HW9

Luke Geel

4/5/2021

2.23. Refer to Grade point average Problem 1.19. b. What is estimated by MSR in your ANOVA table? by MSE? Under what condition do MSR and MSE estimate the same quantity? MSR: 5.542 MSE: 0.751 When β1=0.

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ purrr   0.3.4
## ✓ tibble  3.0.3     ✓ dplyr   1.0.2
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
loadRData <- function(fileName){
  load(fileName)
  get(ls()[ls() != "fileName"])
}
gpa <- loadRData("/Users/lukegeel/Downloads/gpa_spring2021.RData")

n <- nrow(gpa) # Extract number of observations
Y <- gpa$Y # Extract response
X <- gpa$X # Extract predictor
linmod <- lm(Y~X) # Fit linear model
Y.hat <- linmod$fitted.values # Obtain fitted values
# Compute sums of squares
SSR <- sum((Y.hat - mean(Y))^2)
SSE <- sum((Y - Y.hat)^2)
# Compute mean squares
MSR <- SSR/1
MSE <- SSE/(n - 2)
MSR
```

```
## [1] 5.542257
```

```
#MSE
```

c. Conduct an F test of whether or not f31 = O. Control the a risk at .01. State the alternatives, decision rule, and conclusion. Alternatives: Null hypothesis: B1 is 0 Alternative hypothesis: B1 is not 0 Decision rule: If F* < F.statistic we reject the null hypothesis. If F* > F.statistic fail to reject null hypothesis. Conclusion: Reject null hypothesis because F* < F.statistic. Thus the alternative hypothesis is correct and B1 id not 0.

```
n <- nrow(gpa) # Extract number of observations
Y <- gpa$Y # Extract response
X <- gpa$X # Extract predictor
linmod <- lm(Y~X) # Fit linear model
Y.hat <- linmod$fitted.values # Obtain fitted values
# Compute sums of squares
SSR <- sum((Y.hat - mean(Y))^2)
SSE <- sum((Y - Y.hat)^2)
# Compute mean squares
MSR <- SSR/1
MSE <- SSE/(n - 2)
F.statistic <- MSR/MSE
alpha <- 0.01
fquantile <- qf(1 - alpha, 1, n - 2)
pvalue <- 1 - pf(F.statistic, 1, n - 2)
pvalue
```

```
## [1] 0.007589142
```

```
F.statistic
```

```
## [1] 7.379631
```

```
fquantile
```

```
## [1] 6.854641
```

2.24. Refer to Copier maintenance Problem 1.20. b. Conduct an F test to determine whether or not there is a linear association between time spent and number of copiers serviced; use a = .10. State the alternatives, decision rule, and conclusion. Alternatives: Null hypothesis: There is not a linear association between time spent and number of copiers serviced Alternative hypothesis: There is a linear association between time spent and number of copiers serviced Decision rule: If F <* F.statistic we reject the null hypothesis. If F* > F.statistic fail to reject null hypothesis. Conclusion: Fail to reject null hypothesis because F* > F.statistic, so there is not a linear association between time spent and number of copiers serviced

```
library(tidyverse)
loadRData <- function(fileName){
  load(fileName)
  get(ls()[ls() != "fileName"])
}
copier <- loadRData("/Users/lukegeel/Downloads/copier_spring2021.RData")
n <- nrow(copier) # Extract number of observations
Y <- copier$Y # Extract response
X <- copier$X # Extract predictor
linmod <- lm(Y~X) # Fit linear model
Y.hat <- linmod$fitted.values # Obtain fitted values
# Compute sums of squares
SSR <- sum((Y.hat - mean(Y))^2)
SSE <- sum((Y - Y.hat)^2)
# Compute mean squares
MSR <- SSR/1
MSE <- SSE/(n - 2)
F.statistic <- MSR/MSE
alpha <- 0.1
fquantile <- qf(1 - alpha, 1, n - 2)
pvalue <- 1 - pf(F.statistic, 1, n - 2)
pvalue
```

