Housing Price Prediction

Sam Johnson, Kinsey Shaw, Regan Tracy

November 5, 2021

1 Introduction

Homes are always unique to the homeowner. The real estate market is diverse and includes high-end colonial homes, modest one-floor homes, and everything in between. Specific attributes of a home contribute to its price, including amenities such as utilities, style, housing material, and others. Through her studies of sales prices in the housing market, Ewelina Woiciak (2016), an Environmental Engineering faculty member at the AGH University of Science and Technology, found that the diversification in the portfolio of market features is directly related to the cost of the home. In our study, we analyzed the data from the Ames Assessor's Office in Iowa, USA, to predict the price of a home in today's market.

With a subject as vital as the construction of a home, it is crucial to be aware of the price that comes with having a home filled with everything you desire. Our study focused on data collected by the Ames Assessor's Office in Iowa, USA, which comprises 82 variables associated with a home's attributes, collected from 2006 to 2010 (City of Ames, IA 2021). You can find a summary of the variables in the appendix, including zoning fees, real estate appraisals, and legislative permits–all factors that heavily influence a home's sale price.

2 Data Preprocessing

By looking at the Appendix, we can see there are four different types of variables in the data set: continuous, discrete, nominal, and ordinal. Continuous and ordinal variables require no pre-processing, as they are already numerical. In the following subsections we detail how nominal and ordinal variables were transformed into usable numerical information.

2.1 Nominal Variables

We opted to transform the nominal variables by creating $n-1$ new features where n equals the number of different values the variable can take on. We named the new features with the original variable name plus an underscore and the value of the variable.

For example: If the feature is *GarageType* and the value of a given observation is *"Attchd"*, we would create a new dummy feature called *GarageType_Attchd*.

2.2 Ordinal Variables

Ordinal variables are categorical variables that are ordered. Because they are ordered, we can simply replace their values with the numbers $1, 2, 3, ..., n$ where n equals the number of different values the variable can take on. The lowest value will be given a 1 while the highest value will be given n.

An example is given in Table 1, showing the values that the ordinal variable *ExterQual* can take on, their interpretation, and their numerical equivalent:

Name	Interpretation	Numerical Equivalent
Ex	Excellent	5
Gd	Good	4
TA	Average/Typical	3
Fa	Fair	2
Po	Poor	1

Table 1: Example of Ordinal to Numerical Variable

2.3 New Feature Creation

Before running any models we hypothesized that it might be a good idea to add a couple of features to the model. The following features were created:
$age = YrSold - YearBuilt$
$ageRemodel = YrSold - YearRemodAdd$

3 Multicollinearity

We fit an initial multiple linear regression model with all of our variables except $Order$ and PID, which are identifiers. We noted that a number of our variables were experiencing multicollinearity, demonstrated through high variance inflation factor (VIF) values. We dealt with this issue by removing one variable at a time until all of the VIF values were below 10. The following variables were sequentially removed from the model:
$YearBuilt$, $ageRemodel$, $RoofStyle\_Gable$, $GarageType\_Attchd$, $ExteriorFirst\_VinylSd$, $BldgType\_OneFam$, $SecondFlrSF$, $GrLivArea$, $GarageCars$, $BsmtQual$

4 Transformations & Assumptions

We ran an initial linear regression model with all remaining explanatory variables and generated the plots in Figure 5. The residual plot displays non-constant variance and the Q-Q plot demonstrates non-normal residuals.

We generated a Cook's D plot in Figure 1 and a leverage plot in Figure 2 to look for outliers and influential points. We can see in Figure 1 that the model contains no outliers. Figure 2 suggests that there are a few highly influential points.

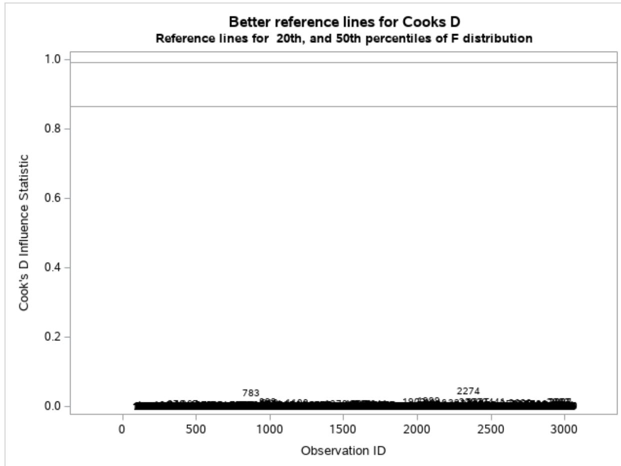


Figure 1: Cook's D Before Transformation

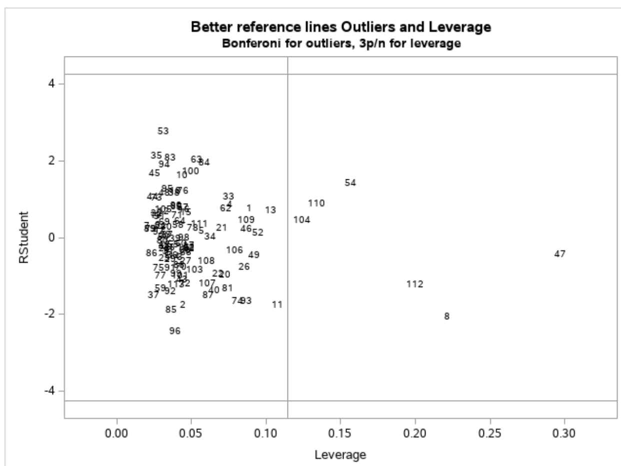


Figure 2: Leverage Plot Before Transformation

Then, we looked at the histograms of our variables. The following variables were right-skewed:
$LotArea$, $BsmtFinSFOne$, $TotalBsmtSF$, $FirstFlrSF$, $GarageArea$, $WoodDeckSF$, $OpenPorchSF$, $SalePrice$, Age

We transformed the right-skewed variables to eliminate issues with non-constant variance, non-normality, and highly influential points. The following variables were created:

$log\_LotArea = \log(LotArea)$
$BsmtFinSFOne\_sqrt = \sqrt{BsmtFinSFOne}$
$TotalBsmtSF\_sqrt = \sqrt{TotalBsmtSF}$
$log\_FirstFlrSF = \log(FirstFlrSF)$
$sqrt\_GarageArea = \sqrt{GarageArea}$
$sqrt\_WoodDeckSF = \sqrt{WoodDeckSF}$

$sqrt\_OpenPorchSF = \sqrt{OpenPorchSF}$
$log\_SalePrice = \log(SalePrice)$
$sqrt\_Age = \sqrt{Age}$

Figure 3 shows the Cook's D after the transformations have been applied. There are still no outliers. Figure 4 shows the leverage plot after the transformations. Unfortunately, there are still a few points with high influence that we were unable to correct. Later, we will employ robust regression to see if we can overcome this issue.

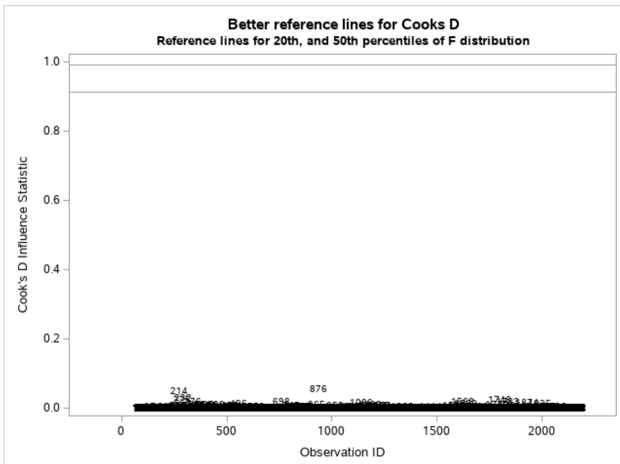


Figure 3: Cook's D After Transformation

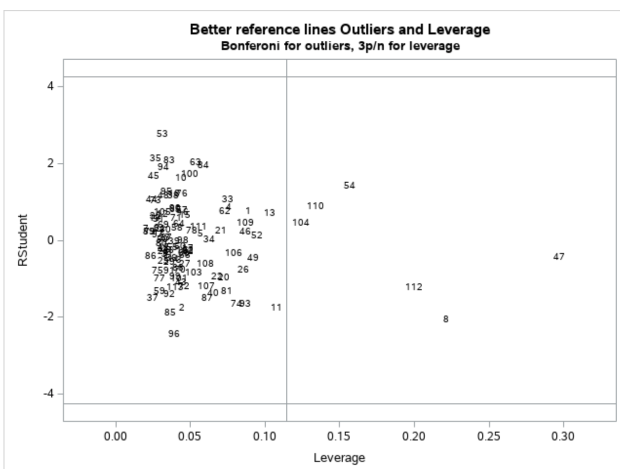
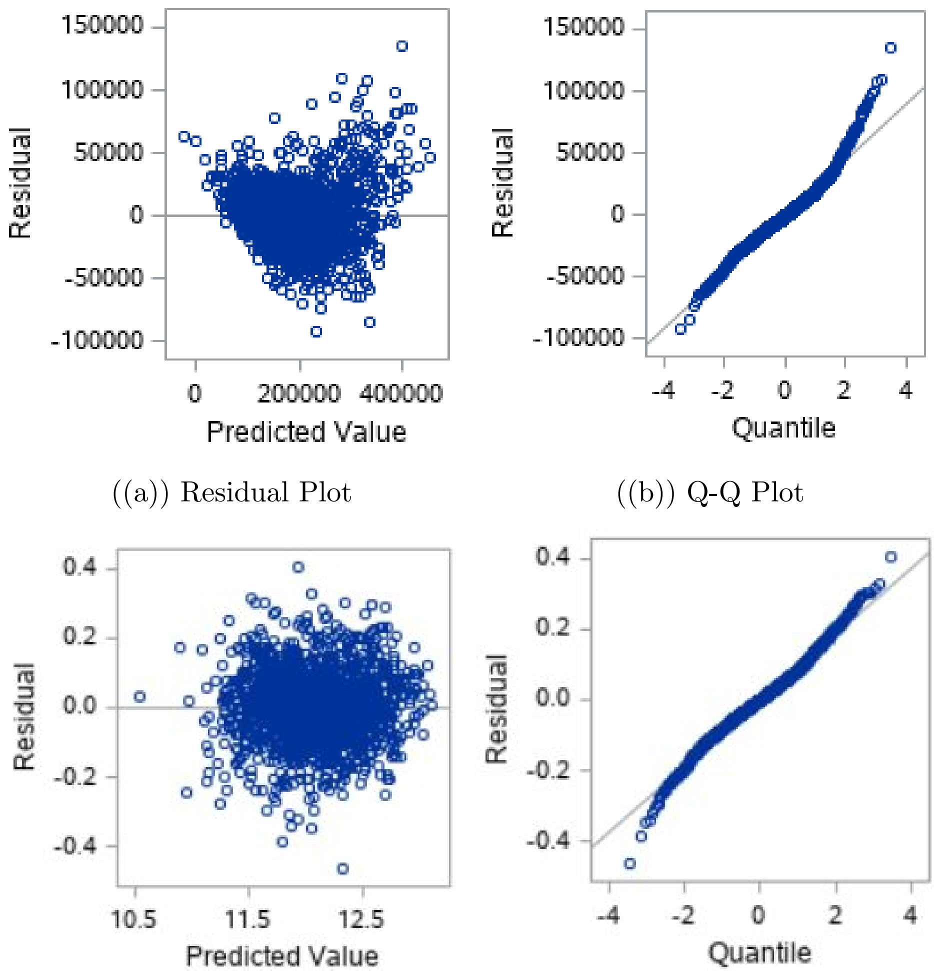


