Advanced Statistics Using R - CA1

Robert Dowd

2023-02-26

Data Description

Sex: 1 for male and 2 for female.
Bwt: Body weight in kg.
Hwt: Heart weight in g.
Height: Height in cm.
Age: Age in years.
Outdoor: Cat kept outdoors; 1 = Always, 2 = Frequently, 3 = Never.

```
cat_data <- read.csv("~/Desktop/DataAnalytics2022/AdvancedStatisticsUsingR/CA1/cat_
hwt.csv")
head(cat_data)
```

```
##   Sex Bwt Hwt Height  Age Outdoor
## 1   2 2.0 7.0   20.7 11.5       1
## 2   2 2.0 7.4   24.0  5.9       1
## 3   2 2.0 9.5   24.4 19.1       1
## 4   2 2.1 7.2   14.8 10.6       2
## 5   2 2.1 7.3   25.9  9.5       1
## 6   2 2.1 7.6   22.8  9.5       1
```

```
str(cat_data)
```

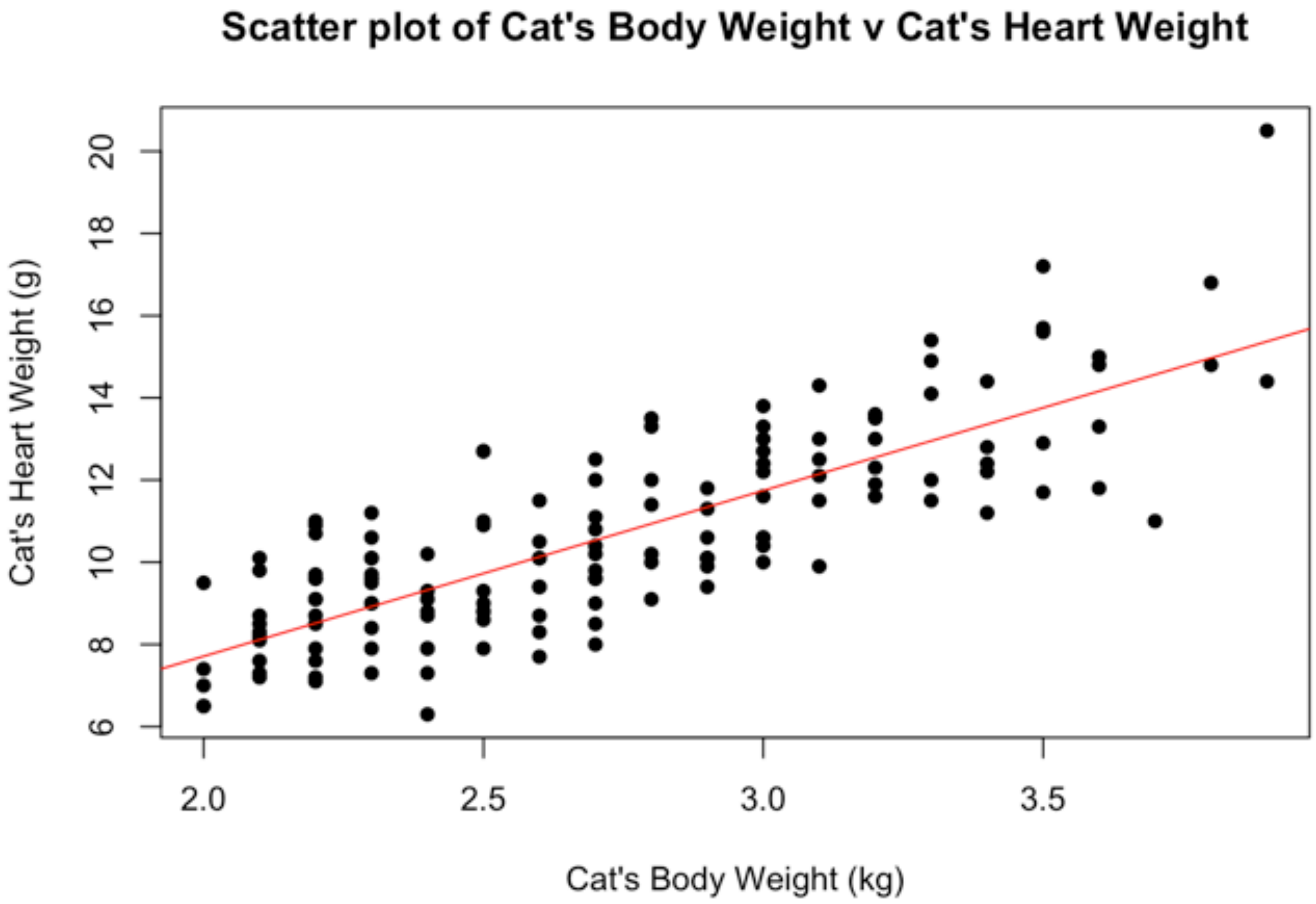
```
## 'data.frame':    144 obs. of  6 variables:
##  $ Sex    : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ Bwt    : num  2 2 2 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
##  $ Hwt    : num  7 7.4 9.5 7.2 7.3 7.6 8.1 8.2 8.3 8.5 ...
##  $ Height : num  20.7 24 24.4 14.8 25.9 22.8 15.5 19.4 16.4 17.8 ...
##  $ Age    : num  11.5 5.9 19.1 10.6 9.5 9.5 8.6 14.7 7.8 7.3 ...
##  $ Outdoor: int  1 1 1 2 1 1 1 2 1 1 ...
```

```
attach(cat_data)
```

1. Relationships

a) Investigate if any of the continuous numerical variables have a linear relationship by producing scatterplots, interpret these plots.

```
#Scatter plot of Cat's Body Weight v Cat's Heart Weight
plot(Bwt,Hwt, xlab="Cat's Body Weight (kg)", ylab="Cat's Heart Weight (g)", main="S
catter plot of Cat's Body Weight v Cat's Heart Weight", pch=16)
#Plot regression line
abline(lm(Hwt~Bwt,data=cat_data),col='red')
```

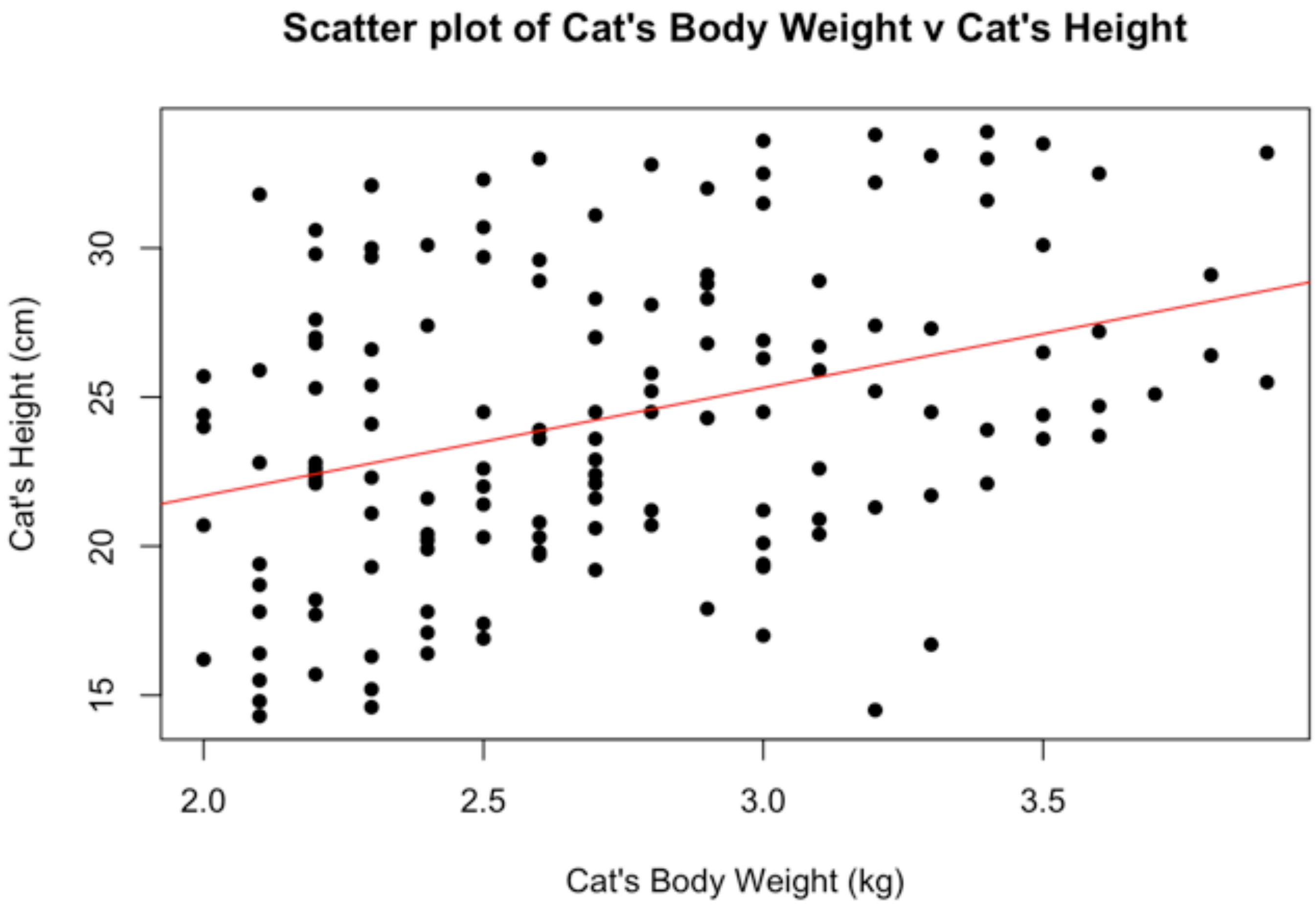


From this scatterplot we can see the following relationship between Cat’s Body Weight v Cat’s Heart Weight:

The data points don’t stray to far away from the regression line.

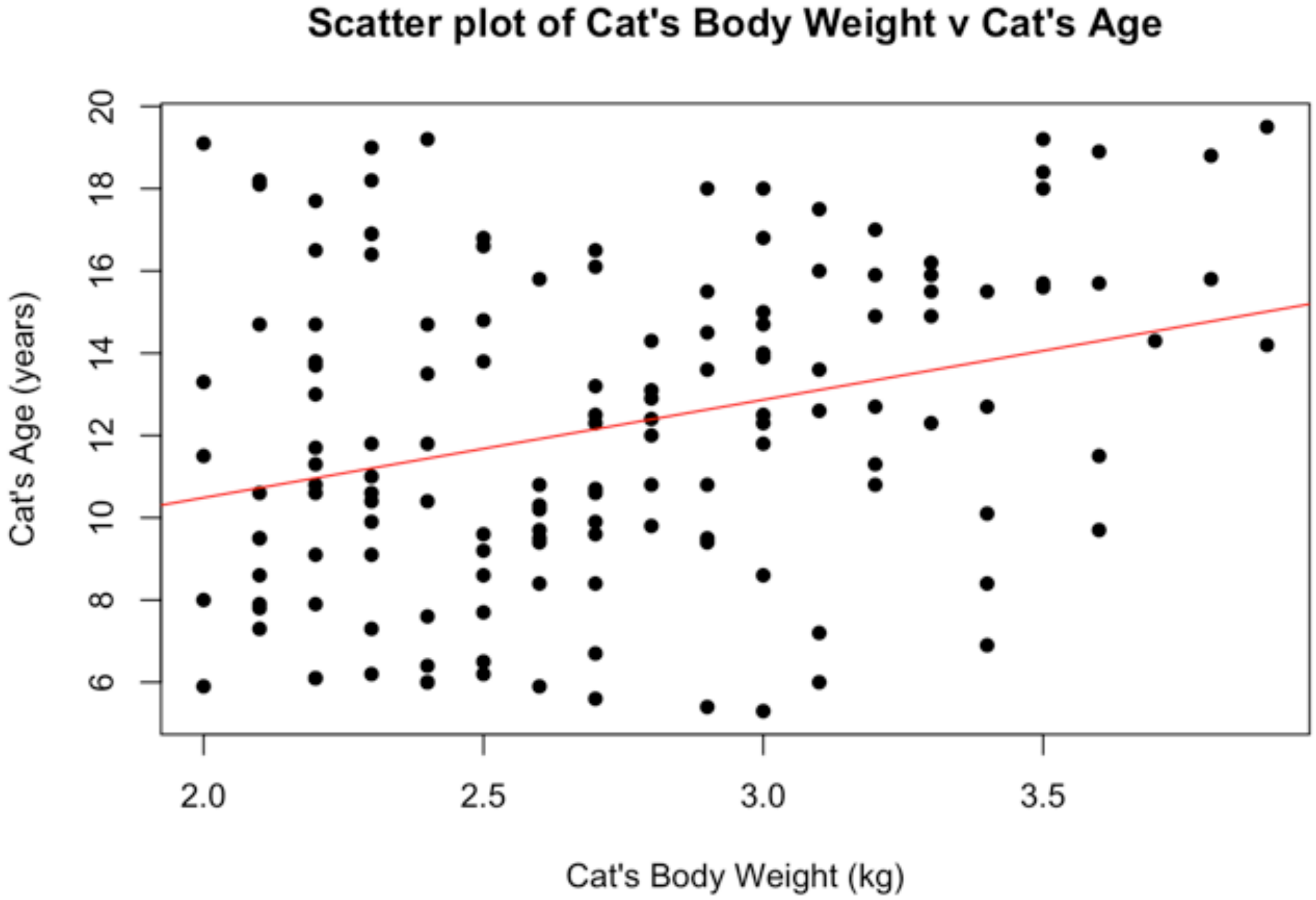
Cat’s Body Weight v Cat’s Heart Weight appear to have a very strong positive association because in general Cat’s Body Weight increases when Cat’s Heart Weight increases.

```
#Scatter plot of Cat's Body Weight v Cat's Height
plot(Bwt,Height, xlab="Cat's Body Weight (kg)", ylab="Cat's Height (cm)", main="Sca
tter plot of Cat's Body Weight v Cat's Height", pch=16)
#Plot regression line
abline(lm(Height~Bwt,data=cat_data),col='red')
```



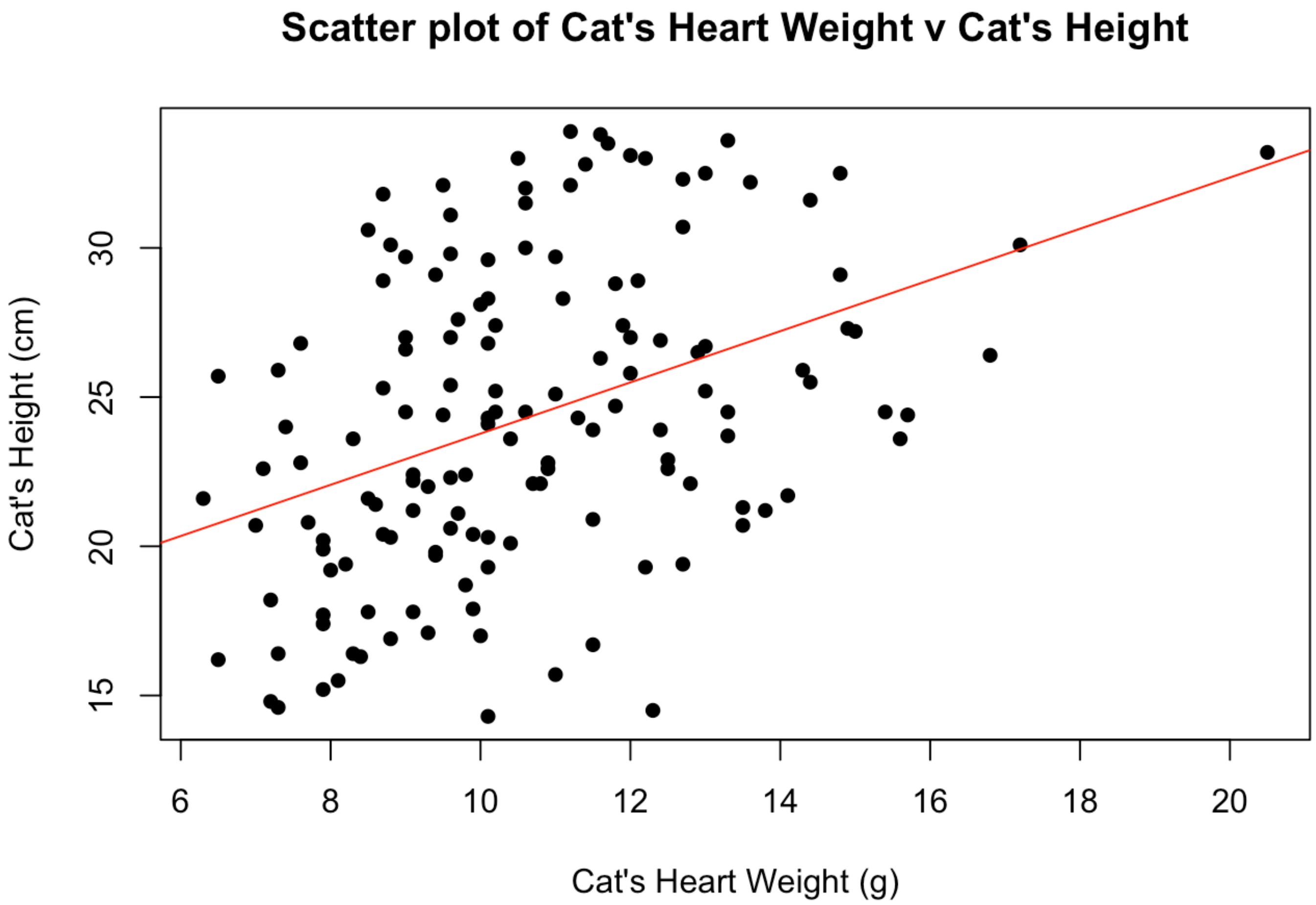
From this scatterplot we can see the following relationship between Cat's Body Weight v Cat's Height:
The data points are very scattered and dispersed but with the regression line we can interpret this scatter plot better.
The data points are greatly dispersed away from the away from the regression line.
Cat's Body Weight v Cat's Height appear to have a weak positive association because in general Cat's Body Weight slightly increases when Cat's Height slightly increases.

```
#Scatter plot of Cat's Body Weight v Cat's Age
plot(Bwt,Age, xlab="Cat's Body Weight (kg)", ylab="Cat's Age (years)", main="Scatter plot of Cat's Body Weight v Cat's Age", pch=16)
#Plot regression line
abline(lm(Age~Bwt,data=cat_data),col='red')
```



From this scatter plot we can see the following relationship between Cat’s Body Weight v Cat’s Age:
The data points are greatly dispersed away from the away from the regression line.
Cat’s Body Weight v Cat’s Age appear to have a weak positive association because in general Cat’s Body Weight slightly increases when Cat’s Age slightly increases.

```
#Scatter plot of Cat's Heart Weight v Cat's Height
plot(Hwt,Height, xlab="Cat's Heart Weight (g)", ylab="Cat's Height (cm)", main="Sca
tter plot of Cat's Heart Weight v Cat's Height", pch=16)
#Plot regression line
abline(lm(Height~Hwt,data=cat_data),col='red')
```

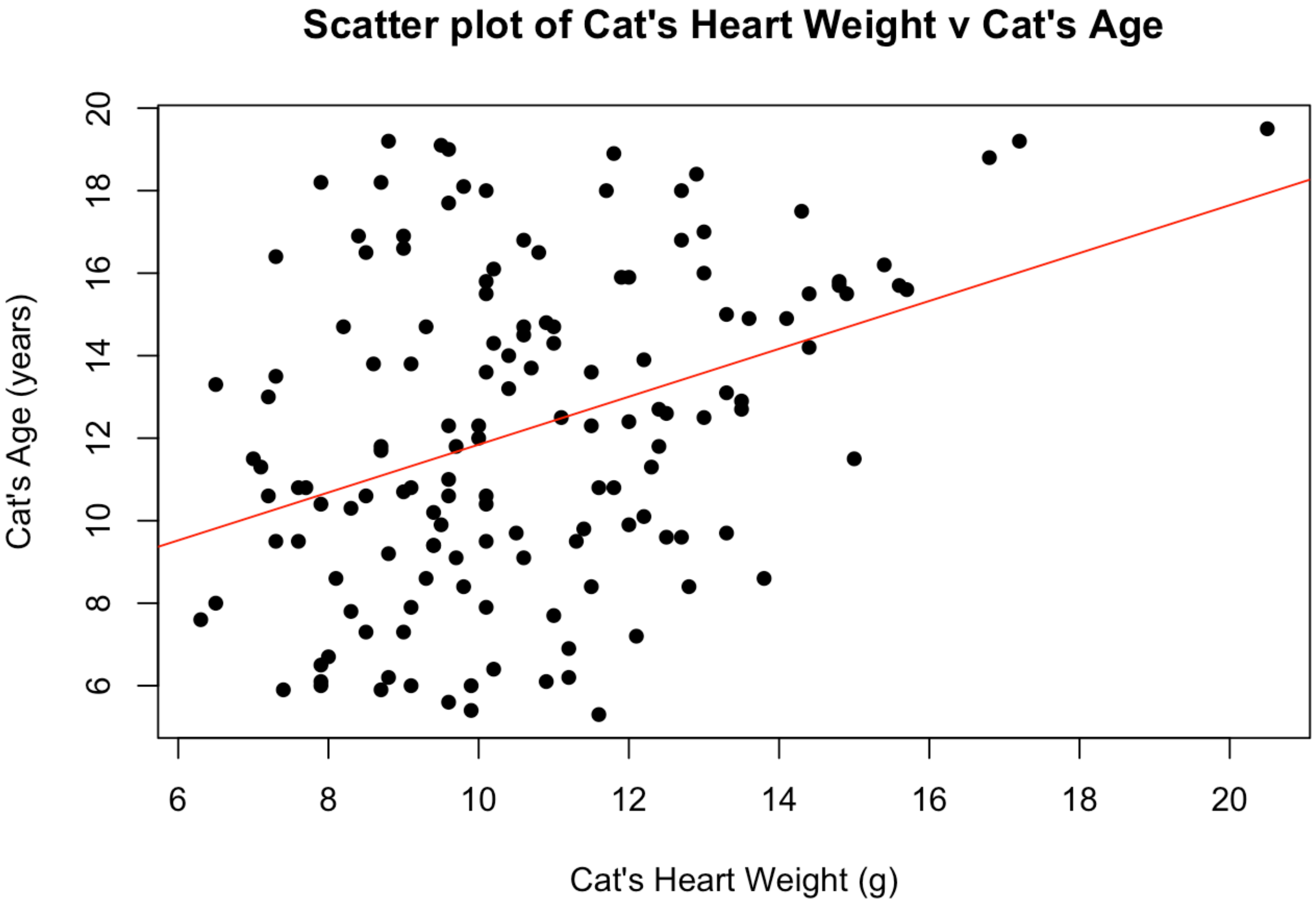


From this scatter plot we can see the following relationship between Cat’s Heart Weight v Cat’s Height:

The data points are moderately dispersed away from the away from the regression line.

