410-57, Assignment #2
Anamitra Bhattacharyya

INTRODUCTION

This assignment specifically deals with building simple linear regression models (LRMs) with individual predictor variables, in addition to multiple linear models (MRMs) with multiple predictor variables. The overall goal of the assignment and project as a whole is to identify suitable predictors of house sale price in Ames, Iowa. The variables are derived from an observational data set from the Ames Assessor's Office used in obtaining values for individual residential properties sold in Ames, Iowa between 2006 and 2010. The LRMs and MRMs are characterized separately and compared, as well as evaluated for goodness-of-fit and adequacy of the models to the data. Furthermore, testing for the significance of the correlations of the fitted models to the variables is performed using analysis of variance (ANOVA) metrics such as the F-statistic and p-value.

RESULTS

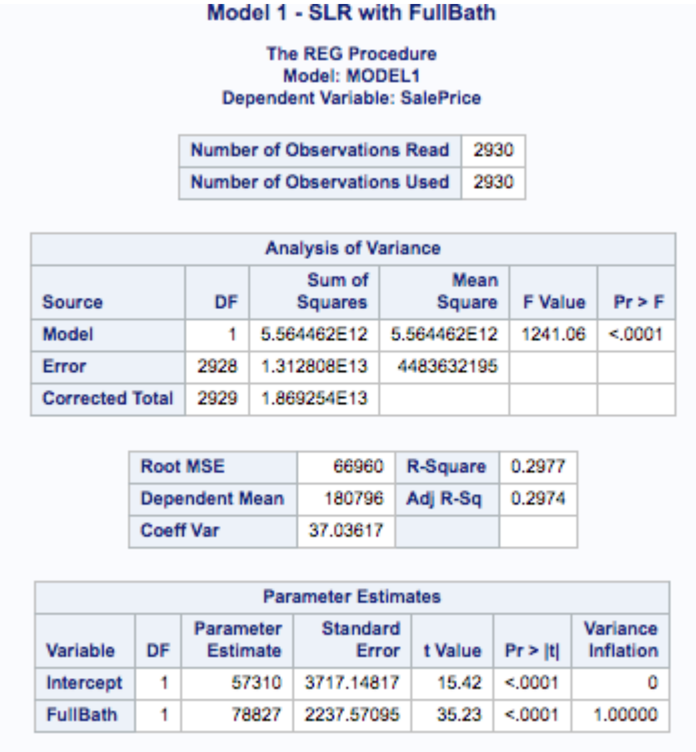
Part A: Simple Linear Regression Models

1. Fitting simple linear model with R2=0.5

For the purposes of this section, Y= SalePrice X=FullBath (R2=0.54560). In the formalism of the linear regression equation this can be stated as:

y=b0 + b1x1 + e, where b0 = y-intercept and b1 = slope and x=FullBath; y=SalePrice; e = error term. Thus,
SalePrice = 57310 + 78827FullBath, that is,
y = 57310 + 78827x

The table of parameter estimates shown below allows one to derive the formula of the linear regression model for SalePrice and FullBath (noted above). From the formula slope for the fitted model (above), we interpret this as: for every 1 FullBath in a property there is an increase in SalePrice of $78,827.



Model 1 - SLR with FullBath

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

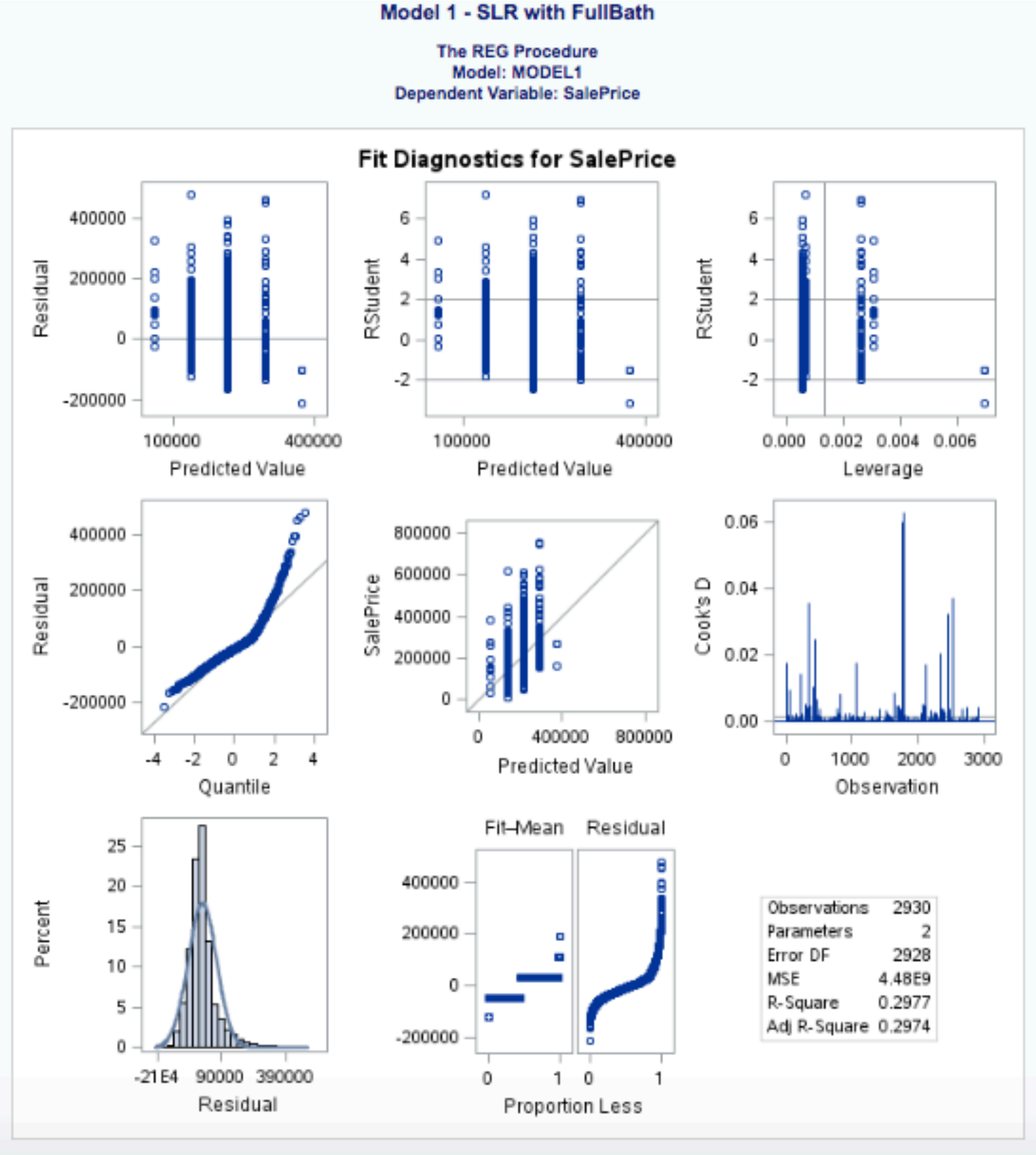
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.564462E12	5.564462E12	1241.06	<.0001
Error	2928	1.312808E13	4483632195		
Corrected Total	2929	1.869254E13			

Root MSE	66960	R-Square	0.2977
Dependent Mean	180796	Adj R-Sq	0.2974
Coeff Var	37.03617		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	57310	3717.14817	15.42	<.0001	0
FullBath	1	78827	2237.57095	35.23	<.0001	1.00000

From the analysis of variance (ANOVA) table, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SalePrice and FullBath. The R-squared value is approximately 0.3, indicating that approximately 30% of the variance in SalePrice is explained by FullBath. Reviewing the diagnostics for goodness-of-fit (see graphic below), the quantile-quantile plot of residuals (QQplot; central left plot), shows deviation away from the ideal line, that is predicted to be obtained if it were a normal distribution. Instead there is a deviation of points, especially at the upper end away from normality.



