Homework 4

Your Name

Due: November 26, 2019 at 11:59pm

Homework Policies:

You are encouraged to discuss problem sets with your fellow students (and with the Course Instructor of course), but you must write your own final answers, in your own words. Solutions prepared "in committee" or by copying someone else's paper are not acceptable. This violates the Brown standards of plagiarism, and you will not have the benefit of having thought about and worked the problem when you take the examinations.

All answers must be in complete sentences and all graphs must be properly labeled.

Recent homework and exams have been filled with a lot of extra material that is not tables or plots. Please only display the plots and tables and not the code. You may have to look up how to hide the code. In addition, if you display multiple plots, make sure they are all labeled properly and that they are arranged in a grid to keep things more orderly. Points will be deducted for not following these instructions.

In this homework you will be required to use .Rmd to do it., you can then knit to a word document of PDF to turn it in.

For the PDF Version of this assignment: PDF

For the R Markdown Version of this assignment: RMarkdown

Turning the Homework in:

Please turn the homework in through canvas. You may use a pdf, html or word doc file to turn the assignment in.

PHP 1510 Assignment Link

PHP 2510 Assignment Link

Diabetes Data

We have some clinical data from a diabetes study. This is a bit larger than some datasets you have been using in the course so far but more realistic of what you might need. This data comes from UCI Machine Learning Repository. I removed some of the columns to make it more manageable. For some uestions below, you will be responsible for filtering out some of the data.

Make the following into a code chunk so you can read the data in.

```
load("diabetes.rda")
```

1. (2 points) How many observations are in this data?

```
dim(diabetes)[1]
```

```
## [1] 101766
```

2. (2 points) How many patients are in this data?

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
diabetes %>%
  select(patient_nbr) %>%
  unique() %>%
  tally()
```

```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.

## # A tibble: 1 x 1
##       n
##   <int>
## 1 71518
```

3. (4 points) What is the highest number of visits for a patient? (Use `dplyr` tools and only print this one patient line.)

```r
diabetes %>%
  group_by(patient_nbr) %>%
  tally(sort=T) %>%
  top_n(1)
```

```
## Selecting by n

## # A tibble: 1 x 2
##   patient_nbr     n
##   <fct>       <int>
## 1 88785891       40
```

When a patient is admitted to a hospital, their admission is generally categorized according to the severity or urgency of their symptoms (e.g., “Elective”, “Emergency”, etc.). We are interested in the different types of hospital admissions for diabetic patients. We will use the data in the file diabetes data clean.csv to answer several questions related to hospital admissions. Use the appropriate statistical tests, plots, or tables to address the questions below.

4. (10 points) Does it appear that admissions type differs by gender?
   - Display appropriate tables or plots
   - Perform a hypothesis test.
   - Interpret the results.

```r
table(diabetes$admission_type, diabetes$gender) %>% prop.table() %>% round(3)
```

```
##
##                   Female  Male Unknown/Invalid
##   Elective         0.102 0.094           0.000
##   Emergency        0.305 0.254           0.000
##   Newborn          0.000 0.000           0.000
##   Not Available    0.027 0.023           0.000
##   Not Mapped       0.002 0.001           0.000
##   Trauma Center    0.000 0.000           0.000
##   Urgent           0.103 0.089           0.000
```

We can see that we have a column with invalid, we must get rid of this as there is no-one in it. We can also see that there are not enough newborn visits or Trauma center to appear as well as Not Available and Not Mapped which do not have the same meanings as other values, so we can drop these.

```
diabetes <- diabetes %>%
  filter(gender != "Unknown/Invalid") %>%
  filter(admission_type != "Newborn" & admission_type != "Not Available" & admission_type != "Not Mappe
  droplevels()
table(diabetes$admission_type, diabetes$gender) %>% prop.table() %>% round(3)
```

```
##
##              Female  Male
##   Elective    0.108 0.099
##   Emergency   0.322 0.269
##   Urgent      0.108 0.094
```

There does not appear to be much of a difference based on percentages. We can perform a Chi-Squared test to be sure.

```
table(diabetes$admission_type, diabetes$gender) %>% chisq.test(correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 33.121, df = 2, p-value = 6.425e-08
```

$$H_0\text{: Rows and Columns are independent vs } H_1\text{: Rows and Columns are dependent}$$

We can see that according to the chi-square test, there is a relationship between gender and admission type. This can be seen with a p-value of 1.4×10^{-8}***. This is most likely due to a large sample size as the percentages do not appear to be all that different.***

5. (10 points) Consider the hospital admissions of `Elective` and `Emergency` categories. Do males and females differ with respect to these?
   - Display appropriate tables or plots
   - Perform a hypothesis test.
   - Interpret the results.

