THE UNIVERSITY OF CHICAGO
HARRIS SCHOOL
OF PUBLIC POLICY

Homework 3: Programming in R Sequence

DPSS 2024

Instructions

- There will be 5 programming in R questions worth 3 points each.
- The R questions will be multiple choice or file upload. However, each R question will have a part a) and part b).
- Part a) will contain the question stem, as well as multiple-choice responses or file upload. Please copy and paste your R code into part b. In order to receive full credit, code must be provided.
- If you get the answer wrong on any of the R questions, but still include code that is close to correct, you may earn partial credit.
- You will not earn extra credit for including your code in part b) of the R questions if you receive full credit in part a).
- While you may discuss approaches with your peers, any copying of code or output will be treated as a case of academic dishonesty. All submitted work should be your own. See the syllabus for details on group collaboration.

You may submit your assignment up to 3 times before the deadline. Only your most recent submission will be saved, so be careful when resubmitting. It is recommended that you write your answers down just in case you accidentally clear your previous submission.The Ed Discussion page is the best place to post questions for a quick response from the teaching assistants. While you may discuss approaches with your peers, any copying of code or output will be treated as a case of academic dishonesty. All submitted work should be your own. See the syllabus for details on group collaboration.

We have a set up section to help prepare you to load the data you will need to use.

Set Up

In the past HW #2, you should have downloaded two datasets entitled Traffic_Crashes-People.csv and Traffic_Crashes-Crashes.csv to a folder on your personal computer.**In this HW #3** assignment, you will work only with the dataset entitled Traffic_Crashes-_-Crashes.csv. Load the following packages and datasets to complete the programming in R questions.

Special note for international students, depending on your location, you may encounter two issues:

- There is a timeout issue that requires you to override default settings and we included code to do so
- API access in the U.S. is blocked :(even when you try a VPN. If this is the case for you then we also have the csv ready for you to load in so please download crash_api.csv from canvas.

```
library(tidyverse)
library(jsonlite) # used to call in data
# setting working directory
folder_path <- "/Users/josemacias/DPSS/Datasets" # use your own path here
setwd(folder_path) # passing the object into the function
# reading in the files
crashes <- read_csv("Traffic_Crashes_-_Crashes.csv")
##### FOR INTERNATIONAL STUDENTS #####
# if you are abroad then use the following to help extend the data pull time
options(timeout = max(1000, getOption("timeout")))
# if the above code did not work, then please download and read in the api csv
crash_api <- read_csv("crash_api.csv")
```

Brief Reminder

In the "Traffic_Crashes-_-Crashes" dataset, we see information about each traffic crash in Chicago for which the Chicago Policy Department responded. Data are available for some police districts in 2015, but citywide data are not available until September 2017. Many crash parameters are reported by the responding police officer, including street condition data, weather condition, and posted speed limits. The data dictionary for review can be found here and **you will need to review the data dictionary to complete this assignment.**

Programming in R Questions

Questions 1 ~ 2

Using tools and code from the API lectures, today we will download some data, you will need following URL for today:

`https://data.cityofchicago.org/resource/85ca-t3if.json?CRASH_MONTH=1&$limit=25000`

The link above lets you download the most recent 25,000 observations about January car crashes from the City of Chicago website, you can tell in the url query strings `CRASH_MONTH=1` plus `&$limit=25000` .

1. Next, do the following:
2. Select only the observations from year 2023.
3. In this filtered sample of car crashes (which occurred in January 2023), what is the most frequent primary contributory cause to the crash, second to "unable to determine"?

Hints: 1) There are different ways to focus only on observations during year 2023. 2) Obtain the date from the column entitled "crash_date," and look at how it is formatted. 3) Consider using tools from string manipulation and/or lubridate. Obtain the primary contributory cause from the variable entitled "prim_contributory_cause."

Select one answer from below:

- FAILING TO YIELD RIGHT-OF-WAY
- IMPROPER BACKING
- None of these answer choices
- IMPROPER OVERTAKING/PASSING

Question 3 ~ 4

For the remainder of this problem set, you will use the crashes dataset you used in HW2 (given in Week 1 materials).

What day of the week has the highest number of crashes?

