HW3

Thursday, May 29, 2025 11:41 AM

Problem 1. The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test scores (x). The data for this study can be found on Canvas, you might find the function read.table("C: file address here GPAdata.txt") useful in reading the data into R. Note, in the analysis that follows you are expected to write all of your own functions for conducting the analysis, and provide them in your write up, of course checking your solutions with the built in functions is a good idea.

(a) Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function. Be sure to interpret the meaning of these estimates in the context of this problem.

(b) Plot the estimated regression function, along with the data. Does the estimated regression function to fit the data well. In addition provide a 95% confidence band around your estimated regression function.

(c) Obtain a point estimate of the mean freshman GPA for students with ACT test scores $x = 30$, also provide a 95% CI and 95% PI for this estimate, with their appropriate interpretation.

(d) Obtain a 99% CI for both least squares estimates, an interpret.

(e) Conduct hypothesis tests about both of the least squares estimates, using a t-statistic at the $\alpha = 0.01$ significance level, be sure to report a p-value. Interpret your results and comment on how these relate to the CIs you constructed in part (d).

(f) Set up the ANOVA table.

(g) Conduct an F-test, at the $\alpha = 0.01$ significance level, of whether or not $\beta_1 = 0$. Be sure to state the null and alternative hypothesis, the decision rule, and interpret your results. Comment on how this relates to one of the tests you performed in part (e).

(h) Obtain R^2 and r, and interpret.

(i) Prepare a box plot for the ACT scores. Are there any noteworthy features?

(j) Plot the residuals against the fitted values, what does this tell you.

(k) Prepare a QQ-plot of the residuals, and discuss appropriately.

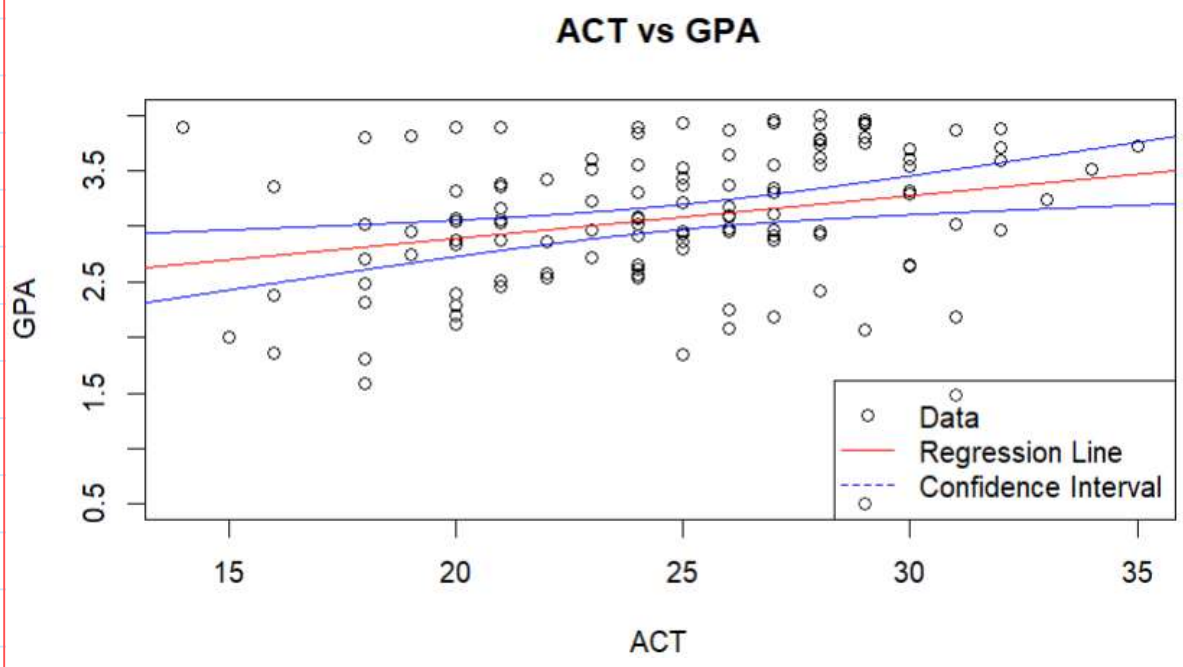
(l) Conduct the Brown Forsythe test to determine whether the assumption of constant error variance is appropriate, use $X = 26$ to split the data into two separate classes.

```
> gpadata<-read.table("C:\\Users\\jacob\\OneDrive\\Documents\\Stats\\8050\\HW3\\GPAdata.txt")
> act<-gpadata[,2]
> gpa<-gpadata[,1]
> actbar<-mean(act)
```

```
> gpadata<-read.table("C:\\Users\\jacob\\OneDrive\\Documents\\Stats\\8050\\HW3\\GPAdata.txt")
> act<-gpadata[,2]
> gpa<-gpadata[,1]
> actbar<-mean(act)
> gpabar<-mean(gpa)
> b1<-sum((act-actbar)*(gpa-gpabar))/sum((act-actbar)^2)
> b0<-gpabar-b1*actbar
> cat("GPA_i=",b0,"+",b1,"ACT_i")
GPA_i= 2.114049 + 0.03882713 ACT_i
```

