HW8

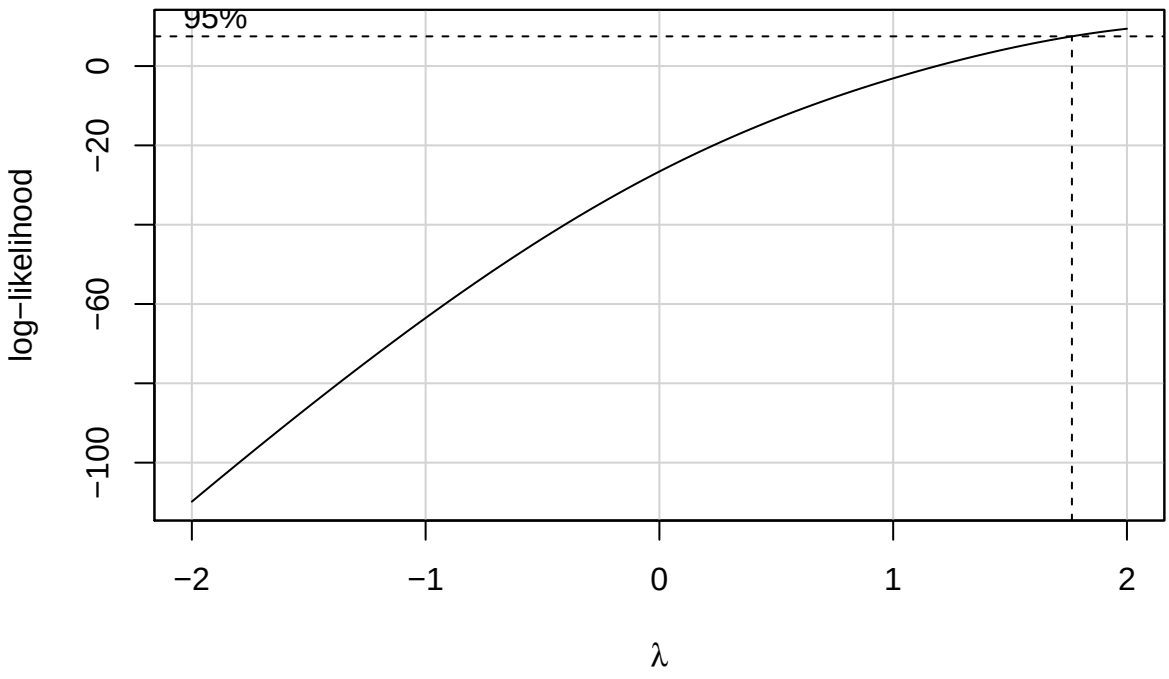
Courtney Schaller

5/4/2021

1. (Ch 5 p.26, Windmill Data) A research engineer is investigating the use of a windmill to generate electricity. He has collected data on the DC output from his windmill (y) and the corresponding wind velocity (x). (40 points)

(a) Identify the most appropriate transformation for the data using Box-Cox method. Pick the best integer λ value. (10 points)

```
car::boxCox(lm(data = windmill, y ~ x), plotit = TRUE)
```



