  0.0229966   -----   ------
Census      0.0010306  0.0034942   0.295   0.7686
Beds        0.0004476  0.0026781   0.167   0.8676
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9926 on 107 degrees of freedom
Multiple R-squared:  ______,    Adjusted R-squared:  ______
F-statistic: 19.48 on 5 and 107 DF,  p-value: 9.424e-14
Analysis of Variance Table
Response: InfctRsk
           Df  Sum Sq Mean Sq F value    Pr(>F)
Stay        1  57.305  57.305 58.1676 1.044e-11 ***
Cultures    1  33.397  33.397 33.8995 6.154e-08 ***
Age         1   0.136   0.136  0.1376   0.71144
Census      1   5.101   5.101  5.1781   0.02487 *
Beds        1   0.028   0.028  0.0279   0.86759
Residuals 107 105.413   0.985
```

a. **What is the value of the estimated coefficient of the variable Stay? Write a sentence that interprets this value.**

The estimated coefficient for `Stay` is 0.2055252. This shows that the percentage of patients who get an infection while hospitalized increases by 0.2055252 for a one unit increase in `Stay`, or the average length of stay, while holding all other variables (`Cultures`, `Age`, `Census`, and `Beds`) constant.

b. **Derive the test statistic, p-value, and critical value for the variable Age. What null and alternative hypotheses are being evaluated with this test statistic? What conclusion should we make about the variable Age?**

$H_0 : \beta_{\text{Age}} = 0$ (We could drop the variable Age from our model in the presence of the other variables.)
$H_A : \beta_{\text{Age}} \neq 0$ (We cannot drop the variable Age from our model in the presence of the other variables.)

```
b_age = 0.0173637
seB_age = 0.0229966
t = b_age/seB_age
t
```

```
## [1] 0.7550551
```

```
critval= qt(1-0.05/2, 113-6)
critval
```

```
## [1] 1.982383
```

```
pval= 2*pt(-t, 113-6)
pval
```

```
## [1] 0.4518747
```

T-statistic = 0.7550551
Critical value = 1.982383
P-Value = 0.4518747

Because the t-statistic is smaller than the critical value, and the p-value is greater than $\alpha = 0.05$, we fail to reject H_0. In other words, we do not have enough evidence to conclude that Age is necessary to keep in the model in the presence of the other variables. We can drop `Age` as it is not necessary in this model.

c. **What is the R2 for this model? Write a sentence that interprets this value in context.**

```
SSr = (57.305 + 33.397 + 0.136 + 5.101 + 0.028)
SSres = 105.413
SSt = SSres + SSr

R2= SSr/SSt
R2
```

```
## [1] 0.4765468
```

$R^2 = 0.4765468$. The coefficient of determination informs us that about 47.65% of the variation in the percentage of patients who get an infection while hospitalized can be explained by `Stay` , the average length of stay, `Cultures` , a ratio of the number of cultures performed per number of patients with no infection (times 100), `Age` , the average patient age, `Census` , the number of patients in the hospital, and `Beds` , the number of beds in the hospital.

d. **Suppose we want to decide between two potential models:**

- Model 1: using x1, x2, x3, x4, x5 as the predictors for InfctRsk
- Model 2: using x1,x2 as the predictors for InfctRsk

Carry out the appropriate hypothesis test to decide which of models 1 or 2 should be used. Be sure to show all steps in your hypothesis test.

$H_0 = \beta_{x3} = \beta_{x4} = \beta_{x5} = 0$ (The parameters in the variables we wish to drop are 0, so the reduced model (model 2) is supported.)
H_A = At least one coefficient in H_0 is not 0 (The full model (model 1) is supported.)

```
SSr1 = (57.305 + 33.397 + 0.136 + 5.101 + 0.028)
SSres1 = 105.413
SSr_red2 = (57.305 + 33.397)
r = 3
n = 113
p = 6

