Assignment 7 (34300)

Seung Chul Lee

11/20/2021

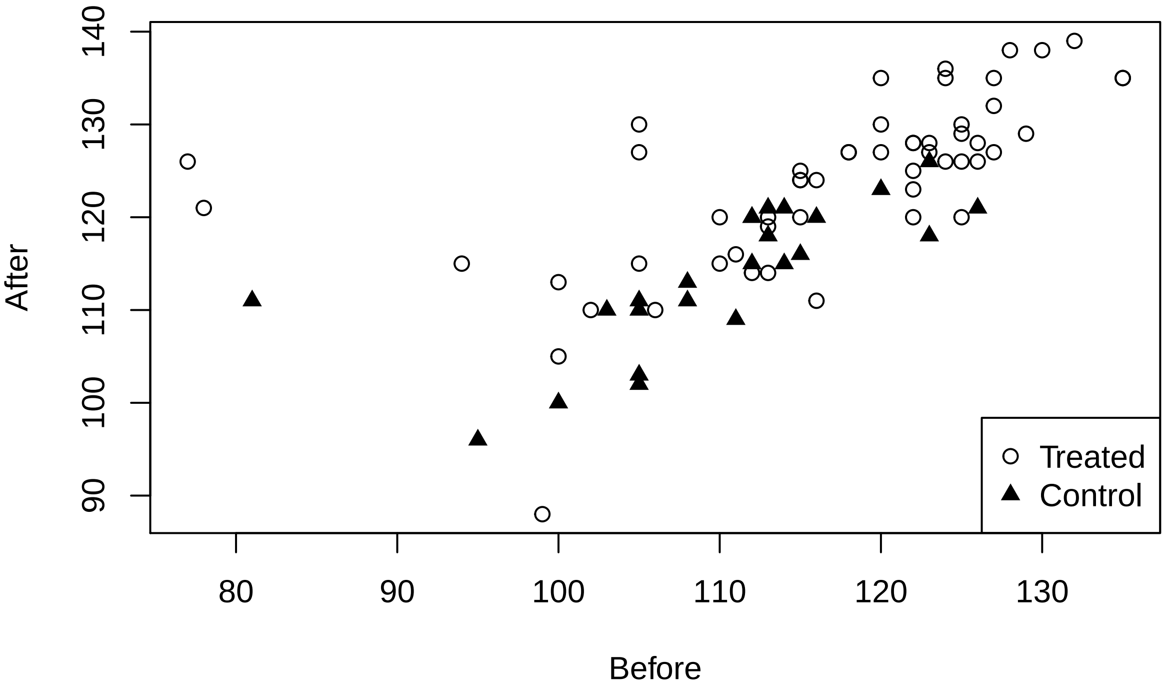
Problem 1

```
require(faraway)
```

```
## Loading required package: faraway
```

(a)

```
plot(hips$fbef[hips$grp == "treat"], hips$faft[hips$grp == "treat"], pch = 1, xlab = "Before", ylab = "
points(hips$fbef[hips$grp == "control"], hips$faft[hips$grp == "control"], pch = 17)
legend("bottomright", legend = c("Treated", "Control"), pch = c(1, 17))
```



The plot seems to suggest that for both groups, there is a linear trend between `fbef` and `faft`. However, there seems to be some outliers to the left corner. This is natural, since the flexion after should depend strongly on the flexion before. I observe that the treated observations are somewhat more bunched up in the higher end, but this needs to be confirmed more rigorously.

(b)

```
model0 = lm(faft ~ fbef + grp, hips)
summary(model0)
```

```
## 
## Call:
## lm(formula = faft ~ fbef + grp, data = hips)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -27.4050  -3.1586   0.0214   3.1850  21.3768 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 59.86141    7.77213   7.702 4.39e-11 ***
## fbef         0.49008    0.06949   7.053 7.39e-10 ***
## grptreat     7.02553    1.77085   3.967 0.000165 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 6.976 on 75 degrees of freedom
## Multiple R-squared:  0.5327, Adjusted R-squared:  0.5202 
## F-statistic: 42.74 on 2 and 75 DF,  p-value: 4.087e-13
```

I set the model as the following:

$$FAFT_i = \beta_0 + \beta_1 FBEF_i + \beta_2 GRP + \epsilon_i$$

This model allows for treatment effect on both the coefficient and the slope between `fbef` and `faft`. Since we have observed that there is a linear relationship between `fbef` and `faft` in general (and that it makes intuitive sense), we would need to see significant coefficient estimates for β_2 if there is a significant treatment effect.

