410-57, Assignment #5
Anamitra Bhattacharyya

INTRODUCTION

This assignment specifically deals with building linear regression models (LRMs), using 'dummy' variables for use with their cognate categorical variables, and variable selection for model building. 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. Using a variety of automated variable selection procedures, in combination with assessment of the quality and significance of these models a more refined fitted model was development and tested using training and test data sets. This resulted in removal of some statistically insignificant variables and operational validation to assess the applicability of the model to develop a business policy relating to prediction of house prices in Ames, Iowa.

RESULTS

Part A: Dummy Coding of Categorical Variables

1. Categorical variable

Choosing HouseStyle as a categorical variable (but not using dummy variables) we generate an equation of the form (from the output table below):
y = 147706 (b0; intercept) + 30045(x) so the fitted y-hat values will be 147706 + 177751 (*i.e.* 147706+30045) + 207796 (*i.e.* 177751 +30045).

Model 1 - SLR with StyleCategory

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	1.280825E12	1.280825E12	215.39	<.0001
Error	2928	1.741171E13	5946623126		
Corrected Total	2929	1.869254E13			

Root MSE	77114	R-Square	0.0685
Dependent Mean	180796	Adj R-Sq	0.0682
Coeff Var	42.65267		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	147706	2667.07709	55.38	<.0001
StyleCategory	1	30045	2047.19942	14.68	<.0001

In a regression model the fitted model regression line should go through the mean values for x and y (see table below). However, these y-hat values (*e.g.* \$147,706, \$177,751, \$207,796) are not an exact match with the mean SalePrice (Y) values (*e.g.* \$146,485, \$176,699, \$206,990).

The MEANS Procedure

StyleCategory=0

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
576	146485.30	48102.08	37900.00	475000.00

StyleCategory=1

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1481	178699.88	81066.94	12789.00	615000.00

StyleCategory=2

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
873	206990.16	85349.91	40000.00	755000.00

Conclusion: When not using dummy variables but assigning random numbers to the categorical variables, the y-hat fitted model does not pass through mean Y-values.

2. Using dummy variables

We need to fit appropriate dummy variables to the categorical variable chosen. When this is done we obtain the equation of the form (from the output table below):

Model 2 - SLR with Style1 & Style2

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	1.283583E12	6.417916E11	107.91	<.0001
Error	2927	1.740895E13	5947712287		
Corrected Total	2929	1.869254E13			

Root MSE	77121	R-Square	0.0687
Dependent Mean	180796	Adj R-Sq	0.0680
Coeff Var	42.65658		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	146485	3213.39219	45.59	<.0001
Style1	1	32215	3787.07016	8.51	<.0001
Style2	1	60505	4139.90910	14.62	<.0001

y = 146485 + 32215*Style1 + 60505*Style2

The fitted model yhat values will be 146485, 178700, 206990; such that,

a) when Style1=0 & Style2=0, then yhat=<u>146485</u>=intercept (baseline)

b) when Style1=1 & Style2=0 then yhat = 146485 + 32215=<u>178700</u>;

c) when Style1=0 & Style2=0 then yhat = 146485 + 60505 = <u>206990</u>;

Coefficient of Style1 is 32215, which is the additional amount for a 1-story building on top of the 'baseline' price. The coefficient of Style2 is 60505, which is the additional amount for a 2-story building on top of the 'baseline' price.

Conclusion: These fitted values from this model with dummy variables are an exact match with the mean SalePrice (Y) values (*e.g.* $146485, $176,699, $206,990) from the second table shown above. Thus, using dummy variables produces a y-hat model plane that passes through the mean Y values exactly.

3. Report on hypothesis tests for betas

The hypothesis being tested the overall (full model) regression models is:

H0: beta1 = beta2 = 0

H1: at least one beta ≠ 0

The overall significance of this regression defined by the F-statistic is 107.91 and the p-value is < 0.0001, which is low. This result indicates to reject the null hypothesis for the full model suggesting that there is a correlation between SalePrice and HouseStyle.