F = ((SSr1-SSr_red2)/r) / ((SSres1)/(n-p))
F
```

```
## [1] 1.781422
```

```
critval = qf(0.95, r, n-p)
critval
```

```
## [1] 2.68949
```

```
p= 1 - pf(F, r, n-p)
p
```

```
## [1] 0.1550925
```

F-statistic = 1.781422
Critical value = 2.68869
P-Value = 0.1550925

Because the F-statistic is smaller than the critical value, and the p-value is greater than $\alpha = 0.05$, we fail to reject H_0. In other words, we do not have enough evidence to conclude that `Age` (x3), `Census` (x4) and `Beds` (x5) are necessary to keep in the model in the presence of the other variables. We can drop `Age` (x3), `Census` (x4) and `Beds` (x5) as they not necessary in this model, so a reduced model (Model 2) is better.

e. **Suppose we want to decide between two potential models:**

- Model 2: using x1,x2 as the predictors for InfctRsk
- Model 3: using x1, x2, x3, x4 as the predictors for InfctRsk

Carry out the appropriate hypothesis test to decide which of models 2 or 3 should be used. Be sure to show all steps in your hypothesis test.

$H_0 = \beta_{x3} = \beta_{x4} = 0$ (The parameters in the variables we wish to drop are 0, so the reduced model (model 2) is supported.)
H_A = At least one coefficient in H_0 is not 0 (The full model (model 3) is supported.)

