Assignment #2

Nate Bitting

Introduction

For this assignment, the objective is to leverage the Ames Housing dataset to develop several regression models using single and/or multiple predictor variables to predict SalePrice. The first step is to create two simple linear regression models using a single predictor variable and SalePrice as the response variable. We will then evaluate the goodness of fit and accuracy of the models in order to determine which model performs better. The second objective is to develop two multiple regression models using several predictor variables to predict SalePrice. We will then compare the results of each model to determine which is superior. Lastly, we will perform a log transformation on SalePrice to determine if it improves model performance.

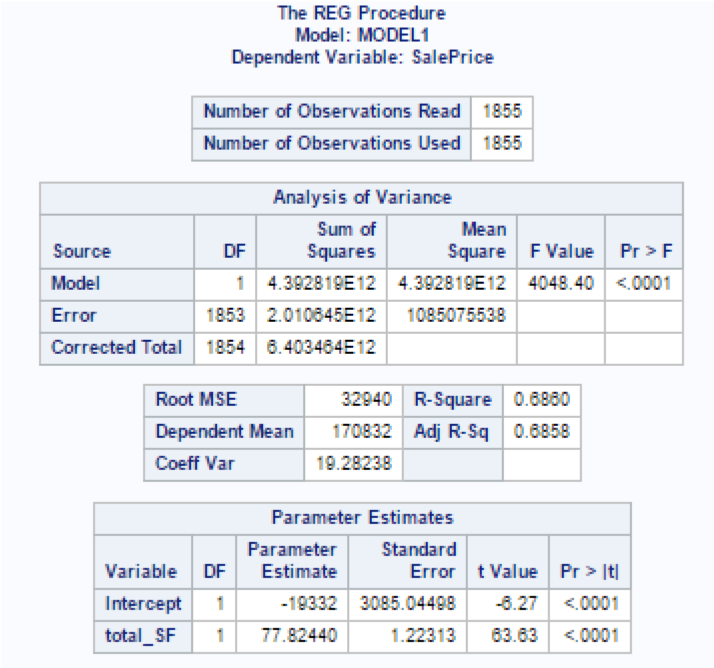
In order to prepare the data, we performed several data cleansing steps to filter out data we did not want included in the sample used to train the regression models for this assignment. We first only included single family homes in residential zoning areas. We also only included homes with paved streets and those that had public utilities. Lastly, we performed a univariate procedure to identify the extreme observations so that we could exclude those outliers from our sample. In the next step before building our regression models, we created dummy variables for many of the categorical variables we were interested in using: Central Air, Fireplaces, Garage, Neighborhood, Kitchen Quality, Basement Quality, Exterior Quality, Fireplace Quality, Foundation type, and Month Sold.

Results

Simple Regression Model 1

Response Variable: *SalePrice* **Predictor Variable**: *total_SF*
Equation: *SalePrice = -19332 + 77.8244(total_SF)*

Model 1 ANOVA Output



The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1855
Number of Observations Used	1855

Analysis of Variance

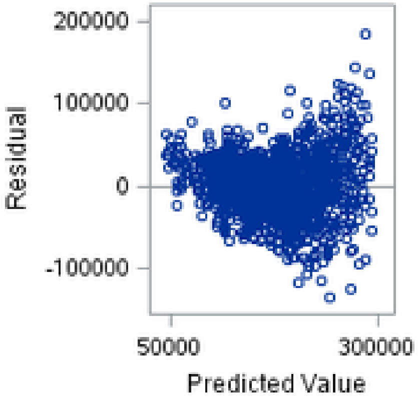
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.392819E12	4.392819E12	4048.40	<.0001
Error	1853	2.010645E12	1085075538		
Corrected Total	1854	6.403464E12			

Root MSE	32940	R-Square	0.6860
Dependent Mean	170832	Adj R-Sq	0.6858
Coeff Var	19.28238		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-19332	3085.04498	-6.27	<.0001
total_SF	1	77.82440	1.22313	63.63	<.0001