Model 2 - SLR with Style1 & Style2

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	1.283583E12	6.417916E11	107.91	<.0001
Error	2927	1.740895E13	5947712287		
Corrected Total	2929	1.869254E13			

Root MSE	77121	R-Square	0.0687
Dependent Mean	180796	Adj R-Sq	0.0680
Coeff Var	42.65658		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	146485	3213.39219	45.59	<.0001
Style1	1	32215	3787.07016	8.51	<.0001
Style2	1	60505	4139.90910	14.62	<.0001

Looking at the individual t-tests in the parameter table, the null hypotheses for the Style 1 and Style2 are:

a) H0: beta1 (Style1) = 0
   H1: beta1 ≠ 0

b) H0: beta2 (Style2) = 0
   H1: beta2 ≠ 0

Conclusion: Since the p-values for each of these t-tests in the parameter table above are low (< 0.0001), the null hypothesis is rejected and conditions satisfy the alternate hypothesis in each case. Therefore, there is a correlation between house style (Style1 and Style2) and SalePrice response variable.

4. Other categorical variable

Added a Zoning categorical variable, the summary table of which is shown below:

Model 2 - SLR with Style1 & Style2

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	1.283583E12	6.417916E11	107.91	<.0001
Error	2927	1.740895E13	5947712287		
Corrected Total	2929	1.869254E13			

Root MSE	77121	R-Square	0.0687
Dependent Mean	180796	Adj R-Sq	0.0680
Coeff Var	42.65658		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	146485	3213.39219	45.59	<.0001
Style1	1	32215	3787.07016	8.51	<.0001
Style2	1	60505	4139.90910	14.62	<.0001

Part B: Automated variable selection

5. Alternate methods of variable selection

Six different methods of variable selection were chosen from a data set of variables comprising: TotalFlrSF, houseage, OverallQual, LotArea, BsmtFinSF1, TotalBath, HouseStyle and Zoning. The six variable selection methods comprised, R-Squared, adjusted R-Squared, Mallow's Cp, forward, backward and Stepwise. The results from the variable selection methods are shown below:

a) R-squared

R-SQUARED

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2928
Number of Observations with Missing Values	2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1.494423E13	1.494423E12	1169.59	<.0001
Error	2917	3.727148E12	1277733245		
Corrected Total	2927	1.867138E13			

Root MSE	35745	R-Square	0.8004
Dependent Mean	180795	Adj R-Sq	0.7997
Coeff Var	19.77119		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-82809	5637.07718	-14.69	<.0001	0
TotalFlrSF	1	64.84336	2.12040	30.58	<.0001	2.58871
houseage	1	-342.00989	33.28871	-10.27	<.0001	2.33013
OverallQual	1	24881	710.80846	35.00	<.0001	2.30376
LotArea	1	0.65258	0.09206	7.09	<.0001	1.19718
BsmtFinSF1	1	23.71716	1.77310	13.38	<.0001	1.49488
TotalBath	1	2089.36112	1171.30954	1.78	0.0746	2.77974
Style1	1	11760	1937.69157	6.07	<.0001	2.15077
Style2	1	-8161.74029	2193.53736	-3.72	0.0002	2.30730
Zone1	1	2767.85231	2722.08430	1.02	0.3093	2.95192
Zone2	1	-2409.55078	3238.73184	-0.74	0.4569	3.18866

b) Adjusted R-squared

ADJUSTED R-SQUARED

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2928
Number of Observations with Missing Values	2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	1.494352E13	1.660392E12	1299.68	<.0001
Error	2918	3.727855E12	1277537735		
Corrected Total	2927	1.867138E13			