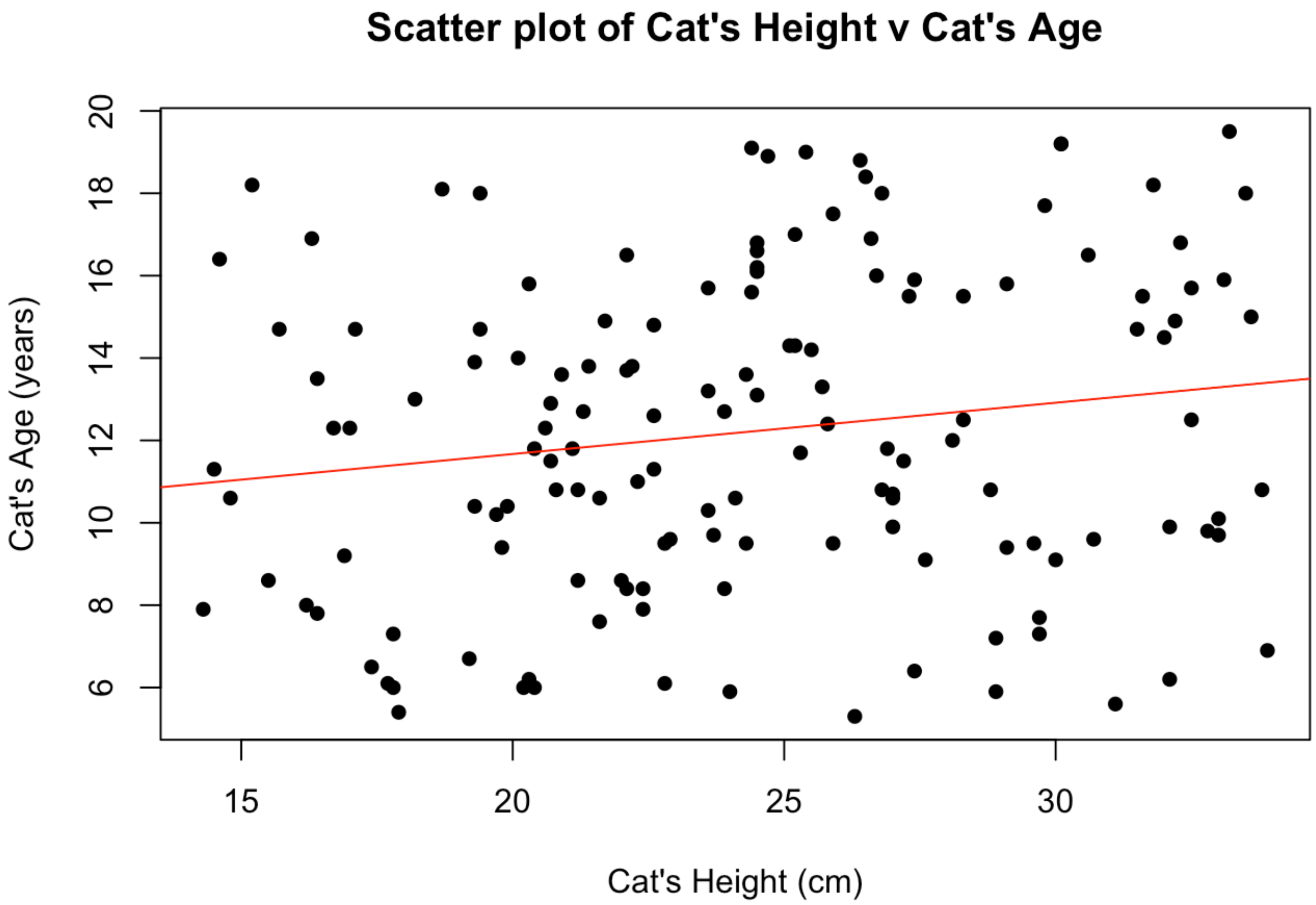
Cat’s Heart Weight v Cat’s Height appear to have a medium positive association because in general Cat’s Heart Weight increases when Cat’s Height increases.

```
#Scatter plot of Cat's Heart Weight v Cat's Age
plot(Hwt,Age, xlab="Cat's Heart Weight (g)", ylab="Cat's Age (years)", main="Scatte
r plot of Cat's Heart Weight v Cat's Age", pch=16)
#Plot regression line
abline(lm(Age~Hwt,data=cat_data),col='red')
```



From this scatter plot we can see the following relationship between Cat’s Heart Weight v Cat’s Age:
The data points are dispersed away from the away from the regression line.
Cat’s Heart Weight v Cat’s Age appear to have a very weak association because in general Cat’s Heart Weight increases when Cat’s Age increases.

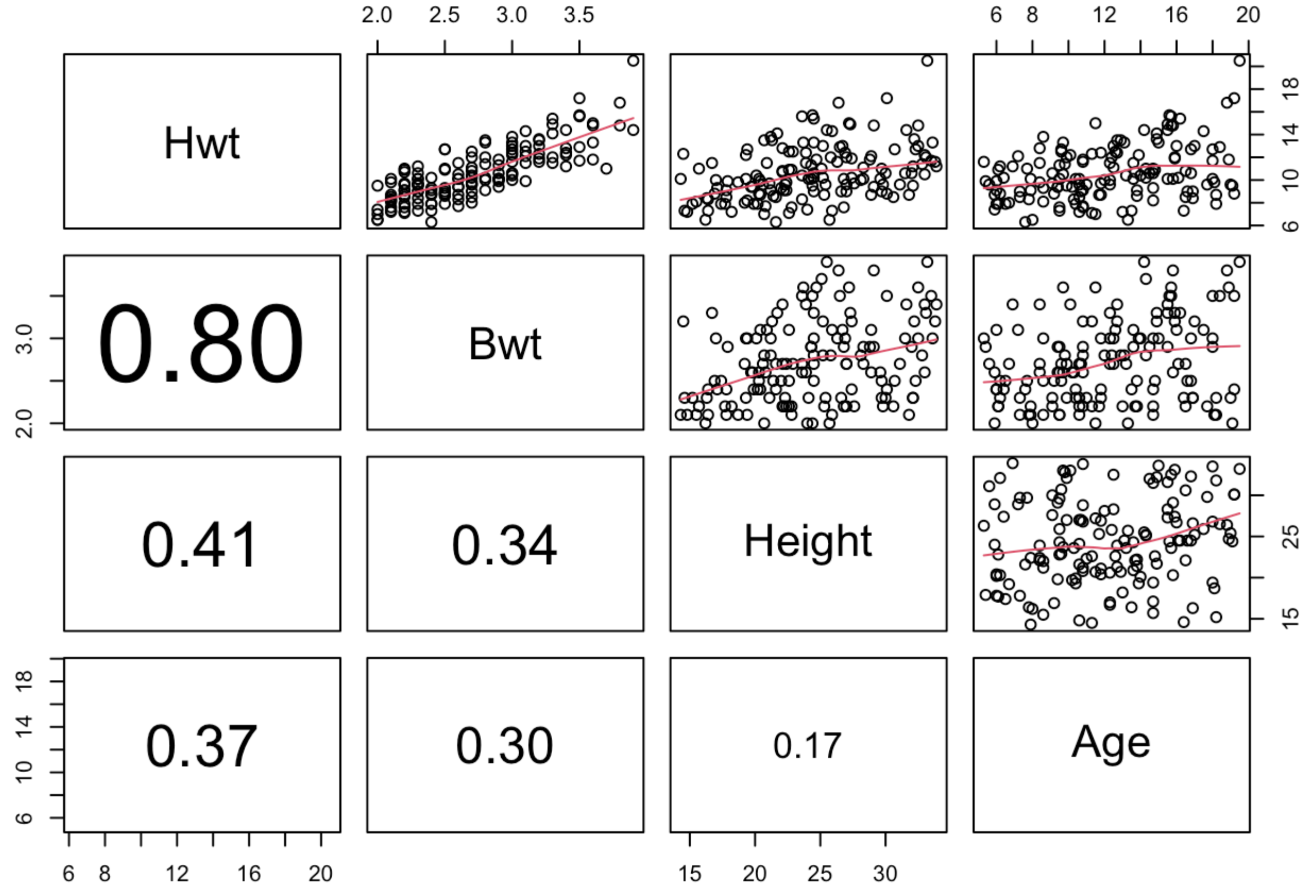
```
#Scatter plot of Cat's Height v Cat's Age
plot(Height,Age, xlab="Cat's Height (cm)", ylab="Cat's Age (years)", main="Scatter
plot of Cat's Height v Cat's Age", pch=16)
#Plot regression line
abline(lm(Age~Height,data=cat_data),col='red')
```



From this scatter plot we can see the following relationship between Cat’s Height v Cat’s Age:8
The data points are greatly dispersed away from the away from the regression line.
Cat’s Height v Cat’s Age appear to have a very weak positive association because in general Cat’s Height slightly increases when Cat’s Age slightly increases.

```
panel.cor <- function(x, y, digits = 2, prefix = "",
                      cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- cor(x, y, use="complete.obs")
  txt <- format(c(r, 0.123456789),
                digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * (abs(r)+0.15))
}
pairs(~Hwt+Bwt+Height+Age,cat_data, upper.panel = panel.smooth, lower.panel= panel.
cor)
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```



Looking at this plot, there are many linear and non-linear relationships:
Hwt & Bwt has a strong positive linear relationship.
Hwt & Height has a weak positive linear relationship.
Hwt & Age has a very weak positive linear relationship.
Bwt & Height has a weak positive linear relationship.
Bwt & Age has a weak positive linear relationship.
Height & Age has a weak positive curved relationship.

b) What is the response variable and why? What is the research question of interest in this dataset?

Response variable is “Hwt” heart weight in grams because this is something that can’t be measured easily.
Research question:
What is the predicted heart weight of a cat?
We are interested in exploring the variables to see if they have a significant impact on a cat’s heart weight and can we use these variables to predict a cat’s heart weight.

c) Investigate if any of the categorical /discrete variables seem to have a relationship with the response variable (selected in part b) using boxplots. Interpret these plots.

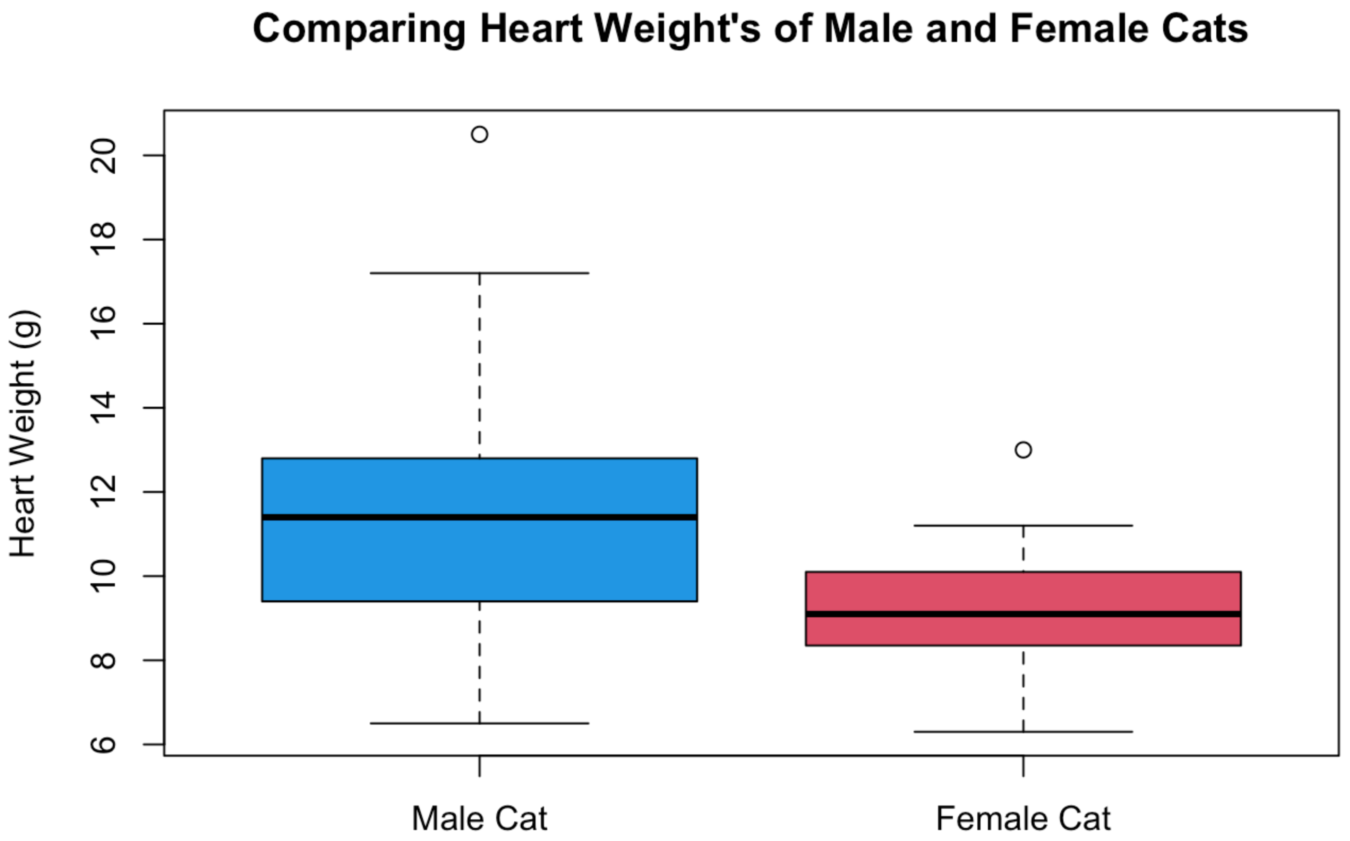
```
#Boxplot of Cat's Heart Weight by Sex
MaleHwt <- Hwt[Sex==1]
summary(MaleHwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.50    9.40   11.40   11.32   12.80   20.50 
```

```
FemaleHwt <- Hwt[Sex==2]
summary(FemaleHwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   6.300   8.350   9.100   9.202  10.100  13.000 
```

```
boxplot(MaleHwt, FemaleHwt, names = c("Male Cat", "Female Cat"), main="Comparing Heart Weight's of Male and Female Cats", ylab="Heart Weight (g)", col=c(4,2))
```



From this box plot of Cat’s Heart Weight by Cat’s Sex, we can see the following information:
Male cats tend to have a higher heart weight than female cats.
As male cat’s heart weight has a higher median and IQR than female cats.
There is a wider spread for male cat’s heart weight than female cats.
There s an outlier present in both male and female groups.

```
#Boxplot of Cat's Heart Weight by Cat's Outdoor Categories
AlwaysHwt <- Hwt[Outdoor==1]
summary(AlwaysHwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   6.500   7.750   8.700   8.767   9.600  11.600
```

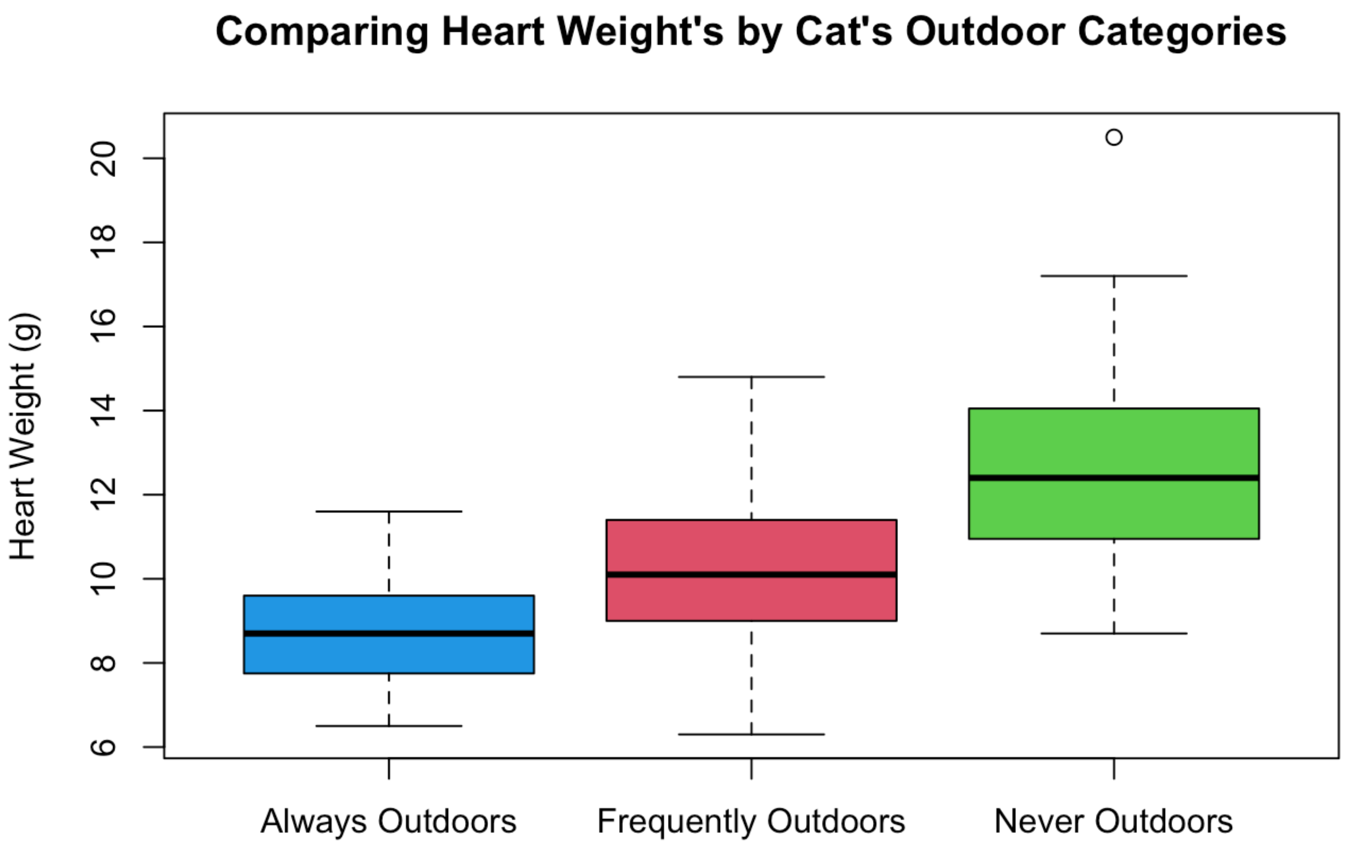
```
FreqHwt <- Hwt[Outdoor==2]
summary(FreqHwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.30    9.00   10.10   10.23   11.40   14.80
```

