James Aniciete

STAT 481

Project 1 – Multiple Regression

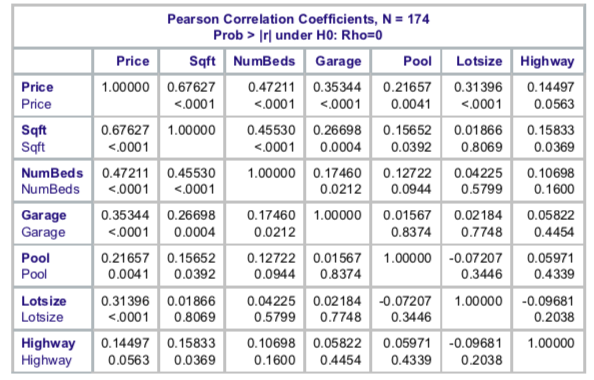
11/2

**Summary Statistics**

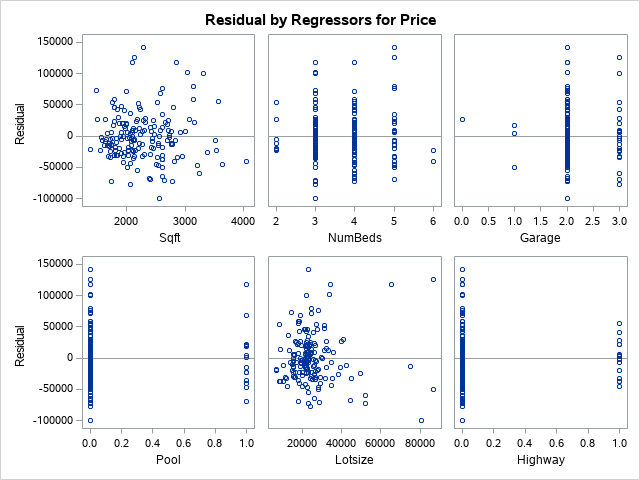
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Summary Statistics | | | | | | |
| Variable | Sample Size | Minimum Value | Median | Mean | Variance (Standard Deviation) | Maximum Value |
| Price | 174 | 156000 | 253450.0 | |  |  | | --- | --- | | 264745.8 |  | | 4146991353  (64397) | 520000 |
| Sqft | 174 | 1384 | |  |  | | --- | --- | | 2156.000 |  | | |  |  | | --- | --- | | 2276.339 |  | | |  |  | | --- | --- | |  | 237307 | |  | (487.14216) | | | 4050 |
| NumBeds | 174 | 2 | 4.000000 | 3.649425 | 0.66829  (0.81749) | 6 |
| Garage | 174 | 0 | 2.000000 | 2.120690 | 0.17610  (0.41964) | 3 |
| Pool | 174 | 0 | 0.000000 | 0.080460 | 0.07441  (0.27279)   |  | | --- | |  | | 1 |
| Lotsize | 174 | 6734.0 | 22516.50 | 24906.77 | 150194537  (12255) | 86830.0 |
| Highway | 174 | 0 | |  |  | | --- | --- | | 0.000000 |  | | |  |  | | --- | --- | | 0.086207 |  | | 0.07923  (0.28148) | 1 |

**First Regression Model**

* Correlation: Since none of the explanatory variables are perfectly correlated, they are all included in the model.



