R-problemset_econometrics

2022-10-19

This is the R-problem set 1 for Econometrics (Fall 2022), New York University Author: Carlos Figueroa (cdf5579)

Let’s begin.

```
#lets load some libraries first

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
```

```
library(magrittr)
library(readxl)
library(tinytex)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(tibble)
library(ggplot2)
```

lets display the data first

```
data = data.frame(read.csv(file = "bertrand2004.csv", header = TRUE , sep =","))

head(data)
```

```
##   education yearsexp sex race call rid sid
## 1         4        6   f    w    0   1   0
## 2         3        6   f    w    0   1   0
## 3         4        6   f    b    0   0   0
## 4         3        6   f    b    0   0   0
## 5         3       22   f    w    0   1   0
## 6         4        6   m    w    0   1   1
```

QUESTION 1:——————————————————————

A. What is the interpretation of C¯ in this situation? Hint: this is a 0/1 variable like a coin flip, so what is the mean of a Bernoulli?

Since C (which stands for a person recieving a call back or not) is between zero and one, it will behave similarly to a bernoulli distribution. That would mean that C bar (the sample mean of C) is going to be equal to p, or the probability of receiving a call back. We can get this p for different subgroups and the group in general.

B. What about M¯ ? Hint: the same as the previous hint.

Since M also follows a bernoulli distribution, its sample mean is going to be p, so such mean is going to be equal to the probability of being a man in this dataset.

c. Let C¯M and C¯F represent the call back rates for males and females respectively. Set up a hypothesis test for whether or not there is discrimination on sex.

Given that we stated that C bar represents the sample distribution, and the probability of recieving a callback, we will want to analyze C bar (M) and C bar (F) to see if one group has higher chances of recieving a callback than the other.

Then, our hypothesis testing will be for a two tail rest:

H0: C¯M = C¯F then C¯M - C¯F = 0 H1: C¯M != C¯F

If we disprove H0, we will say that there is discrimination on sex in this particular dataset.

Here, in this section, we are going to see if both groups have similar qualifications in order to make this hypothesis test viable. If men or women are more qualified, that will change our view of the results of our hypothesis test.

QUESTION 2——————————————————————

Calculate the mean years of experience and education by sex

```
exp_male = mean(data[data$sex == 'm', 'yearsexp'])
exp_female = mean(data[data$sex == 'f', 'yearsexp'])


educ_male = mean(data[data$sex == 'm', 'education'])
educ_female = mean(data[data$sex == 'f', 'education'])


print("Mean of years of Experience for males")
```

```
## [1] "Mean of years of Experience for males"
```

```
exp_male
```

```
## [1] 7.579042
```

```
print("Mean of years of Experience for females")
```

```
## [1] "Mean of years of Experience for females"
```

```
exp_female
```

```
## [1] 7.964218
```

```
print("Mean of years of Education for males")
```

```
## [1] "Mean of years of Education for males"
```

```
educ_male
```

```
## [1] 3.791328
```

```
print("Mean of years of Education for females")
```

```
## [1] "Mean of years of Education for females"
```

```
educ_female
```

```
## [1] 3.611784
```

Female have more experience than male, and male have more education. But only by little, they are pretty close. Let’s test if they are significally different.

QUESTION 3———————————————————————

Calculate the standard errors and do a t-test (by hand or using the test commands) for whether years of experience or education are different by sex

```
print("Standard Deviations for Male (years of experience)")
```

```
## [1] "Standard Deviations for Male (years of experience)"
```

```
sd(data[data$sex == 'm', 'yearsexp'])
```

```
## [1] 5.209866
```

```
print("Standard Deviations for Female (years of experience)")
```

```
## [1] "Standard Deviations for Female (years of experience)"
```

```
sd(data[data$sex == 'f', 'yearsexp'])
```

```
## [1] 5.00232
```

```
print("Standard Deviations for Male (years of education)")
```

```
## [1] "Standard Deviations for Male (years of education)"
```

```
sd(data[data$sex == 'm', 'education'])
```

```
## [1] 0.6412555
```

```
print("Standard Deviations for Female (years of education)")
```

```
## [1] "Standard Deviations for Female (years of education)"
```

```
sd(data[data$sex == 'f', 'education'])
```

```
## [1] 0.6135367
```

```
#lets confirm our results with proper t-test package

t.test(data[data$sex == 'm', 'yearsexp'], data[data$sex == 'f', 'yearsexp'], var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  data[data$sex == "m", "yearsexp"] and data[data$sex == "f", "yearsexp"]
## t = -2.2273, df = 4822, p-value = 0.02597
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.72420773 -0.04614426
## sample estimates:
## mean of x mean of y
##  7.579042  7.964218
```

```
t.test(data[data$sex == 'm', 'education'], data[data$sex == 'f', 'education'], var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  data[data$sex == "m", "education"] and data[data$sex == "f", "education"]
## t = 8.4575, df = 4822, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1379258 0.2211626
## sample estimates:
## mean of x mean of y
##  3.791328  3.611784
```

