Regression Analysis of the Ames, Iowa Dataset

Stuart Miller, Paul Adams, and Chance Robinson

Master of Science in Data Science, Southern Methodist University, USA

1 Introduction

Ramsey and Schafer (3)

2 Ames, Iowa Data Set

The Ames, Iowa Data Set describes the sale of individual residential properities from 2006-2010 in Ames, Iowa (1). The data was retreved from the dataset hosting site Kaggle, where is it listed under a machine learning competition named *House Prices: Advanced Regression Techniques* (2). The data is comprized of 37 numeric features, 43 non-numeric features and an obervation index split between a training set and a testing set, which contain 1460 and 1459 observations, respectively. The response variable (`SalePrice`) is only provided for the training set. The output of a model on the test set can be submitted to the Kaggle competition for scoring the performance of the model in terms of RMSE. The first analysis models property sale prices (`SalePrice`) as the response of living room area (`GrLivArea`) of the property and neighborhood (`Neighborhood`) where it is located. **Add some details on the question 2 variables?**

3 Analysis Question I

3.1 Question of Interest

Restatement of the problem

3.2 Modeling

TODO: Build and fit the model

We will consider two models: (1) the logarithm of sale price as the response of living room area and (2) the logarithm of sale price as the response of living room area accounting for differences in the three neighborhood of interest (Brookside, Northwest Ames, and Edwards) where Edwards will be used as the reference.

Reduced Model

$$\mu\{log(SalePrice)\} = \beta_0 + \beta_1(LivingRoomArea) \tag{1}$$

Full Model

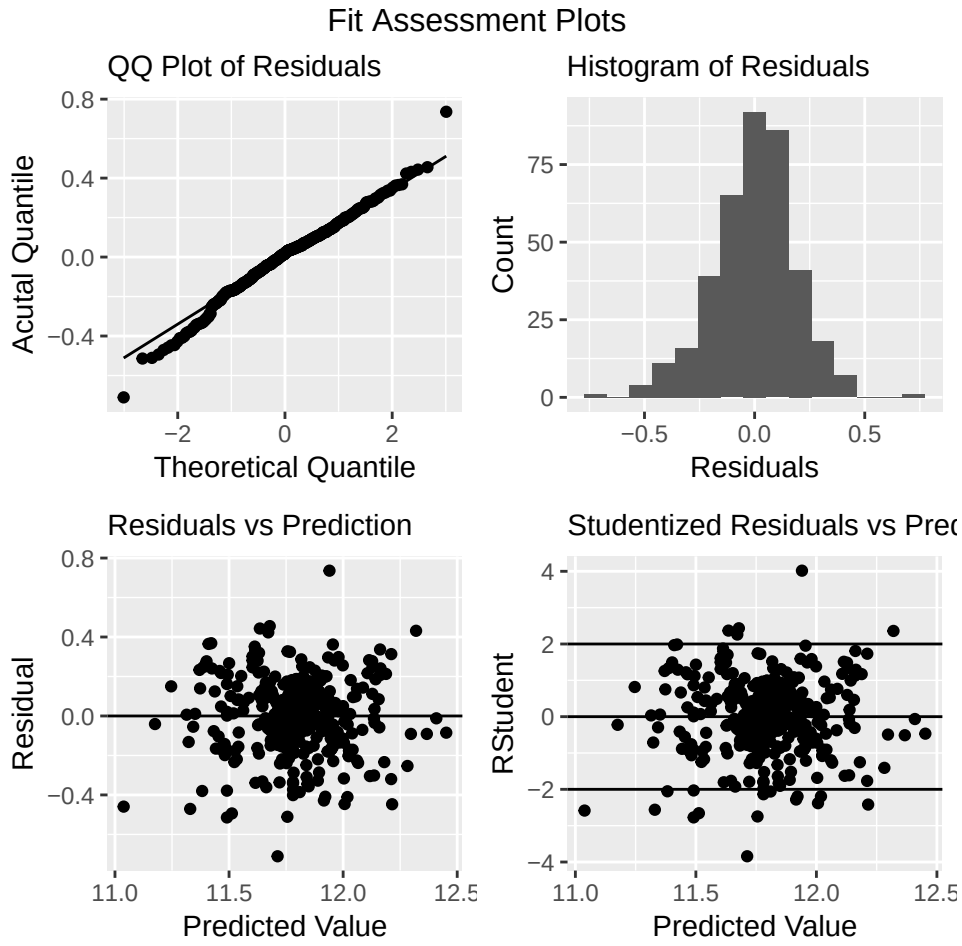
$$\mu\{log(SalePrice)\} = \beta_0 + \beta_1(LivingRoomArea) + \beta_2(Brookside) + \beta_3(NorthwestAmes) + \beta_3(Brookside)(LivingRoomArea) + \beta_4(NorthwestAmes)(LivingRoomArea) \quad (2)$$

We will use an extra sums of square test to verify that the interaction terms are useful for the model. The ESS test provides convincing evidence that the interaction terms are useful for the model (p-value < 0.0001); thus, we will continue with the full model.

```
## Analysis of Variance Table
##
## Model 1: log(SalePrice) ~ (GrLivArea) + Neighborhood_BrkSide + Neighborhood_NAmes
## Model 2: log(SalePrice) ~ (GrLivArea) + Neighborhood_BrkSide + Neighborhood_NAmes +
##     (GrLivArea) * Neighborhood_BrkSide + (GrLivArea) * Neighborhood_NAmes
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    377 14.824
## 2    375 13.441  2    1.3834 19.299 1.053e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3 Model Assumption Assessment

Address each assumption



3.4 Comparing Competing Models

- Adj R^2
- CV Press

Model	Adj R^2	CV PRESS
Reduced Model	0.5	2000
Full Model	0.7	1500

RMSE	CV.Press	Adjused.R.Squared
0.1910566	12.51675	0.5084024

3.5 Parameters

- Estimates
- Influential points
- Residual plots

3.6 Conclusion

A short summary of the analysis

4 Analysis Question II

4.1 Question of Interest

Restatement of the problem

4.2 Modeling

Type of selection

4.3 Model Assumption Assessment

Address each assumption

4.4 Comparing Competing Models

- Adj R^2
- CV Press
- Kaggle score

4.5 Conclusion

A short summary of the analysis

5 Appendix

Include "well commented" `code` in the appendex!

References

[1] Cock, D. D. (2011). Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3).

[2] Kaggle (2016). Ames housing dataset. Data retrieved from the Kaggle website, https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

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