```
NeverHwt <- Hwt[Outdoor==3]
summary(NeverHwt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    8.70   10.97   12.40   12.62   13.93   20.50 
```

```
boxplot(AlwaysHwt, FreqHwt, NeverHwt, names = c("Always Outdoors", "Frequently Outd
oors", "Never Outdoors"), main="Comparing Heart Weight's by Cat's Outdoor Categorie
s", ylab="Heart Weight (g)", col=c(4,2,3))
```



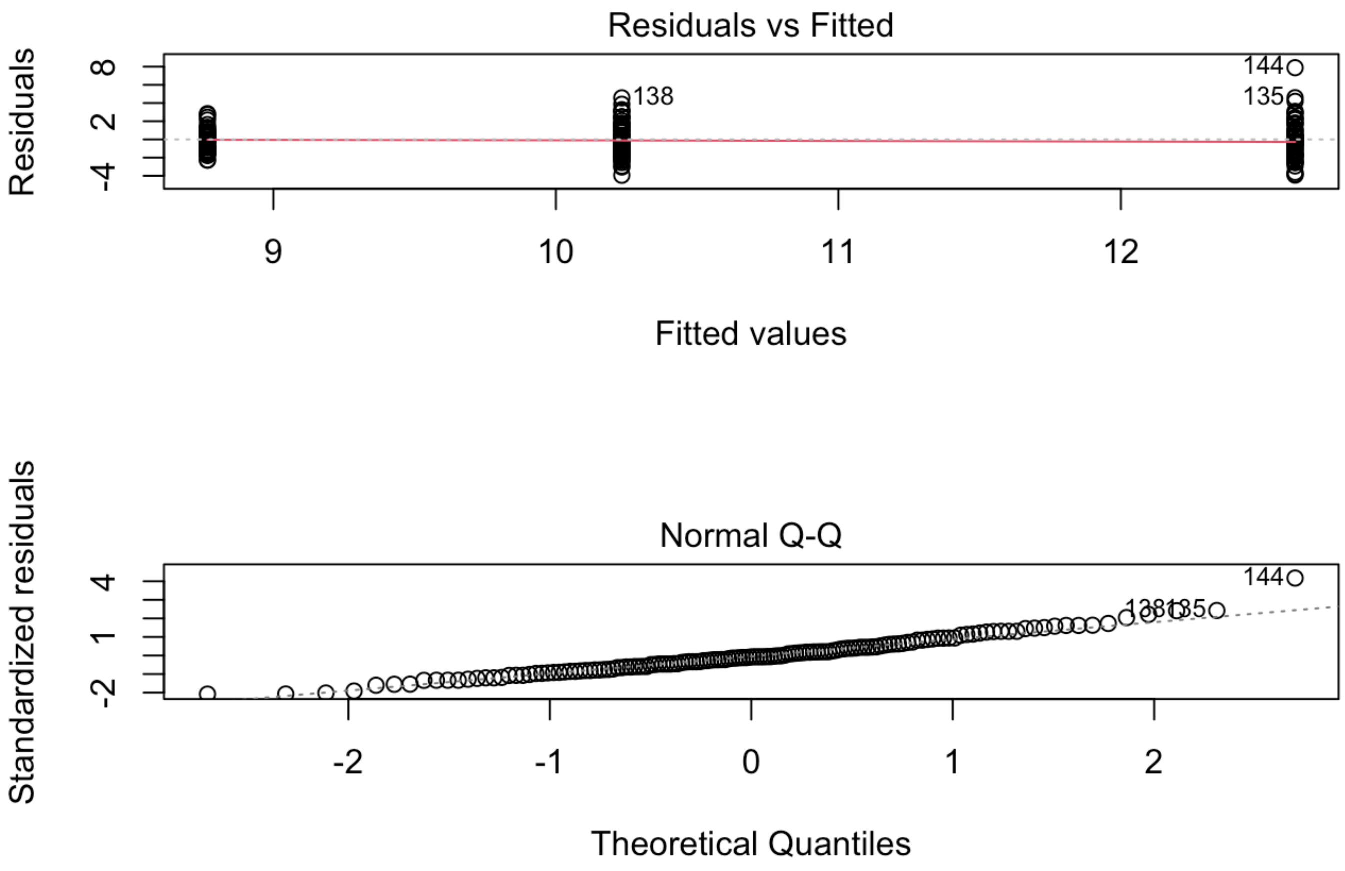
From this box plot of Cat’s Heart Weight by Cat’s Outdoor Categories, we can see the following information:
Cat’s that are kept always outdoors tend to have a lower heart weight.
As cat’s heart weight has a higher median and IQR if the cat was never kept outdoors.
The spreads are similar across all boxplots and median roughly in middle.
There is one outlier present in cat’s that are never kept outdoors.

d) Using part c), choose one categorical /discrete variable with at least 3 categories to test to see if there is any difference between the means of the response variable (selected in part b) using an one-way ANOVA test. (Note: if categorical variables are not set up as factors, so change to factor prior to running analysis (as.factor())

Selecting “Outdoor” variable to run a one way anova on “Hwt” heart weight.
As seen from the boxplot above there seems to be a difference in the means but also the variability between groups seems similar so it is appropriate to run the one way anova.

Null hypothesis is that there is no difference in the true mean of “Hwt” cat’s heart weight for each “Outdoor” group.
H0: mu_1=mu_2=mu_3, where mu_i is the true mean of “Hwt” cat’s heart weight for group i.
Alternative hypothesis is that at least one “Hwt” cat’s heart weight mean is different for each “Outdoor” group.
H1: not all mu_i are the same, where mu_i is the true mean of “Hwt” Cat’s heart weight for group i.

```
res1 <- aov(Hwt~as.factor(Outdoor), data = cat_data)
par(mfrow=c(2,1))
plot(res1, which=1:2)
```



Interpret Residuals vs Fitted:
The spread of the three groups looks roughly the same so we are happy with our equal variance assumption.
The scatter seems randomly distributed within in the 3 groups, so we are happy with our iid assumption.
Interpret Normal Q-Q Plot:
The points fall roughly in a straight line indicating that the residuals are roughly normally distributed.

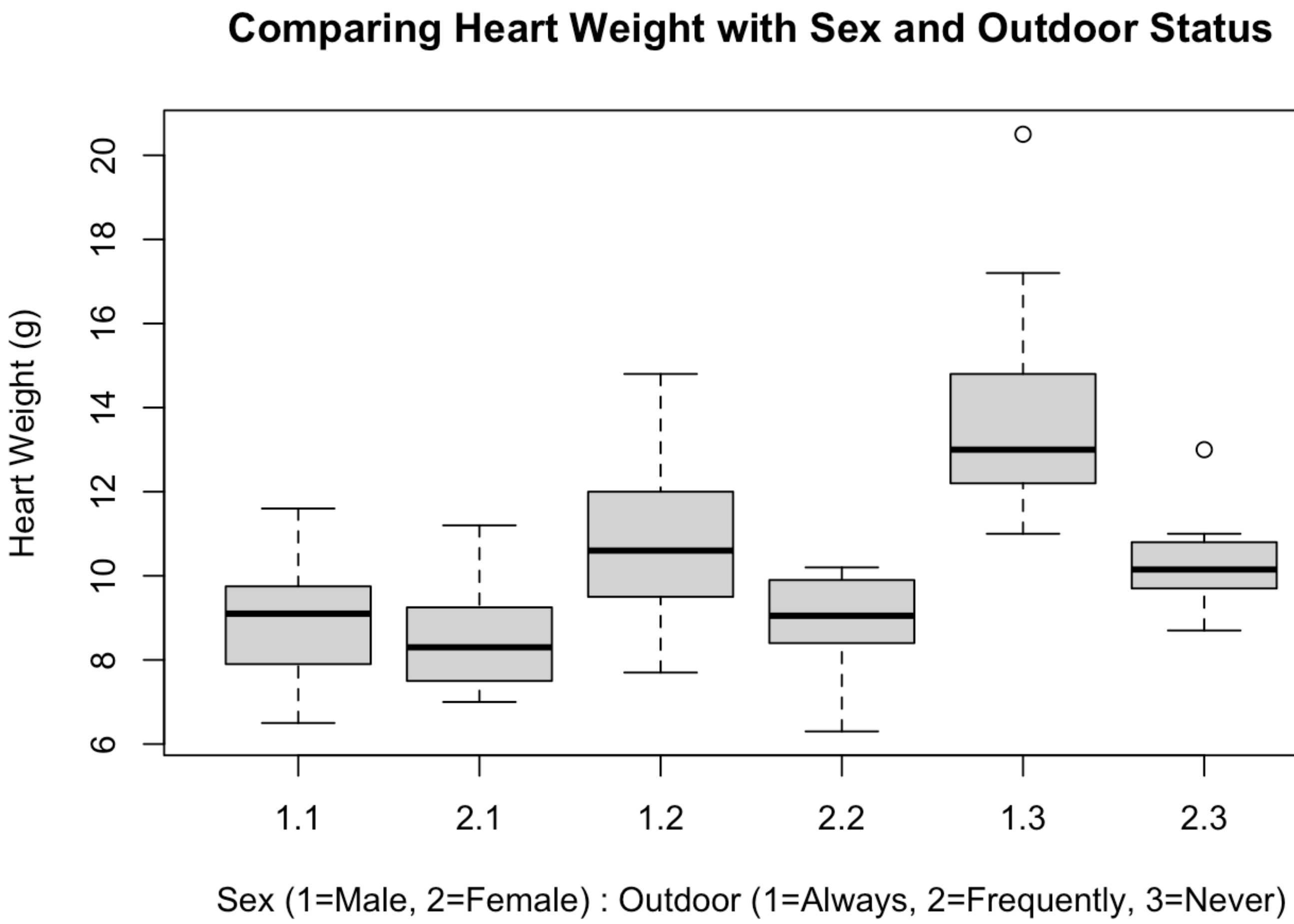
```
summary(res1)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)    
## as.factor(Outdoor)   2  333.8  166.91    45.8 4.71e-16 ***
## Residuals          141  513.8    3.64                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value = 4.71e-16 which is less than 0.05, so we reject the null hypothesis.
We can conclude that at least one of the Outdoor" groups has a different true mean for "Hwt" Cat's heart weight from the others.
i.e. there is a significant relationship between "Outdoor" and Hwt" Cat's heart weight.

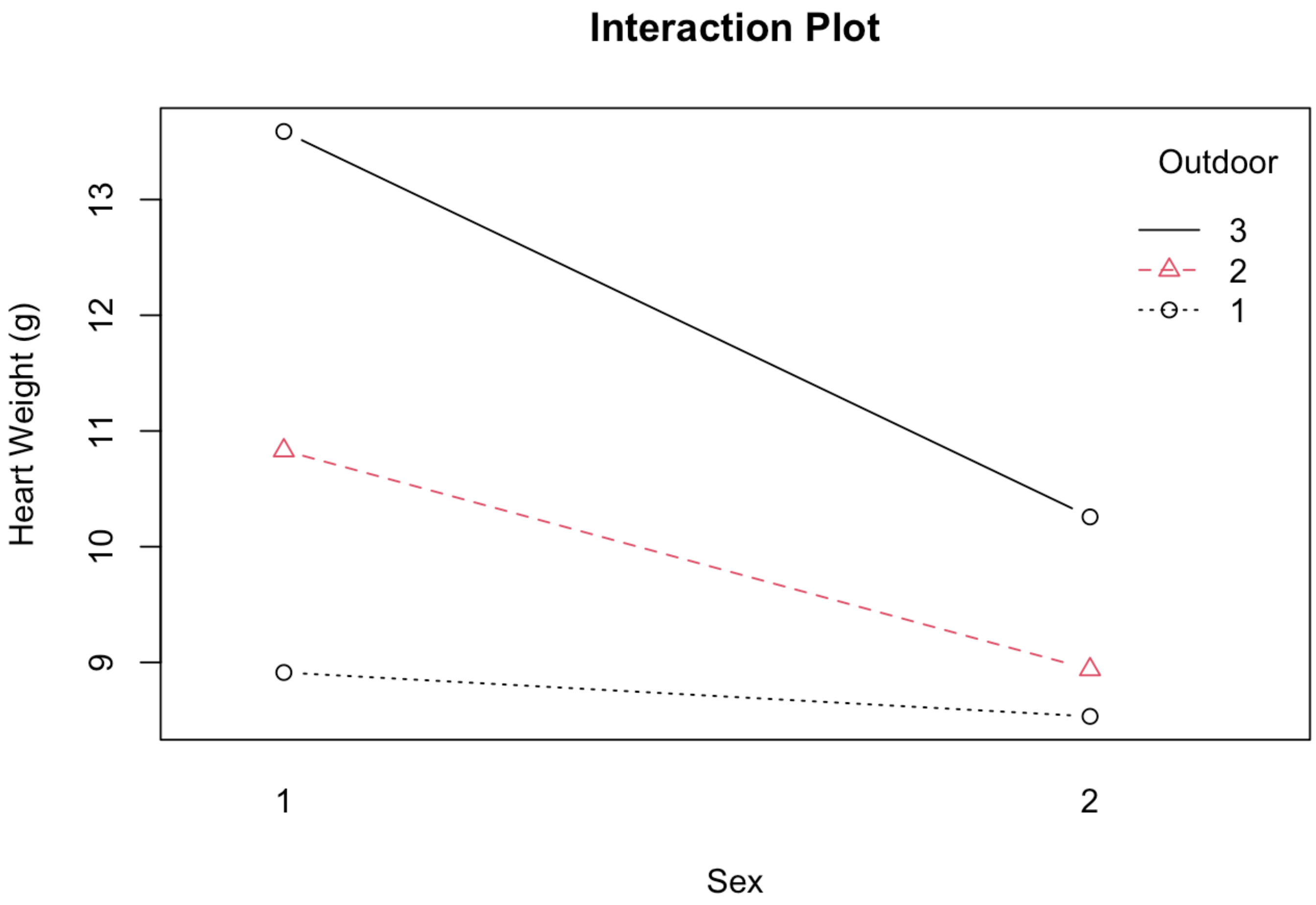
e) Create a boxplot to see if there is any difference between the means of the response variable (selected in part b) across the two variables Sex and Outdoor. Also, create a plot to see if there is an interaction effect for these two variables with the response variable. Interpret these plots. Test to see if any of these factors and/or interactions have a significant relationship with the response variable (selected in part b) using an two-way ANOVA test. (Note: categorical variables are not set up as factors, so change to factor prior to running analysis (as.factor())

```
boxplot(Hwt~Sex+Outdoor,data=cat_data, main="Comparing Heart Weight with Sex and Ou
tdoor Status", xlab="Sex (1=Male, 2=Female) : Outdoor (1=Always, 2=Frequently, 3=Ne
ver)", ylab="Heart Weight (g)")
```



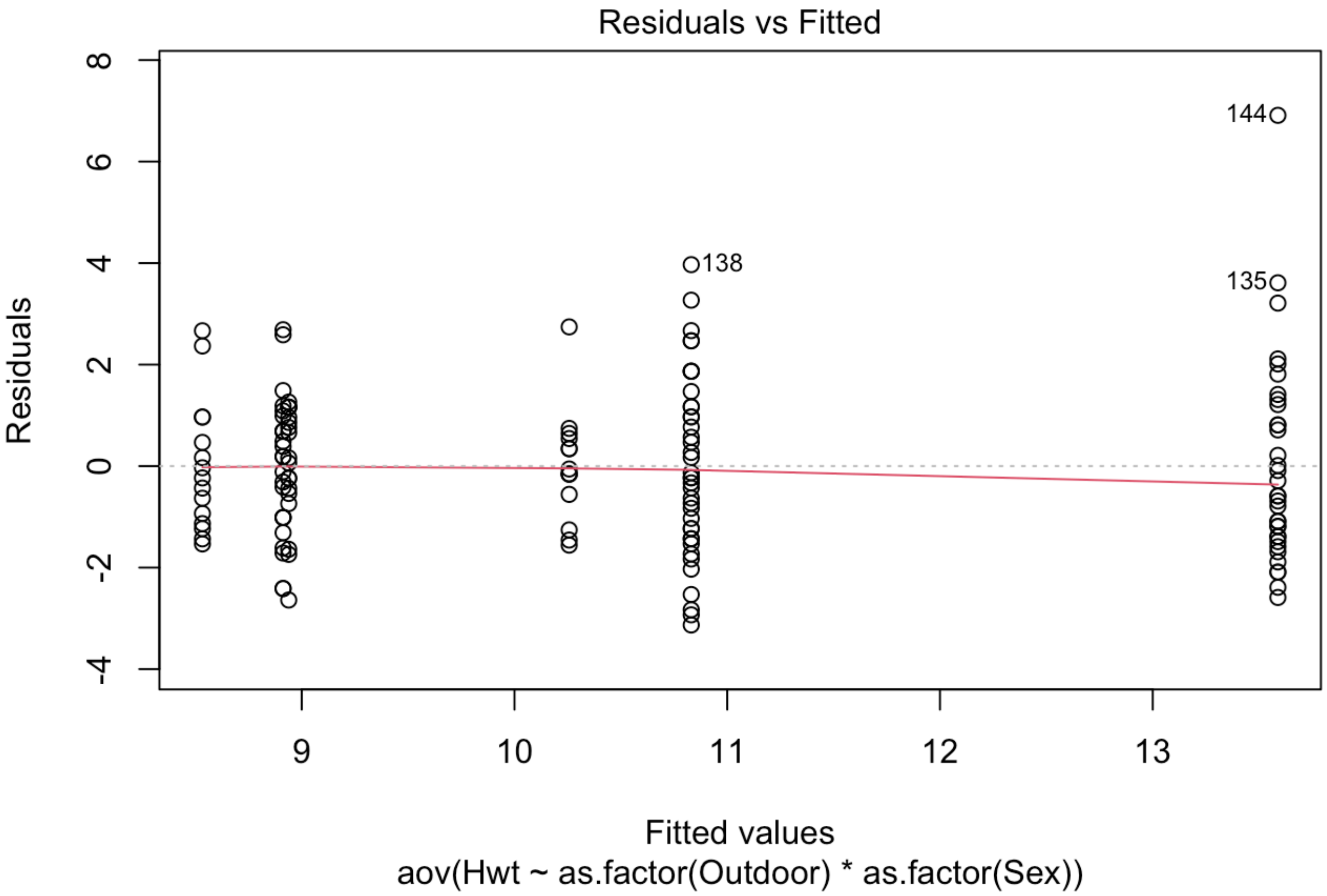
We can see that male cat’s heart weight has a higher median and IQR for all outdoor categories.
We can see that the median heart weight tends to be lower for cats that are kept outdoors.
The spreads are similar across all boxplots and median roughly in middle.
There is an outlier present for a female cat that is never kept outside.
This outlier appears to match the median heart weight of a male cat that was never kept outside so this possible could be an error.
This outlier could have possibly been entered into the data as a female when it was actually a male.