QUESTION 4——————————————————

Now, using the same commands (or any other commands you prefer), calculate the difference in call back rates by sex.

```
print("Probability of recieving a callback by being a male")
```

```
## [1] "Probability of recieving a callback by being a male"
```

```
mean(data[data$sex == 'm', 'call'])
```

```
## [1] 0.07407407
```

```
print("Probability of recieving a callback by being a female")
```

```
## [1] "Probability of recieving a callback by being a female"
```

```
mean(data[data$sex == 'f', 'call'])
```

```
## [1] 0.08259349
```

QUESTION 5——————————————————

Finally, do a t-test to see if these differences are statistically significant. You may do this by hand, with the t.test command, or with any other command.

```
t.test(data[data$sex == 'm', 'call'], data[data$sex == 'f', 'call'], var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  data[data$sex == "m", "call"] and data[data$sex == "f", "call"]
## t = -0.91371, df = 4822, p-value = 0.3609
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.026798626  0.009759796
## sample estimates:
##  mean of x  mean of y
## 0.07407407 0.08259349
```

Not that different indeed.

PART 2 ---------—————————————————----------

Regression analysis

QUESTION 6——————————————————————

First, load the dataset just as before. For this exercise, you will also need the ggplot2 library loaded.

```
df = data.frame(read.csv(file = "stocks.csv", header = TRUE , sep =","))


head(df)
```

