ECON-UB 251

Assignment 1, Fall 2022 (Sebastiano Manzan)

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 (this requires the installation of LaTeX, either MacTeX in Mac or MiKTeX in Windows[1]). 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 Thursday September 29, 2022.

1. **Theory**

1.1 [10%] For the model $Y_i = \beta_1 X_i + u_i$

- Derive the OLS estimator of β_1
- This model is called *regression through the origin*: can you explain why?
- If the intercept β_0 is actually different from zero, do you think this model will provide a *biased* estimate of the slope coefficient? Provide an intuitive argument for why this could be the case (graphical?)

1.2 [10%] What happens to the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ if we multiply the dependent variable by 100, that is, the regression model becomes $(100 * Y_i) = \beta_0 + \beta_1 X_i + u_i$? and what if we multiply the independent variable X_i by 100, that is, $Y_i = \beta_0 + \beta_1(100 * X_i) + u_i$? Justify your answer based on the formula of the OLS estimators

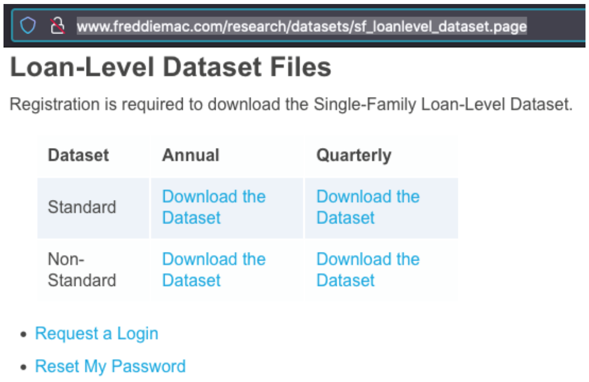
1.3 [10%] What happens to the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ if we standardize both the dependent and independent variables? That is, we define $Y_i^* = (Y_i - \bar{Y})/\sigma_Y$ and $X_i^* = (X_i - \bar{X})/\sigma_X$ and run the regression $Y_i^* = \beta_0 + \beta_1 X_i^* + \epsilon_i^*$? Justify your answer based on the formula of the OLS estimators

2. **Empirical**

In this assignment we analyze a sample of loans that originated in the US between 1999 and 2020 that were purchased by Freddie Mac. The data are available at this link and you are required to register to download the data (click on `Request a login`). Once you have obtained the credentials, click on `Download the dataset` that corresponds to `Standard` and `Annual`. After entering your credentials you will see a list of files that are either all loans originated in a certain year (`historical_data_year.zip`) or a sample

[1] You can also produce a Pdf by knitting to Word and then export to Pdf from Word

(`sample_year.zip`). Download the file assigned to you in the Table below (alternatively, you can download the historical file if you have a computer powerful enough to handle biggish data). Save it to a location in your drive and unzip the file. The unzipped file consists of a folder `sample_year` that contains the origination file (`sample_origination_year.txt`) and the performance file (`sample_svcg_year.txt`). We will use the **origination** file in the analysis.



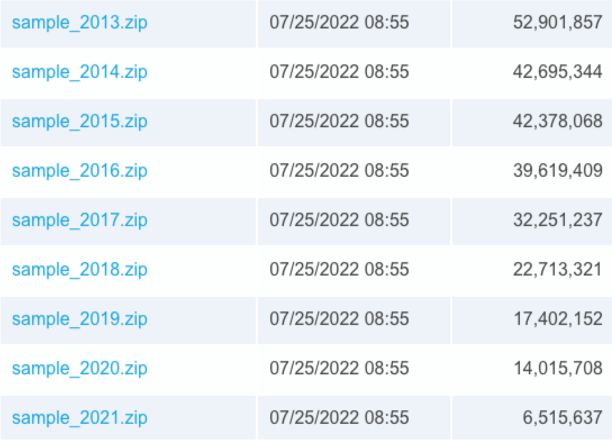
www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page

Loan-Level Dataset Files

Registration is required to download the Single-Family Loan-Level Dataset.

Dataset	Annual	Quarterly
Standard	Download the Dataset	Download the Dataset
Non-Standard	Download the Dataset	Download the Dataset

- Request a Login
- Reset My Password



sample_2013.zip	07/25/2022 08:55	52,901,857
sample_2014.zip	07/25/2022 08:55	42,695,344
sample_2015.zip	07/25/2022 08:55	42,378,068
sample_2016.zip	07/25/2022 08:55	39,619,409
sample_2017.zip	07/25/2022 08:55	32,251,237
sample_2018.zip	07/25/2022 08:55	22,713,321
sample_2019.zip	07/25/2022 08:55	17,402,152
sample_2020.zip	07/25/2022 08:55	14,015,708
sample_2021.zip	07/25/2022 08:55	6,515,637

