Homework 3

Ishita Dutta

4/23/2021

2.22

It is possible for the entire set of 30 to have a correlation coefficient not equal to 0 because the other 20 cases after the first 10 may show a different linearity, leading to a coefficient not equal to zero. In addition, the correlation coefficient for the entire set of 30 values may equal to zero, as the first 10 values could be on one side of the regression line while the rest of the cases can be on the other, making the balance even and the coefficient 0.

2.25

a)

```
library(knitr)
airfreight_breakage = read.table("airfreight+breakage.txt")
Y1 = airfreight_breakage[,1]
X1 = airfreight_breakage[,2]
n1 = length(X1)
fit1 = lm(Y1~X1)
y_hat1 = fit1$fitted.values
SSTO1 = sum((Y1-mean(Y1))^2)
SSE1 = sum((Y1-y_hat1)^2)
SSR1 = sum((y_hat1-mean(Y1))^2)
MSR1 = SSR1/(1)
MSE1 = SSE1/(n1-2)
Fstatistic1 = MSR1/MSE1
pvalue1 = pf(Fstatistic1, 1, n1-2, lower.tail = F)
result1 = data.frame(Source=c("Regression", "Error", "Total"),
 SS=c(SSR1, SSE1, SSTO1),Df=c(1, n1-2,n1-1),
 MS=c(MSR1, MSE1,NA), F_value=c(Fstatistic1,NA,NA),
 p_value=c(pvalue1,NA,NA))
kable(result1)
```

Source	SS	Df	MS	F_value	p_value
Regression	160.0	1	160.0	72.72727	2.75e-05
Error	17.6	8	2.2	NA	NA
Total	177.6	9	NA	NA	NA

SS and df are additive.

b) Hypotheses: Ho –> B1 = 0
   Ha –> B1 != 0
   Decision rule: Ho <= F(1-alpha; 1, n-2) < Ha

```
result1=data.frame(Source=c("Regression"),
F_value1=c(Fstatistic1),
p_value1=c(pvalue1))
kable(result1)
```

Source	F_value1	p_value1
Regression	72.72727	2.75e-05

```
alpha = .05
Fcritical1 = qf(1-alpha, 1, n1-2)
Fcritical1
```

```
## [1] 5.317655
```

```
Fstatistic1 <= Fcritical1
```

```
## [1] FALSE
```

Because F* > F(1-alpha; 1, n-2), there is sufficient evidence to conclude the Ha that B1 is not equal to zero.

c)

```
Fstatistic1
```

```
## [1] 72.72727
```

```
r1 = cor(X1,Y1)
t1 = r1*sqrt(n1-2)/sqrt(1-r1^2)
t1
```

```
## [1] 8.528029
```

```
F1 = t1 ^ 2
F1
```

```
## [1] 72.72727
```

t* is 8.528029. t* ^ 2 is the F statistic.

d)

```
R_sqrt1=summary(fit1)$r1.squared
R_sqrt1
```

```
## NULL
```

```
r1
```

```
## [1] 0.949158
```

```
proportion1 = (SSR1/SSTO1)*100
proportion1
```

```
## [1] 90.09009
```

The proportion of the variation in Y taken into account by introducing X is 90.09%