The linear model using the categorical variable `grp` suggests that there is a significant correlation between `faft` and treatment status. As mentioned before, it is natural to have a strong correlation between `fbef` and `faft`, captured by β_1.

(c)

```
hips$fdiff = hips$faft - hips$fbef
model1 = lm(fdiff ~ grp, hips)
summary(model1)
```

```
## 
## Call:
## lm(formula = fdiff ~ grp, data = hips)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -18.481  -5.231  -0.792   2.441  41.519
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.792      1.854   2.045   0.0443 *
## grptreat       3.690      2.229   1.656   0.1019
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.084 on 76 degrees of freedom
## Multiple R-squared:  0.03481,    Adjusted R-squared:  0.02211
## F-statistic: 2.741 on 1 and 76 DF,  p-value: 0.1019
```

I try modeling the effect of treatment on the difference or the change in flexion. Note that the coefficient on treatment is no longer significant under this model. However, this way of comparison may be more desirable. First of all, we are able to basically model the same thing with less parameters, which reduces the degrees of freedom required. Furthermore, this model is more intuitive, since we can gauge the effect of the treatment on the difference by looking at the single coefficient on `grptreat`. Here, we are effectively assuming that the true coefficient of `fbef` is 1 and subtracting it from the response. With the earlier model, we may be capturing a temporal difference from `fbef` to `faft`. For instance, the severity of arthritis may exacerbate along with time. It would be more appropriate to see the improvement (i.e., the difference), or the lack thereof, after receiving the actual treatment or the control treatment.

(d)

I continue using the model with `fdiff` as it seems to be a more appropriate model.

```
# Studentized residual
rstud = as.data.frame(rstudent(model1))
colnames(rstud) = "rstud"
rstud[abs(rstud$rstud) > abs(qt(.05 / (78*2), 76)), ]
```

```
## [1] 4.397088 5.401276
```

Note that there are extreme outliers with studentized residuals that exceed the Bonferroni corrected bound.

```
hips[which(rstud$rstud > abs(qt(.05 / (78*2), 76))), ]
```

```
##    fbef faft rbef raft   grp  side person fdiff
## 49   78  121   35   34 treat right     25    43
## 50   77  126   30   32 treat  left     25    49
```

I confirm that the outliers are the two observations with `fbef` below 90. The two points happen to come from the same person, with ID number 25. Note that the rotation angles, `rbef` and `raft` have not changed much. Thinking about the context, it is extremely unnatural that a person with AS is suddenly able to flex their hips at a much higher angle (an improvement of 40-50 degrees), but not be able to rotate much.

```
hips[hips$fbef < 90, ]
```

```
##    fbef faft rbef raft     grp  side person fdiff
## 49   78  121   35   34   treat right     25    43
## 50   77  126   30   32   treat  left     25    49
## 70   81  111   14   13 control  left     35    30
```

It is also unnatural that the flexion angle is less than 90, which implies that the person cannot even sit up straight. I strongly suspect that these observations are corrupted (e.g., measurement error) and should be removed from the model.

```
hips_rmout = hips[hips$fbef >= 90, ]
model2 = lm(fdiff ~ grp, hips_rmout)
summary(model2)
```

```
##
## Call:
## lm(formula = fdiff ~ grp, data = hips_rmout)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##    -17     -4      0      3     19
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.652      1.209   2.193   0.0315 *
## grptreat       3.348      1.452   2.305   0.0240 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.799 on 73 degrees of freedom
## Multiple R-squared:  0.06786,    Adjusted R-squared:  0.05509
## F-statistic: 5.314 on 1 and 73 DF,  p-value: 0.024
```

After the removal, I now find that the treatment has a positive effect on improving flexion angles.

(e)

```
confint(model2, "grptreat", level = 0.95)
```

```
##             2.5 %   97.5 %
## grptreat 0.453448 6.242204
```

The estimated size of the effect is between 0.453448 and 6.242204. Obviously, the confidence interval also indicates that the coefficient is significantly above zero at the 5% level. Although it is statistically significant, we must really ask ourselves if the improvement is economically significant as well. For instance, if the cost of conducting (or receiving) a treatment is high, it may not be a good choice to spend the money to gain, at most, 6 degrees. Such qualitative judgment must be done in addition to this.

