Assignment #5
Matthew Dobbin
18/07/2017

Introduction:
The purpose of this report is to document the analysis of the Ames, Iowa housing data and the building of four regression models for the sales price of a typical home. To build the regression models three different automated variable selection methods are used. They are forward, backward and stepwise variable selection. The following sections of the report discuss the in-sample model fit, predictive accuracy and operational validation for each variable selection method.

Section 1: Sample Definition and Data Split:
This section of the report provides documentation on the Ames Iowa housing dataset and the sample that was selected for the analysis.

Section 1.1: Sample Definition
The Ames dataset has 2930 observations and 82 variables containing information from individual residential properties sold in Ames from 2006 to 2010. A review of the data dictionary revealed that there were several observations in the data that should be excluded. The data dictionary recommended that partial sales should be excluded as they do not represent actual market values. All sales that were not normal were excluded. It also recommended that homes with living area above ground greater than 4000 square feet should be excluded as they were not typical of the dataset. Sales of homes greater than 4000 square feet were deemed not typical.

The purpose of the regression model is to predict the sales price of a typical home in Ames. In the Ames dataset, there were several different zoning classifications. Table 1 shows the different classifications and the count of sales for each classification.

Table 1 - Count of Sales by Zoning Classification.

Zoning Classification	Count of Sales
Agriculture	2
Commercial	25
Floating Village Residential	139
Industrial	2
Residential High Density	27
Residential Low Density	2273
Residential Medium Density	462
NA	0

Due to the high count of residential low density sales, all other types of zones will be excluded. It was also expected that a typical home would have at least one bedroom above ground and a garage area. A waterfall of the conditions that was used to drop the observations that were not typical of the dataset are shown in Table 2 as well as a count of the observations that were dropped and the count of the remaining eligible sample.

Table 2-Drop condition waterfall table.

#	Drop Condition	Count of Sales
01	Building Type Not Single Family Detached	505
02	Not Residential Low Density	416
03	Not Normal Sale	330
04	Not Paved Street	1
05	Built Pre-1950	245
06	Above Ground Living Area Greater Than 4000 SQFT	1
07	Above Ground Living Area Less Than 800 SQFT	9
08	Lot Area Greater Than 100,000 SQFT	3
09	Bedrooms Above Ground Less Than 1.	4
10	Total Basement SQFT Less Than 1 SQFT.	25
11	Garage Area Less Than 1 SQFT.	19
12	Lot Frontage Greater Than 300 SQFT.	1
	Remaining Eligible Sample	1051

The dataset was reduced to 1047 observations of 44 variables by eliminating variables deemed not necessary for the regression models and omitting NA values. The list of variables that were kept are shown in Appendix 1.

Several additional variables were calculated from the existing variables. The formulas for the additional variables are shown in Appendix 1. Table 3 shows the pool of candidate predictor variables that are used in the automated variable selection. The response variable is SalePrice.

Table 3 - Pool of candidate predictor variables.

Continous Variables	Calculated Variables	Indicator Variables
LotFrontage	TotalBathCalc	garage1
LotArea	TotalSqftCalc	garage2
MasVnrArea	QualityIndex	garage3
LowQualFinSF		CornerLotInd
GarageArea		FireplaceInd1
BsmtUnfSF		FireplaceInd2
WoodDeckSF		PoolInd
OpenPorchSF		I2007
MiscVal		I2008
PoolArea		I2009
GrLivArea		I2010

Section 1.2: The Train/Test Split

A uniform random number was used to split the sample into training and test data sets. The training set is used for in-sample model development and the test set used for out-of-sample model assessment. Table 4 shows the count of observations in each data set.

Table 4 - Count of observations for the training and test data sets.

Data set	Count of Observations
Training Set	744
Test Set	303

Section:2 Model Identification and In-Sample Model Fit

Using the cleaned training data set the best models were found using automated variable selection techniques. The final model estimates for the forward, backward and stepwise variable selection methods are shown in the following sections.