Root MSE	35743	R-Square	0.8003
Dependent Mean	180795	Adj R-Sq	0.7997
Coeff Var	19.76968		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-84298	5269.70567	-16.00	<.0001	0
TotalFlrSF	1	64.83753	2.12022	30.58	<.0001	2.58868
houseage	1	-347.25380	32.53145	-10.67	<.0001	2.22566
OverallQual	1	24864	710.40624	35.00	<.0001	2.30151
LotArea	1	0.65564	0.09196	7.13	<.0001	1.19479
BsmtFinSF1	1	23.63868	1.76983	13.36	<.0001	1.48959
TotalBath	1	2144.77618	1168.84944	1.83	0.0666	2.76850
Style1	1	11871	1931.73295	6.15	<.0001	2.13789
Style2	1	-8099.93537	2191.79601	-3.70	0.0002	2.30399
Zone1	1	4318.73534	1750.37498	2.47	0.0137	1.22076

c) Mallow's Cp

MALLOWS CP

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2928
Number of Observations with Missing Values	2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	1.494352E13	1.660392E12	1299.68	<.0001
Error	2918	3.727855E12	1277537735		
Corrected Total	2927	1.867138E13			

Root MSE	35743	R-Square	0.8003
Dependent Mean	180795	Adj R-Sq	0.7997
Coeff Var	19.76968		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-84298	5269.70567	-16.00	<.0001	0
TotalFlrSF	1	64.83753	2.12022	30.58	<.0001	2.58868
houseage	1	-347.25380	32.53145	-10.67	<.0001	2.22566
OverallQual	1	24864	710.40624	35.00	<.0001	2.30151
LotArea	1	0.65564	0.09196	7.13	<.0001	1.19479
BsmtFinSF1	1	23.63868	1.76983	13.36	<.0001	1.48959
TotalBath	1	2144.77618	1168.84944	1.83	0.0666	2.76850
Style1	1	11871	1931.73295	6.15	<.0001	2.13789
Style2	1	-8099.93537	2191.79601	-3.70	0.0002	2.30399
Zone1	1	4318.73534	1750.37498	2.47	0.0137	1.22076

d) Forward

FORWARD

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2928
Number of Observations with Missing Values	2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1.494423E13	1.494423E12	1169.59	<.0001
Error	2917	3.727148E12	1277733245		
Corrected Total	2927	1.867138E13			

Root MSE	35745	R-Square	0.8004
Dependent Mean	180795	Adj R-Sq	0.7997
Coeff Var	19.77119		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-82809	5637.07718	-14.69	<.0001	0
TotalFlrSF	1	64.84336	2.12040	30.58	<.0001	2.58871
houseage	1	-342.00989	33.28871	-10.27	<.0001	2.33013
OverallQual	1	24881	710.80846	35.00	<.0001	2.30376
LotArea	1	0.65258	0.09206	7.09	<.0001	1.19718
BsmtFinSF1	1	23.71716	1.77310	13.38	<.0001	1.49488
TotalBath	1	2089.36112	1171.30954	1.78	0.0746	2.77974
Style1	1	11760	1937.69157	6.07	<.0001	2.15077
Style2	1	-8161.74029	2193.53736	-3.72	0.0002	2.30730
Zone1	1	2767.85231	2722.08430	1.02	0.3093	2.95192
Zone2	1	-2409.55078	3238.73184	-0.74	0.4569	3.18866

e) Backward

BACKWARD

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2928
Number of Observations with Missing Values	2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1.494423E13	1.494423E12	1169.59	<.0001
Error	2917	3.727148E12	1277733245		
Corrected Total	2927	1.867138E13			

Root MSE	35745	R-Square	0.8004
Dependent Mean	180795	Adj R-Sq	0.7997
Coeff Var	19.77119		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-82809	5637.07718	-14.69	<.0001	0
TotalFlrSF	1	64.84336	2.12040	30.58	<.0001	2.58871
houseage	1	-342.00989	33.28871	-10.27	<.0001	2.33013
OverallQual	1	24881	710.80846	35.00	<.0001	2.30376
LotArea	1	0.65258	0.09206	7.09	<.0001	1.19718
BsmtFinSF1	1	23.71716	1.77310	13.38	<.0001	1.49488
TotalBath	1	2089.36112	1171.30954	1.78	0.0746	2.77974
Style1	1	11760	1937.69157	6.07	<.0001	2.15077
Style2	1	-8161.74029	2193.53736	-3.72	0.0002	2.30730
Zone1	1	2767.85231	2722.08430	1.02	0.3093	2.95192
Zone2	1	-2409.55078	3238.73184	-0.74	0.4569	3.18866

f) Stepwise

Stepwise

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2928
Number of Observations with Missing Values	2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1.494423E13	1.494423E12	1169.59	<.0001
Error	2917	3.727148E12	1277733245		
Corrected Total	2927	1.867138E13			