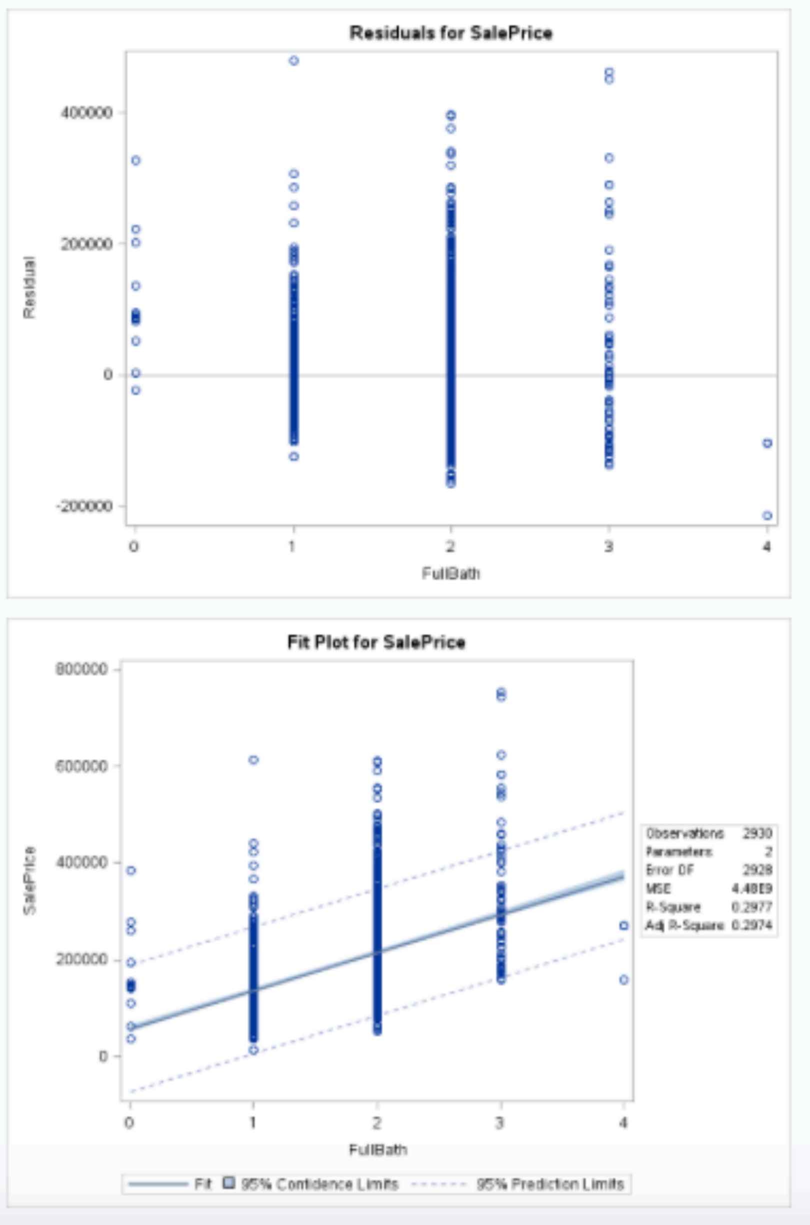
Model 1 - SLR with FullBath

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Fit Diagnostics for SalePrice

Observations	2930
Parameters	2
Error DF	2928
MSE	4.48E9
R-Square	0.2977
Adj R-Square	0.2974

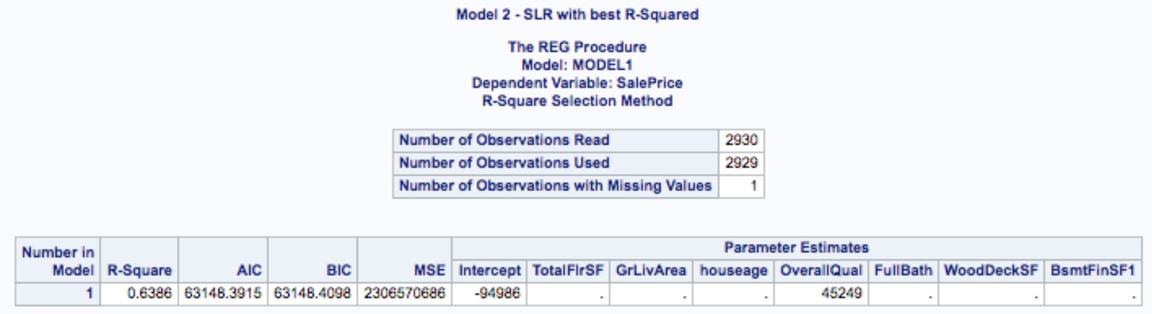
Looking at the residuals (top left panel above) rather than having a random distribution points around the zero level, there seems to be an elliptical distribution of residuals (more so above the zero line than below) and more toward the lower end of the predicted value. The distribution of residuals (plot panel on the bottom left above) shows the distribution is skewed to the right. The Cook's Distance plot (center right panel above) for outliers shows that for all observations the threshold is below 1.0, suggesting that outliers are not a significant issue. Most of the observations fall outside the 95% confidence interval for the fitted regression model, see bottom panel plot below. Plotting the 95% prediction interval indicates a greater proportion of the observations are within the predictions bounds, however, there is a significant number of observations outside the 95% prediction interval. So for a 2-bath home using the prediction interval there is a variation in house price from $20,000 to $250,000, which is not that helpful in the real estate market.



Conclusion: There is a reasonable correlation in the linear regression model (LRM) for FullBath and SalePrice, though only 30% of the variance in SalePrice is explained by FullBath. The LRM for this fitted model does pass through the mean SalePrice (~$180,000, derived from PROC MEANS in Assignment 1) for a house with 2 bathrooms (FullBath mean=1.5), which is good as it is consistent with the underlying observations from Assignment 1. Outliers are less of an issue for FullBath, the goodness-of-fit (GOF) is not so good as the observations deviate from a normal distribution.

2. Fitting simple linear model with best LRM

In attempting to find the 'best' linear model from the following variables, TotalFlrSF, GrLivArea, houseage, OverallQual, FullBath, WoodDeckSF, BsmtFinSF1, all LRMs we scanned and produced OverallQual as the best model (see table below):



Model 2 - SLR with best R-Squared

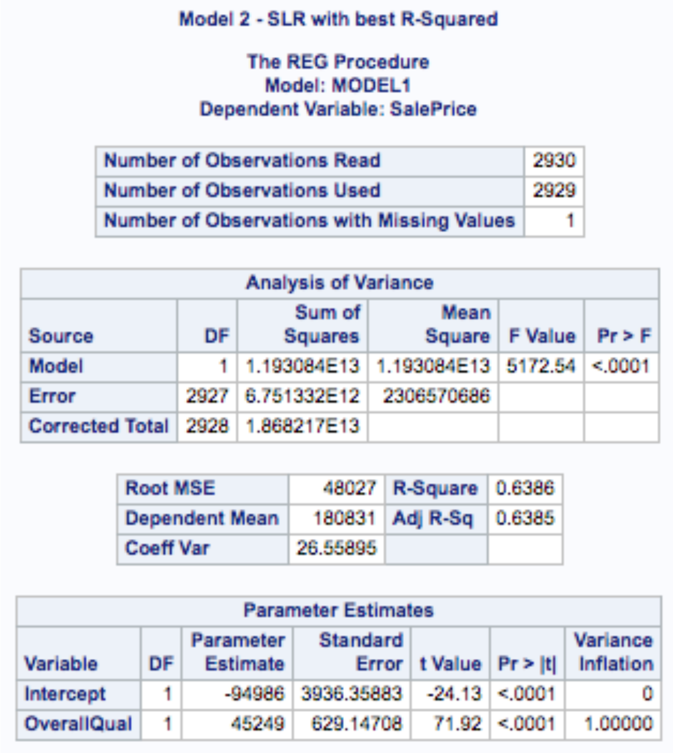
The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice
R-Square Selection Method

Number of Observations Read	2930
Number of Observations Used	2929
Number of Observations with Missing Values	1

Number in					Parameter Estimates							
Model	R-Square	AIC	BIC	MSE	Intercept	TotalFlrSF	GrLivArea	houseage	OverallQual	FullBath	WoodDeckSF	BsmtFinSF1
1	0.6386	63148.3915	63148.4098	2306570686	-94986	.	.	.	45249	.	.	.

