Midterm: Modules 1 - 3

Alex Shapovalov

2024-01-25

Note to students: Make sure to put your name in the “author” section above and do not change what is in the “date” section.

Midterm Instructions

- The exam should be submitted as a pdf file. The easiest way is to knit as an html file and convert to a pdf. There is a video in module 1 about how you can do that!
- Partial credit is awarded based off of the work shown.
- If a question asks for a graph, table, or calculation (like an average), make sure that it appears in your knitted document.
- Your exam should be your own work. While you can use the internet for help, any major deviations to methods seen in this course will be marked incorrect, even if it gives the correct answer.
- The code should be readable and commented. If I’m unsure what your code did, I can’t award partial credit!

Setup

```r
knitr::opts_chunk$set(echo = TRUE)

# Load the tidyverse, skimr, gt, and ggplot2 packages
pacman::p_load(tidyverse, skimr, gt, ggplot2)
#Alternatively:
library(tidyverse)
library(skimr)
library(gt)
library(ggplot2)

#Throughout this assignment we will be looking at a US economic dataset from 1967 to 2015.
#pce represents the personal consumption expenditures in billions of dollars
#pop gives the total population in thousands
#psavert gives the personal savings rate as a percentage
#unemploymed gives the median duration of unemployment measured in weeks
#unemploy gives the number of unemployed in thousands
#Import the dataset below and name it "econ"
econ <- read_csv("economics.csv")
```

1) Understanding the Data

1.1) Cases and Variables **(2 pts)**

How many cases and how many variables are in this dataset? How would you describe a case for this dataset? Classify each variable as either categorical or quantitative. Display any code used to answer these questions below.

```r
tibble(econ)
```

```
## # A tibble: 574 × 7
##    date         pce    pop psavert unemploymed unemploy month
##    <date>     <dbl>  <dbl>   <dbl>       <dbl>    <dbl> <dbl>
##  1 1967-07-01  507. 198712    12.6         4.5     2944     7
##  2 1967-08-01  510. 198911    12.6         4.7     2945     8
##  3 1967-09-01  516. 199113    11.9         4.6     2958     9
##  4 1967-10-01  512. 199311    12.9         4.9     3143    10
##  5 1967-11-01  517. 199498    12.8         4.7     3066    11
##  6 1967-12-01  525. 199657    11.8         4.8     3018    12
##  7 1968-01-01  531. 199808    11.7         5.1     2878     1
##  8 1968-02-01  534. 199920    12.3         4.5     3001     2
##  9 1968-03-01  544. 200056    11.7         4.1     2877     3
## 10 1968-04-01  544  200208    12.3         4.6     2709     4
## # ℹ 564 more rows
```

-[number of cases here] 574 Cases

-[Number of variables here] 7 Variables

-[Describe a case here] 1967-07-01, 506.6, 198712.0, 12.6, 4.5, 2944, 7

-[Classify your variables here] 1 is categorical (date) and the other 6 are numerical

1.2) Data Transformation **(1 pt)**

The month variable is being treated like a number, but it does not make sense to do so. Convert the data for the month column so it is instead being treated as a character and save it into your econ dataset. Check to make sure the column was converted by skimming your dataset.

```r
econ$month <- as.character(econ$month)

skim(econ)
```

Data summary

Name	econ
Number of rows	574
Number of columns	7
_______________________	
Column type frequency:	
character	1
Date	1
numeric	5
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
month	0	1	1	2	0	12	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	1967-07-01	2015-04-01	1991-05-16	574

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
pce	0	1	4820.09	3556.80	506.7	1578.3	3936.85	7626.33	12193.8	▇▅▃▂▂
pop	0	1	257159.65	36682.40	198712.0	224896.0	253060.00	290290.75	320402.3	▇▇▇▇▇
psavert	0	1	8.57	2.96	2.2	6.4	8.40	11.10	17.3	▂▇▇▅▁
unemploymed	0	1	8.61	4.11	4.0	6.0	7.50	9.10	25.2	▇▂▁▁▁
unemploy	0	1	7771.31	2641.96	2685.0	6284.0	7494.00	8685.50	15352.0	▂▇▇▂▁

