Linear Models: Project Proposal

STAT 230A – Linear Models

Huy G. Pham & John V. Manousakis

April 1, 2021

1 Dataset selection

For this project, we propose looking at a dataset containing assessed values for residential properties sold in Ames, Iowa in the late 2000's. Although the Boston housing dataset is commonly used, it is outdated for today's housing market, and contains fewer covariates. By using the Ames data set, updated inference statements can be made about property values, and deeper analyses of variables can be performed. The data set was compiled by prof. Dean DeCock and contains general information about a home's construction and features, such as size, condition, and year built, as well as the target variable: its sale price (DeCock 2011). In general, information such as square footage of the house and its features provide continuous variables to perform linear analysis with; however, the data also contains various categorical classifications, such as roofing material or the presence/absence of features like garages or pools. Non-continuous variables will be assessed to see if the data is usable as binary, or categories/factors that can be used to divide the set into subgroups, then analyzed via methods seen in the course.

2 Inquiries

Overall, we are interested in how each variable affects the final sale price of the homes, which can be seen with ordinary least squares. We are also interested in any significant changes that arise if we assume homoskedasticity, and we propose comparing the EHW variance to identify any differences. We are interested in challenging the assumptions of the EHW variance estimator by comparing its results with a bootstrap estimate for the variance. Moreover, the size of the dataset allows for less significant covariates to be filtered out, and we propose using LASSO as a preliminary method of reducing covariates considered in the fit. Hypothesis testing could also be utilized to infer whether certain covariates have an influence in the sale price or not. We plan to experiment in developing models using transformations of the covariates and the outcome to determine if such models could be more suitable for prediction. Cross-validation can be used to rank the models in terms of accuracy. The dataset is known to contain several outliers, where sale prices do not represent actual market values, and we propose to use leave-one-out methods in order to form an outlier-detecting mechanism and identify such points. Lastly, we propose to analyze the data with respect to sub-groups, say, locations of the neighborhoods, and investigate whether or not there are cases where Simpson's paradox arise using the concept of partial correlation.

References

DeCock, Dean (2011). "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project". In: *Journal of Statistics Education* 19.3, null. DOI: 10.1080/10691898.2011.11889627. eprint: https://doi.org/10.1080/10691898.2011.11889627. URL: https://doi.org/10.1080/10691898.2011.11889627.