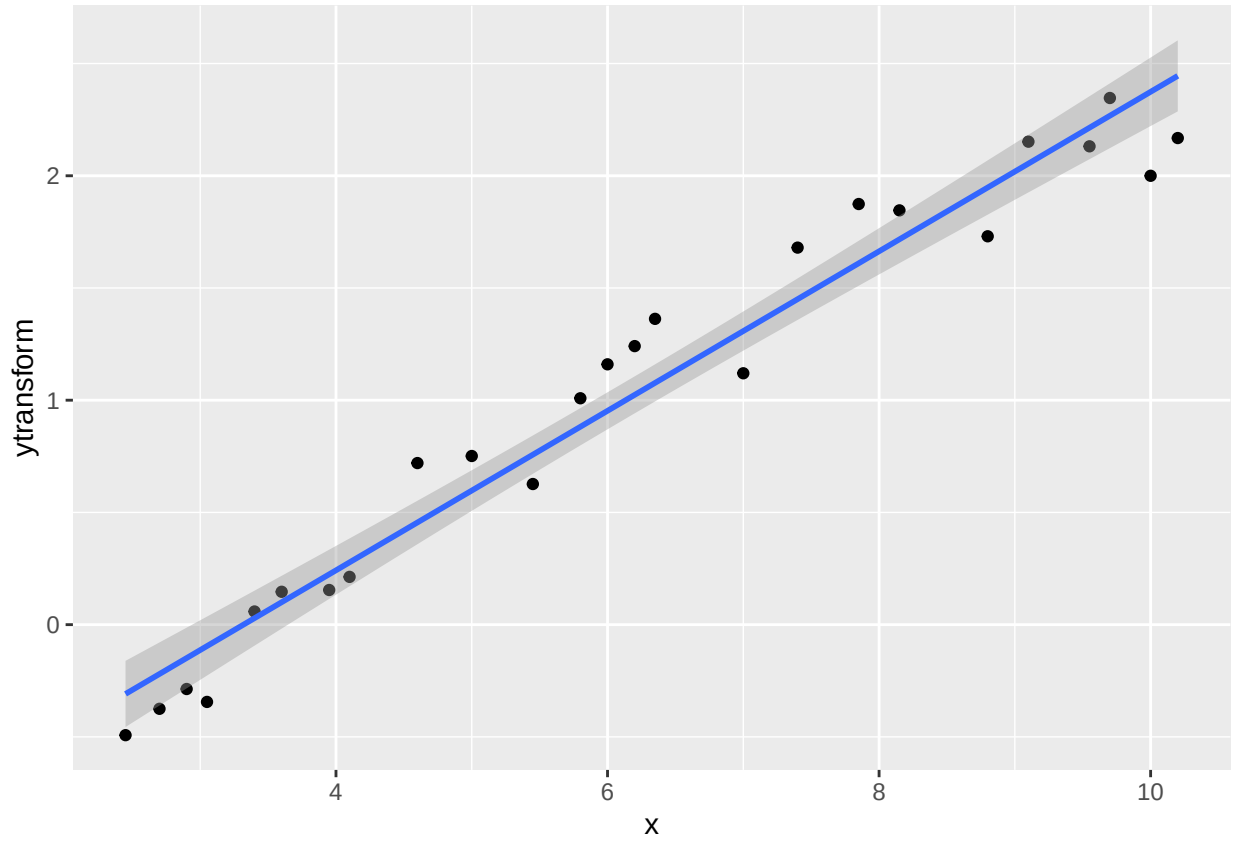
$\lambda = 2$

(b) Perform complete regression analysis with wind velocity (x) on the transformed DC output (y). (30 points)

i. Draw a scatter plot and add a regression line. (5 points)

```
windmill <- windmill %>%
    mutate(ytransform = (y^2 - 1)/2)
ggplot(data = windmill, aes(x = x, y = ytransform)) + geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



ii. Fit a linear regression model. (5 points)

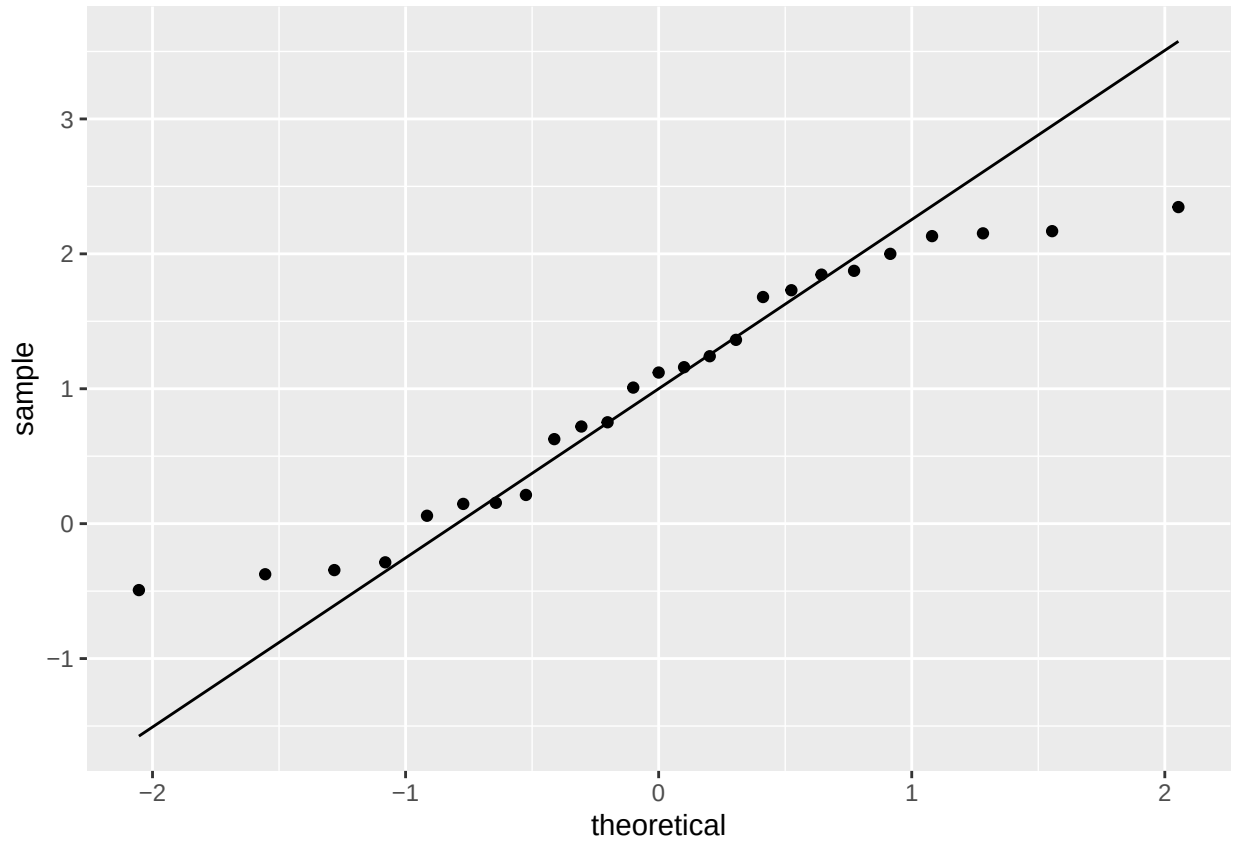
```
lmwindmill <- lm(data = windmill, ytransform ~ x)
summary(lmwindmill)
```

