Assignment #3

Nate Bitting

Introduction

The objective of this assignment is to first build several simple linear and multiple regression models and perform both model adequacy checking and model validation. We will build these models leveraging the original dataset without removal of any outliers and then rebuild the models by removing up to 200 observations of potential outliers. We will develop the criteria for what is deemed extreme observations versus a normal observations by analyzing the Ames Housing dataset's various features such as total *square footage* and *sale price*.

For model adequacy checking, we will perform several goodness-of-fit and residual analysis procedures including to assess the assumptions of regression analysis:

1. The relationship between the response *y* and the regressors is approximately linear
2. The error term ε has a mean of zero
3. The error term ε has constant variance
4. The errors are uncorrelated
5. The errors are normally distributed

Lastly, we will perform transformations on the response variable, *sale price*, and rebuild the models to compare the goodness-of-fit and residual analyses to see if there is any improvement over the original models. The ultimate goal is to thoroughly assess several models to find the optimal model for predictive accuracy when new observations are ran through the model for predicting *sale price* of other homes in the Ames area.

Results

Simple Linear Regression without any Transformation on the Response Variable

The initial step we took was to develop two simple linear regression models leveraging one regressor and one response variable for each model. For Model 1, we decided to use T*otal Square Footage* for our regressor variable and *Sale Price* as our response variable. For Model 2, we selected *Overall Quality* as the regressor variable and *Sale Price* as the response variable. We will then combine the regressors from Model 1 and Model 2 to produce a multiple regression model, which we call Model 3. In the next section we will produce all of the goodness-of-fit results and analysis to compare all three models.

Table 1 - Goodness-of-Fit for Simple Regression vs Multiple Regression Models

Analysis Components	Model 1 SalePrice = Total_SF;	Model 2 SalePrice = OverallQual	Model 3 SalePrice = Total_SF OvearllQual;
R-Squared	0.7390	0.6415	0.8200
Adjusted R-Squared	0.7389	0.6413	0.8198
Residual Analysis	• **Scatter of Residuals vs Predicted Value** – appears to be a funnel shape indicating heteroscedasticity • **Scatter of Residuals vs Regressor –** there also appears to be a funnel shape in this plot and could indicate a transformation or additional predictor variables are required for improved performance • **Normality Assumption –** based on the histogram of the residuals, they appear to be normally distributed and have a mean of zero • **QQ-Plot –** appears to be some skewing towards the end of each tail of the plot, indicating some outliers may exist • **Linear Relationship between Predictor and Response** – there does appear to be an approximate linear relationship between Total_SF and SalePrice	• **Scatter of Residuals vs Predicted Value** – similar to Model 1, a funnel shape is present and indicates heteroscedasticity • **Scatter of Residuals vs Regressor –** a funnel shape of the plot is also present and could indicate additional predictors are required or a transformation is needed • **Normality Assumption –** the residuals appear to have a non-normal distribution and is slightly skewed to the left • **QQ-Plot –** skewing is present towards the right tail of the plot, indicating outliers must exist in the dataset • **Linear Relationship between Predictor and Response** – there appears to be a linear relationship between SalePrice and OverallQual	• **Scatter of Residuals vs Predicted Value** – based on this plot, there appears to be a non-linear pattern, which could indicate a transformation may be required • **Scatter of Residuals vs Regressors –** the variance appears to increase as total_SF and OverallQual increase, indicating a transformation may be required • **Normality Assumption –** residuals appear to be normally distributed • **QQ-Plot –** there is higher variance towards the upper end of the plot indicating some outliers exist in the data

Let us first compare Models 1 and 2 leveraging the goodness-of-fit and residual analyses performed in Table 1. By comparing the R-Squared values of the two simple regression models, we can conclude that Model 1, leveraging *total*

square footage, fits the data much better than Model 2. Assessing the scatter plots of the residuals against the predicted values for both models, we noticed a funnel pattern was present for both Model 1 and 2. Thus, based on the models "as-is" we can conclude that Model 1 is superior to Model 2. We also saw indications of outliers that exist in the data. Removal of these outliers could improve the models, which we will explore in a later section.

For Model 3, we generated a multiple regression model by combining the regressors from Models 1 and 2. The scatter plot of residuals vs the predicted values actually showed a non-linear pattern, which indicates a transformation could be required to produce a model with better fit to the data. As we already learned from Models 1 and 2, outliers most likely exist in the data, which should be removed for improved model performance. As can be seen in Table 1, the Adjusted R-squared value increased to 0.8192 from 0.7389 when comparing Model 3 to Model 1, respectively. Thus, we can conclude that Model 3 has better performance than either Model 1 or Model 2.