(f)

Each (left, right) pair of a subject will inevitably be dependent. For instance, the pair will be subject to the same life patterns (like exercising) and the diet of the person. Then, this will violate the independence assumption between errors. Since the model assumptions do not hold, we cannot trust any inference based on the model.

(g)

```
avg_fbef = aggregate(hips_rmout$fbef, list(hips_rmout$person), FUN = mean)
colnames(avg_fbef) = c("person", "fbef")
avg_faft = aggregate(hips_rmout$faft, list(hips_rmout$person), FUN = mean)
colnames(avg_faft) = c("person", "faft")
hips_avg = merge(avg_fbef, avg_faft, by = "person")
trt = unique(hips_rmout[, c("person", "grp")])
hips_avg = merge(hips_avg, trt, by = "person")
hips_avg$fdiff = hips_avg$faft - hips_avg$fbef
model3 = lm(fdiff ~ grp, hips_avg)
summary(model3)
```

```
##
## Call:
## lm(formula = fdiff ~ grp, data = hips_avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0000 -3.3750 -0.8333  3.0000 12.5000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.667      1.361   1.959   0.0578 .
## grptreat       3.333      1.645   2.026   0.0502 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.715 on 36 degrees of freedom
## Multiple R-squared:  0.1023, Adjusted R-squared:  0.07741
## F-statistic: 4.104 on 1 and 36 DF,  p-value: 0.05023
```

Since I have reason to believe that the observations with `fbef` higher than 90 are corrupted, I exclude them to run the new model. When I run the regression with individual averages, the coefficient is no longer significant. The estimate value is approximately the same. I suspect that the decrease in sample size accounts for this difference.

```
rstud2 = as.data.frame(rstudent(model3))
colnames(rstud2) = "rstud"
rstud2[abs(rstud2$rstud) > abs(qt(.05 / (36*2), 34)), ]
```

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

I find that the resulting regression fit has no significant outlier problems.

Problem 2

```
require(lattice)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:faraway':
##
##     melanoma
```

(a)

```
summary(barley)
```

```
##      yield            variety     year                 site
##  Min.   :14.43   Svansota :12   1932:60   Grand Rapids   :20
##  1st Qu.:26.88   No. 462  :12   1931:60   Duluth         :20
##  Median :32.87   Manchuria:12             University Farm:20
##  Mean   :34.42   No. 475  :12             Morris         :20
##  3rd Qu.:41.40   Velvet   :12             Crookston      :20
##  Max.   :65.77   Peatland :12             Waseca         :20
##                  (Other)  :48
```

Note that there are two levels for `year`.

```
levels(barley$variety)
```

```
##  [1] "Svansota"         "No. 462"          "Manchuria"        "No. 475"
##  [5] "Velvet"           "Peatland"         "Glabron"          "No. 457"
##  [9] "Wisconsin No. 38" "Trebi"
```

There are 10 levels for `variety`.

```
levels(barley$site)
```

```
## [1] "Grand Rapids"    "Duluth"          "University Farm" "Morris"
## [5] "Crookston"       "Waseca"
```

There are 6 levels for `site`.

Using up to the triple interaction term will lead to $(2-1)+(10-1)+(6-1)+(2-1)\times(10-1)+(10-1)\times(6-1)+(2-1)\times(6-1)+(2-1)\times(10-1)\times(6-1) = 119$ degrees of freedom used. That is, $(2-1)+(10-1)+(6-1) = 15$ degrees of freedom for each factor level, $(2-1)\times(10-1)+(10-1)\times(6-1)+(2-1)\times(6-1) = 59$ degrees of freedom for all the two-way interactions, and $(2-1)\times(10-1)\times(6-1) = 45$ degrees of freedom for the three-way interactions. Counting the intercept as an extra parameter will yield a total of 120 degrees of freedom.

Note that the residuals take on the degree of freedom equal to $n-p$, which in this case is $120-119-1 = 0$. That is, including up to the triple interaction terms will lead to a "perfect" fit, with a covariate made to correspond to each data point. There will be no room for residuals left to do any statistics.

