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

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 intial 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**; therefore, analysis only considers properties with living rooms less than 3500 sq. ft. in area.

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

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

Full Model

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

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:

- The residuals of the model appear to be approximately normally distrubited 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.

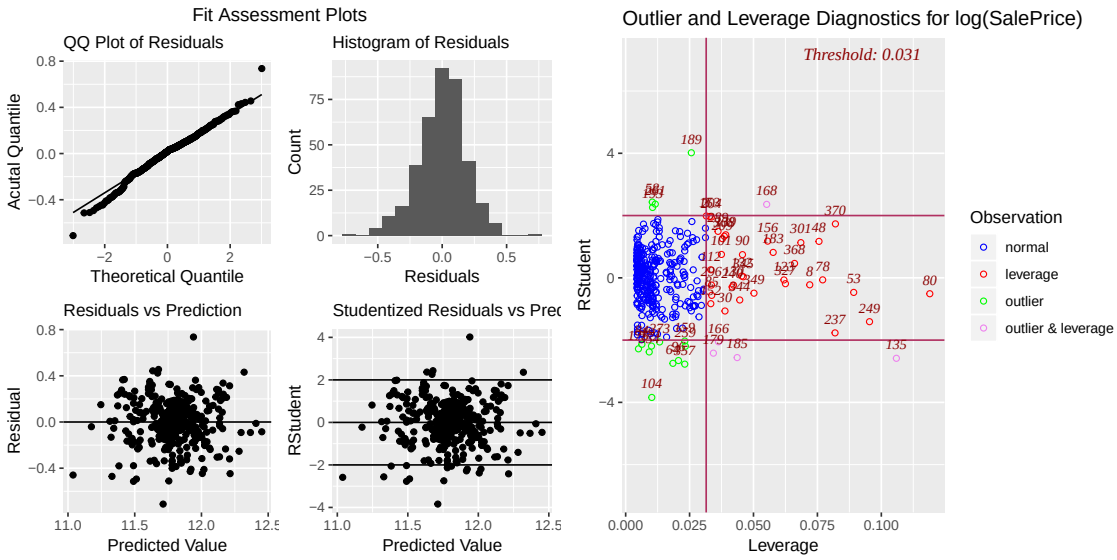

Figure 1: Diagnostic Plots

3.4 Comparing Competing Models

The two models were trained and validated on the training dataset using 10-fold cross validation. The table below summerizes 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

The following table summerizes the parameter estimates for the full model.

Parameter	Estimate	CI.Lower	CI.Upper
Intercept	11.0254845	10.8861855	11.1647836
GrLivArea	0.0005387	0.0004324	11.1647836
Neighborhood_BrkSide	-0.2338906	-0.4468114	-0.0209698
Neighborhood_NAmes	0.4178562	0.2558923	0.5798200
GrLivArea:Neighborhood_BrkSide	0.0001996	0.0000336	0.0003656
GrLivArea:Neighborhood_NAmes	-0.0002145	-0.0003366	-0.0000924

Where `Intercept` is β_0, `GrLivArea` is β_1, `Neighborhood_BrkSide` is β_2, `Neighborhood_NAmes` is β_3, `GrLivArea:Neighborhood_BrkSide` is β_4, and `GrLivArea:Neighborhood_NAmes` is β_5

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 neighorhood with a 95% confidence interval of [1.026 , 1.040]
- 1.077 for the Brookside neighorhood 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

In response to the analysis commissioned by Century 21, the log transform of property sale price was modeled as a linear response to the property living room area for residential properties in Ames, IA. It was determined that it was necessary to include interaction terms to allow for the influence of neighborhood on sale price. Based on the model, there is strong evidence of an associated multiplicative increase in median sale price for an increase in living room area (p-vlue < 0.0001, overall F-test).

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

5.1 Checking for Linearity

The images below show the scatter plots of log sale price vs living room area (Figure 2). In the image on the right, the scatter plot is shown for each neighborhood. In the image on the left the observations for all three neighborhoods are included. In all cases, a linear model appears to be reasonable to model this data.

