

ETF5952 Quantitative Methods for Risk Analysis

Semester 1, 2021

ASSIGNMENT 2

Important Instruction

- This assignment comprises 25% of the assessment for ETF5952. This is an individual, NOT a syndicate, assignment. On the Assignment Cover Sheet, read the references to plagiarism and collusion from University Statute 4.1. Part III-Academic Misconduct.
- Answer all questions, and start from a new page for each question. Your assignment must be typed and you must submit a pdf file (A4 pages) with an Assignment Cover Sheet (from the ASSIGNMENTS section of Moodle). To later confirm your upload was successful, go to the “ASSIGNMENTS” section and click. On the “Assignment 2” uploading link. The uploaded file’s name will be shown.
- If you have a valid reason not to meet the deadline, you will be requested to submit what you have done at the due date and receive your grade relative to opportunity. Without any valid reasons, 10% of Assignments allocated marks will be deducted for each day that it is late.
- Submit one pdf file only. Do NOT submit/attach R scripts or output files. Do not submit your assignment in a folder.
- You should summarize what you obtain to answer questions, instead of providing all codes and outputs. If you provide too many outputs relative to questions, then we will consider that you may not understand the questions and your answers would be subject point deduction.
- If you have questions regarding materials, you are encouraged to use our consultation. The course email should be used only for pointing out typos and personal matters.

Question 1 (20 points: 5 points \times 4)

In this question, you analyze two data sets provided by AirBnB¹, `airbnb20.csv` and `airbnb21.csv`, which contain information on properties in Melbourne in Feb 2020 and 2021, respectively. The data set contains variables: rent price (*price*), the number of bedrooms (*bedrooms*), the number of beds (*beds*), room type (*room_type*), and customers' review score (*review_scores_rating*).

To answer the questions below, you report figure separately for each data set.

1. Suppose that you use the data, `airbnb20`, to estimate a regression model, in which the dependent variable is price and the rest of variables can be used for regressors. As regression specifications, you could consider regressors and their pair-wise interactions. If you have to select a model, how many models you have to estimate?²
2. Set the seed as “2021”. Apply lasso for estimating linear regression models, using the two data sets separately. Report the plot of penalty term and mean square errors from each regression.
3. For each data set, report the estimation results from lasso with penalty terms selected cross-validation. Explain the effect of the number of bedrooms on price (no more 30 words for each result, & round numbers to two decimal places.) Here, you answer only the effect of bedrooms and ignore interaction terms.
4. For each data set, report the estimation results from lasso with penalty terms selected by AICc. Explain the effect of the number of bedrooms on price (no more 30 words for each result, & round numbers to two decimal places.) Here, you answer only the effect of bedrooms and ignore interaction terms.

Question 2 (20 points: 5 points \times 4)

For this question, use a data set, “`ASX.csv`”, which includes daily time series of S&P/ASX 200. As we did in Assignment 1, change “date” variable in the date format and crate a new data set (data frame) for daily return of *price*. We are interested in estimating $AR(p)$ models for $p = 1, \dots, 5$.

1. Create the data set that includes return, its lagged returns up to 5 lags and date and make sure no missing values. More specifically, let $\{r_t\}_{t=1}^T$ be the original data set of returns with the sample size T . Then, the lagged variables for r_t are r_{t-1}, \dots, r_{t-5} . Since time period t in the original data set starts from 1, the new data should starts from $t = 6$ to avoid missing values.

Report summary statistics and head of the new data set (use *summary* and *head*).

2. Estimate five models, $AR(1), \dots, AR(5)$, and report the estimation result.
3. Report values of AIC or BIC for the five models and the best model for each criterion.
4. Report time series plot of return and fitted values from the model selected by AIC.

Question 3 (30 points: 15 points \times 2)

For this question, use a data set, `credit.csv`, on credit card information. See Assignment 1 for explanation of variables included in the data set.