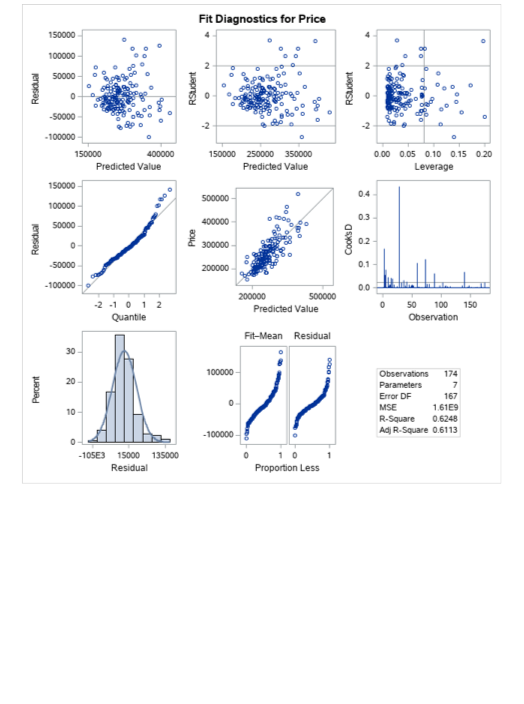
* Model Checking
  + Check for Linearity: The plots of the residuals against Sqft, Numbeds, and Lotsize do not appear to have a pattern, so they satisfy the linearity check. Garage, Pool, and Highway appear to have a pattern where each is mainly concentrated over a single value, so they do not satisfy the linearity check. Since the values of Garage, Pool, and Highway appear to remain relatively constant, excluding said variables may improve the regression model.



* + Check for Independence: Since we are not working with time-series data or data gathered in a sequence such as a geographical location, there is no need to check for independence. We will assume it is satisfied.
  + Check for Normally Distributed Errors: The Shapiro-Wilk Test for Normality yields a p-value < .0001. Using a .05 significance level, we reject the null hypothesis of the errors following a normal distribution. Thus, the model does not have normally distributed errors.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tests for Normality** | | | | |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.956632 | **Pr < W** | <0.0001 |
| **Kolmogorov-Smirnov** | **D** | 0.089805 | **Pr > D** | <0.0100 |
| **Cramer-von Mises** | **W-Sq** | 0.32412 | **Pr > W-Sq** | <0.0050 |
| **Anderson-Darling** | **A-Sq** | 1.980717 | **Pr > A-Sq** | <0.0050 |

* + Check for Equal Variances: As shown in the graph in the top left corner of the set of graphs, the plot of the residuals against the fitted values does not appear to have a pattern. Thus, the model satisfies the test for equal variances.

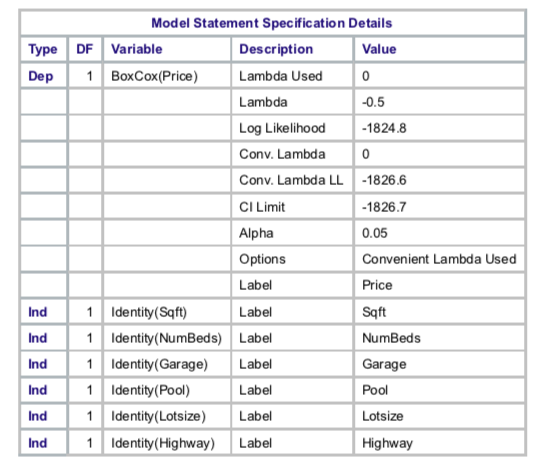


* Regression Model 1:



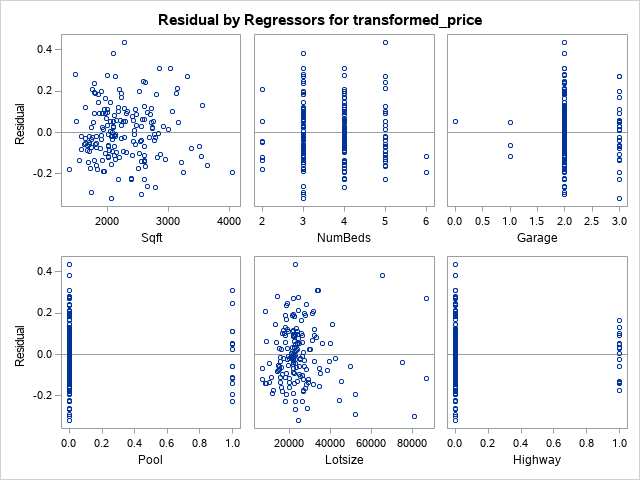
**Transformation**

Let Y = Price. Since we have non-normal error terms, we will transform Y using a Box-Cox Transformation. We obtain that λ = 0, so we will use Y' = loge(Y) = loge(Price).

