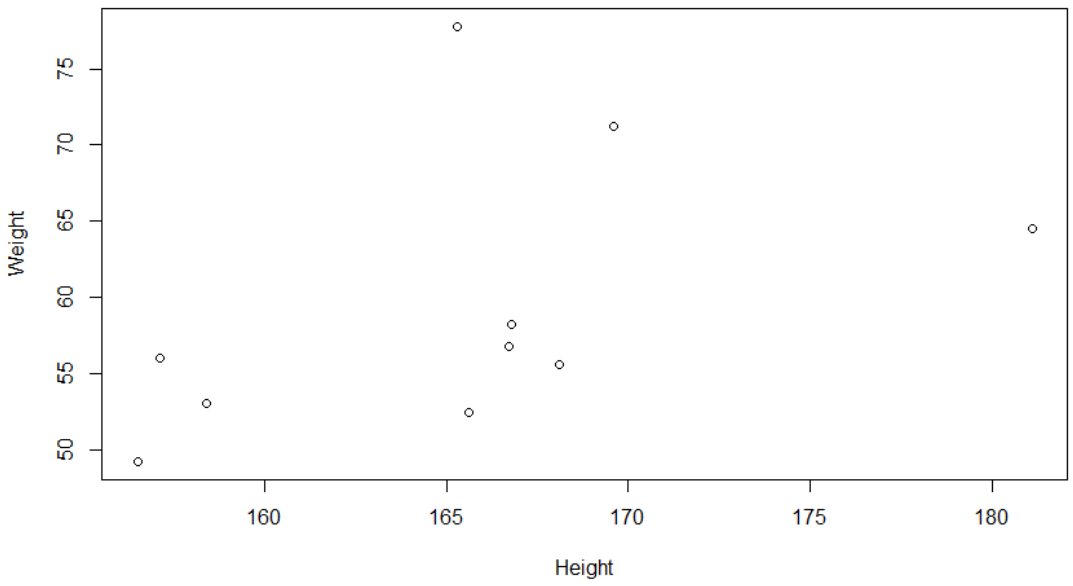
STAT: 8030, Homework: 1

2.1.1 Draw a scatterplot of wt on the vertical axis versus ht on the horizontal axis. On the basis of this plot, does a simple linear regression model make sense for these data? Why or why not?

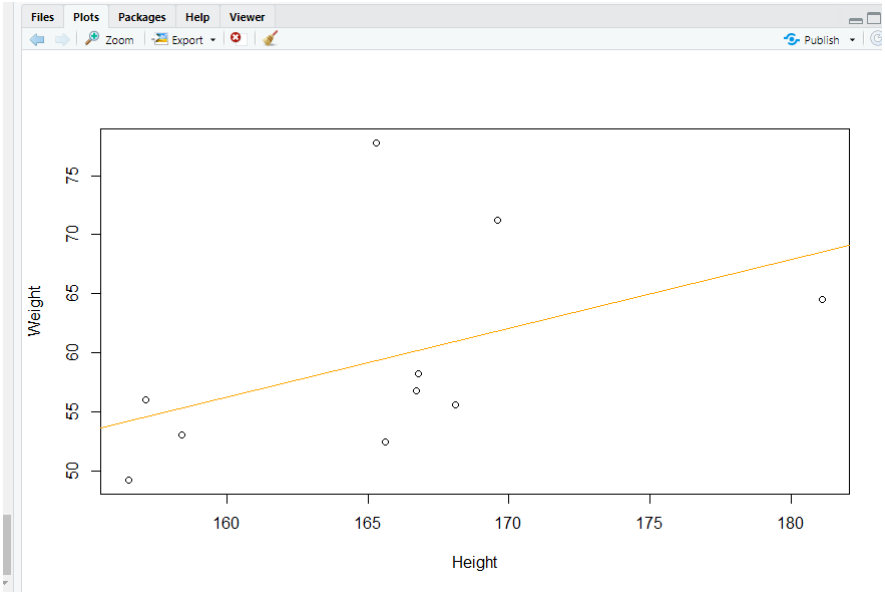
```
> library(alr4)
> data("Htwt")
> Htwt
      ht   wt
1  169.6 71.2
2  166.8 58.2
3  157.1 56.0
4  181.1 64.5
5  158.4 53.0
6  165.6 52.4
7  166.7 56.8
8  156.5 49.2
9  168.1 55.6
10 165.3 77.8
> plot(Htwt$ht, Htwt$wt, xlab="Height", ylab="Weight")
>
```



In our data, height and weight are dependent variables, meaning, one is responsible for the other. Predictor variable in our case here is weight and the response variable being the height. Since the variables are dependent on each other, a simple linear regression model will make sense for this data. However, according to me, the amount of data we have (10) is too less of a data to have an accurate prediction.