Root MSE	35745	R-Square	0.8004
Dependent Mean	180795	Adj R-Sq	0.7997
Coeff Var	19.77119		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-82809	5637.07718	-14.69	<.0001	0
TotalFlrSF	1	64.84336	2.12040	30.58	<.0001	2.58871
houseage	1	-342.00989	33.28871	-10.27	<.0001	2.33013
OverallQual	1	24881	710.80846	35.00	<.0001	2.30376
LotArea	1	0.65258	0.09206	7.09	<.0001	1.19718
BsmtFinSF1	1	23.71716	1.77310	13.38	<.0001	1.49488
TotalBath	1	2089.36112	1171.30954	1.78	0.0746	2.77974
Style1	1	11760	1937.69157	6.07	<.0001	2.15077
Style2	1	-8161.74029	2193.53736	-3.72	0.0002	2.30730
Zone1	1	2767.85231	2722.08430	1.02	0.3093	2.95192
Zone2	1	-2409.55078	3238.73184	-0.74	0.4569	3.18866

Stepwise

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	TotalFlrSF	houseage	OverallQual	LotArea	BsmtFinSF1	TotalBath	Style1	Style2	Zone1	Zone2	SalePrice	_IN_	_P_	_EDF_	_MSE_	_RSQ_	_AIC_	_BIC_
1	MODEL1	PARMS	SalePrice	35745.39	-82809.16	64.8434	-342.010	24880.65	0.65258	23.7172	2089.36	11759.53	-8161.74	2767.85	-2409.55	-1	10	11	2917	1277733245.3	0.80038	61406.32	61408.40

FORWARD

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	TotalFlrSF	houseage	OverallQual	LotArea	BsmtFinSF1	TotalBath	Style1	Style2	Zone1	Zone2	SalePrice	_IN_	_P_	_EDF_	_MSE_	_RSQ_	_AIC_	_BIC_
1	MODEL1	PARMS	SalePrice	35745.39	-82809.16	64.8434	-342.010	24880.65	0.65258	23.7172	2089.36	11759.53	-8161.74	2767.85	-2409.55	-1	10	11	2917	1277733245.3	0.80038	61406.32	61408.40

BACKWARD

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	TotalFlrSF	houseage	OverallQual	LotArea	BsmtFinSF1	TotalBath	Style1	Style2	Zone1	Zone2	SalePrice	_IN_	_P_	_EDF_	_MSE_	_RSQ_	_AIC_	_BIC_
1	MODEL1	PARMS	SalePrice	35745.39	-82809.16	64.8434	-342.010	24880.65	0.65258	23.7172	2089.36	11759.53	-8161.74	2767.85	-2409.55	-1	10	11	2917	1277733245.3	0.80038	61406.32	61408.40

Conclusions: Overall the six different variable selection methods gave the same model results in terms of the summary table and p-values. P-values for the predictor variables in the selection models were all significant except for Zoning. Using MSE, AIC and BIC criteria to evaluate the selection methods, these gave the same results (see above). Thus, the Zoning variables appear to be candidates for removal from the model.