Figure 4: Leveraged Plot After Transformation

Figure 5 shows some additional diagnostics from before and after the transformations were applied. The residual plots show that the previously non-constant variance has been corrected. The Brown-Forsythe test of constant variance backs up this view with a p-value of 0.17694, which is greater than 0.05.

The Q-Q plots show that the non-normality problem has been corrected by the transformations. The correlation test of normality solidifies this view with a value of 0.99464, which is greater than the minimum required value of 0.987 where $n = 100$ and $\alpha = 0.05$.



((a)) Residual Plot ((b)) Q-Q Plot

((c)) Residual Plot for Log ((d)) Q-Q Plot for Log

Figure 5: Before and After Transformation Graphs

5 Interactions

We created three interaction variables to enhance the model's predictive ability. The formulas are given below:

$culda\_con = LotConfig\_CulDSa * LandContour\_Lvl$
$Asphshn\_hip = ExteriorFirst\_AsphShn * RoofStyle\_hip$
$GasW\_CBlock = Heating\_GasW * Foundation\_CBlock$

We created these new features to see if our variables were being influenced by each other.

We first tested the potential interaction between a lot with a cul-da-sac and a level land contour. We did this because it is difficult to build a well functioning cul-da-sac on unlevel land. The resulting p-value was statistically significant at 0.0001.

In our second interaction, we combined a home having an asphalt exterior with a hip style rooftop. The outside look of a home is as important to a homeowner as the inside. Hip style rooftops are some of the most popular rooftops in America, due to their clean look and strong structure. This interaction was statistically insignificant with a p-value of 0.1403.

In our final interaction, we interacted the cinder block foundation to gas heating. We wanted to see if there was a potential interaction that would represent the type of insulation the home would have. This interaction was not significant either, with a p-value of 0.41.

In summary, in a regression with just these three interaction terms, only one of them came out statistically significant and was kept in the model. In the next section we will see if model selection chooses to keep the remaining interactions in the model.

6 Model Selection

Before performing model selection, we put 75% of the data into a training set and the remaining 25% of the data into a testing set. By withholding data from the training set, we can evaluate later how well our model performs on new unseen data.

We will also withhold the $SalePrice$ observations 335 and 295 from the training dataset. Later on, we will fit our model and analyze how well the model performs on these specific observations.

To choose a final model, we ran backwards selection, stepwise selection, and all possible regressions. The results are given below:

Backward selection returned a model with 40 different explanatory variables. It had an R^2_{adj} value of 0.9357.

Stepwise selection returned a model with 40 different explanatory variables. It had an R^2_{adj} value of 0.9357.

The model with the highest R^2_{adj} returned a model with 51 explanatory variables. It had an R^2_{adj} value of 0.9359.

The $C(P)$, AIC, and SBC metrics all selected models with 46 explanatory variables and an R^2_{adj} of 0.9359.

The model with the lowest number of explanatory variables was selected by both backward selection and stepwise selection. Furthermore, it had a high R^2_{adj} value of 0.9357, only slightly lower than the highest value of 0.9359. The 40 variable model is our *best* model. The theoretical model is as follows:

$$
\begin{aligned}
log\_SalePrice = {} & \beta_0 + \beta_1 * sqrt\_Age + \beta_2 * log\_LotArea + \beta_3 * OverallQual \\
& + \beta_4 * OverallCond + \beta_5 * YearRemodAdd + \beta_6 * BsmtCond \\
& + \beta_7 * BsmtExposure + \beta_8 * BsmtFinSFOne\_sqrt + \beta_9 * TotalBsmtSF\_sqrt \\
& + \beta_{10} * log\_FirstFlrSF + \beta_{11} * BsmtFullBath + \beta_{12} * FullBath \\
& + \beta_{13} * HalfBath + \beta_{14} * BedroomAbvGr + \beta_{15} * KitchenAbvGr \\
& + \beta_{16} * KitchenQual + \beta_{17} * TotRmsAbvGrd + \beta_{18} * Fireplaces \\
& + \beta_{19} * sqrt\_GarageArea + \beta_{20} * sqrt\_WoodDeckSF \\
& + \beta_{21} * sqrt\_OpenPorchSF + \beta_{22} * GarageType\_BuiltIn \\
& + \beta_{23} * GarageType\_nan + \beta_{24} * LandContour\_HLS \\
& + \beta_{25} * LotConfig\_CulDSa + \beta_{26} * LotConfig\_Inside \\
& + \beta_{27} * HouseStyle\_OnePointFiveUnf + \beta_{28} * HouseStyle\_OneStory \\
& + \beta_{29} * HouseStyle\_SFoyer + \beta_{30} * HouseStyle\_SLvl \\
& + \beta_{31} * HouseStyle\_TwoPointFiveUnf + \beta_{32} * HouseStyle\_TwoStory \\
& + \beta_{33} * RoofStyle\_Gambr + \beta_{34} * RoofStyle\_Hip \\
& + \beta_{35} * ExteriorFirst\_BrkComm + \beta_{36} * ExteriorFirst\_BrkFace \\
& + \beta_{37} * ExteriorFirst\_MetalSd + \beta_{38} * ExteriorFirst\_PreCast \\
& + \beta_{39} * Foundation\_PConc + \beta_{40} * CentralAir\_Y + \epsilon
\end{aligned}
$$

7 Rechecking Assumptions

Now that we have a final model, we will run through our assumptions one more time to ensure that they are met.

Figure 6 contains the residual and Q-Q plots for the final model. We can see that the variance is constant and that the residuals are normally distributed. The p-value for the Brown-Forsythe test of constant variance is a massive 0.629, further solidifying this view. The correlation test of normality of residuals results in an output of 0.996, satisfying the assumption of normality.

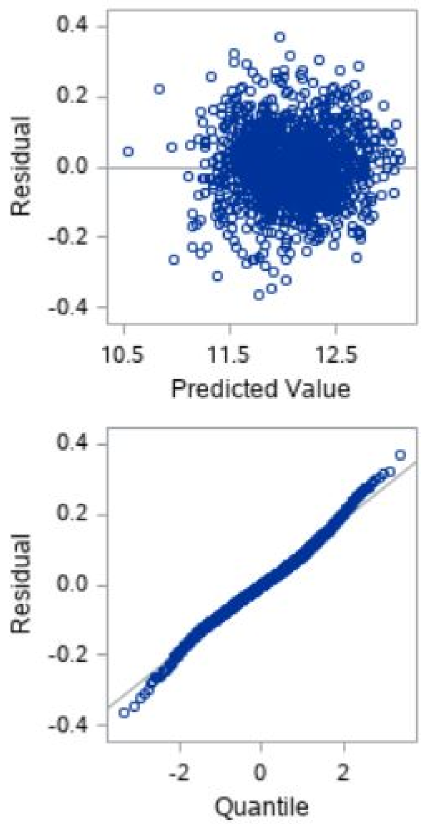


Figure 6: Final Model Residuals & QQ Plot

Figure 7 shows that no outliers exist in the final model.

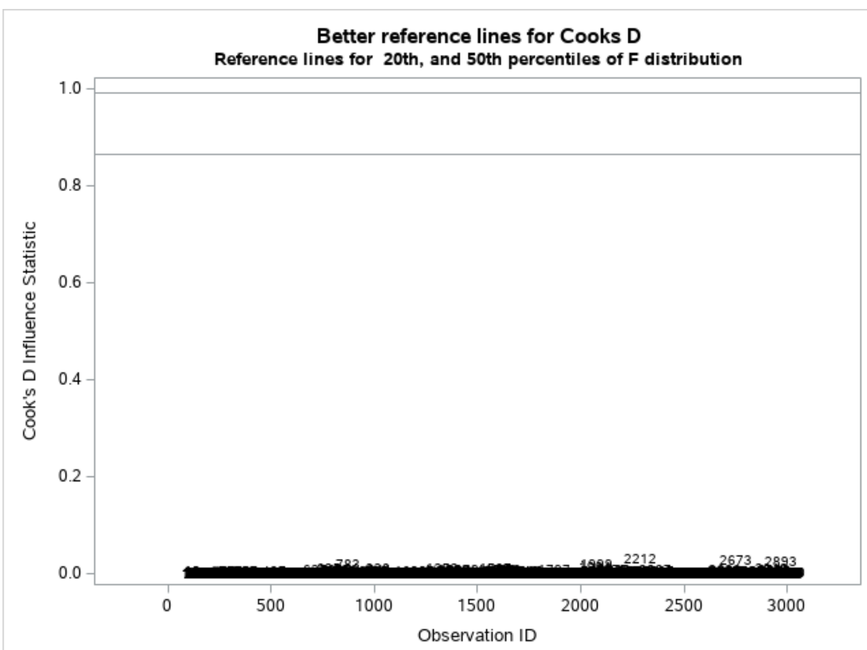


Figure 7: Final Model Cook's D

Figure 8 shows that there are still some highly influential points even after performing various transformations. In our case, we were unable to satisfy all of our assumptions even after performing various transformations. Later we will try robust regression to deal with these influential points.