```
SSr3 = (57.305 + 33.397 + 0.136 + 5.101)
SSres3 = 105.413 + 0.028
SSr_red2 = (57.305 + 33.397)
r = 2
n = 113
p = 5

F = ((SSr3-SSr_red2)/r) / ((SSres3)/(n-p))
F
```

```
## [1] 2.68205
```

```
critval = qf(0.95, r, n-p)
critval
```

```
## [1] 3.080387
```

```
p= 1 - pf(F, r, n-p)
p
```

```
## [1] 0.07297994
```

F-statistic = 2.68205
Critical value = 3.080387
P-Value = 0.07297994

Because the F-statistic is smaller than the critical value, and the p-value is larger than $\alpha = 0.05$, we fail to reject H_0 in favor of H_A. In other words, we do not have enough evidence to conclude that `Age` (x3) and `Census` (x4) are necessary to keep in the model in the presence of the other variables. We can drop `Age` (x3) and `Census` (x4) as they necessary in this model, so a reduced model (Model 2) is better.

---

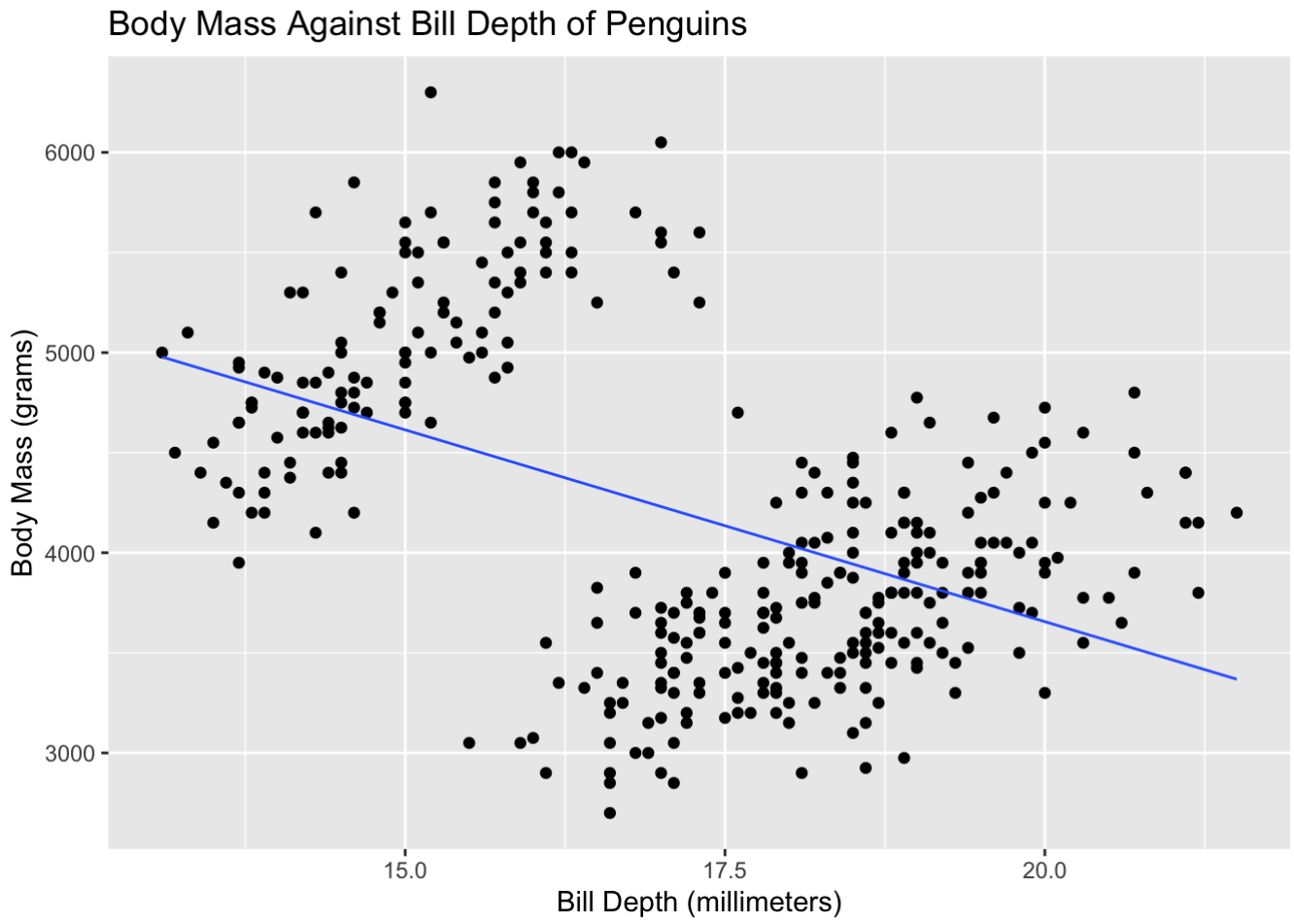
6. We will revisit the data set penguins from the palmerpenguins package. The data set contains size measurements for adult foraging penguins near Palmer Station, Antarctica. In this set of questions, we focus on exploring the relationship between body mass (y) and bill depth (x1) of three species of penguins.