2.1 [10%] Read the file

- Notice that the file is a `txt` file with columns separated by the pipe (`|`) and that there are no column headers (which means we will have to input the column names manually)
- One option to read the file is via the menu `Environment` -> `Import Dataset` ->`From Text (readr)` -> `Browse` -> select the origination file; in the window that appears next select `Other` for `Delimiter` and type `|`, untick `First Row as Names` and finally `Import`
    - Once you import the file you will see in the console the command that `RStudio` executed to import the dataset. Copy the command to your markdown document:

```
setwd("/users/smanzan/Dropbox (Personal)/MyFolders/Teaching/NYU/Econometrics/myassignments/")
library(readr)
sample_orig_2020 <- read_delim("./Data/sample_orig_2020.txt",
     delim = "|", escape_double = FALSE, col_names = FALSE, trim_ws = TRUE, )
```

- The last thing to do is to give a name to the columns. This can be done by creating a vector of column names and assigning those to the columns of the data frame `sample_orig_year`. Copy the vector `orig.names` below and paste in your markdown document:

```
# see "User Guide" and "File Layout" in Resources
orig.names <- c("Credit_Score","First_Payment _ate","First_Time_Homebuyer", "Maturity_Date",
                "MSA","Mortgage_Insurance_Percentage","Number_Units","Occupancy_Status","CLTV",
                "DTI","UPB","LTV","Interest_Rate","Channel","Prepayment_Penalty",
                "Amortization_Type","State","Property_Type","Postal_Code","Sequence_Number",
                "Purpose","Loan_Term","Number_Borrowers","Seller_Name","Servicer_Name",
                "Super_Conforming","Pre-HARP_Loan","Program_Indicator","HARP_Indicator",
                "Valuation_Method","Interest_Only")

colnames(sample_orig_2020) <- orig.names
```

- An alternative is to name the column names while reading the data via the argument `col_names`:

```
sample_orig_2020 <- read_delim("./Data/sample_orig_2020.txt",
     delim = "|", escape_double = FALSE, col_names = orig.names, trim_ws = TRUE)
```

3. **Statistical analysis and plotting** (using `ggplot2`)

3.1 [10%] Plot a histogram of *credit score* and the *interest rate*

- Are there outliers in the distribution of these two variables?
- Freddie Mac uses the number `9999` for missing values: is it possible that the outliers might be such values? If yes, eliminate the rows of the data frame corresponding to these values; for example, filter the rows that are not equal to `9999` as follows

```
library(dplyr)
sample_orig_2020 = filter(sample_orig_2020, Credit_Score != 9999)
```

- discuss the distributional characteristics of the two variables (use the filtered dataset if you had outliers/missing values)

3.2 [10%] Calculate the mean, standard deviation, skewness, and kurtosis of the `Credit_Score` variable and discuss the results

3.3 [10%] Do a scatter plot of the *credit score* and *interest rate* variables and discuss whether the plot shows any dependence between the two variables

4.1 [10%] Estimate the coefficients of the regression model $\text{int\_rate}_i = \beta_0 + \beta_1 * \text{credit\_score}_i + u_i$ using the formulas derived in class for the OLS estimators.

4.2 [10%] Provide an interpretation of the estimate of the slope coefficient and discuss whether the sign of the coefficient is consistent with economic reasoning

4.3 [10%] Estimate the regression model using the `lm()` command. Interpret the R^2 and SER of the regression

File to analyze

Name	file
Saniya	sample_2008.zip
Makhambet	sample_2010.zip
Karthik	sample_2004.zip
James	sample_2003.zip
Christine	sample_2005.zip
Emmanuel	sample_2008.zip
Patricio	sample_2007.zip
Gerry	sample_2007.zip
Mario	sample_2005.zip
Albert	sample_2007.zip
David	sample_2009.zip
Valentin	sample_2007.zip
Hamza	sample_2014.zip
Bin	sample_2006.zip
Eugene	sample_2013.zip
Xin	sample_2017.zip
Antonio	sample_2001.zip
Siddh	sample_2014.zip
Ashley	sample_2010.zip
Zharmakhan	sample_2006.zip
Deven	sample_2005.zip
Thiago	sample_2006.zip
Catherine	sample_2003.zip
Shuyi	sample_2013.zip
Catherine	sample_2009.zip
Kevin	sample_2012.zip
Phil	sample_2006.zip
qiyue	sample_2002.zip
Tony	sample_2018.zip
Jissa	sample_2003.zip
Jieyi	sample_2013.zip
Sandy	sample_2007.zip
Siteng	sample_2001.zip
Jessica	sample_2005.zip
Elaine	sample_2019.zip
Tianchen	sample_2011.zip
Catherine	sample_2017.zip
Vera	sample_2015.zip

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