Code ▾

36-309 / 36-749 Final Exam (Fall 2022)

Due Thursday, December 15, 12:00pm

Sanaz Saadatifar

FINAL EXAM INSTRUCTIONS: MUST READ

- The format of the final is identical to that of the homeworks and midterms: You do everything in RStudio and turn in a PDF on Gradescope by 12:00pm on Thursday, December 15.
- During the exam, you are not allowed to talk with peers about 36-309/36-749 material. You will be asked to sign an academic integrity statement below - you must sign this statement to receive any credit for the exam.
- Although you can’t talk with peers, the exam is still “open everything.” Remember that you can always refer back to previous homework/lab/midterm solutions, lectures, R demos, and the course textbook.
- **Throughout the exam, include all the R code you used to arrive at your answers.** For example, if a question asks you to run a statistical analysis or make a graph, you should include the code that runs that analysis or makes that graph. This is equivalent to “showing your work” in other classes - if you don’t show your work, we won’t know how you arrived at your answers.
- Email Professor Branson if you have any **clarifying** questions about the exam (e.g., if you think there are any typos or ambiguities in the question). However, any emails sent after 3pm on Wednesday, December 14 are not guaranteed to be answered.

Academic Integrity Statement: MUST SIGN TO RECEIVE CREDIT FOR EXAM

“By writing my name below, I certify that I have not talked with anyone else (other than Professor Branson) about 36-309/36-749 material from December 13 to December 15.”

[Sanaz Saadatifar]

Question 1: Full STEM Ahead (49 points)

In this problem, we'll work with data that mimics a study I've worked on in psychology to address disparities in STEM (Science, Technology, Engineering, and Mathematics). The goal of the study was to assess if educational interventions could improve interest in pursuing science among 4th and 5th graders from underrepresented backgrounds. (For the purpose of this study, "underrepresented background" was defined as non-white and non-Asian.) Here's the dataset:

Hide

```
stemData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/main/stemData.csv")
#ensure categorical variables are factors
stemData$underrep = factor(stemData$underrep)
stemData$treatment = factor(stemData$treatment)
```

In the study, 4th and 5th graders at several elementary schools in Pennsylvania were randomized to one of three programs: A "control" program where students completed their typical curriculum; a "standard STEM" program that had previously been designed to increase students' interest in science; and a "new STEM" program developed to increase interest in science for underrepresented students specifically. Students participated in the program for a year. At the end of the study, students took a standardized test measuring their knowledge in science at an elementary-school level. The dataset includes the following variables:

- `testScore` : The student's test score, from 0 to 100.
- `income` : The student's family's annual income before the study started. This was measured in thousands of dollars (i.e., `income = 1` denotes $1000).
- `underrep` : Whether or not the student was from an underrepresented background ("yes" or "no").
- `treatment` : The program the student was randomized to (either "control", "standSTEM", or "newSTEM").

a. (14pts) We'll first perform some initial EDA and analyses for these data. Professor L. Hill (https://en.wikipedia.org/wiki/Lauryn_Hill) makes the following initial statements about the data:

STATEMENT 1: "It's well-known that many societal disparities lead to different income levels between underrepresented and not-underrepresented students, at least on average. I would guess that there is evidence of this difference in our data."

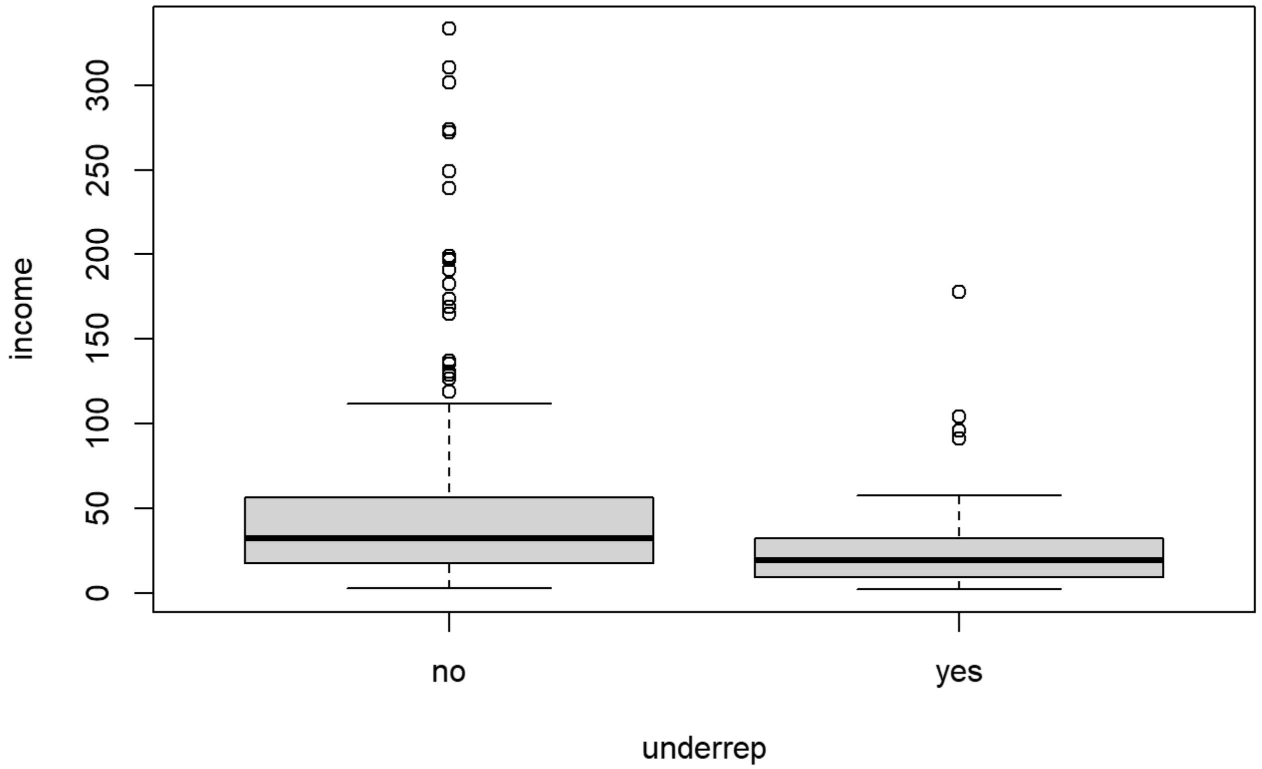
STATEMENT 2: "It'll be important to account for income levels in our analyses, because it's well-known that income tends to be positively associated with test scores. I would guess that there is evidence of this positive association in our data.

We'll assess these statements in this question. For this part, answer the following **FOUR** questions.