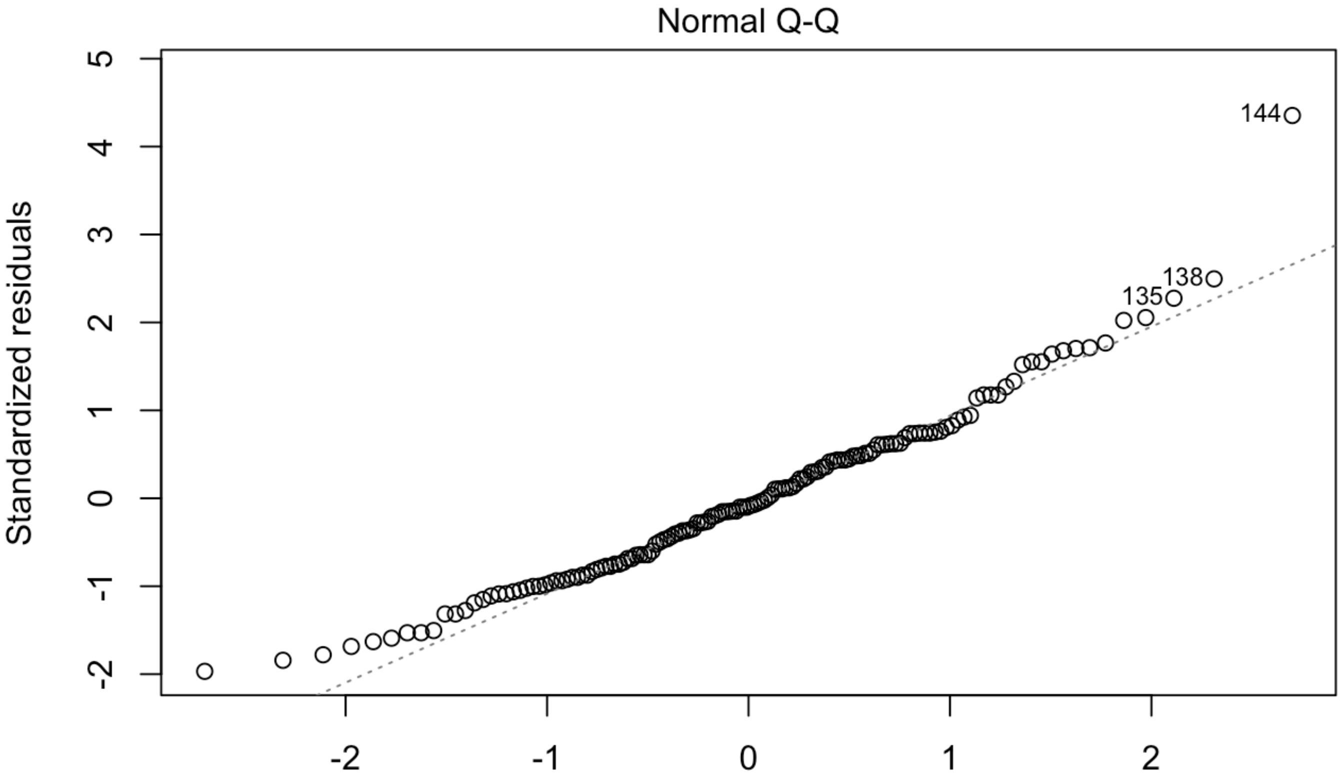
```
#Interaction Plot
interaction.plot(x.factor = Sex,
                 trace.factor = Outdoor,
                 response = Hwt,
                 fun = mean, trace.label ="Outdoor",
                 type = "b", legend = TRUE,
                 main="Interaction Plot",xlab = "Sex", ylab="Heart Weight (g)",
                 pch=c(1,2),col=c(1,2))
```

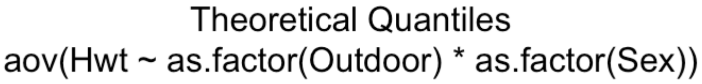


When we look at the interaction plot, the lines seem roughly parallel for cat’s that are frequently and never kept outdoors.
The line for cat’s that are always kept outdoors doesn’t appear to be parallel with the other outdoor groups.
Therefore it appears as if the interaction term may be important between these two variables.

```
res2<-aov(Hwt~as.factor(Outdoor)*as.factor(Sex),data=cat_data)
plot(res2, which=1:2)
```

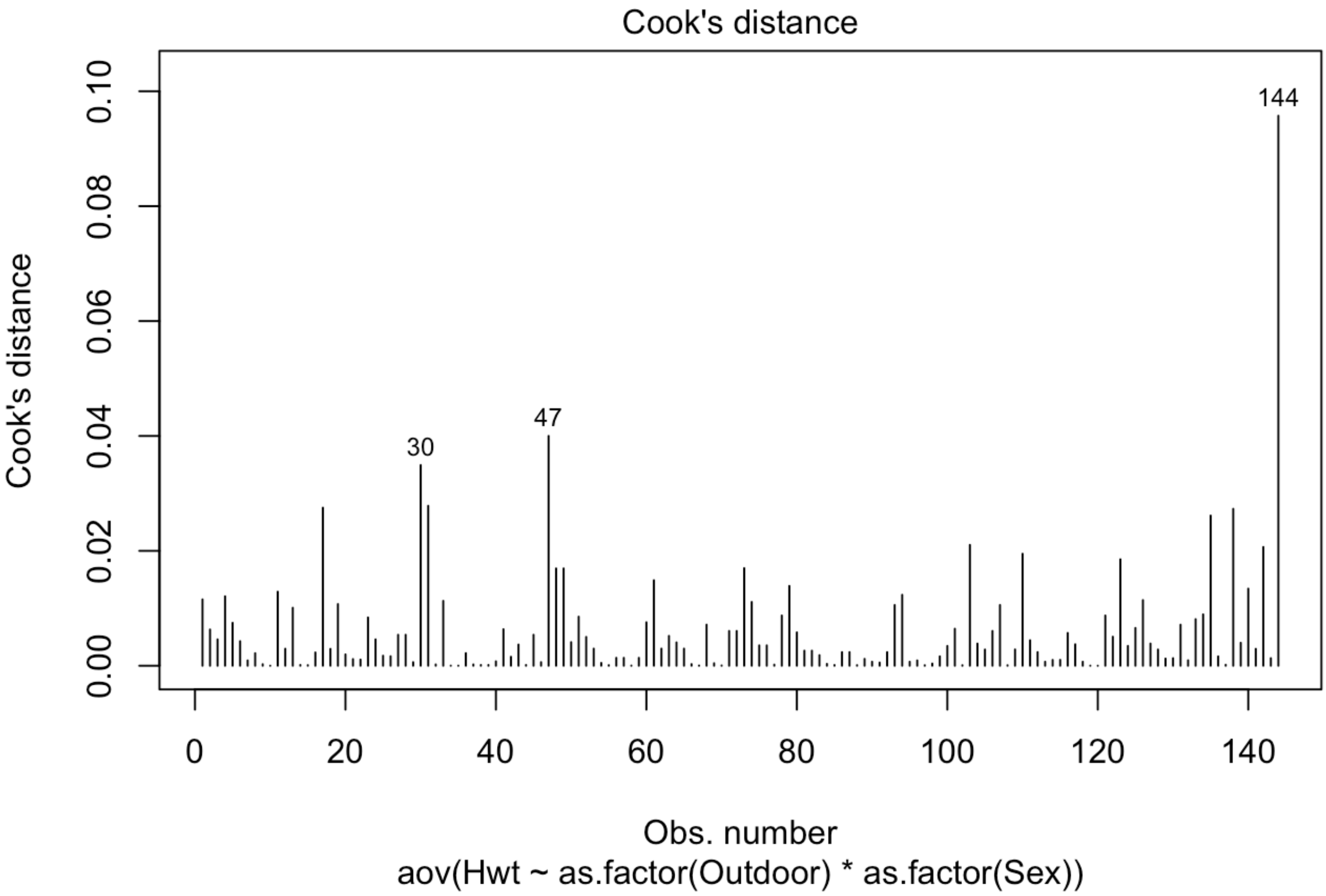

Residuals vs Fitted
Residuals
8
6
4
2
0
-2
-4
144
138
135
9
10
11
12
13
Fitted values
aov(Hwt ~ as.factor(Outdoor) * as.factor(Sex))

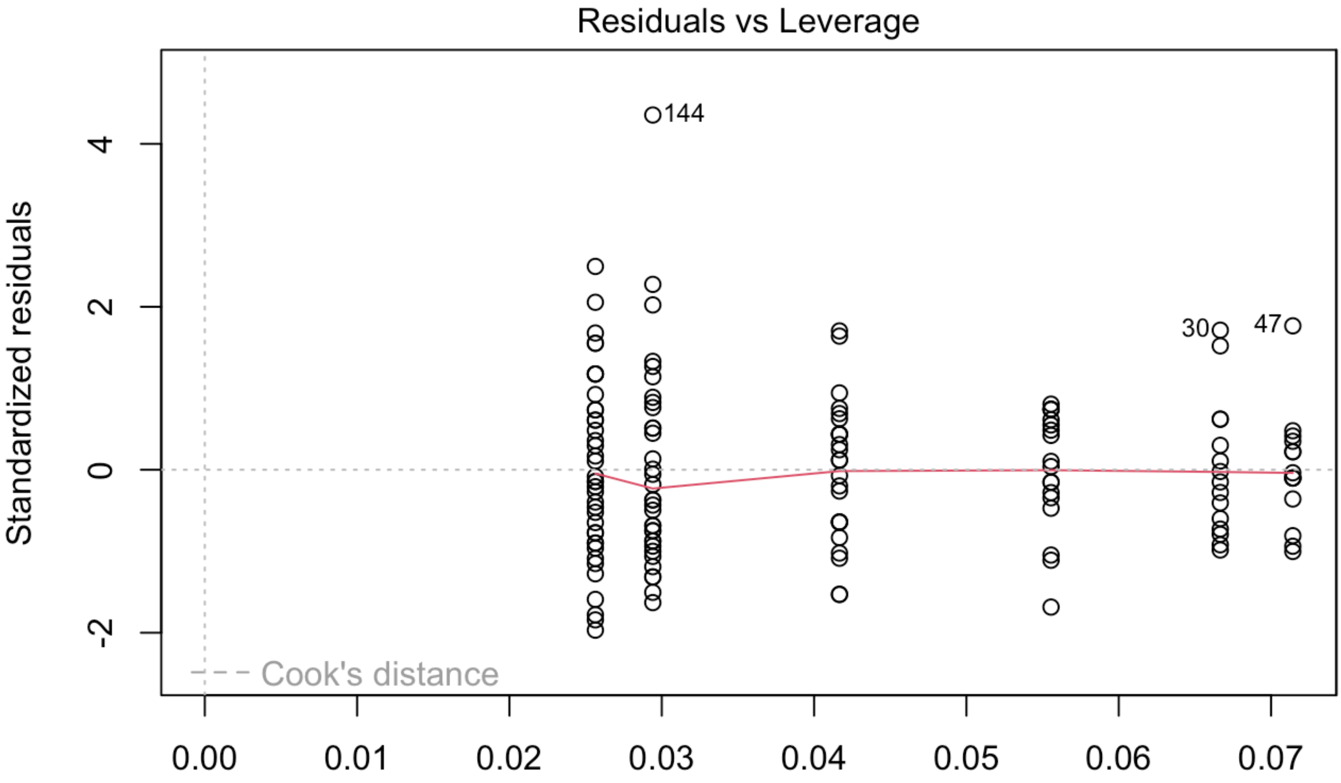

Normal Q-Q
Standardized residuals
5
4
3
2
1
0
-1
-2
144
138
135
-2
-1
0
1
2



Interpret Residuals vs Fitted:
The spread of the groups appear to be roughly the same so we are happy with our equal variance assumption.
There appears to be an outlier in the data represented by data point 144.
The scatter seems randomly distributed within in the groups, so we are happy with our iid assumption.
Interpret Normal Q-Q Plot:
The points fall roughly in a straight line indicating that the residuals are roughly normally distributed.
There appears to be an outlier in the data represented by data point 144.

```
plot(res2, which=4:5)
```


Cook's distance
Cook's distance
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
80
100
120
140
30
47
144
Obs. number
aov(Hwt ~ as.factor(Outdoor) * as.factor(Sex))


Residuals vs Leverage
Standardized residuals
-2
0
2
4
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
144
30
47
Cook's distance

Leverage
aov(Hwt ~ as.factor(Outdoor) * as.factor(Sex))

There is an outlier with data point 144 but there is no need to worry about this outlier as the cook’s distance is below 0.5

```
#Check if the data is balanced
addmargins(table(Outdoor,Sex))
```

```
##        Sex
## Outdoor   1   2 Sum
##     1    24  15  39
##     2    39  18  57
##     3    34  14  48
##     Sum  97  47 144
```

The data appears to be unbalanced as there is nearly twice as much males than females.

```
library(car)
```

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

```
Anova(res2, type = "2")
```

```
## Anova Table (Type II tests)
## 
## Response: Hwt
##                                    Sum Sq  Df F value    Pr(>F)    
## as.factor(Outdoor)                 305.24   2 58.7737 < 2.2e-16 ***
## as.factor(Sex)                     113.78   1 43.8176 7.374e-10 ***
## as.factor(Outdoor):as.factor(Sex)  41.66   2  8.0216 0.0005062 ***
## Residuals                          358.36 138                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing 3 hypothesis here (2 main effects and 1 interaction effect)
1st H0: The true mean for cat’s heart weight is the same across outdoor groups.
1st H1: The true means for cat’s heart weight are not the same across outdoor groups.
2nd H0: The true mean for cat’s heart weight is the same across males and females.
2nd H1: The true means for cat’s heart weight are not the same across males and females.
3rd H0: There is no interaction effect between sex and outdoor groups on cat’s heart weight.
3rd H0: There is an interaction effect between sex and outdoor groups on cat’s heart weight.

Interpretting the three tests:
-1 The p-value < 2.2e-16 which is less than 0.05 so we reject the first null hypothesis.
We can conclude that there is a relationship between cat’s heart weight and outdoor groups.

-2 The p-value = 7.374e-10 which is less than 0.05 so we reject the second null hypothesis.
We can conclude that there is a relationship between cat’s heart weight and sex.

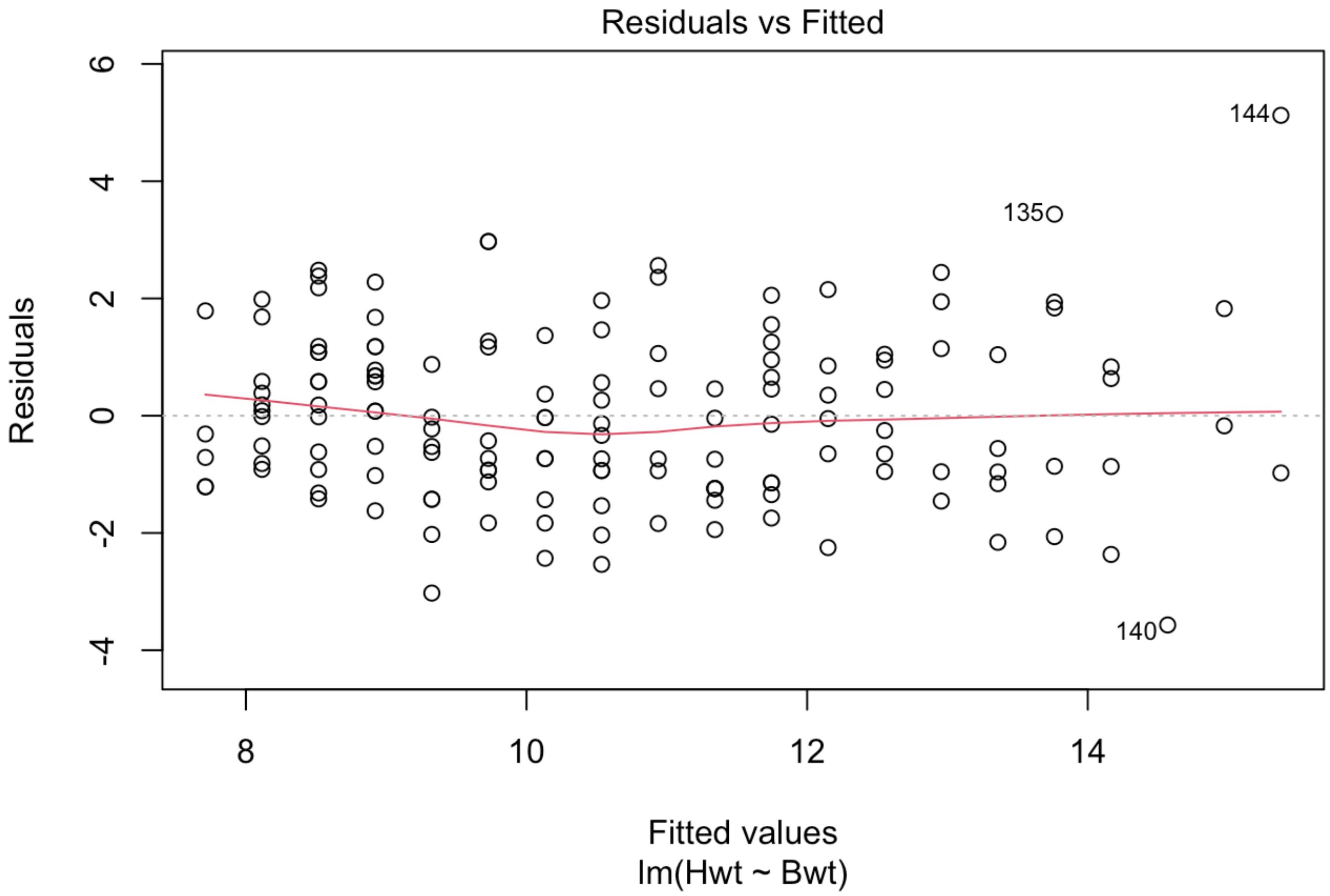
-3 The p-value = 0.0005062 which is less than 0.05 so we reject the third null hypothesis.
There is an interaction effect between sex and outdoor groups on cat’s heart weight.
The relationship between cat’s heart weight, outdoor groups and sex groups depend on each other.

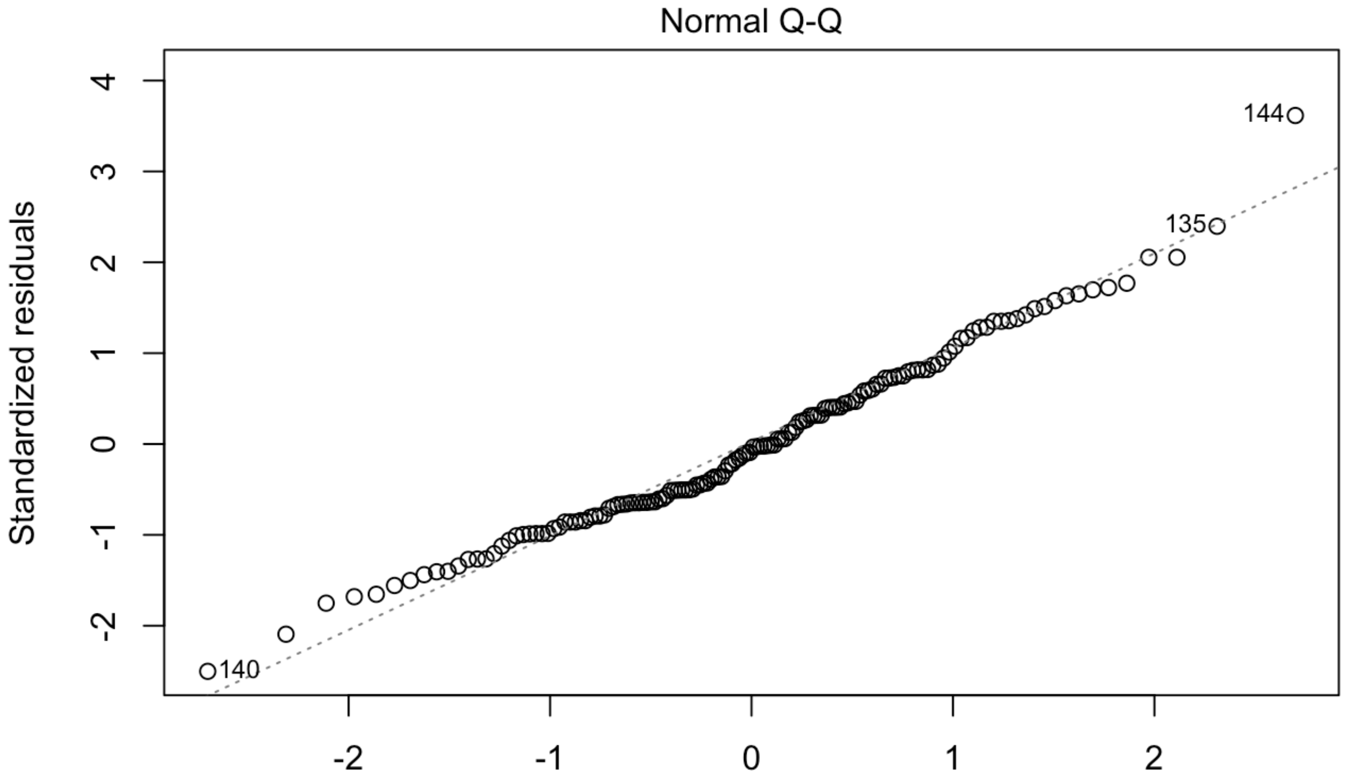
2. Multiple Linear Regression

a) Fit the “best” simple linear regression model, based on your answers from question 1, justify your choice. Comment on whether the assumptions are satisfied, interpretation of the results and the fit of the model.

MODEL1 = cat’s heart weight and cat’s body weight