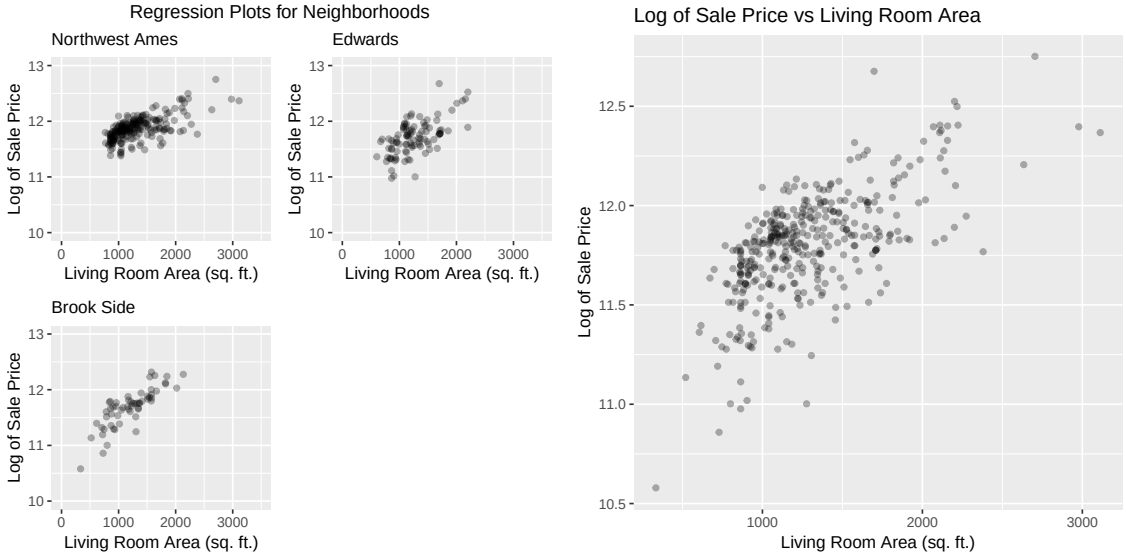

Figure 2: Scatter Plots of Log of Sale Price vs Living Room Area

5.2 Analysis of Influential points

Some text Figure 3 **Add some disucssion on this influential points here**

5.3 R Code For Analysis 1

```
### Compuational Setup
# libraries
library(knitr)
library(kableExtra)
library(tidyverse)
library(olsrr)
library(gridExtra)
library(caret)
library(multcomp)

# set a random seed for repodicibility
```

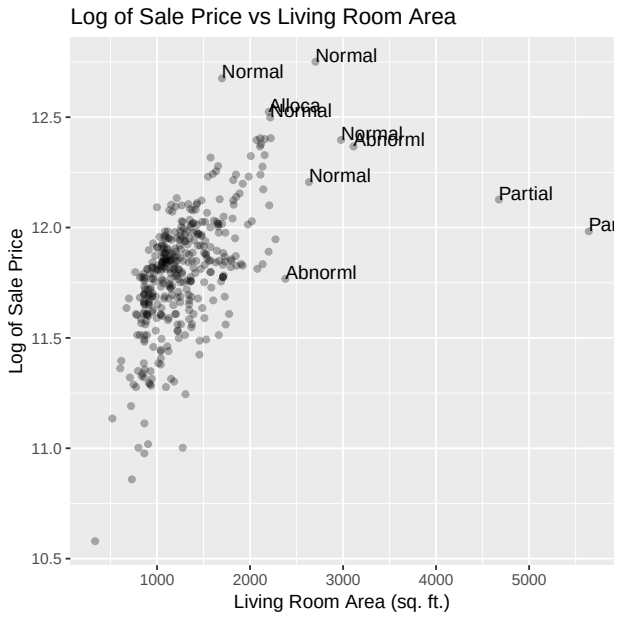

Figure 3: Influential Points

```
set.seed(123)
getwd()
# helper code
source('./helper/visual.R')
source('./helper/data_munging.R')
source('./helper/performance.R')

# load data
train <- read_csv('./data/train.csv')
test <- read_csv('./data/test.csv')

train %>% ggplot(aes(x = (GrLivArea), y = log(SalePrice))) +
  geom_point(alpha = 0.3) +
  labs(title = 'Log of Sale Price vs Living Room Area',
       y = 'Log of Sale Price', x = 'Living Room Area') +
  geom_text(aes(label = ifelse((log(GrLivArea) > 7.75 & log(SalePrice) > 11) |
                                 (log(SalePrice) > 12.45),
                               SaleCondition, '')), hjust=0, vjust=0)
```

5.4 R Code For Analysis 2

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

```
train %>% ggplot(aes(x = (GrLivArea), y = log(SalePrice))) +
  geom_point(alpha = 0.3) +
  labs(title = 'Log of Sale Price vs Living Room Area',
       y = 'Log of Sale Price', x = 'Living Room Area') +
  geom_text(aes(label = ifelse((log(GrLivArea) > 7.75 & log(SalePrice) > 11) |
                                 (log(SalePrice) > 12.45),
                               SaleCondition, '')), hjust=0, vjust=0)
```

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.