1.3) Unemployment percentage **(1 pt)**

Add a column to the dataframe for the **percentage** of unemployed that month in thousands. Call this new column “pctunemploy”. Then display the first 3 cases in the dataframe to check the conversion occured. Hint: You will need to use the number of unemployed and population to calculate the percentage that are unemployed.

```r
pctunemploy <- c((econ$unemploy / econ$pop) * 100)

econ <- cbind(econ, pctunemploy)

head(econ)
```

```
##         date   pce    pop psavert unemploymed unemploy month pctunemploy
## 1 1967-07-01 506.7 198712    12.6         4.5     2944     7    1.481541
## 2 1967-08-01 509.8 198911    12.6         4.7     2945     8    1.480562
## 3 1967-09-01 515.6 199113    11.9         4.6     2958     9    1.485589
## 4 1967-10-01 512.2 199311    12.9         4.9     3143    10    1.576933
## 5 1967-11-01 517.4 199498    12.8         4.7     3066    11    1.536858
## 6 1967-12-01 525.1 199657    11.8         4.8     3018    12    1.511592
```

2) Categorical variables

2.1) Frequency Tables **(1 pt)**

Make and display a frequncy table for the month the data was collected in. Does looking at a distribution of the month provide us any meaningful information? Explain your response.

```r
monthfreq <- xtabs(formula = ~ month, data = econ)

monthfreq
```

```
## month
##  1 10 11 12  2  3  4  5  6  7  8  9
## 48 48 48 48 48 48 48 47 47 48 48 48
```

-[your response here]

No it doesn’t, there’s an equal amount of months and only tells us that there’s data for 47-48 years in the dataset

2.2 Reordering and Barcharts

As you may have noticed in the table from the previous question, R is ordering the months strangely (not in numerical order). Modify the data in R so that R understands that there is a specific order we should think of the month in. Then create and display a bar graph for the months. Hint: We will need to tell R that month is a factor.

```r
econ$month <-
  factor(x = econ$month,
         levels = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"))

monthfreq <- xtabs(formula = ~ month, data = econ)

monthfreq
```

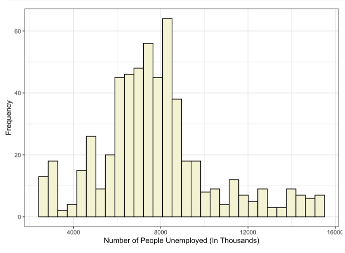
```
## month
##  1  2  3  4  5  6  7  8  9 10 11 12
## 48 48 48 48 47 47 48 48 48 48 48 48
```

3) Quantitative Variable

3.1) Histograms **(1 pt)**

Create and Display a histogram with 30 bins for the unemploy variable. Rename your axes as necessary so they better describe the variable and its units.

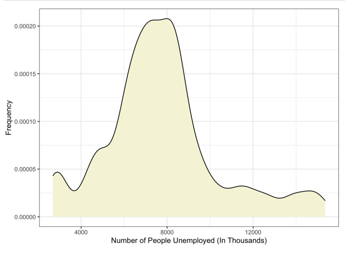
```r
gg_unemploy <- ggplot(data = econ, mapping = aes(x = unemploy)) + theme_bw()

gg_unemploy + geom_histogram(fill = "beige", color = "black", bins = 30) + labs(x = "Number of People Unemployed
(In Thousands)", y = "Frequency")
```



3.2) Density Plot **(1 pt)**

Repeat the graph above, but create a density plot instead of a histogram.

```r
gg_unemploy + geom_density(fill = "beige") + labs(x = "Number of People Unemployed (In Thousands)", y = "Frequenc
y")
```



Describe the shape of the data using the plots (histogram and density plot) above.

- (Your answer here)

The shape is mostly neutral (bell curve)

3.3) Mean vs Median **(1 pt)**

How should the mean and median of unemployment compare to each other? Briefly explain why

-(Your answer here)

The median and average should be similar since it is a normal distribution.