```
## [1] 0
```

```
F.statistic
```

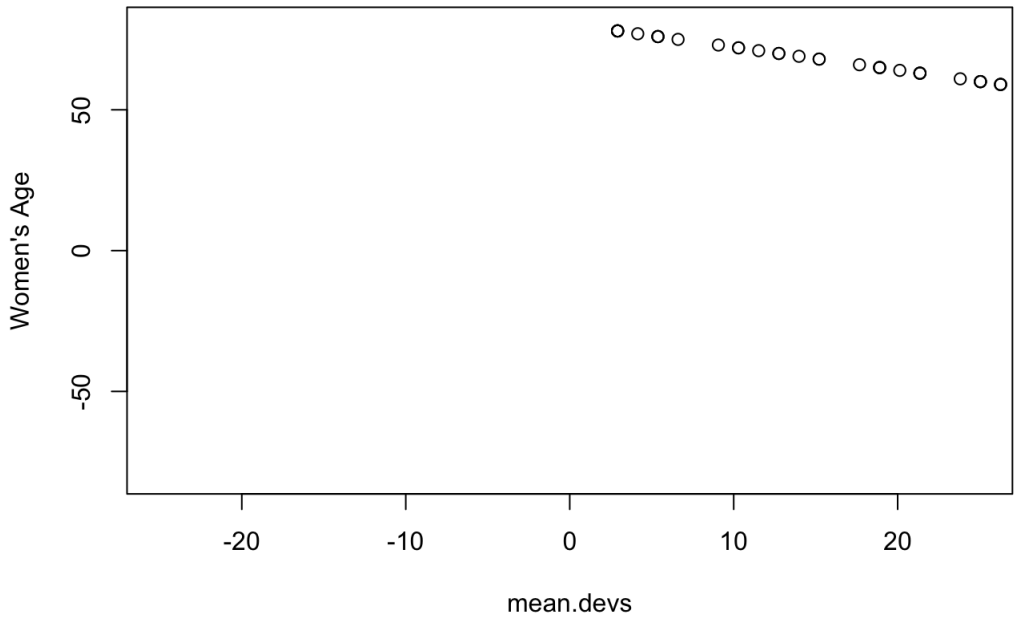
```
## [1] 603.8051
```

```
fquantile
```

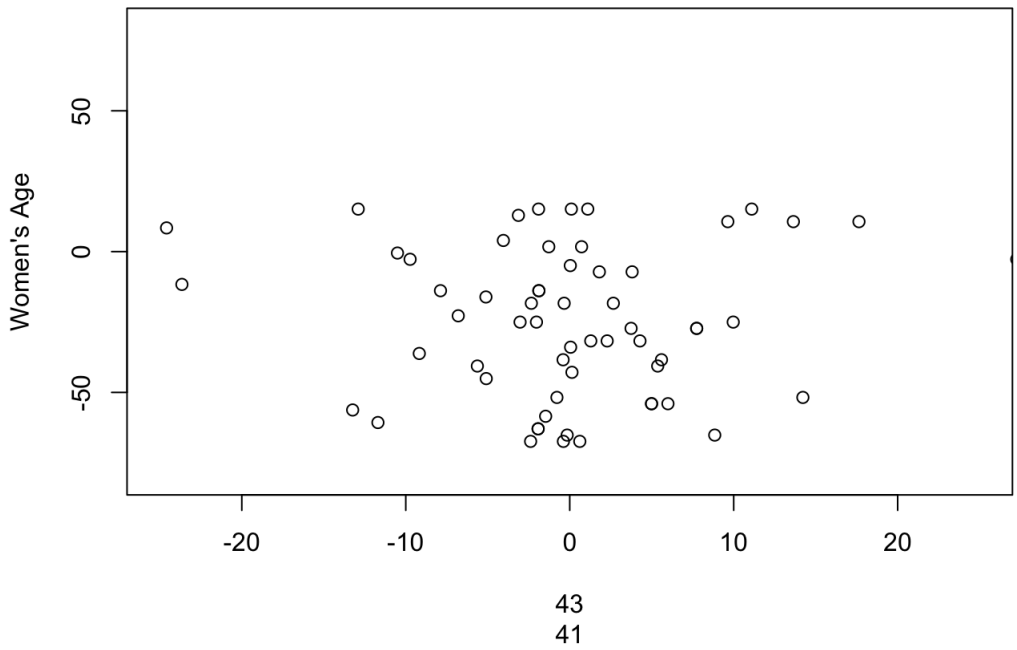
```
## [1] 2.825999
```

2.29. Refer to Muscle mass Problem 1.27. a. Plot the deviations Yi - Yhati against Xi on one graph, Plot the deviations Yhati - Ybar against Xi on another graph, using the same scales as in the first graph. From your two graphs, does SSE or SSR appear to be the larger component of SSTO? What does this imply about the magnitude of R2? SSR appears to be the larger component of SSTO because SSE is the sum of square error which is approximately zero. This means that the magnitude of R2 will be larger because more of the variance is explained by the regression due to the sum of square errors approaching zero.

