Empirical Exercise 1, Stock and Watson Chapter 2

Econ 440 - Introduction to Econometrics

Your Name + CWID + youremail@fullerton.edu

21 March 2022

Tips

Explicitly mark the question or question number in your code. Show the output, not just the code: A few months/years from now, the packages will have been updated and the code may no longer run, so it's a good habit to keep a record of the output. Don't forget to answer the questions!

Load dataset

```
library(readxl)
df1 <- read_xlsx("Age_HourlyEarnings.xlsx",
                 col_names=TRUE,
                 skip=1,
                 trim_ws=TRUE)
```

```
## New names:
## * `` -> ...12
## * `` -> ...13
```

Remove empty rows and columns: simple and symmetric!

```
df1 <- df1[, colMeans(is.na(df1)) != 1]
df1 <- df1[rowMeans(is.na(df1)) != 1,]
```

There are other ways to do this, e.g. using the `Filter()` and `complete.cases()` functions.

Reshape data from wide to long form:

```
library("tidyr")
df2 <- gather(df1, Age, Probability, -AHE)
```

Make Age a numeric or integer value:

```
df2$Age <- as.integer(df2$Age)
```

This is needed after conversion from wide (horizontal) to long (vertical), since the variable `Age` in dataframe `df2` is constructed from the column names in dataframe `df1`. And `colnames` are `strings` (aka `character vectors` in `R`).

Now let's look at our data

```
head(df2)
```

```
## # A tibble: 6 x 3
##     AHE   Age Probability
##   <dbl> <int>       <dbl>
## 1     5    25     0.00298
## 2     6    25     0.00116
## 3     7    25     0.00247
## 4     8    25     0.00240
## 5     9    25     0.00356
## 6    10    25     0.00516
```

(a) Compute the marginal distribution of Age.

This is a situation where data in wide format is convenient! Often data is more convenient in long format.

Sum by column: The marginal distribution of Age

```
colSums(df1[,names(df1) != "AHE"])
```

```
##       25       26       27       28       29       30       31       32
## 0.084890 0.092231 0.085471 0.093393 0.103496 0.104731 0.103932 0.108075
##       33       34
## 0.108802 0.114979
```

Sum by row:

```
rowSums(df1[,names(df1) != "AHE"])
```

```
##  [1] 0.0170797 0.0119195 0.0220219 0.0211498 0.0278363 0.0457155 0.0347409
##  [8] 0.0698452 0.0374300 0.0630133 0.0404099 0.0324878 0.0566902 0.0284178
## [15] 0.0590886 0.0359038 0.0523294 0.0676648 0.0364852 0.0415001 0.0423723
## [22] 0.0486227 0.0312523 0.0214405 0.0177339 0.0085762 0.0097391 0.0044335
## [29] 0.0140999
```

If the data is in long format, we have to work a little harder. Below are several approaches.

With the data in long format:

```
sum(df2[df2$Age == 25,]$Probability)
```

```
## [1] 0.08489
```

```
sum(df2[df2$Age == 26,]$Probability)
```

```
## [1] 0.092231
```

```
sum(df2[df2$Age == 27,]$Probability)
```

```
## [1] 0.085471
```

with a split/apply routine to group by Age

```
sapply(split(df2, df2$Age), function(x) sum(x$Probability))
```

```
##       25       26       27       28       29       30       31       32 
## 0.084890 0.092231 0.085471 0.093393 0.103496 0.104731 0.103932 0.108075 
##       33       34 
## 0.108802 0.114979
```

To see how this works, try this command: `split(df2, df2$Age)`. It returns a list of dataframes. Then `sapply` applies the function we defined as `function(x) sum(x$Probability)` to each dataframe.

or we can always write a loop

```r
for (age in 25:34){
    print(sum(df2[df2$Age == age,]$Probability))
}
```

```
## [1] 0.08489
## [1] 0.092231
## [1] 0.085471
## [1] 0.093393
## [1] 0.1035
## [1] 0.10473
## [1] 0.10393
## [1] 0.10807
## [1] 0.1088
## [1] 0.11498
```

(b) Compute the mean of AHE for each value of Age

```r
## fill this space with your code
```

(c) Compute and plot the mean of AHE versus Age. Are average hourly earnings and age related? Explain.

```r
## fill this space with your code
```

(d) Use the law of iterated expectations to compute the mean of AHE

```r
## fill this space with your code
```

(e) Compute the variance of AHE.

```r
## fill this space with your code
```

(f) Compute the covariance between AHE and Age.

```r
## fill this space with your code
```

(g) Compute the correlation between AHE and Age.

```r
## fill this space with your code
```

(h) Relate your answers in (f) and (g) to the plot you constructed in (c).

```
## fill this space with your code
```