Since we do not have a lot of data around 0, b0 does not mean a lot. Similarly, since b1 is small, it might not be statistically meaningful

```
> a<-0.05
> n<-length(act)
> actvals<-seq(min(act)-1,max(act)+1,length.out=500)
> MSE<-sum((gpa-b0-b1*act)^2)/(n-2)
> flow<-function(x){b0+b1*x-qt(1-a/2,n-2)*sqrt(MSE*(1/n+(x-actbar)^2/(sum((act-actbar)^2))))}
> fhigh<-function(x){b0+b1*x+qt(1-a/2,n-2)*sqrt(MSE*(1/n+(x-actbar)^2/(sum((act-actbar)^2))))}
> plot(act,gpa,col="black",xlab="ACT",ylab="GPA",main="ACT vs GPA")
> abline(b0,b1,col="red")
> lines(actvals,flow(actvals),col="blue")
> lines(actvals,fhigh(actvals),col="blue")
> legend("bottomright",
+        legend = c("Data", "Regression Line", "Confidence Interval"),
+        col = c("black", "red", "blue"),
+        bg = adjustcolor("white", alpha.f = 0.1),
+        lty = c(NA, 1, 2),
+        pch = c(1, NA, NA),
+        lwd = c(NA, 1, 1))
```




```
> gpa30<-b0+b1*30
> YhCIradius<-qt(1-a/2,n-2)*sqrt(MSE*(1/n+(30-actbar)^2/(sum((act-actbar)^2))))
> PIradius<-qt(1-a/2,n-2)*sqrt(MSE*(1+1/n+(30-actbar)^2/(sum((act-actbar)^2))))
> cat("CI=(",gpa30-YhCIradius,",",gpa30+YhCIradius,")")
CI=( 3.104246 , 3.453481 )
> cat("PI=(",gpa30-PIradius,",",gpa30+PIradius,")")
PI=( 2.032612 , 4.525114 )
```

We know that the PI will be wider as the radius is larger than the CI.

```
> a<-0.01
> b0CIradius<-qt(1-a/2,n-2)*sqrt(MSE*(1/n+actbar^2/(sum((act-actbar)^2))))
> b1CIradius<-qt(1-a/2,n-2)*sqrt(MSE/(sum((act-actbar)^2)))
> cat("CI=(",b0-b0CIradius,",",b0+b0CIradius,")")
CI=( 1.273903 , 2.954196 )
> cat("CI=(",b1-b1CIradius,",",b1+b1CIradius,")")
CI=( 0.005385614 , 0.07226864 )
```

Since we do not have a lot of data near 0, the CI for b0 does not mean a lot, because b0 does not mean a lot. I think that we are pretty sure b1 is not 0, but it is pretty small

\beta_0 H0: \beta_0=0
Ha: \beta_0\neq 0

\beta_1 H0: \beta_1=0
Ha: \beta_1\neq 0

```
> t0<-b0/sqrt(MSE*(1/n+actbar^2/(sum((act-actbar)^2))))
> t1<-b1/sqrt(MSE/(sum((act-actbar)^2)))
> p0<-pt(t0,n-2)
> p1<-pt(t1,n-2)
> cat("b0 pvalue=",2*min(p0,1-p0))
b0 pvalue= 1.30445e-09> cat("b1 pvalue=",2*min(p1,1-p1))
b1 pvalue= 0.002916604
```

Since \beta_0 pvlaue<0.01, we fail to reject H0 that \beta_0 is 0
However, \beta_1 pvalue>0.01, so we reject H0 that \beta_1 is 0

```
> SSR<-sum((b0+b1*act-gpabar)^2)
> atable<-data.frame(Source=c("SS","df","MS","F*","pvalue"),
+                    Regression=c(SSR,1,SSR,SSR/MSE,1-pf(SSR/MSE,1,n-2)),
+                    Error=c(MSE*(n-2),n-2,MSE,NA,NA),
+                    Total=c(SSR+MSE*(n-2),n-1,NA,NA,NA))
> print(t(atable))
           [,1]          [,2]          [,3]          [,4]          [,5]
Source     "SS"          "df"          "MS"          "F*"          "pvalue"
Regression "3.587845899" "1.000000000" "3.587845899" "9.240242702" "0.002916604"
Error      " 45.8176078" "118.0000000" "  0.3882848" NA            NA
Total      " 49.40545"   "119.00000"   NA            NA            NA
```

