SAS® GLOBAL FORUM 2018

USERS PROGRAM

Multiple Regression Diagnostics in SAS®: Predicting House Sale Prices

Karen Bui

University of Central Florida

April 8 – 11 | Denver, CO

#SASGF

Multiple Regression Diagnostics in SAS®: Predicting House Sale Prices

Karen Bui, University of Central Florida

Faculty Advisor: Dr. Daoji Li of UCF

ABSTRACT

Being able to predict a variable of interest through multiple regressions is a powerful tool, but how can you tell if the model you have chosen is actually useful? And if there is a potentially better model, how do you know where to begin with adjustments? In this exploration, we look into the different possible analyses in judging a regression model's overall effectiveness using data on house sale prices in King County, Washington. There were a total of 21,613 observations, with twelve potential independent variables and house sale price as the target prediction. Using various SAS® procedures, we are able to examine the significance of a preliminary full linear model, obtain a reduced model through elimination of insignificant predictors, and inspect various diagnostic plots and specific output values for variable collinearity, sample normality, and outliers. All analysis was completed in SAS® Studio.

Our technical aim was to be able to accurately predict house sale prices using a multiple regression model with various characteristics of the resident property acting as the predictor variables, but the bigger goal for this study was to understand the specific methods used to verify sample assumptions made for regression analysis, which involve assessing predictor interaction (collinearity), resemblance of the sample to a normal distribution, and outlier impact to validate model adequacy and improve the prediction tool.

DATA

This dataset comes from Kaggle, which is a website used primarily as a "platform for predictive modelling and analytics competitions" with an abundance of datasets uploaded by companies and other users.
This data consists of house sale prices for properties sold between May 2014 and May 2015 in King County, WA (including Seattle).

The target variable (Y) was the **house sale price,** while twelve predictors (X's) from the data file were considered in the full model; these predictors were various characteristics of the sold property, as listed below:

- Number of **bedrooms**
- Number of **bathrooms**
- **Square footage** of **home**
- **Square footage** of entire **lot**
- Number of **floors**
- Presence of **waterfront** view
- Times the home was **viewed**
- Overall **condition** of the home
- Overall **grade** of the home (by King County grading standards)
- **Square footage** of home **excluding basement**
- **Square footage** of **basement**
- **Year** the home was built

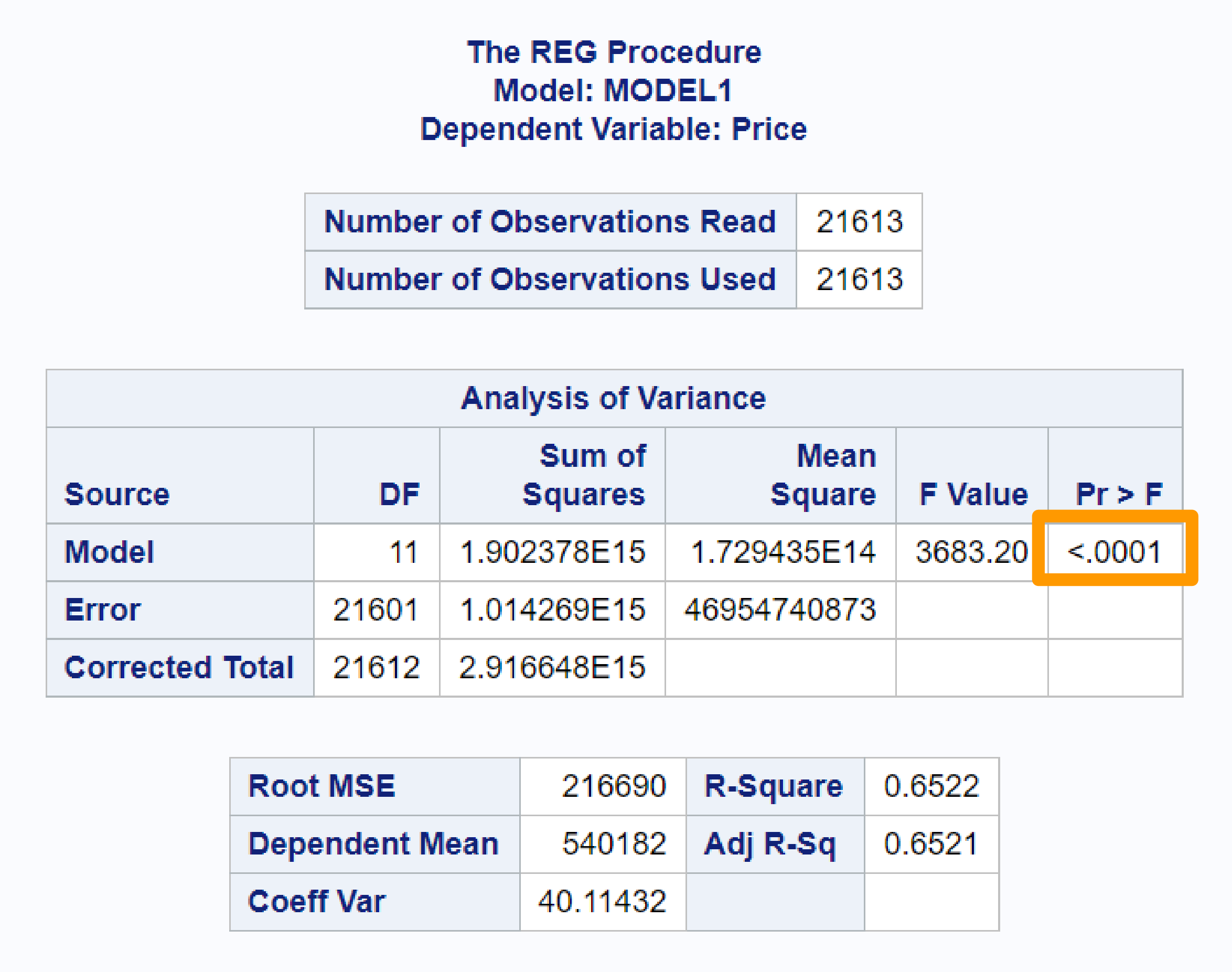
There were a total of 21,613 observations in this dataset.

FULL MODEL EVALUATION

Our first step is to check the significance of the full linear model as well as to briefly examine the significance of each of the predictors to, jointly, give us a better idea of whether these variables are appropriate to use in predicting the house sale price. This is very easily done using a SAS procedure statement called PROC REG; we can specify the model with price as the dependent variable and all twelve previously mentioned home characteristics as the independent variables, as shown below:

```
proc reg data=HouseSales plots(maxpoints=none);
   model Price = Bedrooms Bathrooms Sqft_Home Sqft_Lot Floors Waterfront Times_Viewed
   Condition Grade Sqft_Basement Year;
   *will exclude Sqft_Above since it is the difference of Sqft_Home and Sqft_Basement;
run;
```

*Note that since this particular dataset exceeds 5000 points, ODS graphics will be suppressed, so we need to specify the 'PLOTS' function immediately after the PROC REG statement, with 'maxpoints' set to 'none' to show all 21,613 data points for all graphs.



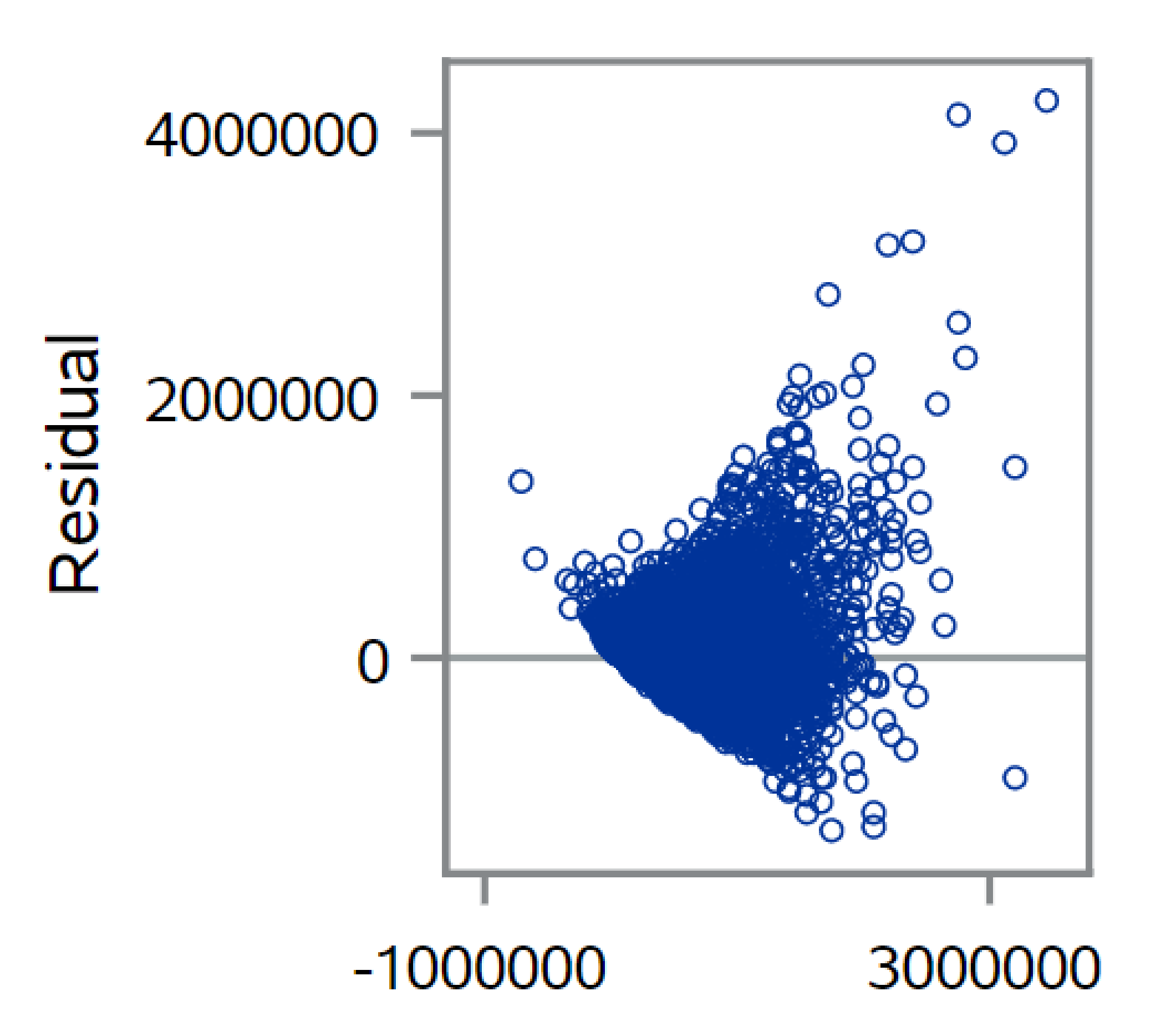
The REG Procedure
Model: MODEL1
Dependent Variable: Price

