ECON-UB 251

Assignment 2, Fall 2022

The learning goals of this assignments are:

1. familiarize with the theory of OLS estimation
2. developing data wrangling skills
3. apply statistical analysis and the linear regression model to analyze data

I *strongly* prefer that you complete the assignment in Rmarkdown and a sample template is provided in Brightspace. You can knit the document to Word or PDF (or export to PDF the Word document; no HTML). Set the `echo` option to `TRUE` so that I can see the code you are using to conduct the analysis. You can discuss the assignment with other students, but each student should submit his/her original work.

Submit in `Brightspace` by 2pm on **Tuesday October 18th, 2022**.

1 Theory

1.1 [10%] Simulating the OV bias

A simulation exercise means generating artificial data to evaluate the behavior of an estimator in that specific controlled environment. We will proceed as follows:

1. Simulate artificial data for for X_1 and X_2 as follows:
   - $X_{1i} \sim N(0,1)$ and $X_{2i} = 0.1X_{1i} + \eta_i$ with $\eta_i \sim N(0,1)$
2. Generate Y_i according to this model: $Y_i = 1.1X_{1i} + 0.5X_{2i} + \epsilon_i$ with $\epsilon_i \sim N(0,1)$
   - both X_1 and X_2 are used to generate Y
3. Estimate by OLS the model $Y_i = \beta_0 + \beta_1 X_{1i} + \xi_i$ with $\xi_i \sim N(0,1)$
   - notice that we are only regressing Y on X_1 and omitting X_2

Repeat steps 1-3 B times and each time store the value of $\hat{\beta}_1$ and the t-statistic calculated as

$$t_1 = (\hat{\beta}_1 - 1.1)/SE(\hat{\beta}_1)$$

where 1.1 represents the true value of β_1 for this simulation exercise. Plot a histogram of the B values of t_1 together with the standard normal distribution. Discuss:

- Under what conditions do we have OV bias? Did we design the simulation in a way to produce biased estimates? Why?
- The simulation results in terms of the distribution of t_1 and whether you find evidence of OV bias
- What would happen to the plot if we would generate X_2 as $X_{2i} = \eta_i$ instead of using $X_{2i} = 0.1X_{1i} + \eta_i$? Discuss you prediction of what (and why) will happen to the OV bias and test the prediction by plotting the histogram of t_1 together with the standard normal in this new simulation environment.

[Python code provided in the Appendix]

```
set.seed(123) # fixing the seed allows to get the same random numbers each time
              # otherwise the clock is used and a different set of random values is used
N = 500       # sample size
B = 1000      # repetitions of the simulation exercise

beta1 <- rep(NA, B) # initialize vectors to store the results of each repetition
tstat <- beta1
```

```r
rho   <- beta1

for (b in 1:B)      # loop repeating B times steps 1-3 above
{
  X1 = rnorm(N)                          # 1) generate artificial X1
  X2 = 0.1 * X1 + rnorm(N)               # 1) generate artificial X2
  Y  = 1.1 * X1 + 0.5 * X2 + rnorm(N)   # 2) generate artificial Y
  rho[b] <- cor(X1, X2)                  # store the correlation

  fit      <- lm(Y ~ X1)                       # 3) regress Y on X1
  beta1[b] <- summary(fit)$coefficients[2,1]  # store the beta_hat and tstat
  tstat[b] <- (beta1[b] - 1.1) / summary(fit)$coefficients[2,2]
}

ggplot(data.frame(tstat = tstat), aes(x = tstat)) +
  geom_histogram(aes(y = ..density..), bins = 100, fill = "tomato4", alpha=0.3) +
  stat_function(fun = dnorm, args=list(mean = 0, sd = 1),  color="dodgerblue3") +
  theme_classic() +
  xlim(min(-5, min(tstat)), max(5, max(tstat)))
```

2 Empirical

- Read the Freddie Mac file that you used in Assignment 1
- Eliminate rows that contain missing values and give a name to the columns

2.1 [10%] Let's investigate the heterogeneity of some key variables across US states in our sample. We will consider the following variables: interest rate, credit score, DTI, UPB, and LTV. Count the number of loans in each state and calculate the sample averages for the variables mentioned above. Show a Table with the top 20 states by number of loans originated and *discuss whether you find large difference across states in these variables.*

- Use the following `dplyr` functions as shown in the code below:
  1. `group_by()`: to group observations by a certain characteristic (e.g., `State`)
  2. `summarize()`: used in combination with `group_by()`calculates a function on each group
  3. `arrange()`: sort the data frame according to a variable

```r
sample_orig_2020 %>%
  group_by(State) %>%
  summarize(N = n(),
            IR = mean(Interest_Rate),
            CS = mean(Credit_Score),
            DTI = mean(DTI),
            UPB = mean(UPB/1000),
            LTV = mean(LTV)) %>%
  arrange(desc(N)) %>%
  head(20) %>% knitr::kable(digits=3)
```

2.2 [10%] Are high credit score borrowers different, in some other dimension than credit score, from low credit score borrowers? Separate borrowers in *prime* and *subprime* based on their credit score being higher/lower relative to 670. Discuss the results.