2.1.2 Show that x = 165.52, y = 59.47, SXX = 472.08, SYY = 731.96, and SXY = 274.79. Compute estimates of the slope and the intercept for the regression of Y on X. Draw the fitted line on your scatterplot

```
>
> # Calculating mean of x i.e., Height
> mean_of_height <- mean(Htwt$ht)
> mean_of_height
[1] 165.52
>
> # Calculating the mean of y i.e., Weight
> mean_of_weight <- mean(Htwt$wt)
> mean_of_weight
[1] 59.47
>
> # Calculating SXX
> sxx <- sum(Htwt$ht^2)-sum(Htwt$ht)^2/10
> sxx
[1] 472.076
>
> # Calculating SYY
> syy <- sum(Htwt$wt^2)-sum(Htwt$wt)^2/10
> syy
[1] 731.961
>
> # Calculating SXY
> sxy <- sum(Htwt$ht*Htwt$wt)-(sum(Htwt$ht)*sum(Htwt$wt))/10
> sxy
[1] 274.786
>
> # Calculating Slope by considering slope as b1
> b1 <- sxy/sxx
> b1
[1] 0.58208
>
> # Calculating the intercept by considering intercept as b0
> b0 <- mean_of_weight - (b1 * mean_of_height)
> b0
[1] -36.87588
>
> # Drawing fitted line on the scatterplot
> abline(a=b0,b=b1,col='orange')
>
```



2.1.3 Obtain the estimate of σ2 and find the estimated standard errors of ˆ β0 and ˆ β1. Also find the estimated covariance between ˆ β0 and ˆ β1. Compute the t-tests for the hypotheses that β0 = 0 and that β1 = 0 and find the appropriate p-values using two-sided tests.

```
> # 2.1.3
> # Calculating the estimate of variance
> RSS <- syy- (sxy^2/sxx)
> var_sigma <- RSS/8
> var_sigma
[1] 71.5017
>
> # Calculating estimated standard errors
> se_b0 <- sqrt(var_sigma * ((1/10)+(mean_of_height^2)/sxx))
> se_b0
[1] 64.4728
>
> se_b1 <- sqrt(var_sigma/sxx)
> se_b1
[1] 0.3891815
>
> # Calculating estimated covariance between b0 and b1
> cov_b0_b1 <- -var_sigma * (mean_of_height/sxx)
> cov_b0_b1
[1] -25.07003
>
> # Computing t-tests
> t_0 = b0/se_b0
> t_0
[1] -0.5719603
> t_1 = b1/se_b1
> t_1
[1] 1.495652
>
> # Calculating p values for the two sided tests
> 2 * pt(-abs(t_0),8)
[1] 0.5830589
> 2 * pt(-abs(t_1),8)
[1] 0.1731089
>

>
> # Let's cross check with the lm function
> m <- lm(ht~wt, data=Htwt)
> summary(m)

Call:
lm(formula = ht ~ wt, data = Htwt)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1173 -5.0462  0.7166  2.5962 13.6917

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 143.1943    15.0807   9.495 1.25e-05 ***
wt            0.3754     0.2510   1.496    0.173
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.791 on 8 degrees of freedom
Multiple R-squared:  0.2185,    Adjusted R-squared:  0.1208
F-statistic: 2.237 on 1 and 8 DF,  p-value: 0.1731

>
```

2.2.1 The line with equation y = x is shown on this plot as the solid line. What is the key difference between points above this line and points below the line?

According to the graph given, we can take away the following points from the data as below:

- y = x is a solid line which has very little points located on it, in comparison with the points as a whole. This indicates that there are few cities were the price of rice hasn't changed in both 2003 and 2009, but this number of cities is less than the total number of cities from the data.
- We can notice the number of cites above the line are more than the number of cities below the line. This indicates that, the price of rice has increased from the year 2003 to 2009 in majority of the cities it is supplied to, i.e the price of rice in the cities above the line is more in 2009 than what it was in 2003 and the price of rice in the cities below the line is less in 2009 than what it was in 2003. This can be considered as a key difference between points above and below the line.

