Assignment 5: Variable Selection and Validation

Andrew G. Dunn[1]

[1]andrew.g.dunn@u.northwestern.edu

Andrew G. Dunn, Northwestern University Predictive Analytics Program

Prepared for PREDICT-410: Regression & Multivariate Analysis.

Formatted using markdown, pandoc, and LaTeX. References managed using Bibtex, and pandoc-citeproc.

Goals & Data Examination

The goal of this report is to design a model which has some predictive explanatory value for the sale price of a home given some contextual information about many observed sales. In examining the data dictionary provided with this data set, we noticed that there were some values that could be used to scope down the data set. Most notably Sale Condition, which can have one of six values:

```
Normal   Normal Sale
Abnorml  Abnormal Sale -  trade, foreclosure, short sale
AdjLand  Adjoining Land Purchase
Alloca   Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family   Sale between family members
Partial  Home was not completed when last assessed (associated with New Homes)
```

Functional, which can have one of eight values:

```
Typ    Typical Functionality
Min1   Minor Deductions 1
Min2   Minor Deductions 2
Mod    Moderate Deductions
Maj1   Major Deductions 1
Maj2   Major Deductions 2
Sev    Severely Damaged
Sal    Salvage only
```

In a formal analysis, we would need to approach the business owner to discuss how these two parameters can be used to scope down the overall data set. We'll take some creative liberty now that we've noticed these two parameters and initially scope down the data set based on the following criteria: Keep observations that are 'Normal' Sale Condition and 'Typ' Functionality. We start with 2930 observations, after our data processing procedure we're down to 2240.

Indicator Variables

Normally we would initially look at categorical variables and only consider the variables that have high independent correlation to the phenomena were interested in modeling (SalePrice). Instead, we will choose categorical variables that are based on observable house features. Many categorical variables within this data set are indications of a level of quality. This quality indication is likely subjective in nature. We recognize that the model we're building is meant to be used for predictive purposes, so we choose to build the model from observable features rather than potentially subjective features.

As such, we've chosen to model HouseStyle and GarageType. HouseStyle can be one of eight different categories:

```
1Story   One story
1.5Fin   One and one-half story: 2nd level finished
1.5Unf   One and one-half story: 2nd level unfinished
2Story   Two story
2.5Fin   Two and one-half story: 2nd level finished
2.5Unf   Two and one-half story: 2nd level unfinished
SFoyer   Split Foyer
SLvl     Split Level
```

Where GarageType can be one of seven different categories:

```
2Types   More than one type of garage
Attchd   Attached to home
Basment  Basement Garage
BuiltIn  Built-In (Garage part of house - typically has room above garage)
CarPort  Car Port
Detchd   Detached from home
NA       No Garage
```

We will code both as indicator variables, and examine one as an independent model in the following section.

Examine Categorical Variable HouseStyle

As stated before, HouseStyle can have one of eight values. Here are the means from our data set:

N	HouseStyle	SalePrice Mean
1108	1Story	174392.88
231	1.5Fin	140087.45
19	1.5Unf	109663.16
690	2Story	201178.07
5	2.5Fin	253000.00
21	2.5Unf	181900.00
63	SFoyer	142558.10
103	SLvl	167385.78

Table 1: Mean SalePrice given HouseStyle

We run a simple linear regression model:

$$\text{SalePrice} = \beta_0 + \beta_1 \text{HouseStyle} + \epsilon$$

Was unable to re-code this variable to perform the regression, for the procedural analysis and interpretation, please see the end of the report where we examined a regression of SalePrice given OverallQual.

Indicator Variables, (Dummy Coding) a Categorical Variable

We will discuss in more detail the dummy coding of HouseStyle. Although HouseStyle is an eight way variable, we can model it with seven dummy coded variables by holding over one variable as the basis for interpretation. This is simpler to consider by examining the table below:

HouseStyle	hs_1	hs_2	hs_3	hs_4	hs_5	hs_6	hs_7
1Story	1	0	0	0	0	0	0
1.5Fin	0	1	0	0	0	0	0
1.5Unf	0	0	1	0	0	0	0
2Story	0	0	0	1	0	0	0
2.5Fin	0	0	0	0	1	0	0
2.5Unf	0	0	0	0	0	1	0
SFoyer	0	0	0	0	0	0	1
SLvl	0	0	0	0	0	0	0

Table 2: Modeling HouseStyle as an indicator Variable

We will use the data procedure to dummy code HouseStyle as an indicator variable. To examine our progress we evaluate a proc freq of HouseStyle:

HouseStyle	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.5Fin	231	10.31	231	10.31
1.5Unf	19	0.85	250	11.16
1Story	1108	49.46	1358	60.63
2.5Fin	5	0.22	1363	60.85
2.5Unf	21	0.94	1384	61.79
2Story	690	30.80	2074	92.59
SFoyer	63	2.81	2137	95.40
SLvl	103	4.60	2240	100.00

Table 3: Frequency HouseStyle

For brevity we examine the hs_1 and hs_2 variables:

hs_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1132	50.54	1132	50.54
1	1108	49.46	2240	100.00

Table 4: Frequency hs_1, Indicator Variable for HouseStyle '1Story'

hs_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2009	89.69	2009	89.69
1	231	10.31	2240	100.00

Table 5: Frequency hs_2, Indicator Variable for HouseStyle '1.5Fin'

We notice that the hs_1 ahd hs_2 variables are properly coded to match up with the corresponding HouseStyle categories in the frequency table, with hs_1 having 1 coded 1108 times and hs_2 have 1 coded 231 times.

We now build a model, but we hold hs_8 to be the basis of interpretation:

$$\text{SalePrice} = \beta_0 + \beta_1 \text{hs\_1} + \beta_2 \text{hs\_2} + \beta_3 \text{hs\_3} + \beta_4 \text{hs\_4} + \beta_5 \text{hs\_5} + \beta_6 \text{hs\_6} + \beta_7 \text{hs\_7} + \epsilon$$

Resulting in the parameter estimations and model diagnostics:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	167386	6752.11484	24.79	<.0001
hs_1	1	7007.10056	7058.98098	0.99	0.3210
hs_2	1	-27298	8119.08454	-3.36	0.0008
hs_3	1	-57723	17110	-3.37	0.0008
hs_4	1	33792	7238.55483	4.67	<.0001
hs_5	1	85614	31381	2.73	0.0064
hs_6	1	14514	16407	0.88	0.3765
hs_7	1	-24828	10960	-2.27	0.0236

Table 6: Model Parameter Estimates for SalePrice = HouseStyle Indicator Variables

Source	
Root MSE	68526
R-Square	0.0811
Adj R-Square	0.0782
F Value	28.13

Table 7: Model Estimator Performance for SalePrice = HouseStyle Indicator Variables

The resulting Model is:

$$\begin{aligned}\text{SalePrice} = 167386 + 7007.10056 \times \text{hs\_1} - 27298 \times \text{hs\_2} - 57723 \times \text{hs\_3} \\ +33792 \times \text{hs\_4} + 85614 \times \text{hs\_5} + 14514 \times \text{hs\_6} - 24828 \times \text{hs\_7}\end{aligned}$$