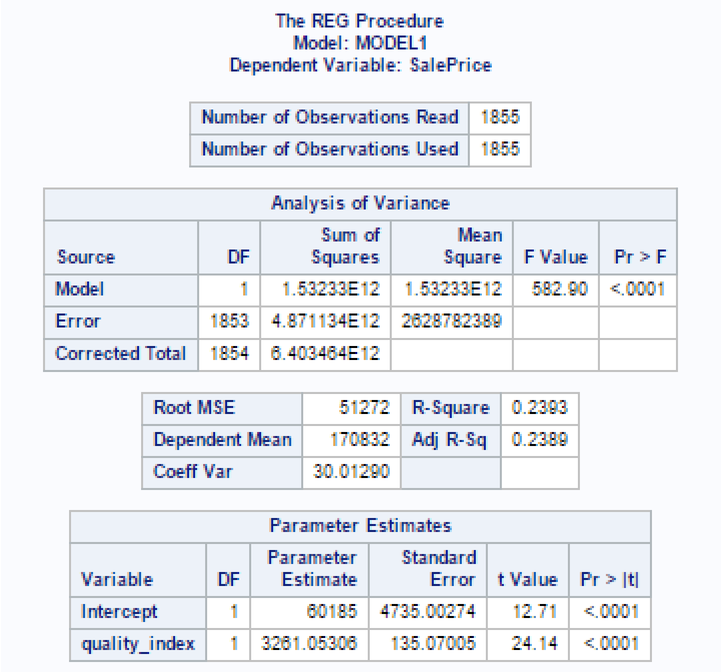
Model 1 Residual Plot



Simple Regression Model 2

Response Variable: *SalePrice* **Predictor Variable**: *quality_index*
Equation: *SalePrice = -19332 + 77.8244(quality_index)*

Model 2 ANOVA Output



The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

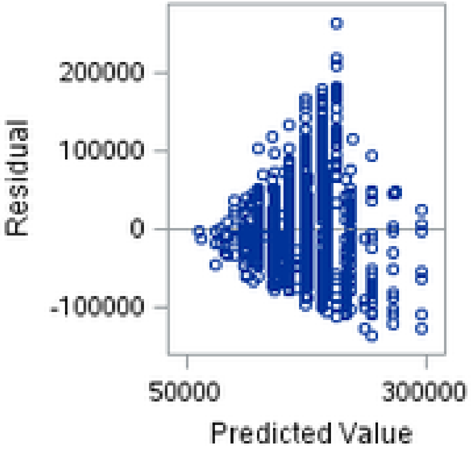
Number of Observations Read	1855
Number of Observations Used	1855

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.53233E12	1.53233E12	582.90	<.0001
Error	1853	4.871134E12	2628782389		
Corrected Total	1854	6.403464E12			

Root MSE	51272	R-Square	0.2393
Dependent Mean	170832	Adj R-Sq	0.2389
Coeff Var	30.01290		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	60185	4735.00274	12.71	<.0001
quality_index	1	3261.05306	135.07005	24.14	<.0001

Model 2 Residual Plot

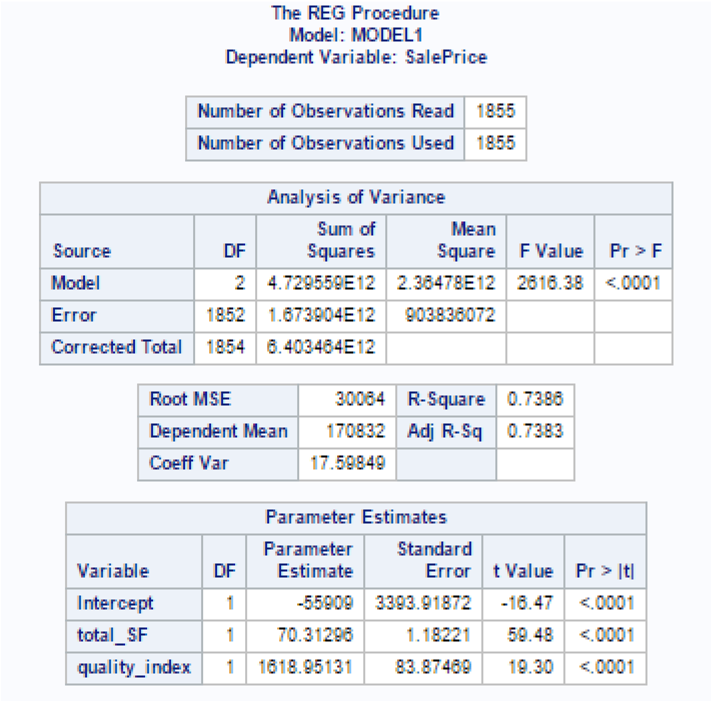


By observing the ANOVA output of both Model 1 and Model 2, you can determine goodness of fit by reviewing the coefficient of determination (R-Square) of both models. Model 1 and Model 2 have an R-Square of 0.686 and 0.2303, respectively. Given Model 1 has a higher R-Square value and the MSE is much lower than Model 2, it is safe to say that Model 1 has a better fit than Model 2. The residual output also give us some insight to the model performance as well. By looking at the *Model 1 Residual Plot*, you will notice the residuals are fairly homoscedastic whereas the *Model 2 Residual Plot* has a fanning out pattern as the residuals increase as the quality_index gets larger. Thus, we can conclude that Model 1 performs better than Model 2.