2.2.2 Which city had the largest increase in rice price? Which had the largest decrease in rice price?

From the Plot we can see "Vilnius" is the city that had the largest increase in rice price (which is around 75-25 = 50) and "Mumbai" is the city that had the largest decrease in the rice price (which is around 100 – 38 = 62).

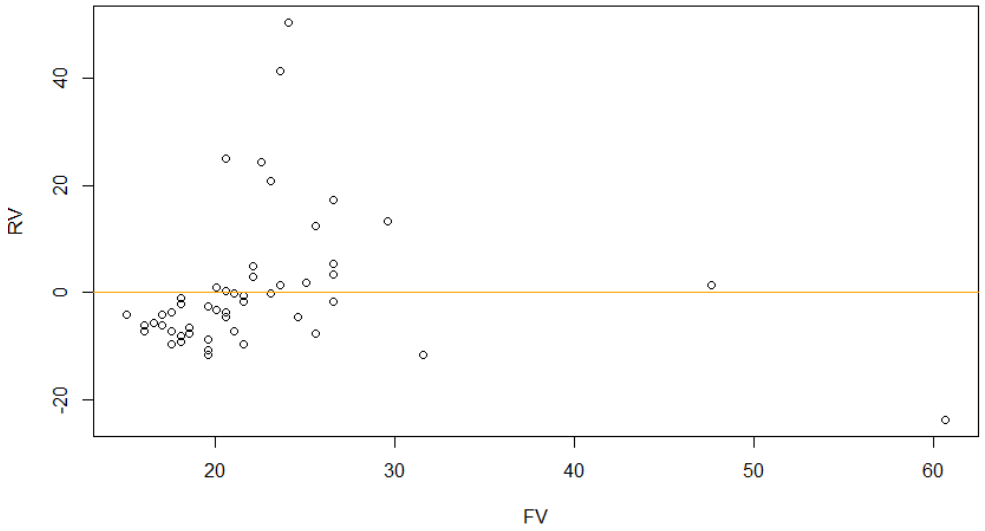
2.2.3 The ols line is shown on the figure as a dashed line, and evidently $\hat{\beta}1 < 1$. Does this suggest that prices are lower in 2009 than in 2003? Explain your answer.

The plot clearly suggests that the prices of rice in 2009 are more than that of 2003 in majority of the cities. This can be inferred by looking at the cities with respect to the y = x line. The slope of the ols line here is clearly less than 1, in our case, these two values plotted (2003 and 2009) cannot be compared using this approach to model the relationship between two variables because they are independent of each other. Hence, we cannot say that prices are lower in 2009 than in 2003.

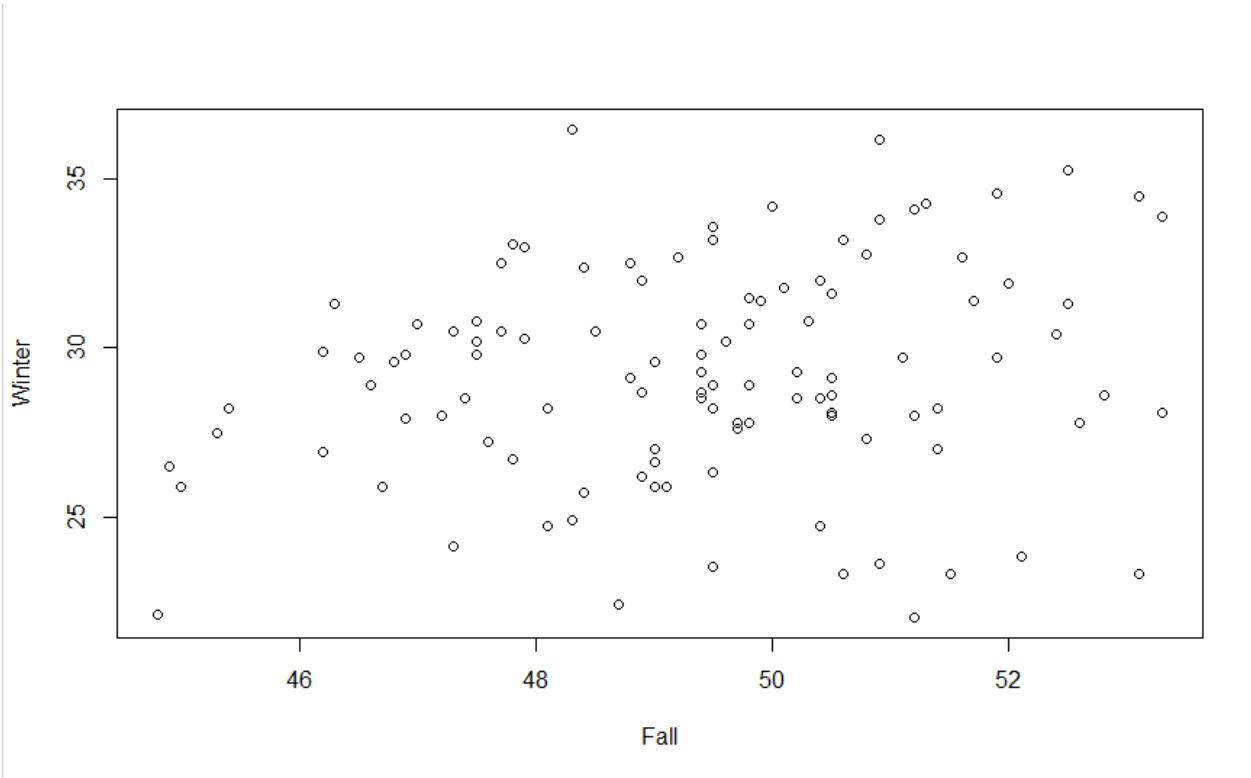
2.2.4 Give two reasons why fitting simple linear regression to the figure in this problem is not likely to be appropriate.