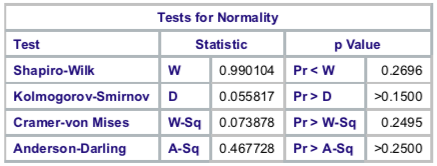


**Second Regression Model**

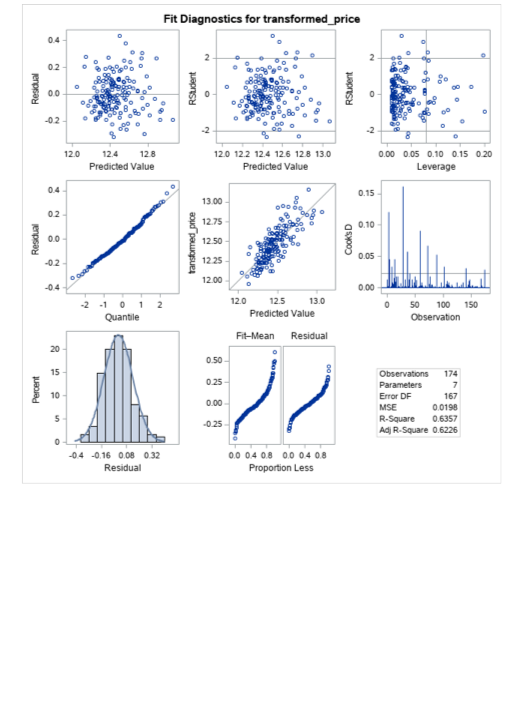
* Model Checking
  + Check for Linearity: The plots of the residuals against Sqft, NumBeds, and Lotsize do not appear to have a pattern, so they satisfy the linearity check. Meanwhile, Garage, Pool, and Highway appear to have a pattern where each is mainly concentrated over a single value, so they do not satisfy the linearity check. Since Garage, Pool, and Highway appear to remain relatively constant, excluding said variables may improve the regression model.



* + Check for Independence: Once again, the check for independence will be assumed to be satisfied since we are not working with time-series data or data gathered in a sequence such as geographic areas.
  + Check for Normally Distributed Errors: The Shapiro-Wilk Test for Normality yields a p-value = .2696. Using a .05 significance level, we do not reject the null hypothesis of the errors following a normal distribution. Thus, the model now has normally distributed errors.



* + Check for Equal Variances: As shown in the graph in the top left corner of the set of graphs, the plot of the residuals against the fitted values does not appear to have a pattern. Thus, the model satisfies the test for equal variances.

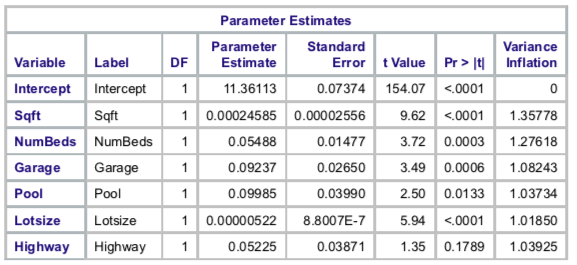


* Regression Model 2:



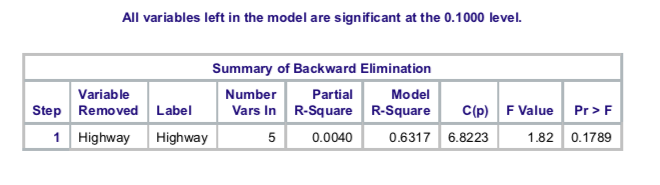
**Multicollinearity**

Since none of the explanatory variables have a VIF > 10, multicollinearity will not be influencing the least square estimates. Thus, none of the variables will be excluded from the model.