```r
library(tidyverse)
loadRData <- function(fileName){
  load(fileName)
  get(ls()[ls() != "fileName"])
}
muscle <- loadRData("/Users/lukegeel/Downloads/muscle_spring2021.RData")
Age <- muscle$X
Mass <- muscle$Y
MuscMass.lm <- lm(Mass~Age)
devs <- muscle - predict(MuscMass.lm)
mean.devs <- predict(MuscMass.lm) - mean(Age)
plot(mean.devs, Age, type = "p", ylab = "Women's Age", ylim= c(-80,80), xlim = c(-25,25))#, xlab= "SSR")
```




```r
plot(devs, Age, type = "p", ylab = "Women's Age", ylim= c(-80,80), xlim = c(-25,25))#, xlab= "Deviation aro
und fitted regression line (SSE)")
```



c. Test whether or not B1, = 0 using an F test with a = .05. State the alternatives, decision rule, and conclusion. The null hypothesis is that β1=0. The alternative hypothesis is that β1≠0. F⋆=164.84, F=4.006873 Conclusion: Because F* > fquantile, we Reject the null hypothesis thus B1 is not 0.

```
load("/Users/lukegeel/Downloads/muscle_spring2021.RData")
n <- nrow(data) # Extract number of observations
Y <- data$Y # Extract response
X <- data$X # Extract predictor
linmod <- lm(Y~X) # Fit linear model
b0 <- linmod$coef[1]
b1 <- linmod$coef[2]

Y.hat <- linmod$fitted.values # b0 + b1*X

SSR <- sum((Y.hat - mean(Y))^2)
SSE <- sum((Y - Y.hat)^2)

MSR <- SSR/1
MSE <- SSE/(n - 2)

F.statistic <- MSR/MSE

alpha <- 0.05
fquantile <- qf(1 - alpha, 1, n - 2)
pvalue <- 1 - pf(MSR/MSE, 1, n - 2)

F.statistic
```

```
## [1] 164.8386
```

```
fquantile
```

```
## [1] 4.006873
```

```
pvalue
```

```
## [1] 0
```

d. What proportion of the total variation in muscle mass remains “unexplained” when age is introduced into the analysis? Is this proportion relatively small or large? The proportion of the total variation in muscle mass that remains “unexplained” when age is introduced is approximately 26%. This proportion is relatively small because about 74% of the variance is explained by age. Therefore, age is good predictor of muscle mass.