```
##
## Call:
## lm(formula = ytransform ~ x, data = windmill)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37420 -0.15514  0.02976  0.15397  0.28536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.17926    0.10620  -11.11 1.02e-10 ***
## x            0.35533    0.01606   22.13  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.199 on 23 degrees of freedom
```

```
## Multiple R-squared:  0.9551, Adjusted R-squared:  0.9532
## F-statistic: 489.7 on 1 and 23 DF,  p-value: < 2.2e-16
```

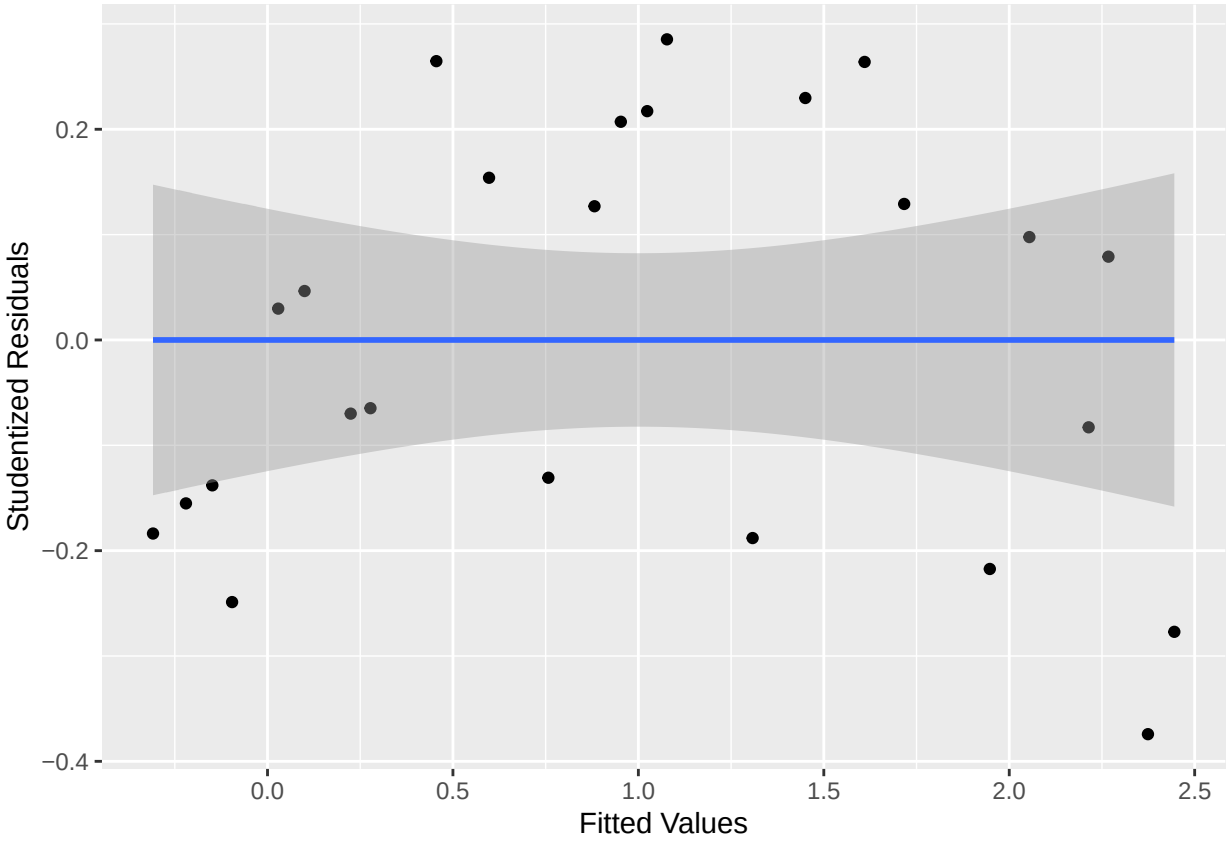
iii. Draw a QQ plot (normal probability plot) and a plot of the studentized residuals versus the fitted values. Interpret them. (10 points)