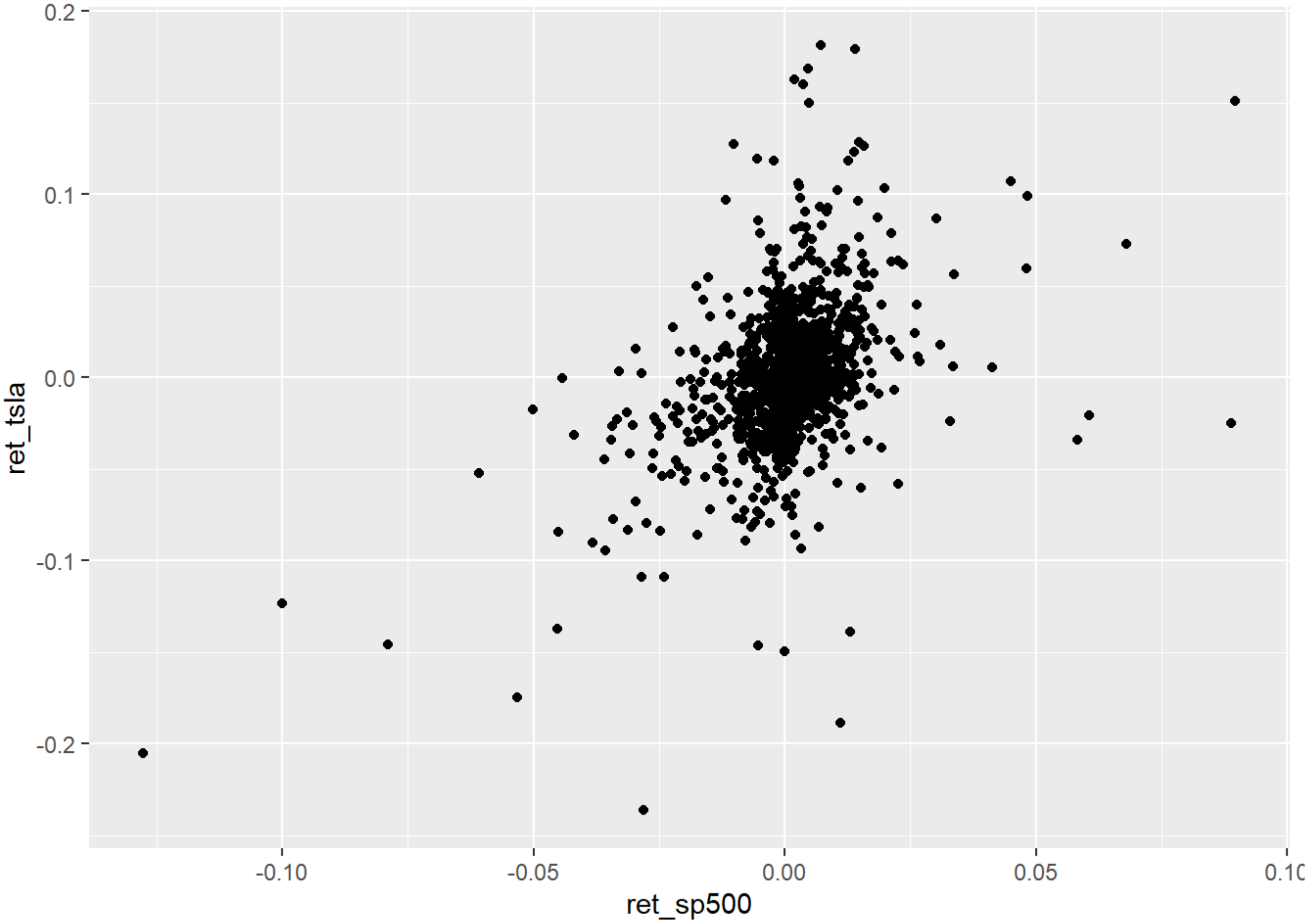
```
##   t  date_s  ret_sp500      ret_f   ret_aapl   ret_msft   ret_tsla
## 1 1 9/20/16  0.0002991 -0.0091249 -0.0000881 -0.0021100 -0.0082730
## 2 2 9/21/16  0.0108580  0.0074720 -0.0001760  0.0165841  0.0028302
## 3 3 9/22/16  0.0064788  0.0074165  0.0093790  0.0010383  0.0058788
## 4 4 9/23/16 -0.0057533 -0.0008215 -0.0168042 -0.0067683  0.0049290
## 5 5 9/26/16 -0.0086248 -0.0132344  0.0015071 -0.0092712  0.0073960
## 6 6 9/27/16  0.0064235 -0.0025007  0.0018587  0.0182851 -0.0153330
```

QUESTION 7——————————————————————

Second, let’s plot the data. Using ggplot2 make a scatter plot of the returns of Tesla as the dependent variable and the returns of market as the independent variable. To this, load the ggplot2 package. For the command, here is a rough schema (that you have to fill in yourself):

The first part tells R to make a ggplot object, using which dataset, x, and y variables are specified. The second part says to create a scatter plot.

```
ggplot(data = df, aes(x = ret_sp500,y = ret_tsla)) + geom_point()
```