Section 2.1 Forward Variable Selection

The model that resulted from the forwarded variable selection contained 16 predictor variables. The model estimates are shown in Table 5. Table 6 shows the variance inflation factors (VIF) for the predictor variables.

Table 5 - Forward variable selection model estimates.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -66684.569   7092.659  -9.402  < 2e-16 ***
GrLivArea        11.705      4.378   2.674 0.007675 **
garage3       40743.532   3760.332  10.835  < 2e-16 ***
TotalSqftCalc    42.310      3.312  12.775  < 2e-16 ***
QualityIndex   1739.323    151.874  11.452  < 2e-16 ***
BsmtUnfSF        33.274      3.361   9.901  < 2e-16 ***
LotArea           2.012      0.281   7.160 1.98e-12 ***
TotalBathCalc  13241.519   1881.803   7.037 4.56e-12 ***
GarageArea       35.239      8.441   4.175 3.35e-05 ***
MasVnrArea       24.346      6.285   3.874 0.000117 ***
PoolInd      176879.701  25125.052   7.040 4.46e-12 ***
PoolArea       -341.340     55.513  -6.149 1.29e-09 ***
OpenPorchSF      50.825     16.752   3.034 0.002500 **
WoodDeckSF       19.172      7.625   2.514 0.012136 *
MiscVal          -4.805      2.904  -1.655 0.098435 .
LotFrontage     -86.537     59.619  -1.452 0.147068
FireplaceInd1 -2871.003   1987.217  -1.445 0.148963
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24720 on 727 degrees of freedom
Multiple R-squared:  0.8953,   Adjusted R-squared:  0.893
F-statistic: 388.5 on 16 and 727 DF,  p-value: < 2.2e-16
```

VIF measures how correlated each independent variable is with the other predictor variables in the model and is used to detect multicollinearity. An issue with the variable selection methods is that if the predictor variable pool contains highly correlated predictor variables the algorithm will select them. A VIF value larger than 20 implies a large inflation of standard errors due to the variable being included in the model. The values shown in Table 6 are all below ten therefore no variables were removed. Multicollinearity does not reduce the predictive ability of the model but it does affect statistical inference.

Table 6 - VIF values for the predictor variables in the foward variable selection model.

PoolInd	PoolArea	TotalSqftCalc	GrLivArea	GarageArea
7.163372	7.134967	6.708535	5.51896	2.551548
BsmtUnfSF	**TotalBathCalc**	**garage3**	**MasVnrArea**	**QualityIndex**
2.548191	2.447455	2.375676	1.683339	1.40325
LotFrontage	**LotArea**	**OpenPorchSF**	**FireplaceInd1**	**WoodDeckSF**
1.305615	1.298458	1.263805	1.201324	1.185547
MiscVal				
1.04269				

Section 2.2 Backward Variable Selection

The model that resulted from the backward variable selection contained 17 predictor variables. The model estimates are shown in Table 7. Table 8 shows the VIF values for the predictor variables. The values are all below ten therefore no variables were removed.

Table 7 - Model estimates for the backward variable selection method.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.477e+04  1.175e+04  -2.108 0.035330 *
LotFrontage   -8.625e+01  5.968e+01  -1.445 0.148867
LotArea        2.010e+00  2.813e-01   7.146 2.18e-12 ***
MasVnrArea     2.436e+01  6.290e+00   3.873 0.000117 ***
MiscVal       -4.820e+00  2.907e+00  -1.658 0.097774 .
BsmtUnfSF      3.326e+01  3.364e+00   9.886  < 2e-16 ***
GrLivArea      1.160e+01  4.424e+00   2.622 0.008912 **
GarageArea     3.409e+01  1.090e+01   3.128 0.001833 **
WoodDeckSF     1.920e+01  7.632e+00   2.516 0.012076 *
OpenPorchSF    5.089e+01  1.677e+01   3.035 0.002491 **
PoolArea      -3.413e+02  5.555e+01  -6.144 1.33e-09 ***
TotalBathCalc  1.316e+04  1.952e+03   6.738 3.26e-11 ***
TotalSqftCalc  4.239e+01  3.352e+00  12.649  < 2e-16 ***
QualityIndex   1.739e+03  1.520e+02  11.442  < 2e-16 ***
garage1       -4.159e+04  6.357e+03  -6.543 1.14e-10 ***
garage2       -4.104e+04  4.166e+03  -9.850  < 2e-16 ***
FireplaceInd1 -2.910e+03  2.002e+03  -1.453 0.146582
PoolInd        1.769e+05  2.514e+04   7.037 4.56e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24740 on 726 degrees of freedom
Multiple R-squared:  0.8953,   Adjusted R-squared:  0.8928
F-statistic: 365.2 on 17 and 726 DF,  p-value: < 2.2e-16
```