We now interpret the model. If the HouseStyle is 1, or '1Story' then the model becomes:

$$\text{SalePrice} = 167386 + 7007.10056$$

That is to say, if the HouseStyle is '1Story', then the SalePrice in this model is $174,387.10, Looking back at our mean table we see that '1Story' had a mean of $174,392.88,

If the HouseStyle is 2, or '1.5Fin' then the model becomes:

$$\text{SalePrice} = 167386 - 27298$$

That is to say, if the HouseStyle is '1.5Fin', then the SalePrice in this model is $140,088.00, Looking back at our mean table we see that '1.5Fin' had a mean of $140,087.45,

If the HouseStyle is 3, or '1.5Unf' then the model becomes:

$$\text{SalePrice} = 167386 - 57723$$

That is to say, if the HouseStyle is '1.5Unf', then the SalePrice in this model is $109,663.00, Looking back at our mean table we see that '1.5Unf' had a mean of $109,663.16,

If the HouseStyle is 4, or '2Story' then the model becomes:

$$\text{SalePrice} = 167386 + 33792$$

That is to say, if the HouseStyle is '2Story', then the SalePrice in this model is $201,178.00, Looking back at our mean table we see that '2Story' had a mean of $201,178.07,

If the HouseStyle is 5, or '2.5Fin' then the model becomes:

$$\text{SalePrice} = 167386 + 85614$$

That is to say, if the HouseStyle is '2.5Fin', then the SalePrice in this model is $253,000.00, Looking back at our mean table we see that '2.5Fin' had a mean of $253,000.00,

If the HouseStyle is 6, or '2.5Unf' then the model becomes:

$$\text{SalePrice} = 167386 + 14514$$

That is to say, if the HouseStyle is '2.5Unf', then the SalePrice in this model is $181,900.00, Looking back at our mean table we see that '2.5Unf' had a mean of $181,900.00,

If the HouseStyle is 7, or 'SFoyer' then the model becomes:

$$\text{SalePrice} = 167386 - 24828$$

That is to say, if the HouseStyle is 'SFoyer', then the SalePrice in this model is $142,558.00, Looking back at our mean table we see that 'SFoyer' had a mean of $142,558.10,

If the HouseStyle is 8 or 'SLvl' then the model becomes:

$$\text{SalePrice} = 167386$$

That is to say, if the HouseStyle is 'SLvl', then the SalePrice in this model is $167,386, Looking back at our mean table we see that 'SLvl' had a mean of $167,385.78,

Dummy Code Hypothesis Testing

$$H_0 : \beta_{1..7} = 0 \text{ versus } H_1 : \beta_{1..7} \neq 0$$

We notice that hs_1, hs_5 and hs_6 are do not yield statistically significant results. Where hs_2, hs_3, hs_4, hs_5, and hs_7 all yield statistically significant results. We believe there is possibly a way to write out the hypothesis testing in a more separate-but-joint fashion, but cannot find a reference to explain. This model, with just indicator variables, is highly uncomfortable to work with and interpret.

Indicator Variables, (Dummy Coding) another Categorical Variable

We will also dummy code the GarageType categorical variable. Here is the respective means and frequency tables:

N	GarageType	SalePrice Mean
13	2Types	169446.15
1342	Attchd	198923.58
21	Basment	154383.33
135	BuiltIn	227725.25
8	CarPort	103943.75
617	Detchd	134038.98
104	NA	106865.14

Table 8: Mean SalePrice given GarageType

GarageType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2Types	13	0.58	13	0.58
Attchd	1342	59.91	1355	60.49
Basment	21	0.94	1376	61.43
BuiltIn	135	6.03	1511	67.46
CarPort	8	0.36	1519	67.81
Detchd	617	27.54	2136	95.36
NA	83	4.64	2240	100.00

Table 9: Frequency GarageType

For brevity we examine the gt_1 and g1_2 variables, which correspond to '2Types' and 'Attchd' respectively:

gt_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2227	99.42	2227	99.42
1	13	0.58	2240	100.00

Table 10: Frequency gt_1, Indicator Variable for GarageType '2Types'

gt_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	898	40.09	898	40.09
1	1342	59.91	2240	100.00

Table 11: Frequency gt_2, Indicator Variable for GarageType 'Attchd'

We notice that the gt_1 and gt_2 variables are properly coded to match up with the GarageType frequency table, with gt_1 have 1 coded 13 times and gt_2 having 1 coded 1342 times respectively.

Automated Variable Selection

From assignment 2 we will consult the results of correlating the continuous variables to SalePrice. We will take the top 5 to incorporate into our automated variable selection strategy. We choose to include MasVnrArea with the knowledge that it will be zero for many of the observations, based on our EDA performed in assignment 2.

Continuous Variable	Correlation to SalePrice	Prob $>\|r\|$ under H_0: ρ=0	Number of Observations
GrLivArea	0.70678	<0.0001	2930
GarageArea	0.64040	<0.0001	2929
TotalBsmtSF	0.63228	<0.0001	2929
FirstFlrSF	0.62168	<0.0001	2930
MasVnrArea	0.50828	<0.0001	2907
BsmtFinSF1	0.43291	<0.0001	2929
BsmtUnfSF	0.18286	<0.0001	2929

Table 12: Continuous variable correlation to SalePrice, top seven.

We run the reg procedure using the selection method; adjrsq, cp, forward, backward, and stepwise. The follow models are found to to be the ranked best by these methods.

Adjusted R-Square Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = &-15885.73 + 70.1238 \times \text{GrLivArea} + 79.4680 \times \text{GarageArea} \\ &+52.6483 \times \text{TotalBsmtSF} + 43.6788 \times \text{MasVnrArea} \\ &-61717.91 \times \text{gt\_1} + 210089.09 \times \text{gt\_4} - 14741.64 \times \text{gt\_5} - 14465.70 \times \text{gt\_6} \\ &-12787.57 \times \text{hs\_2} - 27696.38 \times \text{hs\_5} - 18947.42 \times \text{hs\_6} + 10917.80 \times \text{hs\_7}\end{aligned}$$

Source	
Root MSE	32905.57
C_p	8.9007
R-Square	0.7892
Adj. R-Square	0.7881
AIC	46382.3907
BIC	46384.5916

Table 13: Model Performance

We generally expect that the selection method will result in selection of many dependent variables. We hope moving forward with the other selection methods that they are not as egregious with their incorporation of variables. We notice that of the listed models, the C_p for this model, with this method, was the lowest of the top 5.

Mallow's C_p Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = &-16287.58 + 70.2305 \times \text{GrLivArea} + 79.1755 \times \text{GarageArea} \\ &+52.9160 \times \text{TotalBsmtSF} + 43.6495 \times \text{MasVnrArea} \\ &-61577.31 \times \text{gt\_1} + 10202.10 \times \text{gt\_4} - 14284.93 \times \text{gt\_6} \\ &-12847.76 \times \text{hs\_2} - 27798.90 \times \text{hs\_5} - 19060.10 \times \text{hs\_6} + 11055.61 \times \text{hs\_7}\end{aligned}$$

Source	
Root MSE	32909.82
C_p	8.4711
R-Square	0.7891
Adj. R-Square	0.7880
AIC	46381.9726
BIC	46384.1409

Table 14: Model Performance

We notice that there is a draw by one parameter in this model, and accordingly the C_p criterion is calculated to be lower.

