HW2_6120

Jacqui Unciano

2023-06-26

```
library(MASS)
library(ggplot2)
library(dplyr)
```

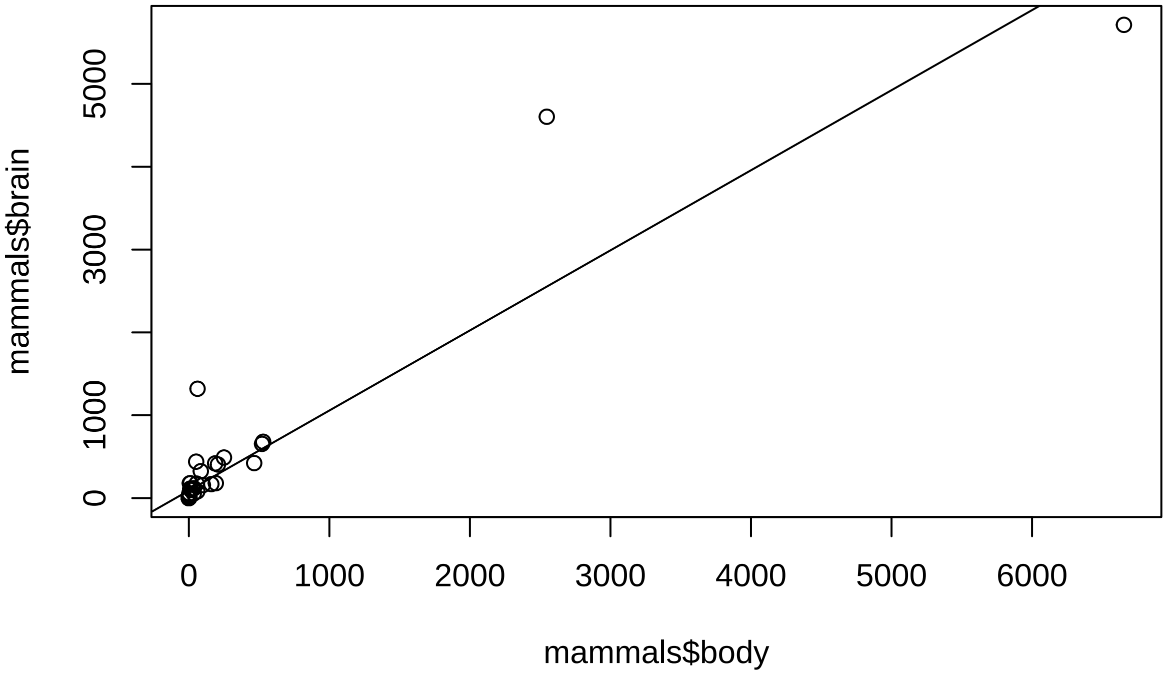
```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(faraway)
library(palmerpenguins)
library(lawstat)
```

Question 1

(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?

```
plot(mammals$body, mammals$brain)
abline(lm(formula=brain~body, data=mammals))
```



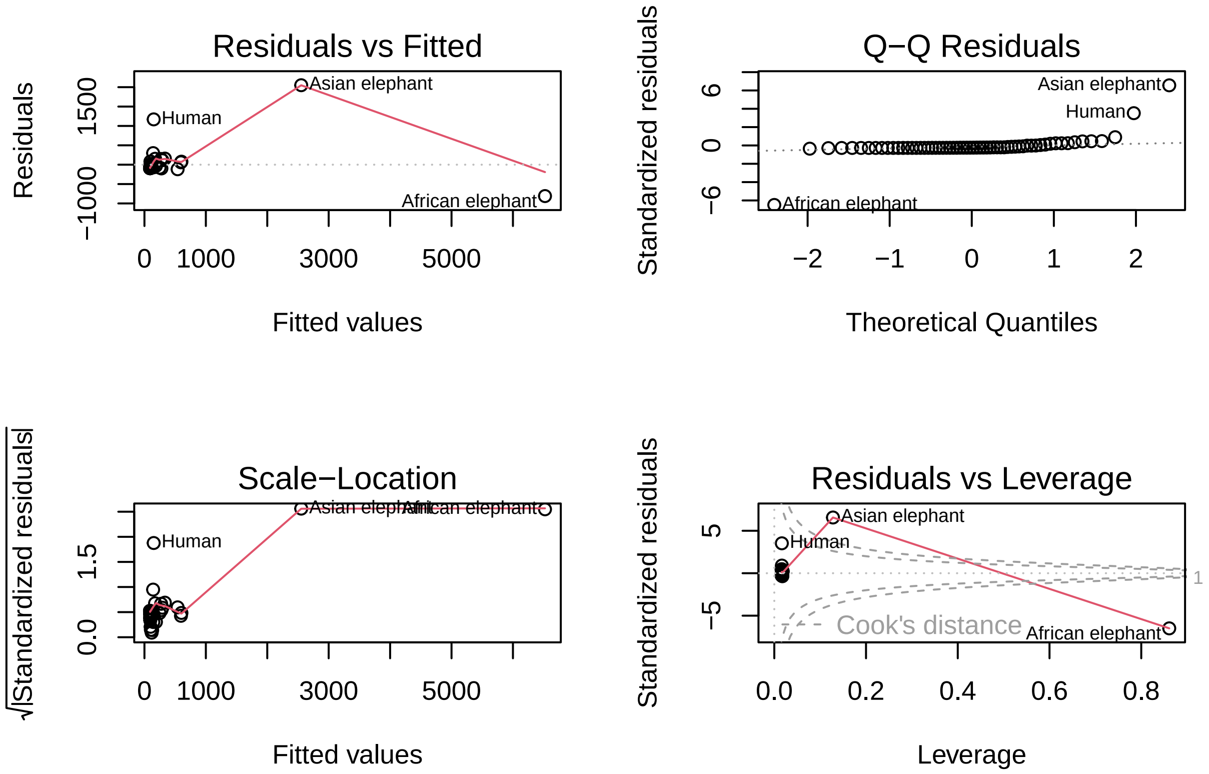
The general pattern of the graph seems to follow a more curvilinear path rather than a linear path. This suggests that we may need to transform one of the variables in order to meet the assumptions for SLR.

Assumption 1 (NOT met): the current regression line seems to either under or over estimate the data points depending on the value of the x.

Assumption 2 (NOT met): the data points seem to increase in varience as x increases (around the lower values of x).

(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?

```
mres = lm(formula=brain~body, data=mammals)
par(mfrow=c(2,2))
plot(mres)
```



Assumption 1 (NOT met): the current regression line seems to either under or over estimate the data points depending on the value of the x. There's no even spread of points above or below the red line.

Assumption 2 (NOT met): the data points seem to increase in variance as x increases (around the lower values of x). But it's hard to assess the vertical spread with the residual plot since there aren't a lot of data points for larger animals with more body mass.

Assumption 3 (met): I believe the data points are independent of each other, so the independence assumption is met.

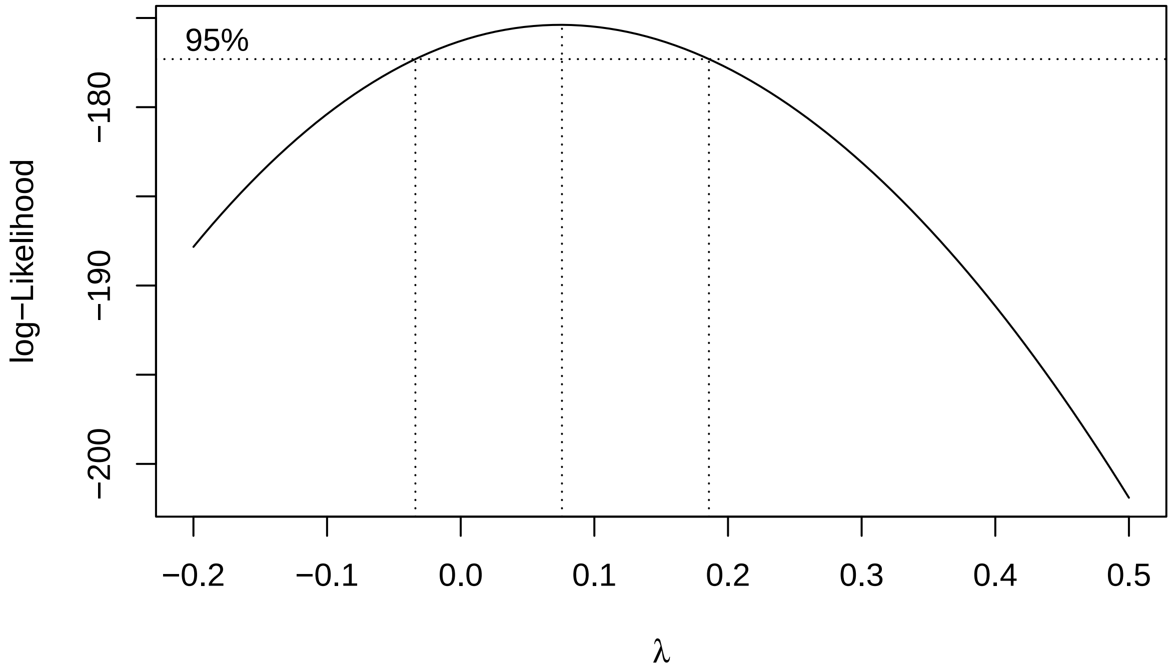
Assumption 4 (NOT met): the QQ plot does not reflect a normal distribution of data points for this dataset because the points are mostly on a horizontal axis rather than a 45-degree axis.