```
ggplot(windmill, aes(sample = ytransform)) + stat_qq() + stat_qq_line()
```




```
ggplot(lmwindmill, aes(x = fitted(lmwindmill), y = residuals(lmwindmill))) + geom_point() +
    geom_smooth(method = "lm") + xlab("Fitted Values") + ylab("Studentized Residuals")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From our residuals vs. fitted plot, we see that our data has a somewhat concave shape; this implies that perhaps a better model exists for our data. Our model overestimates y when x is comparatively high or low, and underestimates y near the average x values. Overall though, the data still fits fairly well to our linear model. Our Q-Q plot is pretty good and supports an assumption of normality, although at the far ends of either side things veer off from expected a bit.

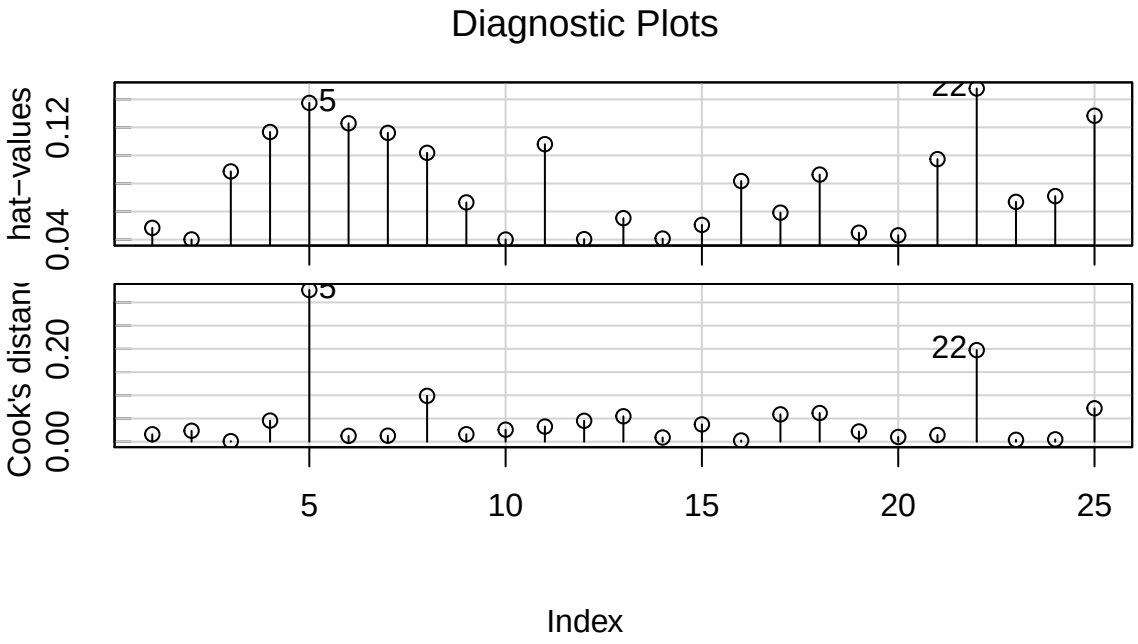
iv. Perform influence analysis. (10 points)

```
summary(influence.measures(lmwindmill))
```