AIC Selection (Analyst Examination of Mallow's C_p results)

We realize that the regression procedure within SAS does not allow for selection by AIC criterion. We therefor examine the output of the C_p selection and choose the model with the lowest value of AIC.

Model Selected:

$$\begin{aligned}\text{SalePrice} = -16287.58 + 70.2305 \times \text{GrLivArea} + 79.1755 \times \text{GarageArea} \\ +52.9160 \times \text{TotalBsmtSF} + 43.6495 \times \text{MasVnrArea} \\ -61577.31 \times \text{gt\_1} + 10202.10 \times \text{gt\_4} - 14284.93 \times \text{gt\_6} \\ -12847.76 \times \text{hs\_2} - 27798.90 \times \text{hs\_5} - 19060.10 \times \text{hs\_6} + 11055.61 \times \text{hs\_7}\end{aligned}$$

Source	
Root MSE	32909.82
C_p	8.4711
R-Square	0.7891
Adj. R-Square	0.7880
AIC	46381.9726
BIC	46384.1409

Table 15: Model Performance

We had initially expected to see the AIC selection criterion result in a model with few parameters due to AIC formulation having a built in penalty as an increasing function of the number of estimated parameters. We are sadly disappointed an have received yet another large model.

Forward Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = -14921 + 67.24173 \times \text{GrLivArea} + 79.24595 \times \text{GarageArea} \\ +51.31153 \times \text{TotalBsmtSF} + 5.47640 \times \text{FirstFlrSF} + 43.47536 \times \text{MasVnrArea} \\ -62045 \times \text{gt\_1} + 10599 \times \text{gt\_4} - 14982 \times \text{gt\_5} - 14275 \times \text{gt\_6} \\ -2680 \times \text{hs\_1} - 13595 \times \text{hs\_2} - 26449 \times \text{hs\_5} - 18783 \times \text{hs\_6} + 8752 \times \text{hs\_8}\end{aligned}$$

Source	
Root MSE	32910.48
C_p	11.5634
R-Square	0.78935
Adj. R-Square	0.78802
F Value	592.60
AIC	46385.04
BIC	46387.29

Table 16: Model Performance

Backward Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = &-16288 + 70.23050 \times \text{GrLivArea} + 79.17552 \times \text{GarageArea} \\ &+52.91604 \times \text{TotalBsmtSF} + 43.64946 \times \text{MasVnrArea} \\ &-61577 \times \text{gt\_1} + 10202 \times \text{gt\_4} - 14285 \times \text{gt\_6} \\ &-12848 \times \text{hs\_2} - 27799 \times \text{hs\_5} - 19060 \times \text{hs\_6} + 11056 \times \text{hs\_7}\end{aligned}$$

Source	
Root MSE	32909.82
C_p	8.47108
R-Square	0.78907
Adj. R-Square	0.78803
F Value	753.98
AIC	46381.97
BIC	46384.14

Table 17: Model Performance

Stepwise Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = -16288 + 70.23050 \times \text{GrLivArea} + 79.17552 \times \text{GarageArea} \\ +52.91604 \times \text{TotalBsmtSF} + 43.64946 \times \text{MasVnrArea} \\ -61577 \times \text{gt\_1} + 10202 \times \text{gt\_4} - 14285 \times \text{gt\_6} \\ -12848 \times \text{hs\_2} - 27799 \times \text{hs\_5} - 19060 \times \text{hs\_6} + 11056 \times \text{hs\_7}\end{aligned}$$

Source	
Root MSE	32909.82
C_p	8.47108
R-Square	0.78907
Adj. R-Square	0.78803
F Value	753.98
AIC	46381.97
BIC	46384.14

Table 18: Model Performance

Comparing Model Performance

We'll make a table to compare the model performance information

Model	Cont.	Ind.	Root MSE	C_p	R-Square	Adj. R-Square	F Value	AIC	BIC
Adj. R-Square	4	8	32905.57	8.9007	0.7892	0.7881	-	46382.3907	46384.5916
Mallow's C_p	4	7	32909.82	8.4711	0.7891	0.7880	-	46381.9726	46384.1409
AIC	4	7	32909.82	8.4711	0.7891	0.7880	-	46381.9726	46384.1409
Forward	5	7	32910.48	11.5634	0.78935	0.78802	592.60	46385.04	46387.29
Backward	4	7	32909.82	8.47108	0.78907	0.78803	753.98	46381.97	46384.14
Stepwise	4	7	32909.82	8.47108	0.78907	0.78803	753.98	46381.97	46384.14

Table 19: Automatic Variable Selection Model Comparison

It seems relevant to mention that models which incorporate more parameters become more complex for interpretation. Going into the variable selection, we had anticipated that Mallow's C_p and AIC would result in models of greatly reduced complexity (parameters), however the results show that these models all performed well by incorporating almost all of the continuous variables in them.

For Mallow's C_p and AIC, we received the same results because we were looking to minimize AIC. The Mallow's C_p method found the models with the lowest AIC, so we naturally used that same model for the AIC selection criteria.

In an appendix to this assignment we included our automatic variable selection using the OverallQual variable. We are overall happier with the two variables that were used above, even if some criterion came out to be worse. We feel one of the most important aspects of modeling is that the model needs to be interpretable. In some fields this is not as necessary, as models become exceedingly complex, in this situation however we feel the needs to reduce model complexity in favor of interpret-ability.

Aside from model complexity there was some concern that OverallQual is a subjective categorical measurement, as opposed to HouseStyle which is an observable categorical measurement. This likely meant that we were vectoring towards building a model that will be more useful for inference than prediction. We say this because in sample we have observations of OverallQual, but out-of- sample there is no systematic way (that we're aware of) of observing and characterizing Overallqual. It is not as obviously measurable as an explicit feature of the premises such as HouseStyle and GarageType.

Overall we see that all models, aside from the forward selection, excluded the FirstFlrSF variable. We notice that the Forward selection logically had the highest C_p value, likely due in part to the inclusion of this model. We're happy to exclude the FirstFlrSF variable as we notice that the overall performance difference when considering the Adj. R-Square, and C_p is no different than Backward and Stepwise methods. We see that Backward and Stepwise have resulted in models that are equivalent.

We'll make the choice to take the model designed by Backward and Stepwise selection.

Indicator Variable Inclusion

We will refit the model incorporating all of the identification variables.

$$\begin{aligned}\text{SalePrice} = -14106 + 70.86187 \times \text{GrLivArea} + 79.56383 \times \text{GarageArea} \\ +52.14204 \times \text{TotalBsmtSF} + 43.57188 \times \text{MasVnrArea} \\ -62056 \times \text{gt\_1} - 151 \times \text{gt\_2} - 4300 \times \text{gt\_3} + 9457 \times \text{gt\_4} \\ -15012 \times \text{gt\_5} - 14633 \times \text{gt\_6} + 0 \times \text{gt\_7} \\ -2063 \times \text{hs\_1} - 15081 \times \text{hs\_2} + 3112 \times \text{hs\_3} - 2752 \times \text{hs\_4} \\ -30202 \times \text{hs\_5} - 21620 \times \text{hs\_6} + 9301 \times \text{hs\_7} + 0 \times \text{hs\_8}\end{aligned}$$

Source	
Root MSE	32934
R-Square	0.7893
Adj. R-Square	0.7877
F Value	487.32

Table 20: Model Performance

It should be noted that both gt_4 and hs_8 were set to zero due to the fact that they make the overall use of each set of indicator variables a linear combination. SAS automatically marks them as zero. Secondarily, every single indicator variable is found to be bias by SAS And only two are considered to be statistically significant (gt_1 and hs_2). All continuous variables included in this model are found to be statistically significant.