```
library(palmerpenguins)
head(penguins)
```

```
## # A tibble: 6 × 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # ℹ 2 more variables: sex <fct>, year <int>
```

a. **Create a scatterplot of the body mass against the bill depth of the penguins. How would you describe the relationship between these two variables?**

```
ggplot(penguins, aes(x=bill_depth_mm, y=body_mass_g))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE, linewidth=0.5)+
  labs(x="Bill Depth (millimeters)", y="Body Mass (grams)", title="Body Mass Against Bil
l Depth of Penguins")
```

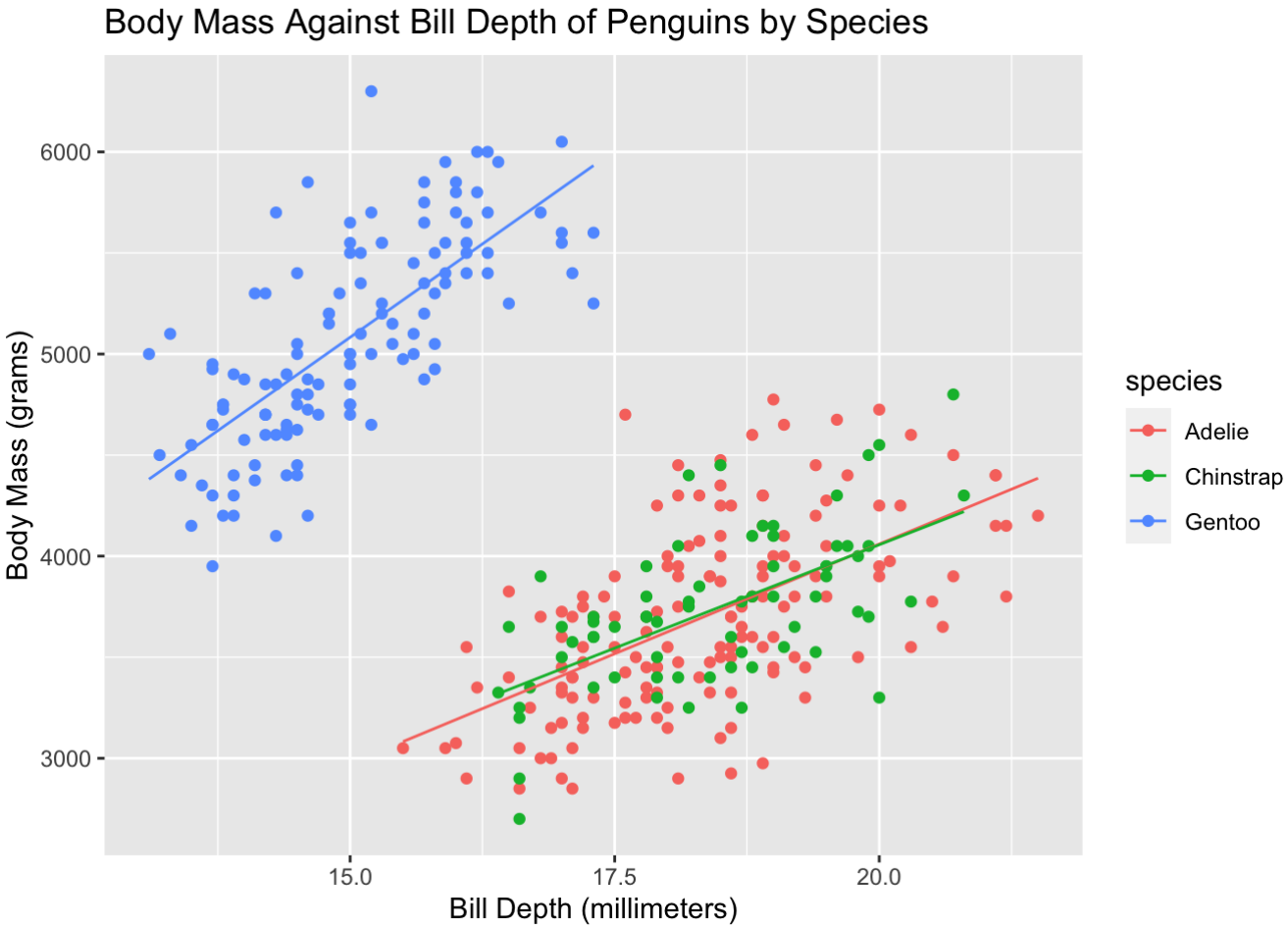


This plot seems to show a negative correlation between penguin body mass and bill depth overall. It is important to note that the points seem to be clustered in two very separate groups.

b. **Create the same scatterplot but now with different colored plots for each species. Also be sure to overlay separate regression lines for each species. How would you now describe the relationship**

between the variables?

```
ggplot(penguins, aes(x=bill_depth_mm, y=body_mass_g, color=species))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE, linewidth=0.5)+
  labs(x="Bill Depth (millimeters)", y="Body Mass (grams)", title="Body Mass Against Bil
l Depth of Penguins by Species")
```



Now that the scatterplot is colored by species, it is clear that there is a positive linear correlation between body mass and bill depth for penguins of each different species. Gentoo penguins seem to have larger body mass and shorter bill depths than the other species of penguins. The slopes are not exactly parallel, indicating that there may exist an interaction between the species of penguin and its bill depth; the impact of bill depth on body mass differs among the species So a regression model with interaction between species and bill depth may be appropriate.

c. **Create a regression with interaction between bill depth and species, i.e.** $y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_1 I_1 + \beta_5 x_1 I_2 + \varepsilon$, **where I1 and I2 are indicator variables where I1 = 1 for Chinstrap penguins and 0 otherwise, and I2 = 1 for Gentoo penguins and 0 otherwise. Write down the estimated regression equation for this model.**

```
contrasts(penguins$species)
```

```
##           Chinstrap Gentoo
## Adelie            0      0
## Chinstrap         1      0
## Gentoo            0      1
```