The R-squared for the fitted LRM was approximately 64% (see table below), which is good and suggests that 64% of the variance in SalePrice is explained by OverallQual. The equation of the fitted, single LRM is described below. In the format y=b0 + b1x1 + e, where x=OverallQual, y=SalePrice and e = error term.
SalePrice = -94986 + 45249*OverallQual that is,
y = -94986 + 45249x

The table of parameter estimates shown below allows one to derive the formula of the linear regression model for SalePrice and OverallQual (noted above). From the formula slope for the fitted model (above), we interpret this as: for every 1 unit increase in OverallQual in a property there is an increase in SalePrice of $45,249.



Model 2 - SLR with best R-Squared

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

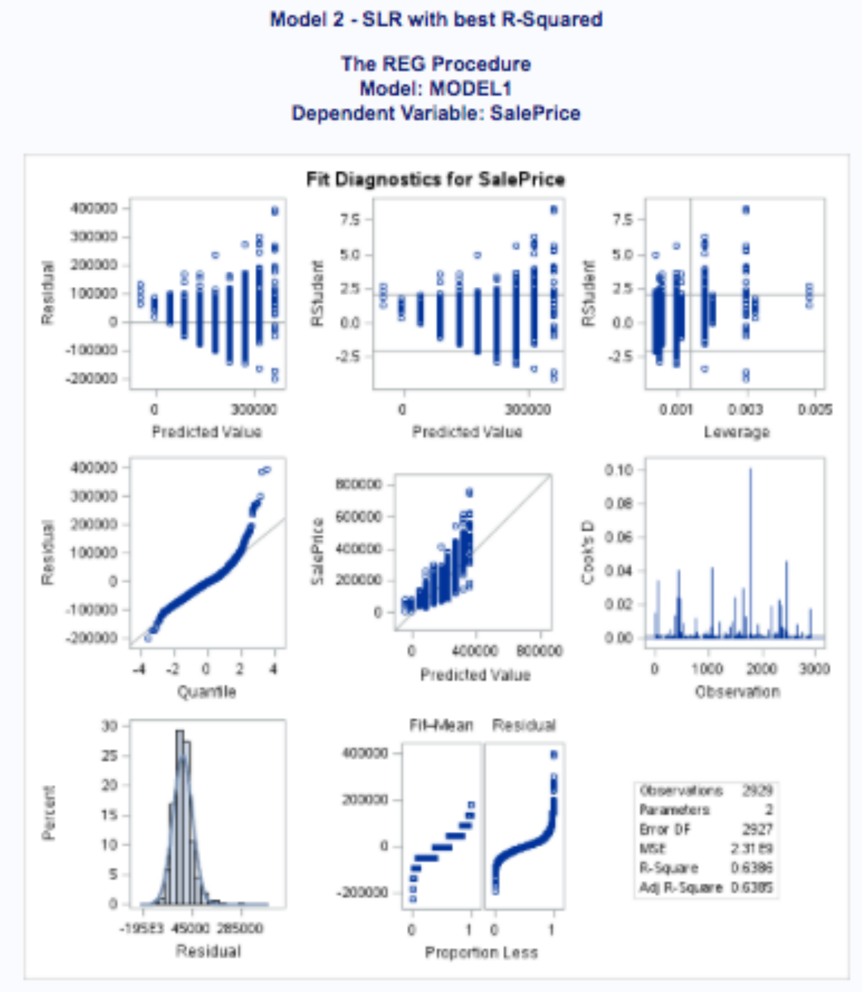
Number of Observations Read	2930
Number of Observations Used	2929
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.193084E13	1.193084E13	5172.54	<.0001
Error	2927	6.751332E12	2306570686		
Corrected Total	2928	1.868217E13			

Root MSE	48027	R-Square	0.6386
Dependent Mean	180831	Adj R-Sq	0.6385
Coeff Var	26.55895		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-94986	3936.35883	-24.13	<.0001	0
OverallQual	1	45249	629.14708	71.92	<.0001	1.00000

From the analysis of variance (ANOVA) table (above), to assess the quality of the fitted model, the F-statistic is significantly high (5172) and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SalePrice and OverallQual. The R-squared value is approximately 0.64 and significant in this single LRM model, this is the basis on which 'best' model is conferred. Reviewing the diagnostics for goodness-of-fit (see graphic below), the quantile-quantile plot of residuals (QQplot; central left plot), shows deviation away from the ideal or normal line. Instead there is a deviation of points, especially at the upper end away from normality.



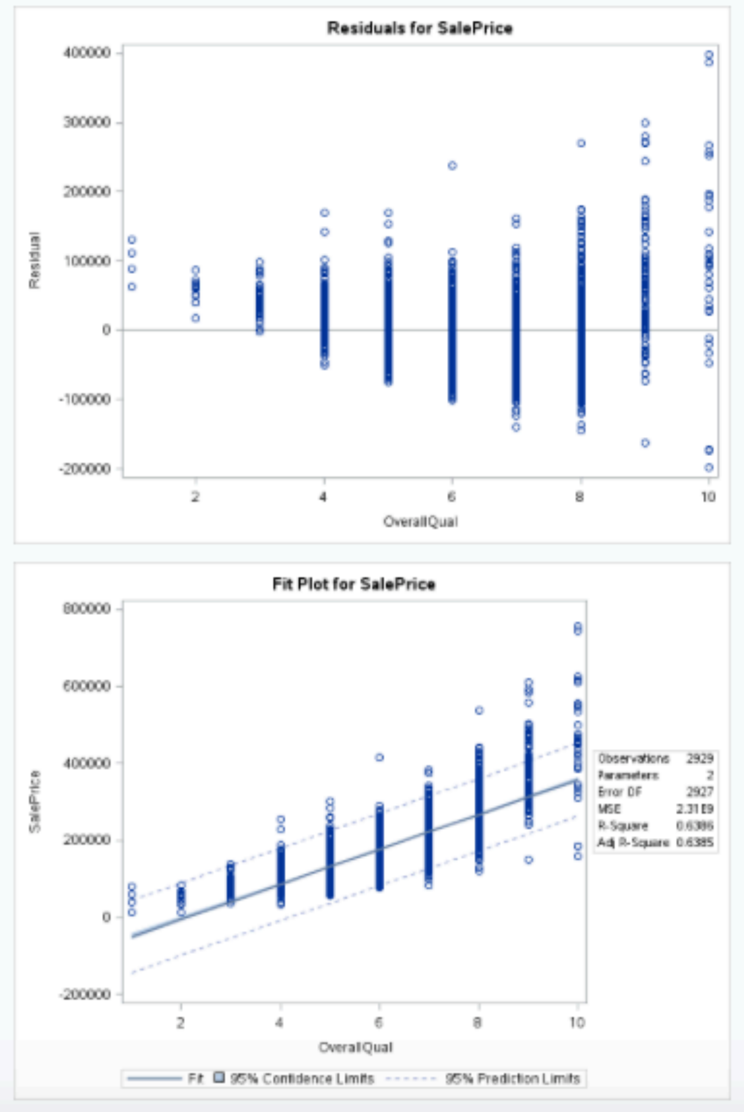
Model 2 - SLR with best R-Squared

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Fit Diagnostics for SalePrice

Observations	2929
Parameters	2
Error DF	2927
MSE	2.31E9
R-Square	0.6386
Adj R-Square	0.6385

Looking at the residuals (top left panel above) rather than having a random distribution points around the zero level, there seems to be funnel-shaped distribution of residuals (lower on the left and increasing toward the right). The distribution of residuals (plot panel on the bottom left above) shows the distribution is skewed to the right. The Cook's Distance plot (center right panel above) for outliers shows that for all observations the threshold is significantly below 1.0, suggesting that outliers are not a significant issue. Most of the observations fall outside the 95% confidence interval for the fitted regression model, see bottom panel plot below. Plotting the 95% prediction interval indicates a greater proportion of the observations are within the predictions bounds (compared to FullBath, for example), however, there are some observations outside the 95% prediction interval. Unfortunately, once again despite being the 'best' model, for an average OveralQual score of 5, there is a variation in SalePrice from $0 to $180,000, which again is not that helpful in the real estate market.