Interpretation of this model is difficult, we have to consider what a single unit increase means for each continuous variable, as well as the combinations for each indicator variable. The model is less interpretable for our business owners, and we worry that for the complexity increase we've not gained justifiable explanatory performance.

The model complexity of interpretation, as well as the relative performance (based on the Adj. R-Square and F Value criteria) are considered, in light of our other forays into this data set, to be poor. Now that its sufficiently bothering us, we decide to run a model to exclude the indicator parameters. We get the following model:

$$\text{SalePrice} = -27086 + 70.65635 \times \text{GrLivArea} + 79.08423 \times \text{GarageArea} + 57.38357 \times \text{TotalBsmtSF} + 49.92249 \times \text{MasVnrArea}$$

Source	
Root MSE	34449
R-Square	0.7682
Adj. R-Square	0.7677
F Value	1842.16

Table 21: Model Performance

We notice that dependent variables are all statistically significant and that the F Value increases significantly compared to the last model.

We also have some discomfort knowing the context of the MasVnrArea, a variable that is only partially applicable to the observations within the data set. If we wanted to, we could remove more observations from the data set, but instead for brevity we'll just run another model where we exclude the MasVnrArea variable:

$$\text{SalePrice} = -35517 + 75.23233 \times \text{GrLivArea} + 84.90909 \times \text{GarageArea} + 61.35713 \times \text{TotalBsmtSF}$$

Source	
Root MSE	35283
R-Square	0.7559
Adj. R-Square	0.7556
F Value	2308.63

Table 22: Model Performance

We see that each of the continuous coefficients are statistically significant and that the F Value has increased significantly for this model.

All in all, its disheartening to see a perceived performance increase when we get rid of the hard-to-wield indicator variables.

Validation Framework

We will use a univariate distribution and sample to select 70% of the observations from our cleansed data set into a training group. For brevity we're not going to report the model parameters. We know this is less than ideal in the sense of the report, however for it takes a great deal of time to export and format.

Obtaining the "Best" Model

We will re-run the models from above, but this time with the training data set.

Adjusted R-Square Selection

Model selected incorporated the following variables:

```
GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_4 gt_7 hs_1 hs_2 hs_5 hs_6 hs_7
```

Source	
Root MSE	33521.67
C_p	11.5843
R-Square	0.77270
Adj. R-Square	0.7706
AIC	31795.7656
BIC	31798.1320

Table 23: Model Performance

Mallow's C_p Selection

Model selected incorporated the following variables:

```
GrLivArea GarageArea TotalBsmtSF MasVnrArea gt_1 gt_4 gt_5 gt_6 hs_2 hs_5 hs_6 hs_7
```

Source	
Root MSE	32909.82
C_p	9.6048
R-Square	0.7724
Adj. R-Square	0.7706
AIC	31793.8094
BIC	31796.0918

Table 24: Model Performance

AIC Selection (Analyst Examination of Mallow's C_p results)

We realize that the regression procedure within SAS does not allow for selection by AIC criterion. We therefor examine the output of the C_p selection and choose the model with the lowest value of AIC.

Model selected incorporated the following variables:

```
GrLivArea GarageArea TotalBsmtSF MasVnrArea gt_1 gt_4 gt_5 gt_6 hs_2 hs_5 hs_6 hs_7
```

Source	
Root MSE	32909.82
C_p	9.6048
R-Square	0.7724
Adj. R-Square	0.7706
AIC	31793.8094
BIC	31796.0918

Table 25: Model Performance

Forward Selection

Model selected incorporated the following variables:

```
GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_3 gt_4 gt_5 gt_6 hs_1 hs_2 hs_5 hs_6 hs_
```

Source	
Root MSE	32910.48
C_p	11.5634
R-Square	0.78935
Adj. R-Square	0.78802
F Value	592.60
AIC	46385.04
BIC	46387.29

Table 26: Model Performance

This method once again selected to use he FirstFlrSF variable.

Backward Selection

Model selected incorporated the following variables:

```
GrLivArea GarageArea TotalBsmtSF MasVnrArea gt_1 gt_4 gt_5 gt_6 hs_2 hs_5 hs_6 hs_7
```

Source	
Root MSE	33533.15
C_p	9.61115
R-Square	0.77209
Adj. R-Square	0.77044
F Value	465.97
AIC	31793.84
BIC	31796.06

Table 27: Model Performance

Stepwise Selection

Source	
Root MSE	33533.15
C_p	9.61115
R-Square	0.77209
Adj. R-Square	0.77044
F Value	465.97
AIC	31793.84
BIC	31796.06

Table 28: Model Performance

Model Comparison

Model	Cont.	Ind.	Root MSE	C_p	R-Square	Adj. R-Square	F Value	AIC	BIC
Adj. R-Square	4	8	32905.57	8.9007	0.7892	0.7881	-	46382.3907	46384.5916
Mallow's C_p	4	7	32909.82	8.4711	0.7891	0.7880	-	46381.9726	46384.1409
AIC	4	7	32909.82	8.4711	0.7891	0.7880	-	46381.9726	46384.1409
Forward	5	7	32910.48	11.5634	0.78935	0.78802	592.60	46385.04	46387.29
Backward	4	7	32909.82	8.47108	0.78907	0.78803	753.98	46381.97	46384.14
Stepwise	4	7	32909.82	8.47108	0.78907	0.78803	753.98	46381.97	46384.14

Table 29: Automatic Variable Selection Model Comparison

Model	Cont.	Ind.	Root MSE	C_p	R-Square	Adj. R-Square	F Value	AIC	BIC
Adj. R-Square	4	8	33521.67	11.5843	0.77270	0.7706	-	31795.7656	31798.1320
Mallow's C_p	4	7	32909.82	9.6048	0.7724	0.7706	-	31793.8094	31796.0918
AIC	4	7	32909.82	9.6048	0.7724	0.7706	-	31793.8094	31796.0918
Forward	5	7	32910.48	11.5634	0.78935	0.78802	592.60	46385.04	46387.29
Backward	4	7	33533.15	9.61115	0.77209	0.77044	465.97	31793.84	31796.06
Stepwise	4	7	33533.15	9.61115	0.77209	0.77044	465.97	31796.06	31793.84

Table 30: Automatic Variable Selection Model Comparison (Training Data)

We notice that forward selection on the full set and training set has the same calculated criterion. Backward and Stepwise produce the same calculated criterion, but the results are different on the training data as compared the overall set.

The techniques seem to have generally come up with the same model given the training data set.

Comparing Models with Training and Test Data

In software, not interpreted in report

Operational Validation

In software, not interpreted in report

Best Model, Revisited with all Dummy coded Variables

Would love to happily report a model.