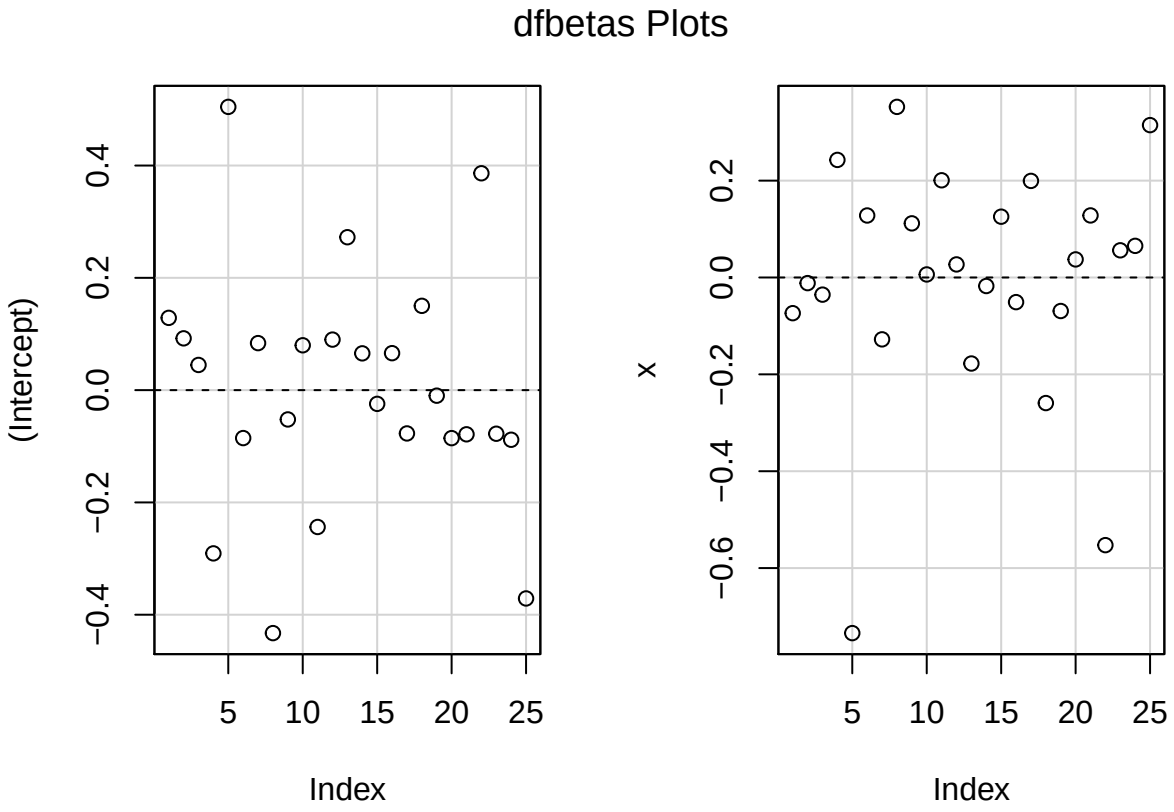
```
## Potentially influential observations of
##   lm(formula = ytransform ~ x, data = windmill) :
## NONE
```

```
## numeric(0)
```

```
car::influenceIndexPlot(lmwindmill, vars = c("hat", "Cook"))
```




```
car::dfbetasPlots(lmwindmill, intercept = T)
```



No points are influential in affecting any regression coefficients and fitted values. None of them are considered leverage points, although points 5 and 22 have the most leverage. No points seem influential to precision of estimation. We need further investigation.

2. (Ch 8 pp.14-15, Patient Data) A hospital is implementing a program to improve service quality and productivity. As part of this program the hospital management is attempting to measure and evaluate patient satisfaction. The data has been collected on a random sample of 25 recently discharged patients. The response variable (y) is satisfaction, a subjective response measure on an increasing scale. The potential aggressor variables are patient age (x1), severity (x2) (an index measuring the severity of the patient's illness), an indicator of whether the patient is a surgical or medical patient (x3) (0=surgical, 1=medical), and an index measuring the patient's anxiety level (x4). (35 points)

(a)

```
patient <- patient %>%
    mutate(x11 = ifelse(Age <= 29, 1, 0), x12 = ifelse(Age <= 39 & Age > 29, 1, 0),
        x13 = ifelse(Age <= 49 & Age > 39, 1, 0), x14 = ifelse(Age <= 59 & Age >
            49, 1, 0), x15 = ifelse(Age > 59, 1, 0))
```