F Test H0: \beta_1=0
Ha: \beta_1\neq 0

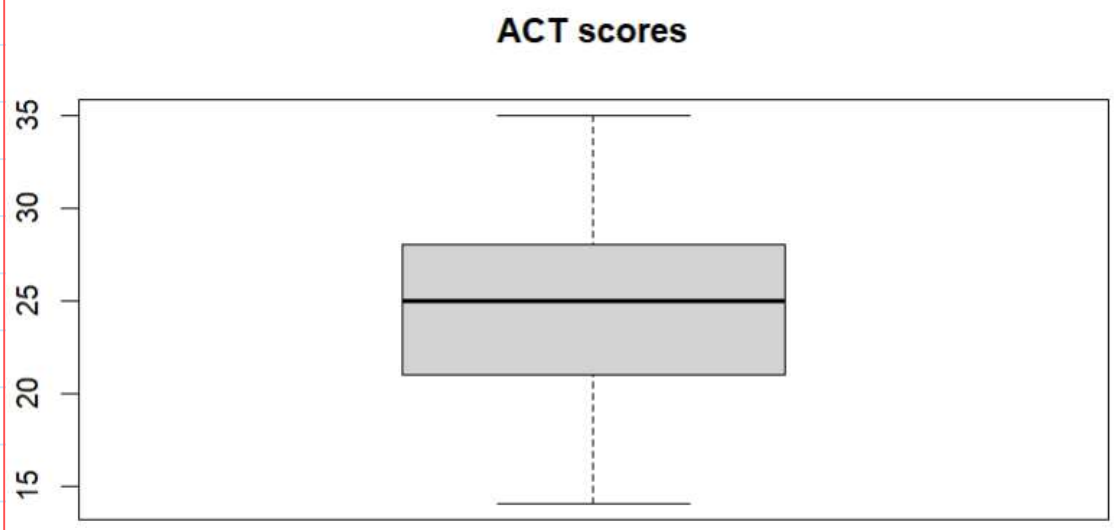
```
> fb1<-(b1/sqrt(MSE/(sum((act-actbar)^2))))^2
> fp<-pf(fb1,1,n-2)
> cat("b1=0? pvalue=",1-fp)
b1=0? pvalue= 0.002916604
```

This pvalue is the same pvalue that we got in the t test done before. This makes sense. Thus we have the same result

```
> R2<-1-(MSE*(n-2))/(SSR+MSE*(n-2))
> r<-sign(b1)*sqrt(R2)
> cat("R^2=",R2,"and r=",r)
R^2= 0.07262044 and r= 0.2694818
```

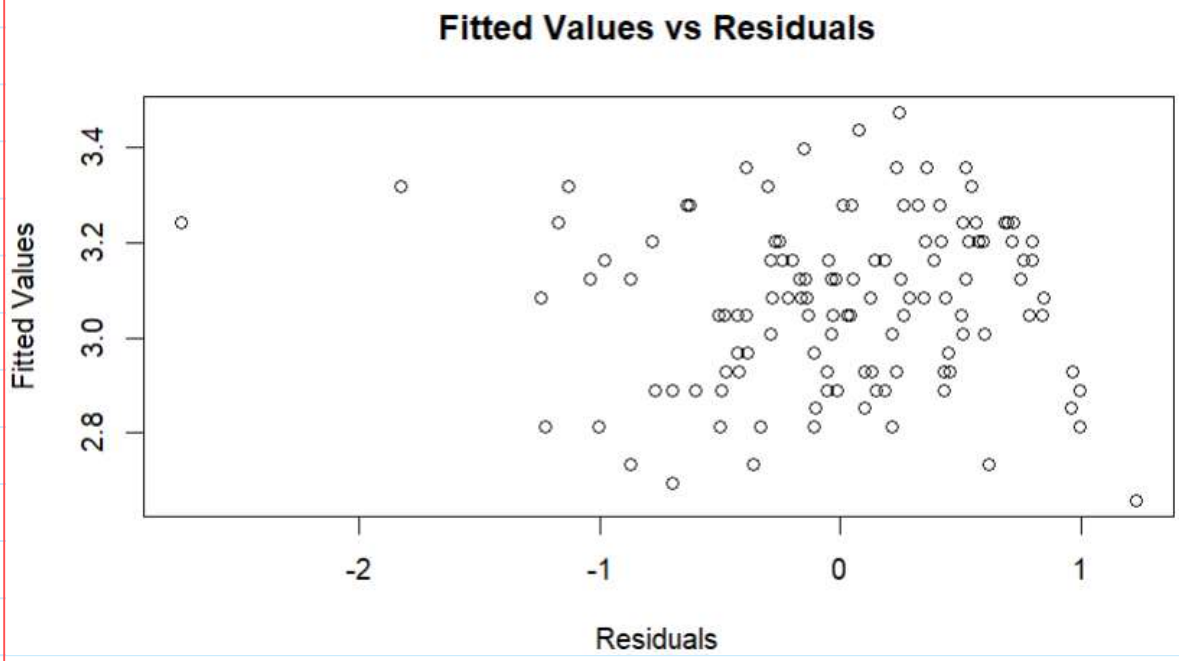
This tells us that GPAs are not very linearly correlated to ACT scores. If there is any relation, it is positive as r>0.

```
> boxplot(act,main="ACT scores")
```