Table 8 - VIF values for the predictor variables in the backward selection method.

garage1	PoolInd	PoolArea	TotalSqftCalc	GrLivArea
8.733225	7.164774	7.135108	6.861457	5.628363
garage2	GarageArea	TotalBathCalc	BsmtUnfSF	MasVnrArea
5.050709	4.249706	2.631011	2.550421	1.683824
QualityIndex	LotFrontage	LotArea	OpenPorchSF	FireplaceInd1
1.403458	1.306729	1.299739	1.264571	1.218067
WoodDeckSF	MiscVal			
1.186348	1.043708			

Section 2.3 Stepwise Variable Selection

The model that resulted from the stepwise variable selection contained 16 predictor variables. The model estimates are shown in Table 9. Table 10 shows the VIF values for the predictor variables. The values are all below ten therefore no variables were removed. The model estimates for the stepwise variable selection are identical to the forward variable selection (although they are arranged in a different order).

Table 9 - Model estimates for the stepwise variable selection model.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -66684.569   7092.659  -9.402  < 2e-16 ***
TotalSqftCalc    42.310      3.312  12.775  < 2e-16 ***
BsmtUnfSF        33.274      3.361   9.901  < 2e-16 ***
garage3       40743.532   3760.332  10.835  < 2e-16 ***
QualityIndex   1739.323    151.874  11.452  < 2e-16 ***
TotalBathCalc 13241.519   1881.803   7.037 4.56e-12 ***
LotArea           2.012      0.281   7.160 1.98e-12 ***
GarageArea       35.239      8.441   4.175 3.35e-05 ***
GrLivArea        11.705      4.378   2.674 0.007675 **
MasVnrArea       24.346      6.285   3.874 0.000117 ***
PoolInd      176879.701  25125.052   7.040 4.46e-12 ***
PoolArea       -341.340     55.513  -6.149 1.29e-09 ***
OpenPorchSF      50.825     16.752   3.034 0.002500 **
WoodDeckSF       19.172      7.625   2.514 0.012136 *
MiscVal          -4.805      2.904  -1.655 0.098435 .
LotFrontage     -86.537     59.619  -1.452 0.147068
FireplaceInd1 -2871.003   1987.217  -1.445 0.148963
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24720 on 727 degrees of freedom
Multiple R-squared:  0.8953,   Adjusted R-squared:  0.893
F-statistic: 388.5 on 16 and 727 DF,  p-value: < 2.2e-16
```

Table 10 - VIF values for the predictor variables in the stepwise model.

PoolInd	PoolArea	TotalSqftCalc	GrLivArea	GarageArea
7.163372	7.134967	6.708535	5.51896	2.551548
BsmtUnfSF	**TotalBathCalc**	**garage3**	**MasVnrArea**	**QualityIndex**
2.548191	2.447455	2.375676	1.683339	1.40325
LotFrontage	**LotArea**	**OpenPorchSF**	**FireplaceInd1**	**WoodDeckSF**
1.305615	1.298458	1.263805	1.201324	1.185547
MiscVal				
1.04269				

Section 2.4 Junk Model

A junk model was created for model comparison purposes. The variables were selected manually and the model estimates are shown in Table 11.

Table 11 - Model estimates for the junk model.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.689e+05  3.812e+04  -9.675  < 2e-16 ***
OverallQual   7.049e+04  6.496e+03  10.851  < 2e-16 ***
OverallCond   4.637e+04  7.066e+03   6.562 9.99e-11 ***
QualityIndex -7.632e+03  1.226e+03  -6.225 8.06e-10 ***
GrLivArea     2.932e+01  4.015e+00   7.302 7.36e-13 ***
TotalSqftCalc 3.923e+01  2.347e+00  16.717  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28430 on 738 degrees of freedom
Multiple R-squared:  0.8594,   Adjusted R-squared:  0.8585
F-statistic: 902.3 on 5 and 738 DF,  p-value: < 2.2e-16
```