Number of Observations Read	21613
Number of Observations Used	21613

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1.902378E15	1.729435E14	3683.20	<.0001
Error	21601	1.014269E15	46954740873		
Corrected Total	21612	2.916648E15			

Root MSE	216690	R-Square	0.6522
Dependent Mean	540182	Adj R-Sq	0.6521
Coeff Var	40.11432		



Though our Global F Test for model usefulness is significant, upon a quick examination of the residual graph (first plot under *Fit Diagnostics for Price*) from the PROC REG output, we can see that the residuals exhibit a fanned pattern. A more technical term for this is "heteroscedasticity", which means that variability in the residuals (difference between the observed value and predicted value) grows as the prediction scale grows. Thus, our linear model is not appropriate; however, we can utilize a common solution to this type of problem, which is to perform a logarithmic transformation on Price.

Multiple Regression Diagnostics in SAS®: Predicting House Sale Prices

Karen Bui, University of Central Florida

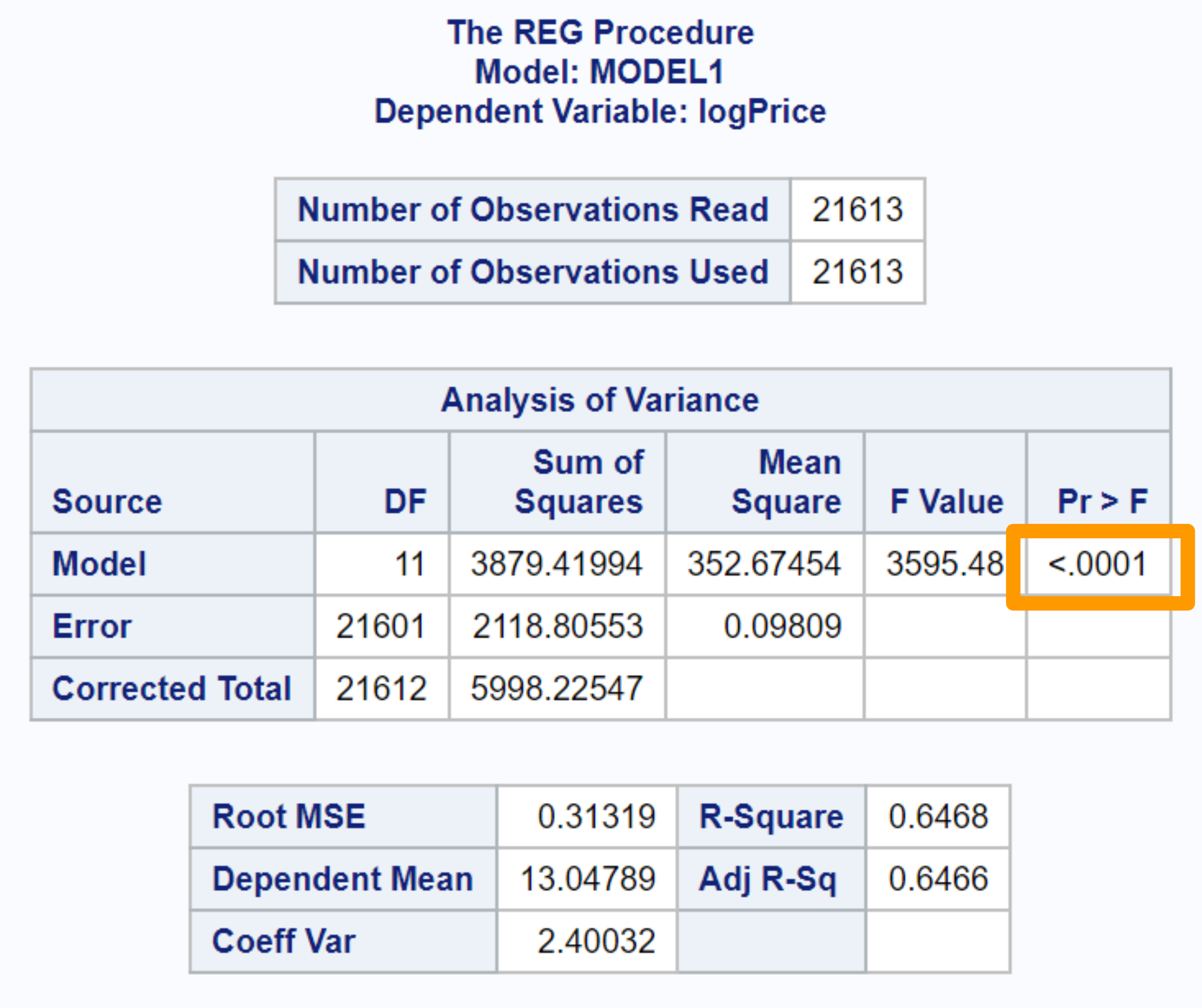
Faculty Advisor: Dr. Daoji Li of UCF

FULL MODEL EVALUATION CONT.

To do this, we can create a new dataset, copy over the original dataset, and use the LOG function to calculate the logarithm of Price. Afterwards, we can simply run the same PROC REG statement as before, this time, with logPrice as the dependent variable.

```
data HouseSalesLog;
   set HouseSales;
   logPrice=log(Price);
   run;
   /*CHECK LOG TRANSFORMATION*/
   proc reg data=HouseSalesLog plots(maxpoints=none);
   model logPrice = Bedrooms Bathrooms...;
run;
```

Once again, we can see that our model is considered useful; furthermore, a closer look at the p-values for our predictors shows that **all of them appear to be significant in predicting house sale price EXCEPT for Sqft_Lot**, suggesting that this could be a variable to eliminate.



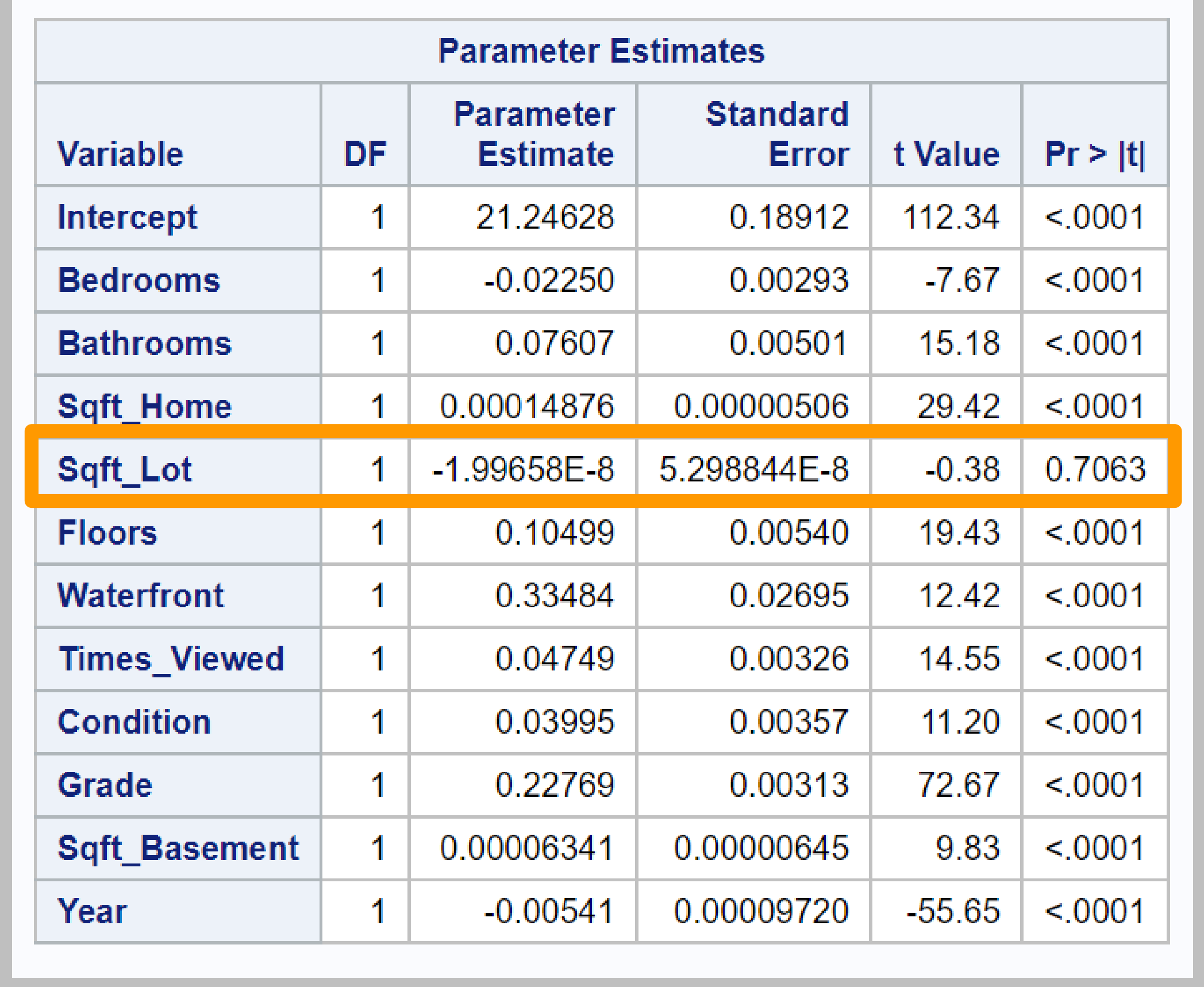
The REG Procedure
Model: MODEL1
Dependent Variable: logPrice