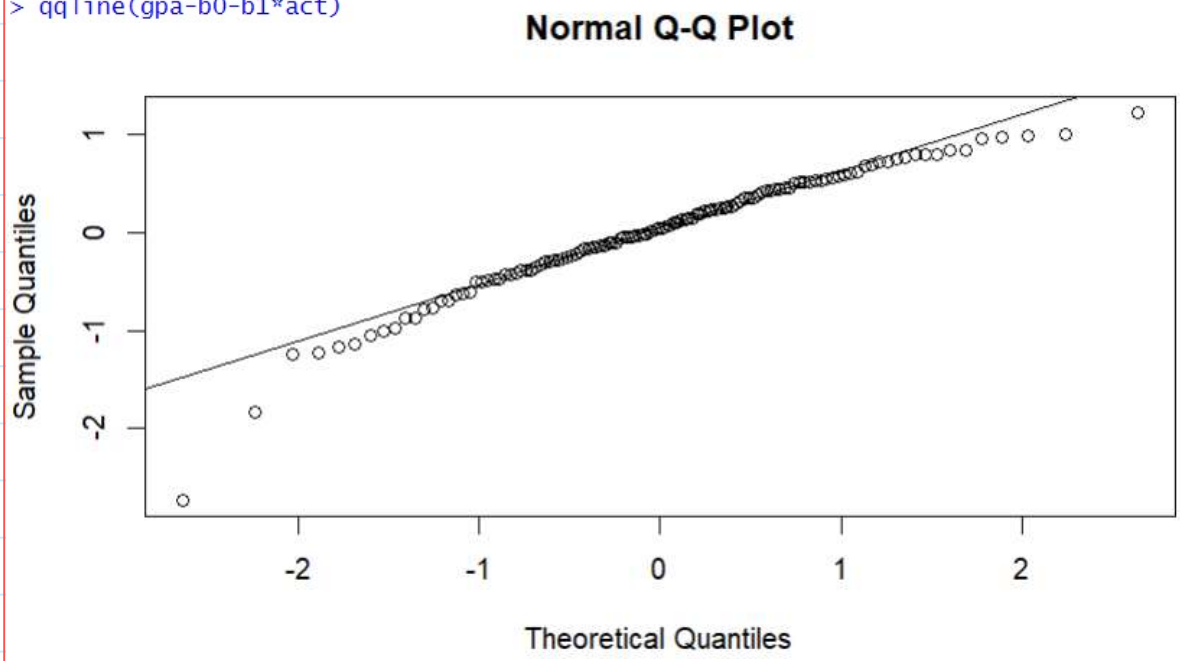
There does not appear to be any deviation from a standard spread

```
> plot(gpa-b0-b1*act,b0+b1*act,main="Fitted Values vs Residuals",
+      xlab="Residuals",ylab="Fitted Values")
```



It seems pretty random and skewed to the right. This skew makes sense as \beta_0 does not have a good interpretation

```
> qqnorm(gpa-b0-b1*act)
> qqline(gpa-b0-b1*act)
```



This seems pretty normal. Near the edges, we are falling off, which is not unordinary

H0: \sigma_{act\leq 26}=\sigma_{act>26}
Ha: \sigma_{act\leq 26}\neq\sigma_{act>26}

```
> threshold<-26
> less<-gpadata[gpadata[,2]<=threshold,]
> great<-gpadata[gpadata[,2]>threshold,]
> actless<-less[,2]
> actgreat<-great[,2]
> gpaless<-less[,1]
> gpagreat<-great[,1]
> resless<-gpaless-b0-b1*actless
> resgreat<-gpagreat-b0-b1*actgreat
> medless<-median(resless)
> medgreat<-median(resgreat)
> d1<-abs(resless-medless)
> d2<-abs(resgreat-medgreat)
> tbf<-(mean(d1)-mean(d2))/sqrt((sum((d1-mean(d1))^2)+sum((d2-mean(d2))^2))*
+                                (1/length(gpaless)+1/length(gpagreat))/(n-2))
> pbf<-pt(tbf,n-2)
> cat("Pvalue for BF=",2*min(pbf,1-pbf))
Pvalue for BF= 0.2817543
```

The Pvalue>0.05 so we reject H0 that the variation in the GPA is the same for those above and below a score of 26 on the ACT

Problem 2. A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10 year age group. Treating age as a predictor variable and muscle mass as your response, complete the following analysis. The data for this study can be found on Canvas. Again, you should use the function that you have written in this analysis.