Multiple Regression Model

Response Variable: *SalePrice* **Predictor Variables**: *total_SF, quality_index*
Equation: *SalePrice = -55909 + 70.31296(total_SF) + 1618.95131(quality_index)*

Multiple Regression Model ANOVA Output



The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1855
Number of Observations Used	1855

Analysis of Variance

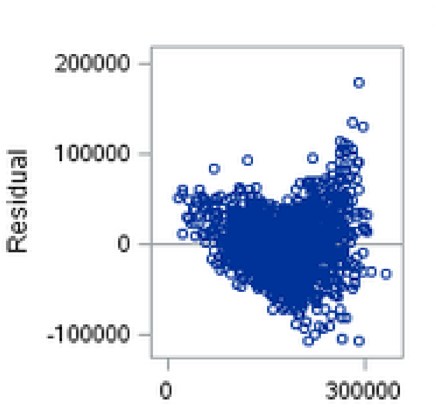
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.729559E12	2.36478E12	2616.38	<.0001
Error	1852	1.673904E12	903836072		
Corrected Total	1854	6.403464E12			

Root MSE	30064	R-Square	0.7386
Dependent Mean	170832	Adj R-Sq	0.7383
Coeff Var	17.59849		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-55909	3393.91872	-16.47	<.0001
total_SF	1	70.31296	1.18221	59.48	<.0001
quality_index	1	1618.95131	83.87469	19.30	<.0001

Multiple Regression Model Residual Plot

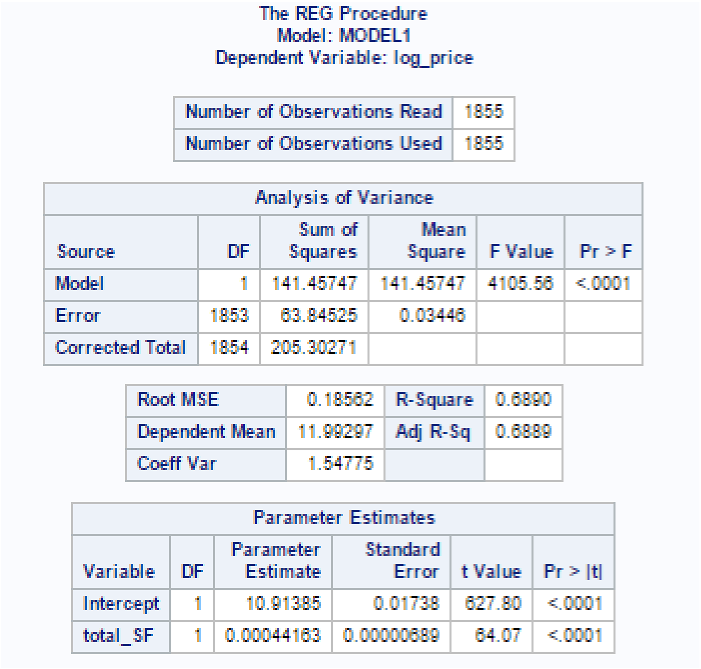


Based on the results in the *Multiple Regression Model ANOVA Output* we can see that the MSE reduced significantly from Model 1 and Model 2 and the R-Square is much higher at 0.7386. As a result of these findings, the Multiple Regression Model has a much better goodness of fit than both Model 1 and Model 2.

Regression Model 1 – Log Transformation on SalePrice

Response Variable: *log(SalePrice)* **Predictor Variable**: *total_SF*
Equation: *log(SalePrice) = 10.91385 + 0.00044163(total_SF)*

Model 1 Log Transformation ANOVA Output



The REG Procedure
Model: MODEL1
Dependent Variable: log_price

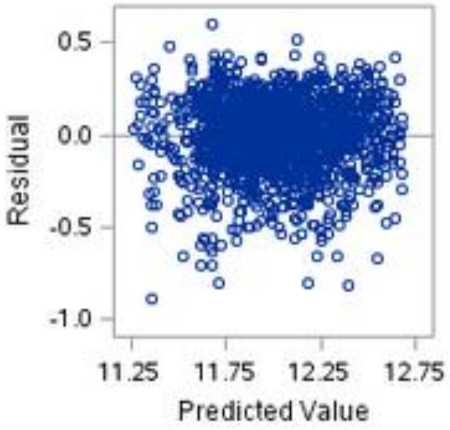
Number of Observations Read	1855
Number of Observations Used	1855

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	141.45747	141.45747	4105.56	<.0001
Error	1853	63.84525	0.03446		
Corrected Total	1854	205.30271			