Number of Observations Read	21613
Number of Observations Used	21613

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	3879.41994	352.67454	3595.48	<.0001
Error	21601	2118.80553	0.09809		
Corrected Total	21612	5998.22547			

Root MSE	0.31319	R-Square	0.6468
Dependent Mean	13.04789	Adj R-Sq	0.6466
Coeff Var	2.40032		



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21.24628	0.18912	112.34	<.0001
Bedrooms	1	-0.02250	0.00293	-7.67	<.0001
Bathrooms	1	0.07607	0.00501	15.18	<.0001
Sqft_Home	1	0.00014876	0.00000506	29.42	<.0001
Sqft_Lot	1	-1.99658E-8	5.298844E-8	-0.38	0.7063
Floors	1	0.10499	0.00540	19.43	<.0001
Waterfront	1	0.33484	0.02695	12.42	<.0001
Times_Viewed	1	0.04749	0.00326	14.55	<.0001
Condition	1	0.03995	0.00357	11.20	<.0001
Grade	1	0.22769	0.00313	72.67	<.0001
Sqft_Basement	1	0.00006341	0.00000645	9.83	<.0001
Year	1	-0.00541	0.00009720	-55.65	<.0001

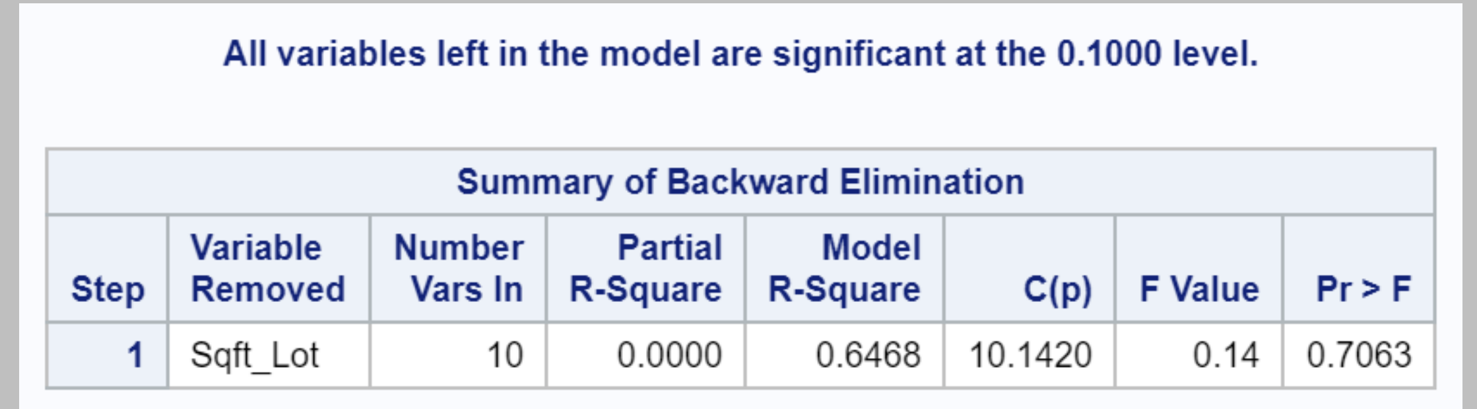
THE STEPWISE PROCEDURE

We can now use PROC STEPWISE to remove insignificant variables and obtain a reduced model, as shown in the following code:

```
proc stepwise;
   model logPrice = Bedrooms Bathrooms Sqft_Home Sqft_Lot Floors Waterfront Times_Viewed
   Condition Grade Sqft_Basement Year/ stepwise forward backward;
run;
```

There are three different selection methods in PROC STEPWISE, that gives us three different suggestions for a reduced model:

backward selection, which starts with the complete model and removes one independent variable at a time;

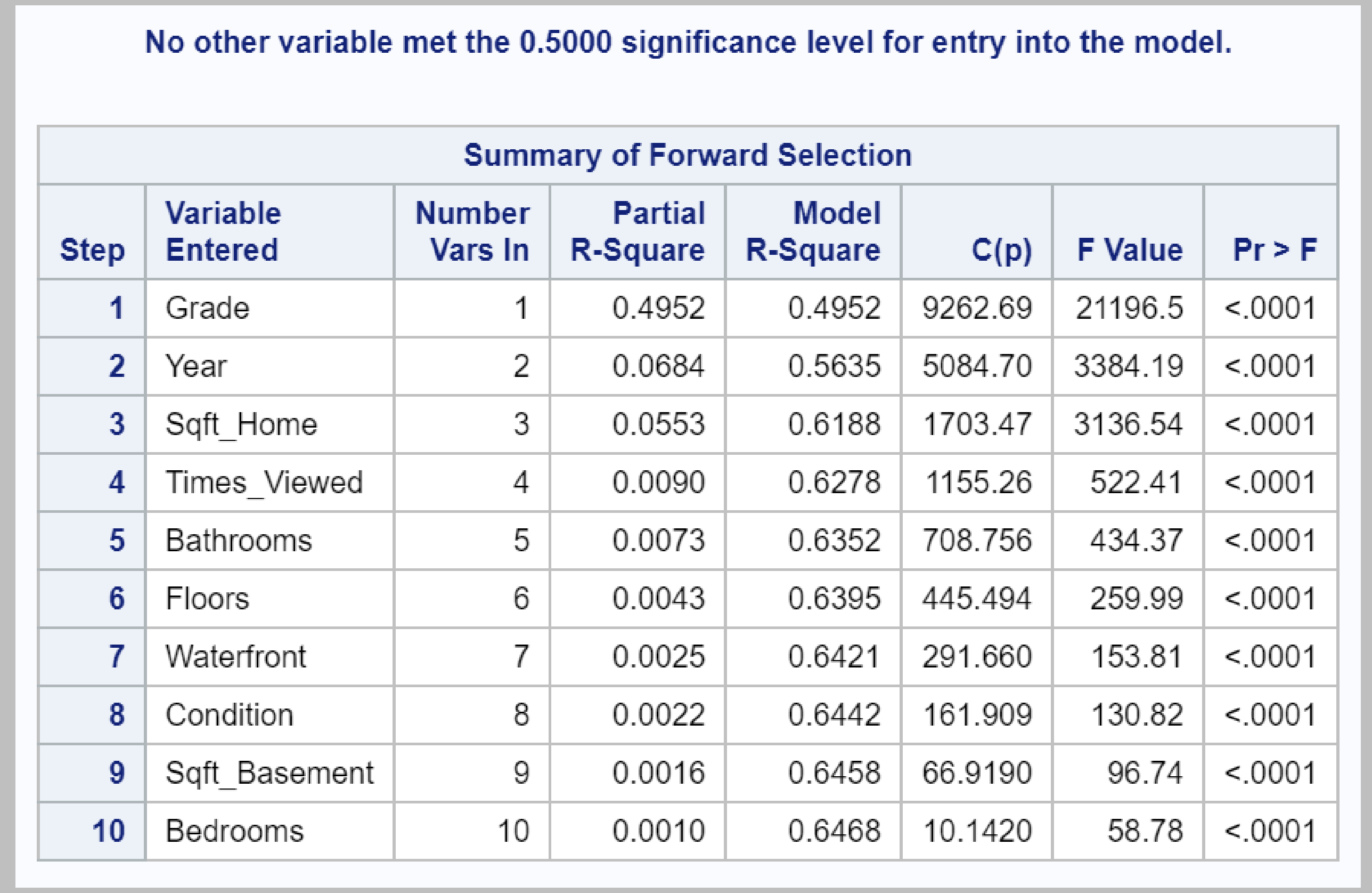


All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Sqft_Lot	10	0.0000	0.6468	10.1420	0.14	0.7063

forward selection, which starts with the first predictor and gradually adds on one independent variable at a time;