```
penguins.int <- lm(body_mass_g~bill_depth_mm*species, data=penguins)
summary(penguins.int)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_depth_mm * species, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -845.89 -254.74  -28.46  228.01 1161.41
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -283.28     437.94  -0.647   0.5182
## bill_depth_mm                   217.15      23.82   9.117   <2e-16 ***
## speciesChinstrap                247.06     829.77   0.298   0.7661
## speciesGentoo                  -175.71     658.43  -0.267   0.7897
## bill_depth_mm:speciesChinstrap  -12.53      45.01  -0.278   0.7809
## bill_depth_mm:speciesGentoo     152.29      40.49   3.761   0.0002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 354.9 on 336 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.807,  Adjusted R-squared:  0.8041
## F-statistic:   281 on 5 and 336 DF,  p-value: < 2.2e-16
```

We note that Adelie penguins are the reference class, which we confirmed with the contracts() function. Knowing this, we can interpret the estimated regression equation below:

The estimated regression equation is
$\hat{y} = -283.28 + 217.15(x) + 247.06(I_1) + -175.71(I_2) + -12.53(x * I_1) + 152.29(x * I_2)$ where y = body mass in grams, x= bill depth in millimeters, I_1= 1 if species is Chinstrap penguins and 0 otherwise, and I_2= 1 if species is Gentoo penguins and 0 otherwise.

d. **Carry out the relevant hypothesis test to see if the interaction terms can be dropped. What is the conclusion?**

$H_0 : \beta_4 = \beta_5 = 0$ (We could drop the interaction terms from our model in the presence of the other variables.)
H_A : At least one of the coefficients in H_0 is not 0 (We cannot drop the interaction terms from our model in the presence of the other variables.)

```
penguins.noint <- lm(body_mass_g~bill_depth_mm + species, data=penguins)
anova(penguins.noint, penguins.int)
```

```
## Analysis of Variance Table
##
## Model 1: body_mass_g ~ bill_depth_mm + species
## Model 2: body_mass_g ~ bill_depth_mm * species
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    338 44399670
## 2    336 42325191  2   2074479 8.2342 0.0003227 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the F-statistic is significant, and the p-value is smaller than $\alpha = 0.05$, we reject H_0 in favor of H_A. In other words, we have enough evidence to conclude that the interaction terns between species and bill depth are necessary to keep in the model in the presence of the other variables.

e. **Based on your answer in part 6d, write out the estimated regression equations relating body mass and bill depth, for each species of the penguins.**

$I_1 = 1$ if species is Chinstrap penguins and 0 otherwise, and $I_2 = 1$ if species is Gentoo penguins and 0 otherwise, so plugging in these indicator values we get the estimated regression equations for each species of penguins as follows:

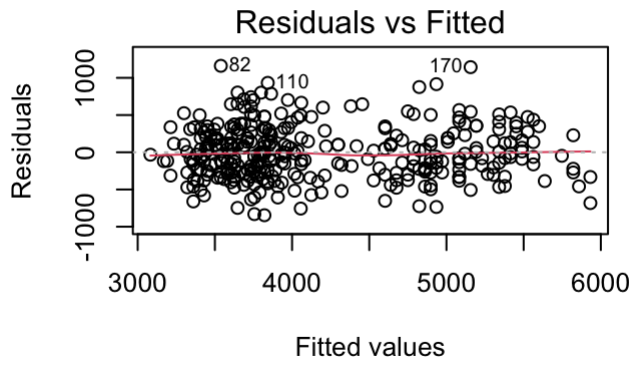
Chinstrap penguins:
$\hat{y} = -283.28 + 217.15(x) + 247.06(1) + -175.71(0) + -12.53(x * 1) + 152.29(x * 0)$
$= (-283.28 + 247.06) + (217.15 + -12.53)(x)$
$= -36.22 + 204.62(x)$

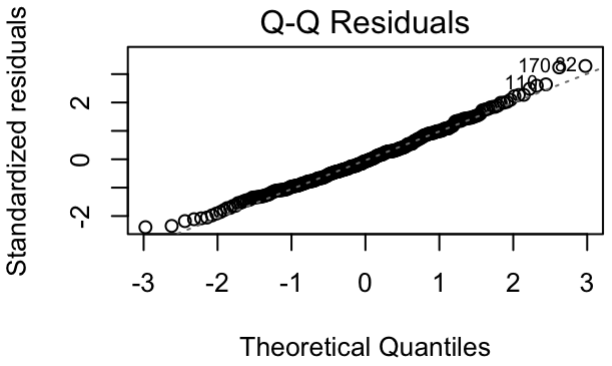
Gentoo penguins:
$\hat{y} = -283.28 + 217.15(x) + 247.06(0) + -175.71(1) + -12.53(x * 0) + 152.29(x * 1)$
$= (-283.28 + -175.71) + (217.15 + 152.29)(x)$
$= -458.99 + 369.44(x)$