Conclusion / Reflection

There were significant challenges with this data set when using categorical variables. We don't suspect these challenges to be unique to this data set, it seems that categorical variables are quite unwieldy when it comes to model complexity and interpretation.

Aside from the relatively conclusions scattered within the text, we have some general remarks from the exercise. Stubborness on the side of the Analyst is not a good thing if there is a deadline, if a model isn't sitting well then trying to make it work through attrition isn't likely the best approach. With software being as expressive and powerful as it is now, there is no need to get hung up on a specific variable. Flexibility in modeling is key, a lot if idiosyncrasies in a data set can be teased out through re-examination and implementation of different models. Using a categorical variable that is a Lickert scale likely means you're relying on observations that are subjective in nature. In some studies this is likely essential, however in other studies where there is observable features for future out-of-sample data, it might be wise to first consider those as variables.

Would a better method for utilizing categorical variables be to perform an extensive EDA with just the continuous variables and then once you have a model that you feel confident about, add in indicator variables and check performance, complexity and interpret-ability.

Procedures

```
title 'Assignment 5';
libname mydata '/scs/crb519/PREDICT_410/SAS_Data/' access=readonly;

* create a temporary variable (data source is read only);
data ames;
  set mydata.ames_housing_data;

ods graphics off;

proc contents data=ames;

data ames_cleaned;
  set ames;
    if (SaleCondition ne 'Normal')
      then delete;
    else if (Functional ne 'Typ')
      then delete;

proc contents data=ames_cleaned;

data ames_indicator;
  set ames_cleaned;
    keep SalePrice GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea OverallQual HouseStyle hs_1 hs
    * Create Indicator Variables from Categorical Variable;
    if HouseStyle in ('1Story' '1.5Fin' '1.5Unf' '2Story' '2.5Fin' '2.5Unf' 'SFoyer' 'SLvl') then do;
      hs_1 = (HouseStyle eq '1Story');
      hs_2 = (HouseStyle eq '1.5Fin');
      hs_3 = (HouseStyle eq '1.5Unf');
      hs_4 = (HouseStyle eq '2Story');
      hs_5 = (HouseStyle eq '2.5Fin');
      hs_6 = (HouseStyle eq '2.5Unf');
      hs_7 = (HouseStyle eq 'SFoyer');
      hs_8 = (HouseStyle eq 'SLvl');
    end;
```

```
    * Recode Categorical Variable so we can use in strait reg;
    if HouseStyle='1Story' then HouseStyle=1;
    if HouseStyle='1.5Fin' then HouseStyle=2;
    if HouseStyle='1.5Unf' then HouseStyle=3;
    if HouseStyle='2Story' then HouseStyle=4;
    if HouseStyle='2.5Fin' then HouseStyle=5;
    if HouseStyle='2.5Unf' then HouseStyle=6;
    if HouseStyle='SFoyer' then HouseStyle=7;
    if HouseStyle='SLvl' then HouseStyle=8;
    * Create Indicator Variables from Categorical Variable;
    if GarageType in ('2Types' 'Attchd' 'Basment' 'BuiltIn' 'CarPort' 'Detchd' 'NA') then do;
      gt_1 = (GarageType eq '2Types');
      gt_2 = (GarageType eq 'Attchd');
      gt_3 = (GarageType eq 'Basment');
      gt_4 = (GarageType eq 'BuiltIn');
      gt_5 = (GarageType eq 'CarPort');
      gt_6 = (GarageType eq 'Detchd');
      gt_7 = (GarageType eq 'NA');
    end;
    * Recode Categorical Variable so we can use in strait reg;
    if GarageType='2Types' then GarageType=1;
    if GarageType='Attchd' then GarageType=2;
    if GarageType='Basment' then GarageType=3;
    if GarageType='BuiltIn' then GarageType=4;
    if GarageType='CarPort' then GarageType=5;
    if GarageType='Detchd' then GarageType=6;
    if GarageType='NA' then GarageType=7;

proc means data=ames_indicator;
  class HouseStyle;
  var SalePrice;

proc freq data=ames_indicator;
  tables HouseStyle hs_1 hs_2 hs_3 hs_4 hs_5 hs_6 hs_7 hs_8;

proc means data=ames_indicator;
  class GarageType;
  var SalePrice;

proc freq data=ames_indicator;
  tables GarageType gt_1 gt_2 gt_3 gt_4 gt_5 gt_6 gt_7;

proc reg data=ames_indicator;
  model SalePrice = HouseStyle;

proc reg data=ames_indicator;
  model SalePrice = hs_1 hs_2 hs_3 hs_4 hs_5 hs_6 hs_7;

proc reg data=ames_indicator outest=reg_rsq_out;
model SalePrice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5 gt_6 g
selection=adjrsq aic bic cp best=5;

proc print data=reg_rsq_out;
```

```
proc reg data=ames_indicator outest=reg_cp_out;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5 gt_
  selection=cp adjrsq aic bic cp best=5;

proc print data=reg_cp_out;

proc reg data=ames_indicator outest=reg_forward_out;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5 gt_
  selection=forward adjrsq aic bic cp best=5;

proc print data=reg_forward_out;

proc reg data=ames_indicator outest=reg_backward_out;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5 gt_
  selection=backward adjrsq aic bic cp best=5;

proc print data=reg_backward_out;

proc reg data=ames_indicator outest=reg_stepwise_out;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5 gt_
  selection=stepwise adjrsq aic bic cp best=5;

proc print data=reg_stepwise_out;

*ods graphics on;

proc reg data=ames_indicator;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5 gt_6 gt_7 hs_1

proc reg data=ames_indicator;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea;

proc reg data=ames_indicator;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF;

*ods graphics off;

data ames_training;
  set ames_indicator;
  u = uniform(123);
  if (u < 0.70) then train = 1;
    else train = 0;
      if (train=1) then train_response=SalePrice;
        else train_response=.;


proc reg data=ames_training outest=reg_rsq_out;
model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5 g
selection=adjrsq aic bic cp best=5;

proc print data=reg_rsq_out;

proc reg data=ames_training outest=reg_cp_out;
  model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_
```

```
  selection=cp adjrsq aic bic cp best=5;

proc print data=reg_cp_out;

proc reg data=ames_training outest=reg_forward_out;
  model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5
  selection=forward adjrsq aic bic cp best=5;

proc print data=reg_forward_out;

proc reg data=ames_training outest=reg_backward_out;
  model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5
  selection=backward adjrsq aic bic cp best=5;

proc print data=reg_backward_out;

proc reg data=ames_training outest=reg_stepwise_out;
  model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5
  selection=stepwise adjrsq aic bic cp best=5;

proc print data=reg_stepwise_out;

proc reg data=ames_training;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea gt_1 gt_2 gt_3 gt_4 gt_5 gt_6 gt_7 hs_1
  output out=reg_indicators_yhat predicted=yhat;

proc reg data=ames_training;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF;
  output out=reg_nonindicators_yhat predicted=yhat;

data indicator_mae;
  set reg_indicators_yhat;
    mae = abs (yhat - train_response);

data non_indicator_mae;
  set reg_nonindicators_yhat;
    mae = abs (yhat - train_response);

proc means data=indicator_mae;
  var mae;

proc means data=non_indicator_mae;
  var mae;

data indicator_performance;
  set reg_indicators_yhat;
    if train_response = . then delete;
    length prediction_grade $7.;
    p_d = abs((yhat - train_response) / train_response);
    if p_d le 0.10 then p_g = 'Grade 1';
      else if p_d gt 0.10 and p_d le 0.15 then p_g = 'Grade 2';
      else p_g = 'Grade 3';

data nonindicator_performance;
```

```
  set reg_nonindicators_yhat;
    if train_response = . then delete;
    length prediction_grade $7.;
    p_d = abs((yhat - train_response) / train_response);
    if p_d le 0.10 then p_g = 'Grade 1';
      else if p_d gt 0.10 and p_d le 0.15 then p_g = 'Grade 2';
      else p_g = 'Grade 3';

proc freq data=indicator_performance;
  tables p_g;

proc freq data=nonindicator_performance;
  tables p_g;

run;
```