- (3pts) First, make one form of **graphical EDA** that allows you to assess STATEMENT 1, and explain in 1-2 sentences how that EDA allows you to assess STATEMENT 1.

Hide

```
boxplot(income~underrep, data = stemData)
```



[The boxplot would be the appropriate EDA as it would show the mean values of income and its dispersion for two different groups of underrep. also th eoutcome here is quantitaive and the Underrep is categorial data so boxplot can be used]

- (4pts) Now conduct a **statistical analysis** that allows you to assess STATEMENT 1. After running your analysis, state whether you Agree or Disagree with STATEMENT 1 for this dataset, and explain in 1-2 sentences. In your answer, be sure to state your **scientific conclusion** for that analysis, and how those scientific conclusions lead you to Agreeing or Disagreeing with STATEMENT 1. (**Hint**: By “statistical analysis,” I mean an analysis that produces a p-value.)

Hide

```
t.test(income~underrep, data = stemData)
```

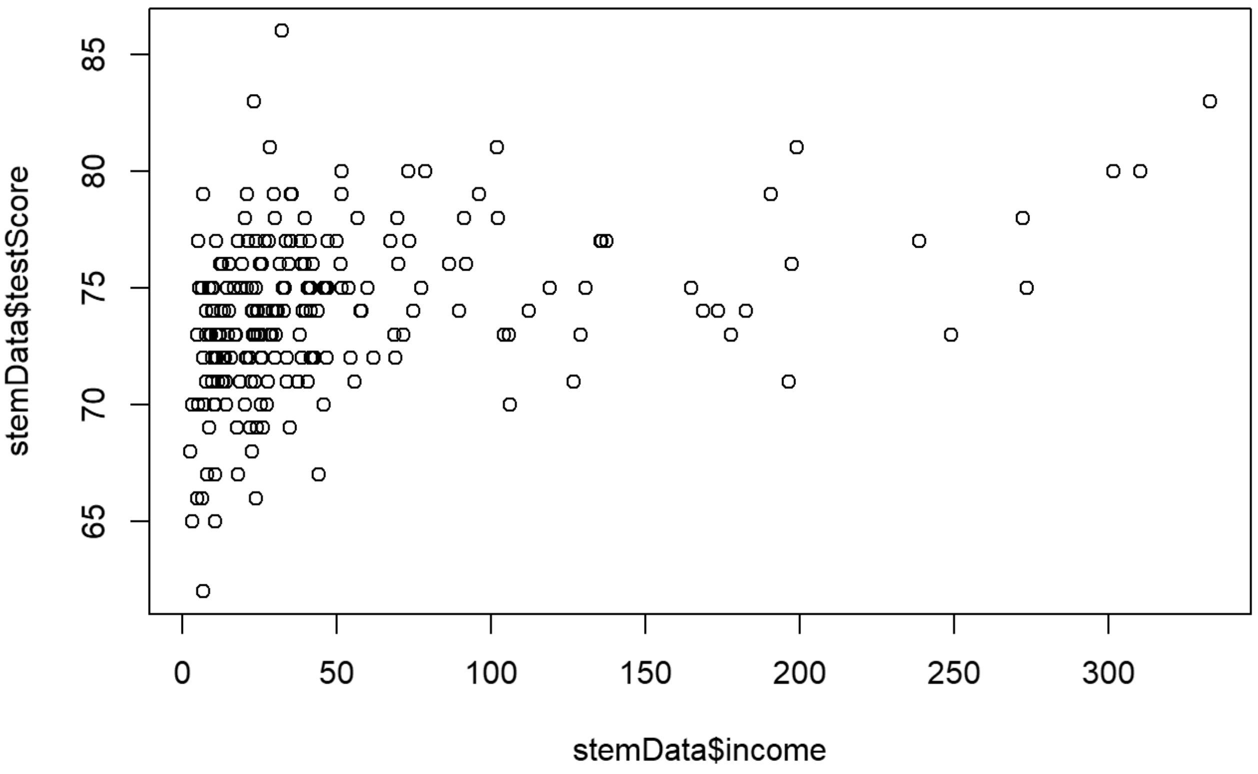
```
##
##  Welch Two Sample t-test
##
## data:  income by underrep
## t = 4.004, df = 164.86, p-value = 9.398e-05
## alternative hypothesis: true difference in means between group no and group yes is not
equal to 0
## 95 percent confidence interval:
##  13.19137 38.85753
## sample estimates:
##  mean in group no mean in group yes
##          54.22345          28.19900
```

[The null hypothesis is that there is no difference between underrepresented and not underrepresented groups' mean income. We can reject the null hypothesis because P-value is less than 0.05 (9.398e-05). Groups who are not underrepresented, have higher income.]

- (3pts) Now make one form of **graphical EDA** that allows you to assess STATEMENT 2, and explain in 1-2 sentences how that EDA allows you to assess STATEMENT 2.

Hide

```
plot(stemData$income, stemData$testScore)
```



[Both outcome and explanatory variables are quantitative so scatterplot can be used. And it can show the relationship and trend between two variables.]

- (4pts) Now conduct a **statistical analysis** that allows you to assess STATEMENT 2. After running your analysis, state whether you Agree or Disagree with STATEMENT 2 for this dataset, and explain in 1-2 sentences. In your answer, be sure to state your **scientific conclusion** for that analysis, and how those scientific conclusions lead you to Agreeing or Disagreeing with STATEMENT 2.

Hide

```
summary(lm(testScore~income, data = stemData))
```

```
##
## Call:
## lm(formula = testScore ~ income, data = stemData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0132  -2.0718  -0.0287   1.9225  12.4552
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.874014   0.276230 263.816  < 2e-16 ***
## income       0.020768   0.003636   5.711 3.53e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.204 on 225 degrees of freedom
## Multiple R-squared:  0.1266, Adjusted R-squared:  0.1227
## F-statistic: 32.62 on 1 and 225 DF,  p-value: 3.532e-08
```

[we see that the p-value for the income slope estimate is very small and certainly less than 0.05 (3.53e-08); thus, we reject the null hypothesis that the slope is zero and conclude that there is a significant positive association between income and score - i.e., we indeed conclude that an increased income is associated with a higher score, on average.]

b. (8pts) Now Professor Hill would like to assess the **internal validity** of this study. For this part, answer the following questions:

- (5pts) First, conduct **two** statistical analyses that allow you to assess internal validity for the `treatment` variable. Each analysis should only involve **one** explanatory variable (other than the `treatment`). After running your analyses, explain in 1-2 sentences how your analyses allow you to assess internal validity. Then, state whether you think this study has Low or High internal validity, and explain in 1-2 sentences.