(b)

As calculated in the previous part, only including up to two-way interactions will lead to $15 + 59 = 74$ degrees of freedom for the pameterizations. Including the intercept gives 75 degrees of freedom in total for the model.

This is better than having a “perfect” fit as in part b.

(c)

```
barley_red = barley[-c(23, 83), ]
mod = lm(yield ~ (site + year + variety)**2, barley_red)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: yield
##              Df Sum Sq Mean Sq  F value    Pr(>F)    
## site          5 6556.4 1311.28 113.4036 < 2.2e-16 ***
## year          1  912.1  912.10  78.8815 2.271e-11 ***
## variety       9 1080.3  120.04  10.3812 2.201e-08 ***
## site:year     5 2164.1  432.83  37.4323 8.807e-15 ***
## site:variety 44 1161.8   26.40   2.2835  0.003615 ** 
## year:variety  9  190.4   21.16   1.8298  0.089547 .  
## Residuals    44  508.8   11.56                       
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From the ANOVA table, I find that the interaction term between `year` and `variety` is not significant at the 5% level.

```
mod1 = lm(yield ~ site + year + variety + site:year + site:variety, barley_red)
mod1_reorder = lm(yield ~ site + year + variety + site:variety + site:year, barley_red)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: yield
##              Df Sum Sq Mean Sq F value    Pr(>F)    
## site          5 6556.4 1311.28 99.3969 < 2.2e-16 ***
## year          1  912.1  912.10 69.1387 3.525e-11 ***
## variety       9 1080.3  120.04  9.0990 3.674e-08 ***
## site:year     5 2164.1  432.83 32.8090 4.377e-15 ***
## site:variety 44 1161.8   26.40  2.0015  0.008104 ** 
## Residuals    53  699.2   13.19                      
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

After removal, `site:variety` still appears to be significant. We cannot judge from this table whether `site:year` should be removed. So I check it with a reordered ANOVA.

```
anova(mod1_reorder)
```

```
## Analysis of Variance Table
##
## Response: yield
##              Df Sum Sq Mean Sq F value    Pr(>F)
## site          5 6556.4 1311.28 99.3969 < 2.2e-16 ***
## year          1  912.1  912.10 69.1387 3.525e-11 ***
## variety       9 1080.3  120.04  9.0990 3.674e-08 ***
## site:variety 44 1161.8   26.40  2.0015  0.008104 **
## site:year     5 2164.1  432.83 32.8090 4.377e-15 ***
## Residuals    53  699.2   13.19
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Both the interaction terms between `site` and `year`, and `site` and `variety` appear significant at the 5% level. It seems better to keep these predictors in the model.

Problem 3

(a)

```
avg_bright = aggregate(pulp$bright, list(pulp$operator), FUN = mean)
colnames(avg_bright) = c("operator", "alphahat")
avg_bright
```

```
##   operator alphahat
## 1        a    60.24
## 2        b    60.06
## 3        c    60.62
## 4        d    60.68
```

The group mean for group A ($\hat{\alpha}_A$) is 60.24; for group B ($\hat{\alpha}_B$) is 60.06; for group C ($\hat{\alpha}_C$) is 60.62; and for group D ($\hat{\alpha}_D$) is 60.68.

(b)

```
alphahat_a = avg_bright$alphahat[avg_bright$operator == "a"]
alphahat_b = avg_bright$alphahat[avg_bright$operator == "b"]
alphahat_c = avg_bright$alphahat[avg_bright$operator == "c"]
alphahat_d = avg_bright$alphahat[avg_bright$operator == "d"]
RSS_a = sum((pulp$bright[pulp$operator == "a"] - alphahat_a)**2)
RSS_b = sum((pulp$bright[pulp$operator == "b"] - alphahat_b)**2)
RSS_c = sum((pulp$bright[pulp$operator == "c"] - alphahat_c)**2)
RSS_d = sum((pulp$bright[pulp$operator == "d"] - alphahat_d)**2)
(sigmahat = sqrt((RSS_a + RSS_b + RSS_c + RSS_d) / (20 - 4)))
```