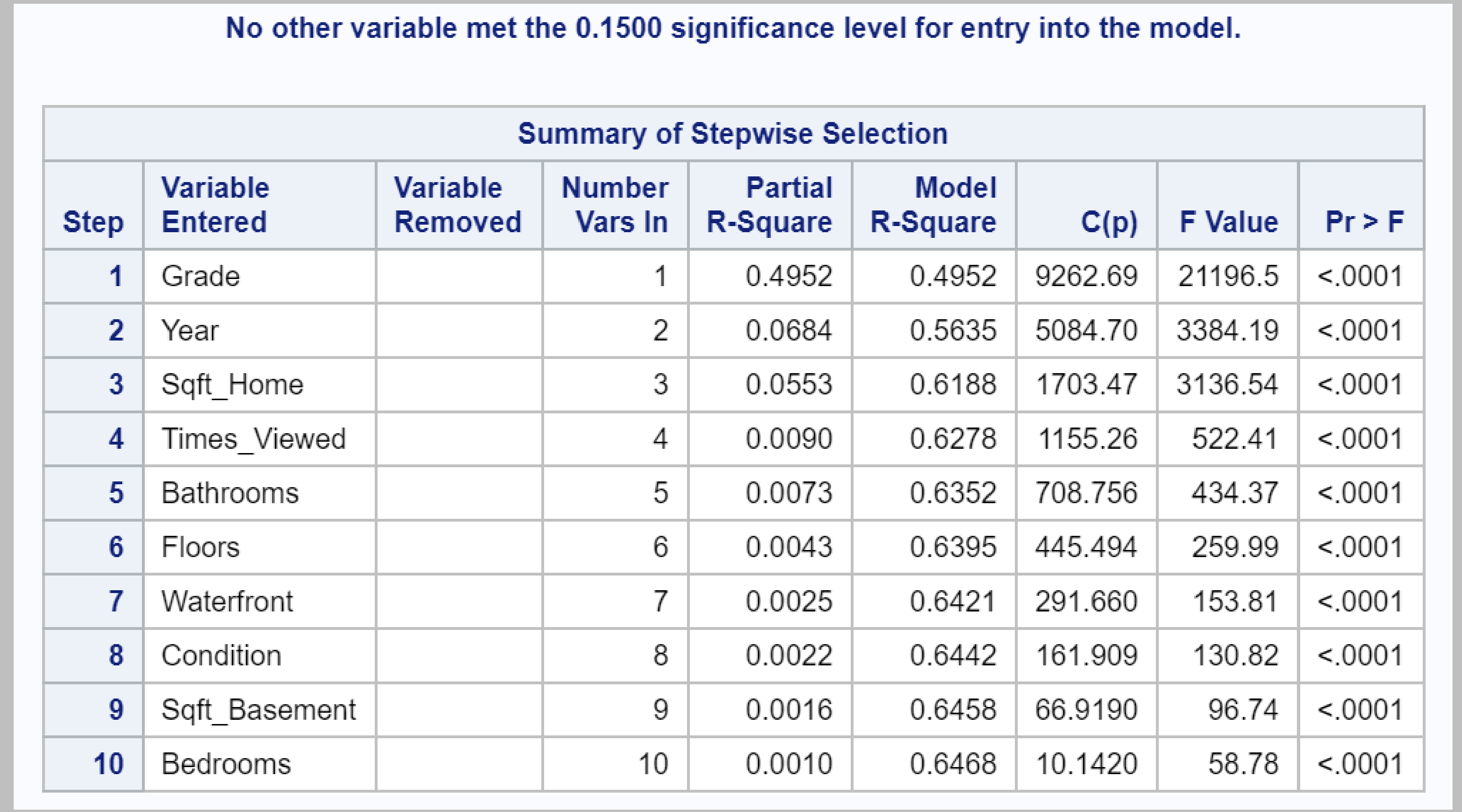
No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Grade	1	0.4952	0.4952	9262.69	21196.5	<.0001
2	Year	2	0.0684	0.5635	5084.70	3384.19	<.0001
3	Sqft_Home	3	0.0553	0.6188	1703.47	3136.54	<.0001
4	Times_Viewed	4	0.0090	0.6278	1155.26	522.41	<.0001
5	Bathrooms	5	0.0073	0.6352	708.756	434.37	<.0001
6	Floors	6	0.0043	0.6395	445.494	259.99	<.0001
7	Waterfront	7	0.0025	0.6421	291.660	153.81	<.0001
8	Condition	8	0.0022	0.6442	161.909	130.82	<.0001
9	Sqft_Basement	9	0.0016	0.6458	66.9190	96.74	<.0001
10	Bedrooms	10	0.0010	0.6468	10.1420	58.78	<.0001

*Sqft_Lot has been removed

and **stepwise selection**, which adds on an independent variable each time but constantly retests the significance of the predictors already present in the model.



No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Grade		1	0.4952	0.4952	9262.69	21196.5	<.0001
2	Year		2	0.0684	0.5635	5084.70	3384.19	<.0001
3	Sqft_Home		3	0.0553	0.6188	1703.47	3136.54	<.0001
4	Times_Viewed		4	0.0090	0.6278	1155.26	522.41	<.0001
5	Bathrooms		5	0.0073	0.6352	708.756	434.37	<.0001
6	Floors		6	0.0043	0.6395	445.494	259.99	<.0001
7	Waterfront		7	0.0025	0.6421	291.660	153.81	<.0001
8	Condition		8	0.0022	0.6442	161.909	130.82	<.0001
9	Sqft_Basement		9	0.0016	0.6458	66.9190	96.74	<.0001
10	Bedrooms		10	0.0010	0.6468	10.1420	58.78	<.0001

*Sqft_Lot has been removed

Running all three variable selection methods coincidentally gives us one common **reduced model that contains all of the original variables in the full model except for Sqft_Lot**. This result corresponds to our observation of Sqft_Lot's insignificant p-value in the PROC REG output.

Multiple Regression Diagnostics in SAS®: Predicting House Sale Prices

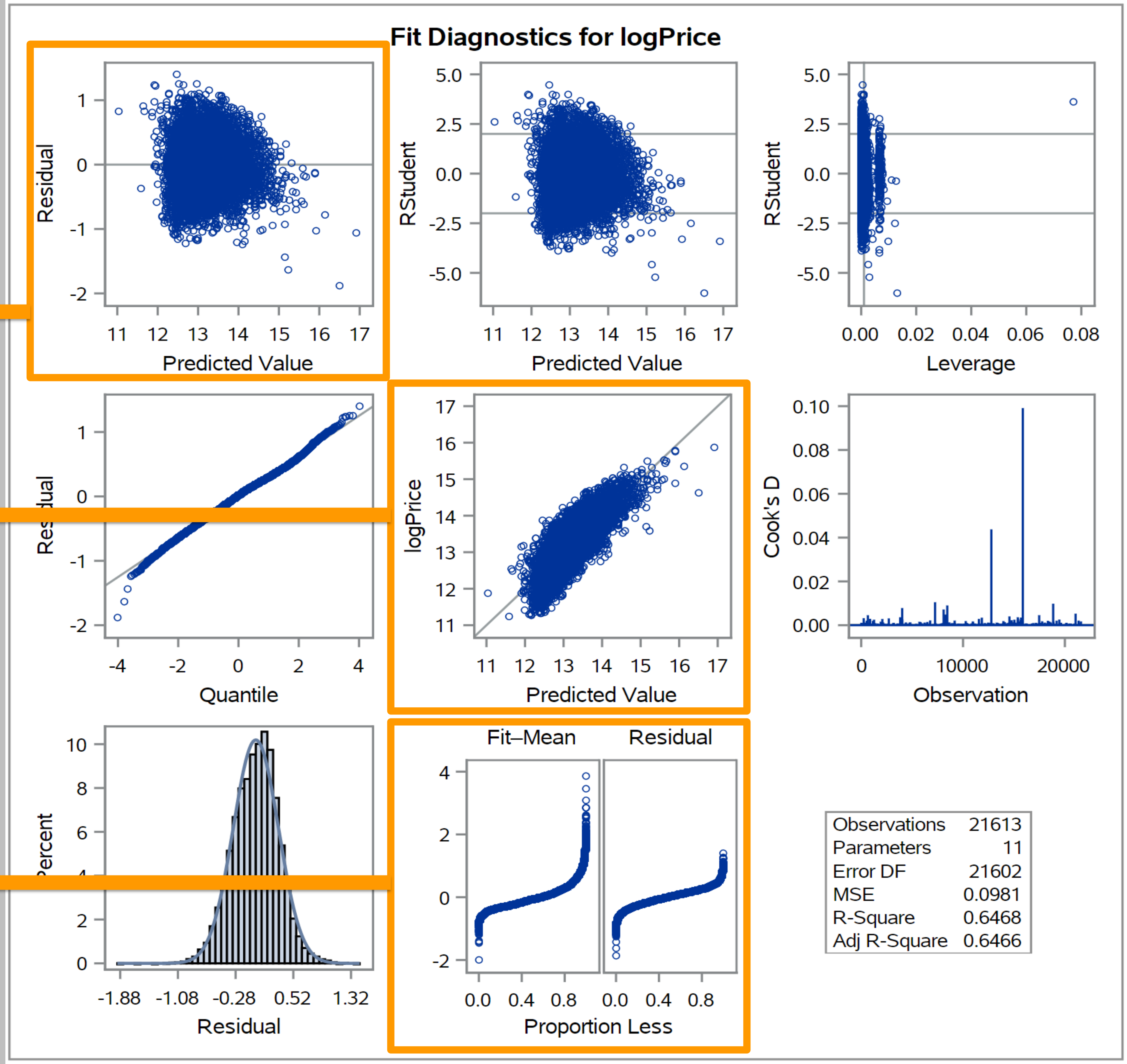
Karen Bui, University of Central Florida