We are interesting in determinants of cigarette consumption. In the following analysis, use all variables except “cigs” as regressors.

1. Estimate the classification tree in which the dependent variable is rating and regressors are Income, Cards, Education, Student, Balance, and Age.
 - Report a plot of the estimation result (use “type=0”).
 - Interpret three numbers at the far-right node and the far-left node (no more than 30 words for each node).

¹Data source: <http://insideairbnb.com/get-the-data.html>

²Here, you do not need to estimate any models. You answer the number of models you consider. For instance, if you have two regressors, you have 4 possible combinations of regressors.

- What is a prediction from the estimated model for an individual with balance = 490, income = 22, and non-student.
2. Estimate the classification tree in which the dependent variable is rating and regressors are Income, Cards, Education, Student, Balance, and Age.
 - Report a plot of the estimation result (use “type=0”).
 - Interpret two numbers at the far-right node and the far-left node (no more than 30 words for each node).
 - What is a prediction from the estimated model for an individual with balance = 990, income = 22, and non-student.

Question 4 (30 points: 5+5+10+10 points)

In this question, we analyze the data used in Kuka et al (2020)³ to quantify the policy effect of the Deferred Action for Childhood Arrivals (DACA) program, which provides temporary work authorization and deferral from deportation for undocumented, high-school-educated youth. The data set, DACA.csv, contains the following variables for individual observations:

- “hs”: a dummy for high-school degree or not (taking 1 for yes or 0 for no)
- “elig_post”: a dummy variable taking 1 if an individual is eligible for DACA and after 2012 or 0 otherwise.
- “year”: year
- “statefp”: states in US
- “age”: age
- “yrimmig”: year of immigration
- “fem”: a dummy for female (taking 1 for female or 0 otherwise)
- “ageimmig”: age of immigration
- “noncit”: a dummy for non-US citizen
- “elig”: a dummy for being in eligible groups
- “bpl”: birth place
- “race”: race

Enacted in August 2012, DACA extended temporary relief from deportation and work authorization two years, initially, subject to renewal to undocumented youth who were in school or had completed high school, and met other criteria based on age and year of arrival. DACA thus generates a discrete increase in the benefits associated with completing high school. Using the cross-sectional and time series variations, we evaluate the effect of DACA program on high-school graduation (“hs”) under the difference-in-difference (DID) framework.

1. Under the DID framework, we need to control for group and time effects. To this end, we use a dummy variable “elig” and use dummy variables for year. Regress “hs” on “elig_post”, “elig” and year dummy variables (notice that if R recognize year as categorical variables, then regression functions include dummy variables for the categorical variable.) Report the result and interpret the effect of DACA on high-school graduation.
2. We need to include control variables to avoid the issue of confounding variables. Under the DID framework considered at the previous questions, report the estimated policy effect (only) and interpret the effect of DACA on high-school graduation (no more than 30 words, round numbers to two decimal places), after including the following regressors:

³“Do Human Capital Decisions Respond to the Returns to Education? Evidence from DACA”, American Economic Journal: Economic Policy 2020, 12(1): 293324. You can find the paper at Moodle.

- female dummy, race dummy, dummy for birth place, interaction between race and year dummy, interaction between state dummy and year dummy, and year of immigration

Notice that if R recognizes birth place, race and state as factor (categorical) variables, then regression functions will include dummy variables for each categories.

3. First, set seed as “1234” and then apply Lasso (single machine learning) for the regression model at the previous question. Report the estimated policy effect (only) and interpret the effect of DACA on high-school graduation (no more than 30 words, round numbers to two decimal places). For the selection of the tuning parameter, use the function *coef* with an option of *select*=”min”.
4. First, set seed as “1234” and then apply double machine learning for the regression model. Report the estimated policy effect (only) and interpret the effect of DACA on high-school graduation (no more than 30 words, round numbers to two decimal places). For the selection of the tuning parameter, use the function *coef* with an option of *select*=”min”.