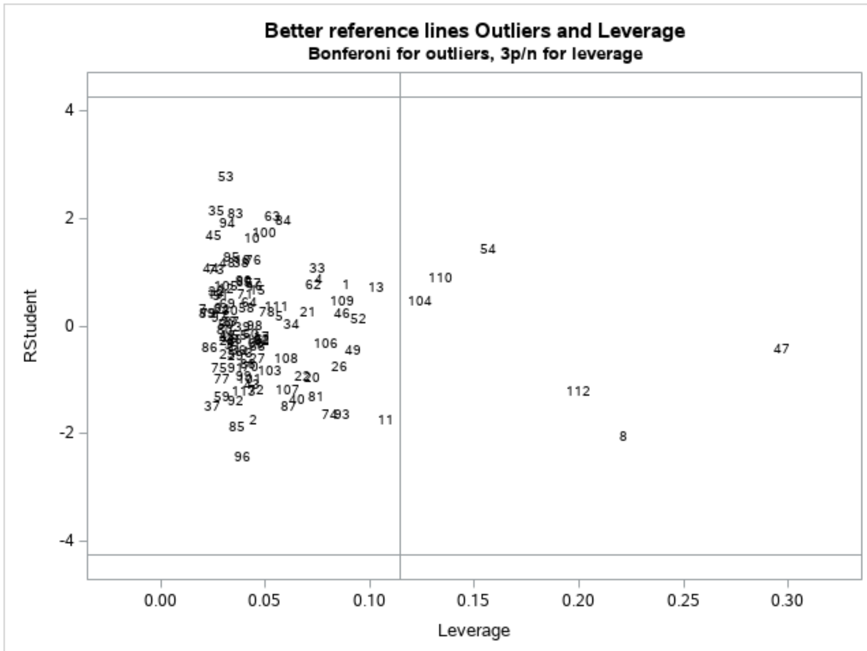


Figure 8: Final Model Leverage Plot

8 Linear Regression Results

We then combined our train and test data and obtained the following equation for our chosen model:

$$\begin{aligned}
\widehat{log_SalePrice} = {} & 7.26245 - 0.02526 * sqrt_Age + 0.09946 * log_LotArea \\
& + 0.07397 * OverallQual + 0.04634 * OverallCond \\
& + 0.00035919 * YearRemodAdd - 0.01293 * BsmtCond \\
& + 0.00985 * BsmtExposure + 0.00282 * BsmtFinSFOne_sqrt \\
& + 0.00371 * TotalBsmtSF_sqrt + 0.2964 * log_FirstFlrSF \\
& + 0.02106 * BsmtFullBath + 0.04736 * FullBath + 0.03995 * HalfBath \\
& - 0.00981 * BedroomAbvGr - 0.08953 * KitchenAbvGr + 0.02928 * KitchenQual \\
& + 0.01383 * TotRmsAbvGrd + 0.03425 * Fireplaces \\
& + 0.00734 * sqrt_GarageArea + 0.0009983 * sqrt_WoodDeckSF \\
& + 0.00149 * sqrt_OpenPorchSF + 0.03097 * GarageType_BuiltIn \\
& + 0.06242 * GarageType_nan + 0.05169 * LandContour_HLS \\
& + 0.02594 * LotConfig_CulDSa + 0.01077 * LotConfig_Inside \\
& - 0.05983 * HouseStyle_OnePointFiveUnf - 0.11818 * HouseStyle_OneStory \\
& - 0.09793 * HouseStyle_SFoyer - 0.10634 * HouseStyle_SLvl \\
& + 0.11241 * HouseStyle_TwoPointFiveUnf + 0.02219 * HouseStyle_TwoStory \\
& + 0.05917 * RoofStyle_Gambr + 0.01274 * RoofStyle_Hip \\
& + 0.10957 * ExteriorFirst_BrkComm + 0.09394 * ExteriorFirst_BrkFace \\
& + 0.02689 * ExteriorFirst_MetalSd + 0.42698 * ExteriorFirst_PreCast \\
& + 0.02462 * Foundation_PConc + 0.0381 * CentralAir_Y
\end{aligned}$$

We will now interpret the coefficients of two of our explanatory variables:

- $sqrt\_Age$: As the square root of the age of the house increases by one year, the $log\_SalePrice$ is expected to decrease, on average, by 0.02526 dollars.

- $OverallQual$: As the overall quality of the home increases by one point on a ten point scale, the $log\_SalePrice$ is expected to increase by 0.07397 dollars.

We ran our selected model on the test set and calculated the Mean Square Prediction Error. We found that it was 0.00929, only slightly higher than the training MSE of 0.00893. This shows that our model generalizes well to unseen data.

Then we checked how our model did on observations 335 and 295. Table 3 summarizes our predictions as well as the 95% prediction intervals with a Bonferroni adjustment.

Observation	$\widehat{log\_SalePrice}$	$\widehat{SalePrice}$	$SalePrice$	Residual	Bonferroni Lower	Bonferroni Upper
335	11.9550	155593.17	157900	2306.83	125605.33	192817.61
295	11.9972	162299.72	215000	-52700.28	130457.14	200606.39

Table 2: Estimated Results for Observations 335 and 295

Our model makes a very accurate prediction on observation 335 and our prediction interval contains the true $SalePrice$. Conversely, our model makes an inaccurate prediction on observation 295 and the prediction interval doesn't contain the true $SalePrice$. We know that our model may occasionally be inaccurate on individual observations but overall it has a high R^2_{adj} value.

The R^2_{adj} value on our final model was 0.9372, meaning that approximately 93.72% of the variance in $log\_SalePrice$ is accounted for by our chosen model variables. This is a great model and is worth our time. The variables in our final model are useful in predicting the sale price of homes.

Figure 9 contains a plot of the predicted $log\_SalePrice$ against the actual $log\_SalePrice$. along with the regression line. We can see that the points follow the line quite closely, showing the high accuracy of our model.

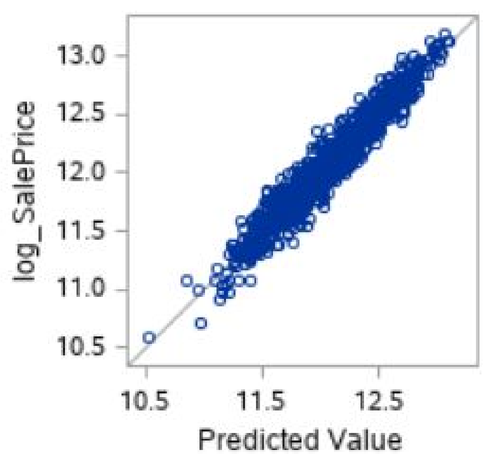


Figure 9: Predicted $log\_SalePrice$ against $log\_SalePrice$

9 Robust Regression

In the prior section we were unable to resolve the issue with several influential points. As a result, we fit an alternate robust regression model on our training set. The MSE on the training dataset was 0.0088504.

Then we used the fitted model to make predictions on the unseen testing dataset. The testing dataset had an MSPR of 0.0141324. Because the MSPR is not ten times greater than the MSE, we can conclude that the model generalizes well to unseen data.

Then, we used our fitted model to make predictions on points 335 and 295, which were withheld from the training dataset. The results are given in the table below:

Observation	$\widehat{log\_SalePrice}$	$\widehat{SalePrice}$	$SalePrice$	Residual
335	11.9525	155204.67	157900	2,695.33
295	12.0024	163145.87	215000	51,854.13

Table 3: Estimated Results for Observations 335 and 295

Similar to the OLS model, we can see that an excellent prediction was made for observation 335 and a poor prediction was made for observation 295.

Then, we combined the training and testing data to fit the best possible model. The fitted equation is given below:

$$\begin{aligned}
\widehat{log_SalePrice} = {} & 7.0932 - 0.024 * sqrt_Age + 0.0964 * log_LotArea + 0.0698 * OverallQuals \\
& + 0.0413 * OverallCond + 0.0004 * YearRemodAdd - 0.0131 * BsmtCond \\
& + 0.0078 * BsmtExposure + 0.0027 * BsmtFinSFOne_sqrt \\
& + 0.0043 * TotalBsmtSF_sqrt + 0.3055 * log_FirstFlrSF \\
& + 0.025 * BsmtFullBath + 0.0436 * FullBath + 0.0366 * HalfBath \\
& - 0.0184 * BedroomAbvGr - 0.0861 * KitchenAbvGr + 0.0288 * KitchenQual \\
& + 0.0158 * TotRmsAbvGrd + 0.0327 * Fireplaces + 0.0077 * sqrt_GarageArea \\
& + 0.0008 * sqrt_WoodDeckSF + 0.0015 * sqrt_OpenPorchSF \\
& + 0.0475 * GarageType_BuiltIn + 0.0695 * GarageType_nan \\
& + 0.0543 * LandContour_HLS + 0.0292 * LotConfig_CulDSa \\
& + 0.013 * LotConfig_Inside - 0.0535 * HouseStyle_OnePointF \\
& - 0.1098 * HouseStyle_OneStory - 0.1063 * HouseStyle_SFoyer \\
& - 0.0873 * HouseStyle_SLvl + 0.0392 * HouseStyle_TwoPointF \\
& + 0.0463 * HouseStyle_TwoStory + 0.0461 * RoofStyle_Gambr \\
& + 0.0164 * RoofStyle_Hip + 0.1035 * ExteriorFirst_BrkCom \\
& + 0.0832 * ExteriorFirst_BrkFac + 0.0251 * ExteriorFirst_MetalS \\
& + 0.4211 * ExteriorFirst_PreCas + 0.0194 * Foundation_PConc \\
& + 0.0386 * CentralAir_Y
\end{aligned}$$

The fitted robust regression model achieved an R^2_{adj} value of 0.7826. The scatterplot of $log\_SalePrice$ against $\widehat{log\_SalePrice}$ is shown in figure 10. We can see that the dots make a line and that the predictions are fairly accurate.