```
diabetes <- diabetes %>%
  filter(admission_type != "Urgent") %>%
  droplevels()
table(diabetes$admission_type, diabetes$gender) %>% prop.table() %>% round(3)
```

```
##
##              Female  Male
##   Elective    0.135 0.124
##   Emergency   0.404 0.337
```

From the above table we can see that there does not appear to be a strong difference between males and females in terms of admission by elective or emergency. If there was a difference it may be that there are $\approx$ ***7% more females admitted by emergency than males.***

```
table(diabetes$admission_type, diabetes$gender) %>% chisq.test(correct=F)
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  .
## X-squared = 32.243, df = 1, p-value = 1.361e-08
```

We can see that according to the chi-square test, there is a relationship between gender and admission type. This can be seen with a p-value of 1.4×10^{-8}***. This is most likely driven by the differences in Emergency admissions.***

6. (10 points) We are interested whether patients differ in the number of medications based on whether their admission type was `Elective` or `Emergency`. Using the appropriate statistical test, determine whether or not patients coming in electively or in the case of an emergency tend to receive the same number of medications.

```r
t.test(num_medications~admission_type, diabetes)
```

```
##
##  Welch Two Sample t-test
##
## data:  num_medications by admission_type
## t = 41.769, df = 27026, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.096598 3.401532
## sample estimates:
##  mean in group Elective mean in group Emergency
##                18.63435                15.38529
```

We can see that the t statistic is 41.769 which means it is 41.769 standard errors away from the mean. With a p-value of 2.2×10^{-16} ***and a confidence interval on the differences between means of*** $(3.10, 3.40)$ ***that there is a difference between the mean number of medications by admissions type. It appears that those who are in the elective group have more medications than those in the emergency.***

7. (8 points) What are the assumptions of the statistical test we use in 6? Are we confident that these assumptions are satisfied?
   - Use bootstrap confidence intervals and compare them to the hypothesis test and t-distribution confidence intervals in 6.

```r
library(rsample)
```

```
## Warning: package 'rsample' was built under R version 3.5.3
```

```
## Loading required package: tidyr
```

```r
library(purrr)
```

```
## Warning: package 'purrr' was built under R version 3.5.3
```

```r
library(ggplot2)
library(dplyr)
diabetes2 <- diabetes %>%
  select(num_medications, admission_type)

bt_data <- bootstraps(diabetes2, times = 1000)
get_diff <- function(splits) {
  d <- analysis(splits)
  mean_ele <- mean(d$num_medications[d$admission_type=="Elective"], na.rm=T)
  mean_eme <- mean(d$num_medications[d$admission_type=="Emergency"], na.rm=T)
  mean_ele - mean_eme
}
bt_data$bt_diff <- map_dbl(bt_data$splits, get_diff)
bt_ci <- round(quantile(bt_data$bt_diff, c(0.025, 0.975)), 3)
bt_ci
```

```
##  2.5% 97.5%
## 3.102 3.398
```

The t-distribution confidence interval was $(3.10, 3.40)$ ***and the bootstrap confidence interval is*** $(3.09, 3.40)$***, these are almost the same and lead to the same conclusion that there are more medications on average for those who are elective vs emergency admittance.***

PHP 2510 Only

Please display your code at the end of this entire question so you may receive partial credit on it.

8. (15 points) Note that the data in diabetes data clean.csv has patients who had repeat visits to the hospital. We know this to be true because `patient_nbr` is the unique patient identfier, and there are instances of recurrence. We would like to see whether the number of medications changes over time. In particular, for those patients who had 2 (or more) visits, we want to see whether the number of medications tends to be different from the first visit to second visit. Use the appropriate statistical test to investigate this. Note that this problem requires a fair bit of data wrangling to get the data prepared. So consider these hints:

- We can assume encounter id is a variable that uniquely identifies each hospital admission.
- We further assume that encounter IDs are only increasing, so if a given patient has two encounters, the first ID will always be less than the second ID.
- The shape of this data set is typically called "long format"; the patients have repeat measurement recorded as new rows. You will want to get these data in to "wide format". There are good examples you can find by Google-ing.
- Consider using the dplyr package. It's not necessary, but it might help.
- The data reshaping is definitely tricky, don't feel bad if you struggle a bit. It's important to get practice with this, because for many of us, data cleaning and reshaping is a big part of the work we have to do prior to model fitting.

```
diabetes <- diabetes %>% arrange(encounter_id)

diabetes_wide <- diabetes %>%
  add_count(patient_nbr) %>%
  filter(n>1) %>%
  group_by(patient_nbr) %>%
  summarise(n_meds1 = num_medications[1], n_meds2 = num_medications[2])


t.test(diabetes_wide$n_meds1, diabetes_wide$n_meds2, paired=T)
```

```
##
##  Paired t-test
##
## data:  diabetes_wide$n_meds1 and diabetes_wide$n_meds2
## t = -4.1448, df = 11365, p-value = 3.426e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5542588 -0.1983366
## sample estimates:
```

```
## mean of the differences
##              -0.3762977
```

Based on the paired t-test, we can see that the p-value is 3.4×10^{-5} ***with a 95% confidence interval of (-0.55, -0.19), this means that there is a difference and it appears there are less medications on the second visit as opposed to the first.***