Root MSE	0.18562	R-Square	0.6890
Dependent Mean	11.99297	Adj R-Sq	0.6889
Coeff Var	1.54775		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.91385	0.01738	627.80	<.0001
total_SF	1	0.00044163	0.00000689	64.07	<.0001

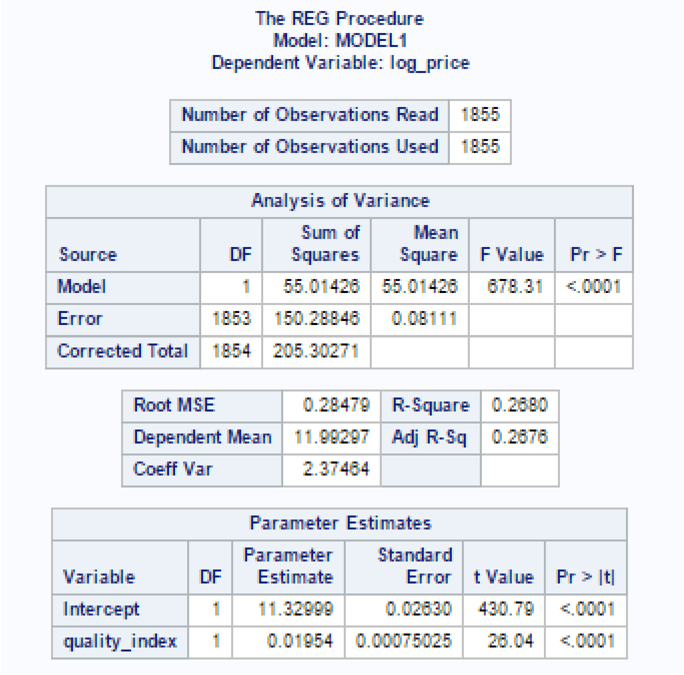
Model 1 Log Transformation Residual Plot



Regression Model 2 – Log Transformation on SalePrice

Response Variable: *log(SalePrice)* **Predictor Variable**: *quality_index*
Equation: *log(SalePrice) = -11.32999 + 0.01954(quality_index)*

Model 2 Log Transformation ANOVA Output



The REG Procedure
Model: MODEL1
Dependent Variable: log_price

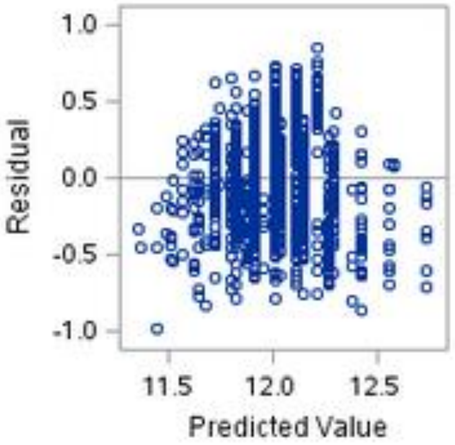
Number of Observations Read	1855
Number of Observations Used	1855

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	55.01426	55.01426	678.31	<.0001
Error	1853	150.28846	0.08111		
Corrected Total	1854	205.30271			

Root MSE	0.28479	R-Square	0.2680
Dependent Mean	11.99297	Adj R-Sq	0.2676
Coeff Var	2.37464		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.32999	0.02630	430.79	<.0001
quality_index	1	0.01954	0.00075025	26.04	<.0001

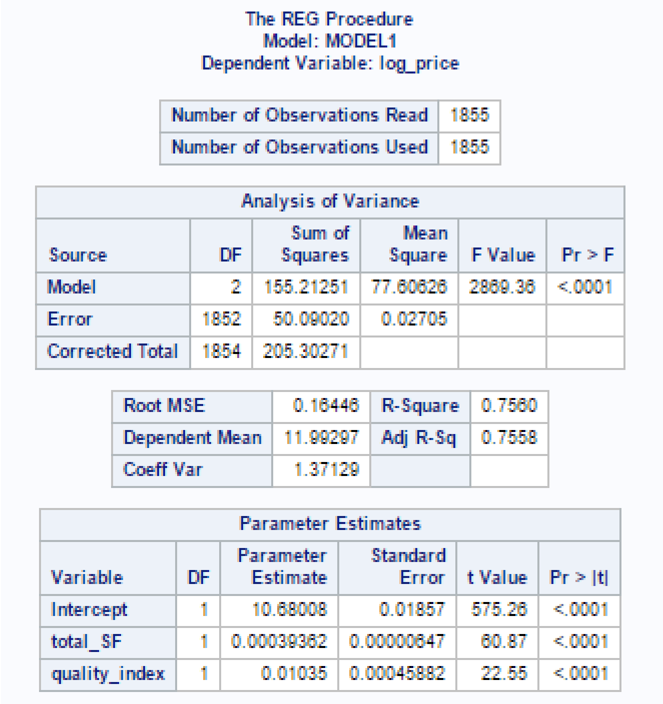
Model 2 Log Transformation Residual Plot