- First reason being that the variables plotted in this graph are independent of each other and simple linear regression works best on the graphs having predictor variable and response variable (i.e. being dependent on each other).

- Second reason being, the vertical distance between the points and the line is more than what is anticipated (from the residual plot below). Linear regression cannot be trusted to be appropriate unless the vertical distance of the points from line is very less.

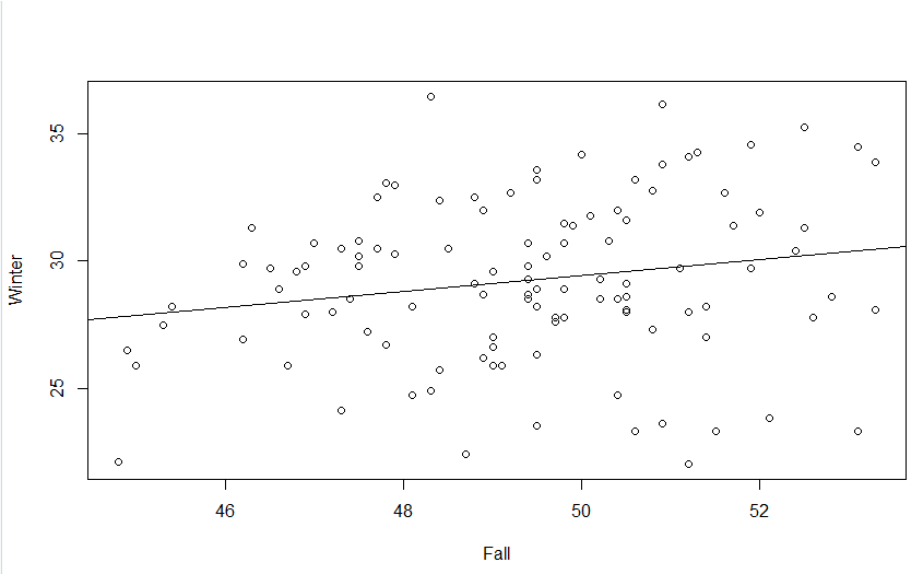


2.6.1 Draw a scatterplot of the response versus the predictor, and describe any pattern you might see in the plot.



From the following most of the scattered points over the graph, we can see that, increase in average temperature in fall over the years increases the average temperature in winter over the years

2.6.2 Use statistical software to fit the regression of the response on the predictor. Add the fitted line to your graph. Test the slope to be 0 against a two-sided alternative, and summarize your results.