Removal of Outliers

In order to perform exploratory data analysis for outlier detection, we first created categorical intervals of both the *Sale Price* and *Total Square Footage* variables. We then plotted a frequency table for each of those new categorical intervals in order to assess the number of observations in each interval. This provides us with a good assessment of where the outliers may exist in the dataset. After reviewing the frequency table, Figure 1, for *Sale Price* it can be concluded that the majority of observations fall between the price of $100,000 and $300,000, representing ~88% of the population. The results of the frequency table, Figure 2, for *Total Square Footage* indicate that most homes are in the range of 1,000 to 3,000 sqft, representing ~83% of the population. We then deleted all observations that did not fall within the bands mentioned above for both *Sale Price* and *Total Square Footage*. This procedure removed 559 observations from the original dataset. We attempted to find a way to reduce the number of observations removed, however, the remaining number of observations in particular categories were minimal. Thus, the outlier removal methods ended up removing far more than the original 200 targeted to be removed.

Figure 1 - Frequency Table for Sale Price

price_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01: [1; 100,000)	112	5.77	112	5.77
02: [100,000; 150,000)	710	36.60	822	42.37
03: [150,000; 200,000)	576	29.69	1398	72.06
04: [200,000; 250,000)	273	14.07	1671	86.13
05: [250,000; 300,000)	140	7.22	1811	93.35
06: [300,000; 350,000)	65	3.35	1876	96.70
07: [350,000; 400,000)	31	1.60	1907	98.30
08: [400,000+]	33	1.70	1940	100.00

Figure 2 – Frequency Table for Total Square Footage

sqft_cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01: [1; 1,000]	322	16.60	322	16.60
02: (1,000; 1,500]	755	38.92	1077	55.52
03: (1,500; 2,000]	590	30.41	1667	85.93
04: (2,000; 2,500]	196	10.10	1863	96.03
05: (2,500; 3,000]	62	3.20	1925	99.23
06: (3,000+]	15	0.77	1940	100.00

Model Performance after Outliers were Removed

The next step is to rebuild the models from the first section now that the outliers have been removed from the dataset. In Table 2, we will provide the comparisons between the three models after rebuilt using the cleaned dataset, which is now free from outliers.

Table 2 - Goodness-of-Fit for Simple Regression vs Multiple Regression Models after Outliers Removed

Analysis Components	Model 4 SalePrice = Total_SF;	Model 5 SalePrice = OverallQual	Model 6 SalePrice = Total_SF OvearllQual;
R-Squared	0.4525	0.4537	0.6299
Adjusted R-Squared	0.4521	0.4533	0.6293
Residual Analysis	• **Scatter of Residuals vs Predicted Value** – After removing the outliers, the residual plot appears to have a funnel shape indicating heteroscedasticity • **Scatter of Residuals vs Regressor –** there also appears to be funnel shaped for the residuals plotted against total_SF • **Normality Assumption –** the residuals do appear to be normally distrusted based on the histogram of residuals • **QQ-Plot –** This plot also supports the residuals are normally distributed • **Linear Relationship between Predictor and Response** – the plot of the regression clearly shows a linear relationship between total_SF and SalePrice	• **Scatter of Residuals vs Predicted Value** – similar to Model 4, the residual plot against the predicted values also show a funnel shape indicating additional regressors are required • **Scatter of Residuals vs Regressor –** the funnel shape is also present in the residuals plotted against OverallQual • **Normality Assumption –** the residuals appear to be normally distributed and have a mean variance of zero • **QQ-Plot –** as with Model 4, the qq-plot shows a straight line, indicating the residuals are normally distributed • **Linear Relationship between Predictor and Response** – there appears to be a linear relationship between OverallQual and SalePrice as depicted from the plot of the regression model	• **Scatter of Residuals vs Predicted Value** – as with the previous two models, the funnel shape appears in the residual plot for Model 6, indicating a transformation is required. • **Scatter of Residuals vs Regressors –** the variance for total_SF appear to increase as total_SF increases whereas OvearllQual appears relatively constant • **Normality Assumption –** the residuals appear to be normally distributed and have a mean variance of zero • **QQ-Plot –** the same is true for Model6 in that the qq-plot shows a straight line and appears to be normally distributed

As can be seen in Table 2, the overall performance of Models 4 – 6 over Models 1-3 resulted in significantly lower R-squared and Adjusted R-Squared values. The residual analysis also indicates unequal variance after analyzing the residual plots against the predictor and response variable in all three models. Thus, the removal of outliers does not seem to have improved the models significantly.