Hide

```
oneWay = aov(income~treatment, data = stemData)
summary(oneWay)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## treatment     2   4677    2338   0.679  0.508
## Residuals   224 771407    3444
```

Hide

```
chisq.test(table(stemData$treatment, stemData$underrep))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(stemData$treatment, stemData$underrep)
## X-squared = 5.5734, df = 2, p-value = 0.06162
```

[The explanatory variables that can be used to assess the internal validity for treatment are income and underrep. So I need to test these two variables across treatment groups. since income is quantitative and treatment is categorial with 3 levels, then ANOVA is used. And since both treatment and underrep are categorial CHI-squared test needed to be used. The results of ANOVA showed that P-value is above 0.05 (0.508) which we cannot reject the null hypothesis of income is same across treatment groups. so in terms of income we conclude that internal validity holds. But for checking internal validity in terms of proportion of underrep across different treatmenmt groups, the CHI.square test results showed that p-value is above 0.05 (0.061) so again we cannot reject the null that treatment and underrep are independent of each other. So again the internal validity holds because the proportion is not necessarily different across treatment group.]

- (3pts) Professor Hill decides to make the following contingency table:

Hide

```
table(stemData$treatment)
```

```
##
##   control   newSTEM standSTEM
##        76        79        72
```

Professor Hill says, “We can see that an unequal number of subjects were assigned to each treatment group, which suggests a lack of internal validity for this study.” Do you Agree or Disagree with this statement? Explain in 1-2 sentences.

[I disagree because there is not a significant difference between them in an extent that would violate the internal validity.]

c. (14pts) Let’s consider three linear regression models, which we’ll refer to as Model 1, Model 2, and Model 3:

Hide

```
model1 = lm(testScore ~ treatment + income, data = stemData)
model2 = lm(testScore ~ treatment + income + underrep, data = stemData)
model3 = lm(testScore ~ treatment + income + underrep + underrep*treatment, data = stemDat
a)
```

For this question, we’ll focus on **Model 2**. Answer the following questions.

- (4pts) Let’s look at the `summary()` output from Model 2:

Hide

```
summary(model2)
```

```
##
## Call:
## lm(formula = testScore ~ treatment + income + underrep, data = stemData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0871  -1.8583  -0.0799   1.9194  11.7614
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        71.580794   0.427503 167.439  < 2e-16 ***
## treatmentnewSTEM    1.615143   0.498896   3.237 0.001391 **
## treatmentstandSTEM  2.022868   0.515192   3.926 0.000115 ***
## income              0.020951   0.003589   5.837 1.87e-08 ***
## underrepyes         0.365926   0.511212   0.716 0.474866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.103 on 222 degrees of freedom
## Multiple R-squared:  0.1913, Adjusted R-squared:  0.1768
## F-statistic: 13.13 on 4 and 222 DF,  p-value: 1.289e-09
```

Given this output, what is the **statistical modeling equation** for Model 2? By "statistical modeling equation," I mean the equation for the **mean** of the Normal distribution involved in the statistical model for Model 2. Please use ß ("beta") notation when writing your statistical model, and clearly define any other notation you need to write the modeling equation. (**Hint**: Your answer should not involve actual numbers, but the `summary()` output may still be helpful in correctly writing out the modeling equation.)

[TestScore mean = βˆ 0 + βˆ I * income + βˆ yes * underrepyes + βˆ ts * treatmentstandSTEM + βˆ tn * treatmentnewSTEM] [TestScore mean = 71.580794 + 0.020951 * income + 0.365926* underrepyes + 2.022868* treatmentstandSTEM + 1.615143* treatmentnewSTEM] [βˆ 0 is intercept coefficient. βˆ I is the slop for income. βˆ yes is the coefficient for the time when underrep is "yes". βˆ ts is coefficient when the treatment is Standard STEM. βˆ tn is coefficient when the treatment is New STEM.]

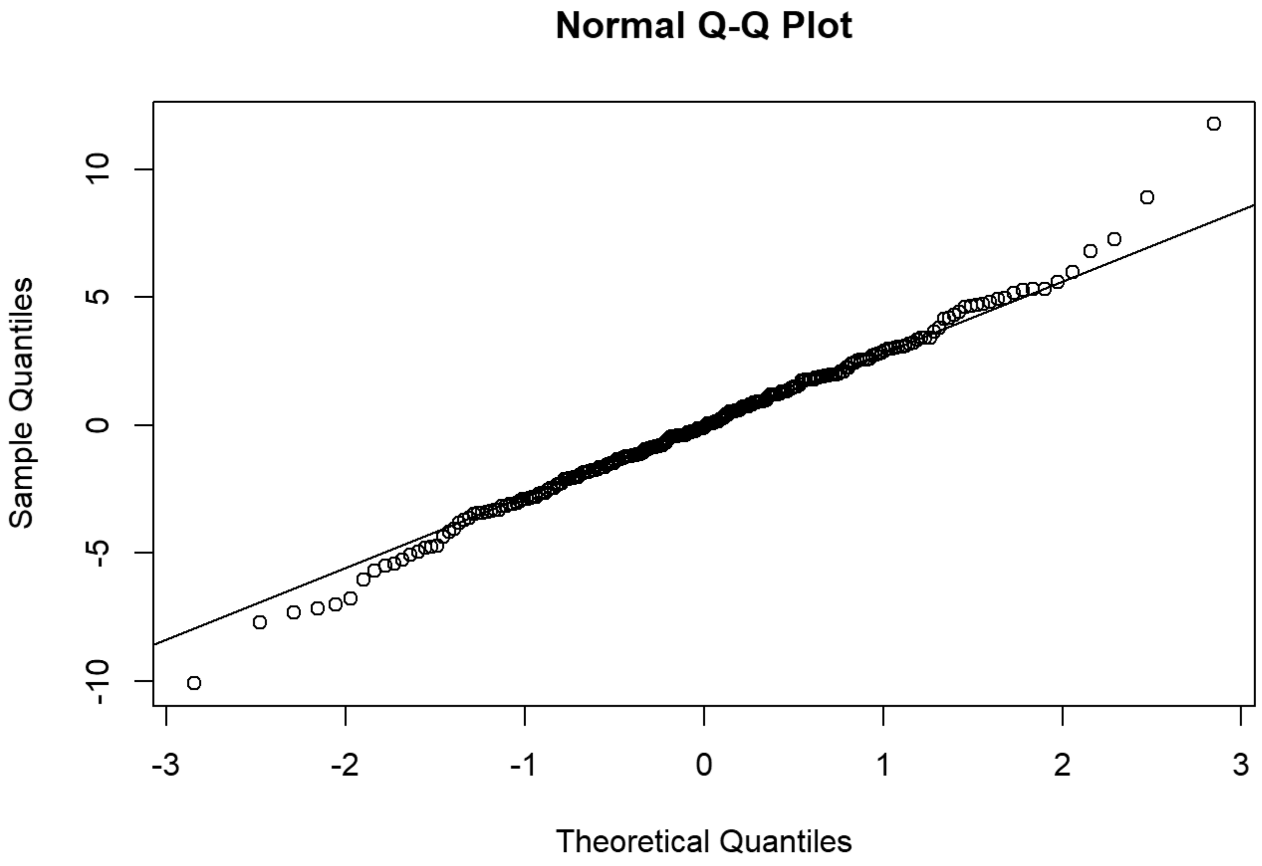
- (4pts) Professor Hill notices that the estimated Intercept for Model 2 is 71.58. What is your interpretation of this estimate, within the context of this dataset? Furthermore, what scientific conclusion can we make based on the p-value for the Intercept, within the context of this dataset?