A Frustrating Direction

Initially we used correlation as a criteria for choosing a categorical variable to model with SalePrice. OveralQual came out at the top:

Categorical Variable	Correlation to SalePrice	Prob > $\|r\|$ under H_0: ρ=0	Number of Observations
OverallQual	0.79926	<0.0001	2930
GarageCars	0.64788	<0.0001	2929
YearBuilt	0.55843	<0.0001	2930
FullBath	0.54560	<0.0001	2930
YearRemodel	0.53297	<0.0001	2930
GarageYrBlt	0.52697	<0.0001	2771
Fireplaces	0.47456	<0.0001	2930

Table 31: Correlation of Categorical Variables to Sale Price

We chose OverallQual to model as an indicator variable. The model became unwieldy over time, and we realize that a primary goal should be in interpretation of the model. We'll include in this section our analysis of OverallQual as an indicator variable so that we have it as reference in the future.

OverallQual as a potential Indicator Variable

We'll choose to examine the categorical variable OverallQual as it was the variable that offered the best performance when doing a simple linear regression to SalePrice. We realize this variable is a Lickert scale [1], 10 being Very Excellent. and 1 being Very Poor. To investigate, we perform a sort and means procedure, obtaining:

N	OverallQual	SalePrice Mean
4	1	48725
13	2	52325.31
40	3	83185.98
226	4	106485.10
825	5	134752.52
732	6	162130.32
602	7	205025.76
350	8	270913.59
107	9	368336.77
31	10	450217.32

Table 32: Sorted Means procedure with OveralQual

We run a simple linear regression model:

$$\text{SalePrice} = \beta_0 + \beta_1 \text{OverallQual} + \epsilon$$

And we get the parameter estimation and diagnostic information:

$$\text{SalePrice} = 45251 \times \text{OverallQual} - 95004$$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-95004	3933.82223	-24.15	<0.0001
OverallQual	1	45251	628.80511	71.96	<0.0001

Table 33: Model Parameter Estimates for SalePrice = OverallQual

Source	
Root MSE	48019
R-Square	0.6388
Adj R-Square	0.6387
F Value	5178.75

Table 34: Model Estimator Performance for SalePrice = OverallQual

We'll compute the pass-through points for the Model $\text{SalePrice} = \beta_0 + \beta_1 \text{OverallQual} + \epsilon$.

$$\text{SalePrice} = 45251 \times 1 - 95004 = -49753$$
$$\text{SalePrice} = 45251 \times 2 - 95004 = -4502$$
$$\text{SalePrice} = 45251 \times 3 - 95004 = 40749$$
$$\text{SalePrice} = 45251 \times 4 - 95004 = 86000$$
$$\text{SalePrice} = 45251 \times 5 - 95004 = 131251$$
$$\text{SalePrice} = 45251 \times 6 - 95004 = 176502$$
$$\text{SalePrice} = 45251 \times 7 - 95004 = 221753$$
$$\text{SalePrice} = 45251 \times 8 - 95004 = 267004$$
$$\text{SalePrice} = 45251 \times 9 - 95004 = 312255$$
$$\text{SalePrice} = 45251 \times 10 - 95004 = 357506$$

The predicted model appears to only be relatively close at points 3, 5, 6, 8.

OverallQual Indicator Variable, Model & Interpretation

OverallQual is an interesting parameter, because it is a 10-way Lickert some would choose to incorporate the parameter into a model as a continuous variable. Above we model it as a continuous parameter, here will we dummy code it and model it as an indicator variable. Although OverallQual ranges from 1-10, we can use nine variables to investigate, this is simpler to consider as a table:

OverallQual	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	$x8$	$x9$
1	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	0	1	0	0
8	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0

Table 35: Modeling 10-Way Lickert OverallQual with 9 Indicator Variables

We will use the data procedure to dummy code OverallQual as an indicator variable. To examine our progress we evaluate a proc freq of the OverallQual:

OverallQual	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	4	0.14	4	0.14
2	13	0.44	17	0.58
3	40	1.37	57	1.95
4	226	7.71	283	9.66
5	825	28.16	1108	37.82
6	732	24.98	1840	62.80
7	602	20.55	2442	83.34
8	350	11.95	2792	95.29
9	107	3.65	2899	98.94
10	31	1.06	2930	100.00

Table 36: Frequency OverallQual

For brevity we examine the oc_1 and oc_2 variables:

oc_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2926	99.86	2926	99.86
1	4	0.14	2930	100.00

Table 37: Frequency oc_1, Indicator Variable for OverallQual 1

oc_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2917	99.56	2917	99.56
1	13	0.44	2930	100.00

Table 38: Frequency oc_2, Indicator Variable for OverallQual 2

We notice that the oc_1 and oc_2 variables are properly coded to match up with the OverallQual frequency table, with oc_1 have 1 coded 4 times and oc_2 having 1 coded 13 times respectively.

We now build a model, but we hold oc_10 to be the basis of interpretation:

$$\text{SalePrice} = \beta_0 + \beta_1 \text{oc\_1} + \beta_2 \text{oc\_2} + \beta_3 \text{oc\_3} + \beta_4 \text{oc\_4} + \beta_5 \text{oc\_5} + \beta_6 \text{oc\_6} + \beta_7 \text{oc\_7} + \beta_8 \text{oc\_8} + \beta_9 \text{oc\_9} + \epsilon$$

Resulting in the parameter estimations and model diagnostics:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	450217	7841.25572	57.42	<.0001
oc_1	1	-401492	23195	-17.31	<.0001
oc_2	1	-397892	14426	-27.58	<.0001
oc_3	1	-367031	10447	-35.13	<.0001
oc_4	1	-343732	8361.76503	-41.11	<.0001
oc_5	1	-315465	7987.21777	-39.50	<.0001
oc_6	1	-288087	8005.57159	-35.99	<.0001
oc_7	1	-245192	8040.61424	-30.49	<.0001
oc_8	1	-179304	8181.14487	-21.92	<.0001
oc_9	1	-81881	8904.98663	-9.19	<.0001

Table 39: Model Parameter Estimates for SalePrice = OverallQual Indicator Variables

Source	
Root MSE	43658
R-Square	0.7023
Adj R-Square	0.7013
F Value	765.22

Table 40: Model Estimator Performance for SalePrice = OverallQual Indicator Variables

We notice that our Adj. R-Square value has increased from 0.6388 to 0.7013.