Multiple Regression Model – Log Transformation on SalePrice

Response Variable: *log(SalePrice)* **Predictor Variables**: *total_SF, quality_index*
Equation: *log(SalePrice) = 10.68008 + 0.00039362(total_SF) + 0.01035(quality_index)*

Multiple Regression Model ANOVA Output



The REG Procedure
Model: MODEL1
Dependent Variable: log_price

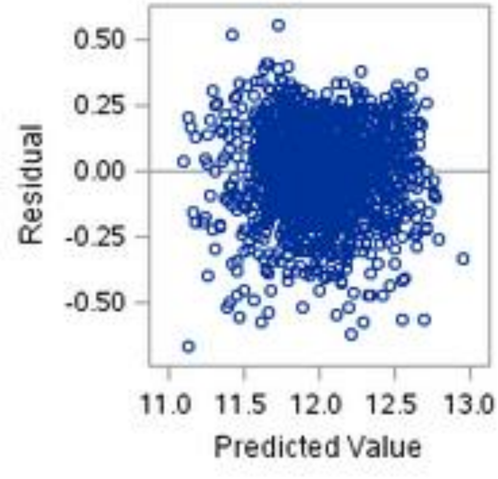
Number of Observations Read	1855
Number of Observations Used	1855

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	155.21251	77.60626	2869.36	<.0001
Error	1852	50.09020	0.02705		
Corrected Total	1854	205.30271			

Root MSE	0.16446	R-Square	0.7560
Dependent Mean	11.99297	Adj R-Sq	0.7558
Coeff Var	1.37129		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.68008	0.01857	575.26	<.0001
total_SF	1	0.00039362	0.00000647	60.87	<.0001
quality_index	1	0.01035	0.00045882	22.55	<.0001