[intercept is the mean test score when the students is not under represented and is in the control group with the income of 0. Scientific conclusion for the P-value is that since P-value is less than 0.05 (< 2e-16), we can reject the null hypothesis saying that intercept coefficient equals to zero.]

- (6pts) Model 2 is a linear regression model; as we've discussed in class, there are particular assumptions implicitly made when running a linear regression model. Make the appropriate residual analysis plot(s) you need to adequately assess three assumptions: The Normality assumption, the Equal Variance assumption, and the Linearity assumption. Then, for each assumption, state whether you believe it is Very Plausible, Somewhat Plausible, Somewhat Not Plausible, or Very Not Plausible for Model 2, based on your plot(s). (**Hint**: For the sake of this question, you do not need to add color to your plot(s).)

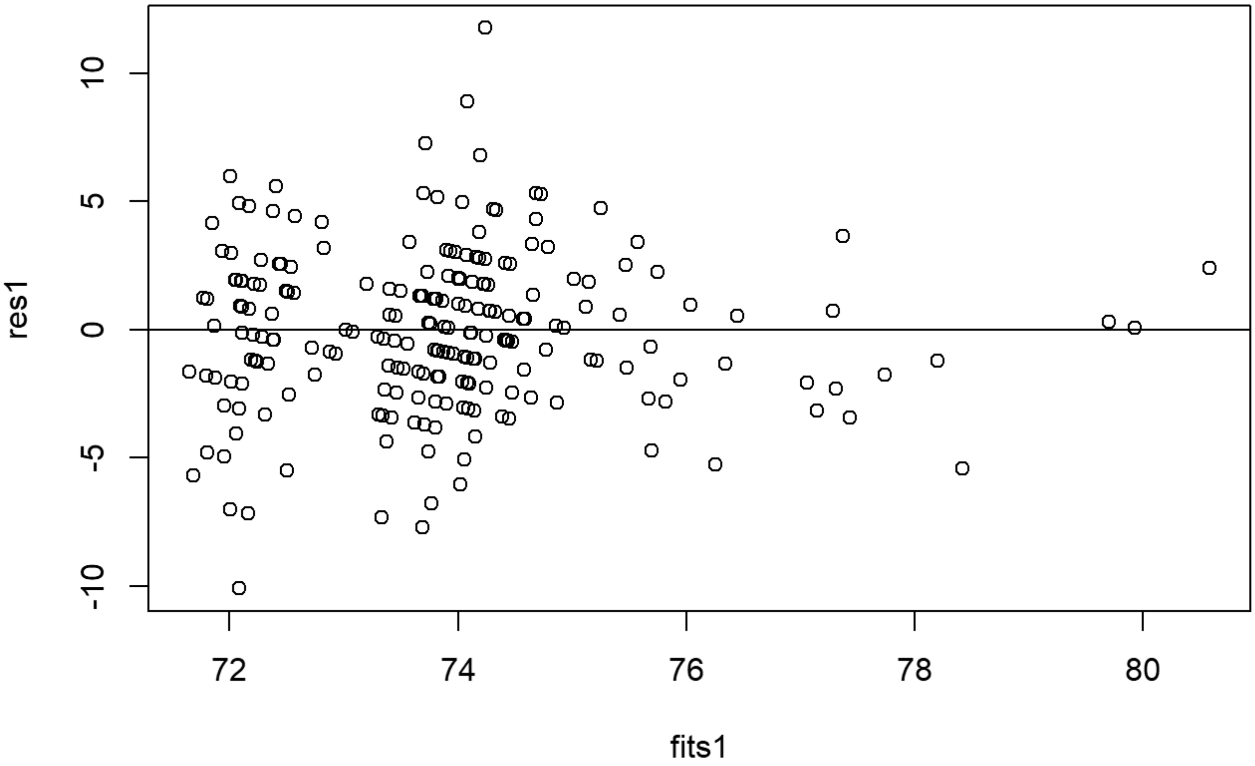
Hide

```
#residuals
res1 = residuals(model2)
#fits
fits1 = fitted(model2)



qqnorm(res1)
qqline(res1)
```



Hide

```
plot(fits1, res1)
abline(h=0)
```



[I think the normality is very plausible because most of point align on the QQ plot line. The equal variance does not hold because the points are concentrated on the left, so it is Somewhat not plausible. The linearity also does not hold and it is Somewhat not plausible because points form a cones shape.]

d. (13pts) Let’s continue considering Models 1, 2, and 3 from Part C. For this part, answer the following questions.

- (4pts) Professor Hill would like to understand which model she should use for inference. In particular, Professor Hill asks the following questions:

1. Should I prefer Model 2 over Model 1?
2. Should I prefer Model 3 over Model 2?
3. Should I prefer Model 3 over Model 1?

Write code that allows you to answer **all** of Professor Hill’s questions. Then, explain in 1-3 sentences how you can use your code’s output to answer Professor Hill’s questions. Please also answer Professor Hill’s questions in your explanation. (**Hint**: It’s okay if you have to write more than one line of code.)

Hide

```
anova(model1, model2, model3)
```

```
## Analysis of Variance Table
##
## Model 1: testScore ~ treatment + income
## Model 2: testScore ~ treatment + income + underrep
## Model 3: testScore ~ treatment + income + underrep + underrep * treatment
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
## 1    223 2142.9                               
## 2    222 2138.0  1     4.934 0.5270 0.46865  
## 3    220 2059.9  2    78.024 4.1664 0.01675 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
anova(model1, model3)
```

```
## Analysis of Variance Table
##
## Model 1: testScore ~ treatment + income
## Model 2: testScore ~ treatment + income + underrep + underrep * treatment
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
## 1    223 2142.9                               
## 2    220 2059.9  3    82.958 2.9533 0.03347 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[Since in the first anova model, the last row 3’s p-value is less than 0.05 (0.01675) then we prefer model 3 over model 2. Also since the p-value for the row 2 is greater than 0.05 (0.46865), then we prefer model 1 over model 2. To test the model 3 to model 1 I write the last line of code, which shows that p-value is less than 0.05 (0.03347) so we prefer model 3 over model 1. Overall model 3 is preferred.]

- (5pts) Regardless of your answer, Professor Hill would like to focus on Model 3. She provides the following `summary()` output:

Hide