```
## [1] 0.3259601
```

(c)

Note that

$$\hat{\alpha}_A \sim N(\alpha_A, \frac{\sigma^2}{J_A}),\ \hat{\alpha}_B \sim N(\alpha_B, \frac{\sigma^2}{J_B})$$

where J_A is the group size for group A. Then,

$$\hat{\alpha}_A - \hat{\alpha}_B \sim N(\alpha_A - \alpha_B, \sigma^2(\frac{1}{J_A} + \frac{1}{J_B}))$$

assuming $\hat{\alpha}_A \perp\!\!\!\perp \hat{\alpha}_B$.

$$var[\hat{\alpha}_A - \hat{\alpha}_B] = \sigma^2(\frac{1}{J_A} + \frac{1}{J_B})$$

$$\begin{align}
\sqrt{var[\hat{\alpha}_A - \hat{\alpha}_B]} &= \sigma\sqrt{\frac{1}{J_A} + \frac{1}{J_B}} \tag{1}\\
&= \sigma\sqrt{\frac{1}{5} + \frac{1}{5}} \tag{2}\\
&= \sigma\sqrt{\frac{2}{5}} \tag{3}
\end{align}$$

Replacing σ with $\hat{\sigma}$ gives:

$$SE(\hat{\alpha}_A - \hat{\alpha}_B) = \hat{\sigma}\sqrt{\frac{2}{5}}$$

Similarly for all other combinations,

$$SE(\hat{\alpha}_B - \hat{\alpha}_C) = SE(\hat{\alpha}_C - \hat{\alpha}_D) = SE(\hat{\alpha}_D - \hat{\alpha}_A) = SE(\hat{\alpha}_A - \hat{\alpha}_C) = SE(\hat{\alpha}_B - \hat{\alpha}_D) = \hat{\sigma}\sqrt{\frac{2}{5}}$$

which is all the 6 possible combinations.

Actually calculating the standard error gives:

```r
(std_error = sigmahat * sqrt(2/5))
```

```
## [1] 0.2061553
```

(d)

```r
diffmat = outer(avg_bright$alphahat, avg_bright$alphahat, '-')
q = qtukey(0.95, 4, 20 - 4)
confint_AB = c(diffmat[1, 2] - q * std_error / sqrt(2), diffmat[1, 2] + q * std_error / sqrt(2))
confint_AC = c(diffmat[1, 3] - q * std_error / sqrt(2), diffmat[1, 3] + q * std_error / sqrt(2))
confint_AD = c(diffmat[1, 4] - q * std_error / sqrt(2), diffmat[1, 4] + q * std_error / sqrt(2))
confint_BC = c(diffmat[2, 3] - q * std_error / sqrt(2), diffmat[2, 3] + q * std_error / sqrt(2))
confint_BD = c(diffmat[2, 4] - q * std_error / sqrt(2), diffmat[2, 4] + q * std_error / sqrt(2))
confint_CD = c(diffmat[3, 4] - q * std_error / sqrt(2), diffmat[1, 2] + q * std_error / sqrt(2))
(confints = rbind(confint_AB, confint_AC, confint_AD, confint_BC, confint_BD, confint_CD))
```