Observations	2929
Parameters	2
Error DF	2927
MSE	2.31E9
R-Square	0.6386
Adj R-Square	0.6385

Conclusion: There is a stronger correlation in the linear regression model (LRM) for OverallQual and SalePrice (compared to FullBath), with approximately 64% of the variance in SalePrice being explained by OverallQual (see table inset in the 'Fix Plot for SalePrice' above). The LRM for this fitted model does pass through the mean SalePrice (~$180,000, derived from PROC MEANS in Assignment 1) for a house with an OverallQual score of 6 (OverallQual mean=6.09), which is good as it is consistent with all of the underlying observations for OverallQual and SalePrice from Assignment 1. Regarding GOF the residuals are not randomly distributed and are skewed slightly. While outliers are less of an issue for OverallQual, most of the observations for OverallQual fall within 95% prediction interval range, and a smaller proportion are outside that range (compared to FullBath, for example).

3. Simple LRM with a categorical variable

I would use KichenQual as a categorical variable. At this stage in the course, however, the fitted model needs to be 'tweaked' into accepting categorical variables, which has not been covered. Thus, this part of assignment 2 will not be addressed here.

Conclusion: pending further analysis

4. Comparison of LRMs: *Of the above 2 models, which one fits better? On what criteria are you assessing the model fit?*

Conclusion: The single, linear regression model (LRM) with OverallQual is better than FullBath based on the R-squared of the two models, that is 0.6385 *versus* 0.2977, respectively. Secondarily, the distribution of the residuals (*e.g.* QQplot) for OverallQual *versus* FullBath is more normal, with more observations falling on the normal distribution control line.

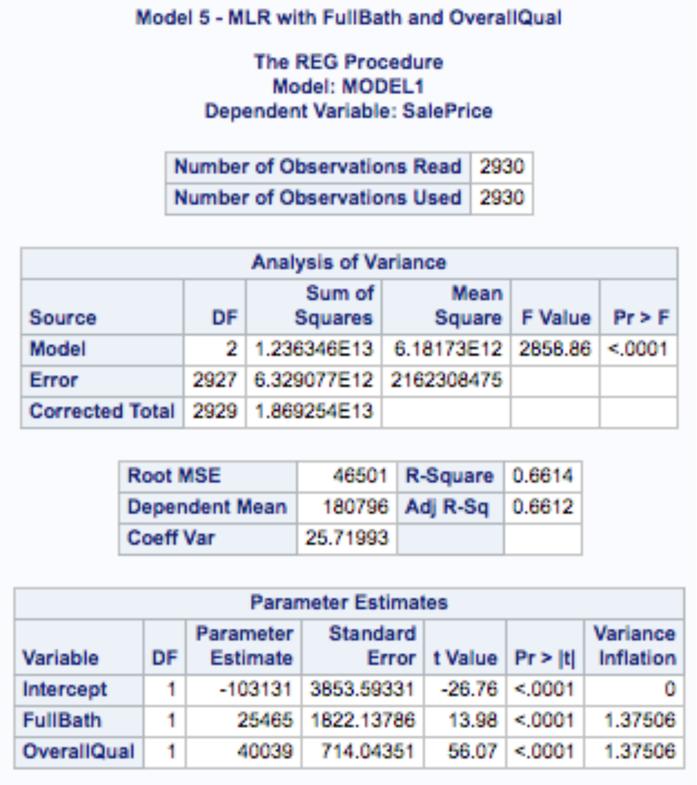
Part B: Multiple Linear Regression Models

5. Fitting a MLR model
The two variables used in parts (1) and (2) earlier were used to fit a multiple linear regression model (MRM) to predict SalePrice. These X variables are FullBath and OverallQual, which will be referred to as x1 and x2, respectively.

In the formalism of the multiple regression equation this can be stated as:

y=b0 + b1x1 + b2x2 where b0 = y-intercept and b1 = slope coefficient (FullBath) and x1=FullBath; b2 = slope coefficient (OverallQual) and x2=OverallQual; y=SalePrice; Thus, the MRM equation is
SalePrice = -103131 + 25465FullBath + 40039OverallQual, that is,
y = -103131 + 25465x1 + 40039x2

The table of parameter estimates shown below allows one to derive the formula of the multiple regression model for SalePrice, FullBath and OverallQual (noted above). From the formula slope coefficients (b1 and b2) for this fitted MRM (above), we interpret this as: for every 1 FullBath in a property there is an increase in SalePrice of $25,465 and for every unit increase in OverallQual there is a $40,039 increase in SalePrice. Interestingly, in the simple fitted models reported earlier, the slope coefficients were FullBath and OverallQual were 78,827 and 45,249, respectively. The slope coefficient of FullBath has decreased disproportionately more than OverallQual in the MRM *versus* the simple LRM.