(a) Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function. Be sure to interpret the meaning of these estimates in the context of this problem.

(b) Plot the estimated regression function, along with the data. Does the estimated regression function to fit the data well. In addition provide a 95% confidence band around your estimated regression function.

(c) Obtain a point estimate of the mean muscle mass for women age $x = 55$, also provide a 95% CI and 95% PI for this estimate, with their appropriate interpretation.

(d) Obtain a 99% CI for both least squares estimates, an interpret.

(e) Conduct hypothesis tests about both of the least squares estimates, using a t-statistic at the $\alpha = 0.01$ significance level, be sure to report a p-value. Interpret your results and comment on how these relate to the CIs you constructed in part (d).

(f) Set up the ANOVA table.

(g) Conduct an F-test, at the $\alpha = 0.01$ significance level, of whether or not $\beta_1 = 0$. Be sure to state the null and alternative hypothesis, the decision rule, and interpret your results. Comment on how this relates to one of the tests you performed in part (e).

(h) Obtain R^2 and r, and interpret.

(i) Prepare a box plot for ages. Are there any noteworthy features?

(f) Plot the residuals against the fitted values, what does this tell you.

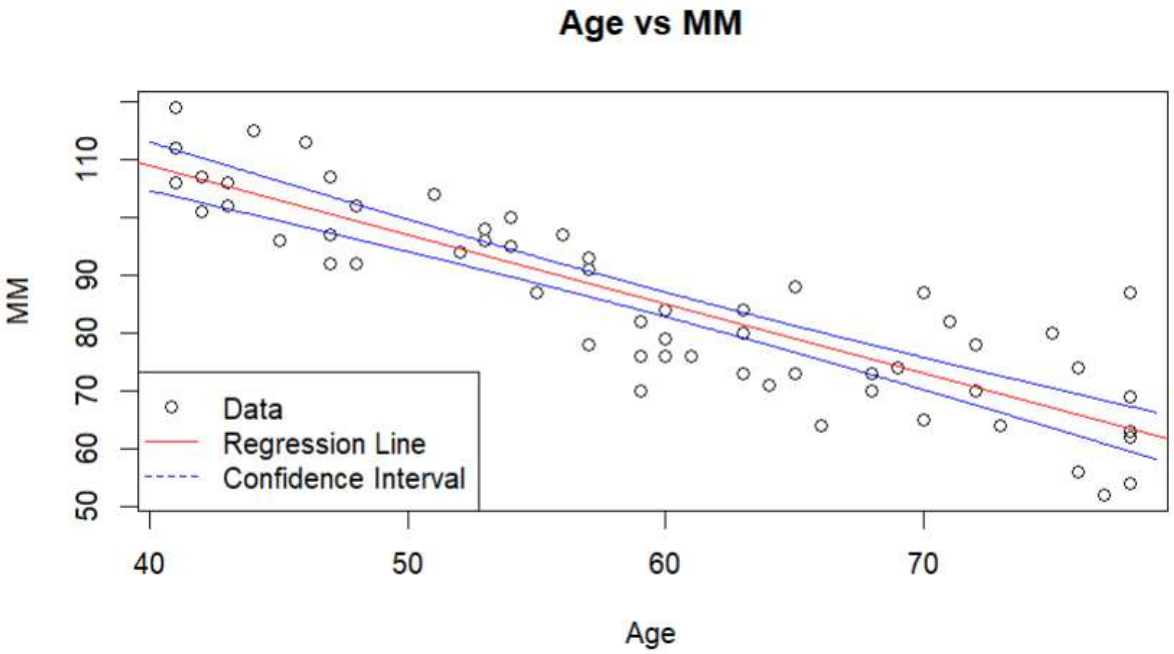
(g) Prepare a QQ-plot of the residuals, and discuss appropriately.