2.29

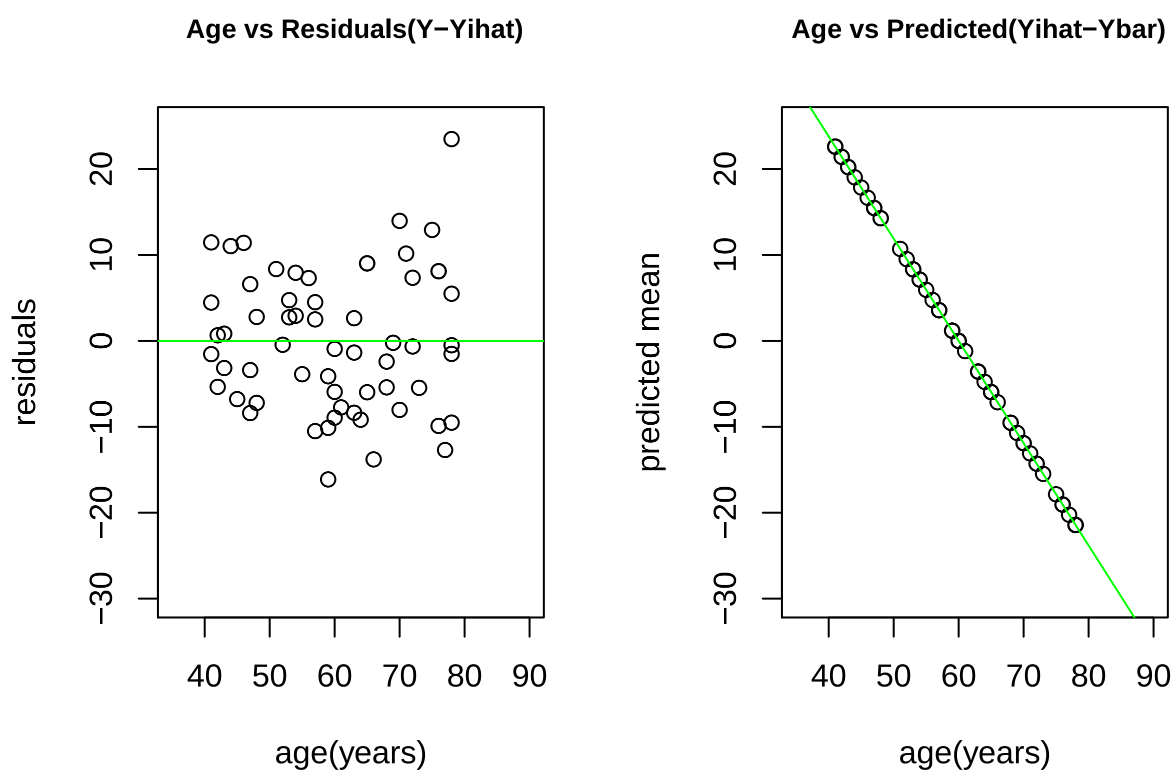
a)

```
musclemass = read.table("muscle+mass.txt")
Y2 = musclemass[,1]
X2 = musclemass[,2]
n2 = length(X2)
fit2 = lm(Y2~X2)
bohat2 = fit2$coefficients[[1]]
b1hat2 = fit2$coefficients[[2]]
Yi_hat2 = bohat2 + b1hat2*(X2)
Y2_1 = Y2-Yi_hat2
SSE2 = sum(Y2_1^2)
Y2_bar = mean(Y2)
Y2_2 = Yi_hat2-Y2_bar
SSR2 = sum(Y2_2^2)
Y2_3 = Y2 - Y2_bar
SSTO2 = sum(Y2_3^2)
result=data.frame(Source=c("Regression", "Error", "Total"),SS=c(SSR2, SSE2, SSTO2))
kable(result)
```

Source	SS
Regression	11627.486
Error	3874.447
Total	15501.933

```
par(mfrow = c(1,2))
fit2_1 = lm(Y2_1~X2)
fit2_2 = lm(Y2_2~X2)
plot(X2,Y2_1,
```

```
 xlim = c(35,90), ylim=c(-30,25),
 main = " Age vs Residuals(Y-Yihat)",
 xlab = "age(years)",
 ylab = "residuals",
 cex.main = .85)
abline(fit2_1, col = "green")
plot(X2,Y2_2,
 xlim = c(35,90), ylim=c(-30,25),
 main = " Age vs Predicted(Yihat-Ybar)",
 xlab ="age(years)",
 ylab = "predicted mean",
 cex.main = .85)
abline(fit2_2, col="green")
```



We can see that SSR is the larger component of SSTO. This implies that R^2 is closer to 1.

b)

```
y_hat2 = fit2$fitted.values
SSTO2 = sum((Y2-mean(Y2))^2)
SSE2 = sum((Y2-y_hat2)^2)
SSR2 = sum((y_hat2-mean(Y2))^2)
MSR2 = SSR2/(1)
MSE2 = SSE2/(n2-2)
Fstatistic2 = MSR2/MSE2
pvalue2 = pf(Fstatistic2, 1, n2-2, lower.tail = F)
```

```r
result2=data.frame(Source=c("Regression", "Error", "Total"),
 SS=c(SSR2, SSE2, SSTO2),Df=c(1, n2-2,n2-1),
 MS=c(MSR2, MSE2,NA), F_value=c(Fstatistic2,NA,NA),
 p_value=c(pvalue2,NA,NA))
kable(result2)
```

Source	SS	Df	MS	F_value	p_value
Regression	11627.486	1	11627.48584	174.062	0
Error	3874.447	58	66.80082	NA	NA
Total	15501.933	59	NA	NA	NA

c) Hypotheses: Ho –> B1 = 0
   Ha –> B1 != 0
   Decision rule: Ho <= F(1-alpha; 1, n-2) < Ha

```r
Fstatistic2 = MSR2/MSE2
pvalue2 = pf(Fstatistic2, 1, n2-2, lower.tail = F)
result2=data.frame(Source=c("Regression"),
 F_value=c(Fstatistic2),
 p_value=c(pvalue2))
kable(result2)
```

Source	F_value	p_value
Regression	174.062	0

```r
alpha2 = .05
Fcritical2 = qf(1-alpha2*2, 1, n2-2)
Fstatistic2 <= Fcritical2
```

```
## [1] FALSE
```

Because F* > F(1-alpha; 1, n-2), there is sufficient evidence to conclude the Ha that B1 is not equal to zero.