Faculty Advisor: Dr. Daoji Li of UCF

REDUCED MODEL: MAIN DIAGNOSTICS PLOTS

We can then analyze our reduced regression model by using the same PROC REG statement once again—this time, excluding the Sqft_Lot variable. In addition, we can add the VIF option for collinearity examination and specify the variables for outlier and normality analysis later on.

```
proc reg data=HouseSalesLog plots(maxpoints=none);
   model logPrice = Bedrooms Bathrooms Sqft_Home Sqft_Basement Floors Waterfront Times_Viewed
   Condition Grade Year/ vif;
   output
      out=HouseSalesRed
      rstudent=jackknife
      h=leverage
      residual=residual;
run;
```



Observations 21613
Parameters 11
Error DF 21602
MSE 0.0981
R-Square 0.6468
Adj R-Square 0.6466

Residual Plot: This plot exhibits random scatter, which indicates appropriate use of a linear model

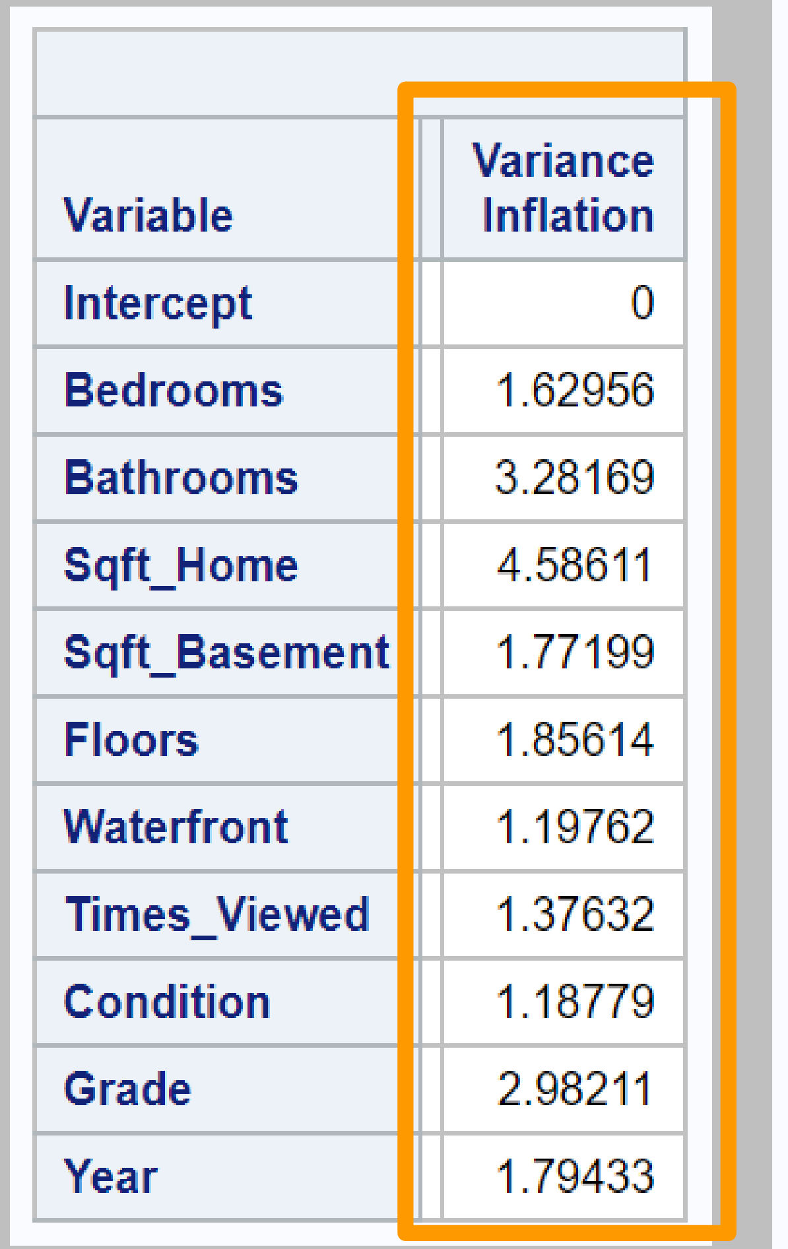
Predicted vs. Observed Plot: The points on this plot align with the reference line to support model adequacy

Residual-Fit Plot: The spread of the centered-fit exceeds that of the residuals, which means that independent variables are responsible for a significant amount of the change in logPrice with little variation remaining, which further supports the adequacy of the linear model

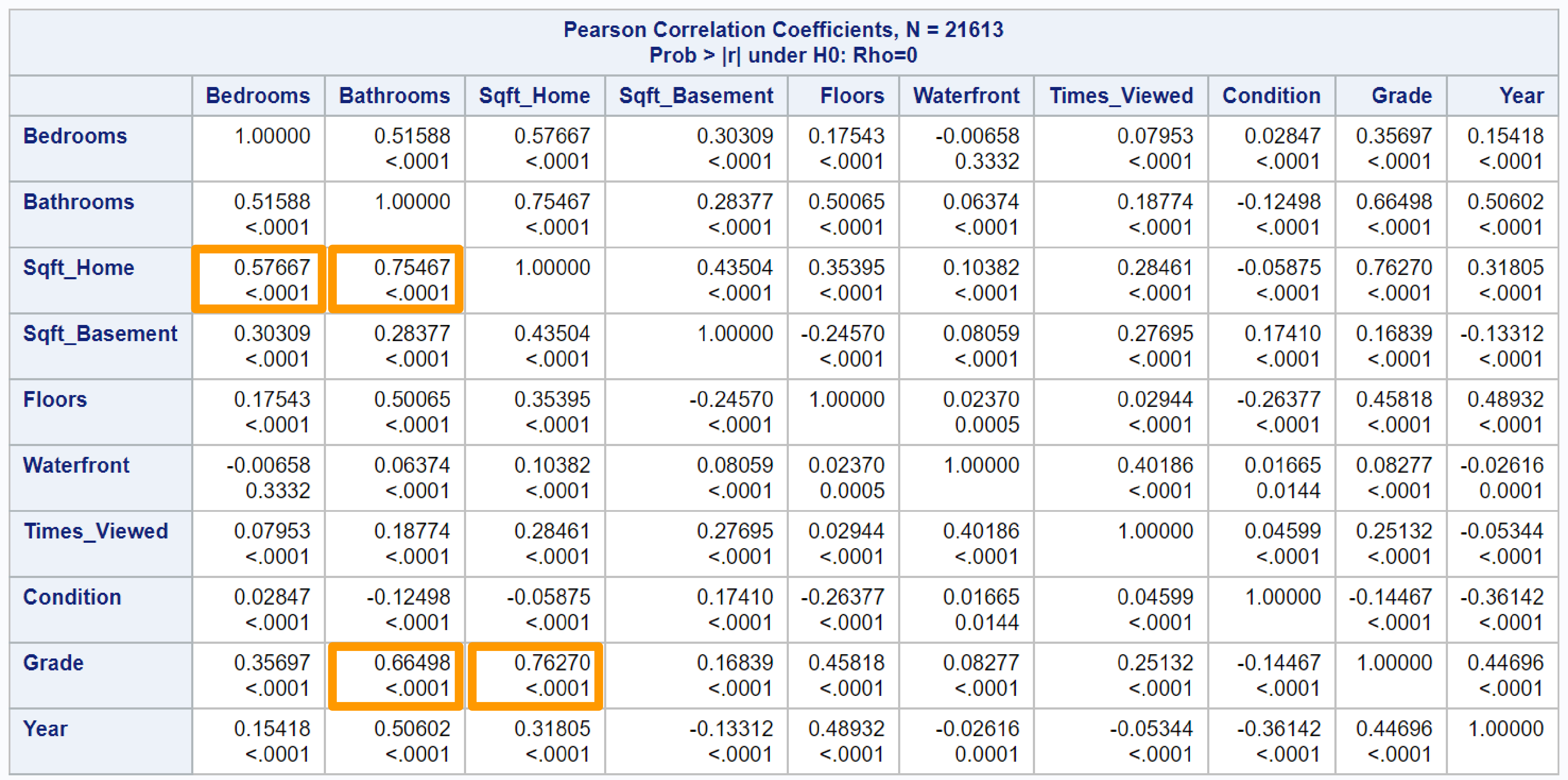
COLLINEARITY

Next, we will look for any signs of collinearity, or correlation between each pair of independent variables in our reduced model. We can refer to the column of "Variable Inflation Factors" for each predictor from our previous PROC REG step (during which we had specified a VIF statement) and also use PROC CORR to do an additional collinearity analysis; we simply need to specify the variables to be examined after the VAR statement:

```
proc corr data=HouseSalesRed;
   var Bedrooms Bathrooms Sqft_Home Sqft_Basement Floors Waterfront Times_Viewed Condition
   Grade Year;
run;
```