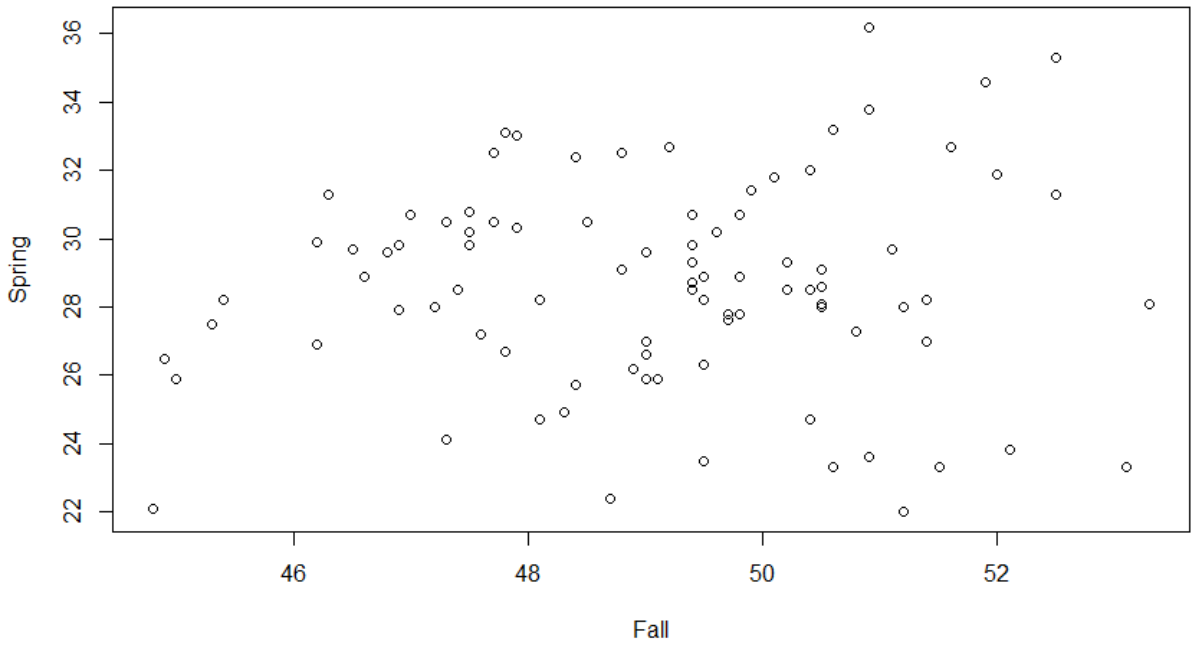


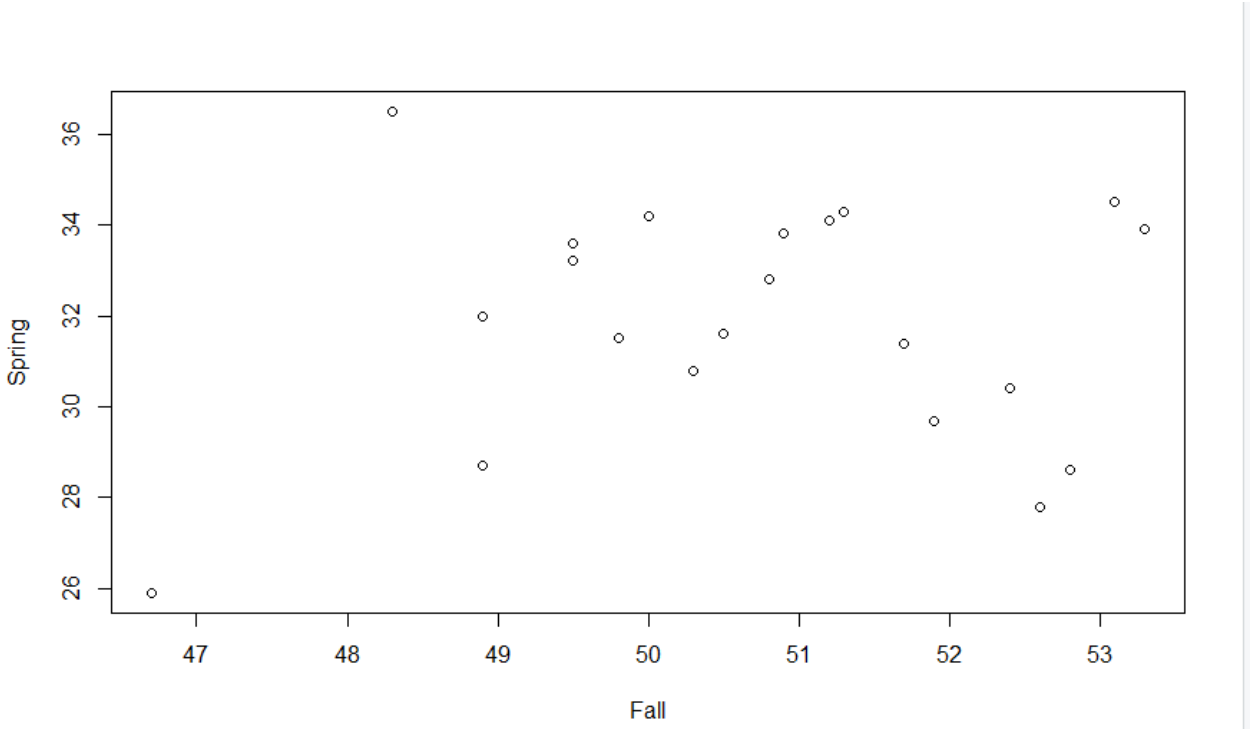

```
> #2.6.2
> lmtemp <- lm(winter ~ fall, data=ftcollinstemp)
> lmtemp

Call:
lm(formula = winter ~ fall, data = ftcollinstemp)

Coefficients:
(Intercept)         fall
    13.7843       0.3132

> abline(lmtemp, col = "black")
>
> summary(lmtemp)

Call:
lm(formula = winter ~ fall, data = ftcollinstemp)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8186 -1.7837 -0.0873  2.1300  7.5896

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7843     7.5549   1.825   0.0708 .
fall          0.3132     0.1528   2.049   0.0428 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.179 on 109 degrees of freedom
Multiple R-squared:  0.0371,    Adjusted R-squared:  0.02826
F-statistic:   4.2 on 1 and 109 DF,  p-value: 0.04284
```

