

## MIDTERM INSTRUCTIONS: MUST READ

Academic Integrity Statement: MUST SIGN TO RECEIVE CREDIT FOR EXAM

Question 1: Multiple Interventions for Cognitive Decline (33 points)

Question 2: Monthly Check-ups (23 points)

Question 3: It's a Me, Logistic Regression (35 points)

Question 4: Did I Make an Error? (12pts)

# 36-309 / 36-749 Midterm 2

Code ▼

Due Friday, December 2, 8pm (via Gradescope)

Sanaz Saadatifar

## MIDTERM INSTRUCTIONS: MUST READ

- The format of the midterm is identical to that of the homework: You do everything in RStudio, and you submit a PDF on Gradescope.
- During the midterm (December 1-2), you are not allowed to talk with anyone 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 midterm.
- Although you can't talk with anyone else, the midterm is still "open everything." Remember that you can always refer back to previous homework/lab solutions, lectures, R demos, and the course textbook.
- **Throughout the midterm, 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. As a result, your midterm should include the code, analyses, and graphs you used to arrive at your answers. 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.
- **Although this exam is out of 100%, there are 103 points available on this exam. Thus, you could miss 3 points and still receive 100%.**

## 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 1 (8am) through December 2 (8pm)."

[Sanaz Saadatifar]

# Question 1: Multiple Interventions for Cognitive Decline (33 points)

Professor T. Reznor ([https://en.wikipedia.org/wiki/Trent\\_Reznor](https://en.wikipedia.org/wiki/Trent_Reznor)) studies Alzheimer's disease, which adversely affects memory over time. Specifically, Professor Reznor conducted an experiment to assess if certain interventions can lessen this cognitive decline. He recruited subjects who all had a similar severity of Alzheimer's disease, and randomized them to one of four interventions:

- “exercise”: subjects perform weekly physical exercises
- “reading”: subjects perform weekly reading exercises
- “discussion”: subjects perform weekly discussion-based exercises
- “control”: no intervention; subjects go about their usual routine

After three months, subjects took a cognitive test to assess their memory. This test was scored from 0 to 50, where a higher number is better.

Here is the dataset produced by this experiment:

Hide

```
memoryData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/main/memoryData.csv")
#ensure that categorical variables are treated as factors
memoryData$intervention = factor(memoryData$intervention)
```

Here are the variables in the dataset:

- `intervention` : The intervention each subject was assigned to (“exercise”, “reading”, “discussion”, or “control”).
- `memory` : The score on the cognitive test meant to assess memory (a higher number is better).

As part of the study's analysis plan, Professor Reznor wants to at least answer the following questions:

0. Is the average `memory` score the same across all four interventions?
1. How does the “control” group compare to the average of the three interventions (“exercise”, “reading”, and “discussion”), in terms of the average `memory` score?
2. How does the “reading” intervention compare to the “discussion” intervention, in terms of the average `memory` score?
3. How does the average of the “reading” and “discussion” interventions compare to the “exercise” intervention, in terms of the average `memory` score?

To address #0, one of Professor Reznor's graduate students ran the following analysis:

Hide

```
onewayModel = aov(memory ~ intervention, data = memoryData)
summary(onewayModel)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## intervention  3   203.8    67.94     5.907 0.00134 **
## Residuals    60   690.1    11.50
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The student correctly rejected the null hypothesis that the average `memory` is the same across all four interventions. Unfortunately, this student dropped out of 36-309/36-749 after the first few weeks, and so they're not sure how to answer Professor Reznor's remaining questions. They're hoping that you'll be able to help.

The above three planned comparisons indeed meet the four conditions for planned comparisons. For the sake of this exam, you **do not** need to check that these conditions hold.

a. (10pts) For this part, answer the following two questions.

- (2pts): First, use the `levels()` function to determine the order of the four `intervention` levels according to R. For this part, you just have to write one line of code that produces output that allows you to determine the order of the `intervention` levels (i.e., which one is listed first, second, third, and fourth).

Hide