The resulting Model is:

$$\begin{aligned}\text{SalePrice} = 450217 &- 401492 \times \text{oc\_1} - 397892 \times \text{oc\_2} \\ &- 367031 \times \text{oc\_3} - 343732 \times \text{oc\_4} - 315465 \times \text{oc\_5} \\ &- 288087 \times \text{oc\_6} - 245192 \times \text{oc\_7} - 179304 \times \text{oc\_8} \\ &- 81881 \times \text{oc\_9}\end{aligned}$$

If the house is of OverallQual 1, then the above model becomes:

$$\text{SalePrice} = 450217 - 401492$$

That is to say, if the OverallQual is 1, then the SalePrice in this model is 48725. Looking back at our sorted mean we see the SalePrice for OverallQual of 1 had a mean of 48725.

If the house is of OverallQual 2, then the above model becomes:

$$\text{SalePrice} = 450217 - 397892$$

That is to say, if the OverallQual is 2, then the SalePrice in this model is 52325 . Looking back at our sorted mean we see the SalePrice for OverallQual of 2 had a mean of 52325.31.

If the house is of OverallQual 3, then the above model becomes:

$$\text{SalePrice} = 450217 - 367031$$

That is to say, if the OverallQual is 3, then the SalePrice in this model is 83186. Looking back at our sorted mean we see the SalePrice for OverallQual of 3 had a mean of 83185.98.

If the house is of OverallQual 4, then the above model becomes:

$$\text{SalePrice} = 450217 - 343732$$

That is to say, if the OverallQual is 4, then the SalePrice in this model is 106485. Looking back at our sorted mean we see the SalePrice for OverallQual of 4 had a mean of 106485.10.

If the house is of OverallQual 5, then the above model becomes:

$$\text{SalePrice} = 450217 - 315465$$

That is to say, if the OverallQual is 5, then the SalePrice in this model is 134752. Looking back at our sorted mean we see the SalePrice for OverallQual of 5 had a mean of 134752.52.

If the house is of OverallQual 6, then the above model becomes:

$$\text{SalePrice} = 450217 - 288087$$

That is to say, if the OverallQual is 6, then the SalePrice in this model is 162130. Looking back at our sorted mean we see the SalePrice for OverallQual of 6 had a mean of 162130.32.

If the house is of OverallQual 7, then the above model becomes:

$$\text{SalePrice} = 450217 - 245192$$

That is to say, if the OverallQual is 7, then the SalePrice in this model is 205025. Looking back at our sorted mean we see the SalePrice for OverallQual of 7 had a mean of 205025.76.

If the house is of OverallQual 8, then the above model becomes:

$$\text{SalePrice} = 450217 - 179304$$

That is to say, if the OverallQual is 8, then the SalePrice in this model is 270913. Looking back at our sorted mean we see the SalePrice for OverallQual of 8 had a mean of 270913.59.

If the house is of OverallQual 9, then the above model becomes:

$$\text{SalePrice} = 450217 - 81881$$

That is to say, if the OverallQual is 1, then the SalePrice in this model is 368336. Looking back at our sorted mean we see the SalePrice for OverallQual of 9 had a mean of 368336.77.

If the house is of OverallQual 10, then the above model becomes:

$$\text{SalePrice} = 450217$$

That is to say, if the OverallQual is 10, then the SalePrice in this model is 450217. Looking back at our sorted mean we see the SalePrice for OverallQual of 10 had a mean of 450217.32.

It seems that we're assuming the dependent SalePrice has a linear relationship with the independent OverallQual, and that the slope does not depend on the OverallQual, but that OverallQual sets the intercept for SalePrice. The variables for $\beta_1, \ldots, \beta_9$ measure the effects of Quality ratings $1, \ldots, 9$ respectively, compared to a Quality rating of 10.For example, in this model $\beta_4 - \beta_2$ reflects the relative difference between OverallQual 4 and 2, respectively on SalePrice.

OverallQual Hypothesis Testing

$$H_0 : \beta_{1..9} = 0 \text{ versus } H_1 : \beta_{1..9} \neq 0$$

For each variable $\beta_{1..9}$ we observe that the model returned results that indicate statistical significance. This model, without a continuous variable, is highly uncomfortable to work with and interpret. Even with the Adj. R-Square value being lower for this model, and all the dependent variables showing statistical significance, it still provides discomfort to the analyst.

Automatic Variable Selection (Using OverallQual)

We ran these models, we figure that even if they were built with improper assumptions of how to use the OverallQual variable that we'd like to keep them for reference. It is interesting to see that the initial models without OverallQual have very low C_p and a 20% decrease in both AIC and BIC criterion.

Adjusted R-Square Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = 6657.59 &+ 69.5822 \times \text{GrLivArea} + 41.8777 \times \text{GarageArea} + 28.0007 \times \text{TotalBsmtSF} \\ &- 24.9846 \times \text{FirstFlrSF} + 16.3510 \times \text{MasVnrArea} + 5.19529 \times \text{BsmtFinSF1} - 16.3582 \times \text{BsmtUnfSF} \\ &+ 12658.92 \times \text{oc\_3} + 23951.92 \times \text{oc\_4} + 36624.84 \times \text{oc\_5} + 52863.67 \times \text{oc\_6} + 79166.05 \times \text{oc\_7} \\ &+ 117691.13 \times \text{oc\_8} + 188655.41 \times \text{oc\_9} + 207986.47 \times \text{oc\_10} - 24331.98 \times \text{hs\_2} - 23028.05 \times \text{hs\_4} \\ &- 58533.55 \times \text{hs\_5} - 46629.36 \times \text{hs\_6} - 7228.92 \times \text{hs\_8}\end{aligned}$$

Source	
Root MSE	33158.70
C_p	19.2110
R-Square	0.8286
Adj. R-Square	0.8274
AIC	60497.5649
BIC	60499.8969

Table 41: Model Performance

We generally expect that the selection method will result in selection of many dependent variables. We hope moving forward with the other selection methods that they are not as egregious with their incorporation of variables. We notice that of the listed models, the C_p for this model, with this method, was the lowest of the top 5.

Mallow's C_p Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = 15358.36 + 69.4728 \times \text{GrLivArea} + 41.9804 \times \text{GarageArea} + 32.6848 \times \text{TotalBsmtSF} \\ -24.8346 \times \text{FirstFlrSF} + 16.5745 \times \text{MasVnrArea} - 20.8651 \times \text{BsmtUnfSF} + 15042.94 \times \text{oc\_4} \\ +27574.85 \times \text{oc\_5} + 43883.85 \times \text{oc\_6} + 70234.08 \times \text{oc\_7} + 108726.48 \times \text{oc\_8} \\ +179964.93 \times \text{oc\_9} + 199334.08 \times \text{oc\_10} - 24111.27 \times \text{hs\_2} - 22835.81 \times \text{hs\_4} \\ -58314.95 \times \text{hs\_5} - 46558.00 \times \text{hs\_6} - 7327.98 \times \text{hs\_8}\end{aligned}$$

Source	
Root MSE	33158.70
C_p	18.7190
R-Square	0.8284
Adj. R-Square	0.8273
AIC	60497.0985
BIC	60499.90

Table 42: Model Performance

AIC Selection (Analyst Examination of Mallow's C_p results)

We realize that the regression procedure within SAS does not allow for selection by AIC criterion. We therefor examine the output of the C_p selection and choose the model with the lowest value of AIC.