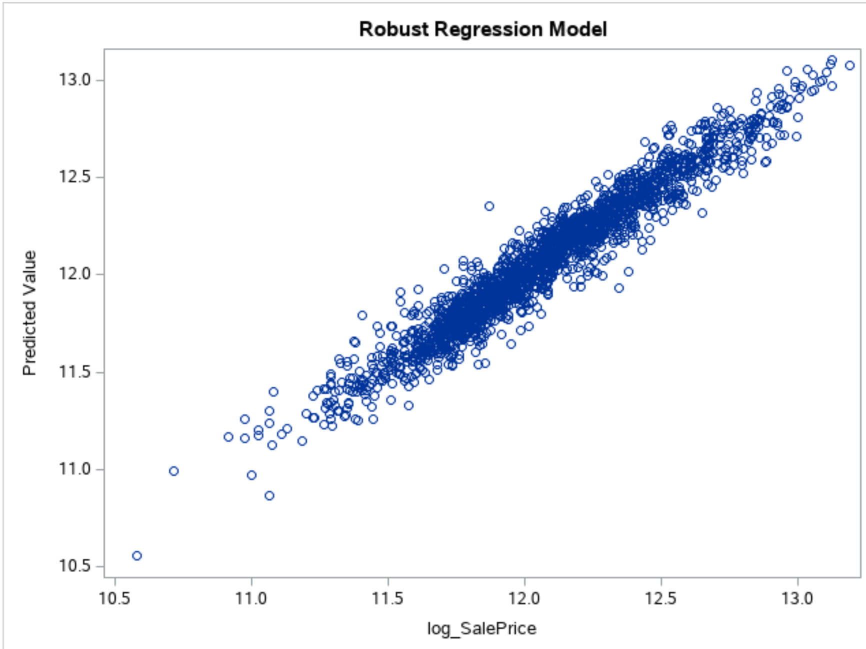


Figure 10: Robust Regression $log\_SalePrice$ against $log\_\widehat{SalePrice}$

10 Final Model Selection

We have now selected an OLS model and a robust regression model as two candidates for our final model. The robust regression model achieved an impressive R^2_{adj} of 0.7826 while the OLS model achieved an even higher R^2_{adj} of 0.9372.

Because OLS vastly outperformed robust regression, we conclude that the influential points didn't have a very detrimental effect after all. Thus, we select the OLS model as our final model. Refer to the Linear Regression Results section for the exact fitted equation.

11 Conclusion

Each home has its own combination of styles and materials, all of which reflect the personality and tastes of the homeowner. Different blends of home styles creates a fluctuation in a home's sales price. Using data given from the Ames Assessor's Office in Iowa, USA, we were able to determine what variables affect and predict a home's sales price. We were also able to account for a massive 93.72% of the variation in the log of the sale price.

In future research, we'd recommend sampling from other states in the country to have a larger scope of how sales price varies across the United States. The state of the economy at the time of investigation should also be accounted for, since the ever-changing real estate market is bound to affect the sales price. We might be able to improve our predictions further by using such info. This improved model would help ensure that people have as much information as possible before becoming homeowners.

12 References

References

City of Ames, IA. City Assessor — City of Ames, IA. (2021). Retrieved from https://www.cityofames.org/government/departments-divisions-a-h/city-assessor.

Wójciak, E. (2016). The essence of equivalent markets in determining the market value of land property for variable planning factors. Real Estate Management and Valuation, 24(3), 71–82. https://doi.org/10.1515/remav-2016-0022

Hip roofs: Hipped roofs installation costs. Retrieved November 26, 2021 from https://modernize.com/roofs/type/hip.

13 Appendix: Variable Summary

Variable	Type	Description
Order	Discrete	Observation number
PID	Nominal	Parcel identification number - Can be used with city web site for parcel review
MSSubClass	Nominal	Identifies the type of dwelling involved in the sale
MSZoning	Nominal	Identifies the general zoning classification of the sale
LotFrontage	Continuous	Linear feet of street connected to property
LotArea	Continuous	Lot size in square feet
Street	Nominal	Type of road access to property
Alley	Nominal	Type of alley access to property
LotShape	Ordinal	General shape of property
LandContour	Nominal	Flatness of the property
Utilities	Ordinal	Type of utilities available
LotConfig	Nominal	Lot configuration
LandSlope	Ordinal	Slope of property
Neighborhood	Nominal	Physical locations within Ames city limits (map available)
Condition1	Nominal	Proximity to various conditions
Condition2	Nominal	Proximity to various conditions (if more than one is present)
BldgType	Nominal	Type of dwelling
HouseStyle	Nominal	Style of dwelling
OverallQual	Ordinal	Rates the overall material and finish of the house
OverallCond	Ordinal	Rates the overall condition of the house
YearBuilt	Discrete	Original construction date
YearRemod/Add	Discrete	Remodel date (same as construction date if no remodeling or additions)
RoofStyle	Nominal	Type of roof
RoofMatl	Nominal	Roof material
Exterior 1	Nominal	Exterior covering on house
Exterior 2	Nominal	Exterior covering on house (if more than one material)
MasVnr Type	Nominal	Masonry veneer type
MasVnr Area	Continuous	Masonry veneer area in square feet
ExterQual	Ordinal	Evaluates the quality of the material on the exterior
ExterCond	Ordinal	Evaluates the present condition of the material on the exterior
Foundation	Nominal	Type of foundation
BsmtQual	Ordinal	Evaluates the height of the basement
BsmtCond	Ordinal	Evaluates the general condition of the basement
BsmtExposure	Ordinal	Refers to walkout or garden level walls

Variable	Type	Description
BsmtFinTypeOne	Ordinal	Rating of basement finished area
BsmtFinSFOne	Continuous	Type 1 finished square feet
BsmtFinTypeTwo	Ordinal	Rating of basement finished area (if multiple types)
BsmtFinSFTwo	Continuous	Type 2 finished square feet
BsmtUnfSF	Continuous	Unfinished square feet of basement area
TotalBsmtSF	Continuous	Total square feet of basement area
Heating	Nominal	Type of heating
HeatingQC	Ordinal	Heating quality and condition
CentralAir	Nominal	Central air conditioning
Electrical	Ordinal	Electrical system
1stFlrSF	Continuous	First Floor square feet
2ndFlrSF	Continuous	Second floor square feet
LowQualFin SF	Continuous	Low quality finished square feet (all floors)
GrLivArea	Continuous	Above grade (ground) living area square feet
BsmtFullBath	Discrete	Basement full bathrooms
BsmtHalfBath	Discrete	Basement half bathrooms
FullBath	Discrete	Full bathrooms above grade
HalfBath	Discrete	Half baths above grade
Bedroom	Discrete	Bedrooms above grade (does NOT include basement bedrooms)
Kitchen	Discrete	Kitchens above grade
KitchenQual	Ordinal	Kitchen quality
TotRmsAbvGrd	Discrete	Total rooms above grade (does not include bathrooms)
Functional	Ordinal	Home functionality (Assume typical unless deductions are warranted)
Fireplaces	Discrete	Number of fireplaces
FireplaceQu	Ordinal	Fireplace quality
GarageType	Nominal	Garage location
GarageYrBlt	Discrete	Year garage was built
GarageFinish	Ordinal	Interior finish of the garage
GarageCars	Discrete	Size of garage in car capacity
GarageArea	Continuous	Size of garage in square feet
GarageQual	Ordinal	Garage quality
GarageCond	Ordinal	Garage condition
PavedDrive	Ordinal	Paved driveway
WoodDeck SF	Continuous	Wood deck area in square feet
OpenPorch SF	Continuous	Open porch area in square feet
EnclosedPorch	Continuous	Enclosed porch area in square feet
3-SsnPorch	Continuous	Three season porch area in square feet
ScreenPorch	Continuous	Screen porch area in square feet
PoolArea	Continuous	Pool area in square feet
PoolQC	Ordinal	Pool quality
Fence	Ordinal	Fence quality
MiscFeature	Nominal	Miscellaneous feature not covered in other categories

Variable	Type	Description
MiscVal	Continuous	Value of miscellaneous feature
MoSold	Discrete	Month Sold (MM)
YrSold	Discrete	Year Sold (YYYY)
SaleType	Nominal	Type of sale
SaleCondition	Nominal	Condition of sale
SalePrice	Continuous	Sale price