```
summary(model3)
```

```
##
## Call:
## lm(formula = testScore ~ treatment + income + underrep + underrep *
##     treatment, data = stemData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0940 -2.0327 -0.0346  2.0756 10.5010
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   71.951646   0.451650 159.309  < 2e-16 ***
## treatmentnewSTEM               0.800160   0.573599   1.395  0.16443
## treatmentstandSTEM             1.770500   0.562305   3.149  0.00187 **
## income                         0.020694   0.003544   5.839 1.87e-08 ***
## underrepyes                   -0.996252   0.801247  -1.243  0.21505
## treatmentnewSTEM:underrepyes    3.075022   1.114817   2.758  0.00630 **
## treatmentstandSTEM:underrepyes  0.526670   1.351957   0.390  0.69724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.06 on 220 degrees of freedom
## Multiple R-squared:  0.2208, Adjusted R-squared:  0.1996
## F-statistic: 10.39 on 6 and 220 DF,  p-value: 3.877e-10
```

Looking at the above output, Professor Hill makes the following two statements:

STATEMENT A: “This regression analysis suggests that, for all students (regardless of whether they are from an underrepresented background or not), the *standard* STEM program increases test scores, compared to the control.”

STATEMENT B: “This regression analysis suggests that, for all students (regardless of whether they are from an underrepresented background or not), the *new* STEM program increases test scores, compared to the control.”

For each statement, state whether you Agree or Disagree. Then, explain in 1-3 sentences. (**Hint**: The above regression analysis has four coefficients involving `treatment` in some way. For this question, you have to consider all four coefficients.)

[I agree with both statement because keeping all other factors same, in statement A and B I calculated the coefficients. In statement A, since the coefficients for standard is positive (1.770500 and 0.526670) even if the negative underpayer’s coefficient (-0.996) is deducted again the result is above 0. So it increases the test scores. For the statement B as well, since the coefficients for New is positive (0.80 and 3.07) even if the negative underpayer’s coefficient (-0.996) is deducted again the result is above 0. So it increases the test scores.]

- (4pts) To wrap up this question, Professor Hill says the following:

“The results of this study are interesting. However, we’ve only found evidence that the STEM programs increase test scores, but I’m not sure that means that the programs increase students’ actual interest in science. Also, it’s worth mentioning that this study only involved 4th and 5th graders; it’s unclear if these programs would work for older students.”

Consider the following 36-309/36-749 concepts:

1. The independence assumption.
2. Correlation does not imply causation.
3. Internal validity.
4. External validity.
5. Construct validity.
6. Sphericity.

Which of the above concepts, if any, are relevant to Professor Hill’s statement here? Explain in 1-3 sentences.

[The construct validity and external validity. Construct validity in terms of test score is not a good criteria to assess the interest of students, and external validity in terms of not being able to generalize the results of this study to students other than 4th and 5th graders.]

Question 2: I See You, ICU (19 points)

For this question we’ll again work with data mimicking real data I’ve worked with in my research. Here’s the dataset:

Hide

```
icuData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/main/icuData.csv")
#ensure categorical variables are factors
icuData$dead28 = factor(icuData$dead28)
icuData$admittedICU = factor(icuData$admittedICU)
```

The dataset contains information on 1000 hospital patients that were recommended to be admitted to the intensive care unit (ICU). However, only some patients were ultimately admitted to the ICU, because there was limited space in the ICU. Here are the variables in this dataset:

- `admittedICU` : Whether or not someone was admitted to the ICU (1 if admitted and 0 if not). This is the treatment variable.
- `dead28` : Equal to 1 if someone died 28 days after they came to the hospital, and 0 if they were alive. In other words, this is the survival status 28 days later (https://en.wikipedia.org/wiki/28_Days_Later). This is the outcome variable.
- `icnarc` : A risk score, where a higher score denotes a higher health risk according to the patient’s physiological measurements taken by their doctors/nurses. It was developed by the Intensive Care National Audit and Research Center (ICNARC), hence the name. This is an explanatory variable.

a. (9pts) First, consider the following logistic regression model:

Hide

```
summary(glm(admittedICU ~ icnarc, family = "binomial", data = icuData))
```

```
##
## Call:
## glm(formula = admittedICU ~ icnarc, family = "binomial", data = icuData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1394  -0.9462  -0.7590   1.1793   1.9633
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.769907   0.169108 -10.466   <2e-16 ***
## icnarc       0.084074   0.009646   8.716   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1339.3  on 999  degrees of freedom
## Residual deviance: 1254.2  on 998  degrees of freedom
## AIC: 1258.2
##
## Number of Fisher Scoring iterations: 4
```

For this part, answer the following two questions about the above output.

- (4pts) The above output includes two coefficient estimates. Write your interpretation of each coefficient estimate on the *odds scale*.

[The first coefficient is intercept's coefficient that represents that odds of addmittedICU for patients with icnarc = 0 is exp(-1.769907). moreover, we see that the icnarc coefficient is 0.084074. The interpretation of this coefficient is: For every one-unit increase in icnarc, the odds of addmittedICU are multiplied by exp(0.084074) = 1.087709 [Thus, the odds of addmittedICU increase slightly for every one-unit increase in icnarc.]]

- (5pts) Does the above output tell us anything about the internal validity of this study? What about external validity? For each, state Yes or No, and explain in 1-2 sentences.

[Yes, it tells us something about both internal and external validity. The above output shows that there is a positive relation between the icnarc and the addmittedICU. So as icnarc gets higher, the probablaity of being admitted to ICU gets higher. So when the addmittedICU=1, this groups mean icnarc will be higher compared to the group with addmittedICU = 0. So this violates the internal validity. About the external validity since this is a regression model, we can generalize that high icnarc gets higher chance generally to be admitted to the ICU.]

b. (10pts) Now let's consider the following logistic regression model:

Hide

```
summary(glm(dead28 ~ admittedICU + icnarc, family = "binomial", data = icuData))
```