3 Statistical analysis and plotting (using `ggplot2`)

3.1 [10%] Plot a histogram of *DTI*, *UPB* (divide this variable by $1,000), and *LTV*.

- Discuss the distribution characteristics of these variables.

- Are there outliers in the distribution of these variables?
- Based on economic reasoning, do you believe these variables should be pricing factors that determine the interest rate of the loan?

3.2 [10%] Calculate the correlation between the credit score, DTI, UPB, and LTV variables and discuss whether you have any concern about the possibility of multicollinearity in the data

4. **Regression Analysis**

4.1 [20%] Estimate a regression of `Interest_Rate` on `Credit_Score`, `DTI`, `UPB` (in thousand of dollars), and `LTV` using heteroskedastic-corrected standard errors.

- Provide an interpretation of the estimated coefficient of `Credit_Score` in this regression
- Do the coefficients of DTI, UPB, and LTV have the sign you expected?
- Discuss the statistical significance of the variables at 5% level
- Compare the $\bar{R}^2$ of this regression to the one with only credit score and discuss which model you consider more accurate
- Did the coefficient estimate of `Credit_Score` change significantly once we added DTI, UPB, and LTV?

4.2 [10%] Test the joint hypothesis $\beta_{DTI} = \beta_{UPB} = \beta_{LTV} = 0$ at 5% significance level

4.3 [20%] The `State` variable represents the state location where the property is located[1]. We could include this variable in the regression to control for state-specific characteristics of the mortgage market.

- Add the `State` variable to the regression model of `Interest_Rate` on `Credit_Score`, `DTI`, `UPB`, and `LTV` (see command below). `R` handles this variable by creating a binary (dummy) variable for each state. For example, the variable `StateNY` takes value 1 for all loans related to a property in NY state and 0 otherwise. You regression will include the four variable from the previous regression plus 52 dummy variables (total 56 regressors).

- Take two states that have a statistically significant coefficient at 10%, in one case a positive value and the other negative. Interpret the two coefficients.

- Did the inclusion of the `State` variable change significantly the estimate of the `Credit_Score` coefficient? what about the coefficients of DTI, UPB, and LTV? Why?

[1] The variable takes 53 values that include the 50 states plus Washington DC (DC), Puerto Rico (PR), and Guyana (GU).

List of Variables

Field	Name	Type
1	Credit Score	Numeric
2	First Payment Date	Date
3	First Time Homebuyer Flag	Alpha
4	Maturity Date	Date
5	Metropolitan Statistical Area (MSA) Or Metropolitan Division	Numeric
6	Mortgage Insurance Percentage (MI %)	Numeric
7	Number of Units	Numeric
8	Occupancy Status	Alpha
9	Original Combined Loan-to-Value (CLTV)	Numeric
10	Original Debt-to-Income (DTI) Ratio	Numeric
11	Original UPB	Numeric
12	Original Loan-to-Value (LTV)	Numeric
13	Original Interest Rate	Numeric - 6,3
14	Channel	Alpha
15	Prepayment Penalty Mortgage (PPM) Flag	Alpha
16	Amortization Type (Formerly Product Type)	Alpha
17	Property State	Alpha
18	Property Type	Alpha
19	Postal Code	Numeric
20	Loan Sequence Number	Alpha Numeric - PYYQnXXXXXXX
21	Loan Purpose	Alpha
22	Original Loan Term	Numeric
23	Number of Borrowers	Numeric
24	Seller Name	Alpha Numeric
25	Servicer Name	Alpha Numeric
26	Super Conforming Flag	Alpha
27	Pre-HARP Loan Sequence Number	Alpha Numeric - PYYQnXXXXXXX
28	Program Indicator	Alpha Numeric
29	HARP Indicator	Alpha
30	Property Valuation Method	Numeric
31	Interest Only (I/O) Indicator	Alpha

Simulationin Python

```
import numpy as np
import scipy.stats as sp
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.formula.api as smf
import statsmodels.api as sm
plt.style.use('seaborn')
```

```
N=500
B=1000
beta1=[]
tstat=[]
```

```
np.random.seed(1)
for b in range(B):                                              #range iterates from 0 to 999 here
    x1=np.random.normal(0,1,N)                                  # 1) generate artificial X1
    x2=0.1*x1+np.random.normal(0,1,N)                           # 1) generate artificial X2
    y=1.1*x1+0.5*x2+np.random.normal(0,1,N)                     # 2) generate artificial Y
    rho=np.corrcoef(x1,y)                                       # store the correlation

    model = sm.OLS(y, x1)                                       # 3) regress Y on X1
    result = model.fit()

    beta1.append(result.params[0])
    tstat.append((beta1[b]-1.1)/(result.bse[0]))                # store the beta_hat and
```

```
myhist=plt.hist(tstat,bins=100, density=True)
x_axis = np.arange(-5, 5, 0.001)
mynorm=plt.plot(x_axis, sp.norm.pdf(x_axis,0,1)) # Mean = 0, SD = 1
plt.xlabel('tstat')
plt.ylabel('density')
plt.show()
```