Table 12 shows the VIF values for the predictor variables. The VIF values indicate that there are highly correlated variables in the model. This is expected as QualityIndex is a calculated variable and is the product of OverallQual and OverallCond.

Table 12 - VIF values for the variables in the junk model.

QualityIndex	OverallQual	OverallCond	GrLivArea	TotalSqftCalc
69.137376	62.78268	37.60615	3.509018	2.546607

Section 2.5 Model Comparison

To compare the in-sample fit of the model the adjusted R-Squared, Akaike's Information Criterion (AIC), Bayes Information Criterion (BIC), mean squared error (MSE) and the mean absolute error (MAE) values were computed. Table 13 shows the values for each of the models.

Table 13 - Computed values to compare in-sample model fit.

Model	R^2adj	AIC	BIC	MSE	MAE
Forward	0.893	17181.84	17264.85	597104663	18425.7
Backward	0.8928	17183.81	17271.44	597082017	18424.72
Stepwise	0.893	17181.84	17264.85	597104663	18425.7
Junk	0.8585	17379.04	17411.32	801687820	20725.21

Each of the metrics represents some concept of fit. Table 14 shows the performance ranking for the models in each metric. The Junk model had the least well fitted values for all the metrics. The forward and stepwise models had the better fit values for adjusted R-squared, AIC and BIC. For MSE and MAE, the backward model had the least error.

Table 14 - Model ranking for each concept of 'fit' metric.

Model	R^2adj	AIC	BIC	MSE	MAE
Forward	1	1	1	2	2
Backward	2	2	2	1	1
Stepwise	1	1	1	2	2
Junk	3	3	3	3	3

Section 3: Predictive Accuracy

The out-of-sample performance (predictive accuracy) of the four models was investigated by computing the mean squared error and the mean absolute error. A comparison of the In-Sample and Out-of-Sample values are shown in Table 15.

Table 15 - Comparison of In-Sample and Out-of-Sample MSE and MAE values.

	MSE		MAE	
Model	**In-Sample**	**Out-of-Sample**	**In-Sample**	**Out-of-Sample**
Forward	597104663	734125655	18425.7	18302.27
Backward	597082017	733231385	18424.72	18280.5
Stepwise	597104663	734125655	18425.7	18302.27
Junk	801687820	824969324	20725.21	20705.86

The junk model has the highest MSE and MAE values. The backward model has the smallest MSE and MAE. Based on using these criteria as the determining factor of performance it could be said the backward model is the best performing model. However there is minimal difference between the errors of the forward, backward and stepwise models. The mean squared error and the mean absolute

error are two different metrics to assess the Out-of-Sample predictive accuracy. The mean absolute error is useful in understanding the magnitude of the error. The mean squared error can be sensitive to large error values due to the squaring process. Except for the junk model MAE, the In-Sample values had less error than the out-of-sample values. When the training data (In-Sample) fits better than the test data (Out-of-Sample) then the models are overfitted.

Section 4: Operational Validation

The predictive accuracy of the models in Section 3 were analysed in the statistical sense. In this section, the predictive accuracy of the models are analysed with a focus on driving business decisions. A variable called PredictionGrade was created. The value was considered Grade 1 if the predicted value is within ten percent of the actual value, Grade 2 if it is not Grade 1 but within fifteen percent of the actual value, Grade 3 if it is not Grade 2 but within twenty-five percent of the actual value, and Grade 4 otherwise.