```
##
## Call:
## glm(formula = dead28 ~ admittedICU + icnarc, family = "binomial",
##     data = icuData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4062  -0.7342  -0.5631  -0.4269   2.2093
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.64345    0.20274 -13.039  < 2e-16 ***
## admittedICU1  0.44688    0.16255   2.749  0.00598 **
## icnarc        0.07351    0.01083   6.785 1.16e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1071.25  on 999  degrees of freedom
## Residual deviance:  998.79  on 997  degrees of freedom
## AIC: 1004.8
##
## Number of Fisher Scoring iterations: 4
```

For this part, answer the following three questions about the above output.

- (3pts) According to this logistic regression model, what is the estimated *log-odds of death* for someone admitted to the ICU and has an ICNARC score of 10? Be sure to show work for how you arrived at your answer.

[log-odds of death = -2.64345 + 0.44688* admittedICU1 + 0.07351* icnarc = -2.64345 + 0.44688* 1 + 0.07351* 10 = -1.46147]

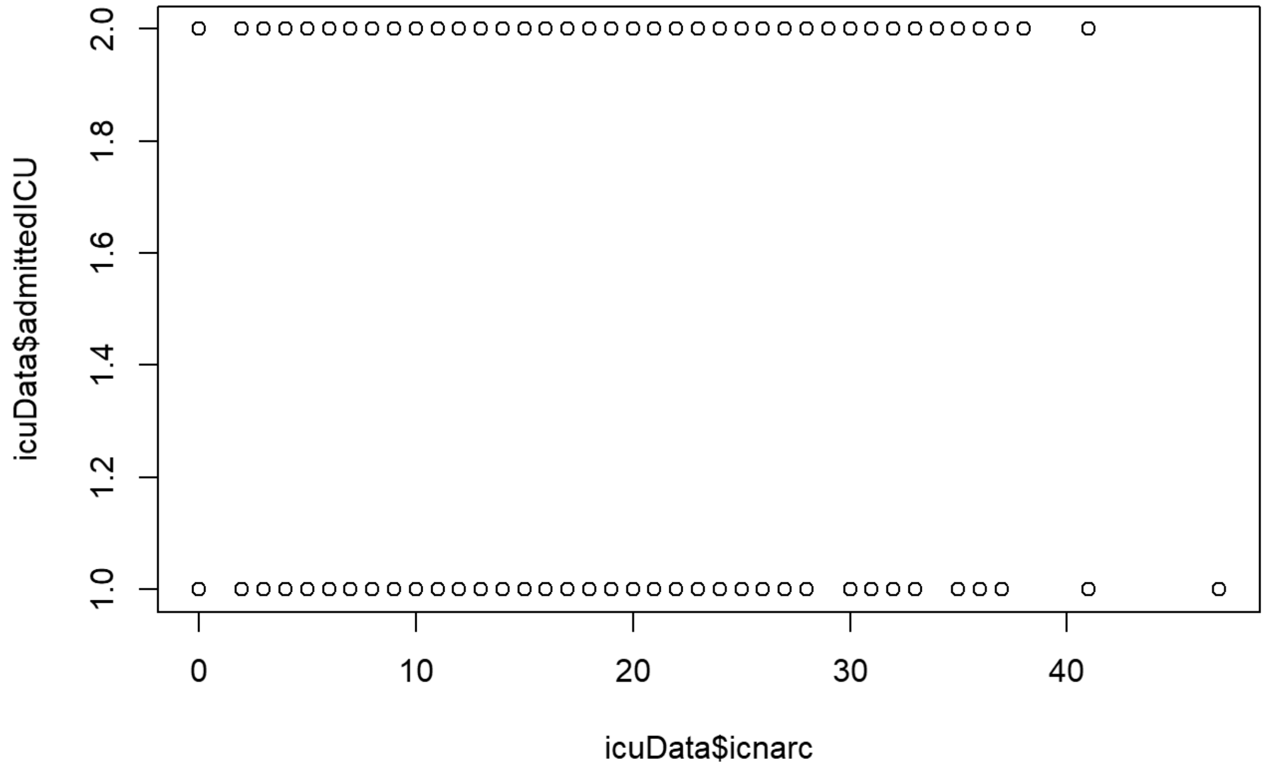
- (3pts) Does the above estimate suggest that someone admitted to the ICU with an ICNARC score of 10 is more likely to survive or not survive, according to this model? Be sure to show work for how you arrived at your answer.

[probability of death is = exp(-2.64345 + 0.44688* admittedICU1 + 0.07351* icnarc)/(1+ exp(-2.64345 + 0.44688* admittedICU1 + 0.07351* icnarc)) = exp(-1.46147)/(1+ exp(-1.46147)) = 0.1882426. since probability of death is 0.1882426, then it is more likely to survive because probability of survive would be higher (1-0.1882426 = 0.8117574)]

- (4pts) Is the above estimate based on an interpolation or extrapolation? Explain in one sentence, and provide any EDA you used to answer this question.

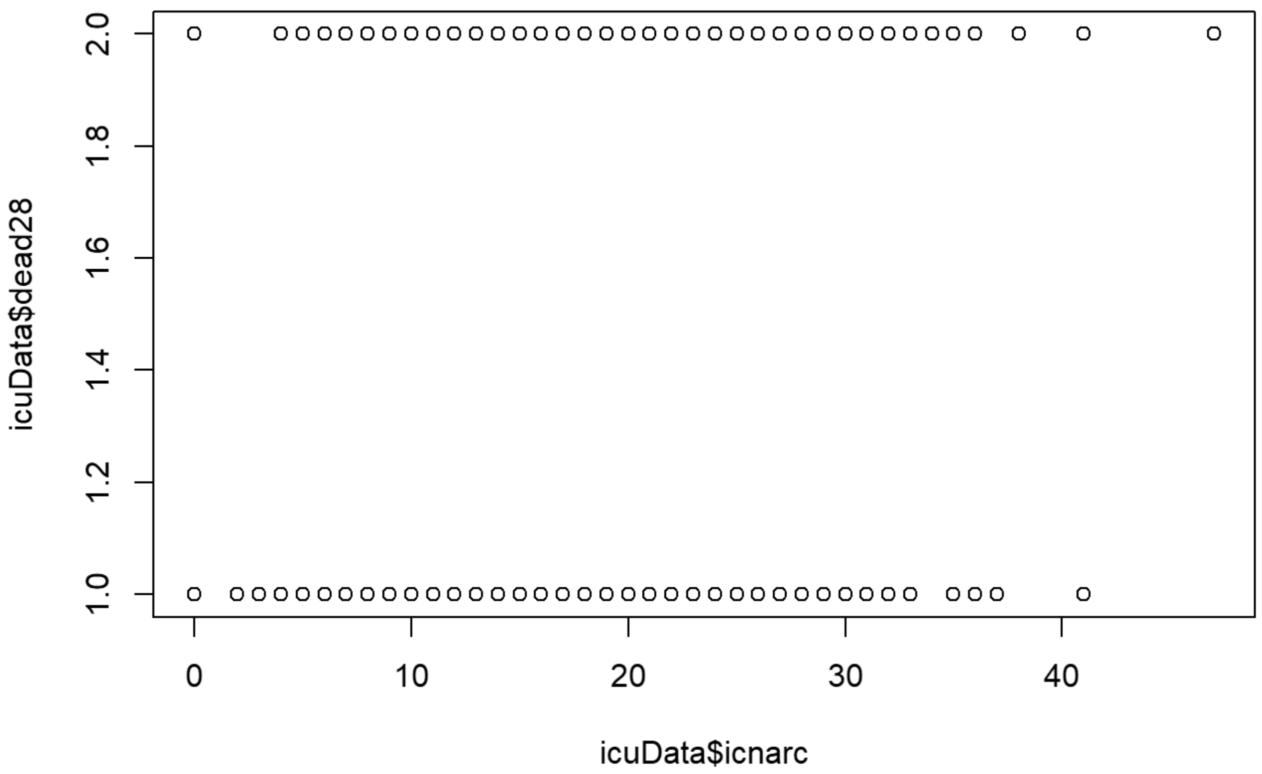
Hide