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

Yes, I would say that both variables need to be transformed because both assumption 1 and 2 are violated. I would transform the y variable first (because the variance is increasing, I would try a root transformation) and if that didn't solve the assumption 1 violation, I would also transform the x variable based on the shape of the scatter plot (in this case, try a log transformation).

(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.

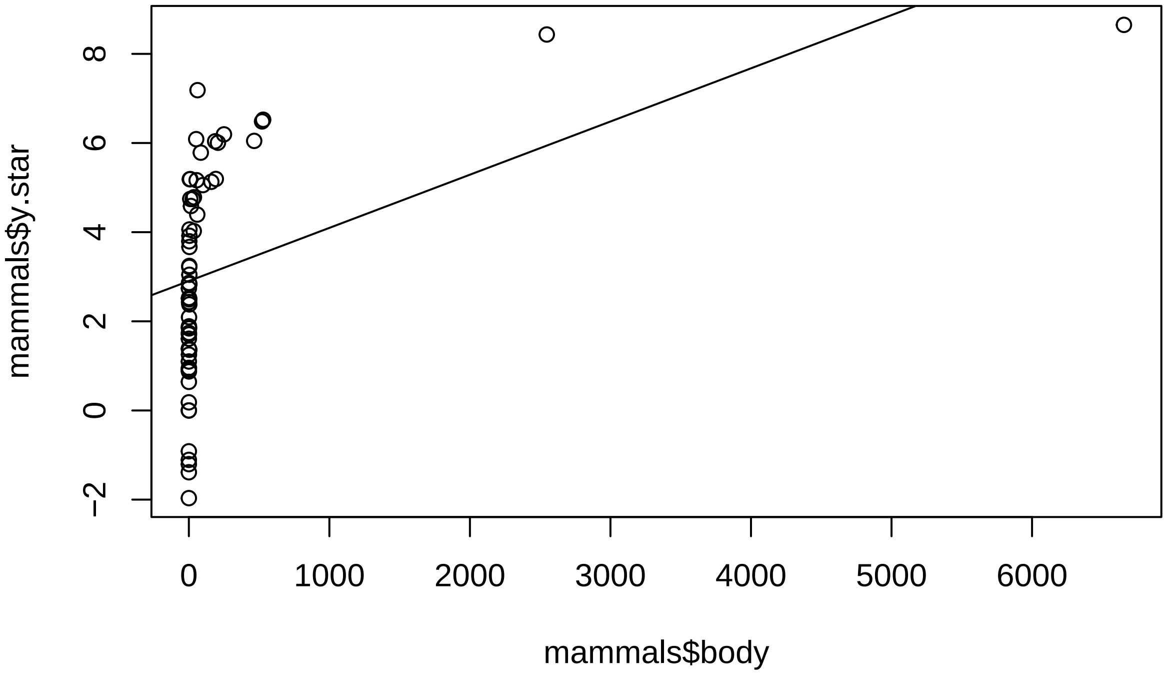
```
MASS::boxcox(mres, lambda=seq(-0.2,0.5,0.1))
```



Because zero is within the 95% confidence interval as seen in the boxcox plot, we can try a log transformation on the response variable (brain weight) and see if that will fix the assumption violations.

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

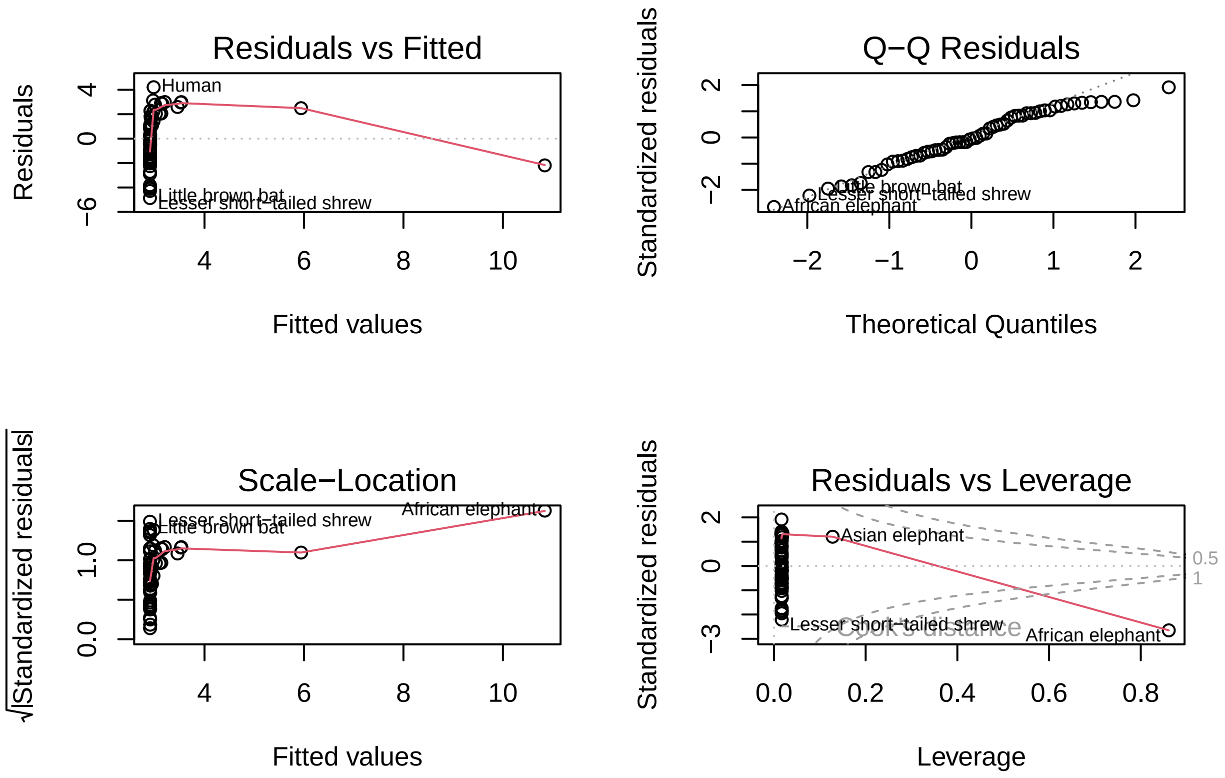
```r
mammals$y.star = log(mammals$brain)
plot(mammals$body, mammals$y.star)
abline(lm(formula=y.star~body, data=mammals))
```



Yes, the error mean zero assumption has been violated, meaning the mean of the residuals for each x value does not equal to zero. We can see this violation because the scatterplot follows a nonlinear trend/curvilinear trend.

(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?

```
ysmres = lm(formula=y.star~body, data=mammals)
par(mfrow=c(2,2))
plot(ysmres)
```