Model 5 - MLR with FullBath and OverallQual

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

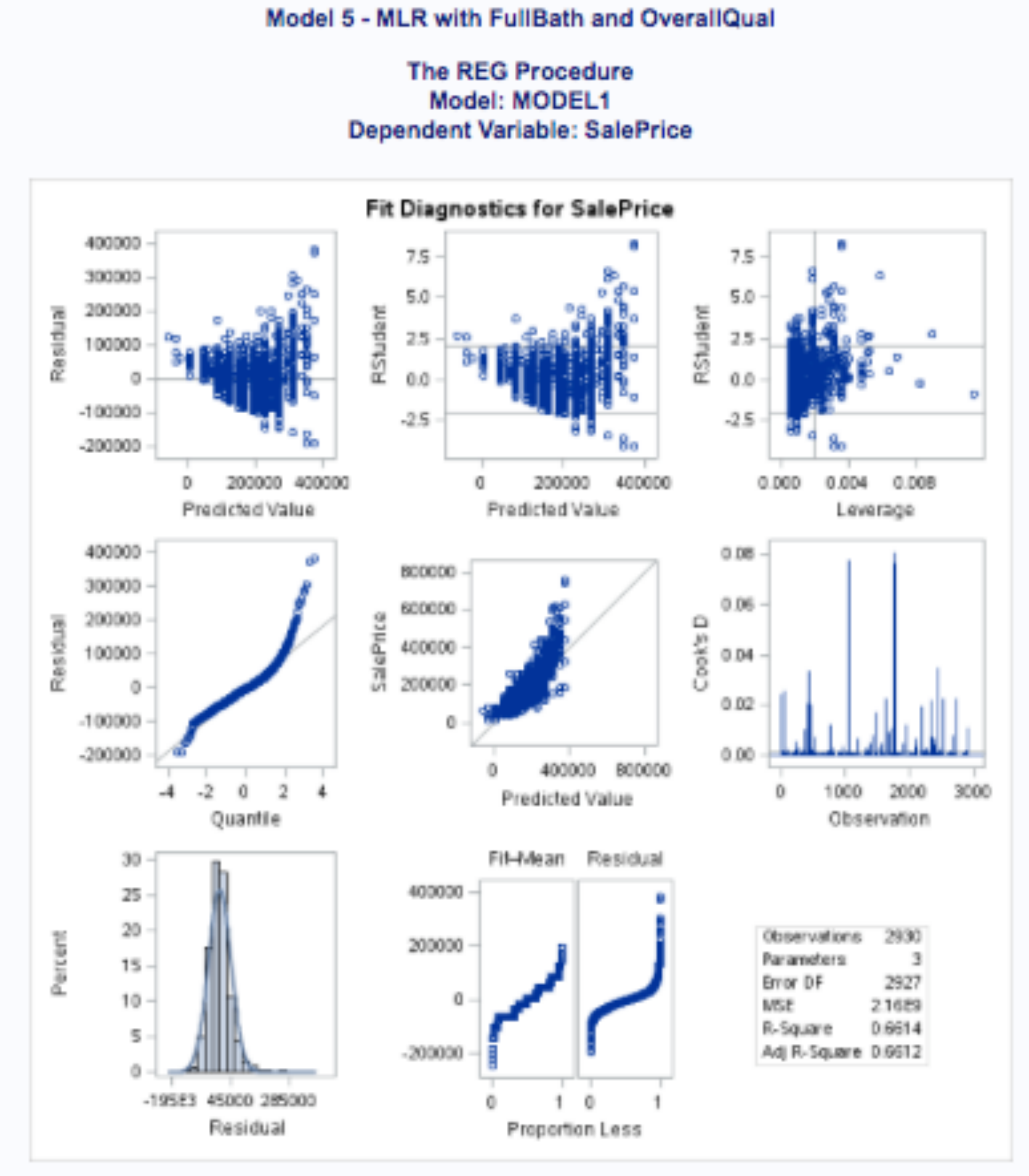
Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.236346E13	6.18173E12	2858.86	<.0001
Error	2927	6.329077E12	2162308475		
Corrected Total	2929	1.869254E13			

Root MSE	46501	R-Square	0.6614
Dependent Mean	180796	Adj R-Sq	0.6612
Coeff Var	25.71993		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-103131	3853.59331	-26.76	<.0001	0
FullBath	1	25465	1822.13786	13.98	<.0001	1.37506
OverallQual	1	40039	714.04351	56.07	<.0001	1.37506

From the analysis of variance (ANOVA) table (above), to assess the quality of the fitted model, the F-statistic is significantly high (2858) and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SalePrice and FullBath + OverallQual. The R-squared value is approximately 0.66 and significant in this MRM model. The adjusted R-squared for this MRM model is slightly more than for the simple LRM for OverallQual, that is 0.6612 *versus* 0.6385, respectively, and significantly more than the single LRM for FullBath (adj-R^2=0.2974). When comparing models of different sizes, we generally use a adjusted R-Squared, which provides a trade-off between model fit and model complexity. Thus, with respect to the adjusted R-squared value, the MRM with FullBath + OverallQual is better than either of the single LRMs for FullBath and OverallQual alone. Reviewing the diagnostics for goodness-of-fit for the MRM (see scatterplot matrix graphic below), the quantile-quantile plot of residuals (QQplot; central left plot), shows deviation away from the ideal or normal line, especially at the upper end away, although at the lower and mid portions the fit to normal is reasonable. The distribution bar chart of residuals (bottom right panel below) confirms that the distribution is skewed slightly to the right. Looking at the scatter of the residuals (upper left panel below), there is again a funnel-shaped distribution increasing to the right so not necessarily random. With regard to the Cook's D plot (middle right panel) the thresholds are below 1.0 suggesting that outliers are less of an issue for this MRM.



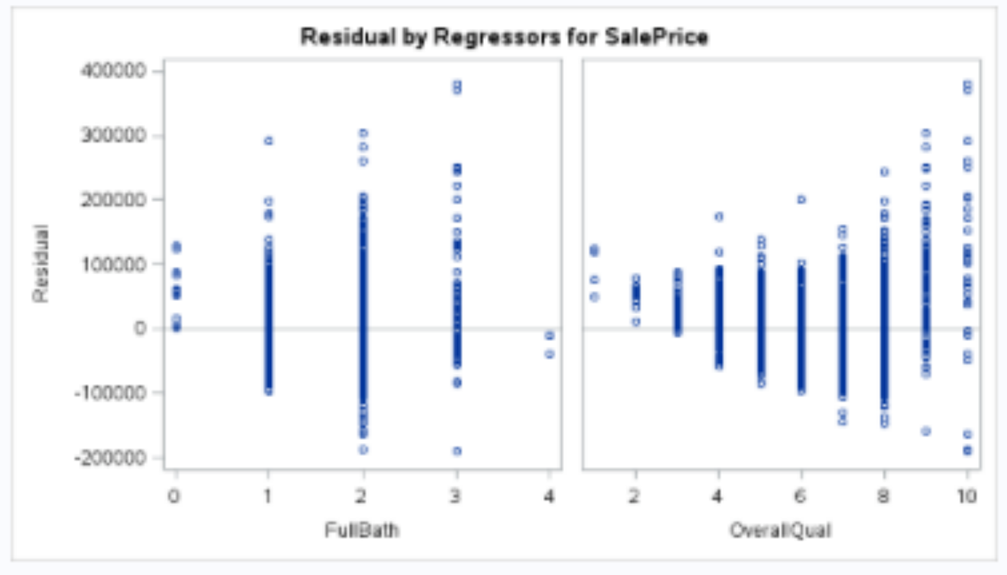
Model 5 - MLR with FullBath and OverallQual

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Fit Diagnostics for SalePrice

Observations	2930
Parameters	3
Error DF	2927
MSE	2.16E9
R-Square	0.6614
Adj R-Square	0.6612

Looking at the residuals for FullBath and OverallQual (see below) in this fitted MRM separately for SalePrice, they display non-random distribution.



Conclusion: The multiple regression model (MRM) for FullBath and OverallQual (adj-R^2=0.6612) is better than either of the single LRMs for FullBath (adj-R^2=0.2974) and OverallQual (adj-R^2=0.6385) alone. Indicating that with FullBath and OverallQual approximately 66% of the variance in SalePrice is explained by these two variables.

6. Adding an third variable to the MLR model

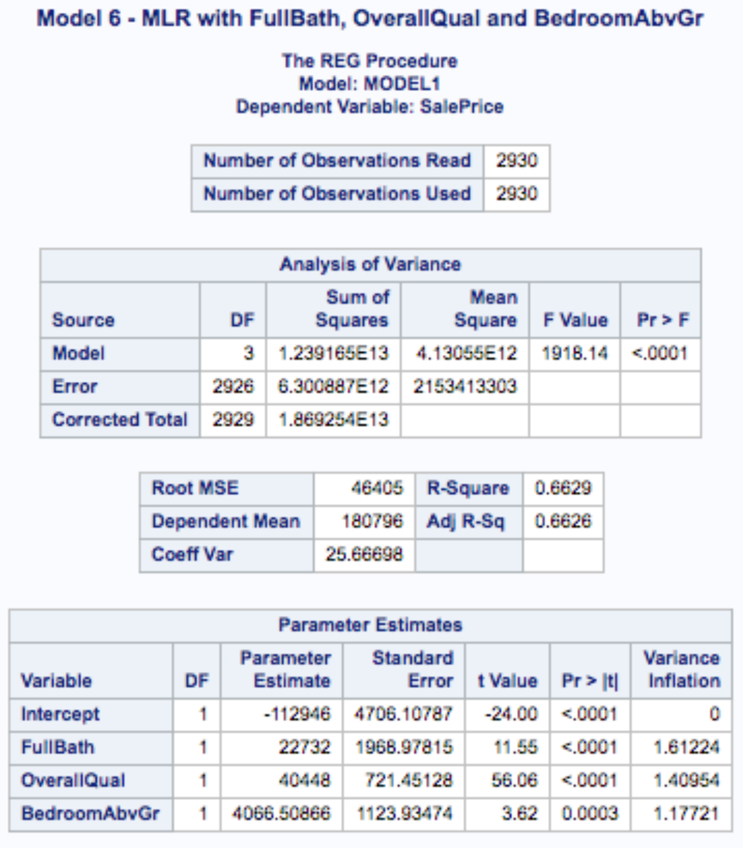
The MRM above was updated to incorporate a variable with the smallest correlation coefficients between X and Y from Assignment 1, which was BedroomAbvGr (0.14391).