```
levels(memoryData$intervention)
```

```
## [1] "control" "discussion" "exercise" "reading"
```

- (8pts) For each of Professor Reznor's planned comparisons (#1, #2, and #3), write out the appropriate **contrast null hypothesis**. In your hypotheses, please use the following notation:  $\mu_E$ ,  $\mu_R$ ,  $\mu_D$ ,  $\mu_C$ , which denote the mean `memory` for the four interventions (exercise, reading, discussion, control). Furthermore, **write your contrast null hypothesis such that the order of the  $\mu$ s is the same as the order of the levels from the previous bullet point**. In your answer, be sure to specify which hypothesis corresponds to which planned comparisons (#1, #2, or #3).

[planned comparison number 1's null hypothesis:  $(1)\mu_C - (1/3)\mu_D - (1/3)\mu_E - (1/3)\mu_R = 0$ , planned comparison number 2's null hypothesis:  $(0)\mu_C + (1)\mu_D + (0)\mu_E - (1)\mu_R = 0$ , planned comparison number 3's null hypothesis:  $(0)\mu_C + (1/2)\mu_D - (1)\mu_E + (1/2)\mu_R = 0$ ]

- b. (9pts) For this part, first write code that correctly tests the three hypotheses you wrote in Part A. After including your code, state your scientific conclusion for each hypothesis, such that you answer Professor Reznor's three planned questions. (**Hint:** For this part you need to appropriately use the `glht()` function, which requires the `multcomp` library. To help you, the `multcomp` library is already loaded below.)

Hide

```
library(multcomp)

glht.fit1 = glht(model = onewayModel,
                  linfct = mcp(intervention = c(1, -1/3, -1/3, -1/3) ))
summary(glht.fit1)
```

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = memory ~ intervention, data = memoryData)
##
## Linear Hypotheses:
##      Estimate Std. Error t value Pr(>|t|)
## 1 == 0    -1.042      0.979  -1.064    0.292
## (Adjusted p values reported -- single-step method)
```

Hide

```
glht.fit2 = glht(model = onewayModel,
                 linfct = mcp(intervention = c(0, 1, 0, -1) ))
summary(glht.fit2)
```

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = memory ~ intervention, data = memoryData)
##
## Linear Hypotheses:
##      Estimate Std. Error t value Pr(>|t|)
## 1 == 0    -4.812      1.199  -4.014 0.000169 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Hide

```
glht.fit3 = glht(model = onewayModel,
                 linfct = mcp(intervention = c(0, 1/2, -1, 1/2) ))
summary(glht.fit3)
```

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = memory ~ intervention, data = memoryData)
##
## Linear Hypotheses:
##           Estimate Std. Error t value Pr(>|t|)
## 1 == 0    0.7187      1.0384   0.692   0.492
## (Adjusted p values reported -- single-step method)
```

[For the planned comparison number 1, we fail to reject the null hypothesis because the p-value is greater than 0.05 (0.292), for the planned comparison number 2, we can reject the null hypothesis because the p-value is less than 0.05 (0.000169), and for the planned comparison number 3, we fail to reject the null hypothesis as the p-value is above 0.05 (0.492).]

c. (14pts) For this part, answer the following three questions.

- (6pts) The purpose of Professor Reznor's first planned comparison was to compare "control" to the other three interventions. Now, Professor Reznor wants to compare "control" to *each* of the three interventions (i.e., "control" vs "reading", control vs "discussion", and "control" vs "exercise"). Given this, the most appropriate procedure to use is Dunnett's procedure. For this part, provide code that appropriately implements Dunnett's procedure. Then, state your scientific conclusions based on the output from that procedure.

Hide

```
glht.dunnett = glht(model = onewayModel, linfct = mcp(intervention = "Dunnett"))
summary(glht.dunnett)
```

```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = memory ~ intervention, data = memoryData)
##
## Linear Hypotheses:
##           Estimate Std. Error t value Pr(>|t|)
## discussion - control == 0  -1.1250      1.1991  -0.938  0.66996
## exercise - control == 0    0.5625      1.1991   0.469  0.93642
## reading - control == 0     3.6875      1.1991   3.075  0.00868 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

[Three null hypotheses are tested here. One: there is no difference between average memory scores of control vs discussion. We fail to reject this one because p-value is above 0.05 (0.67). two: there is no difference between average memory scores of control vs exercise. We fail to reject this one because p-value is above 0.05 (0.93644). Three: there is no difference between average memory scores of control vs reading. We can reject this one because p-value is less than 0.05 (0.00866).]