```
#MODEL1 = cat's heart weight and cat's body weight
#Treating cat's heart weight as response variable
slr_Hwt_Bwt = lm(Hwt~Bwt)
#Check Assumptions
plot(slr_Hwt_Bwt,which=1:2)
```


Residuals vs Fitted
Residuals
6
4
2
0
-2
-4
144
135
140
8
10
12
14
Fitted values
lm(Hwt ~ Bwt)

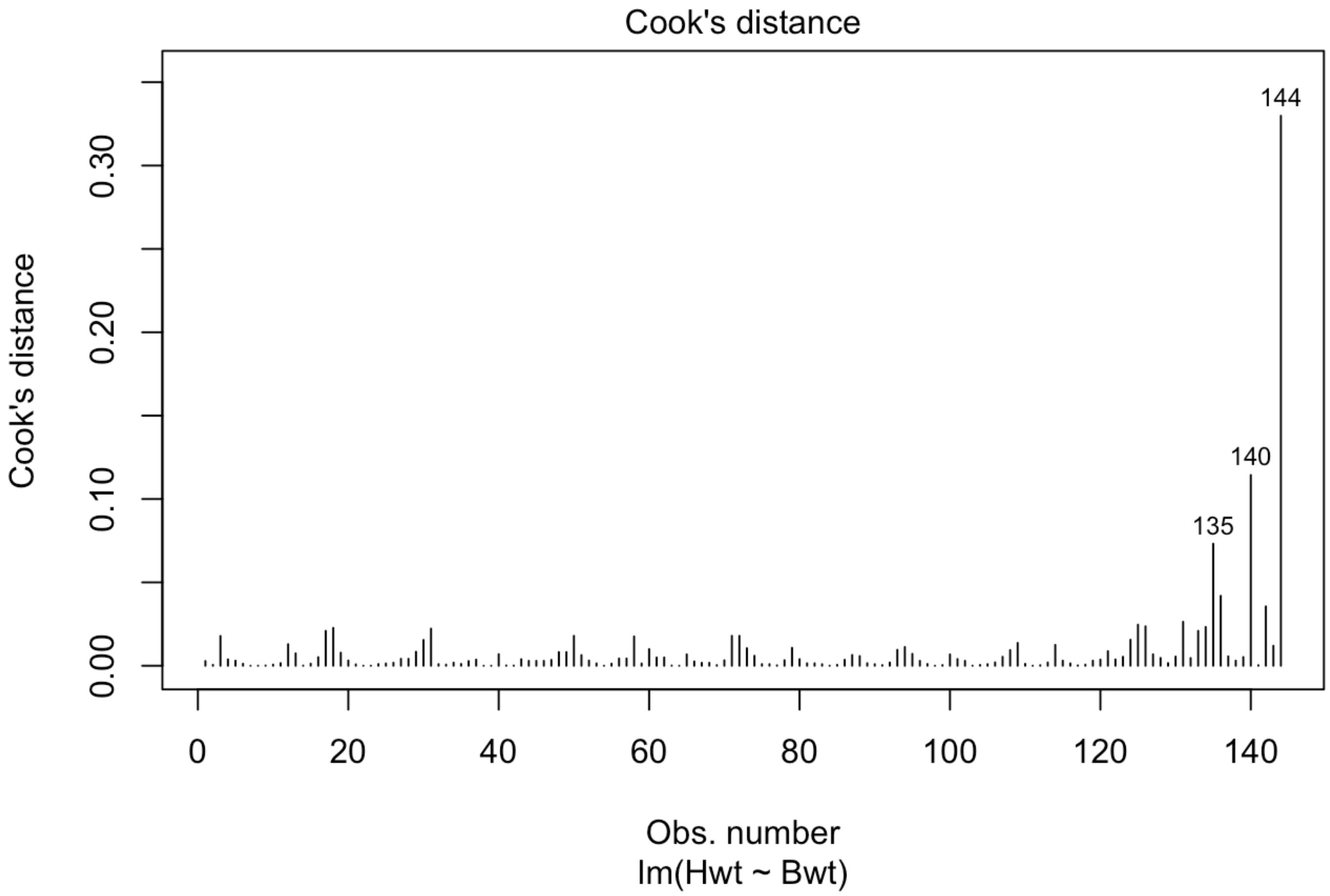

Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
144
135
140
-2
-1
0
1
2

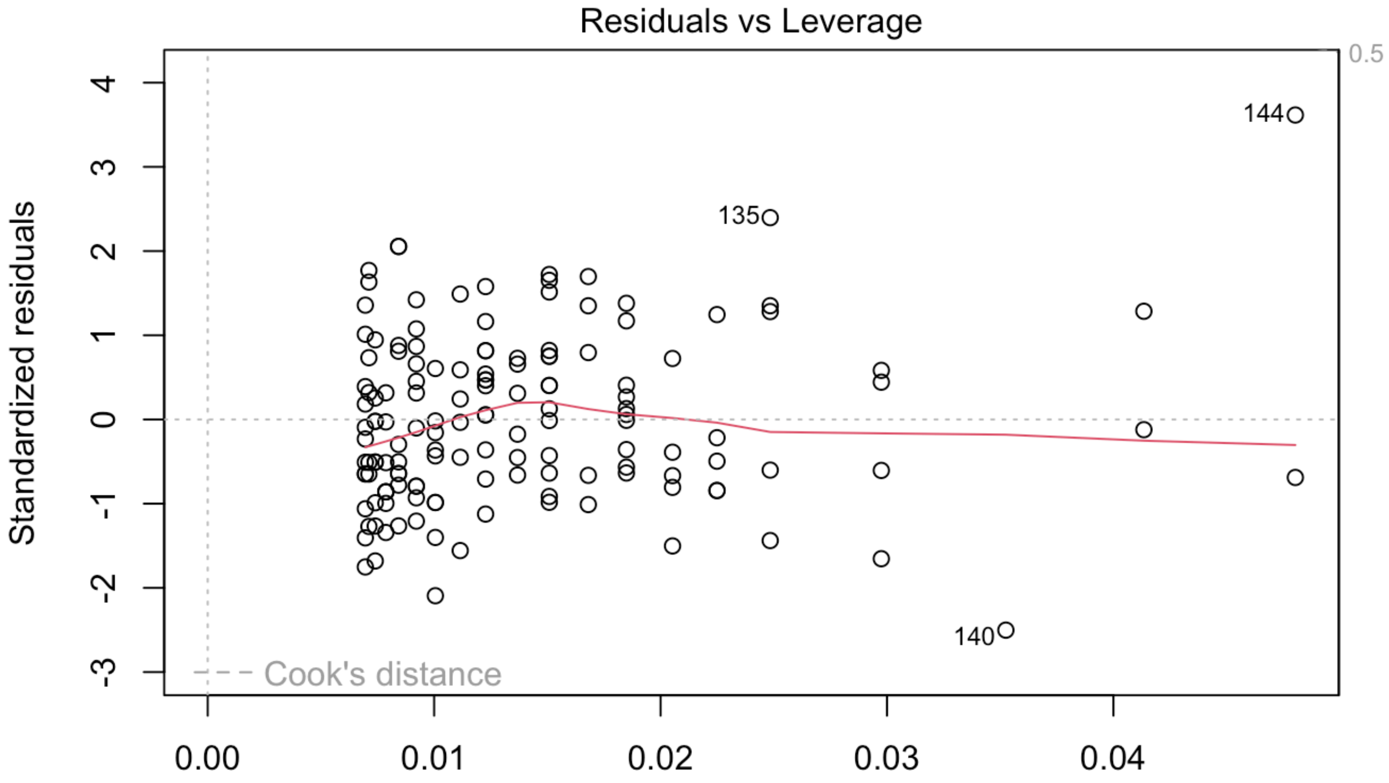
Theoretical Quantiles
lm(Hwt ~ Bwt)

Interpret Residuals vs Fitted:
The spread looks roughly the same, so we are happy with our equal variance assumption.
The scatter seems randomly distributed, so we are happy with our iid assumption.

Interpret Normal Q-Q Plot:
The points fall roughly in a straight line indicating that the residuals are roughly normally distributed.

```
plot(slr_Hwt_Bwt,which=4:5)
```


Cook's distance
Cook's distance
0.00
0.10
0.20
0.30
0
20
40
60
80
100
120
140
144
140
135
Obs. number
lm(Hwt ~ Bwt)


Residuals vs Leverage
Standardized residuals
-3
-2
-1
0
1
2
3
4
0.5
144
135
140
Cook's distance
0.00
0.01
0.02
0.03
0.04

Leverage
lm(Hwt ~ Bwt)

There is no need to worry about outliers as the cook's distance is below 0.5.
There is also no high leverage points.

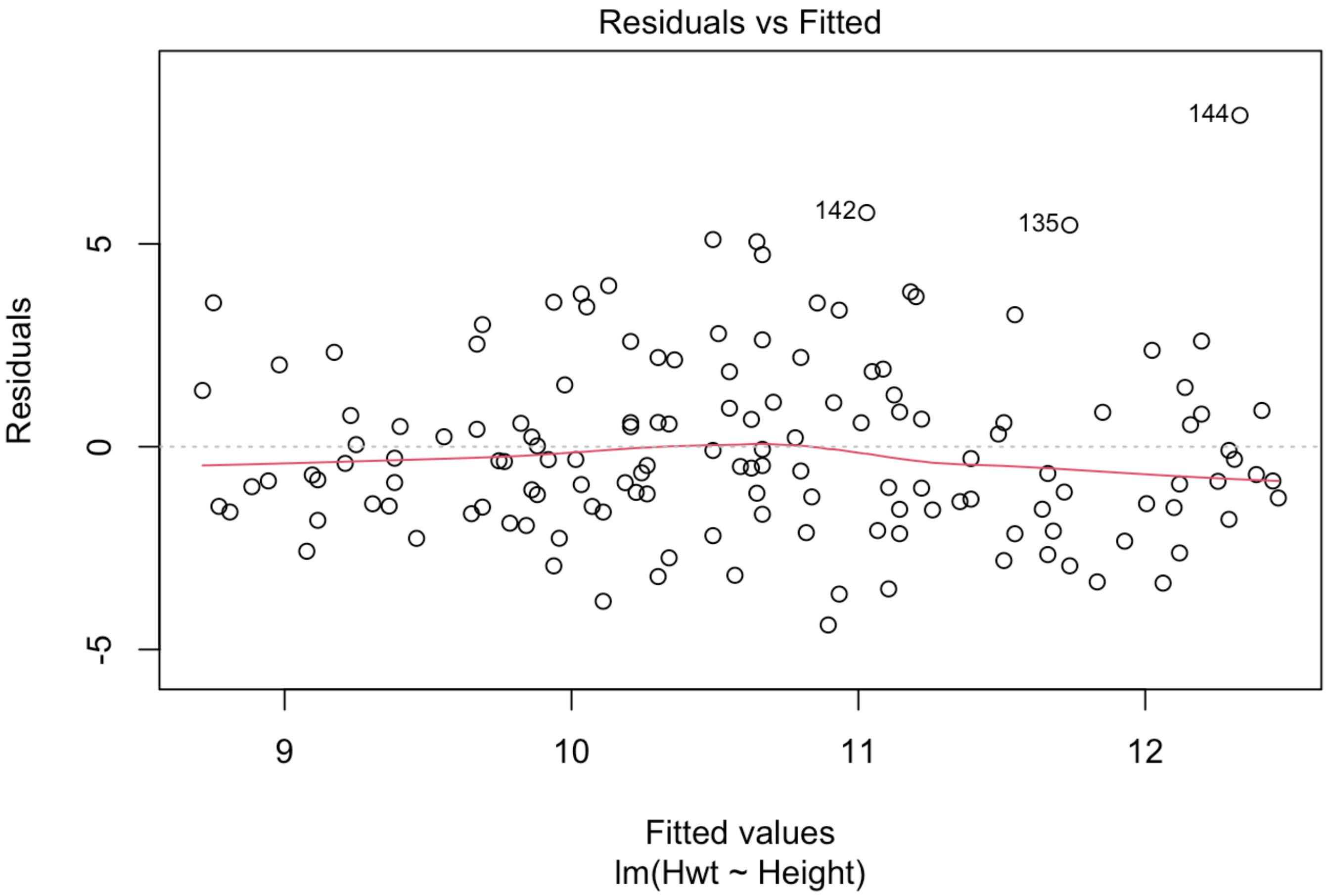
```
summary(slr_Hwt_Bwt)
```

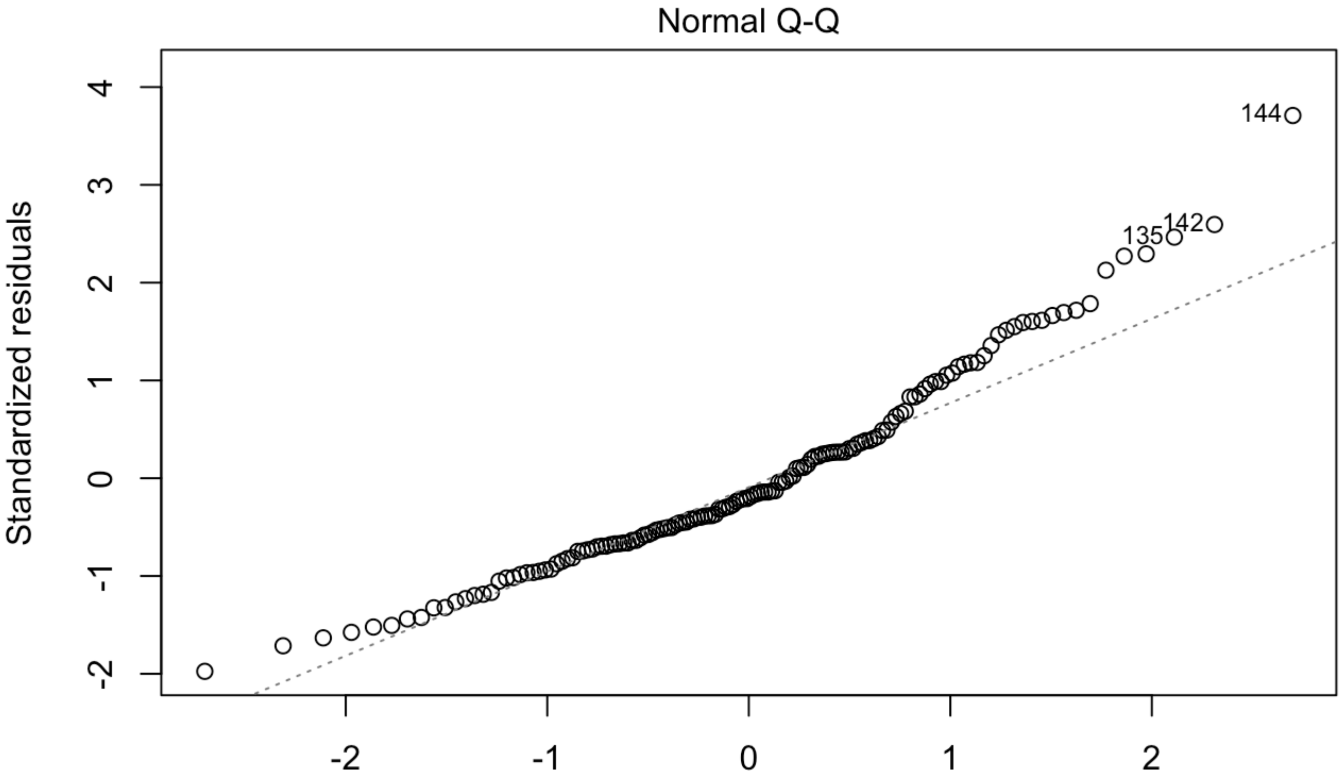
```
##
## Call:
## lm(formula = Hwt ~ Bwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5694 -0.9634 -0.0921  1.0426  5.1238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923  -0.515    0.607
## Bwt           4.0341     0.2503  16.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 142 degrees of freedom
## Multiple R-squared:  0.6466, Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

Cat's heart weight = -0.3567 + 4.0341 (Cat's body weight)
For every 1kg increase of cat's body weight we expect cat's heart weight to increase on average by 4g.
The intercept can't be sensibly interpreted as it means if we have 0kg for cat's body weight then cat's heart weight is -0.4g.
H0: True slope between cat's heart weight and cat's body weight = 0 i.e. No Relationship.
H1: True slope between cat's heart weight and cat's body weight != 0 i.e. Relationship.
As the p-value < 2.2e-16 which is less than 0.05 we reject the null hypothesis.
We can conclude that there is a relationship between cat's heart weight and cat's body weight.
Almost 65% of variation in cat's heart weight is being explained by the model.
The residual standard error = 1.452, therefore there is a small spread around the line.

MODEL2 = cat's heart weight and cat's height

```
#MODEL2 = cat's heart weight and cat's height
#Treating cat's heart weight as response variable
slr_Hwt_Height = lm(Hwt~Height)
#Check Assumptions
plot(slr_Hwt_Height,which=1:2)
```


Residuals vs Fitted
Residuals
-5
0
5
9
10
11
12
144
142
135
Fitted values
lm(Hwt ~ Height)