Hints 1) To learn about CRASH_DAY_OF_THE_WEEK, you can go to the data dictionary listed here, check the bottom of the website and check the different columns until you find the one you need to learn more about. **Please make sure you check the key from the data dictionary! Otherwise, the numbering of the days may not match up with your expectations.**

- Saturday

- Sunday
- Monday
- Friday

Question 5 ~ 6

Use the programming concepts learned in R for loops, ifelse or casewhen, to create a variable which either equals 'Weekend' if day is either Sunday or Saturday, or otherwise equals 'Weekday'. How many crashes occurred on weekends? Select one from below:

- 166,959
- 446,746
- 521,422
- 192,567
- None of the above

Question 7 ~ 8

For question seven we will have you practice running a simple regression, although you have not seen much regression in R yet, this question will ask you to take a moment, self-teach some code, and attempt. This is a great example of leveling up in programming. When we run OLS with a binary dependent variable, we call this the **Linear Probability Model (LPM)**

1.
   - $Y = \beta_0 + \beta_1 X + \mu$
   - $E(\hat{Y}|X) = \beta_0 + \beta_1 X$

Special interpretation of E (Y | X) with binary Y (from law of total probability):

2. $E(Y|X) = 1 \times P(Y = 1|X) + 0 \times P(Y = 0|X) = P(Y = 1|X)$

Combining (1) and (2), we get the LPM interpretation:

$$P(Y = 1|X) = \beta_0 + \beta_1 X$$

Practical Interpretation

- $\hat{Y}$ is the probability that $Y_i = 1$, given X_i
- β_1 is the increase in the probability that $Y = 1$, given a unit increase in X. We can see as:

$$\Delta P(Y = 1|X) = \beta_1 \Delta X$$

  - Probabilities in decimal form e.g. if $\beta = 0.01$ means that a one unit increase in X increases the probability of the outcome by 1 percentage point
  - Note: 1 percentage point is not the same as 1 percent!

- Same interpretation with multiple X's
- t-statistics and F-statistics are as usual

Now then, the task is the following:

1) Please manipulate the data so that you create a new dependent binary variable where the variable is binary reflecting the presence of any injuries (INJURIES_TOTAL) that has a value greater than or equal to 1.

2) Use POSTED_SPEED_LIMIT as your X_1, now, *without correcting for heteroskedasticity,* to estimate the following regression formula in R:

$$Injuries = \beta_0 + \beta_1 \times (PostedSpeedLimit) + \mu$$

3) Please be ready to provide your code, the output and interpret your results by selecting one from below:

- The coefficient on posted speed limit is 4.28*10^-3, but it is not statistically significant. This means that there is no change in the probability of being in a car crash when the posted.
- The coefficient on posted speed limit is 4.28*10^-3, and statistically significant at the 0.001 level. This means that among car crashes, a 1 mph increase in the posted speed limit is associated with an increase in the probability of injury by 0.428 percent, without adjusting for
- The coefficient on posted speed limit is 4.28*10^-3, but it is not statistically significant. This means that conditional on being in a car crash, an increase in the posted speed limit is not associated with a change in the probability of injury.
- The coefficient on posted speed limit is 4.28*10^-3, but it is not statistically significant. This means that there is no change in the probability of being in a car crash when the posted speed limit increases.
- The coefficient on posted speed limit is 4.28*10^-3, and statistically significant at the 0.001 level. This means that among car crashes, a 1 mph increase in the posted speed limit is associated with an increase in the probability of injury by 0.428 percent, without adjusting for any other factors.

Hints 1) You may use the function `lm(y~x, data = yourdata)`, 2) there is sample code to learn more about LPM's in R here. 3) Once you run your lm model and assigned it a name, use the function `summary()` to view the complete output. In the linked example there are a couple of versions, you do not have to do corrections for *heteroskedasticity* in this question.

Question 9

For question nine, we will have you build on your regression in question eight by correcting for heteroskedasticity. As a reminder **Heteroskedasticity** refers to the situation where the variance of the error terms varies across observations, which can affect the reliability of the regression results. LPM models suffer from this issues.

Step 1: Install and the following libraries

```
install.packages("lmtest")
install.packages("sandwich")
library(lmtest)
library(sandwich)
```

Step 2: Let's correct for heteroskedasticity using standard errors by rerunning the regression in question four as shown.

```
# regression formula
lpm <-lm(data =crashes,formula =injury_binary~POSTED_SPEED_LIMIT)
# Correction for heteroskedasticity where
coeftest(lpm,vcov =vcovHC(lpm,type ="HC1"))
```

Code Explanation:

- `coeftest()` is a function from the lmtest package that provides hypothesis tests for model coefficients.
- `vcovHC()` is a function from the sandwich package that computes heteroskedasticity-consistent (robust) covariance matrix estimators.
- `type = "HC1"` specifies the type of robust standard error correction to use.

Step 3: Please interpret your result, what changed if anything? What does this say about our correction?