- (4pts) Professor Reznor knows that another “corrections for unplanned comparisons” procedure is the Tukey procedure. Given this, he decides to run the following code:

Hide

```
TukeyHSD(onewayModel)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = memory ~ intervention, data = memoryData)
##
## $intervention
##          diff          lwr          upr          p adj
## discussion-control -1.1250 -4.29355851  2.043559  0.7844476
## exercise-control    0.5625 -2.60605851  3.731059  0.9655419
## reading-control     3.6875  0.51894149  6.856059  0.0162719
## exercise-discussion  1.6875 -1.48105851  4.856059  0.4998008
## reading-discussion  4.8125  1.64394149  7.981059  0.0009471
## reading-exercise    3.1250 -0.04355851  6.293559  0.0545852
```

Looking at this output, Professor Reznor says: “It looks like the first three point estimates (in the `diff` column) are exactly the same as your point estimates from Dunnett’s procedure, but the p-values (in the `p adj` column) are different than your p-values from Dunnett’s procedure. Why do these things happen?” In 1-3 sentences, explain to Professor Reznor why **both** of these things happen (that the point estimates are the same, but the p-values are different). (**Hint:** If you aren’t getting the same point estimates, you either made a mistake earlier or you’re misinterpreting output.)

[Point estimate are the same because the confidence interval of the same datasets would be similar in all possible comparison methods. (Eventually the datasets are same). However, since Dunnett is specifically correcting errors or pairwise comparisons for specific control group vs other treatment groups, it can better control the errors and P-value is smaller so power is higher here. However, Tukey method is more conservative, and it results in higher p-values. ]

- (4pts) Here we’ve considered three follow-up comparisons (comparing “control” vs each of the other three interventions). We considered two procedures: Dunnett’s and Tukey’s. For these three follow-up comparisons, which of these two procedures has **more statistical power**? Explain your answer in 1-2 sentences.

[As menthoned previously, since Dunnett is specifically correcting errors or pairwise comparisons for specific control group vs other treatment groups, it can better control the errors and P-value is smaller so it has more statistical power]

## Question 2: Monthly Check-ups (23 points)

In this question we will continue to work with Professor Reznor's study from Question 1. In particular, Professor Reznor says: "In case it's helpful, we actually measured each subject's memory each month of the study, such that we also have measurements after one month and after two months." In other words, the outcome we used in Question 1 was the third-month measurement, but Professor Reznor also has several repeated measurements.

Given this information, Professor Reznor would like you to conduct another analysis for his study. To aid you in this task, he provides you with the following dataset:

Hide

```
tallMemoryData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/main/tallMemoryData.csv")
#ensure that categorical variables are treated as factors
tallMemoryData$intervention = factor(tallMemoryData$intervention)
tallMemoryData$subjectID = factor(tallMemoryData$subjectID)
tallMemoryData$month = factor(tallMemoryData$month)
```

This dataset contains four variables: `intervention`, `memory`, `subjectID`, and `month`. The first two variables (`intervention` and `memory`) are the same as in Question 1. Meanwhile, `subjectID` is a number ranging from 1 to 64, denoting each subject in the study, and `month` denotes the month that the `memory` measurement was taken (either "month1", "month2", or "month3"). All of the "month3" measurements are the same `memory` used in Question 1.

You can complete this question without completing Question 1; however, just be sure to read the beginning of Question 1 before starting this question.