From the summary we observe above, we can get the following:

```
>
> beta0 <- 13.7834
> beta1 <- 0.3132
>
> s0 <- 7.5549
> s1 <- 0.1528
>
> t0 <- beta0/s0
> t0
[1] 1.824432
> t1 <- beta1/s1
> t1
[1] 2.049738
>
> 2 * pt(abs(t0),109, lower.tail = FALSE) #P values
[1] 0.07082599
> 2 * pt(abs(t1),109, lower.tail = FALSE)
[1] 0.04279025
```

2.6.4 Divide the data into 2 time periods, an early period from 1900 to 1989, and a late period from 1990 to 2010. You can do this using the variable year in the data file. Are the results different in the two time periods?





From the graphs of two time periods, I can say the average temperature in spring increase as the average temperature in fall increases over time during the 1900-1989 period.

Whereas, first, during the initial years, the average temperature increases in spring as the average temperature in fall increases and then in later years the average temperature in spring decreases as the average temperature in fall increases over time during the 1990-2010 period.

2.8.1 What is the meaning of the parameter α?

Here, this parameter represents the value of y when x is equal to its mean value.

2.14.1 Using the Heights data, create a construction set by selecting approximately 2/3 of the rows of the data file at random. The remaining 1/3 of the rows will comprise the validation set.

```
# 2.14

data("Heights")
Heights

#2.14.1
n <- 458 #1/3 of the rows
vrows <- sample(1:nrow(Heights),n,replace=F)
vset <- Heights[vrows,]
cset <- Heights[-vrows,]
```