6. Tweaking variable selection model

Since zoning was not statistically significant, this was variable was deleted from the set and the model was re-fit using the stepwise selection method. The results from the analysis are shown below:

Stepwise

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2928
Number of Observations with Missing Values	2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1.493575E13	1.866968E12	1458.84	<.0001
Error	2919	3.735632E12	1279764418		
Corrected Total	2927	1.867138E13			

Root MSE	35774	R-Square	0.7999
Dependent Mean	180795	Adj R-Sq	0.7994
Coeff Var	19.78690		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-80953	5096.77853	-15.88	<.0001	0
TotalFlrSF	1	65.37476	2.11085	30.97	<.0001	2.56137
houseage	1	-364.26205	31.82041	-11.45	<.0001	2.12572
OverallQual	1	24680	707.07460	34.90	<.0001	2.27601
LotArea	1	0.69598	0.09057	7.68	<.0001	1.15702
BsmtFinSF1	1	23.69379	1.77123	13.38	<.0001	1.48936
TotalBath	1	2289.90905	1168.38528	1.96	0.0501	2.76148
Style1	1	12423	1920.38887	6.47	<.0001	2.10918
Style2	1	-8497.43480	2187.77114	-3.88	0.0001	2.29155

Stepwise

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	TotalFlrSF	houseage	OverallQual	LotArea	BsmtFinSF1	TotalBath	Style1	Style2	SalePrice	_IN_	_P_	_EDF_	_MSE_	_RSQ_	_AIC_	_BIC_
1	MODEL1	PARMS	SalePrice	35773.80	-80952.73	65.3748	-364.262	24679.60	0.69598	23.6938	2289.91	12423.44	-8497.43	-1	8	9	2919	1279764417.8	0.79993	61408.98	61411.03

Conclusion: Deleting the Zoning variable and re-fitting the model resulted in a small decrease in the adjusted R-squared for the overall regression model, 0.7994 compared 0.7997. However, the other variables selected in the model were all significant based on the p-values of their respective t-tests, including the 'HouseStyle' categorical variable.

Part C: Validation Framework

7. Create Training Set

Created a train/test split of the data for cross validation purposes in a 70%/30% ratio.

8. Model Identification by Automated Variable Selection and Predictive Accuracy

The response variable 'train_response' comprising the LogSalePrice and 70% training set, were used to execute the variable validation regimen from step 5 above. This was performed to find the 'best' models using automated variable selection using the techniques: adjusted R-Squared, AIC, Mallow's Cp, forward, backwards, and stepwise variable selection. The summary tables from each variable selection technique are reported below.

a) Stepwise

Stepwise

The REG Procedure
Model: MODEL1
Dependent Variable: train_response

Number of Observations Read	2930
Number of Observations Used	2038
Number of Observations with Missing Values	892

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	268.96826	33.62103	1312.03	<.0001
Error	2029	51.99338	0.02563		
Corrected Total	2037	320.96165			

Root MSE	0.16008	R-Square	0.8380
Dependent Mean	12.02420	Adj R-Sq	0.8374
Coeff Var	1.33130		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.77331	0.02725	395.40	<.0001	0
TotalFlrSF	1	0.00029317	0.00001127	26.01	<.0001	2.51995
houseage	1	-0.00266	0.00017047	-15.61	<.0001	2.07853
OverallQual	1	0.12360	0.00379	32.65	<.0001	2.18705
LotArea	1	0.00000386	4.665574E-7	8.27	<.0001	1.14956
BsmtFinSF1	1	0.00009656	0.00000976	9.89	<.0001	1.50745
TotalBath	1	0.03136	0.00636	4.93	<.0001	2.88007
Style1	1	0.03024	0.01043	2.90	0.0038	2.16312
Style2	1	-0.06306	0.01173	-5.37	<.0001	2.33611

b) Forward

Forward

The REG Procedure
Model: MODEL1
Dependent Variable: train_response

Number of Observations Read	2930
Number of Observations Used	2038
Number of Observations with Missing Values	892

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	268.96826	33.62103	1312.03	<.0001
Error	2029	51.99338	0.02563		
Corrected Total	2037	320.96165			