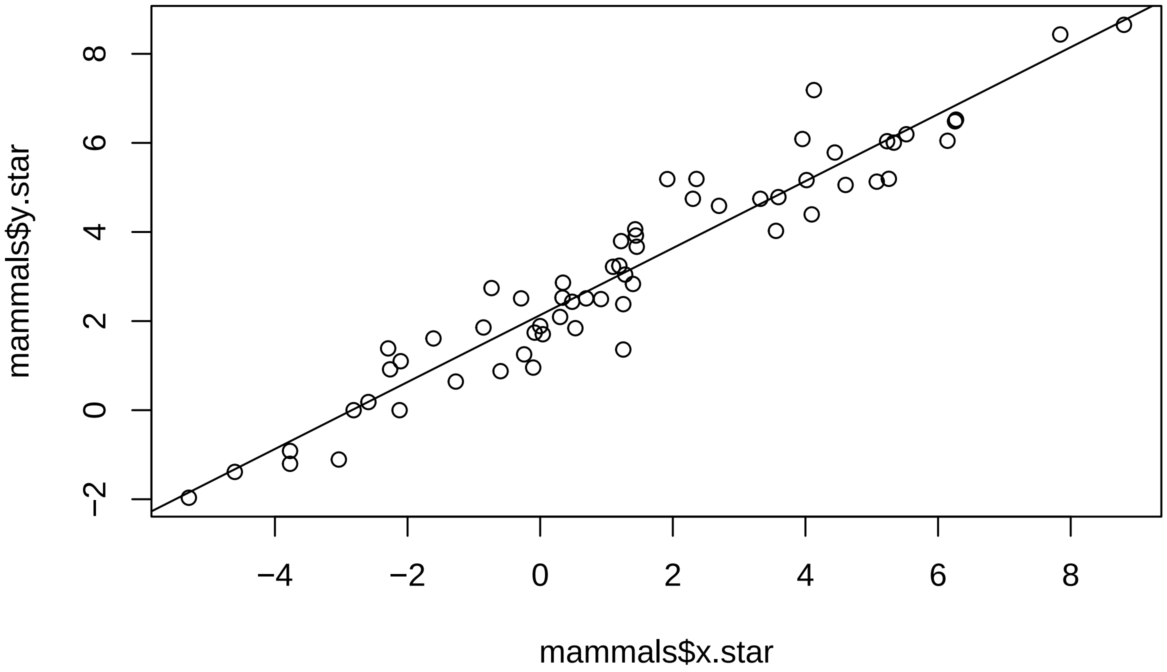


As mentioned before, the mean error zero assumption has been violated. This claim is supported by the residual plot of fitted values and the QQ plot. The variance assumption is not violated as there are no obvious places where the variance is increasing or decreasing as y increases.

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

I would transform the x variable. I would first try a log transformation because of the shape of the plot and see if that helps fix the assumption violation. If not, I would look for other transformations based on the shape of the graph.

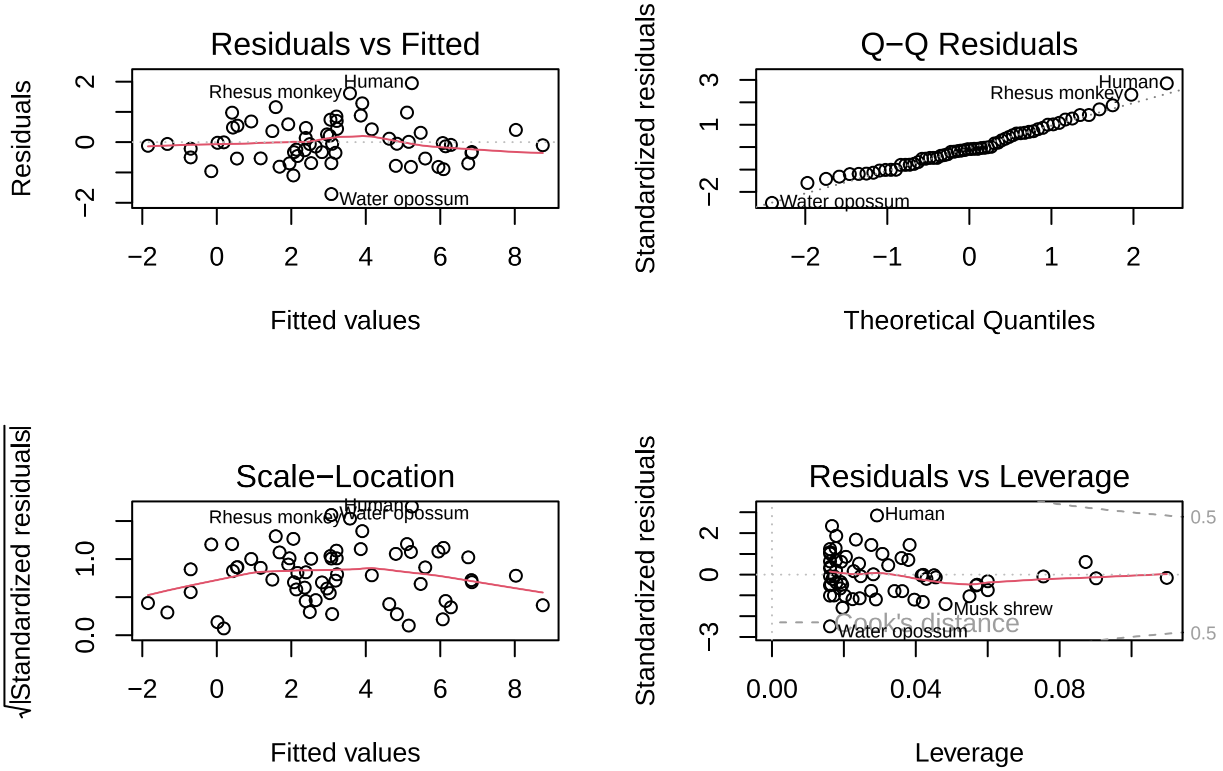
```
mammals$x.star = log(mammals$body)
plot(mammals$x.star, mammals$y.star)
abline(lm(formula=y.star~x.star, data=mammals))
```



There are no obvious signs of assumption violations. The data points are evenly distributed over the regression line and the variance doesn't seem to change for the most part. We could argue that the variance increases a little at the beginning, but it's not too prevalent, so I wouldn't worry about it.

(h) Fit a simple linear regression to y-star against x-star, 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.

```
yxsmres = lm(formula=y.star~x.star, data=mammals)
par(mfrow=c(2,2))
plot(yxsmres)
```



As you can see in the residual plot and QQ plot above, the error mean zero assumption has been fixed. The residual plot shows that the points are more evenly distributed across the residual line and the QQ plot is more linear at a 45-degree angle. The variance doesn't seem to change either (for the most part).