```
summary(MuscMass.lm)
```

```
##
## Call:
## lm(formula = Mass ~ Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.5897  -3.0549  -0.2494   4.4592  27.2637
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.78889    5.85158   27.14   <2e-16 ***
## Age          -1.22932    0.09575  -12.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.676 on 58 degrees of freedom
## Multiple R-squared:  0.7397, Adjusted R-squared:  0.7352
## F-statistic: 164.8 on 1 and 58 DF,  p-value: < 2.2e-16
```

6.11. Refer to Grocery retailer Problem 6.9. Assume that regression model (6.5) fOi’ three predictor variables with independent normal error terms is appropriate. a. Test whether there is a regression relation, using level of significance .05. State the alternatives, decision rule, and conclusion. What does yom’ test result imply about B1, B2, and B3? What is the P-value of the test? Alternatives: Ho: There is no

regression relation Ha: There is a regression relation Decision rule: If the test statistic doesn't exceed the F-quantile, we fail to reject the null hypothesis and conclude that there is not a regression relation. If the test statistic does exceed the F-quantile, then there is a regression relation. Conclusion: Because the test statistic exceeds F (0.95; 1, df), we conclude Ha: β1 does not equal 0, we conclude that there is evidence of a regression relation at level α = 0.05. p-value: 1.371703e-11 This tells us that B1, B2, and B3 are all useful for predicting the response.

```
grocery <- loadRData("/Users/lukegeel/Downloads/grocery_spring2021.RData")
Y <- grocery$Y
X1 <- grocery$X1
X2 <- grocery$X2
X3 <- grocery$X3
linmod <- lm(Y~X1+X2+X3)
Y.hat <- linmod$fitted.values
SSR <- sum((Y.hat - mean(Y))^2)
SSE <- sum((Y - Y.hat)^2)
# Compute mean squares
MSR <- SSR/1
MSE <- SSE/(n - 2)
F.statistic <- MSR/MSE
alpha <- 0.05
fquantile <- qf(1 - alpha, 1, n - 2)
pvalue <- 1 - pf(F.statistic, 1, n - 2)
pvalue
```

```
## [1] 1.371703e-11
```

```
F.statistic
```

```
## [1] 70.36822
```

```
fquantile
```

```
## [1] 4.006873
```

c. Calculate the coefficient of multiple determination R^2. How is this measure interpreted here? R^2 = 0.5482 which means that the model explains 54.82% of the variation in the regression relation.

```
r2 <- summary(linmod)$r.squared
r2
```

```
## [1] 0.5481748
```

*6.16. Refer to Patient satisfaction Problem 6.15. Assume that regression model (6.5) for three predictor variables with independent normal error terms is appropriate. a. Test whether there is a regression relation; use alpha = .10. State the alternatives, decision rule, and conclusion. What does your test imply about B1, B2, and B3? What is the P-value of the test? Alternatives: Ho: There is no regression relation Ha: There is a regression relation Decision rule: If the test statistic doesn't exceed the F-quantile, then there is not a regression relation.If the test statistic does exceed the F-quantile, then there is a regression relation. Conclusion: Because the test statistic exceeds F (0.9; 1, df), we conclude Ha: β1 does not equal 0, we conclude that there is evidence of a regression relation at level α = 0.05. p-value: 0+ This tells us that there is no evidence that at least one of B1, B2, and B3 is not 0.

```
patient <- loadRData("/Users/lukegeel/Downloads/patient_satisfaction_spring2021.RData")
Y <- patient$Y
X1 <- patient$X1
X2 <- patient$X2
X3 <- patient$X3
linmod <- lm(Y~X1+X2+X3)
Y.hat <- linmod$fitted.values
SSR <- sum((Y.hat - mean(Y))^2)
SSE <- sum((Y - Y.hat)^2)
MSR <- SSR/1
MSE <- SSE/(n - 2)
F.statistic <- MSR/MSE
alpha <- 0.1
fquantile <- qf(1 - alpha, 1, n - 2)
pvalue <- 1 - pf(F.statistic, 1, n - 2)
pvalue
```

```
## [1] 0
```

```
F.statistic
```

```
## [1] 162.6271
```

```
fquantile
```

```
## [1] 2.794089
```

c. Calculate the coefficient of multiple determination. What does it indicate here? r2 = 0.7371 which means that the model explains 73.71% of the variation in the regression relation.