- a. (12pts) Professor Reznor would still like to understand how `intervention` affects `memory`, but he also wants to understand how `month` is related to `memory`, because he suspects that subjects' memory may change over time. For this part, answer the following three questions.
  - (4pts) Hopefully it is clear that `intervention` is a *between-subjects* factor and `month` is a *within-subjects* factor. Thus, the most appropriate analysis is a mixed between- and within-subjects two-way ANOVA analysis. For this part, write code that runs the appropriate mixed between- and within-subjects two-way ANOVA analysis. You only have to write code for this part, but your code should produce output that can be used to make scientific conclusions about the effect of `intervention` and `month` on `memory`. (**Hint:** For this part you need to appropriately use the `anova_test()` function, which requires the `rstatix` library. To help you, the `rstatix` library is already loaded below.)

Hide

```
library(rstatix)
anova_test(dv = memory, wid = subjectID, within = month, between = intervention, data = tallMemoryData)
```

```
## ANOVA Table (type II tests)
##
## $ANOVA
##           Effect DFn DFd      F      p p<.05 ges
## 1      intervention    3   60 4.661 0.005000    * 0.092
## 2           month    2  120 9.506 0.000147    * 0.082
## 3 intervention:month    6  120 4.947 0.000148    * 0.122
##
## $`Mauchly's Test for Sphericity`
##           Effect      W      p p<.05
## 1           month 0.686 1.5e-05    *
## 2 intervention:month 0.686 1.5e-05    *
##
## $`Sphericity Corrections`
##           Effect   GGe      DF[GG]   p[GG] p[GG]<.05   HFe      DF[HF]
## 1           month 0.761 1.52, 91.34 0.000626    * 0.777 1.55, 93.23
## 2 intervention:month 0.761 4.57, 91.34 0.000696    * 0.777 4.66, 93.23
##           p[HF] p[HF]<.05
## 1 0.000569    *
## 2 0.000628    *
```

- (4pts) Professor Reznor says: “I would expect that memory measurements in adjacent months are more correlated with each other than the memory measurements made in non-adjacent months. In other words, I hypothesize that the covariance between the month1 and month2 measurements or the covariance between the month2 and month3 measurements is different from the covariance between the month1 and month3 measurements.” Does your output above Support or Not Support Professor Reznor’s hypothesis here? State Support or Not Support, and explain in 1-3 sentences.

**[No this test does not analyze the covariance. To do so we needed to conduct an ANCOVA test. Here the only results we can get is to analyze whether memory measurements is different across different months. ]**

- (4pts) Using your above output, state your scientific conclusions for your analysis, within the context of this dataset. In your answer, please specify the p-values you used to arrive at your conclusions, and explain why you used those p-values specifically.

**[First, I note that I should use the “Sphericity Corrections” output, because the p-value for Mauchly’s Test for Sphericity is 1.5e-05, and thus I reject the null hypothesis that sphericity holds. I decide to use the Huynh-Feldt (HF) p-value. First looking at the p-value for the interaction (the intervention:month row), I see that the p-value is 0.000628, which is less than 0.05. Thus, I reject the null hypothesis that there is no interaction between the month and intervention factors. Within the “Sphericity Corrections” table (and again looking at the HF p-value), I see that the p-value for month is quite small and certainly less than 0.05 (0.000569). Thus, I reject the null hypothesis that month1 = month2 = month3, thereby concluding that the mean memory differs across levels of months. Meanwhile, to understand the main effect of intervention, I can look at the “ANOVA” output; as discussed above, intervention is only displayed in the “ANOVA” output because it is a between-subjects factor. I see that the p-value for intervention, I see that the p-value is 0.005, which is less than 0.05. Thus, I reject the null hypothesis that the mean memory differs across levels of intervention ]**



- b. (11pts) Out of curiosity, Professor Reznor wants to compare the `memory` scores near the beginning of the study (after one month) to the `memory` scores at the end of the study (after three months). Thus, he takes the “month1” and “month3” `memory` measurements from the previous dataset, and puts them into a new, *wide-format* dataset, given here:

Hide

```
wideMemoryData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall12022/main/wideMemoryData.csv")
```

This dataset contains two variables, `memory.month1` and `memory.month3`, which denote each subject's `memory` scores for “month1” and “month3”, respectively. Thus, for this dataset, each subject has a *pair* of outcome measurements. For this part, answer the following questions.