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

```
summary(yxsmres)
```

```
##
## Call:
## lm(formula = y.star ~ x.star, data = mammals)
##
## 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 ***
## x.star       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 as follows: y-star = 2.13479+0.75169x-star, where y-star=log(brain) and x-star=log(body)

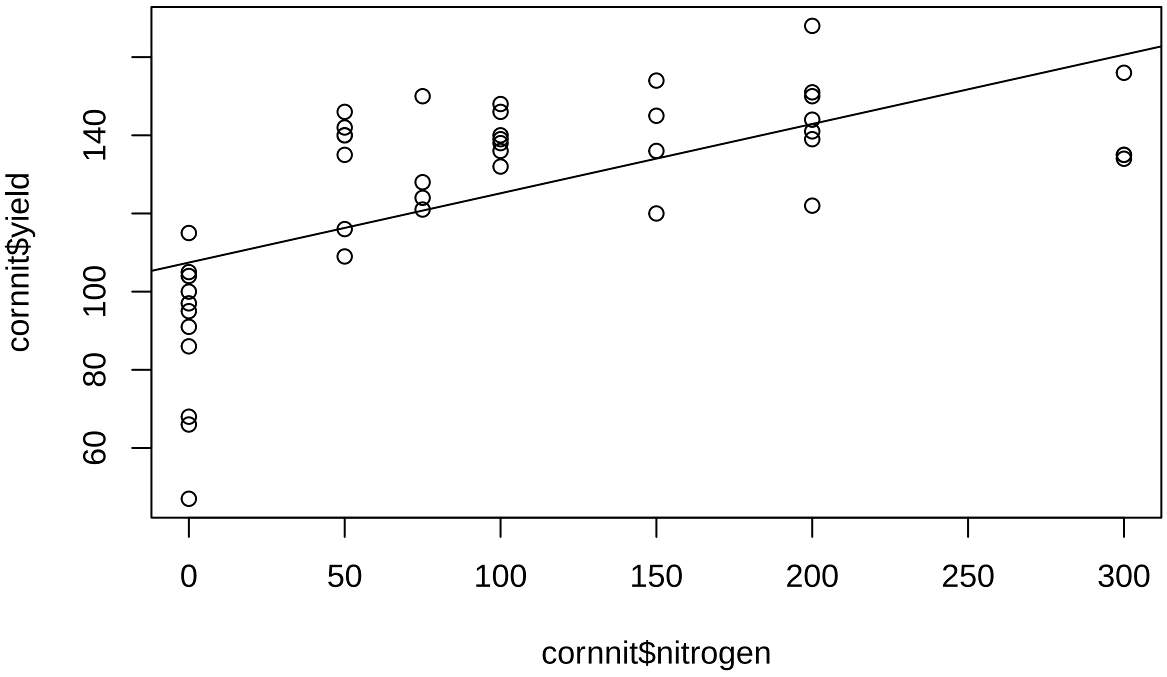
For a 1% increase in the body weight of a land mammal, the predicted brain weight of a land mammal is multiplied by (1+0.01)^0.75169=1.007508.

Question 2

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

The predictor variable is the amount of nitrogen per acre and the predictor variable is the amount of corn yield per arce.

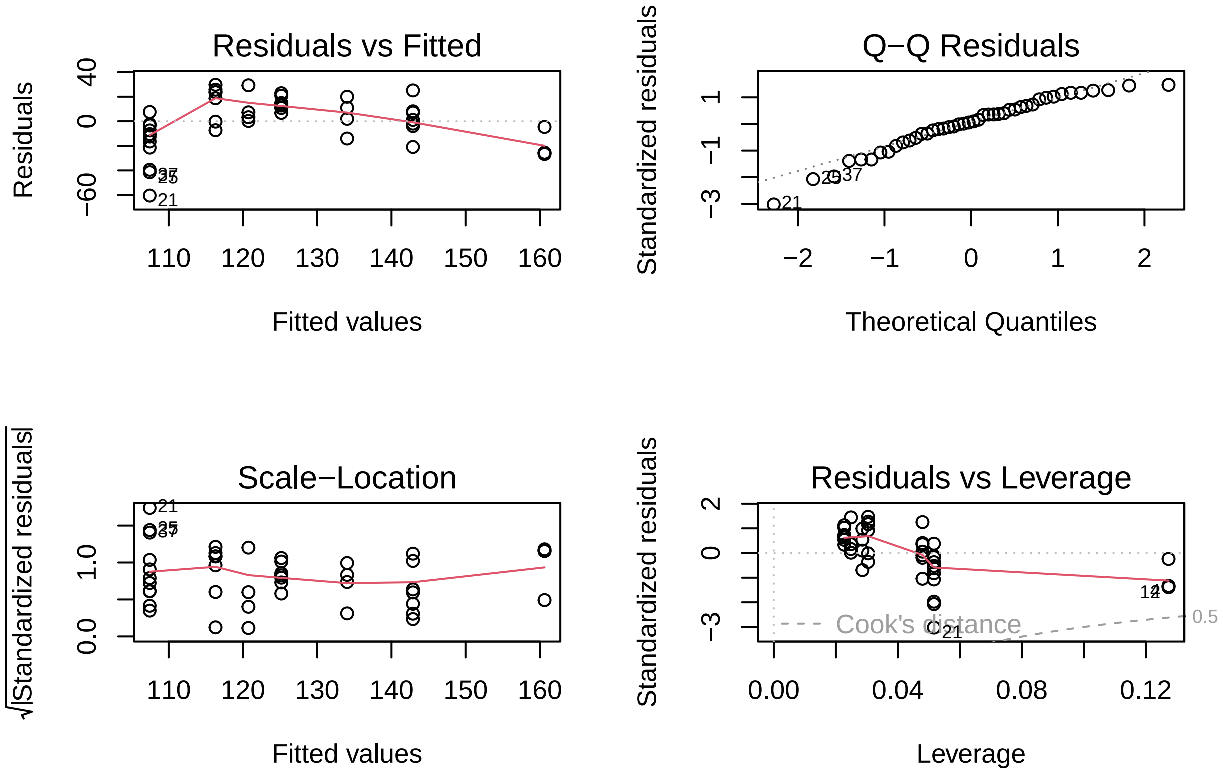
```r
plot(cornnit$nitrogen, cornnit$yield)
abline(lm(yield~nitrogen, data=cornnit))
```



It looks like the data is follow more of a nonlinear trend, suggesting a transformation on the predictor variable. So, the more nitrogen fertilizer added per acre, the more corn is yielded per acre, but there's yield capacity/limitation after reaching a certain level/amount of nitrogen fertilizer.

(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.

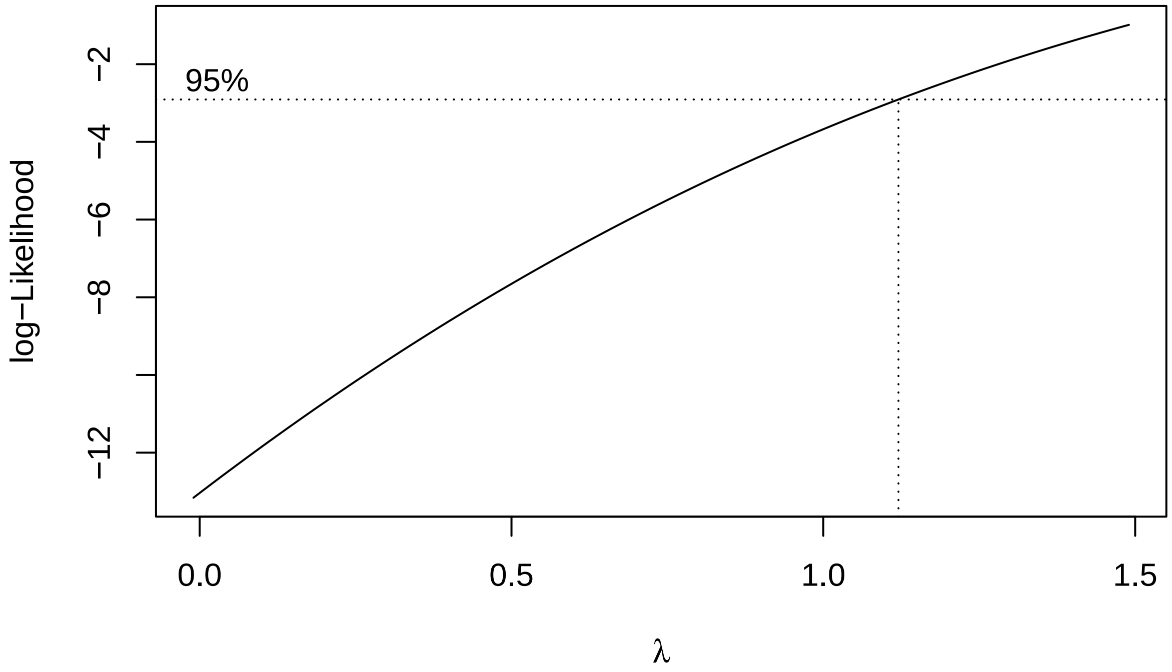
```r
cres = lm(yield~nitrogen, data=cornnit)
par(mfrow=c(2,2))
plot(cres)
```



The vertical variance along the residual plot seems pretty stable, so currently no response variable transformation is needed. However, it seems like the error mean zero assumption is violated, as seem in both the residual and QQ plots. This suggests a transformation on the predictor variable, and based on the shape of the scatterplot in 2a, I would try a log transformation first.

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

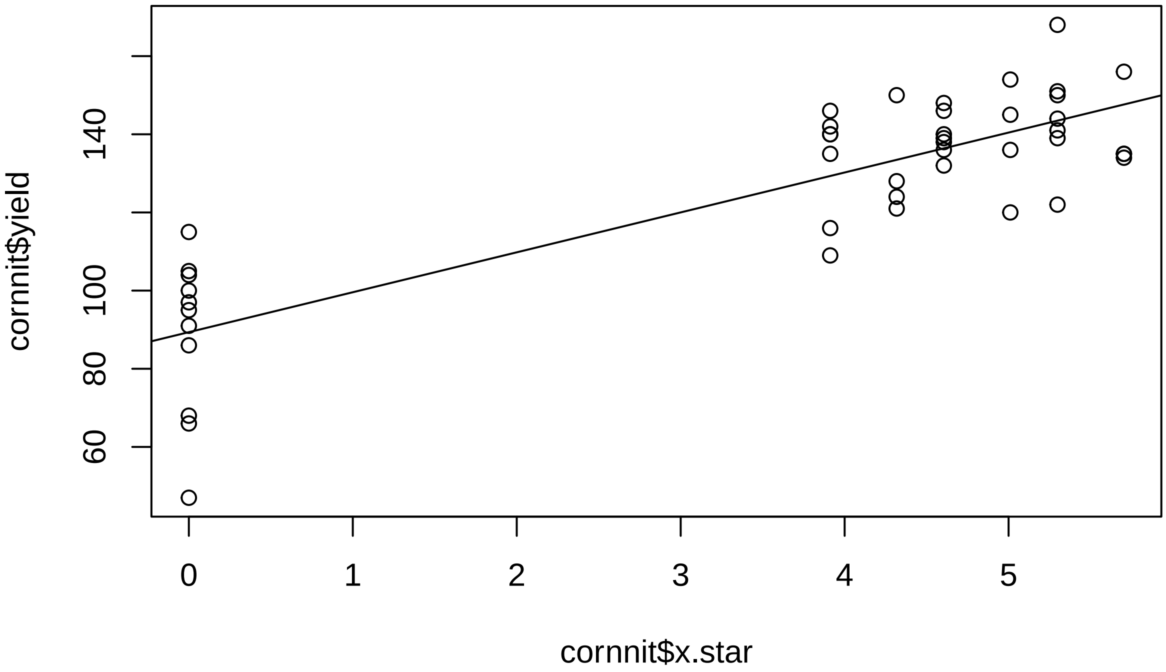
```r
MASS::boxcox(cres, lambda=seq(-0.01,1.5,0.1))
```



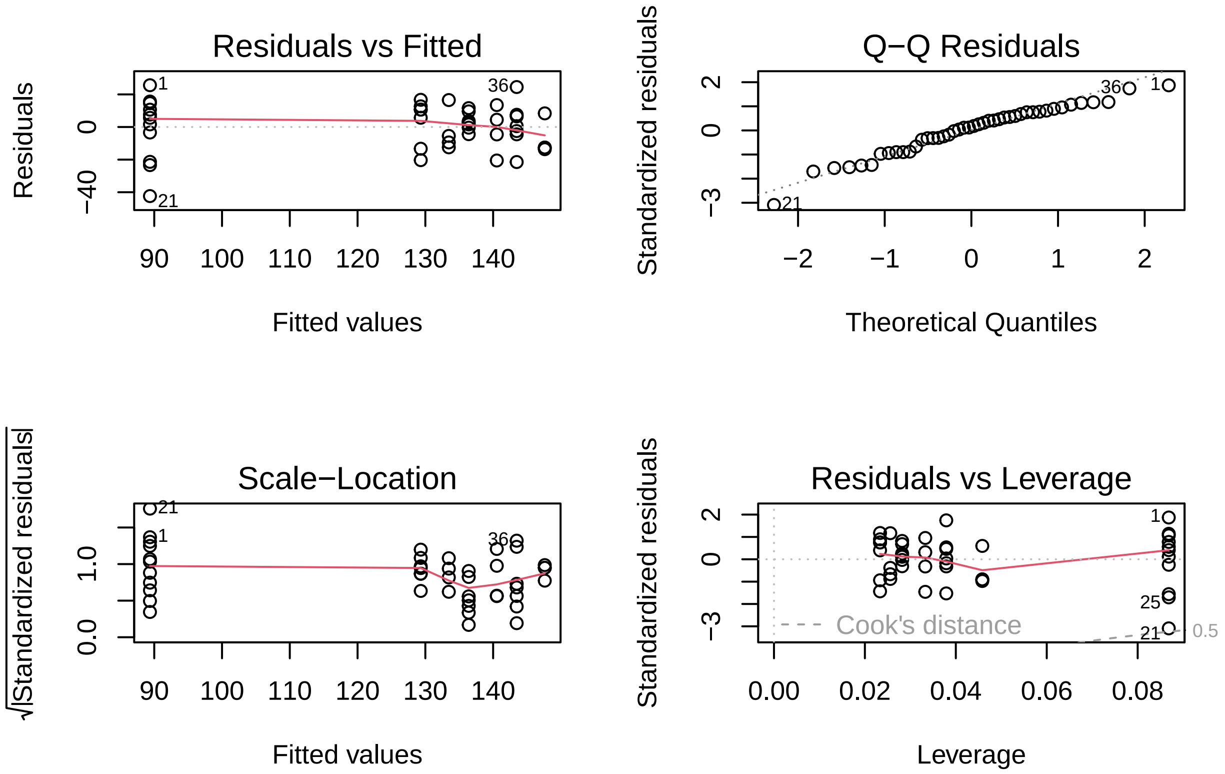
The boxcox plot supports my analysis of not transforming the response variable. 1 is so close to the 95% confidence interval, which means a transformation on the response variable may not be needed. 1.5 is within the 95% CI, so if I really needed too, I would try a transformation on the response variable by y^1.5.

(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?