Test whether or not Age Groups 1, 2 and 3 are different in terms of satisfaction. (15 points)

i. Write down the null and alternative hypotheses. (5 points)

$\beta_{12} = \beta_{13} = \beta_{12} - \beta_{13} = 0$

$H_a : \beta_{13} \neq 0$ or $\beta_{12} \neq 0$ or $\beta_{13} \neq \beta_{12}$

ii. Rewrite the null and alternative hypotheses with appropriate D and d for the general linear hypothesis approach. (5 points)

$satisfaction = \beta_0 + \beta_{12}x_{12} + \beta_{13}x_{13} + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

$$D = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}$$

$$d = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

iii. Find the observed test statistic, p-value, and draw a conclusion. (5 points)

```
lmpatient <- lm(data = patient, Satisfaction ~ x12 + x13 + Severity + Sugical.Medical +
    Anxiety)
summary(lmpatient)
```

```
##
## Call:
## lm(formula = Satisfaction ~ x12 + x13 + Severity + Sugical.Medical +
##     Anxiety, data = patient)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.414  -7.207   1.020   3.873  34.377
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     104.2613    15.4939   6.729 1.98e-06 ***
## x12              21.4999     8.0743   2.663   0.0154 *
## x13               6.8638    10.1581   0.676   0.5074
## Severity         -0.6879     0.2933  -2.345   0.0300 *
## Sugical.Medical  -4.0513     5.9776  -0.678   0.5061
## Anxiety          -2.3104     1.9511  -1.184   0.2509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.49 on 19 degrees of freedom
## Multiple R-squared:  0.6315, Adjusted R-squared:  0.5346
## F-statistic: 6.513 on 5 and 19 DF,  p-value: 0.001097
```

```
test <- multcomp::glht(lmpatient, linfct = D, rhs = d)
summary(test, test = Ftest())
```

```
##
##   General Linear Hypotheses
##
## Linear Hypotheses:
##        Estimate
## 1 == 0   21.500
## 2 == 0    6.864
## 3 == 0   14.636
##
## Global Test:
##       F DF1 DF2 Pr(>F)
## 1 3.625   2  19 0.0464
```

The model finds that $\beta_{12} = 21.500$ and $\beta_{13} = 6.8638$, which therefore means that the difference between β_{12} and β_{13} is 14.636. Under the null hypothesis we expect each of these to be equal to 0. Performing an F-test,

we see that the likelihood of getting these values for our betas is 0.04639546, which is less than $\alpha = 0.05$. Therefore we must reject the null hypothesis and conclude that at age group (below 30, in 30s, or in 40s) does affect satisfaction score.

(b) (20 points)

i. Present the summary table of the fit. (10 points)

```
lmpatient2 <- lm(data = patient, Satisfaction ~ x14 + x15 + Severity + Sugical.Medical +
    Anxiety)
summary(lmpatient2)
```

```
##
## Call:
## lm(formula = Satisfaction ~ x14 + x15 + Severity + Sugical.Medical +
##     Anxiety, data = patient)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.157  -7.407  -1.525   6.006  16.002
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       87.7627     9.3204   9.416 1.37e-08 ***
## x14              -16.2392     5.7674  -2.816   0.0110 *
## x15              -48.8871     8.1199  -6.021 8.59e-06 ***
## Severity          -0.4358     0.2006  -2.172   0.0427 *
## Sugical.Medical   -1.7310     4.0850  -0.424   0.6765
## Anxiety            4.9538     2.0958   2.364   0.0289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.902 on 19 degrees of freedom
## Multiple R-squared:  0.828,  Adjusted R-squared:  0.7827
## F-statistic: 18.29 on 5 and 19 DF,  p-value: 1.123e-06
```

ii. Interpret the regression coefficients. (10 points)

β_0 : For a theoretical surgical patient under 50 with anxiety and severity scores of zero, our model predicts a satisfaction score of 87.7627.

β_{14} : If a patient is in their fifties, we expect their satisfaction to be 16.2392 points lower than someone less than 50 years old.

β_{15} : If a patient is older than 60, we expect their satisfaction to be 48.8871 points lower than someone less than 50 years old.

β_2 : For every unit increase in a patient's illness' severity, we expect a drop in satisfaction by 0.4358 points.

β_3 : If a patient is a medical patient (as opposed to surgical), we expect a drop in satisfaction by 1.7310 points

β_4 : For every unit increase in a patient's anxiety, we expect an increase in satisfaction by 4.9538