Root MSE	0.16008	R-Square	0.8380
Dependent Mean	12.02420	Adj R-Sq	0.8374
Coeff Var	1.33130		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.77331	0.02725	395.40	<.0001	0
TotalFlrSF	1	0.00029317	0.00001127	26.01	<.0001	2.51995
houseage	1	-0.00266	0.00017047	-15.61	<.0001	2.07853
OverallQual	1	0.12360	0.00379	32.65	<.0001	2.18705
LotArea	1	0.00000386	4.665574E-7	8.27	<.0001	1.14956
BsmtFinSF1	1	0.00009656	0.00000976	9.89	<.0001	1.50745
TotalBath	1	0.03136	0.00636	4.93	<.0001	2.88007
Style1	1	0.03024	0.01043	2.90	0.0038	2.16312
Style2	1	-0.06306	0.01173	-5.37	<.0001	2.33611

c) Backward

Backward

The REG Procedure
Model: MODEL1
Dependent Variable: train_response

Number of Observations Read	2930
Number of Observations Used	2038
Number of Observations with Missing Values	892

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	268.96826	33.62103	1312.03	<.0001
Error	2029	51.99338	0.02563		
Corrected Total	2037	320.96165			

Root MSE	0.16008	R-Square	0.8380
Dependent Mean	12.02420	Adj R-Sq	0.8374
Coeff Var	1.33130		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.77331	0.02725	395.40	<.0001	0
TotalFlrSF	1	0.00029317	0.00001127	26.01	<.0001	2.51995
houseage	1	-0.00266	0.00017047	-15.61	<.0001	2.07853
OverallQual	1	0.12360	0.00379	32.65	<.0001	2.18705
LotArea	1	0.00000386	4.665574E-7	8.27	<.0001	1.14956
BsmtFinSF1	1	0.00009656	0.00000976	9.89	<.0001	1.50745
TotalBath	1	0.03136	0.00636	4.93	<.0001	2.88007
Style1	1	0.03024	0.01043	2.90	0.0038	2.16312
Style2	1	-0.06306	0.01173	-5.37	<.0001	2.33611

d) Adjusted R-squared

Adjusted R-squared

The REG Procedure
Model: MODEL1
Dependent Variable: train_response

Number of Observations Read	2930
Number of Observations Used	2038
Number of Observations with Missing Values	892

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	268.96826	33.62103	1312.03	<.0001
Error	2029	51.99338	0.02563		
Corrected Total	2037	320.96165			

Root MSE	0.16008	R-Square	0.8380
Dependent Mean	12.02420	Adj R-Sq	0.8374
Coeff Var	1.33130		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.77331	0.02725	395.40	<.0001	0
TotalFlrSF	1	0.00029317	0.00001127	26.01	<.0001	2.51995
houseage	1	-0.00266	0.00017047	-15.61	<.0001	2.07853
OverallQual	1	0.12360	0.00379	32.65	<.0001	2.18705
LotArea	1	0.00000386	4.665574E-7	8.27	<.0001	1.14956
BsmtFinSF1	1	0.00009656	0.00000976	9.89	<.0001	1.50745
TotalBath	1	0.03136	0.00636	4.93	<.0001	2.88007
Style1	1	0.03024	0.01043	2.90	0.0038	2.16312
Style2	1	-0.06306	0.01173	-5.37	<.0001	2.33611

e) R-squared

R-squared

The REG Procedure
Model: MODEL1
Dependent Variable: train_response

Number of Observations Read	2930
Number of Observations Used	2038
Number of Observations with Missing Values	892

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	268.96826	33.62103	1312.03	<.0001
Error	2029	51.99338	0.02563		
Corrected Total	2037	320.96165			

Root MSE	0.16008	R-Square	0.8380
Dependent Mean	12.02420	Adj R-Sq	0.8374
Coeff Var	1.33130		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.77331	0.02725	395.40	<.0001	0
TotalFlrSF	1	0.00029317	0.00001127	26.01	<.0001	2.51995
houseage	1	-0.00266	0.00017047	-15.61	<.0001	2.07853
OverallQual	1	0.12360	0.00379	32.65	<.0001	2.18705
LotArea	1	0.00000386	4.665574E-7	8.27	<.0001	1.14956
BsmtFinSF1	1	0.00009656	0.00000976	9.89	<.0001	1.50745
TotalBath	1	0.03136	0.00636	4.93	<.0001	2.88007
Style1	1	0.03024	0.01043	2.90	0.0038	2.16312
Style2	1	-0.06306	0.01173	-5.37	<.0001	2.33611

f) Mallow's Cp

Mallows CP

The REG Procedure
Model: MODEL1
Dependent Variable: train_response

Number of Observations Read	2930
Number of Observations Used	2038
Number of Observations with Missing Values	892