```
r2 <- summary(linmod)$r.squared
r2
```

```
## [1] 0.737113
```

7.4 Refer to Grocery retailer Problem 6.9. a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X1; with X3, given X1; and with X2, given X1,and X3.

```
                    SS          Df          MS
```

SSR(x1, x2, x3) 1613306.850 3.000 537768.950 SSR(x1) 51038.553 1.000 51038.553 SSR(x3|x1) 3782.727 1.000 3782.727 SSR(x3|x1, x2) 1558485.571 1.000 1558485.571 SSE 1329745.150 48.000 27703.024 Total 2943052.000 51.000 1641009.874

```
grocery <- loadRData("/Users/lukegeel/Downloads/grocery_spring2021.RData")
#view(grocery)
Y <- grocery$Y
X1 <- grocery$X1
X2 <- grocery$X2
X3 <- grocery$X3
linmod <- lm(Y~X1+X2+X3)
anova(linmod)
```

X1
X2
X3
Residuals

4 rows | 1-1 of 6 columns

```
SSR = sum( anova(linmod)[1:3,2] )  #SSR(X1, X2, X3), by summing above three SSR
MSR = SSR / 3                     #MSR(X1, X2, X3) = SSR / df
SSE = anova(linmod)[4,2]            #SSE(X1, X2, X3)
MSE = anova(linmod)[4,3]            #MSE(X1, X2, X3)

#attain alternate decompositions of Extra Sums of Squares:
#get SSR(X3), SSR(X1|X3) and SSR(X2|X1,X3)

#to get SSR( X2, X3 | X1 ) = SSE( X1 ) - SSE( X1, X2, X3 ),
#use equation (7.4b). You need:
#run a linear model involving only X1 to obtain SSE( X1 ).
linmod1 = lm( Y ~ X1)
anova(linmod1)
```

X1

Residuals

2 rows | 1-1 of 6 columns

```
SSE.x1 = anova(linmod1)[1,3]
#then calculate needed SSR
SSR.x1 <- SSE.x1 - SSE
SSR.x1
```

```
## [1] -1278707
```

```
SSE.x1
```

```
## [1] 51038.55
```

```
SSE
```

```
## [1] 1329745
```

```
linmod.aov <- anova(linmod)
tab <- as.table(cbind(
  'SS' = c("SSR(x1, x2, x3)" = sum(linmod.aov[1:3, 2]),
           "SSR(x1)"         = linmod.aov[1, 2],
           "SSR(x3|x1)"      = linmod.aov[2, 2],
           "SSR(x3|x1, x2)"  = linmod.aov[3, 2],
           "SSE"             = linmod.aov[4, 2],
           "Total"           = sum(linmod.aov[, 2])),

  'Df' = c(                    sum(linmod.aov[1:3, 1]),
                               linmod.aov[1, 1],
                               linmod.aov[2, 1],
                               linmod.aov[3, 1],
                               linmod.aov[4, 1],
                               sum(linmod.aov$Df)),

  'MS' = c(                    sum(linmod.aov[1:3, 2]) / sum(linmod.aov[1:3, 1]),
                               linmod.aov[1, 3],
                               linmod.aov[2, 3],
                               linmod.aov[3, 3],
                               linmod.aov[4, 3],
                               sum(linmod.aov[, 3]))
))

#round(tab, 2)
linmod.aov
```

X1

X2

X3

Residuals

4 rows | 1-1 of 6 columns

```
tab
```

```
##                          SS          Df          MS
## SSR(x1, x2, x3) 1613306.850       3.000  537768.950
## SSR(x1)           51038.553       1.000   51038.553
## SSR(x3|x1)         3782.727       1.000    3782.727
## SSR(x3|x1, x2)  1558485.571       1.000 1558485.571
## SSE             1329745.150      48.000   27703.024
## Total           2943052.000      51.000 1641009.874
```

b. Test whether X2 can be dropped from the regression model given that X1, and X3 are retained. Use the F* test statistic and ex = .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test? Alternatives: H0: B2=0 Ha: B2 does not equal 0 Decision rule: If F<*F(0.95,1,48) then fail to reject H0 and X2 can be dropped. Conclusion: F* <F(0.95,1,48) thus X2 can be dropped P-value: 0.649