Multiple Regression Model Residual Plot

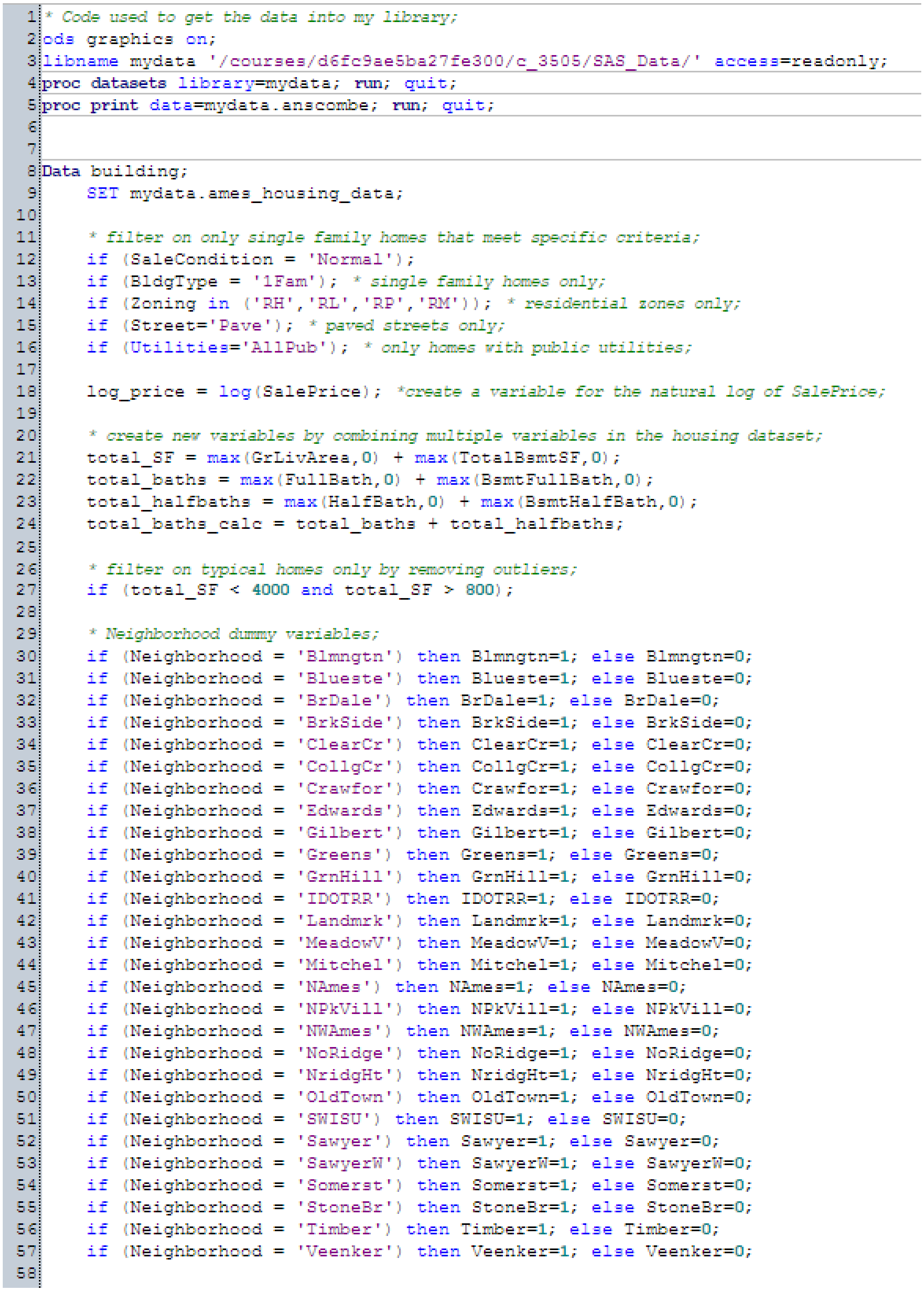


After rebuilding the models leveraging the log transformation of the SalePrice for all three models we can leverage the ANOVA output and residuals to determine if these new models perform better than the original models. For Model 1 and Model 2, we can see there was no improvement in their R-Square values as they remained exactly the same before the transformation on SalePrice. However, after the log transformation on SalePrice, the Multiple Regression Model showed some improvement its R-Square value going from 0.7386 up to 0.7560.

Conclusion

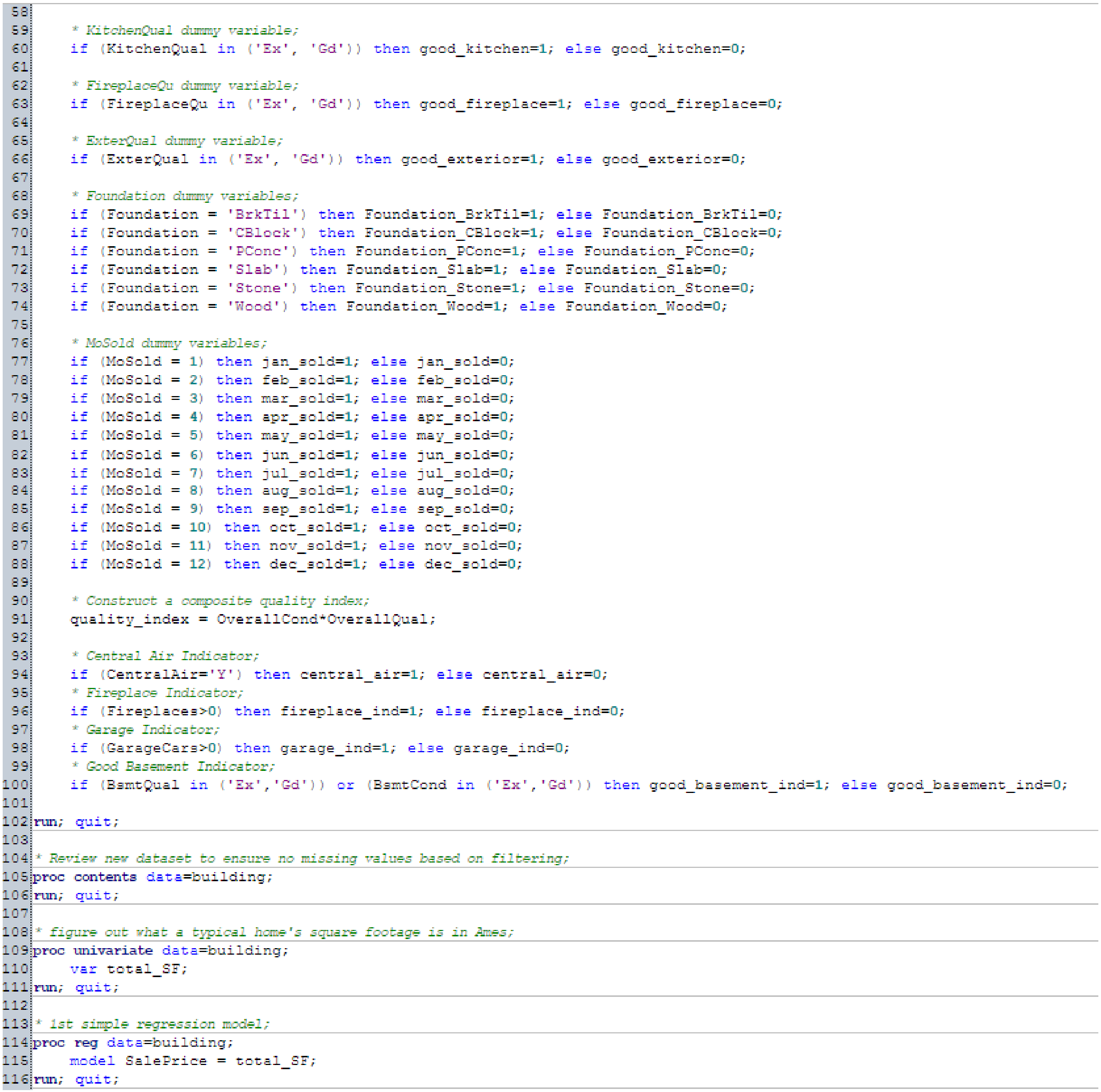
In conclusion, performing the log transformation on SalePrice in the Multiple Regression Model had the best goodness of fit than all other models. Performing the log transformations on the simple linear regression models had no impact on model performance or fit than the original models. By leveraging the ANOVA output from REG procedure, comparison of multiple models is a very straightforward endeavor given the limited number of regressors used in this assignment. As you begin to add more predictor variables to your model, you can continue to tweak and fine tune the models for optimal performance.