Model Performance after performing a Log Transformation on Sale Price

In this section we will analyze

Table 3 - Goodness-of-Fit for Multiple Regression Models before and after a Log Transformation on Sale Price

Analysis Components	Model 7 SalePrice = Total_SF OverallQual total_baths good_kitchen good_exterior;	Model 8 Log(SalePrice) = Total_SF OverallQual total_baths good_kitchen good_exterior
R-Squared	0.7021	0.6949
Adjusted R-Squared	0.7010	0.6938
Residual Analysis	• **Scatter of Residuals vs Predicted Value** – the residual plot still appears to have a funnel shape indicating the variance of residuals is not constant • **Scatter of Residuals vs Regressors –** as with the first residual plot, the funnel shape is still present indicating heteroscedasticity • **Normality Assumption –** the residuals appear to be normally distributed and have a mean of zero • **QQ-Plot –** the straight line in the qq-plot is still present, indicating a normally distributed error term	• **Scatter of Residuals vs Predicted Value** – the funnel shape is not present in Model 8, indicating the variance is constant after the log transformation is performed • **Scatter of Residuals vs Regressors –** there appears to be relatively constant variance in Model 8 vs Model 7 for all regressors • **Normality Assumption –** the residuals appear to be normally distributed and have a mean of zero • **QQ-Plot –** the straight line in the qq-plot is still present, indicating a normally distributed error term

Based on the analysis presented in Table 3, performing the log transformation did not significantly change the Adjusted R-Squared values. However, while assessing the residuals between both models, the residuals from Model 8 are homoscedastic as opposed to the heteroscedasticity shown in Model 7. Thus, we can conclude that Model 8 fits the data much better than Model 7 after the log transformation is performed.

Conclusion

When comparing the original simple and multiple regression models against the models generated after the outlier removal step (Table 1 vs Table 2), we can clearly see that model performance decreased, rather than increased as originally anticipated. This result indicates that what we thought were outliers actually aided, rather than degraded, overall model performance. In order to improve the outcome of Models 4 – 9, further assessment of outlier removal would be required. A potential secondary option is to generate the studentized residuals in the original dataset and identify outliers exceeding a particular absolute value of the studentized residual values. Overall, Model 3 outperformed all other models produced in this analysis in Tables 1 and 2. In the last exercise of comparing a log transformation on the response variable and no transformation at all, we can see that in this particular instance, the transformation was beneficial in improving the model.