14 Appendix: Code

```
/* Final Project SAS File: */

/* You need your own import statement here, as we all have
   different file paths. Make sure you name your dataset
       housing     . */

/* We create some additional features and withhold on some
   observations. */
data housing; set housing;
age = YrSold - YearBuilt;
ageRemodel = YrSold - YearRemodAdd;
ID = order;
/* These two lines  d o n t  work in SAS. You have to change the
   quotation marks. */
If order = 335 then SalePrice =     .    ;
If order = 295 then SalePrice =     .    ;
run;

/* Take a look at the data. */
proc print data=housing(obs=5);
run;

/* Our first regression includes all variables (except ID) and
   checks for multicollinearity. The list of variables can be
   viewed here. */
proc reg data=housing;
model SalePrice = age ageRemodel LotArea LotShape OverallQual
   OverallCond YearBuilt YearRemodAdd BsmtQual BsmtCond
   BsmtExposure BsmtFinSFOne TotalBsmtSF FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd
   Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF YrSold
   GarageType_Attchd GarageType_Basment GarageType_BuiltIn
   GarageType_CarPort GarageType_Detchd GarageType_nan Street_Pave
    LandContour_HLS LandContour_Low LandContour_Lvl
   LotConfig_CulDSa LotConfig_FRThree LotConfig_FRTwo
   LotConfig_Inside BldgType_Duplex BldgType_OneFam BldgType_Twnhs
    BldgType_TwnhsE HouseStyle_OnePointFiveUnf HouseStyle_OneStory
    HouseStyle_SFoyer HouseStyle_SLvl HouseStyle_TwoPointFiveFin
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gable
   RoofStyle_Gambr RoofStyle_Hip RoofStyle_Mansa RoofStyle_Shed
   RoofMatl_Membran RoofMatl_Metal RoofMatl_TarGrv
```

```
   RoofMatl_WdShake RoofMatl_WdShngl ExteriorFirst_AsphShn
   ExteriorFirst_BrkComm ExteriorFirst_BrkFace
   ExteriorFirst_CemntBd ExteriorFirst_HdBoard
   ExteriorFirst_ImStucc ExteriorFirst_MetalSd
   ExteriorFirst_Plywood ExteriorFirst_PreCast
   ExteriorFirst_Stucco ExteriorFirst_VinylSd ExteriorFirst_WdSdng
    ExteriorFirst_WdShing Foundation_CBlock Foundation_PConc
   Foundation_Slab Foundation_Stone Foundation_Wood Heating_GasW
   Heating_Grav Heating_OthW Heating_Wall CentralAir_Y / vif
   collin;
run;

/*

/************************ Initial Model Diagnostics &
   Transformations***************************/

We've systematically used VIF to remove the following variables
   from the model in this order:
YearBuilt, ageRemodel, RoofStyle_Gable, GarageType_Attchd,
   ExteriorFirst_VinylSd, BldgType_OneFam, SecondFlrSF, GrLivArea,
    GarageCars, BsmtQual
The resulting model is given below. All VIF values are now below
   10 and no collinearity diagnostics are above 0.5.
Check outlier diagnostics for the model before transformations.
*/
ods graphics on / imagemap=on;
proc reg data=housing plots(label) = (DFFITS DFBETAS);
        id ID;
        model SalePrice = age LotArea LotShape OverallQual
           OverallCond YearRemodAdd BsmtCond BsmtExposure
           BsmtFinSFOne TotalBsmtSF FirstFlrSF LowQualFinSF
           BsmtFullBath BsmtHalfBath FullBath HalfBath
           BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd
           Fireplaces GarageArea WoodDeckSF OpenPorchSF YrSold
           GarageType_Basment GarageType_BuiltIn
           GarageType_CarPort GarageType_Detchd GarageType_nan
           Street_Pave LandContour_HLS LandContour_Low
           LandContour_Lvl LotConfig_CulDSa LotConfig_FRThree
           LotConfig_FRTwo LotConfig_Inside BldgType_Duplex
           BldgType_Twnhs BldgType_TwnhsE
           HouseStyle_OnePointFiveUnf HouseStyle_OneStory
           HouseStyle_SFoyer HouseStyle_SLvl
           HouseStyle_TwoPointFiveFin HouseStyle_TwoPointFiveUnf
           HouseStyle_TwoStory RoofStyle_Gambr RoofStyle_Hip
```

```
            RoofStyle_Mansa RoofStyle_Shed RoofMatl_Membran
            RoofMatl_Metal RoofMatl_TarGrv RoofMatl_WdShake
            RoofMatl_WdShngl ExteriorFirst_AsphShn
            ExteriorFirst_BrkComm ExteriorFirst_BrkFace
            ExteriorFirst_CemntBd ExteriorFirst_HdBoard
            ExteriorFirst_ImStucc ExteriorFirst_MetalSd
            ExteriorFirst_Plywood ExteriorFirst_PreCast
            ExteriorFirst_Stucco ExteriorFirst_WdSdng
            ExteriorFirst_WdShing Foundation_CBlock
            Foundation_PConc Foundation_Slab Foundation_Stone
            Foundation_Wood Heating_GasW Heating_Grav Heating_OthW
            Heating_Wall CentralAir_Y / partial;
        ods output outputstatistics=out2;
        output out=out3  cookd=CooksD ;
run; quit;
ods graphics / imagemap=off;

/* Alternative thresholds for influential obs. and outlier
   diagnostics */
data temp;
        p=80; /* p = # beta's (incl. intercept */
        n = 2102; /* n = sample size */
        CooksD20 = finv(.20,p,n-p);
        CooksD50 = finv(.50,p,n-p);
        RStudent95Bonf = tinv((1-.05/2/n),(n-p));
        NegRStudent95Bonf=-1*RStudent95Bonf;
        Leverage3 = 3*p/n;
        DFBETAS = 2/n**0.5; if (n <= 30) then DFBETAS = 1;
        DFFITS = 2*(p/n)**0.5; if (n <= 30) then DFFITS = 1;

proc print data=temp;
        var CooksD20 CooksD50 RStudent95Bonf NegRStudent95Bonf
            Leverage3 DFBETAS DFFITS;
        title1 'Alternative thresholds';
run;

data betterplots; set out2 out3 temp;
run;

/* Make Plot with Better Cook's D reference Lines */
proc sgplot data=betterplots;
        scatter x=ID y=cooksD / markerchar=ID;
        xaxis label = 'Observation ID';
        yaxis label = "Cooks D";
        title1 'Better reference lines for Cooks D';
```

```
        title2 'Reference lines for  20th, and 50th percentiles of
           F distribution';
        refline cooksD20 / axis=Y;  /*20th percentile*/
        refline cooksD50 / axis=Y;  /*50th percentile*/
        yaxis max = 1;
run;

/* Make Plot with Better Studentized Deleted Residuals and
   Leverage Lines */
proc sgplot data=betterplots;
        scatter x=HatDiagonal y=RStudent / markerchar=ID;
        xaxis label = 'Leverage';
        yaxis label = 'Studentized Deleted Residuals';
        title1 'Better reference lines Outliers and Leverage';
        title2 'Bonferoni for outliers, 3p/n for leverage ';
        refline RStudent95Bonf / axis=Y;  /*Upper limit outliers*/
        refline NegRStudent95Bonf / axis=Y;  /*lower limit
           outliers*/
        refline Leverage3 / axis=X;  /* limit leverage */
        yaxis max=4.5 min=-4.5;
Run;

/* We make histograms for all remaining variables to see if they
   are skewed. */
proc univariate data=housing noprint;
hist SalePrice age LotArea LotShape OverallQual OverallCond
   YearRemodAdd BsmtCond BsmtExposure BsmtFinSFOne TotalBsmtSF
   FirstFlrSF LowQualFinSF BsmtFullBath BsmtHalfBath FullBath
   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd
   Fireplaces GarageArea WoodDeckSF OpenPorchSF YrSold
   GarageType_Basment GarageType_BuiltIn GarageType_CarPort
   GarageType_Detchd GarageType_nan Street_Pave LandContour_HLS
   LandContour_Low LandContour_Lvl LotConfig_CulDSa
   LotConfig_FRThree LotConfig_FRTwo LotConfig_Inside
   BldgType_Duplex BldgType_Twnhs BldgType_TwnhsE
   HouseStyle_OnePointFiveUnf HouseStyle_OneStory
   HouseStyle_SFoyer HouseStyle_SLvl HouseStyle_TwoPointFiveFin
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip RoofStyle_Mansa RoofStyle_Shed RoofMatl_Membran
   RoofMatl_Metal RoofMatl_TarGrv RoofMatl_WdShake
   RoofMatl_WdShngl ExteriorFirst_AsphShn ExteriorFirst_BrkComm
   ExteriorFirst_BrkFace ExteriorFirst_CemntBd
   ExteriorFirst_HdBoard ExteriorFirst_ImStucc
   ExteriorFirst_MetalSd ExteriorFirst_Plywood
   ExteriorFirst_PreCast ExteriorFirst_Stucco ExteriorFirst_WdSdng
```

```
    ExteriorFirst_WdShing Foundation_CBlock Foundation_PConc
   Foundation_Slab Foundation_Stone Foundation_Wood Heating_GasW
   Heating_Grav Heating_OthW Heating_Wall CentralAir_Y;
title1 'Predictor Variable Histograms';
Run;

/* Here are some right skewed variables.  It might be wise to
   transform them to eliminate outliers.
LotArea, BsmtFinSFOne, TotalBsmtSF, FirstFlrSF, GarageArea,
   WoodDeckSF, OpenPorchSF, SalePrice, Age  /*

/* We check correlations to find out the minimum value of each
   variable in the above list.  This will help us decide on either
    a log transformation or a square root transformation. The
   following table was output. */

Simple Statistics
Variable
N
Mean
Std Dev
Sum
Minimum
Maximum
LotArea
2104
9434
4042
19849050
1300
39104
BsmtFinSFOne
2104
444.55276
432.45806
935339
0
1972
TotalBsmtSF
2104
1069
411.16294
2248502
0
3206
```

