Graded Assignment 1

FIN 403: Econometrics

October 11, 2023

1 General information

- This is the first of the two practical assignments that we will have this semester.
- It is due at noon (12:00) on October 24 — no late submissions will be accepted.
- As described in the course syllabus, you can work on this assignment in teams of up to three students, with the teams remaining fixed across the two assignments. Please enter your team composition on Moodle ("Group Choice for Graded Assignments"). If you are looking for a team, please use the Ed Discussion forum or let us know. Do not share code or solutions with other students outside your team.
- Please submit through Moodle (one submission per team – make sure that all your names and SCIPER numbers are included with your submission). Your submission should consist of ONE zip file containing (i) a concise write-up of your solutions ($\leq$ 3 pages of text, plus tables/figures; don't send full log files), and (ii) the code(s) used to generate the solutions. As discussed in the first lecture, we strongly recommend to use Stata or R.
- Grading: there will be a total of 100 points, distributed as follows:
  – 70 points: Content of the write-up — whether the presented solutions & explanations are correct. Each of the 10 questions below accounts for 7 points.
  – 30 points: Presentation and coding style — whether output from your analysis is presented clearly (e.g. as discussed for the regression tables in Chapter 3 of the lectures) and whether the submitted code is well commented and easy to follow.

2 Task

We will use a dataset on residential real estate transactions from the US. Our goal is to understand how the selling price of a property is related to the property's characteristics (for example, we expect that larger properties will tend to be more expensive). Our version of the dataset contains the following variables:

- *pid* - Parcel Identification Number (a property identifier)
- *saleprice* - Sale (Transaction) Price
- *fullbath* - Number of Bathrooms
- *grlivarea* - General Living Area (in square feet, ft^2)
- *centralair* - Central Air Conditioning (Yes/No)
- *yearbuilt* - Year of Construction
- *overallqual* - Overall Quality of the Property (1-10)
- *salecondition* - Type of Transaction (categorical). Categories are: *Normal*, *Abnormal* (not a market transaction), *Family* (within family)

Note 1 : for all statistical tests, please use a 5% significance level.

Note 2: in your write-up, you should be able to show all the regression results in a single table – there should be four estimated models, for items (c), (d), (f), (h).

(a) Load the dataset and inspect the data. Each *pid* should appear only once – is that the case? Write code to check for duplicate observations. If you find any, delete them from the dataset.

(b) Transform the categorical variable *centralair* into a numerical dummy variable. Prepare a nice table with summary statistics for all variables. Does anything stand out?

(c) Estimate the following simple model with OLS:

$$saleprice_i = \beta_1 + \beta_2 grlivarea_i + \varepsilon_i \tag{1}$$

Interpret the estimated coefficient b_2. Then, illustrate the model graphically—that is, prepare a scatter plot of *saleprice* against *grlivarea* and add a fitted regression line to the plot.[1] Discuss what you see – how do you think the observations with very large living areas (above 4,500ft^2) and low prices affect the estimated coefficient? Are these high-leverage observations? Are they influential observations?

Check the *salecondition* of these observations. Could this explain the low price? From now on, drop observations with *salecondition* other than *Normal*.

(d) One way of dealing with right-skewed distributions/outliers is to use a log-transformation of the variable. Plot a histogram of $ln(saleprice)$ and compare it to the histogram of *saleprice*. Next, estimate model (1) using $ln(saleprice)$ as the dependent variable. What is the predicted price of house with 2500ft^2 surface?

(e) Calculate the residuals for the model in levels (from item c) and in logs (from item d) and plot them against *grlivarea*. Is the homoskedasticity assumption likely to be violated in either of the models?

(f) Estimate the following more complex model:

$$ln(saleprice)_i = \beta_1 + \beta_2 grlivarea_i + \beta_3 yearbuilt_i + \beta_4 fullbath_i + \beta_5 centralair_i + \varepsilon_i \tag{2}$$

How do you interpret the estimated coefficient for the intercept, b_1, in this regression? And how do you interpret the coefficient on $centralair$, b_5?

Test the null hypothesis that having 100ft^2 additional living area has the same effect on log(price) as having been built 9 years later.

(g) Check some of the diagnostics we discussed for (i) collinearity, (ii) functional form (RESET) for model (2). Interpret the output from the diagnostics you perform. What happens to the results of the RESET test if you do not drop transactions with abnormal *salecondition*? (no need to present the output for the comparison, just comment)

(h) Now, add *overallqual* to the model.[2] How does b_2 change? Calculate correlation coefficients between *overallqual* and *ln(saleprice)* and between *overallqual* and *grlivarea*. How do these correlations relate to your observation about the change of b_2?

(i) Generate a dummy *oldhouse* for houses built before 1970, include an interaction between *grlivarea* and *oldhouse* in the model, and drop *yearbuilt* and *overallqual* from the model. What do the estimation results from this model tell us about the relationship between the living area and the price for old houses?

(j) Perform the Chow test to check whether there are differences in the statistical determinants of log salesprices between old houses and 'new' (not old) houses, using *grlivarea*, *fullbath* and *centralair* as covariates. Interpret the results.

[1] Tip: use *twoway lfit Y X* in Stata or *abline(lm(Y=X))* within the *plot* command in R to produce the fitted line.

[2] Treat *overallqual* as a continuous variable in this exercise.