The three variables were used to fit a multiple linear regression model (MRM) to predict SalePrice. These X variables are FullBath (x1) and OverallQual (x2) and BedroomAbvGr (x3). In the formalism of the multiple regression equation this can be stated as:

y=b0 + b1x1 + b2x2 + b3x3 + e, where b0 = y-intercept and b1 = slope coefficient (FullBath), x1=FullBath; b2 = slope coefficient (OverallQual), x2=OverallQual; b3 = slope coefficient (BedroomAbvGr), x3=BedroomAbvGr; y=SalePrice; e= error term.

Thus, the MRM equation is:
SalePrice = -112946 + 22732FullBath + 40448OverallQual + 4066BedroomAbvGr, that is,
y = -112946 + 22732x1 + 40448x2 + 4066x3

The table of parameter estimates shown below allows one to derive the formula of the multiple regression model for SalePrice, FullBath, OverallQual and BedroomAbvGr (noted above). From the formula slope coefficients (b1, b2 and b3) for this fitted MRM (above), we interpret this as: for every 1 FullBath in a property there is an increase in SalePrice of $22,732; for every unit increase in OverallQual there is a $40,448 increase in SalePrice; and for every bedroom above ground the SalePrice increases by $4,066. Interestingly, the slope coefficients from the previous MRM for FullBath + OverallQual reported earlier, did not change much in the MRM with three variables, and the BedroomAbvGr added the least dollar value response in SalePrice (Y), namely $4,066.



Model 6 - MLR with FullBath, OverallQual and BedroomAbvGr

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.239165E13	4.13055E12	1918.14	<.0001
Error	2926	6.300887E12	2153413303		
Corrected Total	2929	1.869254E13			

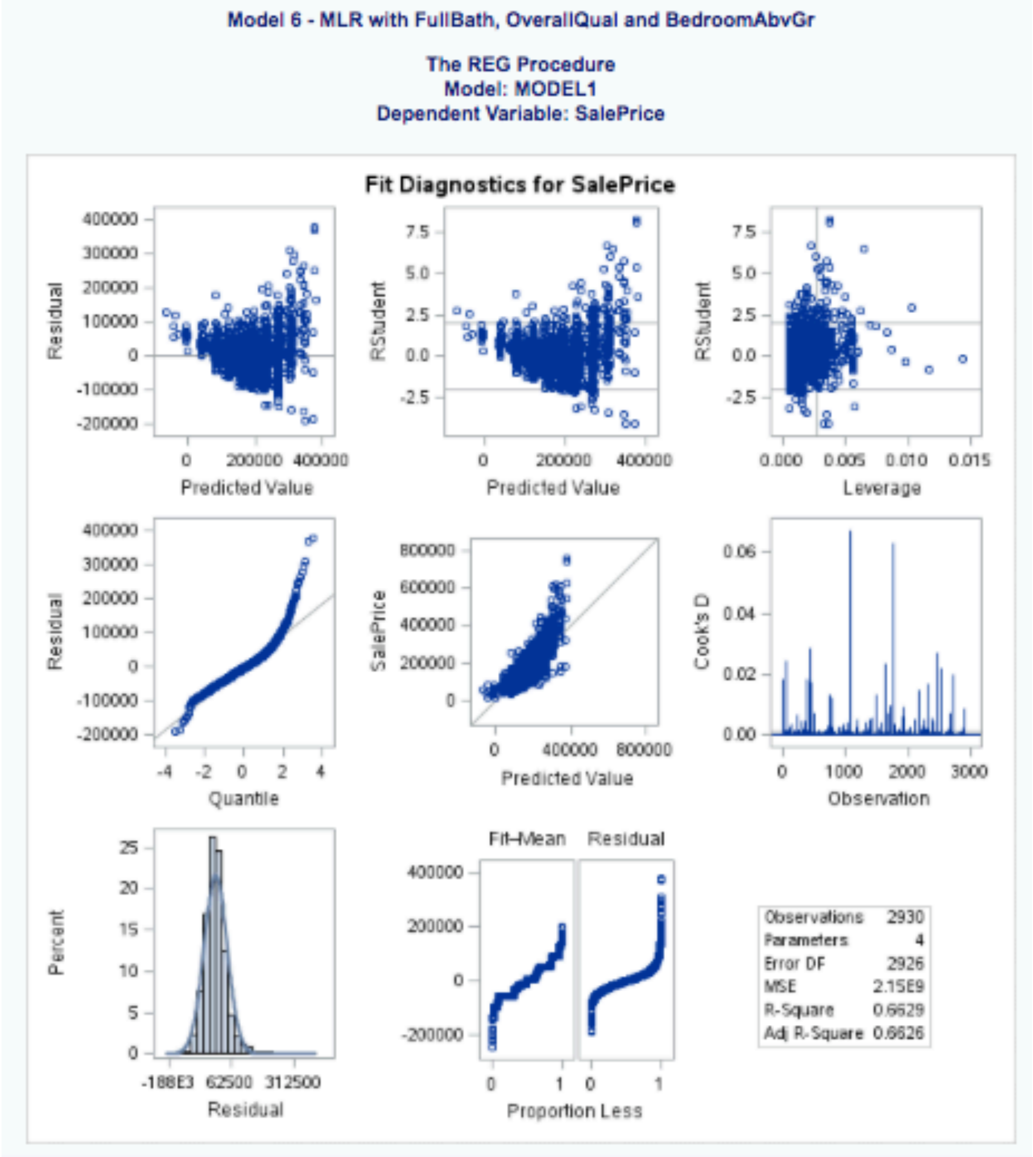
Root MSE	46405	R-Square	0.6629
Dependent Mean	180796	Adj R-Sq	0.6626
Coeff Var	25.66698		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-112946	4706.10787	-24.00	<.0001	0
FullBath	1	22732	1968.97815	11.55	<.0001	1.61224
OverallQual	1	40448	721.45128	56.06	<.0001	1.40954
BedroomAbvGr	1	4066.50866	1123.93474	3.62	0.0003	1.17721

From the analysis of variance (ANOVA) table (above), to assess the quality of the fitted model, the F-statistic is significantly high (1918) and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SalePrice and FullBath + OverallQual + BedroomAbvGr. The adjusted R-squared value for this 3-variable model is 0.6626 which is essentially the same as the previous MRM with only 2-variables (FullBath+OverallQual). Thus, merely adding more variables does not necessarily improve the fit drastically to the model.

Reviewing the diagnostics for goodness-of-fit for the MRM (see scatterplot matrix graphic below), the quantile-quantile plot of residuals (QQplot; central left plot), shows deviation away from the ideal or normal line, especially at the upper end away, although at the lower and mid portions the fit to normal is reasonable. The distribution bar chart of residuals (bottom right panel below) confirms that the distribution is skewed slightly to the right. Looking at the scatter of the residuals (upper left panel below), there is again a funnel-shaped distribution increasing to the right, so it is not necessarily random. With regard to the Cook's D plot (middle right panel) the thresholds are below 1.0 suggesting that outliers are less of an issue for this 3-variable MRM.