FirstFlrSF
2104
1155
370.72028
2429996
334.00000
3228
GarageArea
2104
485.96055
206.56413
1022461
0
1488
WoodDeckSF
2104
96.32510
121.96344
202668
0
1424
OpenPorchSF
2104
47.14116
59.66910
99185
0
382.00000
SalePrice
2102
187409
74414
393933812
39300
535000
age
2104
32.04420
29.24768
67421
0
128.00000

```
/* Perform transformations on select variables */
data housing; set housing;
        log_LotArea = log(LotArea);
BsmtFinSFOne_sqrt = sqrt(BsmtFinSFOne);
TotalBsmtSF_sqrt = sqrt(TotalBsmtSF);
log_FirstFlrSF = log(FirstFlrSF);
sqrt_GarageArea = sqrt(GarageArea);
sqrt_WoodDeckSF = sqrt(WoodDeckSF);
sqrt_OpenPorchSF = sqrt(OpenPorchSF);
        log_SalePrice = log(SalePrice);
        sqrt_Age = sqrt(Age);
Run;


/* Here is our new model after performing transformations and
   replacing our old variables with the new ones. */
proc reg data=housing;
model log_SalePrice = sqrt_Age log_LotArea LotShape OverallQual
   OverallCond YearRemodAdd BsmtCond BsmtExposure
   BsmtFinSFOne_sqrt TotalBsmtSF_sqrt log_FirstFlrSF LowQualFinSF
   BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
   KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
   sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF YrSold
   GarageType_Basment GarageType_BuiltIn GarageType_CarPort
   GarageType_Detchd GarageType_nan Street_Pave LandContour_HLS
   LandContour_Low LandContour_Lvl LotConfig_CulDSa
   LotConfig_FRThree LotConfig_FRTwo LotConfig_Inside
   BldgType_Duplex BldgType_Twnhs BldgType_TwnhsE
   HouseStyle_OnePointFiveUnf HouseStyle_OneStory
   HouseStyle_SFoyer HouseStyle_SLvl HouseStyle_TwoPointFiveFin
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip RoofStyle_Mansa RoofStyle_Shed RoofMatl_Membran
   RoofMatl_Metal RoofMatl_TarGrv RoofMatl_WdShake
   RoofMatl_WdShngl ExteriorFirst_AsphShn ExteriorFirst_BrkComm
   ExteriorFirst_BrkFace ExteriorFirst_CemntBd
   ExteriorFirst_HdBoard ExteriorFirst_ImStucc
   ExteriorFirst_MetalSd ExteriorFirst_Plywood
   ExteriorFirst_PreCast ExteriorFirst_Stucco ExteriorFirst_WdSdng
    ExteriorFirst_WdShing Foundation_CBlock Foundation_PConc
   Foundation_Slab Foundation_Stone Foundation_Wood Heating_GasW
   Heating_Grav Heating_OthW Heating_Wall CentralAir_Y / vif
   collin;
Run;
```

```
/****************** Checking Model Diagnostics After
   Transformations *********************/

/* We create the macro for running diagnostics. */
%macro resid_num_diag(dataset,datavar,label='requested variable',
   predvar=' ',predlabel='predicted variable'); title; data
   shortfourplotdataset; set &dataset; label &datavar = &label;
   if &datavar ne .; run; proc means data=shortfourplotdataset
   noprint; var &datavar; output out=shortfourplotoutset N=nval
   mean=meanval; data shortfourplotoutset; set shortfourplotoutset
   ; xn=nval; CALL SYMPUT('nval',xn); xmean=meanval; CALL
   SYMPUT('meanval',xmean); %global nvalue; %let nvalue=&nval; %
   global meanvalue; %let meanvalue=&meanval; run; %if &predvar ne
    ' ' %then %do; data shortfourplotdataset; set
   shortfourplotdataset; label &predvar = &predlabel;
   proc sort data=shortfourplotdataset out=shortfourplottemp;
        by descending &predvar; data shortfourplottemp; set
   shortfourplottemp; shortfourplotorder = _n_;
   shortfourplotgroup = 1-(shortfourplotorder < ceil(&nvalue/2));
      proc means data=shortfourplottemp median noprint; by
   shortfourplotgroup; var &datavar; output out=
   shortfourplotouttemp median=medresid; run; data
   shortfourplottempnew; merge shortfourplottemp
   shortfourplotouttemp; by shortfourplotgroup; d = abs(&
   datavar-medresid); run; run; proc ttest data=
   shortfourplottempnew plots=none; class shortfourplotgroup
   ; var d; ods output TTests=shortfourplotBFtemp;
   title1 '(Ignore this nuisance output)'; run; run;
   data shortfourplotBFtemp2; set shortfourplotBFtemp; if
   method = 'Pooled'; t_BF = abs(tValue); BF_pvalue =
   probt; keep t_BF BF_pvalue; proc print data=
   shortfourplotBFTemp2; title1 'P-value for Brown-Forsythe
   test of constant variance'; title2 'in ' &label ' vs. ' &
   predlabel; run; %end; proc sort data=shortfourplotdataset
   out=shortfourplottemp; by &datavar; data shortfourplottemp;
    set shortfourplottemp; n=&nvalue; expectNorm = probit((_n_
   -.375)/(n+.25)); proc corr data=shortfourplottemp; var &
   datavar expectNorm; title1 'Output for correlation test of
   normality of ' &label; title2 '(Check text Table B.6 for
   threshold)'; run; title; quit; %mend resid_num_diag;

/* We check the diagnostics of our model. */
%resid_num_diag(dataset=out1, datavar=resid, label='Residual',
predvar=pred, predlabel='Predicted Value');
```

```
/* Check for outliers. */
ods graphics on / imagemap=on;
proc reg data=housing plots(label) = (DFFITS DFBETAS);
        id ID;
        model log_SalePrice = sqrt_Age log_LotArea LotShape
           OverallQual OverallCond YearRemodAdd BsmtCond
           BsmtExposure BsmtFinSFOne_sqrt TotalBsmtSF_sqrt
           log_FirstFlrSF LowQualFinSF BsmtFullBath BsmtHalfBath
           FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual
            TotRmsAbvGrd Fireplaces sqrt_GarageArea
           sqrt_WoodDeckSF sqrt_OpenPorchSF YrSold
           GarageType_Basment GarageType_BuiltIn
           GarageType_CarPort GarageType_Detchd GarageType_nan
           Street_Pave LandContour_HLS LandContour_Low
           LandContour_Lvl LotConfig_CulDSa LotConfig_FRThree
           LotConfig_FRTwo LotConfig_Inside BldgType_Duplex
           BldgType_Twnhs BldgType_TwnhsE
           HouseStyle_OnePointFiveUnf HouseStyle_OneStory
           HouseStyle_SFoyer HouseStyle_SLvl
           HouseStyle_TwoPointFiveFin HouseStyle_TwoPointFiveUnf
           HouseStyle_TwoStory RoofStyle_Gambr RoofStyle_Hip
           RoofStyle_Mansa RoofStyle_Shed RoofMatl_Membran
           RoofMatl_Metal RoofMatl_TarGrv RoofMatl_WdShake
           RoofMatl_WdShngl ExteriorFirst_AsphShn
           ExteriorFirst_BrkComm ExteriorFirst_BrkFace
           ExteriorFirst_CemntBd ExteriorFirst_HdBoard
           ExteriorFirst_ImStucc ExteriorFirst_MetalSd
           ExteriorFirst_Plywood ExteriorFirst_PreCast
           ExteriorFirst_Stucco ExteriorFirst_WdSdng
           ExteriorFirst_WdShing Foundation_CBlock
           Foundation_PConc Foundation_Slab Foundation_Stone
           Foundation_Wood Heating_GasW Heating_Grav Heating_OthW
           Heating_Wall CentralAir_Y / partial;
        ods output outputstatistics=out2;
        output out=out3  cookd=CooksD ;
run; quit;
ods graphics / imagemap=off;

/* Alternative thresholds for influential obs. and outlier
   diagnostics */
data temp;
        p=80; /* p = # beta's (incl. intercept */
        n = 2102; /* n = sample size */
        CooksD20 = finv(.20,p,n-p);
```

```
        CooksD50 = finv(.50,p,n-p);
        RStudent95Bonf = tinv((1-.05/2/n),(n-p));
        NegRStudent95Bonf=-1*RStudent95Bonf;
        Leverage3 = 3*p/n;
        DFBETAS = 2/n**0.5; if (n <= 30) then DFBETAS = 1;
        DFFITS = 2*(p/n)**0.5; if (n <= 30) then DFFITS = 1;

proc print data=temp;
        var CooksD20 CooksD50 RStudent95Bonf NegRStudent95Bonf
           Leverage3 DFBETAS DFFITS;
        title1 'Alternative thresholds';
run;

data betterplots; set out2 out3 temp;
run;

/* Make Plot with Better Cook's D reference Lines */
proc sgplot data=betterplots;
        scatter x=ID y=cooksD / markerchar=ID;
        xaxis label = 'Observation ID';
        yaxis label = "Cooks D";
        title1 'Better reference lines for Cooks D';
        title2 'Reference lines for  20th, and 50th percentiles of
            F distribution';
        refline cooksD20 / axis=Y;  /*20th percentile*/
        refline cooksD50 / axis=Y;  /*50th percentile*/
        yaxis max = 1;
run;

/* Make Plot with Better Studentized Deleted Residuals and
   Leverage Lines */
proc sgplot data=betterplots;
        scatter x=HatDiagonal y=RStudent / markerchar=ID;
        xaxis label = 'Leverage';
        yaxis label = 'Studentized Deleted Residuals';
        title1 'Better reference lines Outliers and Leverage';
        title2 'Bonferoni for outliers, 3p/n for leverage ';
        refline RStudent95Bonf / axis=Y;  /*Upper limit outliers*/
        refline NegRStudent95Bonf / axis=Y;  /*lower limit
           outliers*/
        refline Leverage3 / axis=X;  /* limit leverage */
        yaxis max=4.5 min=-4.5;
Run;

/******************************** Interactions
```

```
   ****************************************/

/*Creating our interactions*/
data housing; set housing;
        culda_con = LotConfig_CulDSa*LandContour_Lvl;
        Asphshn_hip = ExteriorFirst_AsphShn * RoofStyle_hip;
        GasW_CBlock = Heating_GasW *Foundation_CBlock;
run;

/* Find interaction p-values. */
proc reg data = housing;