Model Selected:

$$\begin{aligned}\text{SalePrice} = 15358.36 &+ 69.4728 \times \text{GrLivArea} + 41.9804 \times \text{GarageArea} + 32.6848 \times \text{TotalBsmtSF} \\ &- 24.8346 \times \text{FirstFlrSF} + 16.5745 \times \text{MasVnrArea} - 20.8651 \times \text{BsmtUnfSF} + 15042.94 \times \text{oc\_4} \\ &+ 27574.85 \times \text{oc\_5} + 43883.85 \times \text{oc\_6} + 70234.08 \times \text{oc\_7} + 108726.48 \times \text{oc\_8} \\ &+ 179964.93 \times \text{oc\_9} + 199334.08 \times \text{oc\_10} - 24111.27 \times \text{hs\_2} - 22835.81 \times \text{hs\_4} \\ &- 58314.95 \times \text{hs\_5} - 46558.00 \times \text{hs\_6} - 7327.98 \times \text{hs\_8}\end{aligned}$$

Source	
Root MSE	33158.70
C_p	18.7190
R-Square	0.8284
Adj. R-Square	0.8273
AIC	60497.0985
BIC	60499.90

Table 43: Model Performance

We had initially expected to see the AIC selection criterion result in a model with few parameters due to AIC formulation having a built in penalty as an increasing function of the number of estimated parameters. We are sadly disappointed an have received yet another large model.

Forward Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = -360.76459 &+ 69.68463 \times \text{GrLivArea} + 41.76716 \times \text{GarageArea} + 27.93859 \times \text{TotalBsmtSF} \\ &- 25.34106 \times \text{FirstFlrSF} + 16.33439 \times \text{MasVnrArea} + 5.28855 \times \text{BsmtFinSF1} - 16.31118 \times \text{BsmtUnfSF} \\ &+ 12950 \times \text{oc\_3} + 24209 \times \text{oc\_4} + 336937 \times \text{oc\_5} + 53201 \times \text{oc\_6} + 79464 \times \text{oc\_7} \\ &+ 117993 \times \text{oc\_8} + 189014 \times \text{oc\_9} + 208487 \times \text{oc\_10} + 7253.75366 \times \text{hs\_1} - 17389 \times \text{hs\_2} \\ &- 16106 \times \text{hs\_4} - 51606 \times \text{hs\_5} - 39701 \times \text{hs\_6} + 5452.76119 \times \text{hs\_7}\end{aligned}$$

Source	
Root MSE	33160.21
C_p	20.4741
R-Square	0.82862
Adj. R-Square	0.82737
F Value	663.78
AIC	60498.82
BIC	60501.18

Table 44: Model Performance

Backward Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = 210542.11 &+ 669.0064 \times \text{GrLivArea} + 41.9546 \times \text{GarageArea} + 32.9086 \times \text{TotalBsmtSF} \\ &- 24.8742 \times \text{FirstFlrSF} + 16.5710 \times \text{MasVnrArea} - 21.1202 \times \text{BsmtUnfSF} - 218229 \times \text{oc\_1} \\ &- 205920.52 \times \text{oc\_2} - 196002 \times \text{oc\_3} - 184601 \times \text{oc\_4} - 172059 \times \text{oc\_5} - 155840 \times \text{oc\_6} \\ &- 129483 \times \text{oc\_7} - 90909 \times \text{oc\_8} - 19592 \times \text{oc\_9} + 5212 \times \text{hs\_1} - 18991 \times \text{hs\_2} \\ &- 17496 \times \text{hs\_4} - 52459 \times \text{hs\_5} - 41077 \times \text{hs\_6}\end{aligned}$$

Source	
Root MSE	33171.58
C_p	21.4498
R-Square	0.82844
Adj. R-Square	0.82725
F Value	696.34
AIC	60499.82
BIC	60502.12

Table 45: Model Performance

Stepwise Selection

Model Selected:

$$\begin{aligned}\text{SalePrice} = 10670.55 &+ 68.9622 \times \text{GrLivArea} + 42 \times \text{GarageArea} + 33.0526 \times \text{TotalBsmtSF} \\ &- 24.7990 \times \text{FirstFlrSF} + 16.5238 \times \text{MasVnrArea} - 21.0767 \times \text{BsmtUnfSF} \\ &+ 15150 \times \text{oc\_4} + 27671 \times \text{oc\_5} + 43855 \times \text{oc\_6} + 70189 \times \text{oc\_7} + 108725 \times \text{oc\_8} \\ &+ 179998 \times \text{oc\_9} + 199501 \times \text{oc\_10} + 5102 \times \text{hs\_1} - 18928 \times \text{hs\_2} - 17456 \times \text{hs\_4} \\ &- 52435 \times \text{hs\_5} - 41063 \times \text{hs\_6}\end{aligned}$$

Source	
Root MSE	33172.66
C_p	19.6388
R-Square	0.82831
Adj. R-Square	0.82724
F Value	773.54
AIC	60498.02
BIC	60500.27

Table 46: Model Performance

Remarks on Comparing Performance

We'll make a table to compare the model performance information

Model	Cont.	Ind.	Root MSE	C_p	R-Square	Adj. R-Square	F Value	AIC	BIC
Adj. R-Square	7	13	33158.70	19.2110	0.8286	0.8274	-	60497.5649	60499.8969
Mallow's C_p	6	12	33158.70	18.7190	0.8284	0.8273	-	60497.0985	60499.90
AIC	6	12	33158.70	18.7190	0.8284	0.8273	-	60497.0985	60499.90
Forward	7	14	33160.21	20.4741	0.82862	0.82737	663.78	60498.82	60501.18
Backward	6	14	33171.58	21.4498	0.82844	0.82725	696.34	60499.82	60502.12
Stepwise	6	12	33172.66	19.6388	0.82831	0.82724	773.54	60498.02	60500.27

It seems relevant to mention that models which incorporate more parameters become more complex for interpretation. Going into the variable selection, we had anticipated that Mallow's C_p and AIC would result in models of greatly reduced complexity (parameters), however the results show that these models all performed well by incorporating almost all of the continuous variables in them.

For Mallow's C_p and AIC, we received the same results because we were looking to minimize AIC. The Mallow's C_p method found the models with the lowest AIC, so we naturally used that same model for the AIC selection criteria.

In terms of an interpretable model, we're not a fan of our initial selection of OverallQual based solely on correlation criteria. This variable is large (10-way Lickert) and now that we've performed automated variable selection we'll have to incorporate it into our ultimate model. There is some concern that OverallQual is a subjective categorical measurement, as opposed to HouseStyle which is an observable categorical measurement. This likely means that we are vectoring towards building a model that will be more useful for inference than prediction. we say this because in sample we have observations of OverallQual, but out-of- sample there is no systematic way of observing and characterizing OverallQual, this is something obtained through the survey methodology.

References

[1]Wikipedia, “Likert scale — wikipedia, the free encyclopedia.” 2015 [Online]. Available: http://en.wikipedia.org/w/index.php?title=Likert_scale&oldid=653173542