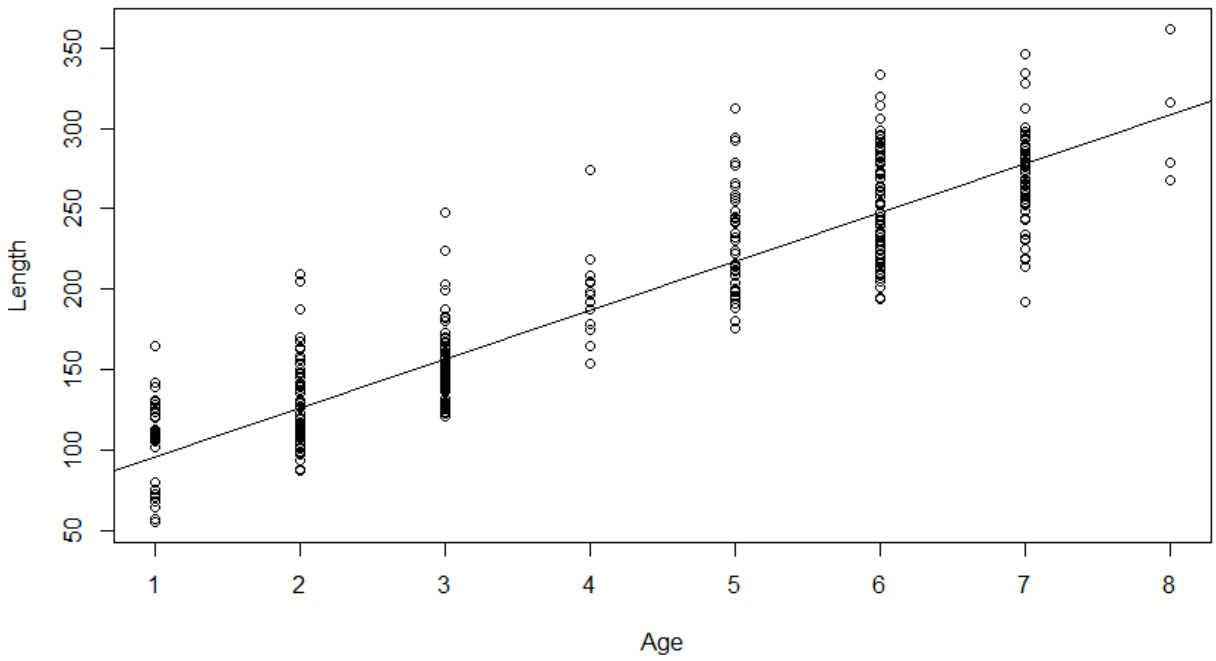
2.14.2 Find the prediction error for the construction set.

```
> #2.14.2
>
> lmHeights <- lm(mheight~dheight, data = cset)
> cp <- predict(lmHeights, newdata = data.frame(cset["dheight"])) #Prediction calculation
>
> ASR <- mean((cset$mheight-cp)^2) #Average squre residual
> ASR
[1] 4.331781
>
> APE <- sqrt(ASR) #Average prediction error
> APE
[1] 2.081293
```

2.14.3 Find the prediction error for the validation set. Think ahead of time: which error do you expect to be larger?

```
> #2.14.3
>
> vp <- predict(lmHeights, newdata = data.frame(vset["dheight"])) #Prediction calculation
>
> ASRV <- mean((vset$mheight-vp)^2) #Average square residual
> ASRV
[1] 3.969833
>
> APEV <- sqrt(ASRV) #Average prediction error
> APEV
[1] 1.992444
```

2.15 Smallmouth bass (Data file: wblake)



2.15.1 Using the West Bearskin Lake smallmouth bass data in the file wblake, obtain 95% intervals for the mean length at ages 2, 4, and 6 years.

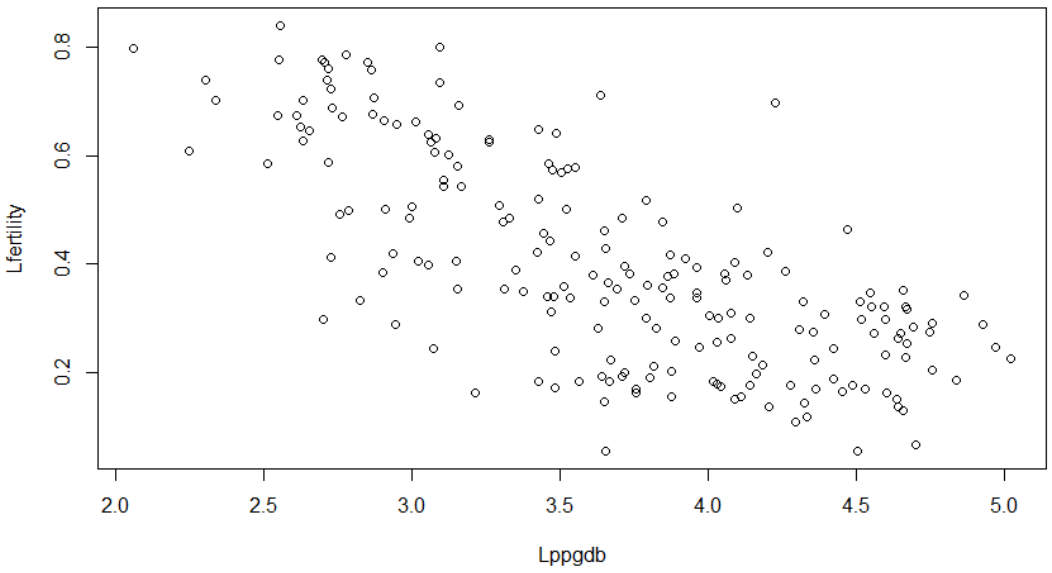
```
> #2.15.1
> predict(m,data.frame(Age=c(2,4,6)), interval="confidence", level=0.95)
       fit      lwr      upr
1 126.1749 122.1643 130.1856
2 186.8227 184.1217 189.5237
3 247.4705 243.8481 251.0929
> ppgdb
```