```
plot(icuData$icnarc, icuData$admittedICU)
```



Hide

```
plot(icuData$icnarc, icuData$dead28)
```



[it is interpolation because we already have data with icnarc=10 that are admitted to the ICU.]

Question 3: Let’s Move It Along, People (32 points)

Professor P. Bridgers (https://en.wikipedia.org/wiki/Phoebe_Bridgers) is studying whether physical therapy can improve muscular movement for stroke recovery patients. For her study, Professor Bridgers recruits patients who experienced either a “minor” stroke or “major” stroke in terms of severity, according to the patient’s doctor. Subjects are then randomized to attend physical therapy either 0, 1, 2, or 3 times a week. After a month, patients are asked to complete 15 simple physical tasks meant to measure their movement ability. The number of tasks they successfully completed is recorded. Here is the data from this study:

Hide

```
strokeData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/main/strokeData.csv")
#ensure categorical variables are factors
strokeData$stroke = factor(strokeData$stroke)
strokeData$treatment = factor(strokeData$treatment)
```

Here are the variables in this study:

- `treatment` : The treatment regime the subject was randomized to (either “no PT”, “PT1”, “PT2”, or “PT3”, denoting 0, 1, 2, or 3 days of physical therapy a week, respectively).

- `stroke` : Whether the subject experienced a “major” stroke or a “minor” stroke.
- `movement` : The number of movement tasks the subject successfully completed, out of 15 tasks total.

a. (9pts) For this part, answer the following two questions.

- (3pts) First, state the **variable type** for each variable in the dataset. As a reminder, we’ve discussed four variable types in class: Quantitative and continuous, quantitative and discrete, categorical and nominal, and categorical and ordinal. In your answer for each variable, briefly explain your reasoning.

[“PT3”: categorical and ordinal. Because 0,1,2,3 forms an order.Stroke: categorial and nominal because major and minor are two distinct categories. Movement: quantitative and discrete because we don’t have a task that is graded as half, like 10.5, or so.]

- (3pts) Professor Bridgers correctly states that this experiment followed a *between-subjects design*. That said, Professor Bridgers wonders if it would have been possible to conduct a within-subjects design for this type of data. Which explanatory variable in this dataset, if any, would be the natural choice for the within-subjects factor in a *new* experimental design? You have four choices: (1) neither `stroke` nor `treatment` would be natural choices; (2) only `stroke` would be a natural choice; (3) only `treatment` would be a natural choice; or (4) both `stroke` and `treatment` would be natural choices. Explain your answer in 1-2 sentences. (**Hint**: By “natural choice,” I mean relatively easy to implement in practice.)

[only treatment would be a natural choice because we can consider three categories for it for a given subject such as” Monday-Tuesday”, “Wednesday-Thursday”, “Friday-Sunday” can be the within-subject for the treatment. Sao one subject that has PT3, would have 1 denoted to all these categories. Or on subject that has no PT can have 0 denoted to all these categories.]

- (3pts) For the sake of this question, let’s say that we had a dataset where both `treatment` and `stroke` were within-subjects factors. For which of these variables, if any, could we have used *counter-balancing* in this hypothetical scenario? Again you have four choices: (1) we could not use counter-balancing for either; (2) we could use counter-balancing only for `stroke` ; (3) we could use counter-balancing only for `treatment` ; or (4) we could use counter-balancing for both. Explain in 1-2 sentences.

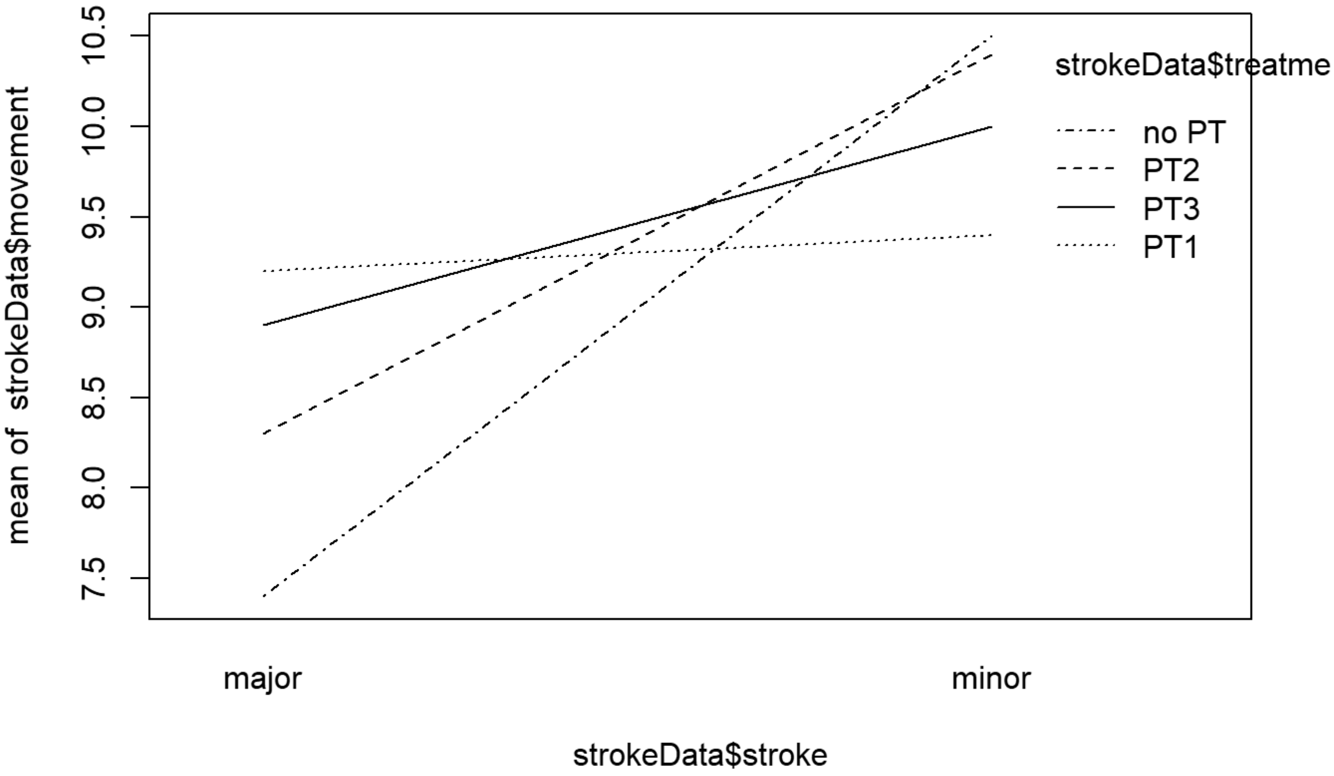
[we could use counter-balancing only for stroke; first of all only stroke had 2 categories. Secondly it is not rational but imaginarily they can first (like half of the particiapants) either grouped into “major” and then “minor” and other half vice versa.]

b. (15pts) Now we will consider several analyses for this dataset. For this part, answer the following four questions.

- (4pts) First, make an **interaction plot** that is appropriate for this dataset. Then, use your interaction plot to interpret the main effect of each explanatory variable, as well as the possible presence and nature of an interaction between the two explanatory variables.

Hide