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	268.96826	33.62103	1312.03	<.0001
Error	2029	51.99338	0.02563		
Corrected Total	2037	320.96165			

Root MSE	0.16008	R-Square	0.8380
Dependent Mean	12.02420	Adj R-Sq	0.8374
Coeff Var	1.33130		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.77331	0.02725	395.40	<.0001	0
TotalFlrSF	1	0.00029317	0.00001127	26.01	<.0001	2.51995
houseage	1	-0.00266	0.00017047	-15.61	<.0001	2.07853
OverallQual	1	0.12360	0.00379	32.65	<.0001	2.18705
LotArea	1	0.00000386	4.665574E-7	8.27	<.0001	1.14956
BsmtFinSF1	1	0.00009656	0.00000976	9.89	<.0001	1.50745
TotalBath	1	0.03136	0.00636	4.93	<.0001	2.88007
Style1	1	0.03024	0.01043	2.90	0.0038	2.16312
Style2	1	-0.06306	0.01173	-5.37	<.0001	2.33611

Conclusions: The results from this step *versus* step 5, indicate the models are very similar to each other, in part because we have removed a non-significant variable – Zoning, in the starting model. From the t-tables we can see from the p-values that they are all significant in the model as the p-values are small.

9.

Model_S

Obs	train	_TYPE_	_FREQ_	MSE_1	MAE_1
1	0	1	891	0.036192	0.11646
2	1	1	2039	0.025512	0.10933

Model_F

Obs	train	_TYPE_	_FREQ_	MSE_1	MAE_1
1	0	1	891	0.036192	0.11646
2	1	1	2039	0.025512	0.10933

Model_B

Obs	train	_TYPE_	_FREQ_	MSE_1	MAE_1
1	0	1	891	0.036192	0.11646
2	1	1	2039	0.025512	0.10933

Model_AdjR2

Obs	train	_TYPE_	_FREQ_	MSE_1	MAE_1
1	0	1	891	0.036192	0.11646
2	1	1	2039	0.025512	0.10933

Model_Rsquared

Obs	train	_TYPE_	_FREQ_	MSE_1	MAE_1
1	0	1	891	0.036192	0.11646
2	1	1	2039	0.025512	0.10933

Model_Mcp

Obs	train	_TYPE_	_FREQ_	MSE_1	MAE_1
1	0	1	891	0.036192	0.11646
2	1	1	2039	0.025512	0.10933

Conclusion: All the models seem to fit equally well according to the output shown above. To evaluate the best model we need to find the one with for example, small values of MSE. The in-sample set (train=1) seemed to display the smallest value (MSE=0.025512 in-sample *versus* MSE=0.03612 for out-of-sample).

10. Operational Validation:

The criteria for evaluating these models, specifically MSE and MAE, do not translate easily to the development of business policy. A set of prediction grades was established: Grade 1 (best, within 10% of actual value), Grade 2 (medium, within 15% of actual value), and Grade 3 (worst), to provide a means of assessing predictive accuracy. Comparing the training (train=1) and test (train=0) sets in the output tables below, shows that there is a reasonable match between training and test sets. For example, for training set Grade 1 the percent frequency is 59.74% compared to 59.37% in the Grade 1 test set, which is close though slightly higher. There are comparable alignments between Grades 2 and 3 for the training and test sets.