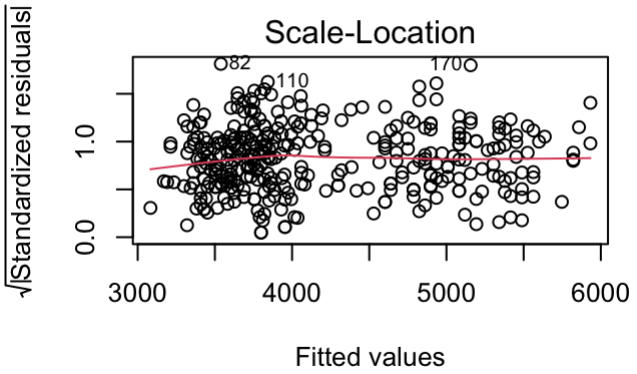
Adelie penguins:
$\hat{y} = -283.28 + 217.15(x) + 247.06(0) + -175.71(0) + -12.53(x * 0) + 152.29(x * 0)$
$= -283.28 + 217.15(x)$

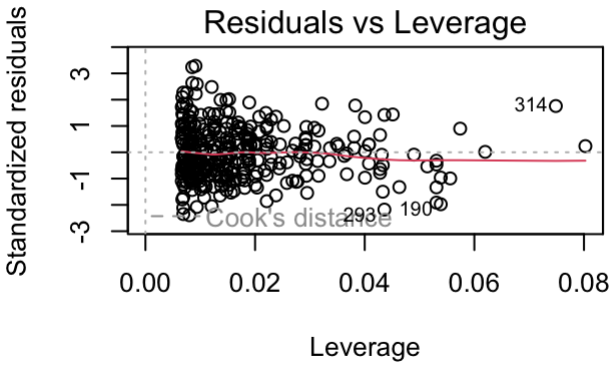
f. **Assess if the regression assumptions are met, for the model you will recommend to use (based on part 6d).**

```
par(mfrow = c(2, 2))
plot(penguins.int)
```

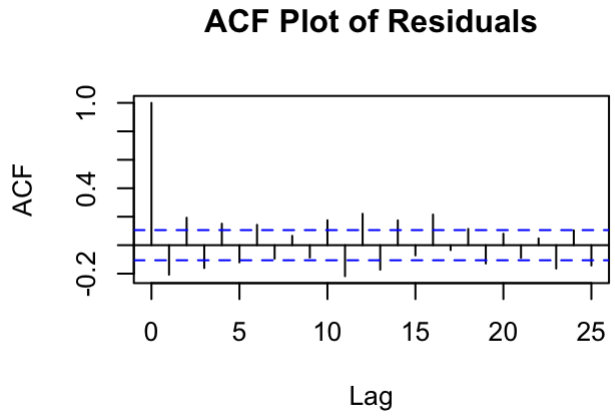









```
acf(penguins.int$residuals, main="ACF Plot of Residuals")
```



All of the assumptions appear to be met. This is clear because the residuals are scattered randomly around 0 in the residual plot with constant variance.

When looking at the ACF plot, we see quite a few significant lag values, but this is because the data is sorted by species, and because gentoo penguins have the largest body mass it is somewhat sorted by weight.

g. **Briefly explain if we can conduct pairwise comparisons for the difference in mean body mass among all pairs of species for given values bill depth, i.e.,**

i. Adelie and Chinstrap,
ii. Adelie and Gentoo,
iii. Chinstrap and Gentoo.

If we are able to, conduct Tukey's multiple comparisons and contextually interpret the results of these hypothesis tests.

We cannot conduct pairwise comparisons for the difference in mean body mass among all pairs of species for given values bill depth because there is a significant interaction between penguin species and bill depth. This is clear when looking at the slopes of the regression equations for each species (6e). For this reason we cannot conduct Tukey's multiple comparisons as it will produce an error due to interaction.

---

7. (You may only use R as a simple calculator or to find p-values or critical values) This question is based on data about teacher salaries from the 50 states plus DC (so n = 51) in the mid 1980s. The variables are:

- `PAY` , y: average annual public school teacher salary, in dollars.

- `SPEND` , x1: Spending on public schools per student, in dollars.
- `AREA` : Region (North, South, West).