d)

```r
prop_tot_var2 = (SSE2/SSTO2)
prop_tot_var2
```

```
## [1] 0.2499332
```

```r
prop_tot_var2 * 100
```

```
## [1] 24.99332
```

e)

```
R_sqrt2=summary(fit2)$r.squared
R_sqrt2
```

```
## [1] 0.7500668
```

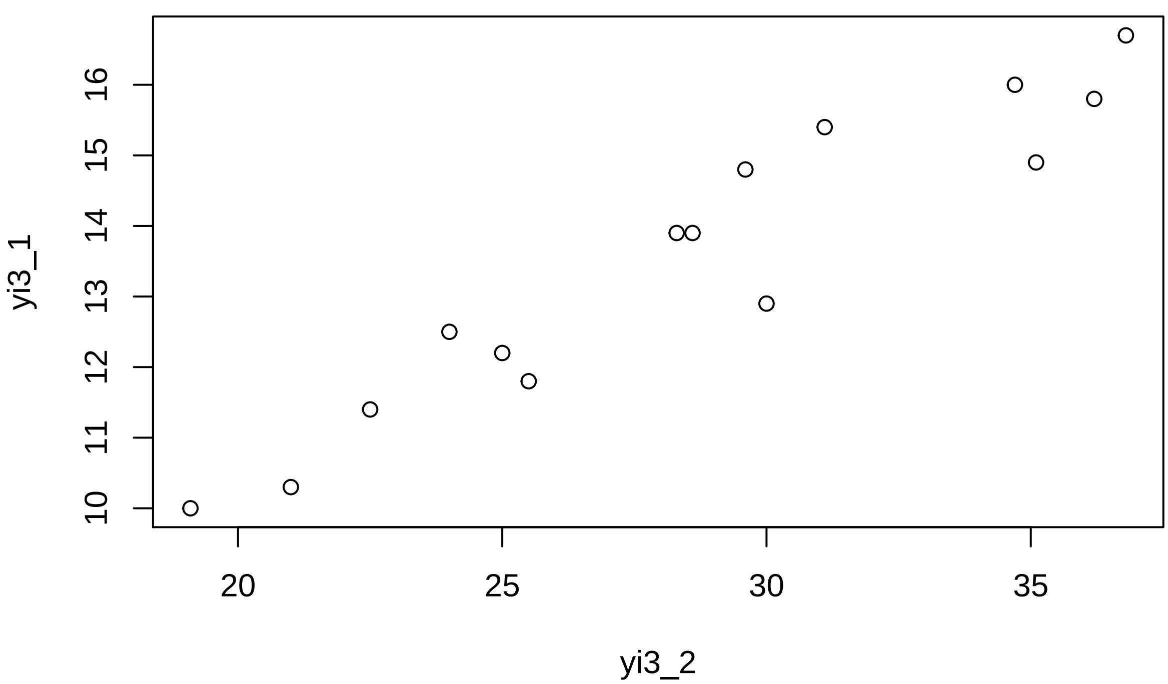
```
r2 = cor(X2,Y2)
r2
```

```
## [1] -0.866064
```

2.42

a)

```
property = read.table("property+assessments.txt")
yi3_1 = property[,1]
yi3_2 = property[,2]
n3 = length(yi3_2)
plot(yi3_2,yi3_1)
```



The bivariate normal model seems appropriate as the points seem to form a line in the scatterplot.

b)

```r
yi3_1_bar = mean(yi3_1)
yi3_2_bar = mean(yi3_2)
s3_1 = sum((yi3_1 - yi3_1_bar)*(yi3_2 - yi3_2_bar))
s3_2 = sum((yi3_1 - yi3_1_bar)^2)
s3_3 = sum((yi3_2 - yi3_2_bar)^2)
r3_12 = s3_1/sqrt(s3_2*s3_3)
r3_12
```

```
## [1] 0.9528469
```

r12 is 0.952847, and stands as an estimator for p12. When this is near 1, it means that there is a strong positive linear association between Y1 and Y2.

c) Hypotheses Ho: p12 = 0
   Ha: p12 != 0 Decision rule: Ho <= t(1 - alpha; n-2) < Ha where t* = r12 * sqrt(n-2) / sqrt(1 - r12^2)

```r
alpha3 = .01
t_stat3 = r3_12*sqrt(n3-2)/sqrt(1 - r3_12^2)
t_stat3
```

```
## [1] 11.32154
```

```r
critical_value3 = qt(1-alpha3/2,n3-2)
critical_value3
```

```
## [1] 3.012276
```

```r
t_stat3 > critical_value3
```

```
## [1] TRUE
```

Because |t*| > t(1 - alpha; n-2), there is sufficient evidence to conclude the Ha that B1 p12 != 0.

d) No, we should not test with p12 = 0.6 vs p12 != 0.6.

2.51

b0 = y-bar - b1 * x-bar E{b0} = E{y-bar} - E{b1 * x-bar} E{b0} = E{sum((1/n) * Yi)} - x-bar * E{b1} E{b0} = (1/n) * sum(E{Yi}) E{b0} = (1/n) * sum(b0 + b1 * Xi - b1 * x-bar) b0 = b0 + b1 * Xi - b1 * x-bar b0 = b0 –> proved.