Variable	Variance Inflation
Intercept	0
Bedrooms	1.62956
Bathrooms	3.28169
Sqft_Home	4.58611
Sqft_Basement	1.77199
Floors	1.85614
Waterfront	1.19762
Times_Viewed	1.37632
Condition	1.18779
Grade	2.98211
Year	1.79433



Pearson Correlation Coefficients, N = 21613
Prob > |r| under H0: Rho=0

	Bedrooms	Bathrooms	Sqft_Home	Sqft_Basement	Floors	Waterfront	Times_Viewed	Condition	Grade	Year
Bedrooms	1.00000	0.51588 <.0001	0.57667 <.0001	0.30309 <.0001	0.17543 <.0001	-0.00658 0.3332	0.07953 <.0001	0.02847 <.0001	0.35697 <.0001	0.15418 <.0001
Bathrooms	0.51588 <.0001	1.00000	0.75467 <.0001	0.28377 <.0001	0.50065 <.0001	0.06374 <.0001	0.18774 <.0001	-0.12498 <.0001	0.66498 <.0001	0.50602 <.0001
Sqft_Home	0.57667 <.0001	0.75467 <.0001	1.00000	0.43504 <.0001	0.35395 <.0001	0.10382 <.0001	0.28461 <.0001	-0.05875 <.0001	0.76270 <.0001	0.31805 <.0001
Sqft_Basement	0.30309 <.0001	0.28377 <.0001	0.43504 <.0001	1.00000	-0.24570 <.0001	0.08059 <.0001	0.27695 <.0001	0.17410 <.0001	0.16839 <.0001	-0.13312 <.0001
Floors	0.17543 <.0001	0.50065 <.0001	0.35395 <.0001	-0.24570 <.0001	1.00000	0.02370 0.0005	0.02944 <.0001	-0.26377 <.0001	0.45818 <.0001	0.48932 <.0001
Waterfront	-0.00658 0.3332	0.06374 <.0001	0.10382 <.0001	0.08059 <.0001	0.02370 0.0005	1.00000	0.40186 <.0001	0.01665 0.0144	0.08277 <.0001	-0.02616 0.0001
Times_Viewed	0.07953 <.0001	0.18774 <.0001	0.28461 <.0001	0.27695 <.0001	0.02944 <.0001	0.40186 <.0001	1.00000	0.04599 <.0001	0.25132 <.0001	-0.05344 <.0001
Condition	0.02847 <.0001	-0.12498 <.0001	-0.05875 <.0001	0.17410 <.0001	-0.26377 <.0001	0.01665 0.0144	0.04599 <.0001	1.00000	-0.14467 <.0001	-0.36142 <.0001
Grade	0.35697 <.0001	0.66498 <.0001	0.76270 <.0001	0.16839 <.0001	0.45818 <.0001	0.08277 <.0001	0.25132 <.0001	-0.14467 <.0001	1.00000	0.44696 <.0001
Year	0.15418 <.0001	0.50602 <.0001	0.31805 <.0001	-0.13312 <.0001	0.48932 <.0001	-0.02616 0.0001	-0.05344 <.0001	-0.36142 <.0001	0.44696 <.0001	1.00000

Since none of the VIF values exceed 10.0, we do not appear to have a problematic beta.
From the correlation coefficients table, we can see that a large majority of the predictor variable combinations have small p-values, which tend to be indication of collinearity; however, their corresponding r (correlation) values seem to all be small as well, so the small p-values should not present a significant problem.

On the other hand, the **Bedrooms-Sqft_Home, Bathrooms-Sqft_Home, Bathrooms_Grade, Grade-Sqft_Home** combinations all yielded moderately large r values, so it is likely that the **grade is correlated with the home size and number of bathrooms, and the home size is correlated with number of bedrooms and bathrooms**—which is logical, considering the fact that the size of the home is a strong factor in how many bedrooms and bathrooms are possible.