SAS Code Output

```
1  *-------------------------------------------------------------------------------------
2  *  Nate Bitting
3  *  Assignment 3
4  *-------------------------------------------------------------------------------------;
5
6  * Code used to get the data into my library;
7  ods graphics on;
8  libname mydata '/courses/d6fc9ae5ba27fe300/c_3505/SAS_Data/' access=readonly;
9  proc datasets library=mydata; run; quit;
10
11 *-------------------------------------------------------------------------------------
12 *  Create original dataset by filtering out unnecessary data and adding new categorical
13    features to the dataset
14 *-------------------------------------------------------------------------------------;
15
16 * Smart data formats for sale price and square footage;
17 proc format;
18 value price_sfmt
19     . = '10: Missing'
20     1 -< 100000 = '01: [1; 100,000)'
21     100000 -< 150000 = '02: [100,000; 150,000)'
22     150000 -< 200000 = '03: [150,000; 200,000)'
23     200000 -< 250000 = '04: [200,000; 250,000)'
24     250000 -< 300000 = '05: [250,000; 300,000)'
25     300000 -< 350000 = '06: [300,000; 350,000)'
26     350000 -< 400000 = '07: [350,000; 400,000)'
27     400000 - high = '08: [400,000+]'
28     other = '09: Invalid Value'
29     ; * use a semi-colon to end each format in the proc format statement;
30     * Note on how we use the < to create open intervals.
31    * The dash - will create closed intervals, and
32    hence the dash should only be used with discrete values;
33 value sqft_sfmt
34     . = '08: Missing'
35     1 - 1000 = '01: [1; 1,000]'
36     1000 <- 1500 = '02: (1,000; 1,500]'
37     1500 <- 2000 = '03: (1,500; 2,000]'
38     2000 <- 2500 = '04: (2,000; 2,500]'
39     2500 <- 3000 = '05: (2,500; 3,000]'
40     3000 - high = '06: (3,000+]'
41     other = '07: Invalid Value'
42     ;
43 run;
44
45 * Dataset before removing outliers;
46 Data building;
47     SET mydata.ames_housing_data;
48
49     * filter on only single family homes that meet specific criteria;
50     if (SaleCondition = 'Normal');
51     if (BldgType = '1Fam'); * single family homes only;
52     if (Zoning in ('RH','RL','RP','RM')); * residential zones only;
53     if (Street='Pave'); * paved streets only;
54     if (Utilities='AllPub'); * only homes with public utilities;
```

```
55
56      log_price = log(SalePrice); *create a variable for the natural log of SalePrice;
57
58      * create new variables by combining multiple variables in the housing dataset;
59      total_SF = max(GrLivArea,0) + max(TotalBsmtSF,0);
60      total_baths = max(FullBath,0) + max(BsmtFullBath,0);
61      total_halfbaths = max(HalfBath,0) + max(BsmtHalfBath,0);
62      total_baths_calc = total_baths + total_halfbaths;
63
64      * Neighborhood dummy variables;
65      if (Neighborhood = 'Blmngtn') then Blmngtn=1; else Blmngtn=0;
66      if (Neighborhood = 'Blueste') then Blueste=1; else Blueste=0;
67      if (Neighborhood = 'BrDale') then BrDale=1; else BrDale=0;
68      if (Neighborhood = 'BrkSide') then BrkSide=1; else BrkSide=0;
69      if (Neighborhood = 'ClearCr') then ClearCr=1; else ClearCr=0;
70      if (Neighborhood = 'CollgCr') then CollgCr=1; else CollgCr=0;
71      if (Neighborhood = 'Crawfor') then Crawfor=1; else Crawfor=0;
72      if (Neighborhood = 'Edwards') then Edwards=1; else Edwards=0;
73      if (Neighborhood = 'Gilbert') then Gilbert=1; else Gilbert=0;
74      if (Neighborhood = 'Greens') then Greens=1; else Greens=0;
75      if (Neighborhood = 'GrnHill') then GrnHill=1; else GrnHill=0;
76      if (Neighborhood = 'IDOTRR') then IDOTRR=1; else IDOTRR=0;
77      if (Neighborhood = 'Landmrk') then Landmrk=1; else Landmrk=0;
78      if (Neighborhood = 'MeadowV') then MeadowV=1; else MeadowV=0;
79      if (Neighborhood = 'Mitchel') then Mitchel=1; else Mitchel=0;
80      if (Neighborhood = 'NAmes') then NAmes=1; else NAmes=0;
81      if (Neighborhood = 'NPkVill') then NPkVill=1; else NPkVill=0;
82      if (Neighborhood = 'NWAmes') then NWAmes=1; else NWAmes=0;
83      if (Neighborhood = 'NoRidge') then NoRidge=1; else NoRidge=0;
84      if (Neighborhood = 'NridgHt') then NridgHt=1; else NridgHt=0;
85      if (Neighborhood = 'OldTown') then OldTown=1; else OldTown=0;
86      if (Neighborhood = 'SWISU') then SWISU=1; else SWISU=0;
87      if (Neighborhood = 'Sawyer') then Sawyer=1; else Sawyer=0;
88      if (Neighborhood = 'SawyerW') then SawyerW=1; else SawyerW=0;
89      if (Neighborhood = 'Somerst') then Somerst=1; else Somerst=0;
90      if (Neighborhood = 'StoneBr') then StoneBr=1; else StoneBr=0;
91      if (Neighborhood = 'Timber') then Timber=1; else Timber=0;
92      if (Neighborhood = 'Veenker') then Veenker=1; else Veenker=0;
93
94      * KitchenQual dummy variable;
95      if (KitchenQual in ('Ex', 'Gd')) then good_kitchen=1; else good_kitchen=0;
96
97      * FireplaceQu dummy variable;
98      if (FireplaceQu in ('Ex', 'Gd')) then good_fireplace=1; else good_fireplace=0;
99
100     * ExterQual dummy variable;
101     if (ExterQual in ('Ex', 'Gd')) then good_exterior=1; else good_exterior=0;
102
103     * Foundation dummy variables;
104     if (Foundation = 'BrkTil') then Foundation_BrkTil=1; else Foundation_BrkTil=0;