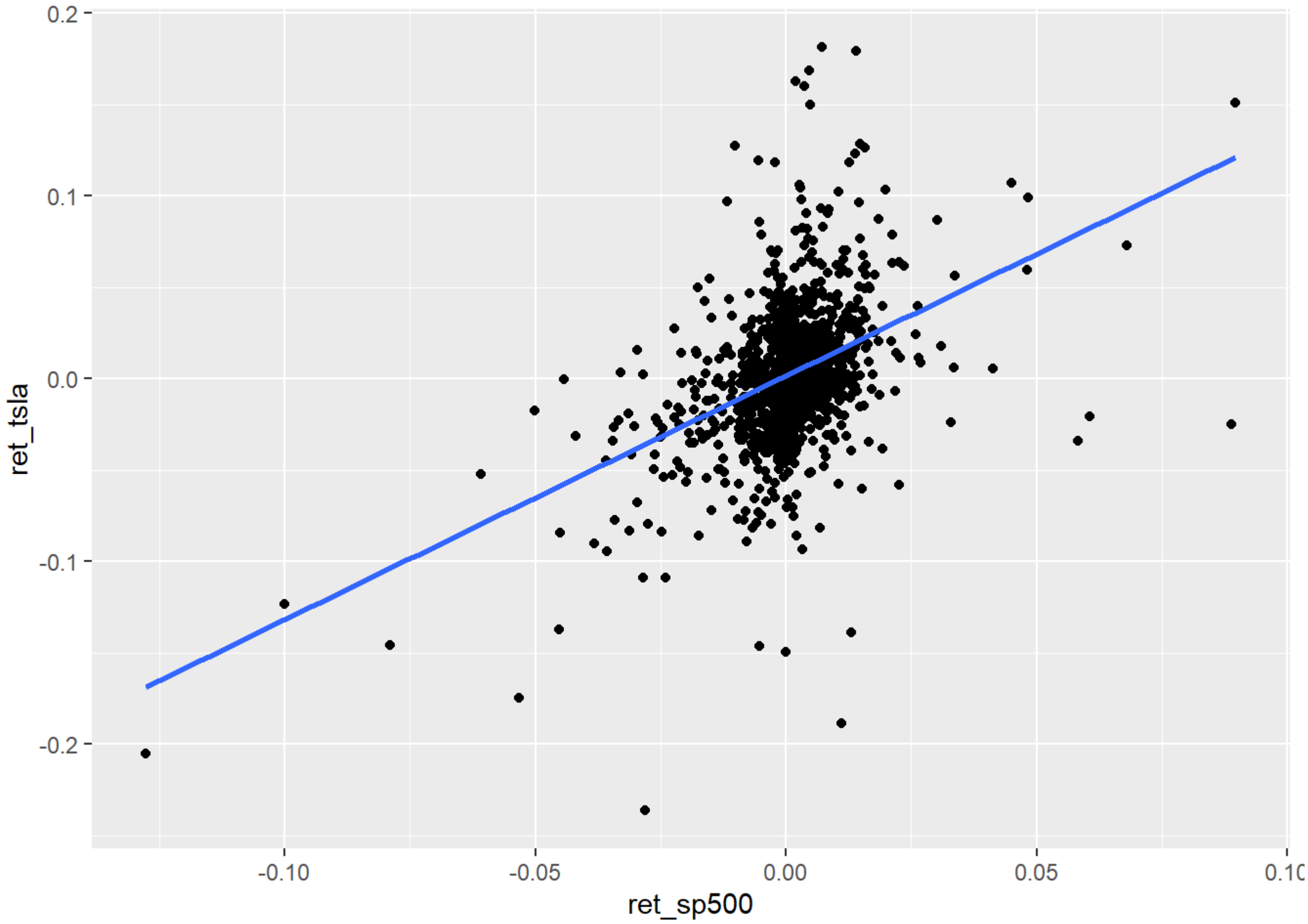


QUESTION 8———————————————————————

Now let’s add a best fit line. To do this we copy and paste the same opening as above with an additional command telling R to add the line:

```
ggplot(data = df, aes(x = ret_sp500,y = ret_tsla)) + geom_point() + geom_smooth(method = 'lm',se
=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



QUESTION 9——————————————————————

Now let’s run a regression to calculate the beta of Tesla stock. To do this we need to use the lm command. The schematic as below, where you need to fill in the y, x, and data entries yourself:

```
lm(ret_tsla ~ ret_sp500, data = df)
```

```
##
## Call:
## lm(formula = ret_tsla ~ ret_sp500, data = df)
##
## Coefficients:
## (Intercept)    ret_sp500
##    0.001531     1.334472
```

QUESTION 10——————————————————————

What is the β of Tesla stock? Calculate the β for each of the other three stocks in the data. Fill in a table that looks like the one at the end of the homework.

1.334472

```
lm(ret_tsla ~ ret_sp500, data = df)
```

```
##
## Call:
## lm(formula = ret_tsla ~ ret_sp500, data = df)
##
## Coefficients:
## (Intercept)    ret_sp500
##    0.001531     1.334472
```