```
cornnit = cornnit%>%
  mutate(x.star = ifelse(nitrogen==0, 0, log(nitrogen)))
xcres = lm(yield~x.star, data=cornnit)
plot(cornnit$x.star, cornnit$yield)
abline(lm(yield~x.star,data=cornnit))
```



```r
par(mfrow=c(2,2))
plot(xcres)
```



There is a significant improvement by transforming the predictor variable (nitrogen). The scatterplot shows the linear trend of the data points and the somewhat constant variance (or negligible variance fluctuation). The residual plots show that the data points are evenly distributed and the horizontal band doesn't show any obvious variance fluctuation.

```
summary(xcres)
```

```
##
## Call:
## lm(formula = yield ~ x.star, data = cornnit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.368 -10.214   2.115  10.643  25.632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   89.368      4.227   21.14  < 2e-16 ***
## x.star        10.213      1.019   10.02 1.05e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.34 on 42 degrees of freedom
## Multiple R-squared:  0.7052, Adjusted R-squared:  0.6982
## F-statistic: 100.5 on 1 and 42 DF,  p-value: 1.048e-12
```

Therefore, the regression equation is as follows: y = 89.368 + 10.213x-star.

Question 3

```r
n = read.table("C:\\Users\\jacqu\\Downloads\\nfl.txt", header=TRUE)
```

(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.

```r
nres = lm(y~x2+x7+x8, data=n)
summary(nres)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = n)
##
## 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
```

y = -1.808372 + 0.003598x2 + 0.193960x7 - 0.004816x8

(b) Interpret the estimated coefficient for the predictor x7 in context.

For every 1% increase in rushing, the predicted number of games won increase by 0.193960, while keeping the other predictor variables 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.

```r
newdat = data.frame(x2=2000, x7=48, x8=2350)
predict(nres, newdat, interval="prediction")
```

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

The predicted number of games won for a team with 2000 passing yardage, 48% rushing plays, and 2350 rushing yardage on the opposing team is 3.381448 (or 3 games contextually). We are 95% confident that the predicted number of games won for a team with 2000 passing yardage, 48% rushing plays, and 2350 rushing yardage on the opposing team is between -0.5163727 and 7.279268. Or to make sense contextually, between 0 and 7 games (can't win 7.279268 games or negative 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.

H0: beta-hat2,7,8=0, none of the variables contribute to predicting the number of wins

HA: beta-hat2,7,8=/=0, at least one of the variables contribute to predicting the number of wins

F-stat: 29.44, F-crit: 3.008787, p-value: 3.273e-08

Conclusion: Yes, there is evidence suggesting that at least one of the predictor variables contributes to predicting the number of wins.

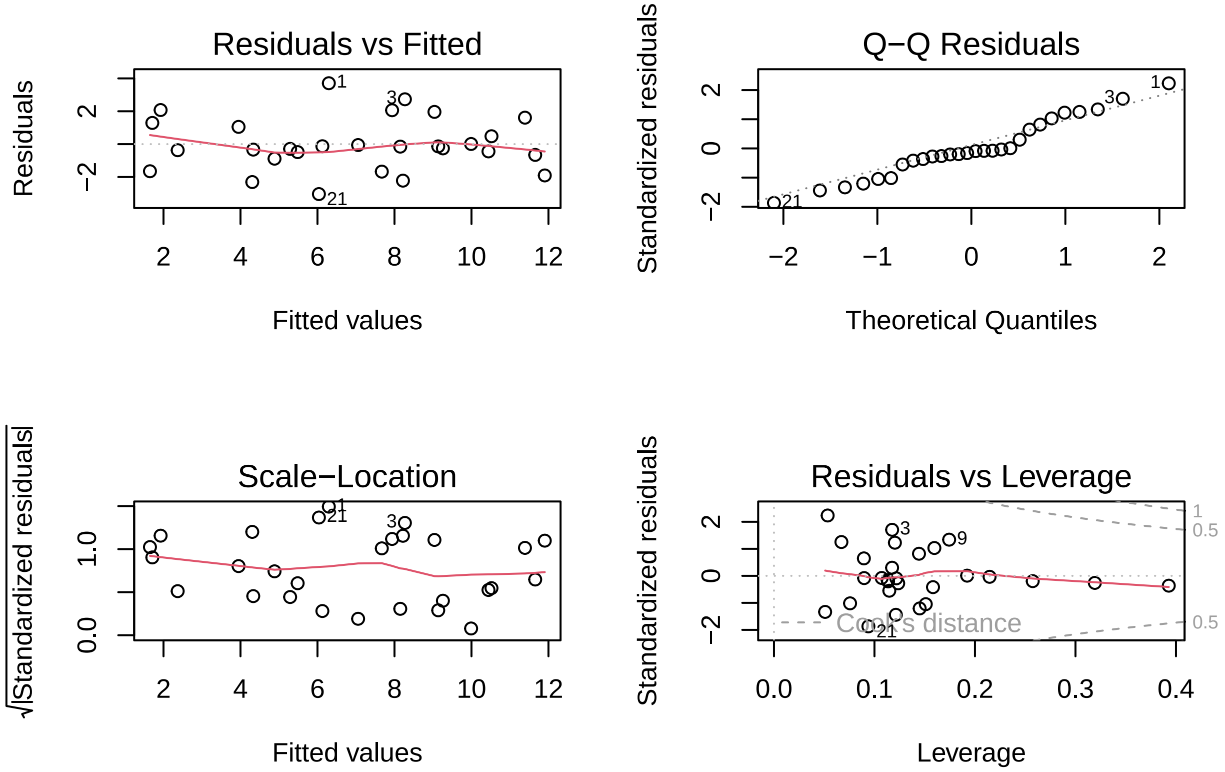
(e) Report the value of the t statistic for the predictor x7. What is the relevant conclusion from this t statistic?

t-stat: 2.198, t-crit = 2.063899

Conclusion: There is evidence to suggest that we should not drop the predictor x7.

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

```r
par(mfrow=c(2,2))
plot(nres)
```



The residual plot shows that there is not obvious change in variance as the fitted values increase and that the residuals for the fitted values are evenly distributed across the horizontal band. The QQ plot isn't perfect, but it's not too skewed from the 45-degree angle, suggesting no violation in normality.

(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?