The prediction grades for the In-Sample training data and the Out-of-Sample test data are shown in Table 16 and Table 17 respectively. The values are shown in distribution form. Similar to the accuracy results in Section 3, there is minimal difference between the accuracy of the forward, backward and stepwise models. The junk model has the smallest percentage of Grade 1 predictions. All four models are of underwriting quality. A model is deemed as 'underwriting quality' if the model is accurate to within ten percent more than fifty percent of the time (Fannie Mae and Freddie Mac definition).

Table 16 - Prediction grades for In-Sample test data.

Model	Grade 1 (0-10%)	Grade 2 (10-15%)	Grade 3 (15-25%)	Grade 4 (25%+)
Forward	0.6142	0.1788	0.1653	0.0417
Backward	0.6129	0.1788	0.1667	0.0417
Stepwise	0.6142	0.1788	0.1653	0.0417
Junk	0.5524	0.2056	0.1653	0.0766

Table 17 - Prediction grades for Out-of-Sample test data.

Model	Grade 1 (0-10%)	Grade 2 (10-15%)	Grade 3 (15-25%)	Grade 4 (25%+)
Forward	0.6568	0.1287	0.1650	0.0495
Backward	0.6568	0.1254	0.1683	0.0495
Stepwise	0.6568	0.1287	0.1650	0.0495
Junk	0.5710	0.1947	0.1551	0.0792

Appendix 1: Sample Defintion

#Variables kept for analysis.

```
keep.vars <- c('SID','PID','LotFrontage','LotArea','LotConfig','Neighborhood','MasVnrArea',
        'LowQualFinSF', 'MiscVal', 'BsmtUnfSF',
        'HouseStyle','OverallQual','OverallCond','YearBuilt','YearRemodel','Exterior1',
        'BsmtFinSF1','BsmtFinSF2','CentralAir','GrLivArea','BsmtFullBath','BsmtHalfBath',
        'FullBath','HalfBath','BedroomAbvGr','TotRmsAbvGrd','Fireplaces','GarageCars',
        'GarageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','ThreeSsnPorch',
        'ScreenPorch','PoolArea','MoSold',
        'YrSold','SaleCondition','SalePrice', 'Heating', 'BldgType', 'TotalBsmtSF',
        'GarageType', 'Condition1');
```

#Create calculated variable

```
sample.df$TotalBathCalc <- sample.df$BsmtFullBath + 0.5*sample.df$BsmtHalfBath +
sample.df$FullBath + 0.5*sample.df$HalfBath;
sample.df$TotalSqftCalc <- sample.df$BsmtFinSF1+sample.df$BsmtFinSF2+sample.df$GrLivArea;
sample.df$QualityIndex <- sample.df$OverallQual*sample.df$OverallCond;
```

#Create indicator variables

```
sample.df$garage1 <- ifelse(sample.df$GarageCars==1,1,0);
sample.df$garage2 <- ifelse(sample.df$GarageCars==2,1,0);
sample.df$garage3 <- ifelse(sample.df$GarageCars>=3,1,0);
sample.df$CornerLotInd <- ifelse(sample.df$LotConfig=='Corner',1,0);
sample.df$FireplaceInd1 <- ifelse((sample.df$Fireplaces>0)&(sample.df$Fireplaces<2),1,0);
sample.df$FireplaceInd2 <- ifelse((sample.df$Fireplaces>1),1,0);
sample.df$PoolInd <- ifelse(sample.df$PoolArea>0,1,0);
sample.df$I2007 <- ifelse(sample.df$YrSold==2007,1,0);
sample.df$I2008 <- ifelse(sample.df$YrSold==2008,1,0);
sample.df$I2009 <- ifelse(sample.df$YrSold==2009,1,0);
sample.df$I2010 <- ifelse(sample.df$YrSold==2010,1,0);
```