```
#anova(update(linmod, . ~ . - x3), linmod)
drop1(linmod, test = "F")
```

<none>
X1
X2
X3

4 rows | 1-1 of 7 columns

c. Does SSR(X1) +SSR(X2|X1) equal SSR(X2) +SSR(X1|X2) here? Must this always be the case? No, they are not equal in this case. This isn't always the case as sometimes they can be equal.

```
#tab
linmod.aov
```

X1
X2
X3
Residuals

4 rows | 1-1 of 6 columns

```
SSRX1 <- linmod.aov[1,2]
SSRX1.X3 <- tab[3,1]
SSRX3 <- linmod.aov[3,2]
SSRX3.X1 <-  tab[3,1]
left <- SSRX1 + SSRX1.X3
right <-SSRX3 + SSRX3.X1
#SSRX1
#SSRX3
left
```

```
## [1] 54821.28
```

```
right
```

```
## [1] 1562268
```

7.5 Refer to Patient satisfaction Problem 6.15. a Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X2; with X1 given X2; and with X3, given X2 and X1. SS Df MS SSR(x1, x2, x3) 1613306.850 3.000 537768.950 SSR(x2) 51038.553 1.000 51038.553 SSR(x1|x2) 3782.727 1.000 3782.727 SSR(x3|x1, x2) 1558485.571 1.000 1558485.571 SSE 1329745.150 48.000 27703.024 Total 2943052.000 51.000 1641009.874

```
patient <- loadRData("/Users/lukegeel/Downloads/patient_satisfaction_spring2021.RData")
Y <- patient$Y
X1 <- patient$X1
X2 <- patient$X2
X3 <- patient$X3
linmod <- lm(Y~X1+X2+X3)
tab <- as.table(cbind(
  'SS' = c("SSR(x1, x2, x3)" = sum(linmod.aov[1:3, 2]),
           "SSR(x1)"         = linmod.aov[1, 2],
           "SSR(x3|x1)"      = linmod.aov[2, 2],
           "SSR(x3|x1, x2)"  = linmod.aov[3, 2],
           "SSE"             = linmod.aov[4, 2],
           "Total"           = sum(linmod.aov[, 2])),

  'Df' = c(                    sum(linmod.aov[1:3, 1]),
                               linmod.aov[1, 1],
                               linmod.aov[2, 1],
                               linmod.aov[3, 1],
                               linmod.aov[4, 1],
                               sum(linmod.aov$Df)),

  'MS' = c(                    sum(linmod.aov[1:3, 2]) / sum(linmod.aov[1:3, 1]),
                               linmod.aov[1, 3],
                               linmod.aov[2, 3],
                               linmod.aov[3, 3],
                               linmod.aov[4, 3],
                               sum(linmod.aov[, 3]))
))

tab
```

```
##                           SS          Df          MS
## SSR(x1, x2, x3) 1613306.850       3.000  537768.950
## SSR(x1)           51038.553       1.000   51038.553
## SSR(x3|x1)         3782.727       1.000    3782.727
## SSR(x3|x1, x2)  1558485.571       1.000 1558485.571
## SSE             1329745.150      48.000   27703.024
## Total           2943052.000      51.000 1641009.874
```

b. Test whether X3 can be dropped from the regression model given that X1, and X2 are retained. Use the F* test statistic and level of significance .025. State the alternatives, decision rule, and conclusion. What is the P-value of the test? Alternatives: H0: B2=0 Ha: B2 does not equal 0 Decision rule: If F<*F(0.975,1,48) then fail to reject H0 and X3 can be dropped. Conclusion: F* >F(0.975,1,48) thus X3 cannot be dropped P-value: 0.03312