Code:



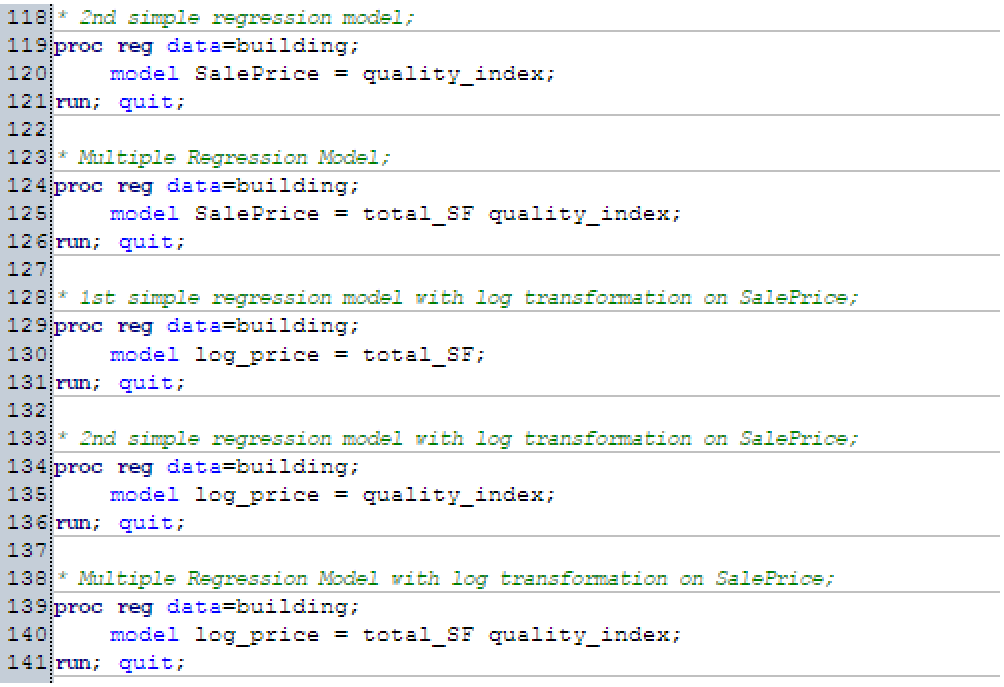
```
* Code used to get the data into my library;
ods graphics on;
libname mydata '/courses/d6fc9ae5ba27fe300/c_3505/SAS_Data/' access=readonly;
proc datasets library=mydata; run; quit;
proc print data=mydata.anscombe; run; quit;


Data building;
    SET mydata.ames_housing_data;

    * filter on only single family homes that meet specific criteria;
    if (SaleCondition = 'Normal');
    if (BldgType = '1Fam'); * single family homes only;
    if (Zoning in ('RH','RL','RP','RM')); * residential zones only;
    if (Street='Pave'); * paved streets only;
    if (Utilities='AllPub'); * only homes with public utilities;

    log_price = log(SalePrice); *create a variable for the natural log of SalePrice;

    * create new variables by combining multiple variables in the housing dataset;
    total_SF = max(GrLivArea,0) + max(TotalBsmtSF,0);
    total_baths = max(FullBath,0) + max(BsmtFullBath,0);
    total_halfbaths = max(HalfBath,0) + max(BsmtHalfBath,0);
    total_baths_calc = total_baths + total_halfbaths;

    * filter on typical homes only by removing outliers;
    if (total_SF < 4000 and total_SF > 800);

    * Neighborhood dummy variables;
    if (Neighborhood = 'Blmngtn') then Blmngtn=1; else Blmngtn=0;
    if (Neighborhood = 'Blueste') then Blueste=1; else Blueste=0;
    if (Neighborhood = 'BrDale') then BrDale=1; else BrDale=0;
    if (Neighborhood = 'BrkSide') then BrkSide=1; else BrkSide=0;
    if (Neighborhood = 'ClearCr') then ClearCr=1; else ClearCr=0;
    if (Neighborhood = 'CollgCr') then CollgCr=1; else CollgCr=0;
    if (Neighborhood = 'Crawfor') then Crawfor=1; else Crawfor=0;
    if (Neighborhood = 'Edwards') then Edwards=1; else Edwards=0;
    if (Neighborhood = 'Gilbert') then Gilbert=1; else Gilbert=0;
    if (Neighborhood = 'Greens') then Greens=1; else Greens=0;
    if (Neighborhood = 'GrnHill') then GrnHill=1; else GrnHill=0;
    if (Neighborhood = 'IDOTRR') then IDOTRR=1; else IDOTRR=0;
    if (Neighborhood = 'Landmrk') then Landmrk=1; else Landmrk=0;
    if (Neighborhood = 'MeadowV') then MeadowV=1; else MeadowV=0;
    if (Neighborhood = 'Mitchel') then Mitchel=1; else Mitchel=0;
    if (Neighborhood = 'NAmes') then NAmes=1; else NAmes=0;
    if (Neighborhood = 'NPkVill') then NPkVill=1; else NPkVill=0;
    if (Neighborhood = 'NWAmes') then NWAmes=1; else NWAmes=0;
    if (Neighborhood = 'NoRidge') then NoRidge=1; else NoRidge=0;
    if (Neighborhood = 'NridgHt') then NridgHt=1; else NridgHt=0;
    if (Neighborhood = 'OldTown') then OldTown=1; else OldTown=0;
    if (Neighborhood = 'SWISU') then SWISU=1; else SWISU=0;
    if (Neighborhood = 'Sawyer') then Sawyer=1; else Sawyer=0;
    if (Neighborhood = 'SawyerW') then SawyerW=1; else SawyerW=0;
    if (Neighborhood = 'Somerst') then Somerst=1; else Somerst=0;
    if (Neighborhood = 'StoneBr') then StoneBr=1; else StoneBr=0;
    if (Neighborhood = 'Timber') then Timber=1; else Timber=0;
    if (Neighborhood = 'Veenker') then Veenker=1; else Veenker=0;