```
lm(ret_aapl ~ ret_sp500, data = df)
```

```
##
## Call:
## lm(formula = ret_aapl ~ ret_sp500, data = df)
##
## Coefficients:
## (Intercept)    ret_sp500
##   0.0006651    1.1987471
```

```
lm(ret_msft ~ ret_sp500, data = df)
```

```
##
## Call:
## lm(formula = ret_msft ~ ret_sp500, data = df)
##
## Coefficients:
## (Intercept)    ret_sp500
##   0.0007002    1.1883818
```

```
lm(ret_f ~ ret_sp500, data = df)
```

```
##
## Call:
## lm(formula = ret_f ~ ret_sp500, data = df)
##
## Coefficients:
## (Intercept)    ret_sp500
##  -0.0003673    1.0576747
```

So the table will go:

For Tesla: 1.334472, Apple: 1.1987471, Microsoft: 1.1883818, Ford: 1.0576747

QUESTION 11——————————————————————————

The results from the base lm command are a bit bare. To learn more, we use the summary command. To do this, we need to save our model and then feed it to the summary command:

Once you’ve run this command, you can report the variance of the error term from your regressions, σ2U, by squaring the residual standard error. Run the command for each of the four stocks in our data—add these results to your table containing the β’s. Which stock has the most volatile returns net of market comovement (highest σ2U)?.

```
model = lm(ret_tsla ~ ret_sp500, data = df)

summary(model)
```

```
##
## Call:
## lm(formula = ret_tsla ~ ret_sp500, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.204911 -0.017303 -0.000874  0.015555  0.170268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0015311  0.0009473   1.616    0.106
## ret_sp500   1.3344723  0.0784592  17.008   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03354 on 1255 degrees of freedom
## Multiple R-squared:  0.1873, Adjusted R-squared:  0.1867
## F-statistic: 289.3 on 1 and 1255 DF,  p-value: < 2.2e-16
```

```
model_1 = lm(ret_aapl ~ ret_sp500, data = df)

summary(model_1)
```

```
## 
## Call:
## lm(formula = ret_aapl ~ ret_sp500, data = df)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.075539 -0.006274 -0.000161  0.005880  0.089739 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.0006651  0.0003482    1.91   0.0564 .  
## ret_sp500   1.1987471  0.0288428   41.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.01233 on 1255 degrees of freedom
## Multiple R-squared:  0.5792, Adjusted R-squared:  0.5789 
## F-statistic:  1727 on 1 and 1255 DF,  p-value: < 2.2e-16
```

```
model_2 = lm(ret_msft ~ ret_sp500, data = df)

summary(model_2)
```

```
## 
## Call:
## lm(formula = ret_msft ~ ret_sp500, data = df)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.038636 -0.005017 -0.000121  0.004611  0.051891 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.0007002  0.0002674   2.618  0.00894 ** 
## ret_sp500   1.1883818  0.0221479  53.657  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.009469 on 1255 degrees of freedom
## Multiple R-squared:  0.6964, Adjusted R-squared:  0.6962 
## F-statistic:  2879 on 1 and 1255 DF,  p-value: < 2.2e-16
```

```
model_3 = lm(ret_f ~ ret_sp500, data = df)

summary(model_3)
```

```
##
## Call:
## lm(formula = ret_f ~ ret_sp500, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.111033 -0.008254 -0.000070  0.007213  0.116108
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0003673  0.0005145  -0.714    0.475
## ret_sp500    1.0576747  0.0426167  24.818   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01822 on 1255 degrees of freedom
## Multiple R-squared:  0.3292, Adjusted R-squared:  0.3287
## F-statistic: 615.9 on 1 and 1255 DF,  p-value: < 2.2e-16
```

```
0.03354^2
```

```
## [1] 0.001124932
```

```
0.01233^2
```

```
## [1] 0.0001520289
```

```
0.009469^2
```

```
## [1] 8.966196e-05
```

```
0.01822^2
```

```
## [1] 0.0003319684
```

For Tesla: 1.334472 and 0.001124932, Apple: 1.1987471 and 0.0001520289, Microsoft: 1.1883818 and 8.966196e-05, Ford: 1.0576747 and 0.0003319684

The highest u variance is for tesla, with 0.001124932. So Tesla is the most volatile stock.