**Best Model**





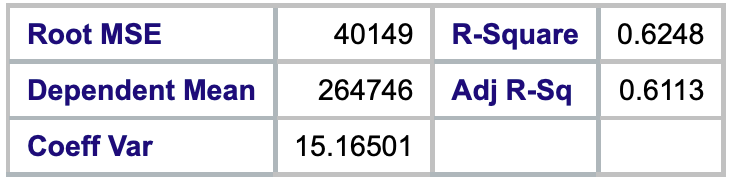
Using backward selection with the criteria for inclusion as having a significance of .10 or lower, we remove the variable Highway since it had a p-value > .10 and keep the other variables.

*loge(Price)* = 11.35525 + .00024981*Sqft* + .05570*NumBeds* + .09297*Garage*

+ .10126*Pool* + .00000510*Lotsize*

**Conclusions Drawn**

* Should there be a house with zero square feet of finished residence, no bedrooms, a garage that can hold zero cars, no pool, and a lot size of zero square feet, the expected log price of a home would be $11.35525.
* If the square feet of finished residence rises by 1 square foot, other variables held constant, the expected log price of a home should increase by $0.00024981.
* If the number of total bedrooms rises by 1 bedroom, other variables held constant, the expected log price of a home should increase by $0.05570.
* If the number of cars that a home’s garage will hold increases by 1 car, other variables held constant, the expected log price of a home should increase by $0.09297.
* If a pool were added to a home, other variables held constant, the expected log price of a home should increase by $0.10126.
* If the lot size of a home was increased by 1 square foot, other variables held constant, the expected log price of a home should increase by $0.00000510.
* As seen at the top of this page, the R2 of the best model is 0.6317. This means that the best model explains 63.17% of the variation in the log price of a home, which is better than the R2 of 62.48% from the original model.



/\* Project 1 \*/

PROC IMPORT DATAFILE="/folders/myfolders/Dataset\_175-348.xlsx"

OUT=data

DBMS=XLSX

REPLACE;

RUN;

PROC PRINT DATA = data;

RUN;

/\* Summary Statistics \*/

PROC UNIVARIATE DATA = data;

VAR price sqft numbeds garage pool lotsize highway;

RUN;

RUN;

/\* Regression Model 1 - code includes the correlation coefficients and the linearity check \*/

PROC REG DATA = data;

model price = sqft numbeds garage pool lotsize highway / corrb;

title 'Model 1';

output out = out1 r = resid1;

RUN;

/\* Normality Check \*/

PROC UNIVARIATE NORMAL PLOT DATA = out1;

VAR resid1;

RUN;

/\* Box-Cox Transformation (Transforming Y) \*/

PROC TRANSREG DATA = data DETAIL;

model boxcox (price / convenient lambda = -3 to 3 by .125) = identity(sqft numbeds garage pool lotsize highway);

OUTPUT out = out2;

RUN;

/\* Transformation Y' = log\_e(Y) \*/

DATA data;

SET data;

newcolumnname = transformed\_price;

transformed\_price = log(price);

RUN;

PROC PRINT DATA = data;

RUN;

/\* Regression Model 2 (with transformation) - includes the correlation coefficients, linearity check, and collinearity (VIF) \*/

PROC REG DATA = data;

model transformed\_price = sqft numbeds garage pool lotsize highway / corrb VIF;

title 'Model 2';

output out = out2 r = resid2;

RUN;

/\* Normality Check \*/

PROC UNIVARIATE NORMAL PLOT DATA = out2;

VAR resid2;

RUN;

/\* "Forward Selection Model" with inclusion for significance <= .10 \*/

PROC REG DATA = data;

model transformed\_price = sqft numbeds garage pool lotsize highway / corrb selection = forward SLENTRY = .10;

title 'Forward Selection Model';

output out = out3 r = resid3;

RUN;

/\* "Backward Selection Model" with inclusion for significance <= .10 \*/

/\* Backward Selection has fewer steps - choose to use this one. \*/

PROC REG DATA = data;

model transformed\_price = sqft numbeds garage pool lotsize highway / selection = backward SLENTRY = .10;

title 'Backward Selection Model';

output out = out3 r = resid3;

RUN;

/\* Normality Check \*/

PROC UNIVARIATE NORMAL PLOT DATA = out3;

VAR resid3;

RUN;