Multiple Regression Diagnostics in SAS®: Predicting House Sale Prices

Karen Bui, University of Central Florida

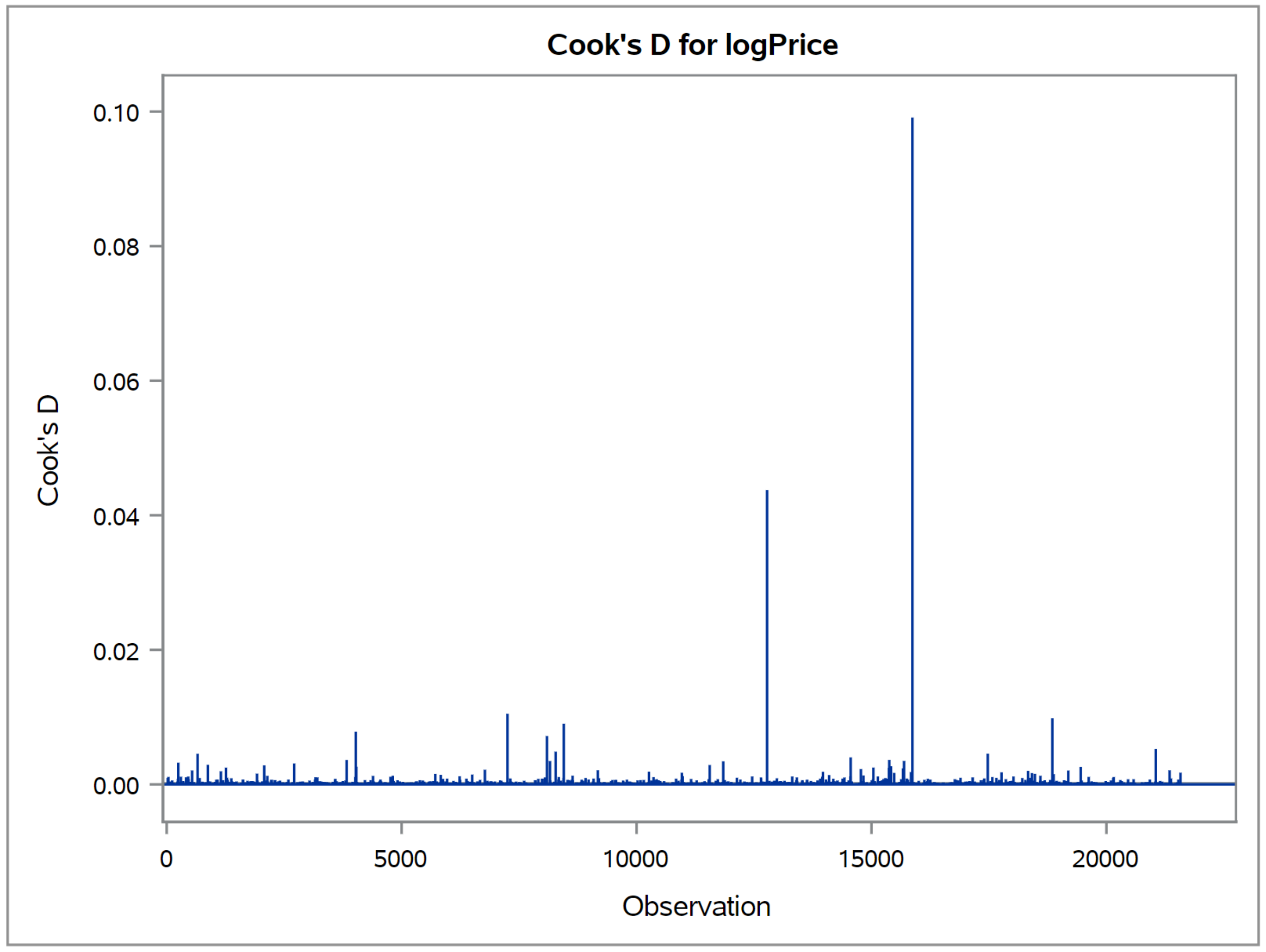
Faculty Advisor: Dr. Daoji Li of UCF

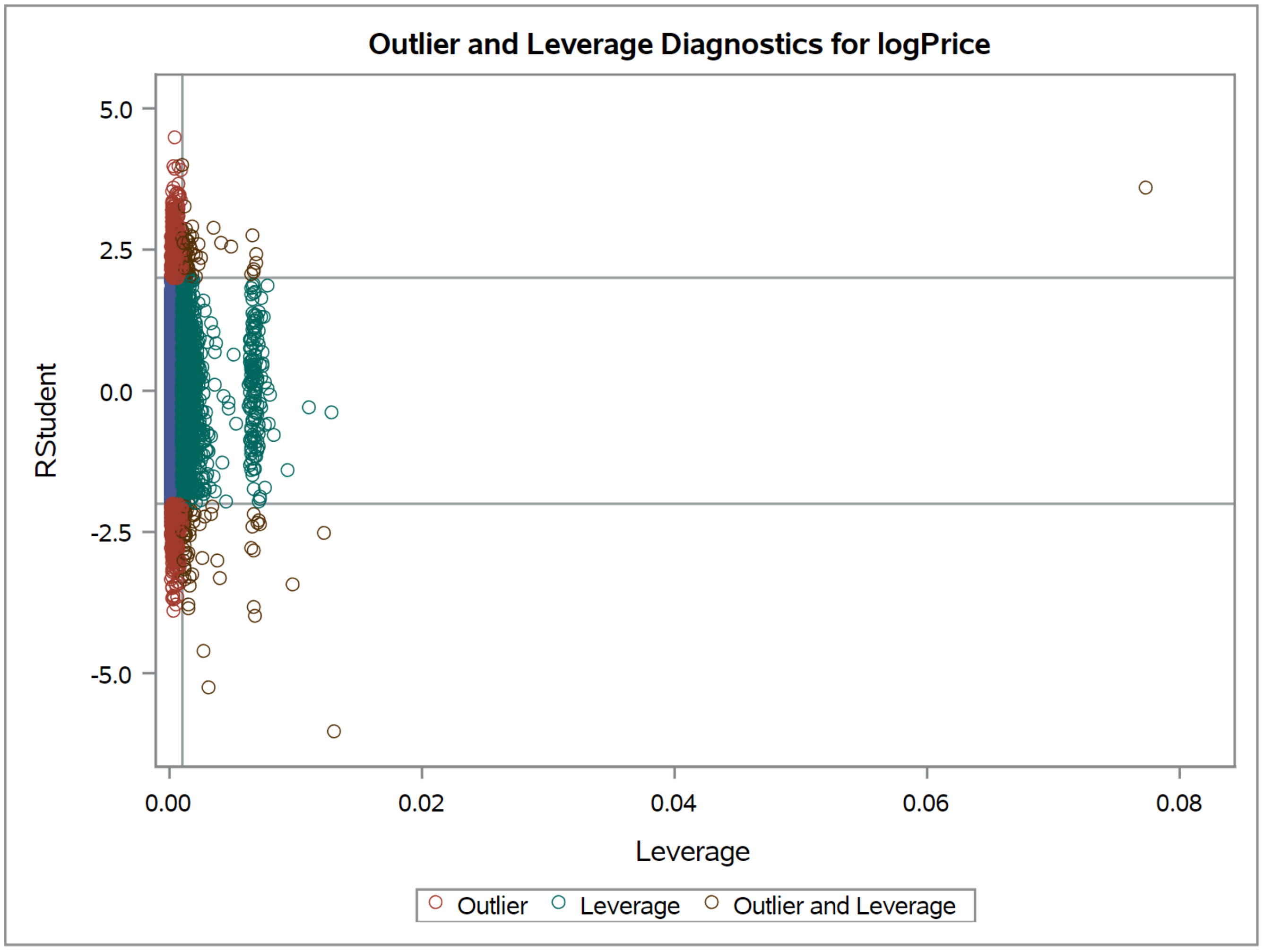
OUTLIERS

From a PROC REG output, leverage values (identified as Hat Diag H) will reveal outliers with respect to independent variables, while Jackknife (RStudent) values will reveal outliers with respect to the dependent or independent variables; in addition, Cook's D values will highlight observations that might be influential in estimating the betas (parameters).

Using the same previous PROC REG procedure as before, specifying an additional PLOTS statement will give us enlarged Cook's D and outlier (Jackknife x leverage) plots:

```
proc reg data=HouseSalesLog plots=(CooksD RStudentByLeverage) plots(maxpoints=none);
   ...
run;
```



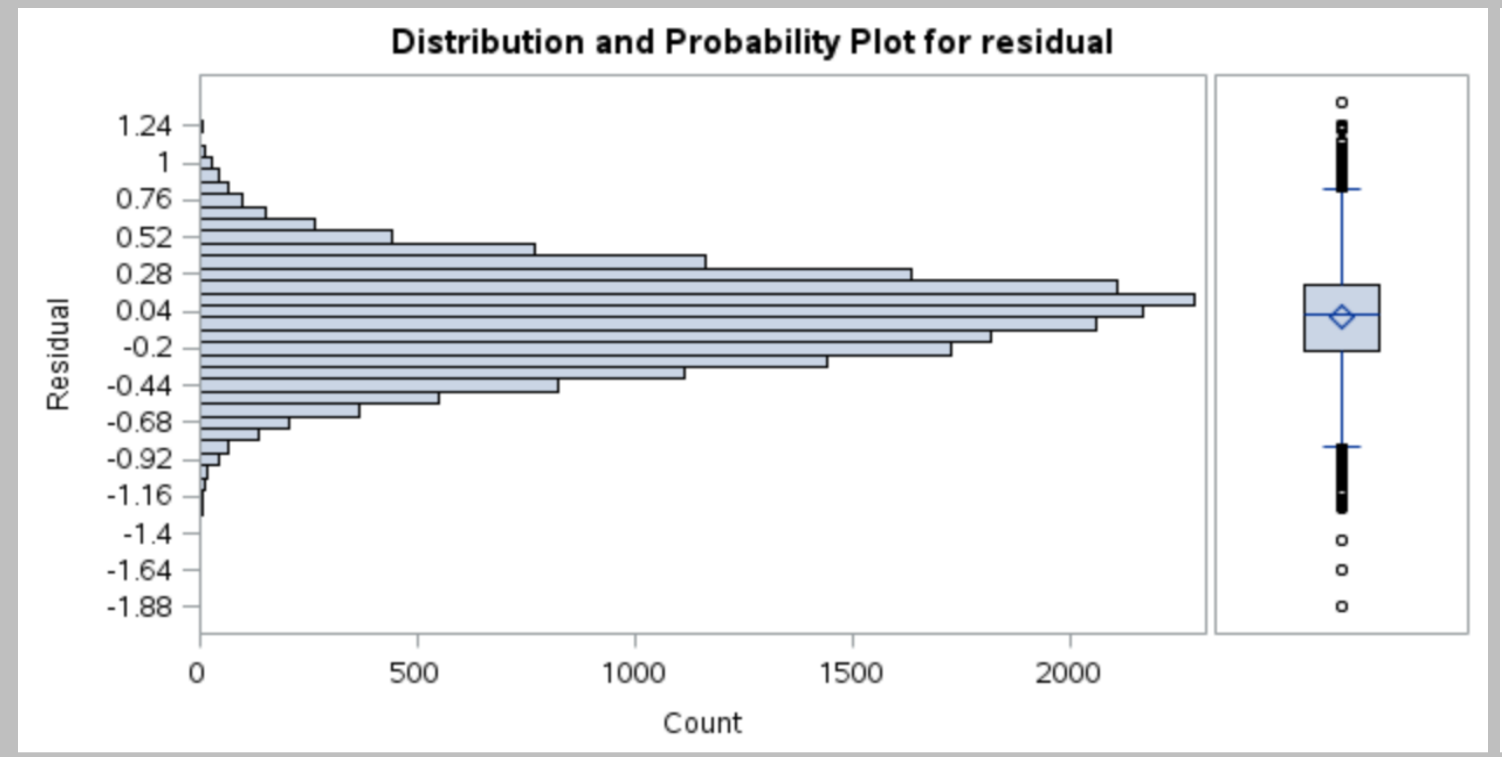


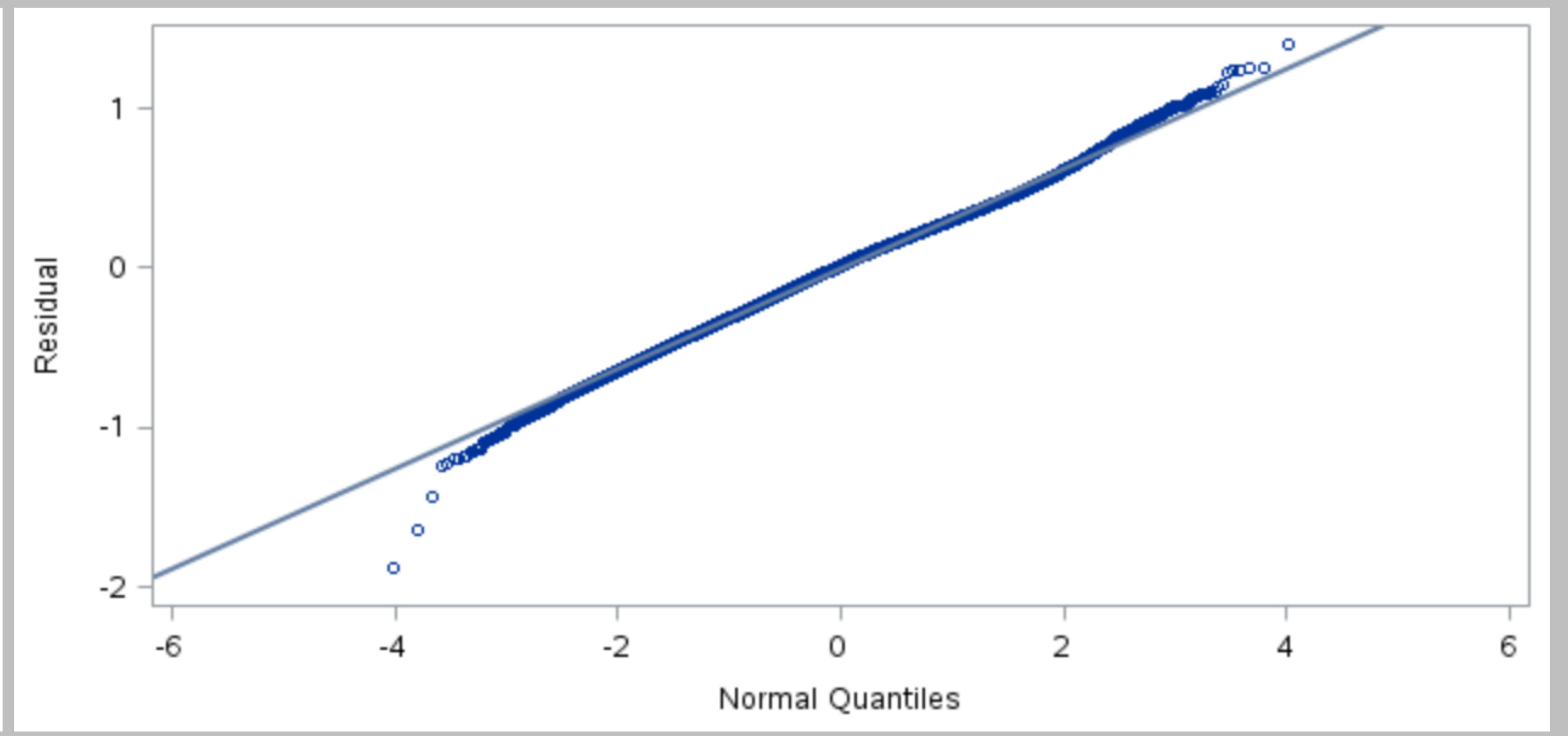
Examining the two plots, we can see that we have at least **two extremely influential observations** and a **large number of outliers with respect to logPrice AND the predictors**, which could be problematic for the normality of the data.