        model log_SalePrice = culda_con Asphshn_gambr GasW_CBlock
           ;
Run;

/* Only one of the interactions was significant. We will let model
    selection decide which ones to include. */


/*************************************** Model Selection
   ****************************************/

/* Split the data into training and testing sets. */
proc surveyselect data = housing seed=5000 out=housing2 rate=0.20
   outall;
run;
data train; set housing2;
        if Selected=0;
data test; set housing2;
        if Selected=1;
proc print data = train (obs=5);
        title1 'Training Data Set';
proc print data = test (obs=5);
        title1 'Test Data Set';
run;


/* Backwards Selection. Note: All variables including interactions
    are included below. */
proc reg data=train;
model log_SalePrice = sqrt_Age log_LotArea LotShape OverallQual
   OverallCond YearRemodAdd BsmtCond BsmtExposure
   BsmtFinSFOne_sqrt TotalBsmtSF_sqrt log_FirstFlrSF LowQualFinSF
   BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
   KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
```

```
    sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF YrSold
    GarageType_Basment GarageType_BuiltIn GarageType_CarPort
    GarageType_Detchd GarageType_nan Street_Pave LandContour_HLS
    LandContour_Low LandContour_Lvl LotConfig_CulDSa
    LotConfig_FRThree LotConfig_FRTwo LotConfig_Inside
    BldgType_Duplex BldgType_Twnhs BldgType_TwnhsE
    HouseStyle_OnePointFiveUnf HouseStyle_OneStory
    HouseStyle_SFoyer HouseStyle_SLvl HouseStyle_TwoPointFiveFin
    HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
    RoofStyle_Hip RoofStyle_Mansa RoofStyle_Shed RoofMatl_Membran
    RoofMatl_Metal RoofMatl_TarGrv RoofMatl_WdShake
    RoofMatl_WdShngl ExteriorFirst_AsphShn ExteriorFirst_BrkComm
    ExteriorFirst_BrkFace ExteriorFirst_CemntBd
    ExteriorFirst_HdBoard ExteriorFirst_ImStucc
    ExteriorFirst_MetalSd ExteriorFirst_Plywood
    ExteriorFirst_PreCast ExteriorFirst_Stucco ExteriorFirst_WdSdng
     ExteriorFirst_WdShing Foundation_CBlock Foundation_PConc
    Foundation_Slab Foundation_Stone Foundation_Wood Heating_GasW
    Heating_Grav Heating_OthW Heating_Wall CentralAir_Y culda_con
    Asphshn_hip GasW_CBlock / selection=backward slstay=.10;
Run;



/* Stepwise Selection */
proc reg data=train;
model log_SalePrice = sqrt_Age log_LotArea LotShape OverallQual
    OverallCond YearRemodAdd BsmtCond BsmtExposure
    BsmtFinSFOne_sqrt TotalBsmtSF_sqrt log_FirstFlrSF LowQualFinSF
    BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
    KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
    sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF YrSold
    GarageType_Basment GarageType_BuiltIn GarageType_CarPort
    GarageType_Detchd GarageType_nan Street_Pave LandContour_HLS
    LandContour_Low LandContour_Lvl LotConfig_CulDSa
    LotConfig_FRThree LotConfig_FRTwo LotConfig_Inside
    BldgType_Duplex BldgType_Twnhs BldgType_TwnhsE
    HouseStyle_OnePointFiveUnf HouseStyle_OneStory
    HouseStyle_SFoyer HouseStyle_SLvl HouseStyle_TwoPointFiveFin
    HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
    RoofStyle_Hip RoofStyle_Mansa RoofStyle_Shed RoofMatl_Membran
    RoofMatl_Metal RoofMatl_TarGrv RoofMatl_WdShake
    RoofMatl_WdShngl ExteriorFirst_AsphShn ExteriorFirst_BrkComm
    ExteriorFirst_BrkFace ExteriorFirst_CemntBd
    ExteriorFirst_HdBoard ExteriorFirst_ImStucc
```

```
   ExteriorFirst_MetalSd ExteriorFirst_Plywood
   ExteriorFirst_PreCast ExteriorFirst_Stucco ExteriorFirst_WdSdng
    ExteriorFirst_WdShing Foundation_CBlock Foundation_PConc
   Foundation_Slab Foundation_Stone Foundation_Wood Heating_GasW
   Heating_Grav Heating_OthW Heating_Wall CentralAir_Y culda_con
   Asphshn_hip GasW_CBlock /
selection=stepwise slstay=.1 slentry=.1;
run;

/* All Possible Regressions */
proc reg data=train;
model log_SalePrice = sqrt_Age log_LotArea LotShape OverallQual
   OverallCond YearRemodAdd BsmtCond BsmtExposure
   BsmtFinSFOne_sqrt TotalBsmtSF_sqrt log_FirstFlrSF LowQualFinSF
   BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
   KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
   sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF YrSold
   GarageType_Basment GarageType_BuiltIn GarageType_CarPort
   GarageType_Detchd GarageType_nan Street_Pave LandContour_HLS
   LandContour_Low LandContour_Lvl LotConfig_CulDSa
   LotConfig_FRThree LotConfig_FRTwo LotConfig_Inside
   BldgType_Duplex BldgType_Twnhs BldgType_TwnhsE
   HouseStyle_OnePointFiveUnf HouseStyle_OneStory
   HouseStyle_SFoyer HouseStyle_SLvl HouseStyle_TwoPointFiveFin
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip RoofStyle_Mansa RoofStyle_Shed RoofMatl_Membran
   RoofMatl_Metal RoofMatl_TarGrv RoofMatl_WdShake
   RoofMatl_WdShngl ExteriorFirst_AsphShn ExteriorFirst_BrkComm
   ExteriorFirst_BrkFace ExteriorFirst_CemntBd
   ExteriorFirst_HdBoard ExteriorFirst_ImStucc
   ExteriorFirst_MetalSd ExteriorFirst_Plywood
   ExteriorFirst_PreCast ExteriorFirst_Stucco ExteriorFirst_WdSdng
    ExteriorFirst_WdShing Foundation_CBlock Foundation_PConc
   Foundation_Slab Foundation_Stone Foundation_Wood Heating_GasW
   Heating_Grav Heating_OthW Heating_Wall CentralAir_Y culda_con
   Asphshn_hip GasW_CBlock
/ selection=AdjRSq Cp AIC SBC;
run;

/* This is the model selected by backward selection. */
proc reg data=train;
Model log_SalePrice = sqrt_Age log_LotArea OverallQual OverallCond
    YearRemodAdd BsmtCond BsmtExposure BsmtFinSFOne_sqrt
   TotalBsmtSF_sqrt log_FirstFlrSF BsmtFullBath FullBath HalfBath
   BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
```

```
   sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF
   GarageType_BuiltIn GarageType_nan LandContour_HLS
   LotConfig_CulDSa LotConfig_Inside HouseStyle_OnePointFiveUnf
   HouseStyle_OneStory HouseStyle_SFoyer HouseStyle_SLvl
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip ExteriorFirst_BrkComm ExteriorFirst_BrkFace
   ExteriorFirst_MetalSd ExteriorFirst_PreCast Foundation_PConc
   CentralAir_Y;
Output out=out1 r=resid p=pred;
run;

/* This is the model selected by stepwise selection. */
proc reg data=train;
Model log_SalePrice = sqrt_Age log_LotArea OverallQual OverallCond
    YearRemodAdd BsmtCond BsmtExposure BsmtFinSFOne_sqrt
   TotalBsmtSF_sqrt log_FirstFlrSF BsmtFullBath FullBath HalfBath
   BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
   sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF
   GarageType_BuiltIn GarageType_nan LandContour_HLS
   LotConfig_CulDSa LotConfig_Inside HouseStyle_OnePointFiveUnf
   HouseStyle_OneStory HouseStyle_SFoyer HouseStyle_SLvl
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip ExteriorFirst_BrkComm ExteriorFirst_BrkFace
   ExteriorFirst_MetalSd ExteriorFirst_PreCast Foundation_PConc
   CentralAir_Y;
Output out=out1 r=resid p=pred;
run;

/* This is the model selected by adjusted R-squared using all
   possible regressions. */
proc reg data=train;
Model log_SalePrice = sqrt_Age log_LotArea LotShape OverallQual
   OverallCond YearRemodAdd BsmtCond BsmtExposure
   BsmtFinSFOne_sqrt TotalBsmtSF_sqrt log_FirstFlrSF BsmtFullBath
   FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual
   TotRmsAbvGrd Fireplaces sqrt_GarageArea sqrt_WoodDeckSF
   sqrt_OpenPorchSF YrSold GarageType_BuiltIn GarageType_nan
   LandContour_HLS LotConfig_CulDSa LotConfig_Inside
   BldgType_Duplex HouseStyle_OnePointFiveUnf HouseStyle_OneStory
   HouseStyle_SFoyer HouseStyle_SLvl HouseStyle_TwoPointFiveUnf
   HouseStyle_TwoStory RoofStyle_Gambr RoofStyle_Hip
   RoofStyle_Mansa RoofMatl_Membran RoofMatl_WdShake
   ExteriorFirst_BrkComm ExteriorFirst_BrkFace
   ExteriorFirst_MetalSd ExteriorFirst_Plywood
   ExteriorFirst_PreCast ExteriorFirst_Stucco ExteriorFirst_WdSdng
```

```
    Foundation_CBlock Foundation_PConc Heating_Wall CentralAir_Y;
Output out=out1 r=resid p=pred;
Run;

/*************************** Assumptions Check on Final Model
   **************************************/

/* Check for outliers on selected model. */
ods graphics on / imagemap=on;
proc reg data=housing plots(label) = (DFFITS DFBETAS);
        id ID;
Model log_SalePrice = sqrt_Age log_LotArea OverallQual OverallCond
    YearRemodAdd BsmtCond BsmtExposure BsmtFinSFOne_sqrt
   TotalBsmtSF_sqrt log_FirstFlrSF BsmtFullBath FullBath HalfBath
   BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
   sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF
   GarageType_BuiltIn GarageType_nan LandContour_HLS
   LotConfig_CulDSa LotConfig_Inside HouseStyle_OnePointFiveUnf
   HouseStyle_OneStory HouseStyle_SFoyer HouseStyle_SLvl
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip ExteriorFirst_BrkComm ExteriorFirst_BrkFace
   ExteriorFirst_MetalSd ExteriorFirst_PreCast Foundation_PConc