Table 1 below provides some summary statistics of the data:

Region	n	Mean PAY	Mean SPEND
North	21	$24424	$3902
South	17	$22894	$3274
West	13	$26159	$3919

Table 1: Summary Statistics of Teacher Pay

a. **Based only on Table 1, briefly comment on the relationship between geographic area and mean teacher pay.**

Mean teacher pay seems to be highest in the West region and lowest in the South.

b. **Based only on Table 1, briefly comment on the relationship between mean public school expenditure (per student) and mean teacher pay.**

There seems to be a positively correlated relationship between the public school expenditure per student and the mean teacher pay. The larger the value teachers are paid seems to correlate postitvely with larger amounts spent per student on average.

c. **Briefly explain why using a multiple linear regression model with teacher pay as the response variable with geographic area and public school expenditure (per student) can give further insight into the relationship(s) between these variables.**

A multiple linear regression model with teacher pay as the response variable with geographic area and public school expenditure (per student) can give further insight into the relationship(s) between these variables because we will be able to address how region and expenditure interact to impact mean teacher pay. This will show if there is a significant difference between mean pay based on region when holding expenditures constant.

Use the following info to answer the rest of question 7.

We want to see if geographic region and spending on public schools affect the average public teacher pay. A regression with no interactions was fitted, i.e.,

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 I_2 + \beta_3 I_3,$$

where I2 and I3 are the dummy codes for AREA. I2 = 1 if AREA = South, 0 otherwise, and I3 = 1 if AREA = West, 0 otherwise. The following output from R is shown below

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.160e+04  1.334e+03   8.690 2.43e-11 ***
SPEND       3.289e+00  3.176e-01  10.354 1.03e-13 ***
AREASouth   5.294e+02  7.669e+02   0.690   0.4934
AREAWest    1.674e+03  8.012e+02   2.089   0.0422 *
##########################################
##Variance-Covariance matrix for beta hats##
##########################################
             (Intercept)        SPEND     AREASouth      AREAWest
(Intercept) 1780535.6980 -393.5597348 -491859.07243 -2.381145e+05
SPEND          -393.5597    0.1008967      63.18227 -1.870101e+00
AREASouth   -491859.0724   63.1822716  588126.71689  2.442380e+05
AREAWest    -238114.5499   -1.8701007  244238.02959  6.418738e+05
```

d. **What is the estimate of β2? Give an interpretation of this value.**

β_2 = 529.4. The estimated difference in the mean annual public school teacher salary between public schools in the South and North regions is 529.4, for given spending on public schools per student. We interpret this as the average annual public school teacher salary for public schools in the South region is $529.4 higher than schools in the North, when controlling for public school expenditure per student.

e. **Using the Bonferroni procedure, compute the 95% family confidence intervals for the difference in mean response for PAY between teachers in the**

i. North region and the South region;
ii. North region and the West region;
iii. South region and the West region,
while controlling for expenditure.

```
# beta_j +- t_(1-alpha/(2*g)),(n-p) * se(beta_j)
g= 3
p= 4
n= 51
t= qt(1-0.05/(2*g), n-p)

b_2 = 529.4
se_b2 = 7.669e+02
cat("North - South: [", b_2 - (t*se_b2), ",", b_2 + (t*se_b2),"] \n")
```

```
## North - South: [ -1374.578 , 2433.378 ]
```

```
b_3 = 1674
se_b3 = 8.012e+02
cat("North - West: [", b_3 - (t*se_b3), ",", b_3 + (t*se_b3),"] \n")
```

```
## North - West: [ -315.1348 , 3663.135 ]
```

```
#(b_2 - b_3)
b2b3 = (529.4 - 1674)
se_b2b3 = sqrt(588126.71689+ 641873.8 - 2*244238)
cat("South - West: [", b2b3 - (t*se_b2b3), ",", b2b3 + (t*se_b2b3),"] ")
```

```
## South - West: [ -3282.493 , 993.2934 ]
```

The difference in mean pay for teachers in the North region and the South region is between -$1374.578 and $2433.378. The difference in mean pay for teachers in the North region and the West region is between -$315.1348 and $3663.135. The difference in mean pay for teachers in the South region and the West region is between -$3282.493 and $993.2934

f. **What do your intervals from part 7e indicate about the effect of geographic region on mean annual salary for teachers (while controlling for expenditure)?**

All three CIs include 0, so there is NOT a significant difference in mean annual public school teacher salary of schools between any pairs of regions, when controlling for public school expenditure per student.