- (7pts) *Using the `wideMemoryData` dataset*, perform the appropriate *paired t-test*. After running your paired t-test, state your scientific conclusion based on the paired t-test, within the context of this dataset. In your answer, be sure to mention what null hypothesis is being tested by your paired t-test, as well as a point estimate. (**Note:** For the purposes of this question, please ignore any conclusions you made about this data in Part A. Just focus on making conclusions about this `wideMemoryData` dataset specifically.)

Hide

```
t.test(x = wideMemoryData$memory.month1, y = wideMemoryData$memory.month3, paired = TRUE)
```

```
##
## Paired t-test
##
## data: wideMemoryData$memory.month1 and wideMemoryData$memory.month3
## t = -3.1235, df = 63, p-value = 0.0027
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -2.9464721 -0.6472779
## sample estimates:
## mean of the differences
## -1.796875
```

**[The null hypothesis is that average memory of month 1 is the same as average memory of month 2. Since the p-value is less than 0.05(0.0027), we reject the null. Point estimate is -1.796875 which is the mean month1 minus mean month 3.]**

- (4pts) To wrap up this study, Professor Reznor says: “In this study we have repeated measurements of `memory` over several months. I’ve heard that for studies with repeated measurements, it’s preferable to use counter-balancing.” Would it have been possible to use counter-balancing in this study? State Yes or No, and explain your answer in 1-2 sentences. Specifically: If you state Yes, explain how you would implement counter-balancing in this study. And if you state No, explain why counter-balancing cannot be implemented for this study.

**[No, because this is a longitudinal study and you cannot change the timing of month1 with month 3. No one can start from its third month. Every one should start from the first month. ]**

# Question 3: It's a Me, Logistic Regression (35 points)

In this question, we'll work with a real education dataset:

Hide

```
dropoutData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/main/italyDataDropout.csv")
```

The dataset consists of 15984 first-year college students at the University of Pisa or the University of Florence (both in Italy). When students enroll in college, their `income` in euros is measured; then, after their first year, it's determined whether or not the student dropped out of college. Here are the variables in this dataset:

- `income` : A measure of the student's annual family income (in euros).
- `university` : Whether the student attended the University of Florence or the University of Pisa.
- `dropout` : Whether the student dropped out of university (1 if yes, 0 if no).

The goal of this study is to assess how `income` and `university` are related to the chance that a student drops out of college.

(**Note:** Here the outcome is `dropout`, which is binary—so, we'll eventually use logistic regression. I used this dataset to construct a similar dataset in Homework4, where `dropout` was not binary. Now you're ready to analyze the real data! The real data has many variables, but for this exam, I've given you only three variables. You **do not** need to look back at Homework4, but I wanted to note this connection.)

a. (12pts) First we will explore the dropout rates in the two universities in this dataset. For this part, answer the following two questions.

- (7pts) Professor A. Bocelli ([https://en.wikipedia.org/wiki/Andrea\\_Bocelli](https://en.wikipedia.org/wiki/Andrea_Bocelli)) is interested in comparing dropout rates at the University of Florence with dropout rates at the University of Pisa. For this part, as EDA, make a table that allows you to see the dropout **rate** for the University of Florence and the dropout **rate** for the University of Pisa. You should construct your table such that (1) it lists four proportions, and (2) you're able to see what proportion of University of Florence students dropped out **and** what proportion of University of Pisa students dropped out. After making your table, state what these two proportions are (for Florence and Pisa), and state which university seems to have a higher dropout rate.

Hide

```
prop.table(table(dropoutData$university, dropoutData$dropout), margin = 1)
```

```
##
##           0           1
##  florence 0.6921583 0.3078417
##    pisa    0.5699214 0.4300786
```

[Dropout proportion for Florence is 30.78%. Dropout proportion for Pisa is 43% . it seems that Pisa university has higher dropout rate. ]

- (5pts) Now, using only the variables `university` and `dropout`, perform the appropriate chi-squared test. After running your chi-squared test, state your scientific conclusion based on the test's output. In your answer, be sure to mention what null hypothesis you are testing here.