2.15.2 Obtain a 95% interval for the mean length at age 9. Explain why this interval is likely to be untrustworthy.

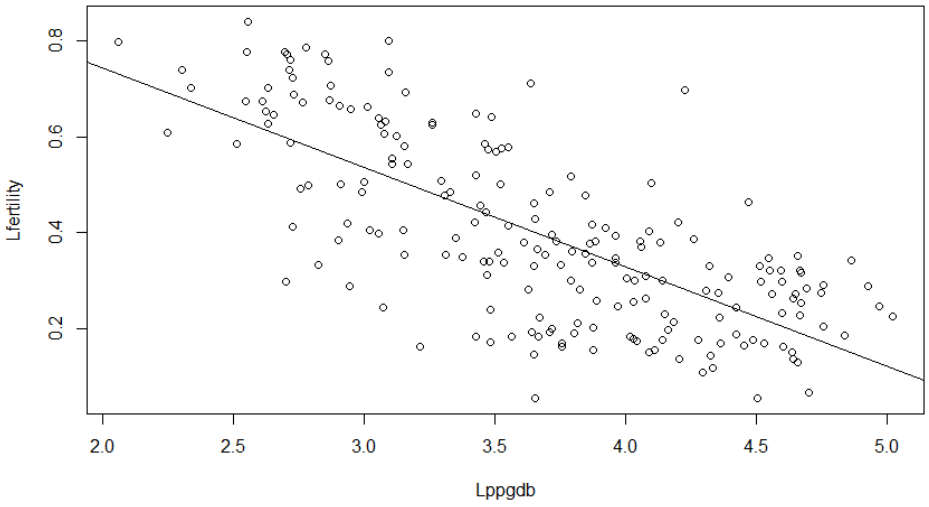
```
> predict(m,data.frame(Age=c(9)), interval="confidence", level=0.95)
       fit      lwr      upr
1 338.4422 331.4231 345.4612
```

The mean length at the age of 9 falls out of range of the graph plotted. Which makes it quite untrustworthy, since it cannot be inferred from the plotted graph

2.16.1 Use a software package to compute the simple linear regression model corresponding to the graph in Problem 1.1.3.



2.16.2 Draw a graph of log(fertility) versus log(ppgdp), and add the fitted line to the graph.



2.16.3 Test the hypothesis that the slope is 0 versus the alternative that it is negative (a one-sided test). Give the significance level of the test and a sentence that summarizes the result.

```
> #2.16.3
> summary(m)

Call:
lm(formula = LFertility ~ LPPgdb, data = UN11)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34669 -0.09398  0.01159  0.10173  0.41517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.15762    0.05236   22.11   <2e-16 ***
LPPgdb      -0.20715    0.01401  -14.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1334 on 197 degrees of freedom
Multiple R-squared:  0.526,     Adjusted R-squared:  0.5236
F-statistic: 218.6 on 1 and 197 DF,  p-value: < 2.2e-16

>
> NOR <- 199 #Number of rows
>
> sbeta1 <- -0.20715
>
> sesbeta1 <- 0.01401 #Standard error
>
> tt <- sbeta1/sesbeta1 #T-test
> tt
[1] -14.78587
>
> pv <- pt(-abs(tt), df= NOR-2, lower.tail = T) #P-value
> pv
[1] 4.506246e-34
```

2.16.4 Give the value of the coefficient of determination, and explain its meaning.

```
> #2.16.4
> summary(m)

Call:
lm(formula = LFertility ~ LPPgdb, data = UN11)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34669 -0.09398  0.01159  0.10173  0.41517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.15762    0.05236   22.11   <2e-16 ***
LPPgdb      -0.20715    0.01401  -14.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1334 on 197 degrees of freedom
Multiple R-squared:  0.526,     Adjusted R-squared:  0.5236
F-statistic: 218.6 on 1 and 197 DF,  p-value: < 2.2e-16
```

Looking at the summary of the following data file, we can see that it's value is 0.52 and there is 52% chance that it can make an impact on the response