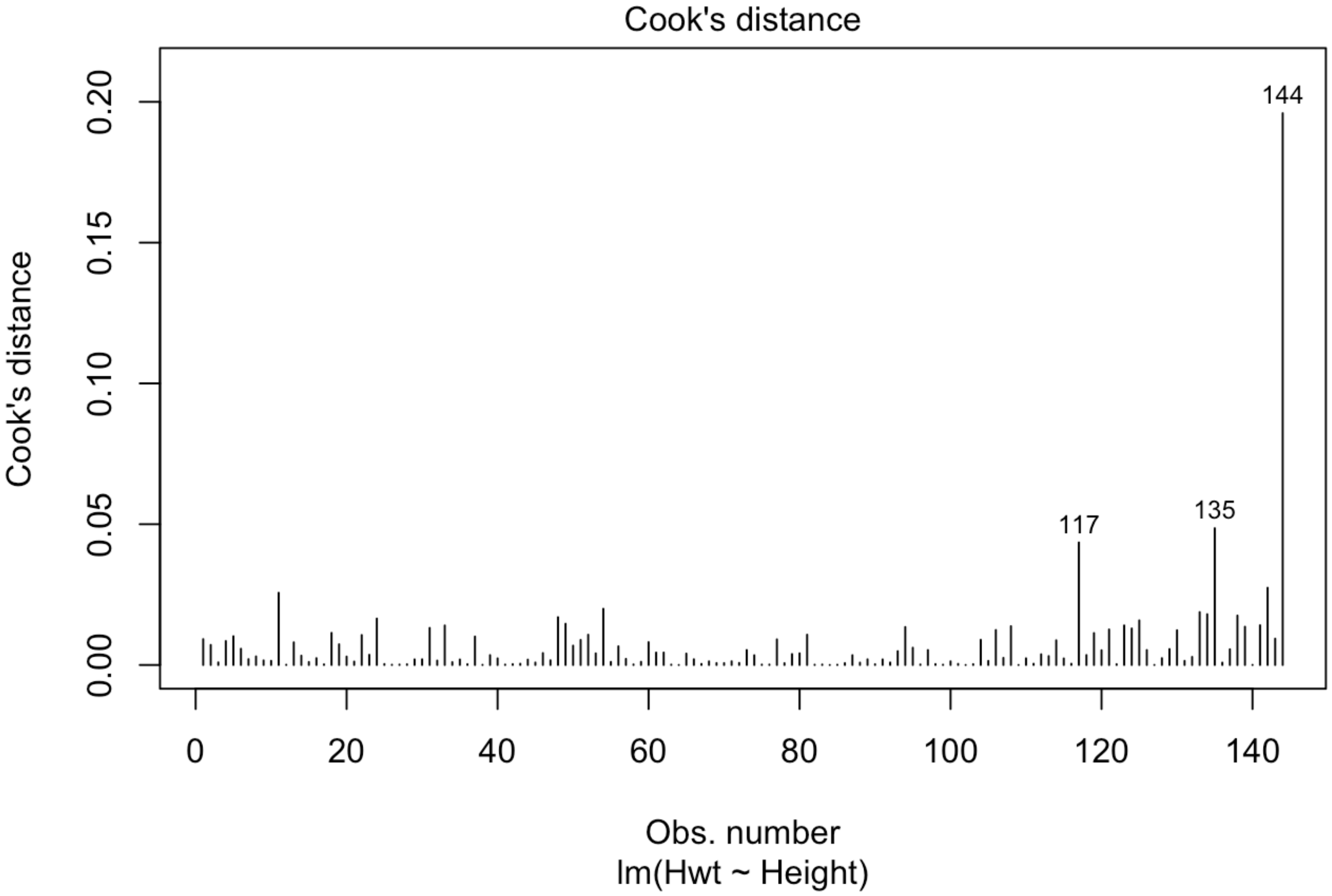

Normal Q-Q
Standardized residuals
-2
-1
0
1
2
3
4
144
135
142

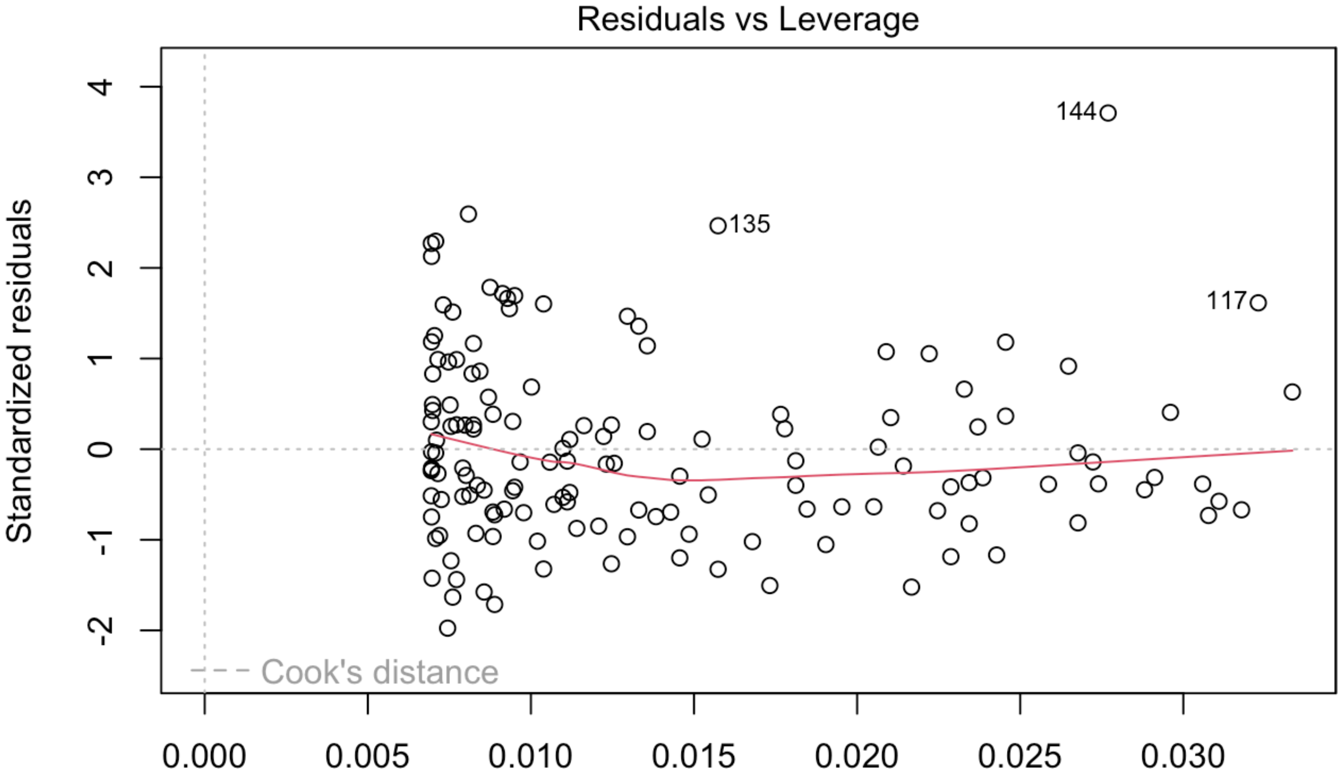
Theoretical Quantiles
lm(Hwt ~ Height)

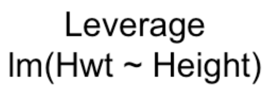
Interpret Residuals vs Fitted:
The spread looks roughly the same, so we are happy with our equal variance assumption.
The scatter seems randomly distributed, so we are happy with our iid assumption.

Interpret Normal Q-Q Plot:
The points fall roughly in a straight line indicating that the residuals are roughly normally distributed.

```
plot(slr_Hwt_Height,which=4:5)
```


Cook's distance
Cook's distance
0.00
0.05
0.10
0.15
0.20
144
117
135
0
20
40
60
80
100
120
140
Obs. number
lm(Hwt ~ Height)


Residuals vs Leverage
Standardized residuals
4
3
2
1
0
-1
-2
144
135
117
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030



There is no need to worry about outliers as the cook's distance is below 0.5.
There is also no high leverage points.

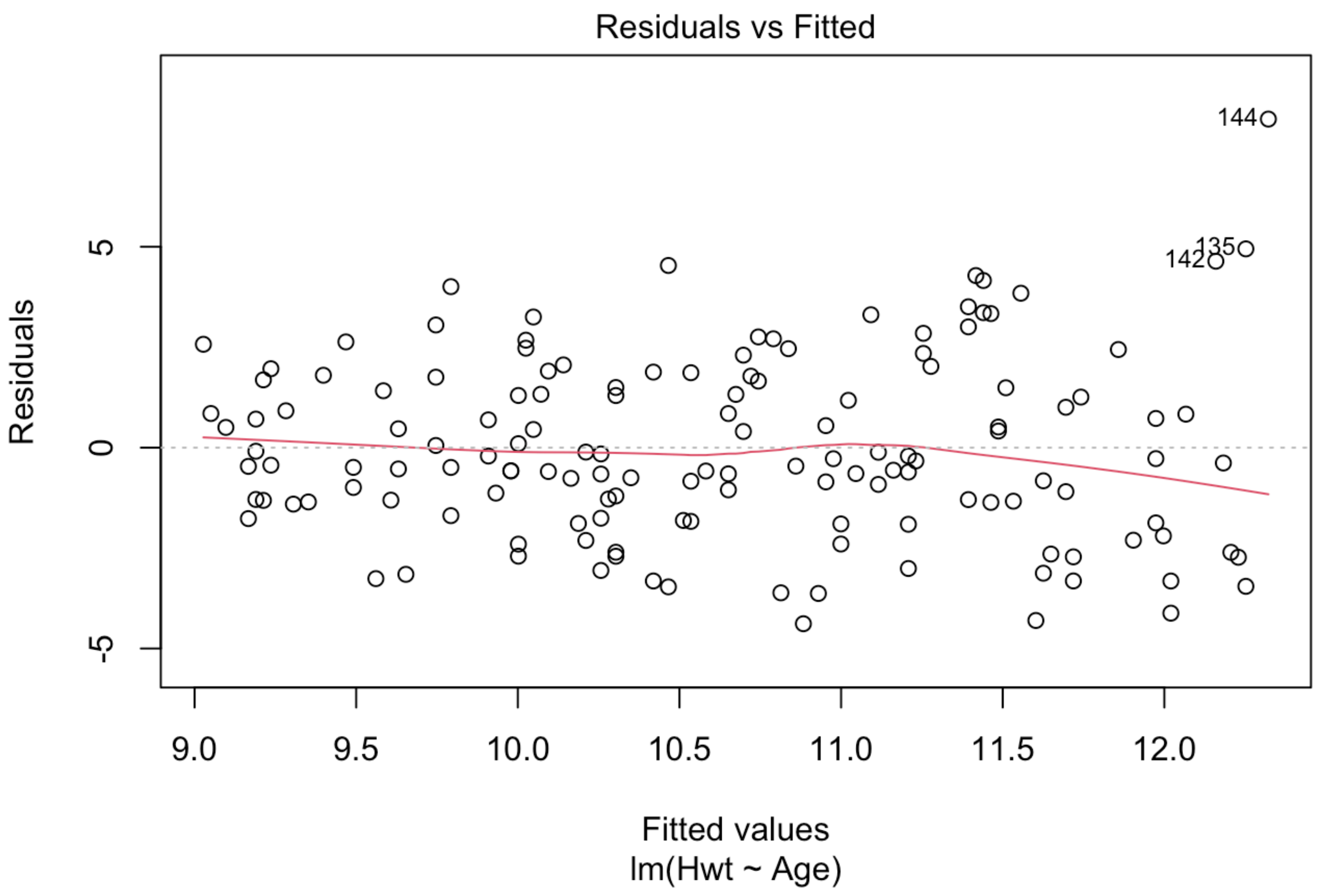
```
summary(slr_Hwt_Height)
```

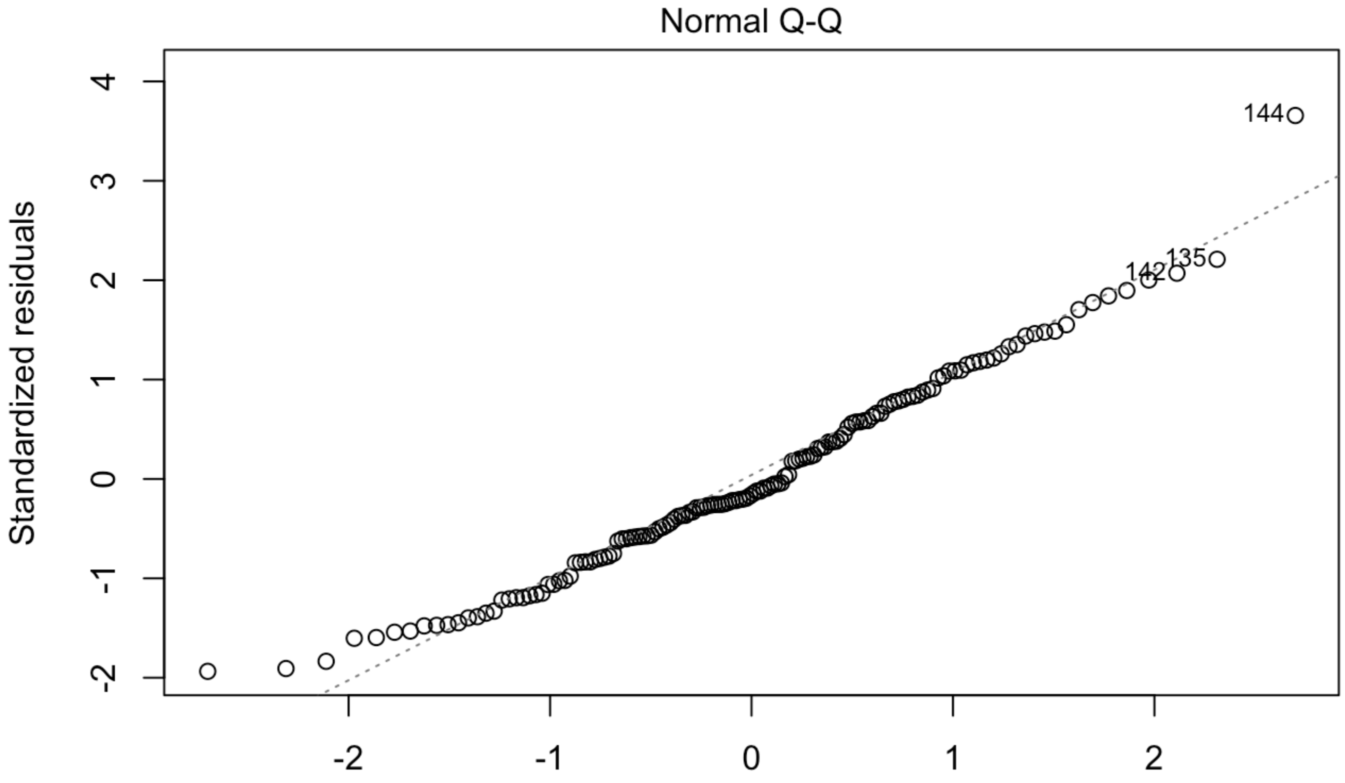
```
##
## Call:
## lm(formula = Hwt ~ Height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3950 -1.4922 -0.4374  1.0885  8.1699
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.97745    0.90035   6.639 6.21e-10 ***
## Height       0.19134    0.03622   5.282 4.69e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.234 on 142 degrees of freedom
## Multiple R-squared:  0.1642, Adjusted R-squared:  0.1583
## F-statistic:  27.9 on 1 and 142 DF,  p-value: 4.693e-07
```

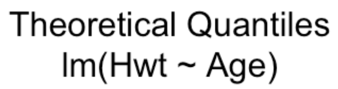
Cat's heart weight = 5.97745 + 0.19134 (Cat's height)
For every 1cm increase of cat's height we expect cat's heart weight to increase on average by 0.2g.
The intercept can't be sensibly interpreted as it means if we have 0cm for cat's height then cat's heart weight is 6g.
H0: True slope between cat's heart weight and cat's height = 0 i.e. No Relationship.
H1: True slope between cat's heart weight and cat's height != 0 i.e. Relationship.
As the p-value = 4.693e-07 which is less than 0.05 we reject the null hypothesis.
We can conclude that there is a relationship between cat's heart weight and cat's height.
16% of variation in cat's heart weight is being explained by the model.
The residual standard error = 2.234, therefore there is a bigger spread around the line than we seen in Model1.

MODEL3 = cat's heart weight and cat's age

```
#MODEL3 = cat's heart weight and cat's age
#Treating cat's heart weight as response variable
slr_Hwt_Age = lm(Hwt~Age)
#Check Assumptions
plot(slr_Hwt_Age,which=1:2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Hwt ~ Age)
144
135
142
5
0
-5
9.0
9.5
10.0
10.5
11.0
11.5
12.0

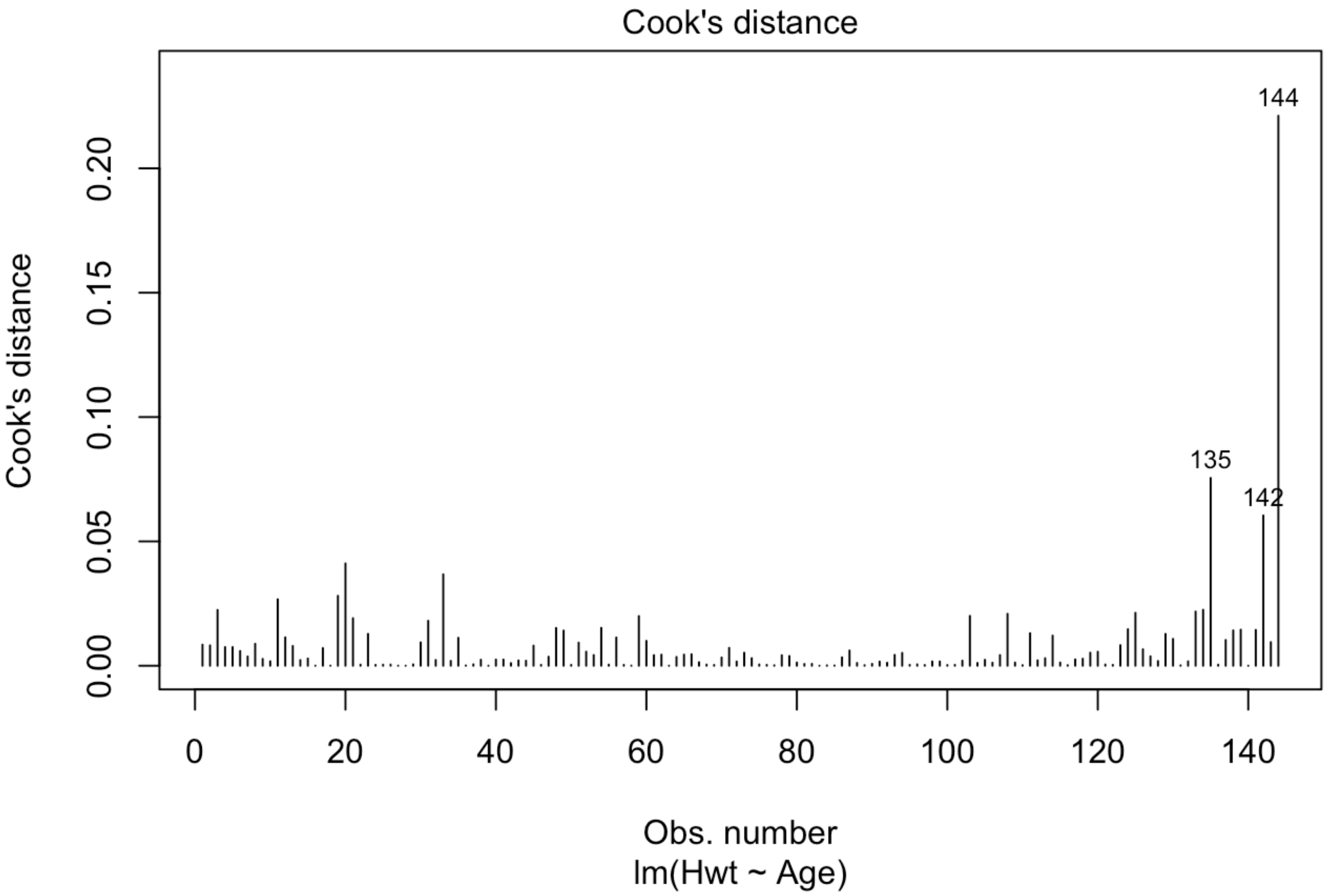

Normal Q-Q
Standardized residuals
144
135
142
4
3
2
1
0
-1
-2
-2
-1
0
1
2

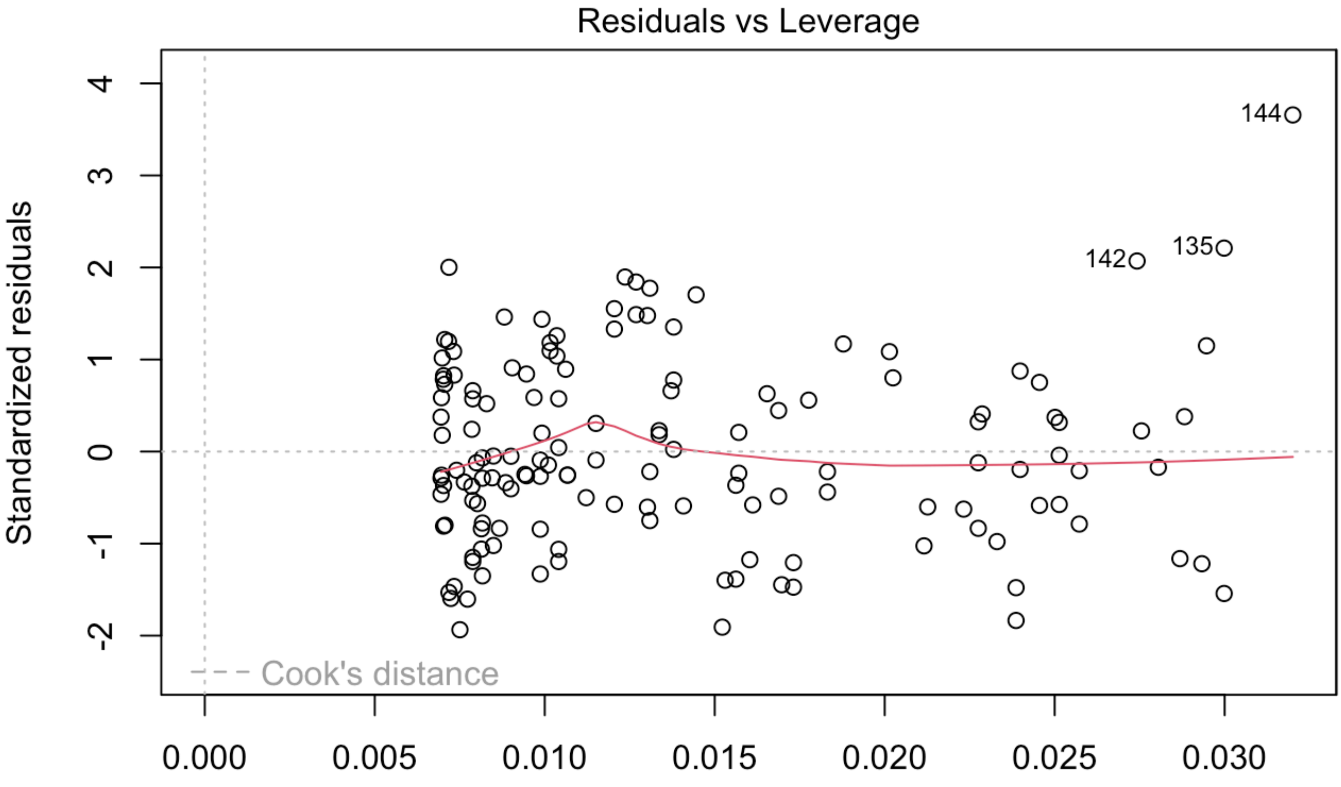


Interpret Residuals vs Fitted:
The spread looks roughly the same, so we are happy with our equal variance assumption.
The scatter seems randomly distributed, so we are happy with our iid assumption.

Interpret Normal Q-Q Plot:
The points fall roughly in a straight line indicating that the residuals are roughly normally distributed.