Hide

```
chisq.test(table(dropoutData$university, dropoutData$dropout))
```

```
##  
##  Pearson's Chi-squared test with Yates' continuity correction  
##  
## data:  table(dropoutData$university, dropoutData$dropout)  
## X-squared = 246.52, df = 1, p-value < 2.2e-16
```

**[Null hypothesis is that dropout rate is independent from the university. Since the p-value is less than 0.05 (2.2e-16) we reject the null.]**

- b. (14pts) Now we will use logistic regression to analyze this dataset. For this part, answer the following questions.
- (4pts) First, fit an *additive* logistic regression model using `dropout` as the outcome and `university` and `income` as explanatory variables. **Do not** include any interactions in the model. For this part, you just need to include `summary()` output from your logistic regression model.

Hide

```
summary(glm(dropout ~ income + university, data = dropoutData, family = "binomial"))
```

```
##
## Call:
## glm(formula = dropout ~ income + university, family = "binomial",
##      data = dropoutData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.177  -1.018  -0.841   1.303   1.632
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.365e-01  4.544e-02 -11.808  < 2e-16 ***
## income       -2.446e-05  3.329e-06  -7.348  2.02e-13 ***
## universitypisa 5.351e-01  3.384e-02  15.813  < 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: 21211  on 15983  degrees of freedom
## Residual deviance: 20907  on 15981  degrees of freedom
## AIC: 20913
##
## Number of Fisher Scoring iterations: 4
```

- (5pts) If you ran your logistic regression model correctly, you should see coefficients for the `university` variable and the `income` variable. Write your interpretation for each of these coefficient *estimates* (which are real numbers, provided by your `summary()` output) on the *odds scale* within the context of this dataset.

[First, we see that the estimate for the income coefficient is  $-2.446e-05$ . The interpretation of this coefficient is, For every one-unit increase in income, the odds of dropouts are multiplied by  $\exp(-2.446e-05) = 0.9999755$  for any fixed university. Thus, the odds of dropouts decreases slightly for every one-unit increase in income. Meanwhile, we see that the estimate for the university of Pisa coefficient is  $5.351e-01$ . Thus, for any fixed number of income (i.e., holding income fixed), the odds of dropout for university of Pisa students is  $\exp(5.351e-01) = 1.707619$  times the odds of dropout for university of Florence. Thus, the odds of dropout are estimated to be higher when the student is studying in university of Pisa.]

- (5pts) Looking at your logistic regression model, Professor Bocelli asks: “Based on your model, what is the estimated probability that a student drops out if their `income` is 10,000 euros and they attend the University of Florence? What about if their income is 10,000 euros and they attend the University of Pisa?” For this part, show any work you used to arrive at your answer; specifically, please write code that shows how you compute the estimated probability for these two cases.

$$P(\text{dropout} = 1) = \frac{\exp(\beta_0 + \beta_1 * \text{income} + \beta_p * \text{UPisa})}{1 + (\exp(\beta_0 + \beta_1 * \text{income} + \beta_p * \text{UPisa}))} = \frac{\exp(-5.365e-01 - 2.446e-05 * \text{income} + 5.351e-01 * \text{UPisa})}{1 + (\exp(-5.365e-01 - 2.446e-05 * \text{income} + 5.351e-01 * \text{UPisa}))}$$

[If `income = 10000` and student studies in university of Florence, the  $\beta_p = 0$ . This results in the  $P(\text{dropout} = 1) = 0.3140829 = 31.4\%$ ] [If `income = 10000` and student studies in university of Pisa, the  $\beta_p = 1$ . This results in the  $P(\text{dropout} = 1) = 0.4388083 = 43.8\%$ ]

- c. (9pts) To wrap up this question, let's compare the inference you conducted for the chi-squared test in Part A with the inference you conducted for the logistic regression in Part B. For this part, answer the following two questions.
- (5pts) Your logistic regression output in Part B should include results for a `universitypisa` row. In particular, there should be a p-value in this row. What null hypothesis is this p-value testing? Please write your answer in words, rather than mathematical notation. Furthermore, based on your output, do you Reject or Fail to Reject this null hypothesis? Explain in one sentence.