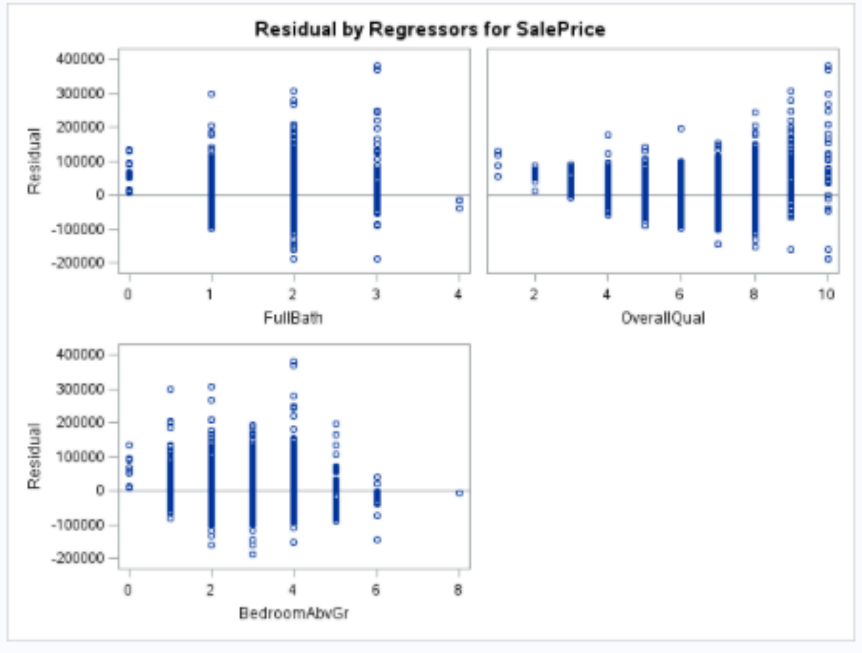
Model 6 - MLR with FullBath, OverallQual and BedroomAbvGr

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Fit Diagnostics for SalePrice

Observations	2930
Parameters	4
Error DF	2926
MSE	2.15E9
R-Square	0.6629
Adj R-Square	0.6626

Looking at the residuals for FullBath, OverallQual and BedroomAbvGr (see below) in this fitted 3-variable MRM separately for SalePrice, they all display non-random distributions.



Conclusion: The 3-variable multiple regression model (MRM) for FullBath, OverallQual and BedroomAbvGr (adj-R^2=0.6626) is basically the same as the 2-variable fitted model for FullBath and OverallQual (adj-R^2=0.6612). Consequently just by adding more variables does not necessarily improve the fit to the model.

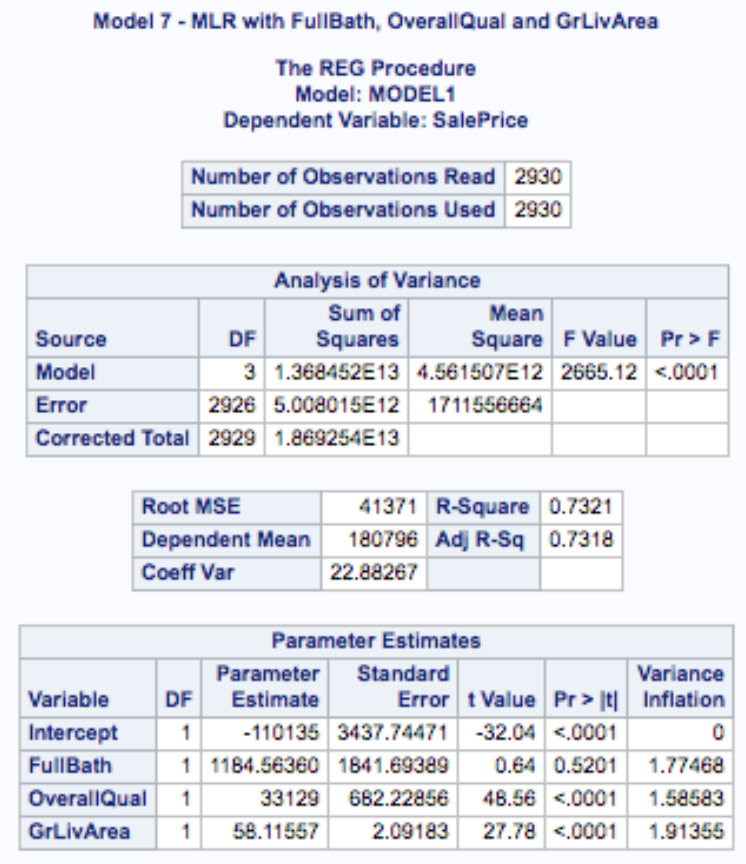
7. Adding an different third variable to the MLR model (Additional)

The MRM above was updated to incorporate a variable with one of the largest correlation coefficients between X and Y from Assignment 1, which was GrLivArea (0.70678). The three variables were used to fit a multiple linear regression model (MRM) to predict SalePrice. These X variables are FullBath (x1) and OverallQual (x2) and GrLivArea (x3). In the formalism of the multiple regression equation this can be stated as:

y=b0 + b1x1 + b2x2 + b3x3 + e, where b0 = y-intercept and b1 = slope coefficient (FullBath), x1=FullBath; b2 = slope coefficient (OverallQual), x2=OverallQual; b3 = slope coefficient (GrLivArea), x3= GrLivArea; y=SalePrice; e= error term.

Thus, the MRM equation is:
SalePrice = -110135 + 1184FullBath + 33129OverallQual + 58GrLivArea, that is,
y = -110135 + 1184x1 + 33129x2 + 58x3

The table of parameter estimates shown below allows one to derive the formula of the multiple regression model for SalePrice, FullBath, OverallQual and GrLivArea (noted above). From the formula slope coefficients (b1, b2 and b3) for this fitted MRM (above), we interpret this as: for every 1 FullBath in a property there is an increase in SalePrice of $1,184; for every unit increase in OverallQual there is a $33,129 increase in SalePrice; and for 1 sq. foot increase in Above Ground Living Area the SalePrice increases by $58.



Model 7 - MLR with FullBath, OverallQual and GrLivArea

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1.368452E13	4.561507E12	2665.12	<.0001
Error	2926	5.008015E12	1711556664		
Corrected Total	2929	1.869254E13			