(l) Conduct the Brown Forsythe test to determine whether the assumption of constant error variance is appropriate, use $X = 60$ to split the data into two separate classes.

```
> muscledata<-read.table("C:\\Users\\jacob\\OneDrive\\Documents\\Stats\\8050\\HW3\\Muscledata.txt")
> age<-muscledata[,2]
> mm<-muscledata[,1]
> agebar<-mean(age)
> mmbar<-mean(mm)
> b1<-sum((age-agebar)*(mm-mmbar))/sum((age-agebar)^2)
> b0<-mmbar-b1*agebar
> cat("MM_i=",b0,"+",b1,"Age_i")
MM_i= 156.3466 + -1.189996 Age_i
```

Since we do not have data of babies, our b0 does not have a lot of meaning. b1 is negative which suggests a negative linear relation.

```
> a<-0.05
> n<-length(age)
> agevals<-seq(min(age)-1,max(age)+1,length.out=500)
> MSE<-sum((mm-b0-b1*age)^2)/(n-2)
> flow<-function(x){b0+b1*x-qt(1-a/2,n-2)*sqrt(MSE*(1/n+(x-agebar)^2/(sum((age-agebar)^2))))}
> fhigh<-function(x){b0+b1*x+qt(1-a/2,n-2)*sqrt(MSE*(1/n+(x-agebar)^2/(sum((age-agebar)^2))))}
> plot(age,mm,col="black",xlab="Age",ylab="MM",main="Age vs MM")
> abline(b0,b1,col="red")
> lines(agevals,flow(agevals),col="blue")
> lines(agevals,fhigh(agevals),col="blue")
> legend("bottomleft",
+        legend = c("Data", "Regression Line", "Confidence Interval"),
+        col = c("black", "red", "blue"),
+        bg = adjustcolor("white", alpha.f = 0.1),
+        lty = c(NA, 1, 2),
+        pch = c(1, NA, NA),
+        lwd = c(NA, 1, 1))
```



```
> mm55<-b0+b1*55
> YhCIradius<-qt(1-a/2,n-2)*sqrt(MSE*(1/n+(55-agebar)^2/(sum((age-agebar)^2))))
> PIradius<-qt(1-a/2,n-2)*sqrt(MSE*(1+1/n+(55-agebar)^2/(sum((age-agebar)^2))))
> cat("CI=(",mm55-YhCIradius,",",mm55+YhCIradius,")")
CI=( 88.60104 , 93.19258 )
> cat("PI=(",mm55-PIradius,",",mm55+PIradius,")")
PI=( 74.37613 , 107.4175 )
```

We know that the PI will be wider as the radius is larger than the CI.

```
> a<-0.01
> b0CIradius<-qt(1-a/2,n-2)*sqrt(MSE*(1/n+agebar^2/(sum((age-agebar)^2))))
> b1CIradius<-qt(1-a/2,n-2)*sqrt(MSE/(sum((age-agebar)^2)))
> cat("CI=(",b0-b0CIradius,",",b0+b0CIradius,")")
CI=( 141.6658 , 171.0273 )
> cat("CI=(",b1-b1CIradius,",",b1+b1CIradius,")")
CI=( -1.430217 , -0.9497743 )
```

We know b0 does not have a good interpretation. We are pretty sure that there will be a negative relation between age and muscle mass from this CI

\beta_0 H0: \beta_0=0
Ha: \beta_0\neq 0

\beta_1 H0: \beta_1=0
Ha: \beta_1\neq 0

```
> t0<-b0/sqrt(MSE*(1/n+agebar^2/(sum((age-agebar)^2))))
> t1<-b1/sqrt(MSE/(sum((age-agebar)^2)))
> p0<-pt(t0,n-2)
> p1<-pt(t1,n-2)
> cat("b0 pvalue=",2*min(p0,1-p0))
b0 pvalue= 0
> cat("b1 pvalue=",2*min(p1,1-p1))
b1 pvalue= 4.123987e-19
```

We are very sure that each \beta is nonzero even though \beta_0 does not mean a lot. We fail to reject in both cases.

```
> SSR<-sum((b0+b1*age-mmbar)^2)
> atable<-data.frame(Source=c("SS","df","MS","F*","pvalue"),
+                    Regression=c(SSR,1,SSR,SSR/MSE,1-pf(SSR/MSE,1,n-2)),
+                    Error=c(MSE*(n-2),n-2,MSE,NA,NA),
+                    Total=c(SSR+MSE*(n-2),n-1,NA,NA,NA))
> print(t(atable))
           [,1]          [,2]          [,3]          [,4]         [,5]
Source     "SS"          "df"          "MS"          "F*"         "pvalue"
Regression "11627.486"   "    1.000"   "11627.486"   "  174.062"  "    0.000"
Error      "3874.44750"  "  58.00000"  "  66.80082"  NA           NA
Total      "15501.93"    "   59.00"    NA            NA           NA
```