105     if (Foundation = 'CBlock') then Foundation_CBlock=1; else Foundation_CBlock=0;
106     if (Foundation = 'PConc') then Foundation_PConc=1; else Foundation_PConc=0;
107     if (Foundation = 'Slab') then Foundation_Slab=1; else Foundation_Slab=0;
108     if (Foundation = 'Stone') then Foundation_Stone=1; else Foundation_Stone=0;
109     if (Foundation = 'Wood') then Foundation_Wood=1; else Foundation_Wood=0;
```

```
    * MoSold dummy variables;
    if (MoSold = 1) then jan_sold=1; else jan_sold=0;
    if (MoSold = 2) then feb_sold=1; else feb_sold=0;
    if (MoSold = 3) then mar_sold=1; else mar_sold=0;
    if (MoSold = 4) then apr_sold=1; else apr_sold=0;
    if (MoSold = 5) then may_sold=1; else may_sold=0;
    if (MoSold = 6) then jun_sold=1; else jun_sold=0;
    if (MoSold = 7) then jul_sold=1; else jul_sold=0;
    if (MoSold = 8) then aug_sold=1; else aug_sold=0;
    if (MoSold = 9) then sep_sold=1; else sep_sold=0;
    if (MoSold = 10) then oct_sold=1; else oct_sold=0;
    if (MoSold = 11) then nov_sold=1; else nov_sold=0;
    if (MoSold = 12) then dec_sold=1; else dec_sold=0;

    * Construct a composite quality index;
    quality_index = OverallCond*OverallQual;

    * Central Air Indicator;
    if (CentralAir='Y') then central_air=1; else central_air=0;
    * Fireplace Indicator;
    if (Fireplaces>0) then fireplace_ind=1; else fireplace_ind=0;
    * Garage Indicator;
    if (GarageCars>0) then garage_ind=1; else garage_ind=0;
    * Good Basement Indicator;
    if (BsmtQual in ('Ex','Gd')) or (BsmtCond in ('Ex','Gd')) then good_basement_ind=1; else good_basement_ind=0;

    *apply the put function to create the categorical variables for the various scales of price and sqft;
    price_cat = put(SalePrice,price_sfmt.);
    sqft_cat = put(total_SF,sqft_sfmt.);

run; quit;


* Review new dataset to ensure no missing values based on filtering;
proc contents data=building;
run; quit;

*-------------------------------------------------------------------------------------
*  Models 1 - 3 are built before outliers are removed
*-------------------------------------------------------------------------------------;

* 1st simple regression model - Model 1;
proc reg data=building;
    model SalePrice = total_SF;
run; quit;

* 2nd simple regression model - Model 2;
proc reg data=building;
    model SalePrice = OverallQual;
run; quit;

* Multiple Regression Model - Model 3;
proc reg data=building;
    model SalePrice = total_SF OverallQual;
run; quit;
```

```
168 *-------------------------------------------------------------------------------------
169 *  Analysis to determine outliers to remove from the dataset
170 *-------------------------------------------------------------------------------------;
171
172 * Create frequency tables for each category;
173 proc freq data=building;
174 tables price_cat sqft_cat;
175 run; quit;
176
177 * Dataset after removing outliers;
178 Data building_no_outliers;
179     SET building;
180
181     *remove outliers for total square footage;
182     if sqft_cat in ('01: [1; 1,000]','06: (3,000+]', '07: Invalid Value','08: Missing') then do;
183         outlier_score = 1;
184     end;
185     if price_cat in ('01: [1; 100,000)','06: [300,000; 350,000)',
186                      '07: [350,000; 400,000)','08: [400,000+]','09: Invalid Value') then do;
187                      outlier_score = 2;
188     end;
189
190     if (outlier_score>0) then delete;
191
192 run; quit;
193
194 * Create frequency tables for each category;
195 proc freq data=building_no_outliers;
196 tables price_cat * sqft_cat;
197 run; quit;
198
199
200 *-------------------------------------------------------------------------------------
201 *  Models 4 - 6 include a log transformation on the response variable, SalePrice
202 *-------------------------------------------------------------------------------------;
203
204 * 1st simple regression model without outliers - Model 4;
205 proc reg data=building_no_outliers;
206     model SalePrice = total_SF;
207 run; quit;
208
209 * 2nd simple regression model without outliers - Model 5;
210 proc reg data=building_no_outliers;
211     model SalePrice = OverallQual;
212 run; quit;
213
214 * Multiple Regression Model - Model 6;
215 proc reg data=building_no_outliers;
216     model SalePrice = total_SF OverallQual;
217 run; quit;
```

```
220 *-------------------------------------------------------------------------------------
221 *  Models 7 - 8 compare two models with the same predictors, but the 2nd model
222 *  has a log transformation on the response variable, SalePrice
223 *-------------------------------------------------------------------------------------;
224
225 * 1st multipe regression model WITHOUT a log transformation on SalePrice - Model 7;
226 proc reg data=building_no_outliers;
227     model SalePrice = total_SF OverallQual total_baths good_kitchen good_exterior;
228 run; quit;
229
230 * 2nd simple regression model WITH a log transformation on SalePrice - Model 8;
231 proc reg data=building_no_outliers;
232     model log_price = total_SF OverallQual total_baths good_kitchen good_exterior;
233 run; quit;
```