BM01BAM - Advanced Statistics & Programming

Individual Assignment 1:

Analyzing US Housing Prices

The first individual assignment is about applying your newly-gained skills in statistics to explore the factors associated with US housing prices. The *Housing Price Dataset from Kaggle* can be downloaded from `https://www.kaggle.com/c/home-data-for-ml-course/data?select=train.csv`. The assignment consists of five exercises. All parts of the assignment can be made with the material taught in the lectures and the tutorials of this and the previous week. Much of the tutorial code can be re-used. This individual assignment is due on **Friday September 16, 2022, 11:59AM**. [Expected length of your submission is around 5 pages, all inclusive.]

Before starting the assignment, it is suggested to make a designated folder with subfolders: *Data*, for all data sets; *Programs*, for all computer code;, and *Results*, for all results, e.g., figures and tables. This will enhance transparency of the work process, adds to developing hygienic coding skill, and helps us to help you in the case of questions.

Submit your individual answers as a pdf-file on Canvas. The manuscript has the R-script in the appendix (in a not too large font). Alternatively, you may opt to write your assignment using R Markdown (or R Sweave), which integrates the code and your answers. The use of Latex for this submission, as stipulated in the course manual, is highly encouraged (see *this* or *this* link for helpful resources). A decent layout is part of the assessment, as well as a proper motivation of your claims and findings.

Assignment

As newly-minted experts in statistics, you are tasked with the challenge of analyzing the determinants of house prices in the US. The intended results are of high-order importance to various stakeholders, such as real-estate agents, customers, and local governments. In the exercises of this assignment, you

will delineate the business problem and research directions yourself, along the guidelines outlined in each exercise. You will gather and prepare the data yourself that can provide insight into this problem, make assumptions where necessary, apply suitable regression techniques, and in the end provide insights and recommendations in clear, well-written, well-motivated, convincing terms.

1 Collect and prepare data

Download the data from Kaggle. Prepare the data for analysis and make your own selection of variables to analyze, keeping in mind the below exercises. You are allowed to use different variables in different exercises. There is a large number of continuous and discrete variables in the data set, which lend themselves naturally to regression analyses. Convert the categorical variables in your analyses to type *factor*, before usage. Present summary statistics of the data set with the variables of your choice using function `stargazer` (remember, function `stargazer` produces readily-formatted tables for Latex) and present a concise description of the main characteristics of your variables. Use function `ggplot` to make three plots of your choice where in each plot the dependent variable *SalesPrice* is related to another variable. For subsequent reporting, it is advised to use function `ggsave` to export these figures in *pdf* or *png* format. The choice of appropriate plot types, e.g., scatter plot, bar chart, etc., is part of the exercise. [0.5-1 pages]

2 Theoretical model and OLS assumptions

Develop a (mini-)theory to explain variation in the main dependent quantity *SalesPrice*. Present a causal relationship diagram, and formulate three research hypotheses as illustrated in class. Formulate the population regression model, including appropriate specification of categorical variables, interactions and non-linearities (if present). Elaborate on each of the six linear regression model assumptions discussed in class and explain why which assumptions could be violated; use appropriate example variables to explain why you think which assumptions might be violated. [0.5-1 pages]

3 OLS regression and model fit

Estimate the model developed in the preceding exercise, both with and without interaction and non-linear terms. Report your results using tables generated with function `stargazer`, as practiced in the tutorial. Present and interpret the estimation results; an example has been shown in class. Also determine which of the independent variables in your model have the larger effect sizes; motivate the criterion you used. [1-1.5 pages]

4 Diagnostic checking

Perform a diagnostic analysis of the estimated model, i.e., systematically analyze if OLS model assumptions have been violated, and if remedies for these violations seriously affect model outcomes. For instance, check if inference is robust for transformations of the explanatory variables and for heteroskedastic standard errors. Also, analyze the (standardized) residuals to assess whether their behavior is consistent with the assumptions of the linear regression model, and check if multicollinearity problems adversely affect the interpretation of the estimation results. Re-run updated models to determine if any remedies or adjustments have been effective to counter the observed modeling issues; discuss consequences of your model interventions by comparing the results with the originally estimated models. [0.75-1.25 pages]

5 Subset analyses

The last exercise is about the consequences of sub-sample analyses for your findings. Sub-samples can be defined based on existing categorical variables or discretized quantitative variables of your liking. Analyze whether differences exist between the estimation results obtained for your model in the different sub-samples. Examples of questions that might be of interest are the following:

- Does the relationship between *SalesPrice* and the independent variables depend on *SalesCondition*?
- Do very large houses bias your result?
- Does the relationship between *SalesPrice* and the independent variables differ between one-story homes, multi-story homes, and homes without a basement?
- How important is the masonry veneer type?

This list with questions is by no means exhaustive and just serves as inspiration. Also, it is not suggested that you are obliged to limit your sub-sample analyses to the mentioned categorical data.

Run different models to deliver insights into meaningful subsets of your data. Present your results in tables, and concisely discuss your findings. Pay special attention to accurately describing the different groups you analyze and how the results change between your groups. Please present additional figures clarifying the differences you might find. [1-1.5 pages]