

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 properties from 2006-2010 in Ames, Iowa (1). The data was retrieved from the dataset hosting site Kaggle, where it is listed under a machine learning competition named *House Prices: Advanced Regression Techniques* (2). The data is comprised of 37 numeric features, 43 non-numeric features and an observation 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

Century 21 has commissioned an analysis of this data to determine how the sale price of property is related to living room area of the property in the Edwards, Northwest Ames, and Brookside neighborhoods of Ames, IA.

3.2 Modeling

Linear regression will be used to model sale price as a response of the living room area. From the initial exploratory data analysis, it was determined that sale prices should be log transformed to meet the model assumptions **Add appendix reference**. Additionally, two observations were removed as they appeared to be from a different population than the other observations in the dataset **Add appendix reference**.

We will consider two models: the logarithm of sale price as the response of living room area (1), the reduced model, and 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 (2), the full model. An extra sums of square (ESS) test will be used to verify that the addition of *Neighborhood* improves the model.

Reduced Model

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

Full Model

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

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 Assumptions Assessment

The following assessments for model assumptions are made based on Figure 1 and Figure 2 (**TODO**):

- The residuals of the model appear to be approximately normally distributed based on the QQ plot of the residuals and histogram of the residuals, suggesting the assumption of normality is met.
- No patterns are evident in the scatter plots of residuals and studentized residuals vs predicted value, suggesting the assumption of constant variance is met.
- While some observations appear to be influential and have high leverage, removing these observations does not have a significant impact on the result of the model fit.
- Based on the scatter plot of the log transform of **SalePrice** vs **GrLivArea**, it appears that a linear model is reasonable.

The sampling procedure is not known. We will assume the independence assumption is met.

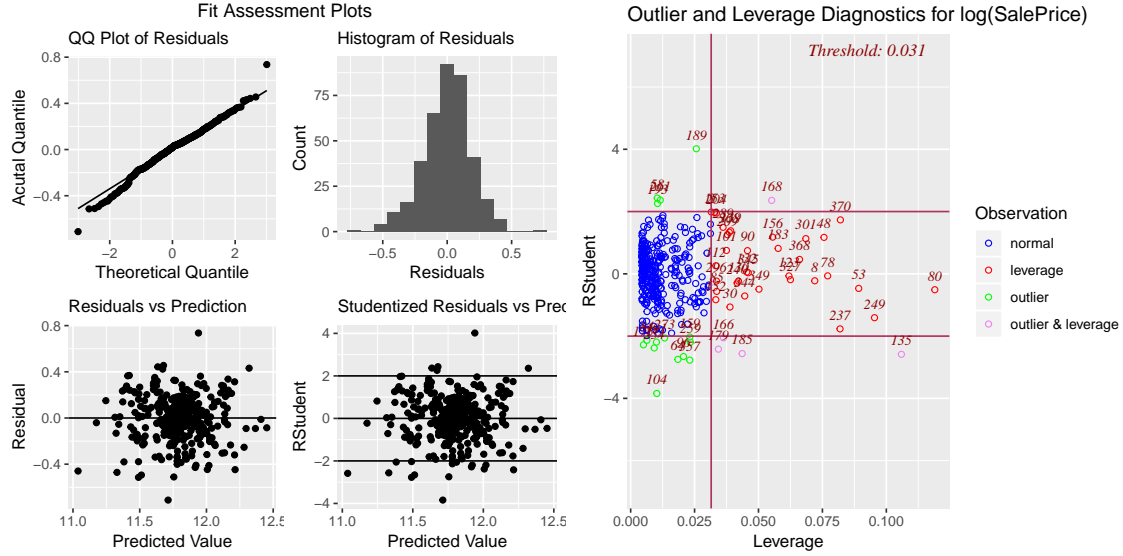


Figure 1: Diagnostic Plots

3.4 Comparing Competing Models

The two models were trained and validated on the training dataset using 10-fold cross validation. The table below summarizes the performance of the models with RMSE, adjusted R^2 , and PRESS. These results show that the full model is an improvement over the reduced model, which is consistent with the result of the ESS test.

Model	RMSE	CV.Press	Adjused.R.Squared
Full Model	0.1910566	12.51675	0.5084024
Reduced Model	0.1988473	13.55835	0.4750767

3.5 Parameters

- Estimates **TODO**
- CI

3.6 Model Interpretation

We estimate that for increase in 100 sq. ft., there is associated multiplicative increase in median price of

- 1.055 for the Edwards neighborhood with a 95% confidence interval of [1.044 , 1.066]
- 1.033 for the Northwest Ames neighborhood with a 95% confidence interval of [1.026 , 1.040]
- 1.077 for the Brookside neighborhood with a 95% confidence interval of [1.063 , 1.090]

Since the sampling procedure is not known and this is an observational study, the results only apply to this data.

3.7 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 appendix!

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.