```
nx1res = lm(y~x1+x2+x7+x8, data=n)
summary(nx1res)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x7 + x8, data = n)
##
## 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
```

t-stat: 0.549, t-crit: 2.068658

Conclusion: There is evidence that we should drop the predictor variable x1.

"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."

I disagree with this statement. An insignificant t test for a coefficient in MLR does not automatically mean that the predictor doesn't have a linear relationship with the number of wins. We would need to run a SLR to test the linear relationship between the x and y variables.

Question 4

(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.

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

```
fullres = lm(hipcenter~., data=seatpos)
summary(fullres)
```

```
##
## 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
```

We have a very significant ANOVA F test (pval: 1.306e-05) but a lot of the t tests are insignificant. This discrepancy is an indication that multicollinearity may be present within the data set.

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

```r
cor(seatpos[,c(1:8)])
```

```
##                 Age     Weight     HtShoes          Ht     Seated       Arm
## Age      1.00000000 0.08068523 -0.07929694 -0.09012812 -0.1702040 0.3595111
## Weight   0.08068523 1.00000000  0.82817733  0.82852568  0.7756271 0.6975524
## HtShoes -0.07929694 0.82817733  1.00000000  0.99814750  0.9296751 0.7519530
## Ht      -0.09012812 0.82852568  0.99814750  1.00000000  0.9282281 0.7521416
## Seated  -0.17020403 0.77562705  0.92967507  0.92822805  1.0000000 0.6251964
## Arm      0.35951115 0.69755240  0.75195305  0.75214156  0.6251964 1.0000000
## Thigh    0.09128584 0.57261442  0.72486225  0.73496041  0.6070907 0.6710985
## Leg     -0.04233121 0.78425706  0.90843341  0.90975238  0.8119143 0.7538140
##              Thigh         Leg
## Age     0.09128584 -0.04233121
## Weight  0.57261442  0.78425706
## HtShoes 0.72486225  0.90843341
## Ht      0.73496041  0.90975238
## Seated  0.60709067  0.81191429
## Arm     0.67109849  0.75381405
## Thigh   1.00000000  0.64954120
## Leg     0.64954120  1.00000000
```

The pair of predictor variables Arm, Weight, Seated, Ht, HtShoes, Leg, and Thigh have high correlations, suggesting a degree of multicollinearity. Only the pairs involving the predictor variable Age had lower correlations.

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

```r
vif(fullres)
```

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

The VIF is incredibly high for HtShoes and Ht, and fairly high for Seated and Leg. This is a good indication that multicollinearity is present between the variables.

(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.

The correlation between the length of body parts is very high.

(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.

I would like to keep the variables with the highest VIF because the height is the most correlated to the other body part variables. This means our regression will be more reliable.

(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.

```
redres = lm(hipcenter~Age+Weight+Ht, data=seatpos)
vif(redres)
```

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

There is still multicollinearity present, but it's not going to be too much of a concern now. This is proven by the VIFs that are small (less than 5).

(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.

```
anova(redres, fullres)
```

```
## 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
```

H0: (HtShoes, Steated, Arm, Thigh, Leg) == 0, supports the reduced model

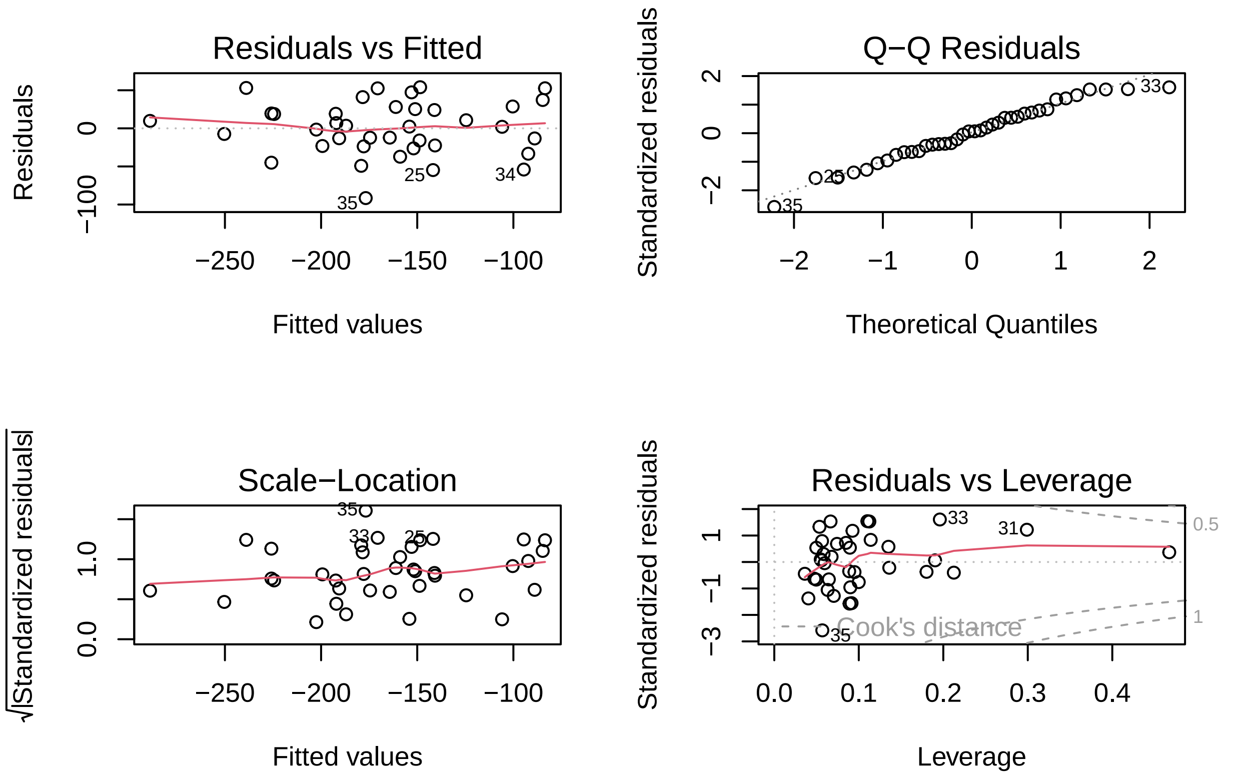
HA: at least one (HtShoes, Steated, Arm, Thigh, Leg) =/= 0, supports the full model

F-stat: 0.5623, pval: 0.7279 ==> fail to reject the reduced model

Conclusion: There is not enough evidence to support the full model, so we should drop the predictors (HtShoes, Steated, Arm, Thigh, Leg) and keep the reduced model.

(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(redres)
```



The residual plot could argue that there is an increase in variation as the fitted values increase, but I would argue that it is negligible.

As for the error mean 0 assumption, the mean of residuals for each fitted value seems to be around 0, so it's not violated. Obviously, it's not perfect, but I don't think the assumption is violated.

(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.

```r
summary(redres)
```

```
##
## 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(fullres)$r.squared
```

```
## [1] 0.6865535
```

```
summary(fullres)$adj.r.squared
```

```
## [1] 0.6000855
```

The estimated regression equation is as follows:

HipCenter = 528.297729 + 0.519504Age + 0.0042711Weight - 4.211905Height

The R-sq value for the reduced model is 0.6562 and for the full model, it's 0.6865535. This makes sense, as the R-sq value will always increase as we add more and more predictors to the model.

The adjusted R-sq value for the reduced model is 0.6258 and for the full model, it's 0.6000855. That's about a 2% difference, and to me, this indicates that there are more variables that we should look into to add to the reduced model. The higher adj. R-sq for the reduced model also supports that the reduced model is a better fit for predicting hip center than the full model.

Question 5

```
n=113
```

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

beta-hat1 = 0.2055252

For every unit increase in the average length of stay, the percentage of patients who get an infection while hospitalized increases by 0.2055252 percent, holding all other variables 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?

```
## test stat = (beta-hatj)/se(beta-hatj)
tstat = 0.0173637/0.0229966
tstat
```

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

```r
## p-val = 2\ast(1-pt(test stat, n-p))
2*(1-pt(tstat,113-6))
```

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

```r
## crit value = qt(1-alpha/2, n-p)
qt(0.975, 113-6)
```

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

H0: Age==0, the variable Age can be dropped from the model

HA: Age=/=0, the variable Age cannot be dropped from the model

t-stat: 0.7550551, t-crit: 1.982383, pval: 0.4518747

Conclusion: We do not have enough evidence to suggest that the variable Age is significant to the model. Therefore, we should drop the variable Age from the model.