   CentralAir_Y / partial;
        ods output outputstatistics=out2;
        output out=out3  cookd=CooksD ;
run; quit;
ods graphics / imagemap=off;

/* Alternative thresholds for influential obs. and outlier
   diagnostics */
data temp;
        p=80; /* p = # beta's (incl. intercept */
        n = 2102; /* n = sample size */
        CooksD20 = finv(.20,p,n-p);
        CooksD50 = finv(.50,p,n-p);
        RStudent95Bonf = tinv((1-.05/2/n),(n-p));
        NegRStudent95Bonf=-1*RStudent95Bonf;
        Leverage3 = 3*p/n;
        DFBETAS = 2/n**0.5; if (n <= 30) then DFBETAS = 1;
        DFFITS = 2*(p/n)**0.5; if (n <= 30) then DFFITS = 1;

proc print data=temp;
        var CooksD20 CooksD50 RStudent95Bonf NegRStudent95Bonf
           Leverage3 DFBETAS DFFITS;
        title1 'Alternative thresholds';
```

```
run;

data betterplots; set out2 out3 temp;
run;

/* Make Plot with Better Cook's D reference Lines */
proc sgplot data=betterplots;
        scatter x=ID y=cooksD / markerchar=ID;
        xaxis label = 'Observation ID';
        yaxis label = "Cooks D";
        title1 'Better reference lines for Cooks D';
        title2 'Reference lines for  20th, and 50th percentiles of
           F distribution';
        refline cooksD20 / axis=Y;  /*20th percentile*/
        refline cooksD50 / axis=Y;  /*50th percentile*/
        yaxis max = 1;
run;

/* Make Plot with Better Studentized Deleted Residuals and
   Leverage Lines */
proc sgplot data=betterplots;
        scatter x=HatDiagonal y=RStudent / markerchar=ID;
        xaxis label = 'Leverage';
        yaxis label = 'Studentized Deleted Residuals';
        title1 'Better reference lines Outliers and Leverage';
        title2 'Bonferoni for outliers, 3p/n for leverage ';
        refline RStudent95Bonf / axis=Y;  /*Upper limit outliers*/
        refline NegRStudent95Bonf / axis=Y;  /*lower limit
           outliers*/
        refline Leverage3 / axis=X;  /* limit leverage */
        yaxis max=4.5 min=-4.5;
Run;

/*********************************** Evaluate our Chosen Model
   ***********************************/

/* We choose the backward selection model as the best model.  Fit
   the model and store its parameters. */
proc reg data=train plots=none;
Model log_SalePrice = sqrt_Age log_LotArea OverallQual OverallCond
    YearRemodAdd BsmtCond BsmtExposure BsmtFinSFOne_sqrt
   TotalBsmtSF_sqrt log_FirstFlrSF BsmtFullBath FullBath HalfBath
   BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
   sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF
   GarageType_BuiltIn GarageType_nan LandContour_HLS
```

```
   LotConfig_CulDSa LotConfig_Inside HouseStyle_OnePointFiveUnf
   HouseStyle_OneStory HouseStyle_SFoyer HouseStyle_SLvl
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip ExteriorFirst_BrkComm ExteriorFirst_BrkFace
   ExteriorFirst_MetalSd ExteriorFirst_PreCast Foundation_PConc
   CentralAir_Y;
store backwardSelection;
title 'Backward Selection Model';
run;

/* Make predictions on the test set. */
proc plm restore=backwardSelection;
        show parameters;  /*display parameters; double check right
            variables*/
        score data=test out=Preds1 predicted;
run;

/* Evaluate accuracy on the test set. */
data Errors; set Preds1;
        sqerror = (log_SalePrice - predicted)**2;
run;

/* Print mean square prediction error. */
proc means data=errors mean;
        var sqerror;
        title1 'Mean Square Prediction Error (MSPR)';
Run;

/* Evaluating predictions on previously withheld points 335 and
   295 and prediction intervals */
proc reg data=housing noprint;
model log_SalePrice = sqrt_Age log_LotArea OverallQual OverallCond
    YearRemodAdd BsmtCond BsmtExposure BsmtFinSFOne_sqrt
   TotalBsmtSF_sqrt log_FirstFlrSF BsmtFullBath FullBath HalfBath
   BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
   sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF
   GarageType_BuiltIn GarageType_nan LandContour_HLS
   LotConfig_CulDSa LotConfig_Inside HouseStyle_OnePointFiveUnf
   HouseStyle_OneStory HouseStyle_SFoyer HouseStyle_SLvl
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip ExteriorFirst_BrkComm ExteriorFirst_BrkFace
   ExteriorFirst_MetalSd ExteriorFirst_PreCast Foundation_PConc
   CentralAir_Y;
output out=out1 p=Yhat stdi=seYhatnew;
/* KEY: stdi is SE of individual prediction */
```

```
data out1; set out1;
alpha = 0.05; /* 1-alpha is simult. pred. level */
p = 40; /* # of beta's (including intercept) */
n = 1681; /* sample size */
g = 2; /* number of simultaneous intervals */
S = sqrt(g*finv(1-alpha,g,n-p)); /* Scheffe crit val */
t = tinv(1-alpha/(2*g),n-p); /* Bonf. crit. val. */
S_upper = Yhat + S*seYhatnew;
S_lower = Yhat - S*seYhatnew;
B_upper = Yhat + t*seYhatnew;
B_lower = Yhat - t*seYhatnew;

proc print data=out1;where order = 335 or order = 295;

var order log_SalePrice Yhat seYhatnew S_lower S_upper
B_lower B_upper;
title1 'Simultaneous 95% interval estimation of
individual prediction';
title2 'at two X-levels, using Scheffe and Bonferroni';
run;


/******************************************* Robust Regression
   ************************************/

/* We evaluate our robust regression model.  Fit the model and
   output its predictions. */
proc robustreg data=train method=M (wf=bisquare) plots=none;
model log_SalePrice = sqrt_Age log_LotArea OverallQual OverallCond
    YearRemodAdd BsmtCond BsmtExposure BsmtFinSFOne_sqrt
   TotalBsmtSF_sqrt log_FirstFlrSF BsmtFullBath FullBath HalfBath
   BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
   sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF
   GarageType_BuiltIn GarageType_nan LandContour_HLS
   LotConfig_CulDSa LotConfig_Inside HouseStyle_OnePointFiveUnf
   HouseStyle_OneStory HouseStyle_SFoyer HouseStyle_SLvl
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip ExteriorFirst_BrkComm ExteriorFirst_BrkFace
   ExteriorFirst_MetalSd ExteriorFirst_PreCast Foundation_PConc
   CentralAir_Y;
output out=out1 p=pred;
title 'Robust Regression Model';
run;

/* Calculate the MSE on the training set. */
```

```
data out1; set out1;
mse = (log_SalePrice - pred)**2;
run;

/* Print the MSE for the training set. */
proc means data=out1;
vars mse;
title 'Robust Regression Training MSE';
run;

/* Now we calculate the MSPR for the testing dataset using our
   estimated equation. */
data test; set test;
pred = 7.0263 - 0.0238 * sqrt_Age + 0.0947 * log_LotArea + 0.0696
   * OverallQual + 0.0413 * OverallCond + 0.0005 * YearRemodAdd -
   0.0172 * BsmtCond + 0.007 * BsmtExposure + 0.0027 *
   BsmtFinSFOne_sqrt + 0.0041 * TotalBsmtSF_sqrt + 0.3136 *
   log_FirstFlrSF + 0.0251 * BsmtFullBath + 0.0492 * FullBath +
   0.0411 * HalfBath - 0.0154 * BedroomAbvGr - 0.0881 *
   KitchenAbvGr + 0.0283 * KitchenQual + 0.0155 * TotRmsAbvGrd +
   0.0306 * Fireplaces + 0.0073 * sqrt_GarageArea + 0.0009 *
   sqrt_WoodDeckSF + 0.0018 * sqrt_OpenPorchSF + 0.0441 *
   GarageType_BuiltIn + 0.0616 * GarageType_nan + 0.0536 *
   LandContour_HLS + 0.0292 * LotConfig_CulDSa + 0.0158 *
   LotConfig_Inside - 0.0454 * HouseStyle_OnePointFiveUnf - 0.1044
    * HouseStyle_OneStory - 0.0979 * HouseStyle_SFoyer - 0.0752 *
   HouseStyle_SLvl + 0.0327 * HouseStyle_TwoPointFiveUnf + 0.046 *
    HouseStyle_TwoStory + 0.0692 * RoofStyle_Gambr + 0.0216 *
   RoofStyle_Hip + 0.098 * ExteriorFirst_BrkComm + 0.0819 *
   ExteriorFirst_BrkFace + 0.0271 * ExteriorFirst_MetalSd - 0.0 *
   ExteriorFirst_PreCast + 0.0177 * Foundation_PConc + 0.0391 *
   CentralAir_Y;
mspr = (log_SalePrice - pred)**2;
run;

/* Print the MSPR for the testing set. */
proc means data=test;
vars mspr;
title 'Robust Regression Testing MSPR';
run;

/* Output the predictions for the points that were previously
   withheld. */
proc print data=out1;
where order=335 or order=295;
```

```
var order pred;
run;

/* Use all of our data to output an estimated equation. */
proc robustreg data=housing method=M (wf=bisquare);
model log_SalePrice = sqrt_Age log_LotArea OverallQual OverallCond
    YearRemodAdd BsmtCond BsmtExposure BsmtFinSFOne_sqrt
   TotalBsmtSF_sqrt log_FirstFlrSF BsmtFullBath FullBath HalfBath
   BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces
   sqrt_GarageArea sqrt_WoodDeckSF sqrt_OpenPorchSF
   GarageType_BuiltIn GarageType_nan LandContour_HLS
   LotConfig_CulDSa LotConfig_Inside HouseStyle_OnePointFiveUnf
   HouseStyle_OneStory HouseStyle_SFoyer HouseStyle_SLvl
   HouseStyle_TwoPointFiveUnf HouseStyle_TwoStory RoofStyle_Gambr
   RoofStyle_Hip ExteriorFirst_BrkComm ExteriorFirst_BrkFace
   ExteriorFirst_MetalSd ExteriorFirst_PreCast Foundation_PConc
   CentralAir_Y;
output out=out1 p=pred;
title 'Robust Regression Model';
run;

/* Create a scatterplot of our predictions. */
proc sgplot data=out1;
scatter x=log_SalePrice y=pred;
run;
```