```
interaction.plot(x.factor = strokeData$stroke, trace.factor = strokeData$treatment,
                 response = strokeData$movement)
```



[From this plot, we can see that the mean movement of minor groups might be higher than major group but without statistical analysis we cannot tell. Moreover, no PT, PT1 , PT2, and PT3 lines collapse. So it suggests (although not formally) that there might be an interaction present between these two factors.]

- (5pts) Now produce `summary()` output for three types of models:

1. A one-way ANOVA model using `stroke` as the explanatory variable.
2. The appropriate *additive* two-way ANOVA model.
3. The appropriate *interactive* two-way ANOVA model.

Then, using your `summary()` output, answer the following: Which model appears to have the *most* statistical power, and which model appears to have the *least* statistical power? Explain in 1-2 sentences.

Hide

```
#1
summary(aov(movement ~ stroke, data = strokeData))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## stroke       1  52.81   52.81   25.96 2.37e-06 ***
## Residuals   78 158.68    2.03                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
#2
summary(aov(movement ~ stroke+treatment, data = strokeData))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## stroke       1  52.81   52.81  25.417 3.11e-06 ***
## treatment    3   2.84    0.95   0.455    0.714    
## Residuals   75 155.84    2.08                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
#3
summary(aov(movement ~ stroke*treatment, data = strokeData))
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)    
## stroke            1  52.81   52.81  28.741 9.53e-07 ***
## treatment         3   2.84    0.95   0.515  0.67343    
## stroke:treatment  3  23.54    7.85   4.270  0.00784 ** 
## Residuals        72 132.30    1.84                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[The statistical power of the third model , the interactive model, is the most, because not only its shows that there is an interaction between stroke and treatment (P-value less than 0.05_ 0.00784), but also its P-value is less compared to the other two models. Also the second model , the additive model, has the least statistical power as it has the higher p-value (3.11e-06), and the p-value of treatment itself is above 0.05 (0.714).]

- (3pts) Regardless of your answer above, state your scientific conclusions according to the *interactive two-way ANOVA model*, within the context of this dataset. Be sure to explain how you arrived at your scientific conclusion.

[Given the P-value for the stroke that is less than 0.05 (9.53e-07), we reject the null that mean movement is same across different stroke groups. moreover, the p-value for the treatment is 0.67343 which is above 0.05 so we fail to reject the null that mean movement does not differ across different treatment groups. finally since the p-value of the interaction between stroke and treatment is less than 0.05 (0.0078), we reject the null that there is no interaction between these two variables.]

- (3pts) Looking at your interactive two-way ANOVA **and** interaction plot, Professor Bridgers concludes that the mean `movement` score is greater for those with a minor stroke than those with a major stroke. Based on your interactive two-way ANOVA **and** interaction plot, would you say that Professor Bridgers's conclusion is misleading? State whether it Is Misleading or Is Not Misleading, and explain in 1-3 sentences.

[No it's not misleading, because based on the P value of interactive two way ANOVA, the mean movement is different between "major" and "minor" groups (p-value is less than 0.05 - 9.53e-07). However just based on this p-value we can't tell which one has higher mean movement. The

interaction plot clearly shows that mean movement for minor groups is much higher compared to the major group.]

c. (8pts) Regardless of your answer in Part B, let’s say that Professor Bridgers wants to use the *additive* two-way ANOVA model for inference. For this part, answer the following.

- (4pts) Using the *additive* two-way ANOVA model, Professor Bridgers is interested in comparing *every pair* of treatments in terms of their mean `movement` (i.e., the “no PT” group versus “PT1” group, the “PT1” group versus the “PT2” group, and so on). Run the correction for unplanned comparisons that is *most* appropriate for this, using the additive two-way ANOVA model. Then, state your scientific conclusions based on your correction fur unplanned comparisons.

Hide

```
AdditiveTwo_way_ANOVA = aov(movement ~ stroke+treatment, data = strokeData)
TukeyHSD(AdditiveTwo_way_ANOVA)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = movement ~ stroke + treatment, data = strokeData)
##
## $stroke
##              diff      lwr      upr   p adj
## minor-major 1.625 0.982901 2.267099 3.1e-06
##
## $treatment
##           diff        lwr      upr     p adj
## PT1-no PT 0.35 -0.8477364 1.547736 0.8686785
## PT2-no PT 0.40 -0.7977364 1.597736 0.8164732
## PT3-no PT 0.50 -0.6977364 1.697736 0.6925204
## PT2-PT1   0.05 -1.1477364 1.247736 0.9995235
## PT3-PT1   0.15 -1.0477364 1.347736 0.9876057
## PT3-PT2   0.10 -1.0977364 1.297736 0.9962411
```

[Based on the results for Stroke, it is clear that with 95% confidence the mean minor is greater than mean major. The point estimate is 1.625. and the P-value is less than 0.05 (3.1e-06). In terms of treatment, none of the pairs have a significant difference. We cannot reject the null that their means are same for all pairs because all p-values are greater than 0.05.]

- (4pts) After looking at your results, Professor Bridgers says, “I wish I had just planned to make all of these paired comparisons ahead of time. If I had, I could have controlled the Type 1 error rate without having to use this correction for unplanned comparisons!” Do you Agree or Disagree with Professor Bridgers’ statement here? Explain your answer in 1-2 sentences. (**Hint**: For this question, you only need to consider the additive two-way ANOVA model, NOT the other ANOVA models.)

[I disagree because these corrections already control the type 1 error rate.]