```
plot(slr_Hwt_Age,which=4:5)
```


Cook's distance
Cook's distance
0.00
0.05
0.10
0.15
0.20
144
135
142
0
20
40
60
80
100
120
140
Obs. number
lm(Hwt ~ Age)


Residuals vs Leverage
Standardized residuals
4
3
2
1
0
-1
-2
144
142
135
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030

Leverage
lm(Hwt ~ Age)

There is no need to worry about outliers as the cook’s distance is below 0.5.
There is also no high leverage points.

```
summary(slr_Hwt_Age)
```

```
##
## Call:
## lm(formula = Hwt ~ Age)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.383 -1.478 -0.357  1.664  8.178
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.79764    0.63173  12.343  < 2e-16 ***
## Age          0.23201    0.04936   4.701 6.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.273 on 142 degrees of freedom
## Multiple R-squared:  0.1347, Adjusted R-squared:  0.1286
## F-statistic:  22.1 on 1 and 142 DF,  p-value: 6.072e-06
```

Cat’s heart weight = 7.79764 + 0.23201 (Cat’s age)
For every 1 year increase of cat’s age we expect cat’s heart weight to increase on average by 0.2g.
The intercept can’t be sensibly interpreted as it means if we have 0 years for cat’s age then cat’s heart weight is 7.8g.
H0: True slope between cat’s heart weight and cat’s age = 0 i.e. No Relationship.
H1: True slope between cat’s heart weight and cat’s age != 0 i.e. Relationship.
As the p-value = 6.072e-06 which is less than 0.05 we reject the null hypothesis.
We can conclude that there is a relationship between cat’s heart weight and cat’s age.
13% of variation in cat’s heart weight is being explained by the model.
The residual standard error = 2.273, therefore there is a bigger spread around the line than we seen in Model1 and it is very similar to Model2.

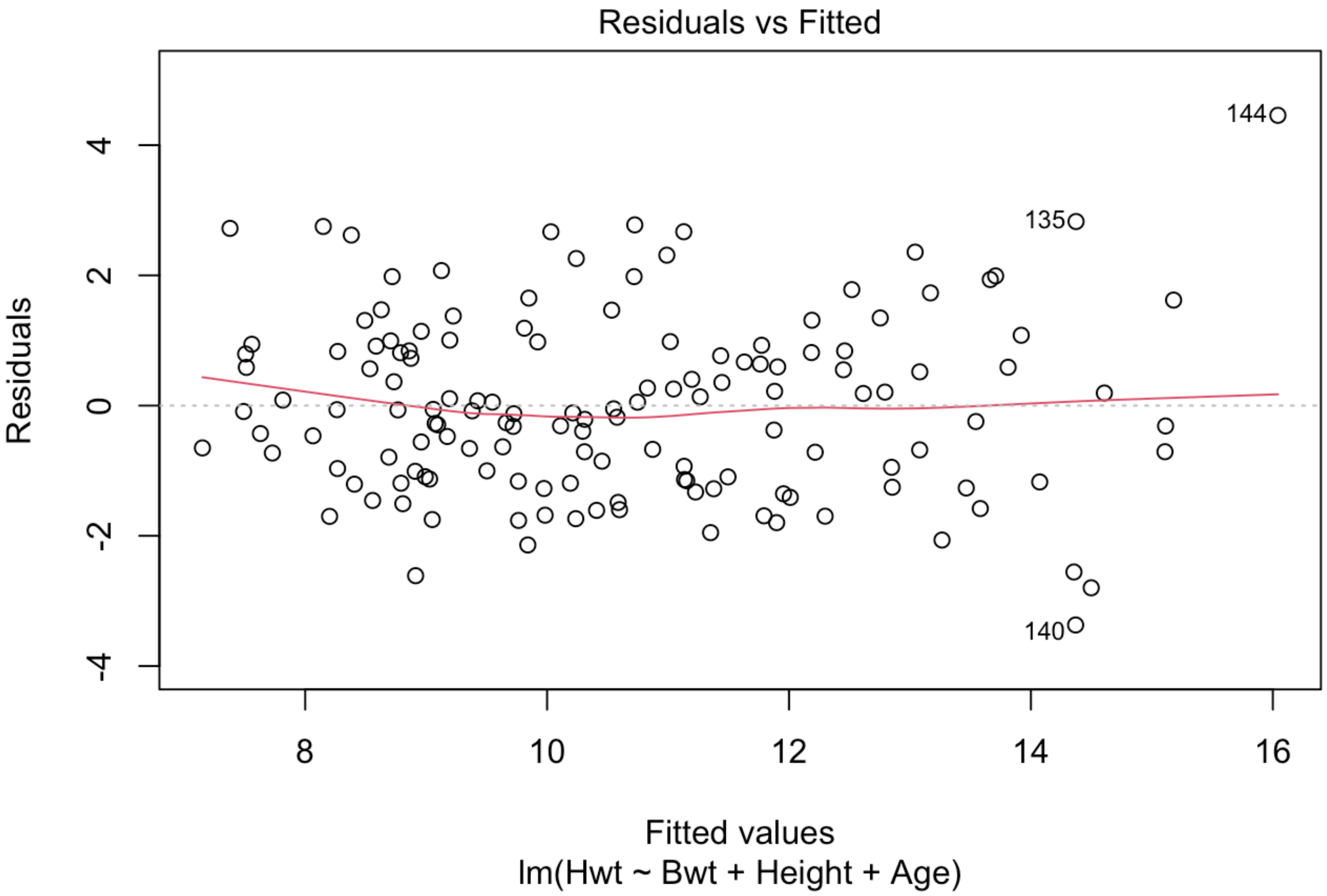
MODEL1 which contains cat’s heart weight and cat’s body weight is our best model
Almost 65% of variation in cat’s heart weight is being explained by the model.
The residual standard error = 1.452, therefore there is a small spread around the line.

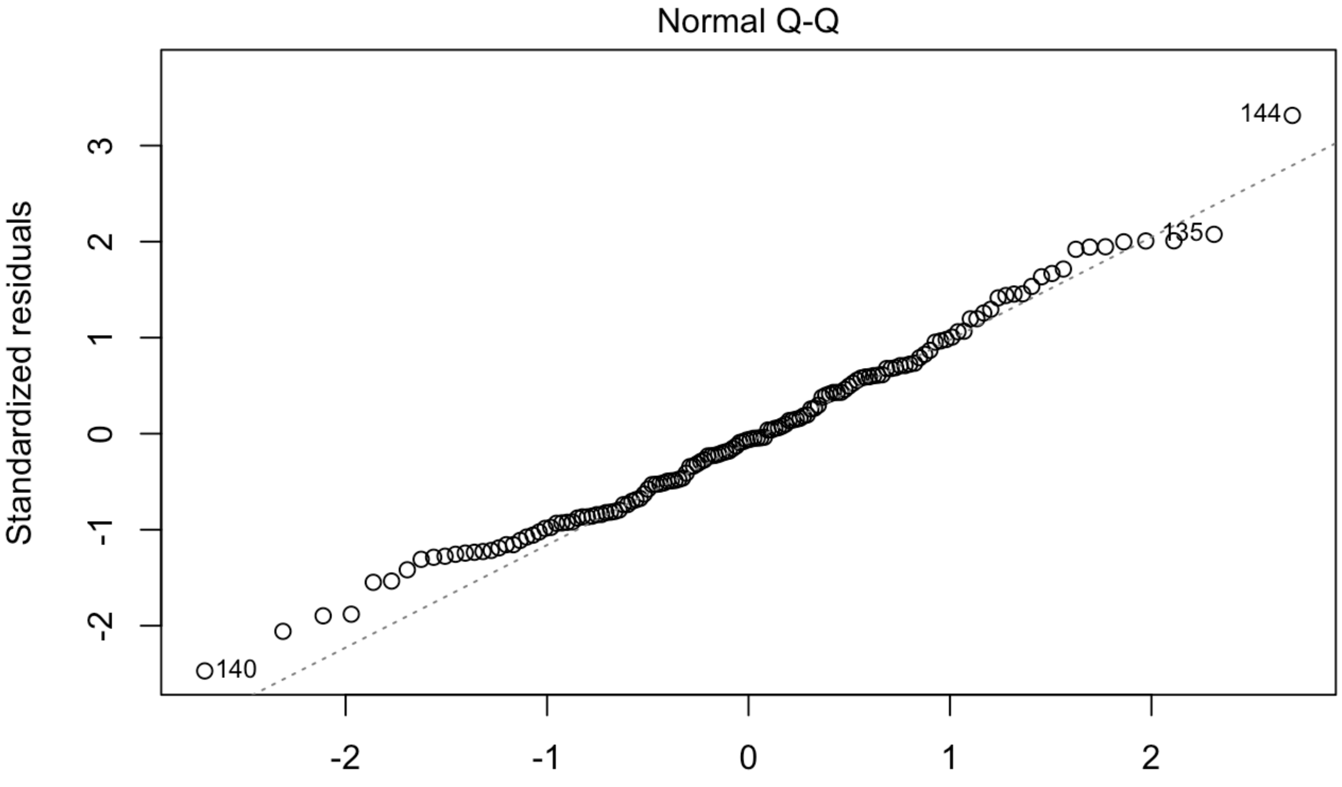
b) Fit a suitable multiple linear regression model, based on your answers from question 1, justify your choice. Comment on whether the assumptions are satisfied, interpretation of the results and the fit of the model.

Possible independent numerical variables in the model are cat’s body weight, height and age as none of these variables are collinear with each other and have linear relationships with cat’s heart weight.

MODEL1: Multiple linear regression model with cat’s body weight, height and age

```
lm1<-lm(Hwt~Bwt+Height+Age,data=cat_data)
plot(lm1,which=1:2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Hwt ~ Bwt + Height + Age)
4
2
0
-2
-4
8
10
12
14
16
144
135
140

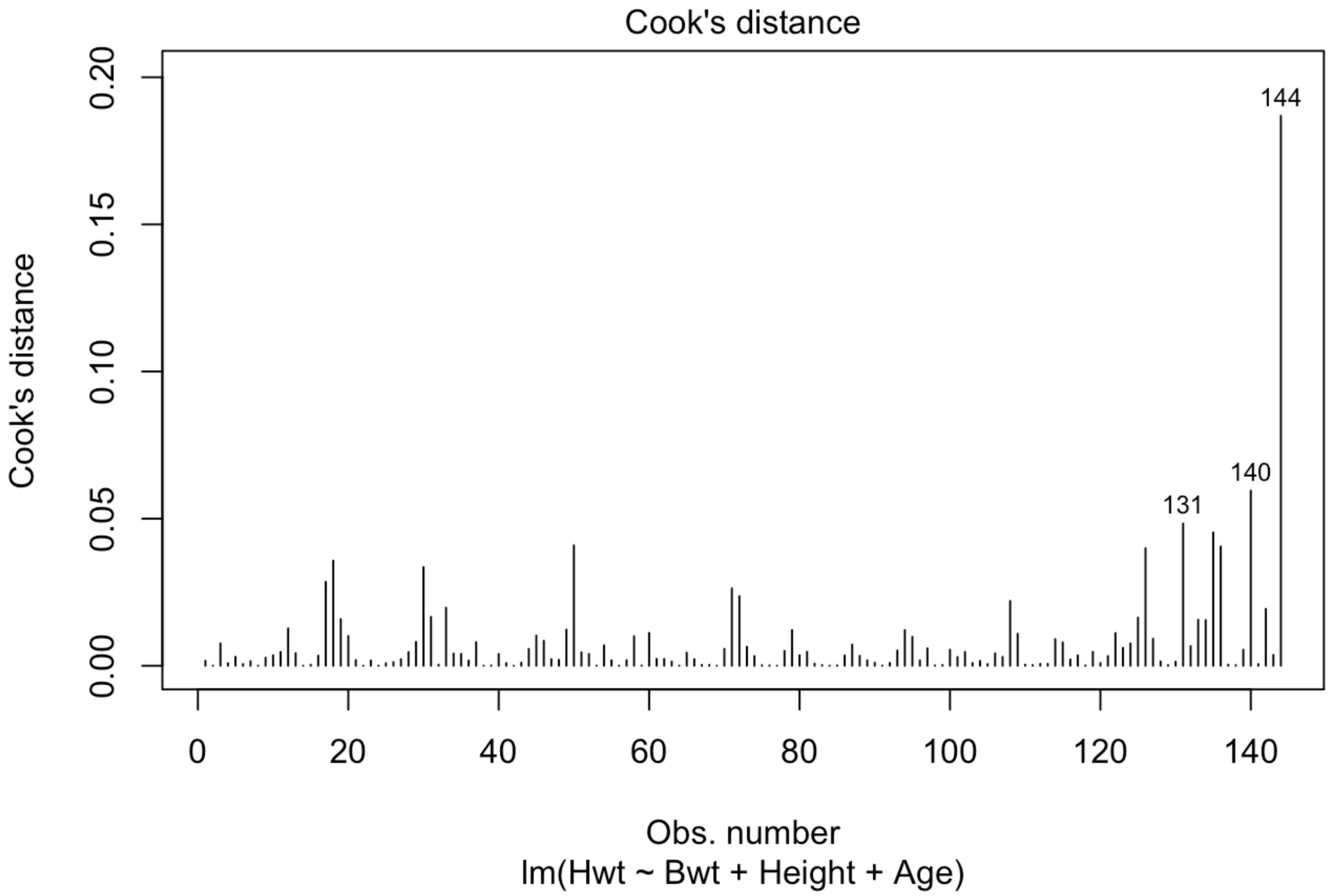

Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-2
-1
0
1
2
144
135
140

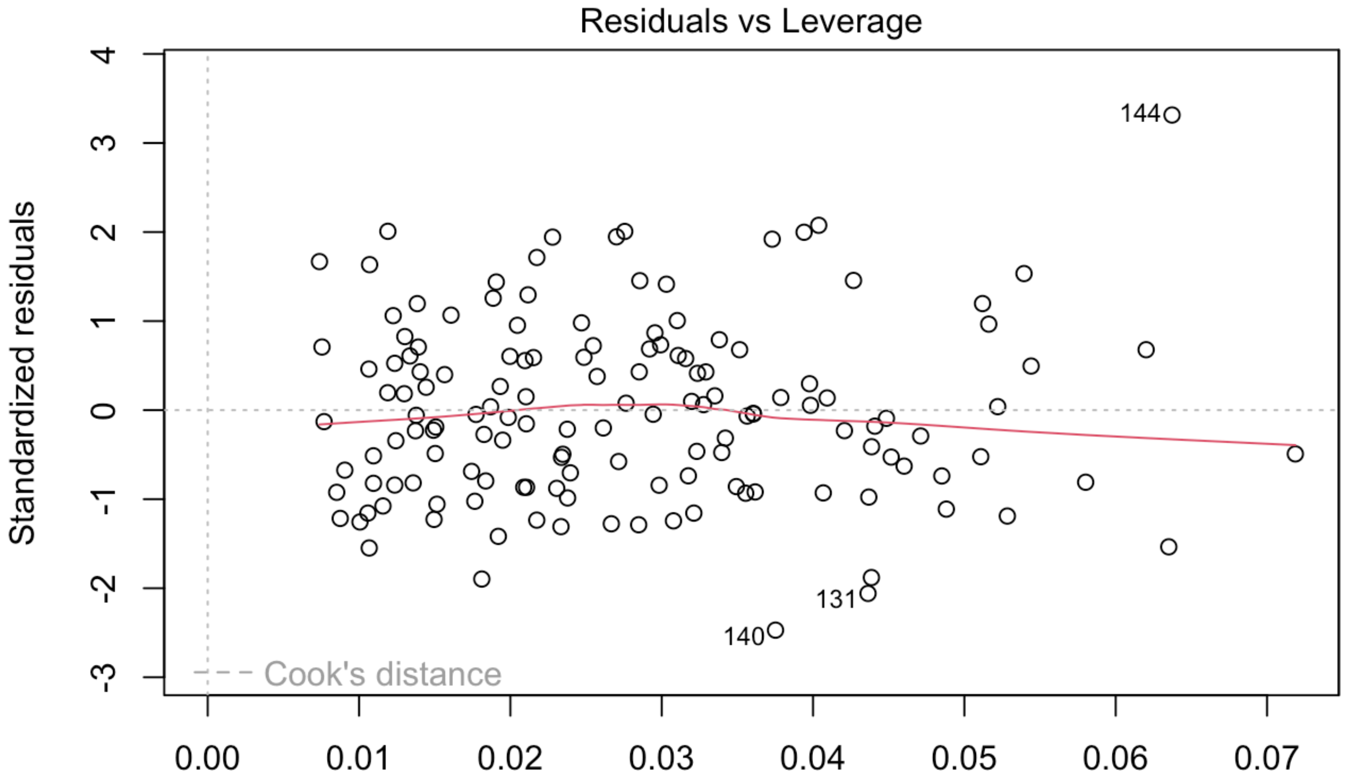
Theoretical Quantiles
lm(Hwt ~ Bwt + Height + Age)

Interpret Residuals vs Fitted:
The spread looks roughly the same, so we are happy with our equal variance assumption.
The scatter seems randomly distributed, so we are happy with our iid assumption.

Interpret Normal Q-Q Plot:
The points fall roughly in a straight line indicating that the residuals are roughly normally distributed.

```
plot(lm1,which=4:5)
```


Cook's distance
Cook's distance
0.00
0.05
0.10
0.15
0.20
144
140
131
0
20
40
60
80
100
120
140
Obs. number
lm(Hwt ~ Bwt + Height + Age)


Residuals vs Leverage
Standardized residuals
4
3
2
1
0
-1
-2
-3
144
131
140
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07

Leverage
lm(Hwt ~ Bwt + Height + Age)

There is no need to worry about outliers as the cook’s distance is below 0.5.
There is also no high leverage points.

```
summary(lm1)
```

```
##
## Call:
## lm(formula = Hwt ~ Bwt + Height + Age, data = cat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3696 -1.1037 -0.0857  0.8583  4.4581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.76574    0.76387  -2.312   0.0223 *
## Bwt          3.60366    0.26409  13.646   <2e-16 ***
## Height       0.06547    0.02405   2.723   0.0073 **
## Age          0.08102    0.03173   2.553   0.0117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.39 on 140 degrees of freedom
## Multiple R-squared:  0.6808, Adjusted R-squared:  0.674
## F-statistic: 99.55 on 3 and 140 DF,  p-value: < 2.2e-16
```

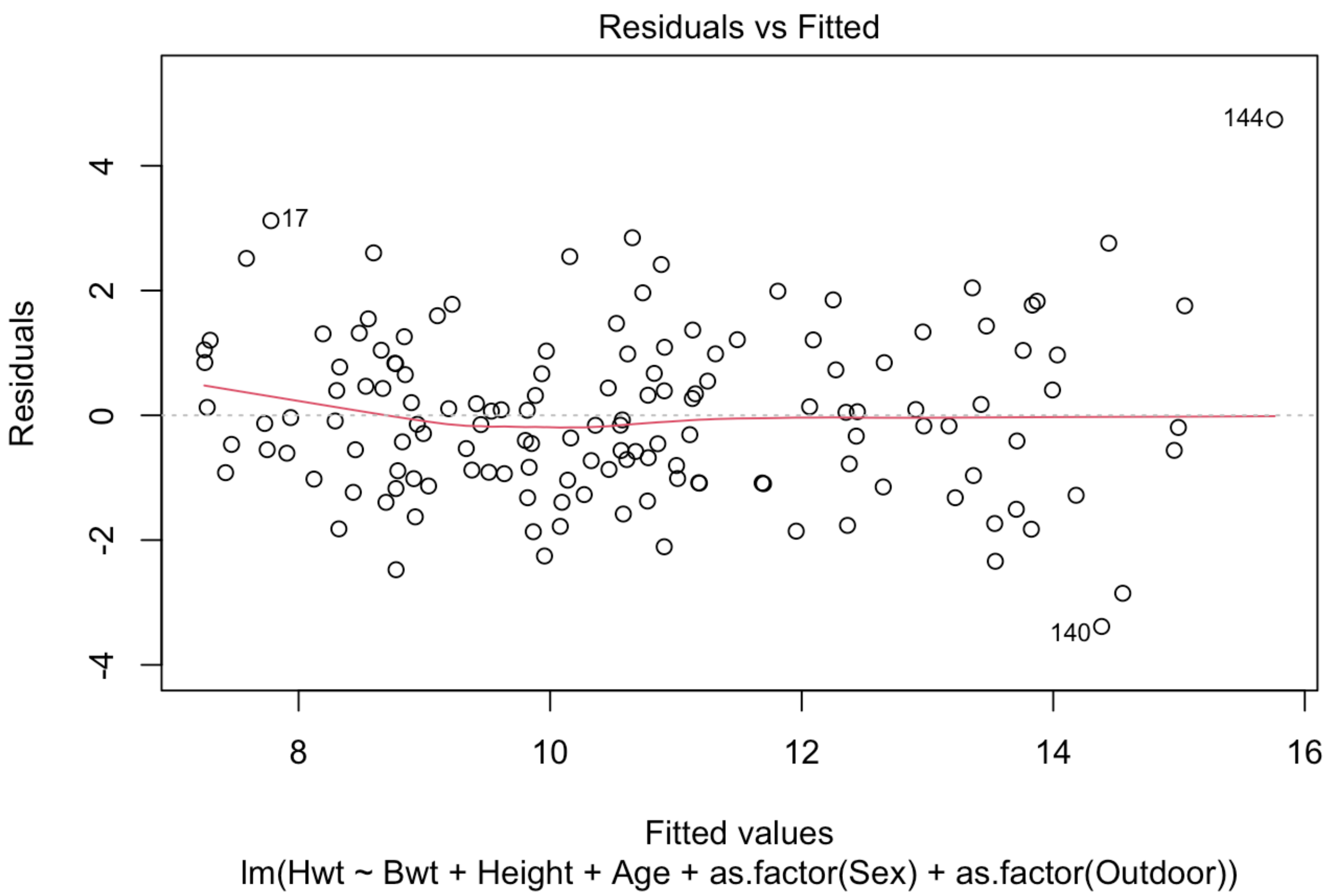
Hwt = -1.76574 + 3.60366 (Bwt) + 0.06547 (Height) + 0.08102 (Age)
For every 1kg in cat’s body weight, cat’s heart weight increases by 3.6g on average keeping cat’s height and age constant.
For every 1cm in cat’s height, cat’s heart weight increases by 0.06547g on average keeping cat’s body weight and age constant.
For every 1year in cat’s age, cat’s heart weight increases by 0.08102g on average keeping cat’s body weight and height constant.
Cat’s body weight has a more significant relationship with cat’s heart weight than cat’s height and age.
68% of variation in cat’s heart weight is being explained by the model and it is better than our best simple linear regression model.
The residual standard error = 1.39, therefore there is a small spread around the line and it is better than our best simple linear regression model.