NORMALITY

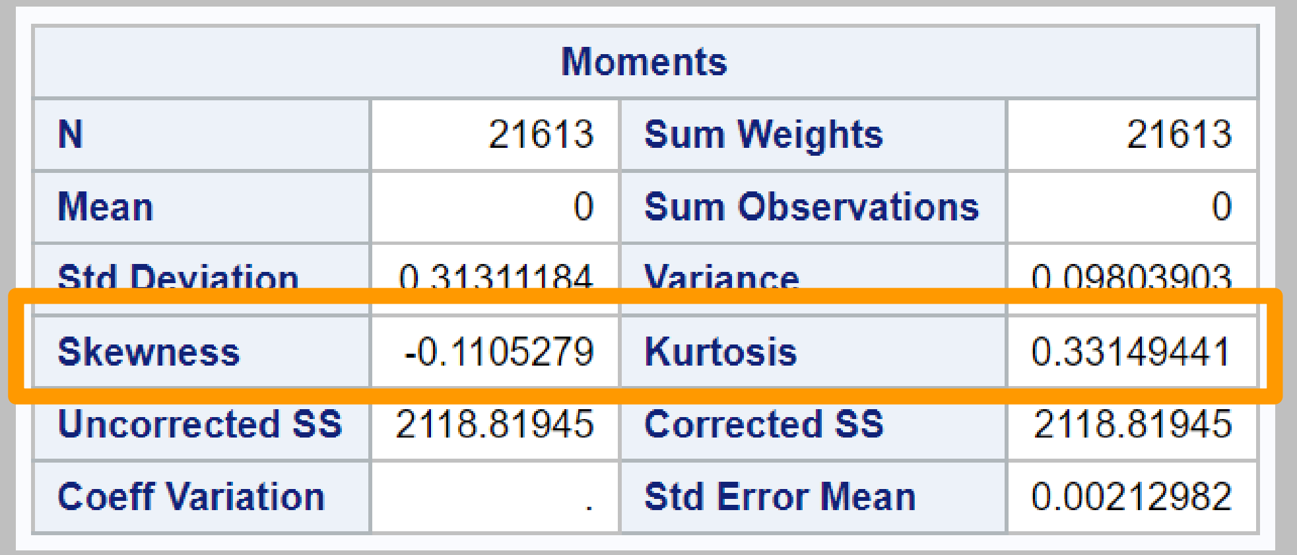
Lastly, we will examine the normality assumption by inspecting the box-plot, histogram, normal probability plot, and the Kolmogorov-Smirnov p-value. We will continue the comparison of the three datasets to see the effect of outlier-removal; using PROC UNIVARIATE, we simply specify the dataset name, a NORMAL PLOT statement, and the variable "residual" (from earlier PROC REG steps).

```
proc univariate data=HouseSalesRed normal plot;
   var residual;
run;
```



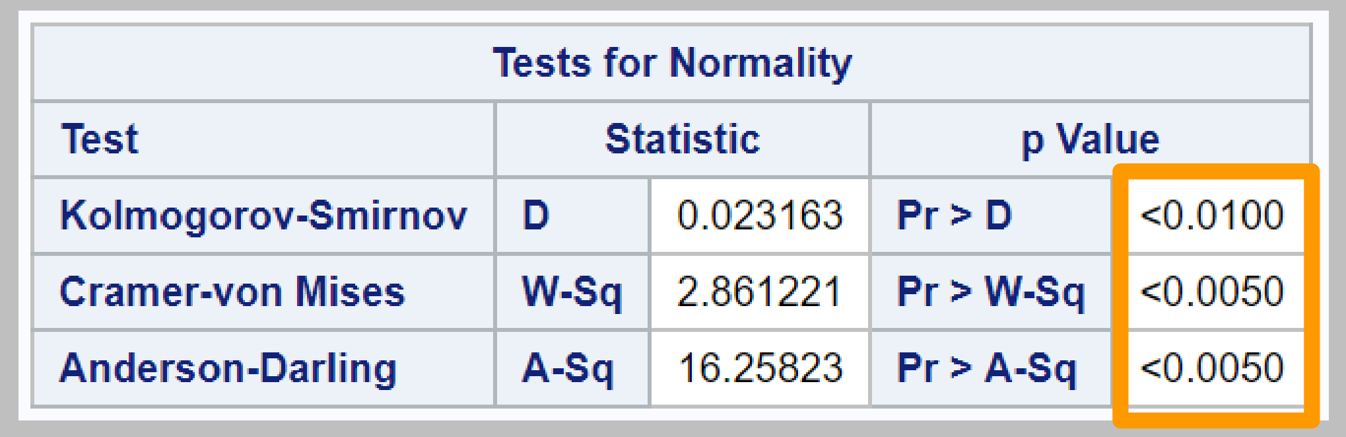


Looking at the output, we can clearly recognize the data's approximate normality (perhaps with a very slight left-tail skew); **the residual histogram and box-plot both appear to be nearly symmetric**—thought the box-plot does show the abundance of outliers previously mentioned. The slight left-tail skew indicates the model's tendency to under-predict the house sale price. The **normal quantile-quantile plot shows the majority of the points residing on the line**, with the ends tailing off a bit, indicating the outliers in the distribution.



Moments

N	21613	Sum Weights	21613
Mean	0	Sum Observations	0
Std Deviation	0.31311184	Variance	0.09803903
Skewness	-0.1105279	Kurtosis	0.33149441
Uncorrected SS	2118.81945	Corrected SS	2118.81945
Coeff Variation	.	Std Error Mean	0.00212982



Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.023163	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.861221	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	16.25823	Pr > A-Sq	<0.0050

The **skewness value for this dataset is approximately -0.11**, and since normally distributed data has a skewness near zero, **this statistic further supports our dataset's approximate normality**. In addition, the **kurtosis value** is observed to be about 0.33, which is considered to be extremely low kurtosis (measured against a standard range of -2 to +2); this **indicates standard heaviness in the left and right histogram tails, so the outlier impact is not too great**. In contrast, all p-values in the "Tests for Normality" table are shown to be less than a standard alpha of 0.05, which rejects the null hypothesis in all four tests for normality.

Multiple Regression Diagnostics in SAS®: Predicting House Sale Prices

Karen Bui, University of Central Florida

Faculty Advisor: Dr. Daoji Li of UCF

CONCLUSION

In this residual analysis, we mainly saw flaws in regards to collinearity, since it makes sense for characteristics of the home to impact each other. This could hint at an exploration into implementation of interaction terms. Besides correlated independent variables, the model after the logarithmic transformation was shown to be useful in predicting house sale price, with residual plots supporting the appropriateness and adequacy of the linear regression. Upon examination of the Cook's D and RStudent by Leverage plots, we saw very many outliers; however, the data appeared to be approximately normally distributed, as shown through the normal probability plot, box-plot, histogram and skewness/kurtosis statistics, although the tests for normality contradicted these results (possibly due to the large number of outliers).

Overall, this model should be considered suitable to use in the prediction of home sale price. Additionally from this exploration, we can use these regression diagnostics to think of other, possibly more complex models to try and predict house sale price, such as models with interaction terms (given the collinearity results) or models with even more predictor variables. It is always possible investigate more deeply, and a good understanding of multiple regression diagnostics techniques will definitely advance the process, but additional knowledge of how to apply those methods in a statistical platform such as SAS will yield much more accurate results and a higher-quality understanding, since the manipulation of big data becomes possible.

REFERENCES

”House Sales in King County, USA”. *Kaggle*. 2016. Available at https://www.kaggle.com/harlfoxem/housesalesprediction

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF