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

What is the price of a home in Ames, Iowa? Our inaugural project for the Statistical Foundations for Data Science course in the Southern Methodist University Master of Science in Data Science (MSDS) program was to compete in an online Kaggle competition utilizing linear regression techniques we've learned in this course to date. Our team elected to use R as the preferred analysis platform under the consensus that it has more broad applicability for use in industry, including data gathering and wrangling, in addition to advanced visualization and analytic tools. The project objective was to apply various predictive models in order to assess the suitability of our parameter selections for determining the sales prices of homes in Ames. The measures of accuracy were applied in terms of the Root Mean Square Error, or RMSE, as well as other comparison models such as cross-validation, and the adjusted R-squared. Our approach outlined in this research document is limited in that we were not permitted to use more advanced algorithms we will be exposed to later in the MSDS program; rather, in conjunction with the aforementioned linear regression techniques, we were directed to apply the exploratory data analysis and data cleaning methods we have learned, which will surely be of use to us in our future personal and academic endeavors.

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 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. In the second analysis, variable selection techniques are used to determine which explanatory varaibles are associated with `SalePrice` to find a predictive model.

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 for linearity (see section 6.1), thus improving our models fit and reducing standard error. Additionally, two observations were removed as they appeared to be from a different population than the other observations in the dataset (see section 6.2); 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

$$\mu\{log(SalePrice)\} = \hat{\beta}_0 + \hat{\beta}_1(LivingRoomArea) \tag{1}$$

Full Model

$$\begin{aligned}\mu\{log(SalePrice)\} = \hat{\beta}_0 + \hat{\beta}_1(LivingRoomArea) + \hat{\beta}_2(Brookside) + \hat{\beta}_3(NorthwestAmes) + \\ \hat{\beta}_3(Brookside)(LivingRoomArea) + \hat{\beta}_4(NorthwestAmes)(LivingRoomArea)\end{aligned} \tag{2}$$

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.

```
## [1] "Neighborhood_BrkSide"
## [1] "Neighborhood_NAmes"
```

```
## 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 4:

- 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 (see section 6.1).

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

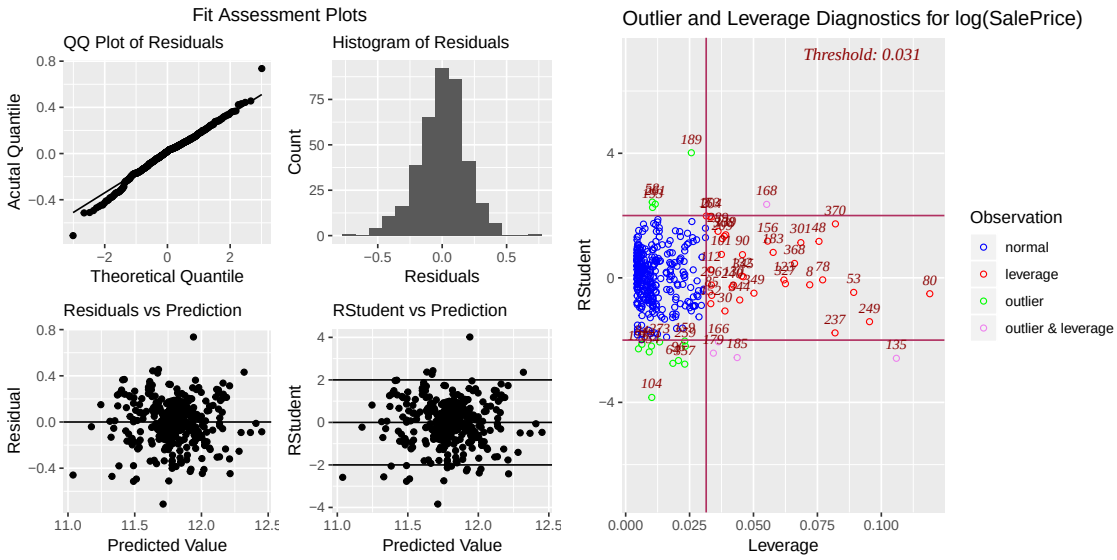


Figure 1: Diagnostic Plots

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

Century 21 has commissioned a second analysis using the same dataset for the creation of a very predictive model of `SalePrice`. The analysis will be expanded to include as many of the 80 total features as required to determine the sale price of residential properties across all neighborhoods of Ames, Iowa, beyond only the three - Edwards, Northwest Ames, and Brookside - previously commissioned for analysis.

4.2 Modeling

Through analyzing our variable selection and cross-validation processes - along with our nascant domain knowledge of residential real estate - we ultimately arrived at a multiple linear regression model featuring 11 linear predictor variables and two interaction terms. Specifically, our variable selection process included direct analysis of a correlation plot and a correlation matrix as well as performing forward selection, backward elimination, and stepwise regression.

Regarding missing data, we imputed NA values for 19 variables using a combination of the data dictionary provided by Century 21 as well as our domain knowledge. After building models with and without transformations applied to variables, we noted no significaznt difference in variable selection from our selection process so elected to use non-transformed predictor variables. We did, however, use the log-transformed `SalePrice` applied in the first analysis.

5 Here

Forward Selection

Forward selection is a variable selection methodology that begins with a constant mean and adds explanatory variables one-by-one until no further additonal predictor variables significantly improve the model's fit. This employess the "F-to-enter" method from the extra-sum-of-squares F-statistic. This was the first method we employed. For this process, we provided the test a starting model with no predictor variables and a model from which terms can be selected, which included all predictor variables available. The process worked forward with selecting one parameter. The suggested model shown in section 6.3.1.

Backward Elimination

Backward elimination is a variable selection methodology that begins with all possible predictor variables and works backward, eliminating variables using all possible combinations until only the best for the fit are provided. This employess the "F-to-remove" method from the extra-sum-of-squares F-statistic. For this process, we provided the test a model with all available predictor variables from which insignificant variables were eliminated. The suggested model shown in section 6.3.2.

Stepwise Regression

Stepwise regression is a variable selection methodology that performs one step of forward selection for each step of backward elimination. The steps are repeated, concurrently, until no further predictor variables can be added or removed. This is the third model approach we used. The suggested model shown in section 6.3.3.

Custom Variable Selection

To develop the custom model, we employed a combination of a correlation matrix for quantitative data, analysis of the summarization of the suggested model from stepwise selection, and through direct analysis of the pairs plots. As previously mentioned, our final model included 11 linear terms and two interaction terms. We removed all variables suggested to be removed by the stepwise regression and backward elimination tests, then reprocessed the updated models until forward selection, backward elimination, and stepwise regression were in agreeance with respect to the linear terms. Once this trial-and-error process was completed, we added interaction terms based on domain knowledge and re-applied the forward selection, backward elimination, and stepwise regression methods until only significant terms - both linear and interactive - remained. We then used graphical analysis to visually confirm interaction between the interactive terms remaining. The custom model shown in section 6.3.4.

5.1 Model Assumption Assessment

The assumption assessment plots were similar for all four models. The assumption assessment plots and discussion for the custom model are provided here with Figure 2. The assumption assessment plots for the other three models are provided for reference in section 6.5.

Based on the diagnostic plots below, the custom model appears to reasonably meet the assumptions of linear regression. The standardized residuals do not appear to exhibit a discernible pattern, indicating constant variance along the regression, or homoscedasticity. While there are some outliers, this does not appear to be an egregious violation. Based on the QQ plot, there is a small level of deviation on the ends of the distribution of the errors, but for the most part, the errors adhere to normality. The sample size should be sufficient to protect against this non-normality. Based on the standardized residuals vs. leverage plot, only a few values have high leverage and are outlying. However, these violations do not appear to be egregious.

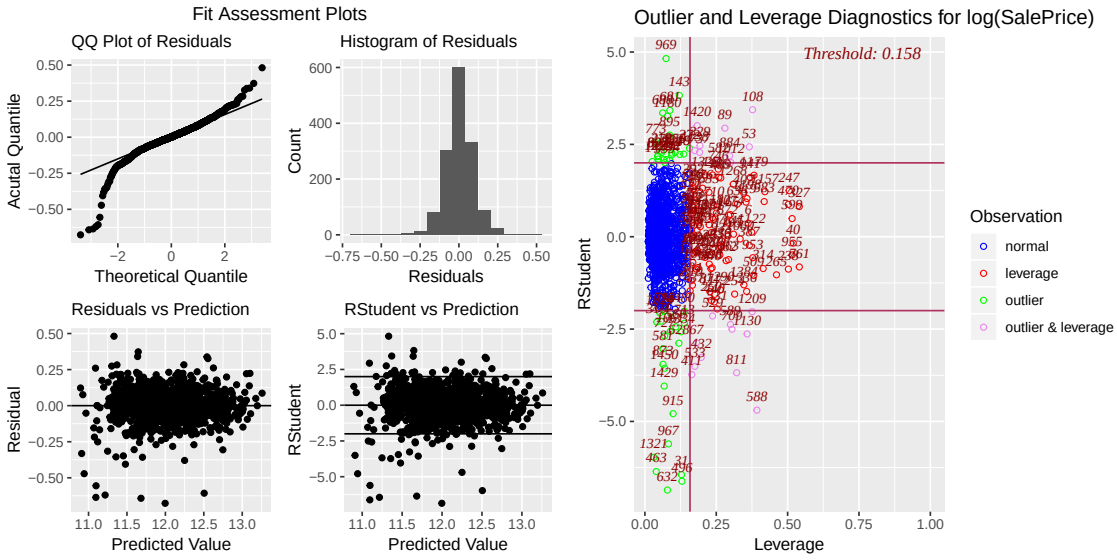


Figure 2: Custom Assumption Assessment Plots

5.2 Comparing Competing Models

While the models from forward, backward, and stepwise selection produce higher adjusted R^2 values on the training data, the yield much higher errors when applied to the Kaggle test set. These selection methods appear to be overfitting to the training data, thus fail to generalize to the Kaggle test set. Undisputedly, the custom model outperformed the model built strictly on the output of the forward selection, backward elimination, and stepwise regression variable selection procedures when applied to a new dataset.

Model	Kaggle.Score	CV.Press	Adjused.R.Squared
Custom	0.13290	17.00007	0.9303195
Forward Selection	0.13476	17.44976	0.9327050
Backward Selection	0.13475	17.23015	0.9327050
Stepwise Regression	0.13476	16.57710	0.9327050

5.3 Conclusion

In an effort to produce a highly accurate and repeatable predictive model using linear regression, all explanatory variables were considered with three types of variable selection techniques: forward selection, backward elimination, and stepwise regression. Additionally, a custom model was initially produced by eliminating variables suggested by the automatic selection processes, visually exploring the data with pairwise scatter plots, and adding interaction terms based on graphical analysis and domain knowledge. Automatic selection was reapplied to suggest terms from the initial custom model, which was then again adjusted for final inspection by the automatic techniques. The final models suggested strictly by the automatic techniques produced high R^2 values, but performed poorly on the Kaggle test set. This suggests the automatic techniques alone were overfitting to the training data. The final custom model, however, produced a high R^2 value and performed remarkably well on the Kaggle test set (see section 6.4). This suggests that the custom model is not overfitting to the training data and generalizes well to an unseen dataset. Ultimately, we determined the best approach is a combination of automatic selection, visual and analytic inspection, and the application of domain knowledge.

6 Appendix

6.1 Checking for Linearity in `SalePrice` vs `GrLivArea`

The scatter plot in Figure 3 shows relationship of `SalePrice` vs `GrLivArea` for all three neighborhoods of interest to Century 21. Based on this plot, it does not appear that this relationship meets the assumptions of linear regression, specifically the constant varaince assumption. The response will be transformed to attempt to handle the changing variance.

Figure 3: Scatter Plot of Sale Price vs Living Room Area

The images below show the scatter plots of log sale price vs living room area (Figure 4). 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.

6.2 Analysis of Influential points

The two outlying observations with living room areas greater than 4000 sq. ft. appear to be from a different distribution than the main dataset. Since these are partial sales, it is possible that the sale prices do not reflect market value. For this reason, we will limit the analysis to properities with less than 3500 sq. ft. (Figure 5)

Figure 4: Scatter Plots of Log of Sale Price vs Living Room Area

Figure 5: Influential Points

6.3 Models Suggested by Automated Selection

6.3.1 Forward Selection

The model suggested by forward selection.

```
log(SalePrice) ~
              OverallQual + GrLivArea + Neighborhood + BsmtFinSF1 +
              OverallCond + YearBuilt + TotalBsmtSF + GarageCars + MSZoning +
              SaleCondition + BldgType + Functional + LotArea + KitchenQual +
              BsmtExposure + CentralAir + Condition1 + ScreenPorch + BsmtFullBath +
              Heating + Fireplaces + YearRemodAdd + Exterior1st + GarageQual +
              WoodDeckSF + SaleType + OpenPorchSF + HeatingQC + LotConfig +
              EnclosedPorch + ExterCond + PoolQC + Foundation + LandSlope +
              RoofMatl + GarageArea + MasVnrType + HalfBath + PoolArea +
              `3SsnPorch` + Street + KitchenAbvGr + GarageCond + FullBath +
              BsmtQual + BsmtFinSF2
```

6.3.2 Backward Selection

```
log(SalePrice) ~
           MSZoning + LotArea + Street + LotConfig + LandSlope +
           Neighborhood + Condition1 + Condition2 + BldgType + OverallQual +
           OverallCond + YearBuilt + YearRemodAdd + RoofStyle + RoofMatl +
           Exterior1st + MasVnrType + ExterCond + Foundation + BsmtQual +
           BsmtCond + BsmtExposure + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
           Heating + HeatingQC + CentralAir + `1stFlrSF` + `2ndFlrSF` +
           LowQualFinSF + BsmtFullBath + FullBath + HalfBath + KitchenAbvGr +
           KitchenQual + Functional + Fireplaces + GarageCars + GarageArea +
           GarageQual + GarageCond + WoodDeckSF + OpenPorchSF + EnclosedPorch +
           `3SsnPorch` + ScreenPorch + PoolArea + PoolQC + SaleType +
           SaleCondition
```

6.3.3 Stepwise Selection

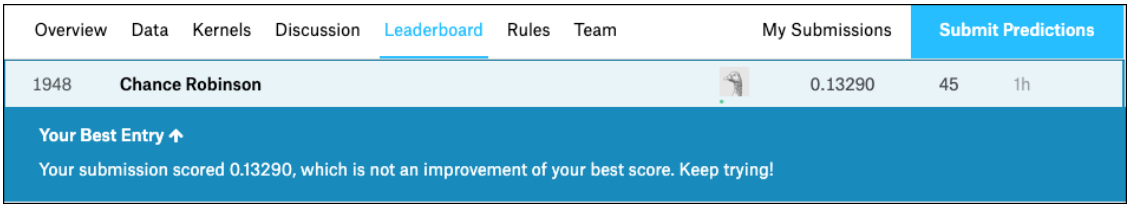
```
log(SalePrice) ~
            MSZoning + LotArea + Street + LotConfig + LandSlope +
            Neighborhood + Condition1 + Condition2 + BldgType + OverallQual +
            OverallCond + YearBuilt + YearRemodAdd + RoofStyle + RoofMatl +
            Exterior1st + MasVnrType + ExterCond + Foundation + BsmtQual +
            BsmtCond + BsmtExposure + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
            Heating + HeatingQC + CentralAir + `1stFlrSF` + `2ndFlrSF` +
            LowQualFinSF + BsmtFullBath + FullBath + HalfBath + KitchenAbvGr +
            KitchenQual + Functional + Fireplaces + GarageCars + GarageArea +
            GarageQual + GarageCond + WoodDeckSF + OpenPorchSF + EnclosedPorch +
            `3SsnPorch` + ScreenPorch + PoolArea + PoolQC + SaleType +
            SaleCondition
```

6.3.4 Custom Model

```
log(SalePrice) ~
             BsmtUnfSF + CentralAir + HalfBath + KitchenQual + Neighborhood +
             OverallCond + OverallQual + RoofMatl + `1stFlrSF` + `2ndFlrSF` +
             YearBuilt + MSZoning:Neighborhood + YearBuilt:Neighborhood
```

6.4 Kaggle Score

The following image shows the result on Kaggle for the custom model.



6.5 Assumption Assessment Plots for Automatic Selection Models

The following discussion applies to the assumption assessment for the three models produced by automatic selection.

Generally, based on the diagonstic plots, these models appear to reasonably meet the assumptions for linear regression. The standardized residuals do not appear to exhibit a discernible pattern, indicating constant variance along the regression, or homoscedasticity. However, there are a small number of observations - relative to the overall sample size - with unusually high residuals. Nonetheless, this is not enough to add detrimental impact to the model. Based on the QQ plot, there is a small level of deviation on the ends of the distribution of the errors, but for the most part, the errors adhere to normality. The sample size should be sufficient to protect against this non-normality. Based on the standardized residuals vs. leverage plot, only a few values have high leverage and are outlying. Compared to the custom model (Figure 2), these diagnostic plots for these models show a few more influential observations with high leverage. However, these observations cannot be excluded from the model.

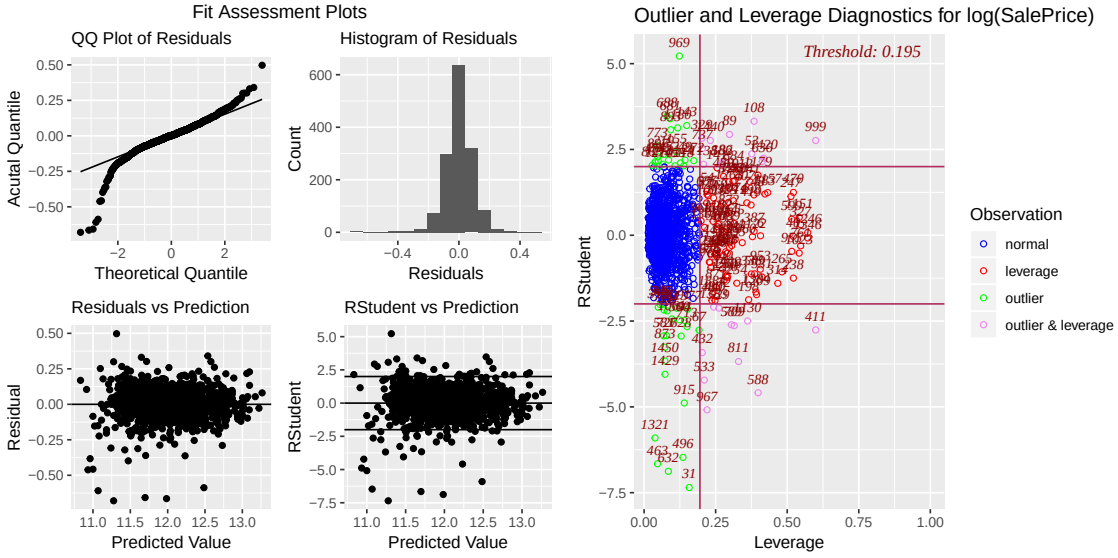


Figure 6: Forward Selection Assumption Assessment Plots

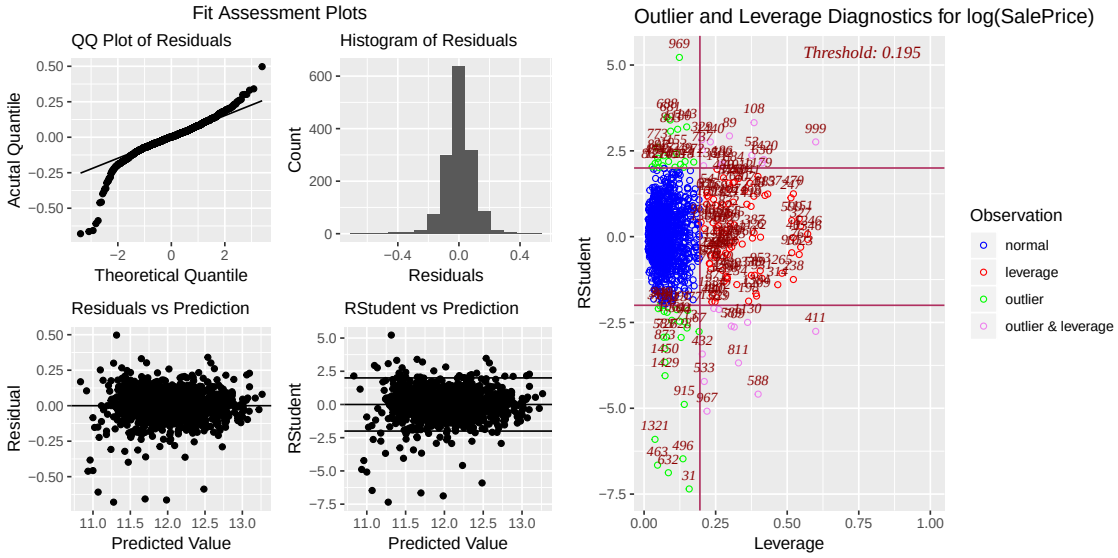


Figure 7: Backward Selection Assumption Assessment Plots

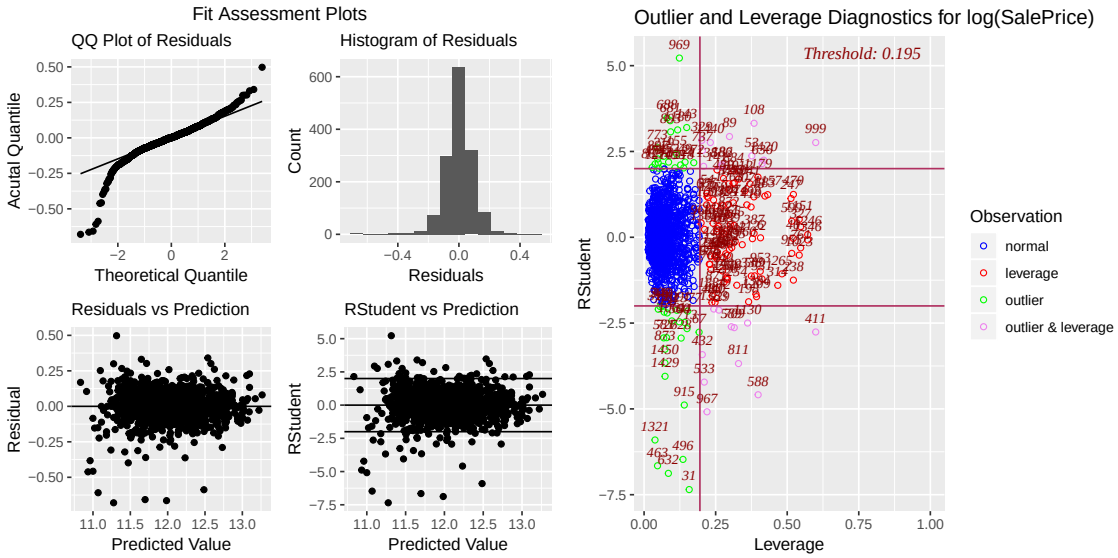


Figure 8: Stepwise Selection Assumption Assessment Plots

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.