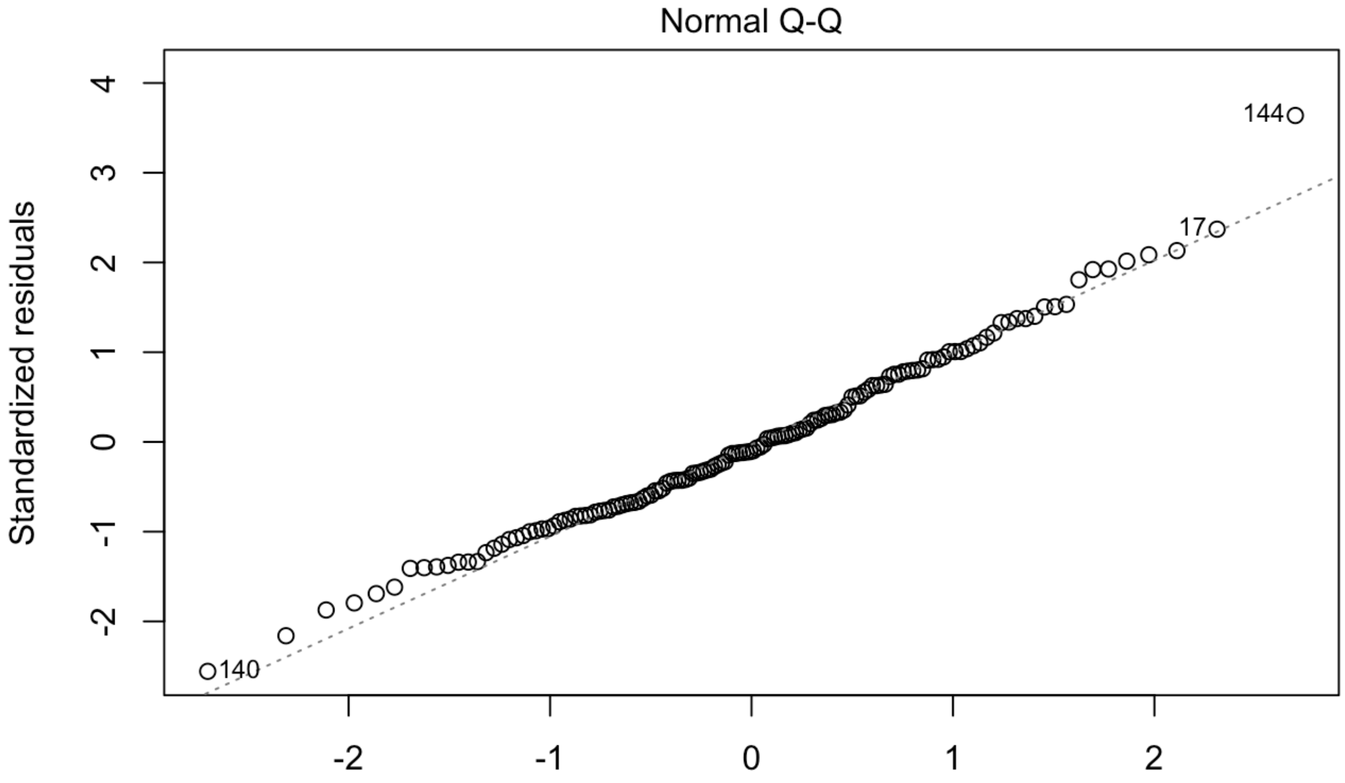
c) Are there any variables you would like to add/remove from the model and why? Re-run the multiple linear regression model with these variables, if any, added/removed. (You can do this more than once, in a stepwise process if you think appropriate). Compare the fit of this model to the model in part b. Pe (rform an F-test to compare the two models, stating your hypothesis and the conclusion of this test. (Hint, use the anova() command). If no variables are removed compare model ran in part a with model ran in part b.

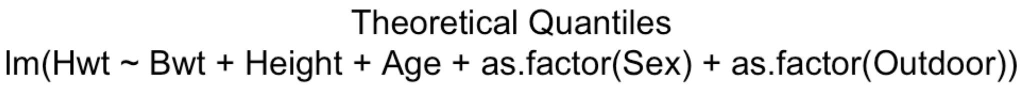
Adding “Sex” and “Outdoor” categories because they have a relationship with “Hwt” cat’s heart weight and may improve our first multiple linear regression model.

MODEL1: Multiple linear regression model with cat’s body weight, height, age, sex and outdoor

```
lm2<-lm(Hwt~Bwt+Height+Age+as.factor(Sex)+as.factor(Outdoor),data=cat_data)
plot(lm2,which=1:2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Hwt ~ Bwt + Height + Age + as.factor(Sex) + as.factor(Outdoor))
4
2
0
-2
-4
8
10
12
14
16
144
17
140

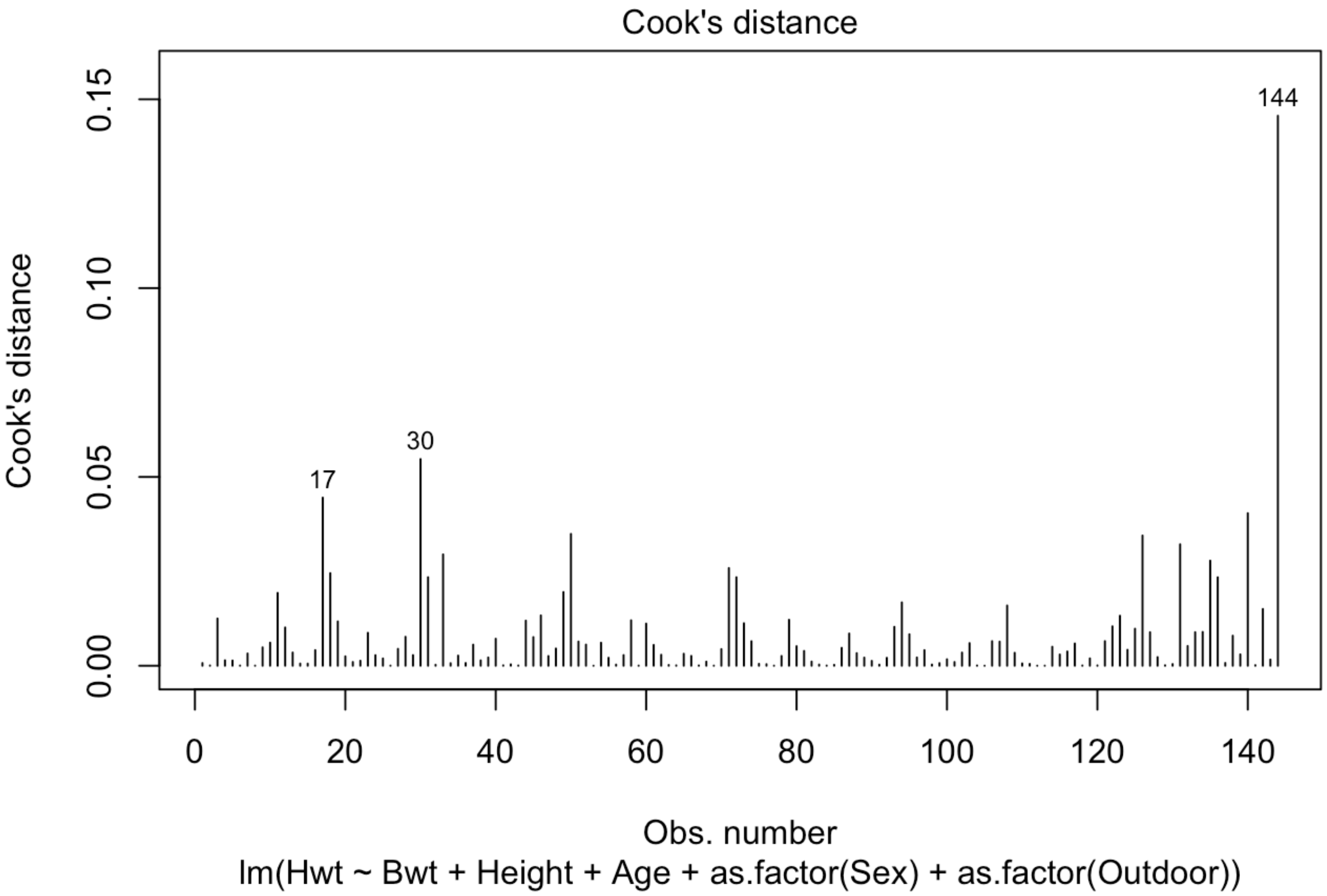

Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
-2
-1
0
1
2
144
17
140

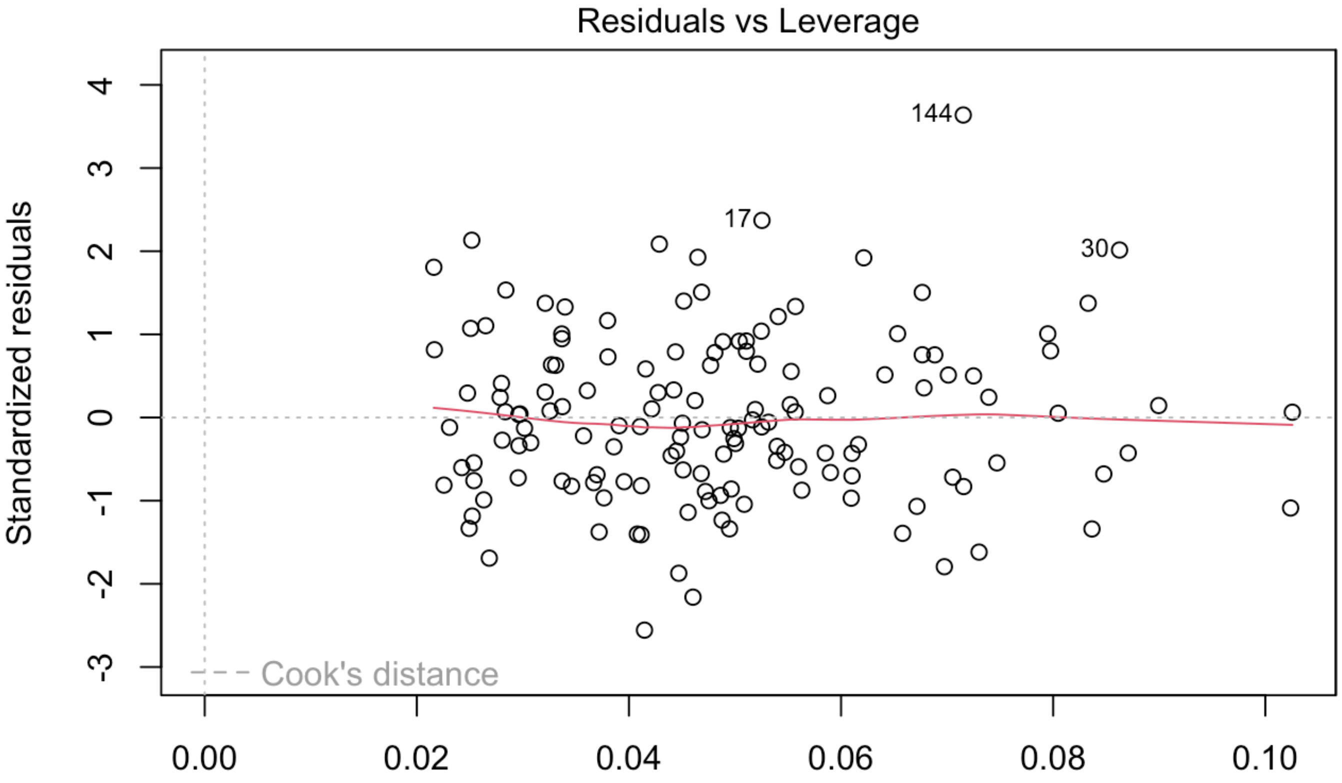


Interpret Residuals vs Fitted:
The spread looks roughly the same, so we are happy with our equal variance assumption.
The scatter seems randomly distributed, so we are happy with our iid assumption.

Interpret Normal Q-Q Plot:
The points fall roughly in a straight line indicating that the residuals are roughly normally distributed.

```
plot(lm2,which=4:5)
```


Cook's distance
Cook's distance
0.00
0.05
0.10
0.15
144
30
17
0
20
40
60
80
100
120
140
Obs. number
lm(Hwt ~ Bwt + Height + Age + as.factor(Sex) + as.factor(Outdoor))


Residuals vs Leverage
Standardized residuals
4
3
2
1
0
-1
-2
-3
144
17
30
Cook's distance
0.00
0.02
0.04
0.06
0.08
0.10

Leverage
lm(Hwt ~ Bwt + Height + Age + as.factor(Sex) + as.factor(Outdoor))

There is no need to worry about outliers as the cook’s distance is below 0.5.
There is also no high leverage points.

```
vif(lm2)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## Bwt                 2.640386  1        1.624926
## Height              1.159160  1        1.076643
## Age                 1.160644  1        1.077332
## as.factor(Sex)      1.608144  1        1.268126
## as.factor(Outdoor) 1.890792  2        1.172630
```

As expected there is no values greater than 10 in the variance inflation factor results.
Therefore we are not worried about collinearity.

```
summary(lm2)
```

```
## 
## Call:
## lm(formula = Hwt ~ Bwt + Height + Age + as.factor(Sex) + as.factor(Outdoor), 
##     data = cat_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.3846 -0.9431 -0.1374  0.8760  4.7397 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          0.34088    1.04651   0.326  0.74512    
## Bwt                  2.80654    0.37859   7.413 1.16e-11 ***
## Height               0.05686    0.02361   2.409  0.01735 *  
## Age                  0.06803    0.03163   2.151  0.03327 *  
## as.factor(Sex)2     -0.44678    0.30474  -1.466  0.14491    
## as.factor(Outdoor)2  0.39979    0.30011   1.332  0.18502    
## as.factor(Outdoor)3  1.25937    0.39680   3.174  0.00186 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.352 on 137 degrees of freedom
## Multiple R-squared:  0.7045, Adjusted R-squared:  0.6916 
## F-statistic: 54.44 on 6 and 137 DF,  p-value: < 2.2e-16
```

Hwt = 0.34088 + 2.80654(Bwt) + 0.05686(Height) + 0.06803(Age) + -0.44678(Sex) + 0.39979(Outdoor) + 1.25937(Outdoor)
For every 1kg in cat’s body weight, cat’s heart weight increases by 2.8g on average keeping all other

variables constant.
For every 1cm in cat’s height, cat’s heart weight increases by 0.05686g on average keeping all other variables constant.
For every 1year in cat’s age, cat’s heart weight increases by 0.06803g on average keeping all other variables constant.
Going from male to female, cat’s heart weight decreases by -0.44678g on average keeping all other variables constant.
Going from always outdoor to outdoor frequently, cat’s heart weight increases by 0.39979g on average keeping all other variables constant.
Going from always outdoor to never outdoor, cat’s heart weight increases by 1.25937g on average keeping all other variables constant.

70% of variation in cat’s heart weight is being explained by the model and it is slightly better than our previous multiple linear regression model.
The residual standard error = 1.352, therefore there is a small spread around the line and it is slightly better than our previous multiple linear regression model.
We can also see that cat’s body weight and going from from always outdoor to never outdoor is very significant.
Cat’s height and age are also significant and it seems like going from male to female is less significant.

F-test
H0: The first multiple linear regression model without sex and outdoor is the preferred model
H1: The last multiple linear regression model with sex and outdoor is the preferred model

```
anova(lm1,lm2)
```

```
## Analysis of Variance Table
##
## Model 1: Hwt ~ Bwt + Height + Age
## Model 2: Hwt ~ Bwt + Height + Age + as.factor(Sex) + as.factor(Outdoor)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    140 270.53
## 2    137 250.47  3    20.054 3.6563 0.01416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value = 0.01416 which is less than 0.05 we reject the null hypothesis.
We can conclude that adding at least sex or outdoor to the model improves the fit of the model.

```
AIC(lm1,lm2)
```

```
##     df      AIC
## lm1  5 509.4544
## lm2  8 504.3632
```

```
BIC(lm1,lm2)
```

```
##     df      BIC
## lm1  5 524.3035
## lm2  8 528.1217
```

Akaike information criterion suggests that the second multiple linear regression model with sex and outdoor is the better model.
Bayesian information criterion suggests that the first multiple linear regression model without sex and outdoor is the better model.

```
#Likelihood Ratio Test
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
lrtest(lm2, lm1)
```

```
## Likelihood ratio test
##
## Model 1: Hwt ~ Bwt + Height + Age + as.factor(Sex) + as.factor(Outdoor)
## Model 2: Hwt ~ Bwt + Height + Age
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -244.18
## 2   5 -249.73 -3 11.091    0.01124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0: The second model and the first model fit the data equally well. i.e. first model is a better fit.
H1: The second model and the first model do not fit the data equally well. i.e. second model is a better fit
As the p-value = 0.01124 which is less than 0.05 we reject the null hypothesis.
This indicates that the second model and the first model do not fit the data equally well. As a result, we should employ the second model.

d) Conclude your overall results. Use your preferred model to predict the fitted values for your response variable and calculate the residual term if you are given the following data:
Sex = 2, Bwt = 2.6, Hwt = 11.2, Height = 23.5, Age = 10, Outdoor = 2
Sex = 1, Bwt = 3.1, Hwt = 12.5, Height = 25.4, Age = 13.2, Outdoor = 2

```
#Hwt = 0.34088 + 2.80654(Bwt) + 0.05686(Height) + 0.06803(Age) + -0.44678(Sex) + 0.39979(Outdoor) + 1.25937(Outdoor)
AnsHwt1 = 0.34088 + (2.80654 * 2.6) + (0.05686 * 23.5) + (0.06803 * 10) + (-0.44678 * 2) + (0.39979 * 2) + (1.25937 * 2)
AnsHwt1
```

```
## [1] 12.07915
```

```
AHwt1 <- 11.2
#Difference between actual and predicted cat's heart weight
Diff1 <- AHwt1-AnsHwt1
Diff1
```

```
## [1] -0.879154
```

```
AnsHwt2 = 0.34088 + (2.80654 * 3.1) + (0.05686 * 25.4) + (0.06803 * 13.2) + (-0.44678 * 1) + (0.39979 * 2) + (1.25937 * 2)
AnsHwt2
```

```
## [1] 14.25493
```

```
AHwt2 <- 12.5
Diff2 <- AHwt2-AnsHwt2
Diff2
```

```
## [1] -1.754934
```