Predicting House Prices with a Linear Regression Model

Kevin Thompson, Sterling Beason, & Brandon Croom

Data Science Program, Southern Methodist University, USA

Abstract

Price prediction is pivotal for real estate. Homeowners on the sell-side want to know when to sell, what to renovate, and how much profit they can expect from their efforts. Homebuyers want to know whether they are getting a fair price, where to look for homes in their budget, and the various trade-offs that accompany a purchasing decision. Real estate companies navigate both sides of real estate; hence, they too are a key stakeholder. In the first part of our analysis, we estimate the relationship between house prices, the square footage, and neighborhood location in Ames, Iowa. In the second part of our analysis, we train a linear regression model to predict house prices in Ames, Iowa.

1 Introduction

Price prediction is pivotal for real estate. Homeowners on the sell-side want to know when to sell, what to renovate, and how much profit they can expect from their efforts. Homebuyers want to know whether they are getting a fair price, where to look for homes in their budget, and the various trade-offs that accompany a purchasing decision. Real estate companies navigate both sides of real estate; hence, they too are a key stakeholder. These stakeholders utilize multiple factors related to real estate to determine the fair price for the property. These same factors can be built into a model for price prediction that assists in taking some of the guess work out of property pricing.

The analysis performed for this paper leverages a data set focused on the housing market in Ames, Iowa. This data set is from the "House Price" Kaggle competition (Kaggle (2016)). In this competition Kaggle competitors attempt build models that best predict housing prices based on the Ames, Iowa data set. Over the course of this paper a similar approach to this specific Kaggle competition will be taken. The first part of the paper will focus on the Ames, Iowa dataset, provide the reader with detailed information about the data and any additional variables the team created for model building. Analysis of this data will then be performed in two parts. In the first part of our analysis, we estimate the relationship between house prices, the square footage, and neighborhood location in Ames, Iowa. In the second part of our analysis, we train a linear regression model to predict house prices in Ames, Iowa.

The intent of this paper is to provide the reader with an understanding of how linear regression can be applied to data and the analysis that must be undertaken to ensure a linear regression model is adequately developed.

Ramsey and Schafer (2013)

2 Ames, Iowa Data

The data used for this analysis, described in the sections below, comes from the Kaggle Competition "House Prices: Advanced Regression Techniques" (Kaggle (2016)). The data set for this competition contains housing

related data for Ames, Iowa. The total data set contains 2919 observations and 80 features or variables. Although too numerous to describe here (see the Kaggle website for full descriptors (Kaggle (2016))), these 80 features relate to quantity and quality based attributes of a physical property that may interest any of the key stakeholders (prospective home buyer, home seller, real estate company/agent). For example the data provides answers to questions such as: "How many rooms in the property?", "What is the condition of the kitchen?", "What is the location of the property?", "Is there a basement?". Delving deeper, the data set breaks down into 46 categorical variables and 34 numeric variables.

The categorical variables break down into a relatively equal mix of nominal and ordinal values (23 nominal and 23 ordinal). The ordinal variables indicate a grading of various property related components such as the overall property quality, overall property condition, room specific quality and room specific conditions. The nominal values provide information on various conditions of the property such as building materials used and dwelling type. Cock (2011)

The numeric variables contain both continuous and discrete values (20 continuous and 14 discrete). The continuous variables indicate information a prospective home buyer would like to understand such as lot size, total square footage, and specific square footage for living spaces. The discrete variables provide the prospective home buyer with an understanding of items such as number of bedrooms, number of bathrooms, etc. Cock (2011)

In reviewing the data it was determined that a few variables could be removed for reasons noted below:

- ID - this field is a record ID field and is not informational for analysis
- Pool Quality - this field does not contain enough variation to be useful
- Miscellaneous Feature - this field contains minimial information

Data quality checks were performed across all remaining variables to address missing values, inconsistent variable names, and to ensure consistent ordering of ordinal variables across all variables of the same type. Comparison analysis was also performed to ensure the variable ordering was appropriate, as shown in figures X & X

3 Analysis Question I

3.1 Problem Statement

A local real estate company, Century 21, would like to understand the relationship between living area of a home and sale prices. Specifically, the company would like to focus only on sales in certain neighborhoods (NAmes, Edwards, and BrkSide). An estimate (or estimates) of this information as well as required confidence intervals should be provided, along with verification of the model assumptions and addressing of suspicious observations. In the conclusion quantify the relationship between living area and sale price with respect to these three neighborhoods.

3.2 Assumption Checks

3.3 Competing Model Comparison

3.4 Parameters

3.5 Conclusion

Ramsey and Schafer (2013)

Pearl (2009)

Ruppert and Matteson (2015)

4 Analysis Question II

4.1 Problem Statement

Build a predictive model, leveraging techniques learned in DS 6371 only, to predict sales prices of homes in all of Ames, Iowa. The goal is to produce four models: a forward selection model, a backward selection model, a stepwise selection model and a custom model. Each model should have an adjusted R^2, CV Press and Kaggle Score. In the conclusion describe which model is best at predicting future sale prices of homes in Ames, Iowa.

4.2 Assumption Checks

4.3 Competing Model Comparison

4.4 Parameters

4.5 Conclusion

Hastie et al. (2009) Trefethen and Bau (1997)

5 Appendix

References

Cock, D. (2011). Ames, iowa: Alternative to the boston housing data as an end of semester regression projects. reference article on Ames, Iowa data set http://www.amstat.org/publications/jse/v19n3/decock.pdf.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Kaggle (2016). House prices. data retrieved from the Kaggle website, https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Ramsey, F. and Schafer, D. (2013). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Brooks/Cole Publishing Company.

Ruppert, D. and Matteson, D. (2015). *Statistics and Data Analysis for Financial Engineering.* Springer.

Trefethen, L. and Bau, D. (1997). *Numerical Linear Algebra.* Society for Industrial and Applied Mathematics.