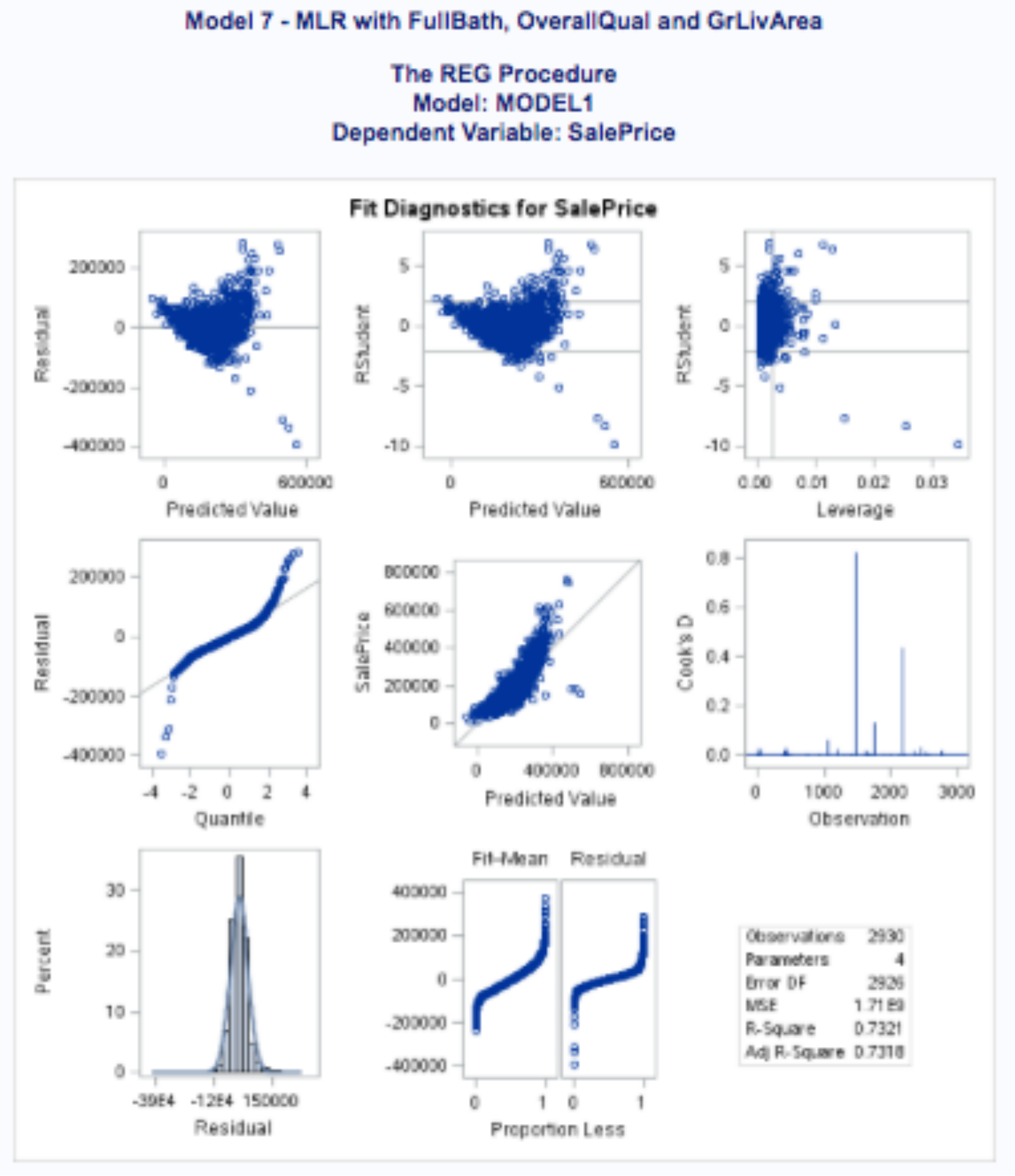
Root MSE	41371	R-Square	0.7321
Dependent Mean	180796	Adj R-Sq	0.7318
Coeff Var	22.88267		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-110135	3437.74471	-32.04	<.0001	0
FullBath	1	1184.56360	1841.69389	0.64	0.5201	1.77468
OverallQual	1	33129	682.22856	48.56	<.0001	1.58583
GrLivArea	1	58.11557	2.09183	27.78	<.0001	1.91355

The R-squared value for this 3-variable model is 0.7329 (which is better than the previous 3-variable MRM with (FullBath+OverallQual+BedroomAbvGr). This indicates that for these 3-variables 73% of the variance in SalePrice is accounted for. Thus, the model can improve by adding additional variables but it depends on what they are.

Reviewing the diagnostics for goodness-of-fit for the MRM (see scatterplot matrix graphic below), the quantile-quantile plot of residuals (QQplot; central left plot), shows deviation away from the ideal or normal line, especially at the upper end away, although at the lower and mid portions the fit to normal is reasonable. The distribution bar chart of residuals (bottom right panel below) indicates a reasonable normal distribution. Looking at the scatter of the residuals (upper left panel below), there is better distribution though there is some skew over to the left side with some observations on the lower right side, so it is not entirely random. With regard to the Cook's D plot (middle right panel), the threshold is just below 1.0, suggesting that outliers may becoming a potential issue for this 3-variable MRM.



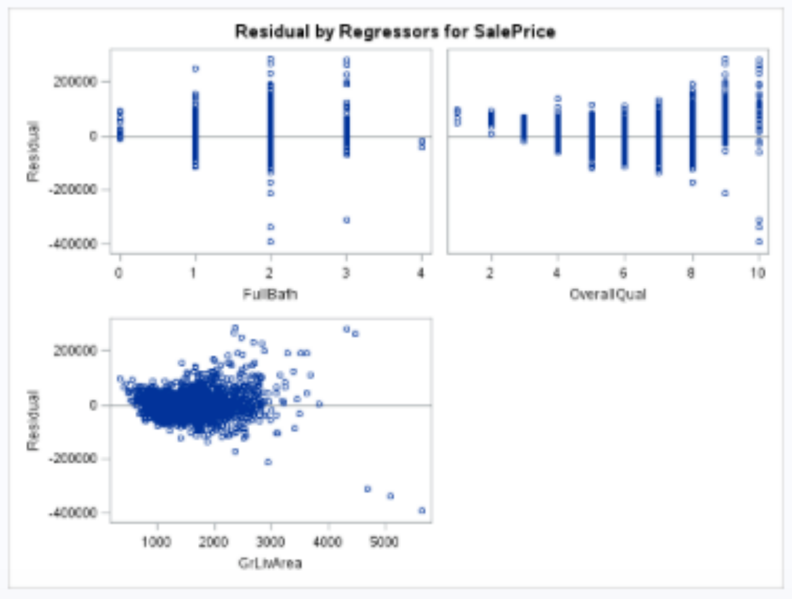
Model 7 - MLR with FullBath, OverallQual and GrLivArea

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Fit Diagnostics for SalePrice

Observations	2930
Parameters	4
Error DF	2926
MSE	1.71E9
R-Square	0.7321
Adj R-Square	0.7318

Looking at the residuals for FullBath, OverallQual and GrLivArea (see below) in this fitted 3-variable MRM separately for SalePrice, they all display some non-random distributions.



Conclusion: The R-squared value for this 3-variable model is 0.7329, which is one of the best correlations to the fitted models we have seen. Consequently, 2-variable MRM can be improved with one containing 2-variables (based on assessment of the R-squared value) but it depends on what variable is added.

CONCLUSIONS

The assignment showed that while simple, linear regression models using single predictor variable is useful, the ability to add more than one predictor can add value to the overall goal which is develop suitable metrics for accurately predicting SalePrice in the Ames, Iowa housing market. The R^2 coefficient was used to compare, among other criteria, how single LRMs could be improved into MRMs. Single regression models had fitted models for FullBath (R^2=0.2977) and OverallQual (R^2=0.6386). By combining these continuous variables together in a 2-variable MRM a fitted model of FullBath and OverallQual could be generated with an R^2 coefficient of 0.6614. Moreover, depending on the nature of the additional variable being added, by creating a 3-variable MRM, such as with FullBath, OverallQual, and GrLivArea, a better-fit model could be produced (R^2=0.7329). Thus, with the latter model approximately 73% of the variance in SalePrice could be account for using these 3-variables.

What conditions or situations would make you think a model is not appropriately specified?

(a) Models where correlations between predictor and response variables are poor such that the ANOVA indicates (*e.g.* small F-statistic and high p-values) to accept the null hypothesis (no correlation) would be inappropriate.

(b) There may be situations where the plots of residuals *versus* predicted values display an unusual format (*e.g.* sinusoidal)

(c) Situations where most of the observations are outside the prediction interval and randomly scattered with regard response and predictor variables.

What do you consider to be next steps in the modeling process?

(a) In this assignment we did not deal with incorporating categorical variables, these also need to be evaluated in addition to continuous variables.

(b) Systematically explore all the combinations of variables used before both individually to build the simple LRMs and from the results build more and complex MLR models with more continuous and categorical variables

(c) Apply some sort of transformation on the variables to reduce scatter and improve the fit of the model, as well as have confidence/prediction ranges such that the observations fall within these limits.

(d) Address whether there are some outliers that are adversely affecting the fitted models in the simple LRMs or the MRMs. If so, should these outliers be removed to produce a better fitted model?