Stepwise model assessment- part 10

The FREQ Procedure

train=0

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	529	59.37	529	59.37
Grade 2	155	17.40	684	76.77
Grade 3	207	23.23	891	100.00

Stepwise model assessment- part 10

The FREQ Procedure

train=1

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	1218	59.74	1218	59.74
Grade 2	341	16.72	1559	76.46
Grade 3	480	23.54	2039	100.00

Conclusion: Using this approach to assess predictive accuracy, we find that the percent of observations in each quality grade of the prediction is similar between training and test sets. This suggests a reasonable regression model has been created for the purposes of making reasonable business decisions.

11. Reporting final model

Task 11: Reporting Final Model

The REG Procedure
Model: MODEL1
Dependent Variable: LogSalePrice

Number of Observations Read	2930
Number of Observations Used	2928
Number of Observations with Missing Values	2

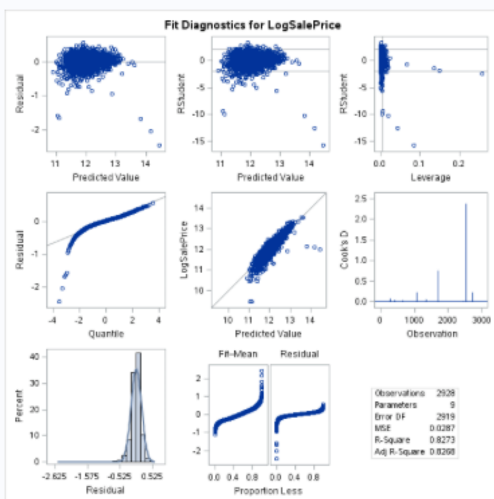
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	401.83422	50.22928	1747.36	<.0001
Error	2919	83.90915	0.02875		
Corrected Total	2927	485.74337			

Root MSE	0.16955	R-Square	0.8273
Dependent Mean	12.02104	Adj R-Sq	0.8268
Coeff Var	1.41041		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.74575	0.02416	444.85	<.0001	0
TotalFlrSF	1	0.00028375	0.00001000	28.36	<.0001	2.56137
houseage	1	-0.00264	0.00015081	-17.53	<.0001	2.12572
OverallQual	1	0.13051	0.00335	38.94	<.0001	2.27601
LotArea	1	0.00000376	4.292601E-7	8.75	<.0001	1.15702
BsmtFinSF1	1	0.00007902	0.00000839	9.41	<.0001	1.48936
TotalBath	1	0.03356	0.00554	6.06	<.0001	2.76148
Style1	1	0.03156	0.00910	3.47	0.0005	2.10918
Style2	1	-0.06335	0.01037	-6.11	<.0001	2.29155

Task 11: Reporting Final Model

The REG Procedure
Model: MODEL1
Dependent Variable: LogSalePrice



Conclusion: All the continuous variables picked out by the automated variable selection process were included in the final model and were statistically significant. While one of the categorical variables, Zoning, was dropped earlier in the process (step 6) since it was not statistically significant, another (*e.g.* 'HouseStyle') remained in the final model. Although the final model was derived from various variable selection procedures, and produced a reasonable correlation between training and test sets, there still remains some goodness-of-fit issues with the regression model.

CONCLUSIONS

After working on this problem and this data for several weeks, what are the challenges presented by the data? What are your recommendations for improving predictive accuracy?

Although the statistical significance of the final regression model was good (overall F-statistic), and the inclusion of the chosen regressor variables in the model was sound (based on the individual t-tests), there remain some concerns. First, from the goodness-of-fit metrics shown in the final model shown above, it is apparent that the randomness of the residuals in the plot above (top left) is in some question – some reverse funnel shape observed. Second, this final model from the assignment does not have SalePrice outliers removed from the earlier assignments. Thus, with the removal of further outliers it is likely that further improvement in the disposition of the residuals, GOF and adjusted-R-squared will be observed. Finally, pruning of the some of the outliers for each of the regressor variables chosen in the final model to will lead to further improvements in the final model, in combination with a transformation of the SalePrice response variable. These comments suggest that the model can be improved further. That said the 'best' regression model shown above has an adjusted-R-squared value >82% indicating that more than 82% of the variance in the SalePrice is explained by the regressor variables in the model.