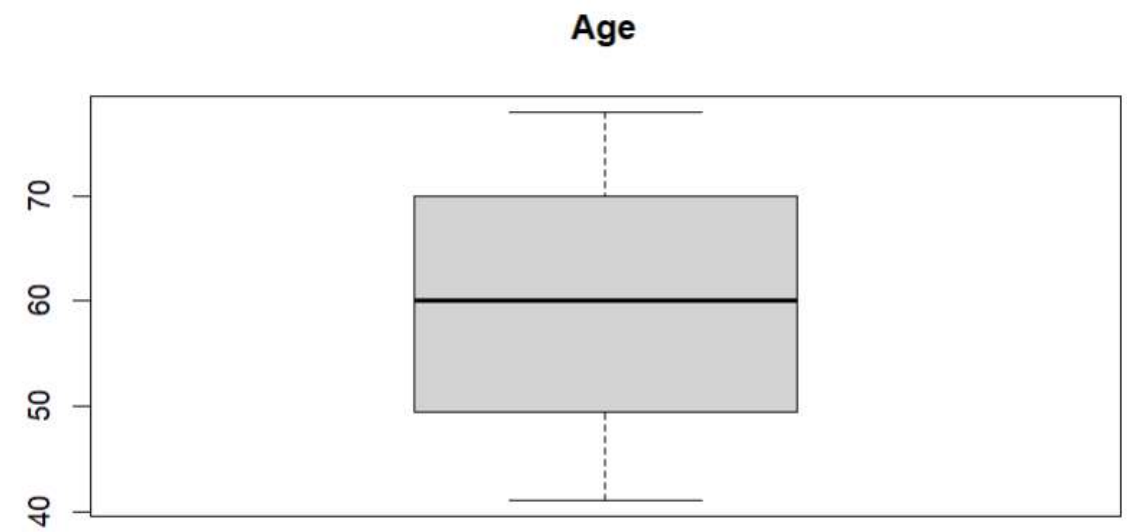
F Test H0: \beta_1=0
Ha: \beta_1\neq 0

```
> fb1<-(b1/sqrt(MSE/(sum((age-agebar)^2))))^2
> fp<-pf(fb1,1,n-2)
> cat("b1=0? pvalue=",1-fp)
b1=0? pvalue= 0
```

When taking round off error into consideration, this is the same value for all intents and purposes and we have the same result.

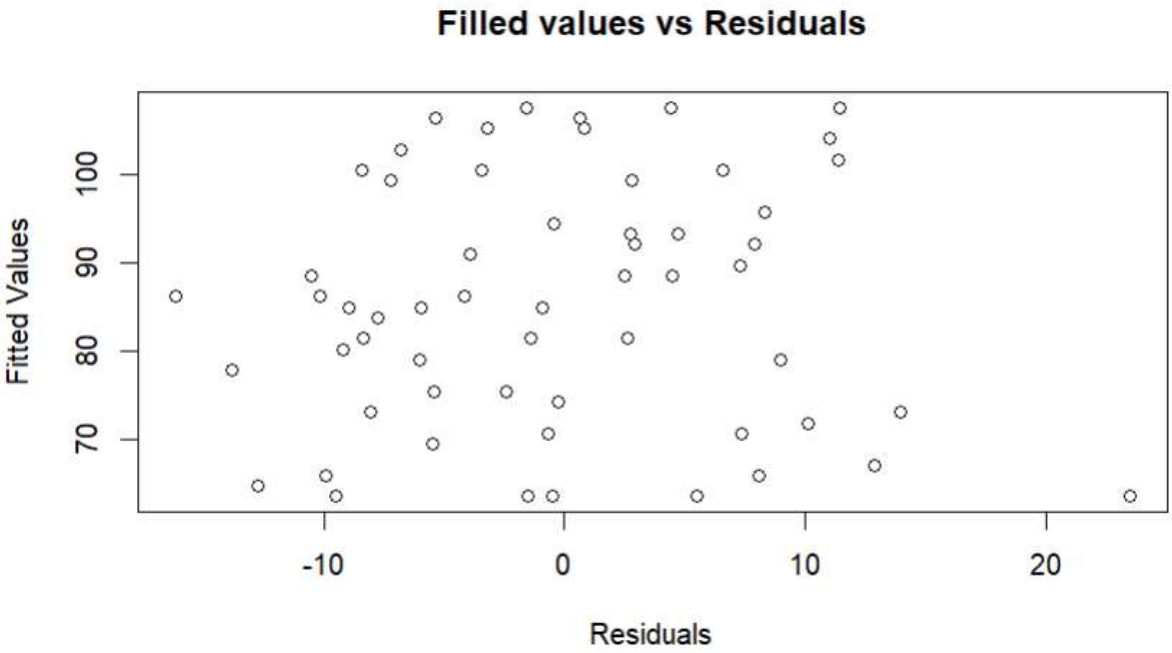
```
> R2<-1-(MSE*(n-2))/(SSR+MSE*(n-2))
> r<-sign(b1)*sqrt(R2)
> cat("R^2=",R2,"and r=",r)
R^2= 0.7500668 and r= -0.866064
```