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

```r
## R-sq = SSR/SST or 1-(SSres/SST)
SST = 57.305+33.397+ 0.136+5.101+ 0.028+105.413
SSres = 105.413
1-(SSres/SST)
```

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

The R-sq is 0.4765468, suggesting that only 47.65% of the variability in the percentage of patients who get an infection while hospitalized by the model with Stay, Cultures, Age, Census, and Beds.

(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.

```r
## F-stat = ([SSR(F)-SSR(R)]/r)/(SSres(F)/(n-p))
Fstat = ((0.136+5.101+ 0.028)/3)/( 105.413/107)
Fstat
```

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

```r
## F-crit = qf(1-alpha, r, n-p)
qf(0.95, 3, 107)
```

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

```r
## pval = 1-pf(Fstat, r, n-p)
1-pf(Fstat, 3, 107)
```

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

H0: (x3, x4, x5)==0, the predictors can be dropped from the model

HA: at least one (x3, x4, x5)=/=0, the predictors can not be dropped from the model

F-stat: 1.781422, F-crit: 2.68949, pval: 0.1550925

Conclusion: We do not have enough evidence to suggest that at least one of the predictors (x3, x4, x5) is significant to keep in the model. Therefore, we can drop the predictors (x3, x4, x5) and use model 2.

(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.

```r
## F-stat = ([SSR(F)-SSR(R)]/r)/(SSres(F)/(n-p))
Fstat2 = ((0.136+5.101)/2)/((0.028+105.413)/108)
Fstat2
```

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

```r
## F-crit = qf(1-alpha, r, n-p)
qf(0.95, 2, 108)
```

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

```r
## pval = 1-pf(Fstat, r, n-p)
1-pf(Fstat2, 2, 108)
```

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

H0: (x3, x4)==0, the predictors can be dropped from the model

HA: at least one (x3, x4)=/=0, the predictors can not be dropped from the model

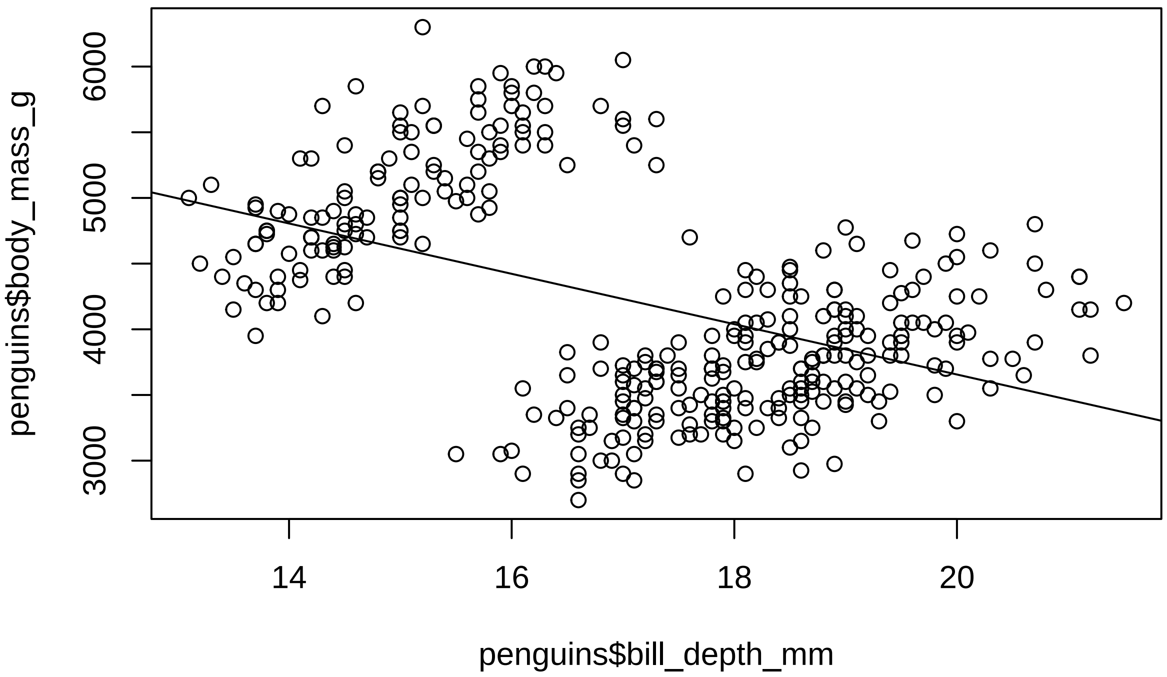
F-stat: 2.68205, F-crit: 3.080387, pval: 0.07297994

Conclusion: We do not have enough evidence to suggest that at least one of the predictors (x3, x4) is significant to keep in the model. Therefore, we can drop the predictors (x3, x4) and use model 2.

Question 6

(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?

```
plot(penguins$bill_depth_mm, penguins$body_mass_g)
abline(lm(formula=body_mass_g~bill_depth_mm, data=penguins))
```

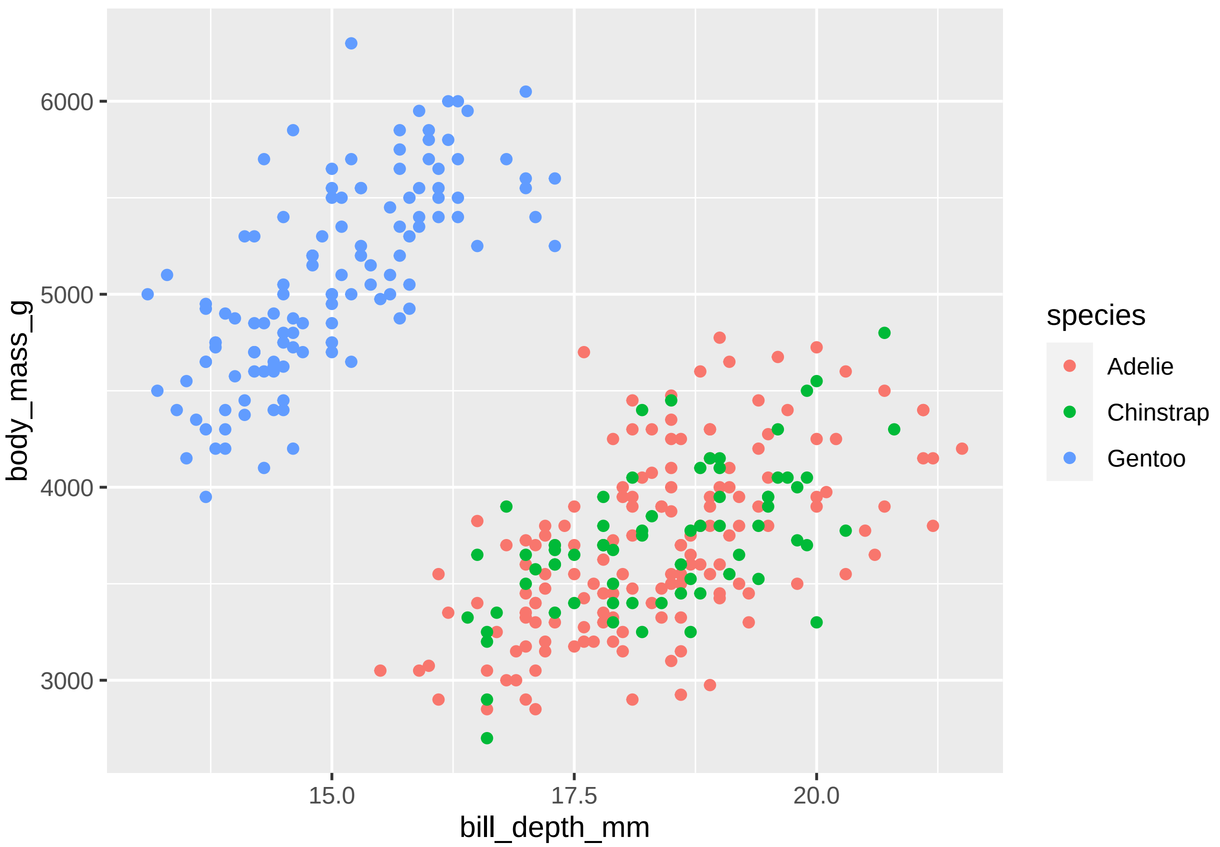


It looks like there are 2 different groupings that have positive, linear relationships with body mass. Without grouping in mind, it does look more random though.

(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)+
  geom_point(aes(x=bill_depth_mm, y=body_mass_g, color=species))
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



All three species of pengiuns have a positive, linear relationship with body mass. Adelie and Chinstrp penguins seem to have similar relationship between bill depth and body mass compared to Gentoo penguins.