**[ It tests whether coefficient for university (for example university of Pisa) is equal to zero. In other words it tests whether university relates to the dropout. The null hypothesis is that this coefficient is equal to zero. However, since  $p\text{-value} < 2e-16$  (less than 0.05), we reject the null. ]**

- (4pts) In Part A, your chi-squared test should have also provided a p-value, and you should have used this p-value to reject or fail to reject a null hypothesis. Consider the scientific conclusion you made with the chi-squared test in Part A, as well as the scientific conclusion you just made in the previous bullet point (about the `universitypisa` row in the logistic regression). Do these two scientific conclusions Agree, Disagree, or Neither agree or disagree with each other? State Agree, Disagree, or Neither, and then explain in 1-3 sentences. Specifically: If you state Agree or Disagree, explain why the conclusions agree or disagree with each other. Meanwhile, if you state Neither, explain why the conclusions aren't related to each other.

**[They both agree with each other. Because in Chi-squared test, we test whether university and dropout are independent of each other, which means if they relate to each other which also means If there is a coefficient between these two which is the same as the null hypothesis for the logistic regression. Both p-values are less than 0.05 and they both reject these null hypotheses. ]**

## Question 4: Did I Make an Error? (12pts)

Professor L. Jordan ([https://en.wikipedia.org/wiki/Snail\\_Mail\\_\(musician\)](https://en.wikipedia.org/wiki/Snail_Mail_(musician))) is performing 100 independent studies. For each study, Professor Jordan compares a treatment group to a control group with a t-test.

**Professor Jordan does not know this**, but for 50 of these studies, the null hypothesis is true; for 20 of these studies, the null hypothesis is false and the power is 40%; and for 30 of these studies, the null hypothesis is false and the power is 80%. For each hypothesis, Professor Jordan uses  $\alpha = 0.05$ .

For this part, answer the following three questions. **Throughout, be sure to show your work for how you arrived at your answers.**

- (4pts) Out of these 100 studies, how many is Professor Jordan expected to get a p-value less than 0.05 (i.e., reject the null hypothesis)? And how many is she expected to get a p-value greater than 0.05 (i.e., fail to reject the null hypothesis)?

**[reject the null hypothesis for 50 studies with true  $H_0$  is  $50 \cdot 0.05 = 2.5$ , for 20 studies with false  $H_0$  is  $20 \cdot 0.4 = 8$ , for 30 studies with False  $H_0$  is  $30 \cdot 0.8 = 24$ .  $\ggggg$   $2.5 + 8 + 24 = 34.5$ . fail to reject the null on the other hand is  $\ggggg$   $(50 - 2.5) + (20 - 8) + (30 - 24) = 47.5 + 12 + 6 = 65.5$ ]**

- (4pts) Out of these 100 studies, how many Type 1 errors is Professor Jordan expected to make? How many Type 2 errors is she expected to make?

**[considering the calculation of last question: Type 1 error is 2.5 because it is when we incorrectly reject the null so it should be when we reject the null when it is true (2.5 out of 50). On the other hand, the type 2 error is  $12 + 6 = 18$  because It is when we incorrectly fail to reject the null, so we should select the times when null is false but we fail to reject it. ]**

- (4pts) Out of these 100 studies, what *percent* of her “reject the null hypothesis” conclusions are expected to be correct? What *percent* of her “fail to reject the null hypothesis” conclusions are expected to be correct? Please write your answers in terms of percents and round to the nearest decimal point (e.g., write 51.4%, not 0.514 or 51.473%)

**[Considering the calculations done for the first question of this part: percent of correctly “reject the null hypothesis” =  $(8+24)/(2.5+8+24) = 0.9275362 = 92.7\%$ , percent of correctly “fail to reject the null hypothesis” =  $47.5/(47.5+12+6) = 0.7251908 = 72.5\%$ . ]**