```
##                   [,1]        [,2]
## confint_AB -0.4098143  0.76981435
## confint_AC -0.9698143  0.20981435
## confint_AD -1.0298143  0.14981435
## confint_BC -1.1498143  0.02981435
## confint_BD -1.2098143 -0.03018565
## confint_CD -0.6498143  0.76981435
```

Using confidence intervals based on Tukey's honestly significant difference, I can conclude that only groups B and D are substantially different from each other at the 5% level.

Problem 4

Assuming there are patients that receive no medication as a control group, their average blood pressure should be 150. Thus, the intercept $\beta_0 = 150$.

Since we are given that no one drug on its own will have a significant effect, the coefficients on the single factor levels will all be zero. That is,

$$\beta_{A1} = \beta_{B1} = \beta_{C1} = 0$$

However, using A with B, and A with C, should reduce the average blood pressure by 10. That is,

$$\beta_{A1:B1} = \beta_{A1:C1} = -10,$$

where as using B and C together is equivalent using just one of the either: $\beta_{B1:C1} = 0$.

Also, using all A, B, and C in combination is not different from using just A and B or A and C. That is, it does not improve upon the two-way interactions (or has no marginal effect). That is,

$$\beta_0 + \beta_{A1} + \beta_{B1} + \beta_{A1:B1} = \beta_0 + \beta_{A1} + \beta_{B1} + \beta_{C1} + \beta_{A1:B1} + \beta_{B1:C1} + \beta_{C1:A1} + \beta_{A1:B1:C1}$$

$$\Rightarrow 150 + 0 + 0 - 10 = 150 + 0 + 0 + 0 - 10 + 0 - 10 + \beta_{A1:B1:C1}$$

$$\therefore \beta_{A1:B1:C1} = 10$$

This makes intuitive sense because we have to adjust for having all three drugs. Since `A1:B1` and `A1:C1` are both included, we are exaggerating the effect of the combination. Hence, we must readjust using the three-way interaction term.

In summary,

$$\beta_{A1} = 0 \quad (4)$$

$$\beta_{B1} = 0 \quad (5)$$

$$\beta_{C1} = 0 \quad (6)$$

$$\beta_{A1:B1} = -10 \quad (7)$$

$$\beta_{A1:C1} = -10 \quad (8)$$

$$\beta_{A1:B1:C1} = 10 \quad (9)$$

Problem 5

(a)

We are given that `S` has 3 levels and `T` has 2 levels. `I` is a continuous variable that will consume one degree of freedom as usual. The number of parameters for `S` is $3 - 1 = 2$ and for `T` is $2 - 1 = 1$. For the two-way interaction terms, `S:T` will have $(3-1)\times(2-1)=2$, `S:I` will have $(3-1)\times 1 = 2$, and `T:I` will have $(2-1)\times 1 = 1$. Thus, $p = 1+1+2+1+2+2+1 = 10$, where the first 1 is accounting for the intercept.

From the residual degree of freedom, $n - p = 65$. Combining the two pieces of information gives: $n = 65 + 10 = 75$.

(b)

Note that

$$F = \frac{(RSS_{Reduced} - RSS_{Full})/k}{RSS_{Full}/(n-p)} \sim F_{k,n-p}$$

Here we are removing 5 covariates (or degrees of freedom) to go from the full model to the reduced model. Hence, $k = 5$. Also, we have $n - p = 65$ from part a. Then, we have

$$F \sim F_{5,65}$$

From the first ANOVA output, we are given that the sum of squared residuals for the full model (i.e. RSS_{Full}) is 2496. From the ANOVA output given in this part of the problem, we can find that the sum of squared residuals for the reduced model (i.e., $RSS_{Reduced}$) is 5013. Calculating our test statistic gives the following:

```
rss_full = 2496
rss_reduced = 5013
k = 5
n_p = 65
(F_stat = ((rss_reduced - rss_full) / k) / (rss_full / n_p))
```

```
## [1] 13.10938
```

(c)

For the sake of this problem, I will focus on `I` and `S` and ignore other covariates for now. The simplified model will look like the following:

$$Y_i = \beta_0 + \beta_I I_i + \beta_{S2} 1_{i,S2=1} + \beta_{S3} 1_{i,S3=1} + \beta_{I:S2} I_i : 1_{i,S2=1} + \beta_{I:S3} I_i : 1_{i,S3=1} + \epsilon_i$$

Consider the case when $S2 = 1$ or $1_{i,S2=1} = 1$. Then, the model boils down to the following:

$$Y_i = \beta_0 + \beta_I I_i + \beta_{S2} + \beta_{I:S2} I_i + \epsilon_i$$

$$\Rightarrow Y_i = (\beta_0 + \beta_{S2}) + (\beta_I + \beta_{I:S2}) I_i + \epsilon_i$$

Since we have $\beta_I \approx -\beta_{I:S2}$,

$$\Rightarrow Y_i = (\beta_0 + \beta_{S2}) + \epsilon_i$$

That is, the variable `I` has no explanatory power on our response when $S2 = 1$. To put this into context, the parental income does not explain a child's reading ability when the type of school the child attends is 2. There can be many different scenarios under which this can happen. Based on the coefficient estimates, my best guess is that the type 2 schools are poor schools, which provide relatively less rigorous programs. In that case, the poor quality of schooling will lead to the student having a less competent level of reading proficiency (inferred from the negative coefficient on `S2`), regardless of how much the parents make. That is, the parental income does not contribute any additional explanatory power of significance when the student goes to a low-quality school.