(c) Create a regression with interaction between bill depth and species 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.

```
newpen = penguins%>%
  mutate(I1 = ifelse(species=="Chinstrap", 1, 0))%>%
  mutate(I2 = ifelse(species=="Gentoo", 1, 0))
regpen = lm(body_mass_g~bill_depth_mm+I1+I2+bill_depth_mm*I1+bill_depth_mm*I2, data=newpen)
summary(regpen)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_depth_mm + I1 + I2 + bill_depth_mm *
##     I1 + bill_depth_mm * I2, data = newpen)
##
## 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 ***
## I1                  247.06     829.77   0.298   0.7661
## I2                 -175.71     658.43  -0.267   0.7897
## bill_depth_mm:I1    -12.53      45.01  -0.278   0.7809
## bill_depth_mm:I2    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
```

The regression equation is as follows:

body_mass_g = -283.28 + 217.15bill_depth_mm + 247.06I1 - 175.71I2 - 12.53bill_depth_mm:I1 + 152.29bill_depth_mm:I2

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

```
redpen = lm(body_mass_g~bill_depth_mm+I1+I2, data=newpen)
anova(redpen, regpen)
```

```
## Analysis of Variance Table
##
## Model 1: body_mass_g ~ bill_depth_mm + I1 + I2
## Model 2: body_mass_g ~ bill_depth_mm + I1 + I2 + bill_depth_mm * I1 +
##     bill_depth_mm * I2
##   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
```

```
qf(0.95,2,336)
```

```
## [1] 3.022601
```

H0: beta-hat4,5==0, we can drop the interaction terms

HA: at least one beta-hat4,5 =/=0, we cannot drop the interaction terms

F-stat: 8.2342, F-crit: 3.022601, pval: 0.0003227 ==> reject H0

Conclusion: There is enough evidence to suggest that at least one of the interaction terms is significant. Therefore, we should keep the interaction terms in the model.

(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.

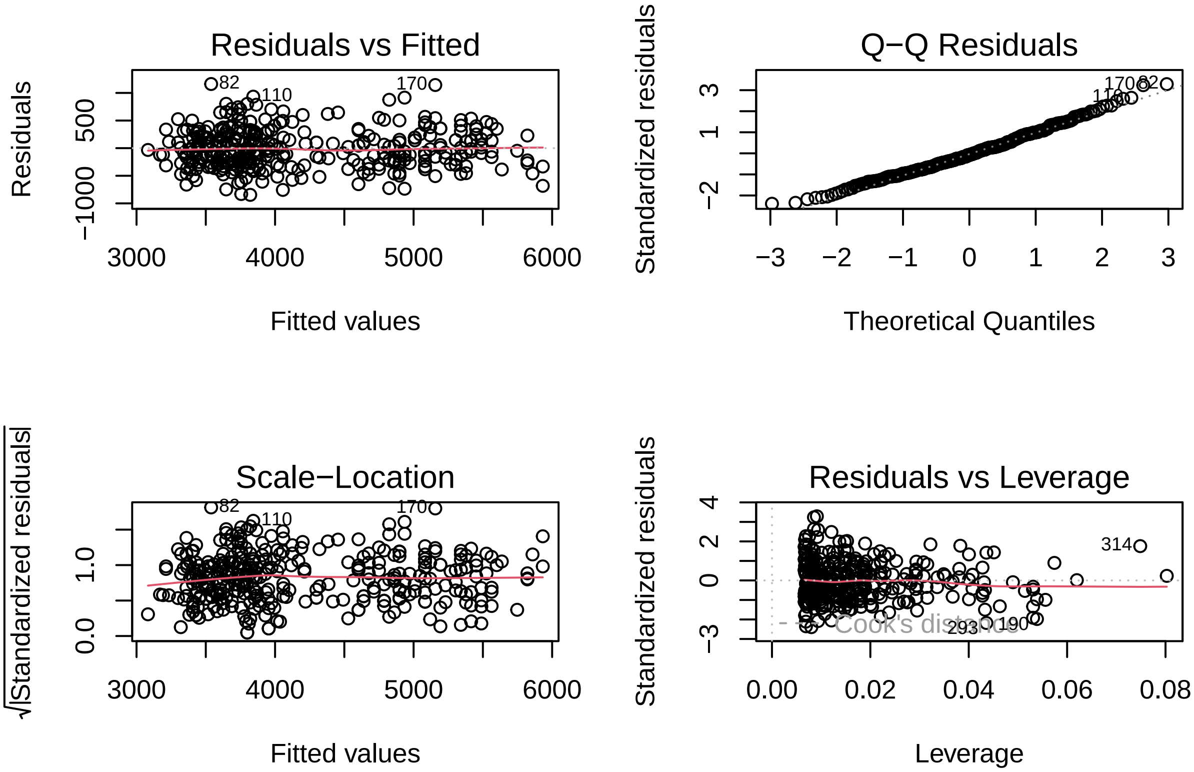
Regression equation (Chinstrap): I1=1, I2=0 ==> -36.22 + 204.62bill_depth_mm

Regression equation (Gentoo): I1=0, I2=1 ==> -458.99 + 369.44bill_depth_mm

Regression equation (Adelie): I1=0, I2=0 ==> -283.28 + 217.15bill_depth_mm

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

```r
par(mfrow=c(2,2))
plot(regpen)
```



I would say that these regression assumptions are met based on the residual and QQ plots above. The residuals are evenly spread along the horizontal band and the variance seems stable. In addition, the QQ plot shows that the data is normally distributed.

```r
levene.test(newpen$body_mass_g, newpen$species)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  newpen$body_mass_g
## Test Statistic = 2.8123, p-value = 0.09468
```

Using the Levene's test, we see that we do not have enough evidence to say that the variance is not the same across all species. Therefore, we fail to reject the H0 that states that the variances are equal across species.

(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 perform a Tukey's multiple comparisons test because the slopes for each species for the regression of body mass against bill depth are different. As seen in the scatterplot in part b, the steepness of the slopes are different, meaning that the interactions may be significant. This is also supported by the partial F test done that checked the significance between species and bill depth.

Question 7

```r
q7 = data.frame("Region"=c("North", "South", "West"),
                "n"=c(21, 17, 13),
                "mean PAY"=c(24424, 22894, 26159),
                "mean SPEND"=c(3901, 3274, 3919))
q7
```

```
##   Region  n mean.PAY mean.SPEND
## 1  North 21    24424       3901
## 2  South 17    22894       3274
## 3   West 13    26159       3919
```

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

The the order South-North-West, the mean teacher pay seems to increase by $2000 each time. However, the South seem to spend about $7000 per student less than both North and West. This is important because of the mean SPEND was to follow the same trend as the mean PAY, the difference in mean SPEND between South-North, North-West would be more equal. This suggests that there might be an interaction between region and mean SPEND.

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

The mean SPEND seems to increase as mean PAY increases.

(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 MLR model will help us determine if region influences mean PAY while holding mean SPEND constant, or if mean SPEND influences mean PAY while holding region constant, or if both mean SPEND and region are associated with mean PAY.

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 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.

(d) What is the estimate of beta2? Give an interpretation of this value.

For teachers employed in the South, the estimated annual salary increases by 529.4 USD while holding all other predictors constant.

(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.

```
## bon CI = beta-hatj +/- t\astse(beta-hatj) where t\ast is 1-(alpha/2g), n-p
## and g is the number of intervals
## Ref Class: North

## CI for beta2 (North vs South)
529.4-qt(1-(0.05/6), 51-4)*766.9
```

```
## [1] -1374.578
```

```
529.4+qt(1-(0.05/6), 51-4)*766.9
```

```
## [1] 2433.378
```

```
## CI for beta3 (North vs West)
1.674e+03-qt(1-(0.05/6), 51-4)*8.012e+02
```

```
## [1] -315.1348
```

```
1.674e+03+qt(1-(0.05/6), 51-4)*8.012e+02
```

```
## [1] 3663.135
```

```
## se(beta2-beta3) = sqrt(var(beta2)+var(beta3)-2cov(beta2,beta3))
seb2b3 = sqrt(588126.71689+ 6.418738e+05-(2*244238.02959))
## CI for beta2-beta3 (South vs West)
(529.4-1.674e+03)-qt(1-(0.05/6), 51-4)*seb2b3
```

```
## [1] -3282.493
```

```
(529.4-1.674e+03)+qt(1-(0.05/6), 51-4)*seb2b3
```

```
## [1] 993.2933
```

North region and the South region:(-1374.578, 2433.378)

North region and the West region: (-315.1348, 3663.135)

South region and the West region: (-3282.493, 993.2933)

(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)?

Since 0 is within all three confidence intervals, that means that the region has no significant effect on the mean PAY while holding mean SPEND constant.