R^2 is a pretty large value and r<0 so we would expect that there be a negative linear relation between these two variables



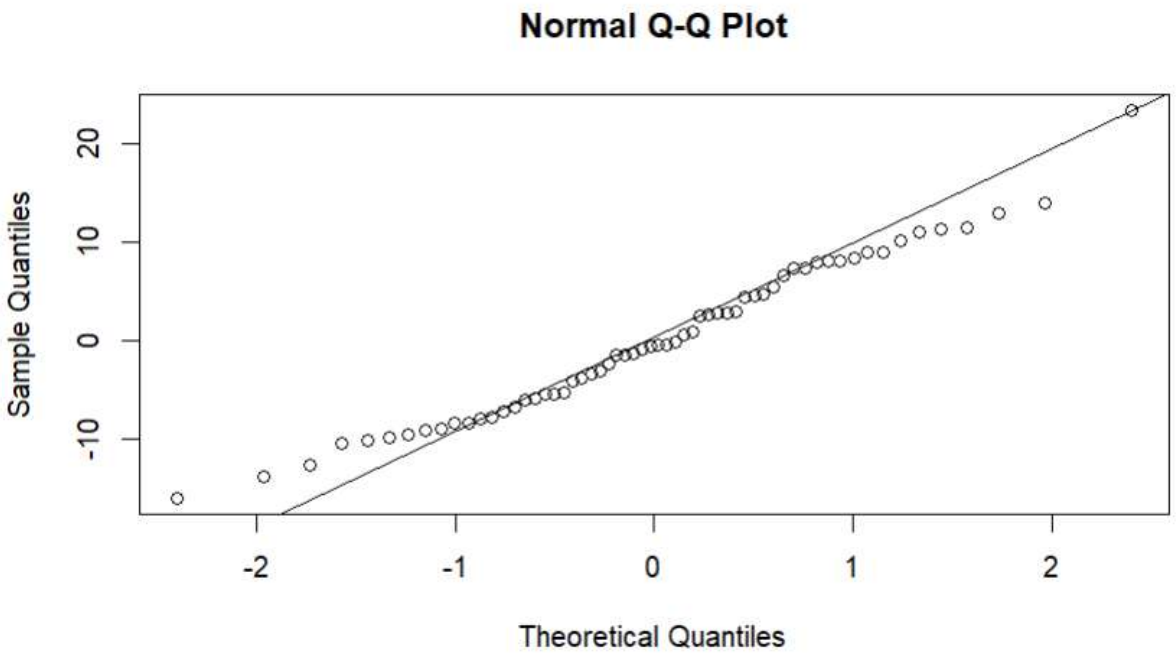
The data looks pretty well spread out

```
> plot(mm-b0-b1*age,b0+b1*age,main="Filled values vs Residuals",
+      xlab="Residuals",ylab="Fitted Values")
```



This data seems pretty random without any obvious bias

```
> qqnorm(mm-b0-b1*age)
> qqline(mm-b0-b1*age)
```



The normal assumption is a little dubious as it seems a majority of the data is off of our qqline. However, it does seem to follow roughly.

H0: \sigma_{age\leq 60}=\sigma_{age>60}

Ha: \sigma_{age\leq 60}\neq\sigma_{age>60}

```
> threshold<-60
> less<-muscledata[muscledata[,2]<=threshold,]
> great<-muscledata[muscledata[,2]>threshold,]
> ageless<-less[,2]
> agegreat<-great[,2]
> mmless<-less[,1]
> mmgreat<-great[,1]
> resless<-mmless-b0-b1*ageless
> resgreat<-mmgreat-b0-b1*agegreat
> medless<-median(resless)
> medgreat<-median(resgreat)
> d1<-abs(resless-medless)
> d2<-abs(resgreat-medgreat)
> tbf<-(mean(d1)-mean(d2))/sqrt((sum((d1-mean(d1))^2)+sum((d2-mean(d2))^2))*
+                                (1/length(mmless)+1/length(mmgreat))/(n-2))
> pbf<-pt(tbf,n-2)
> cat("Pvalue for BF=",2*min(pbf,1-pbf))
Pvalue for BF= 0.1714619
```

The Pvalue>0.05 so we reject H0 that the variation in the muscle mass is the same for those above and below 60 years of age