```

Code Continued:



```

    * KitchenQual dummy variable;
    if (KitchenQual in ('Ex', 'Gd')) then good_kitchen=1; else good_kitchen=0;

    * FireplaceQu dummy variable;
    if (FireplaceQu in ('Ex', 'Gd')) then good_fireplace=1; else good_fireplace=0;

    * ExterQual dummy variable;
    if (ExterQual in ('Ex', 'Gd')) then good_exterior=1; else good_exterior=0;

    * Foundation dummy variables;
    if (Foundation = 'BrkTil') then Foundation_BrkTil=1; else Foundation_BrkTil=0;
    if (Foundation = 'CBlock') then Foundation_CBlock=1; else Foundation_CBlock=0;
    if (Foundation = 'PConc') then Foundation_PConc=1; else Foundation_PConc=0;
    if (Foundation = 'Slab') then Foundation_Slab=1; else Foundation_Slab=0;
    if (Foundation = 'Stone') then Foundation_Stone=1; else Foundation_Stone=0;
    if (Foundation = 'Wood') then Foundation_Wood=1; else Foundation_Wood=0;

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

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

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

run; quit;

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

* figure out what a typical home's square footage is in Ames;
proc univariate data=building;
    var total_SF;
run; quit;

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

Code Continued:



```
* 2nd simple regression model;
proc reg data=building;
    model SalePrice = quality_index;
run; quit;

* Multiple Regression Model;
proc reg data=building;
    model SalePrice = total_SF quality_index;
run; quit;

* 1st simple regression model with log transformation on SalePrice;
proc reg data=building;
    model log_price = total_SF;
run; quit;

* 2nd simple regression model with log transformation on SalePrice;
proc reg data=building;
    model log_price = quality_index;
run; quit;

* Multiple Regression Model with log transformation on SalePrice;
proc reg data=building;
    model log_price = total_SF quality_index;
run; quit;
```