```
patient <- loadRData("/Users/lukegeel/Downloads/patient_satisfaction_spring2021.RData")
Y <- patient$Y
X1 <- patient$X1
X2 <- patient$X2
X3 <- patient$X3
linmod <- lm(Y~X1+X2+X3)
drop1(linmod, test = "F")
```

<none>
X1
X2
X3

4 rows | 1-1 of 7 columns

7.6 Refer to Patient satisfaction Problem 6.15. Test whether both X2 and X3 can be dropped from the regression model given that X1 is retained. Use ex ~ .025. State the alternatives, decision rule, and conclusion. What is the P-value of the test? Alternatives: H0: X2 and X3 are not significant and can be dropped. Ha: X2 and X3 are significant and can't be dropped. Decision rule: If p-value<0.025 then fail to reject H0 and X3 and X2 can be dropped. Conclusion: p-value<0.025 thus X3 and X2 can be dropped P-value: 2.976e-12

```
patient <- loadRData("/Users/lukegeel/Downloads/patient_satisfaction_spring2021.RData")
Y <- patient$Y
n <- nrow(patient)
X1 <- patient$X1
X2 <- patient$X2
X3 <- patient$X3
summary(lm(Y~X1+X2+X3))
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.370  -5.146  -0.693   3.691  31.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 162.4755    17.0611   9.523 4.73e-12 ***
## X1           -1.2470     0.2022  -6.168 2.28e-07 ***
## X2           -0.4068     0.4631  -0.879   0.3846
## X3          -14.7226     6.6826  -2.203   0.0331 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.467 on 42 degrees of freedom
## Multiple R-squared:  0.7371, Adjusted R-squared:  0.7183
## F-statistic: 39.25 on 3 and 42 DF,  p-value: 2.976e-12
```

```
linmod <- lm(Y~X1+X2+X3)
linmod1 <- lm(Y~X1)
#anova(lm(Y~X1), patient)
```

7.9 Refer to Patient satisfaction Problem 6.15. Test whether B1 = -1.0 and B2 = 0; use ex = .025 State the alternatives, full and reduced models, decision rule, and conclusion. Full model: Yi = B0 + B1Xi1 + B2Xi2 + B3Xi3 + error Reduced model: Yi + Xi1 = B0 + B3Xi3 + error Alternatives: H0: B1 = -1.0 and B2 = 0 Ha: Not both equalities hold Decision rule: If F* < F(0.975,2,42) fail to reject H0 If F* > F(0.975,2,42) reject H0 Conclusion: Since F* < F(0.975,2,42), we fail to reject the null hypothesis and thus, B1 = -1.0 and B2 = 0

```
patient <- loadRData("/Users/lukegeel/Downloads/patient_satisfaction_spring2021.RData")
Y <- patient$Y
n <- nrow(patient)
X1 <- patient$X1
X2 <- patient$X2
X3 <- patient$X3
#summary(lm(Y~X1+X2+X3))
linmod <- lm(Y~X1+X2+X3)
linmod1 <- lm(Y~X3)
#summary(linmod)
anova(linmod)
```

X1
X2
X3
Residuals

4 rows | 1-1 of 6 columns

```
Y.hat <- linmod$fitted.values # Obtain fitted values
Yi.hat <- linmod1$fitted.values
# Compute sums of squares
SSR1 <- sum((Yi.hat - mean(Y))^2)
SSR <- sum((Y.hat - mean(Y))^2)
SSE1 <- sum((Y - Yi.hat)^2)
SSE <- sum((Y - Y.hat)^2)
SSE
```

```
## [1] 3764.319
```

```
SSE1
```

```
## [1] 7926.323
```

```
Fstar <- ((SSE-SSE1)/2)/(SSE/42)
Fstar
```

```
## [1] -23.21856
```