Find and display the mean and the median for the unemploy variable in the code chunk below:

```r
mean(econ$unemploy)
```

```
## [1] 7771.31
```

```r
median(econ$unemploy)
```

```
## [1] 7494
```

4) One Categorical and One Quantitative Variable

4.1) Boxplots of unemployment and month **(2.5 pts)**

Create box plots of unemployment by month. Make sure your graph has the following properties:

-Do not create it using small multiples.

-Have unemploy appear on the x-axis.

-Have month appear on the y-axis

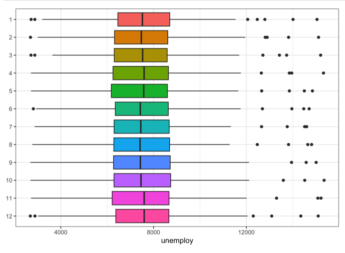
-Have the first month appear at the top of the y-axis and the twelfth month appear at the bottom of the y-axis.

-Have the x-axis be called “Unemployment (in thousands)”

-Have each boxplot have a unique color.

-Make sure the legend does not show.

```r
gg_unemploy + geom_boxplot(mapping = aes(y = fct_rev(month), fill = month), show.legend = F) + labs(y = NULL)
```



Is there a relationship between the month and the unemployment? If so, what is it?

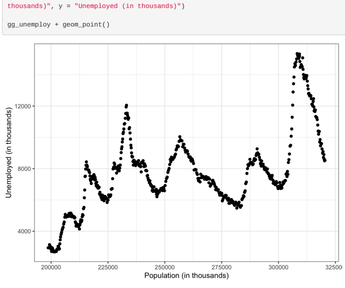
-[Your answer here] I don’t see much of a relationship, there are slight fluctuations but it is mostly consistent.

5) Two Quantitative Variables

5.1) Scatterplot **(2 pts)**

Create and display a scatterplot of unemploy vs population with the population (pop) on the x-axis and unemployment (unemploy) on the y-axis. Have the x-axis labeled as “Population (in thousands)” and the y-axis labeled as “Unemployed (in thousands)”.

```r
gg_unemploy <-  ggplot(data = econ, mapping = aes(x = pop, y = unemploy)) + theme_bw() + labs(x = "Population (in
thousands)", y = "Unemployed (in thousands)")

gg_unemploy + geom_point()
```



5.2) Describing Scatterplots **(2 pts)**

Describe the 4 characteristics of the association between the population and unemployment. Find (and display) the correlation between these two variables as well (as it will be needed in describing the association). Does this correlation make sense? Why or why not?

```r
cor(x = econ$pop,
    y = econ$unemploy,
    use = "complete.obs")
```

```
## [1] 0.6337165
```

- [Describe the 4 characteristics here] No association, no outliers, no curvature (wavy), strong relationship

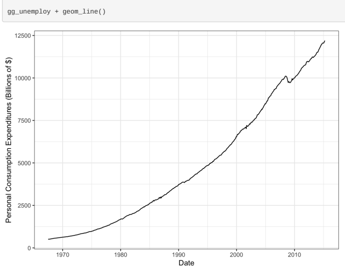
-[Does the correlation make sense?] Theoretically yes, population growth over time shouldn’t lead to too great of a jump in unemployment

6) Date vs PCE

6.1) Line Graph **(2 pts)**

Create and display a linegraph of date vs pce with the date on the x-axis and Personal Consumption Expenditures in billions of dollars (pce) on the y-axis. Have the x-axis labeled as “Date” and the y-axis labeled as “Personal Consumption Expenditures (Billions of $)”.

```r
gg_unemploy <-  ggplot(data = econ, mapping = aes(x = date, y = pce)) + theme_bw() + labs(x = "Date", y = "Person
al Consumption Expenditures (Billions of $)")

gg_unemploy + geom_line()
```



6.2) Describing Line Graphs **(1 pt)**

Describe the 4 characteristics of the association between the date and pce.

- [Describe the 4 characteristics here] Positive association, nearly no outliers, curves up, strong relationship