ISE 5103 Intelligent Data Analytics

Homework 5 - Modeling

Daniel Carpenter & Sonaxy Mohanty

October 2022

Contents

Packages

```
# Data Wrangling
library(tidyverse)

# Modeling
library(MASS)
library(caret) # Modeling variants like SVM
library(earth) # Modeling with Mars
library(pls)  #Modeling with PLS
library(glmnet) #Modeling with LASSO

# Aesthetics
library(knitr)
library(cowplot)  # multiple ggplots on one plot with plot_grid()
library(scales)
library(kableExtra)
library(ggplot2)

#Hold-out Validation
library(caTools)

#Data Correlation
library(GGally)
library(regclass)

#RMSE Calculation
library(Metrics)

#p-value for OLS model
library(broom)

#ncvTest
library(car)
```

General Data Prep

For general data preparation, please see conceptual steps below. See `.rmd` file for detailed code.

(i) Read Data

```
# Convert all character data to factor
hd <- read.csv('housingData.csv', stringsAsFactors = TRUE) %>%

# creates new variables age, ageSinceRemodel, and ageofGarage and
  dplyr::mutate(age = YrSold - YearBuilt,
                ageSinceRemodel = YrSold - YearRemodAdd,
                ageofGarage = ifelse(is.na(GarageYrBlt), age, YrSold - GarageYrBlt)) %>%

# remove some columns used in the above calculations
```

```
  dplyr::select(!c(Id,YrSold ,
                   MoSold, YearBuilt, YearRemodAdd))
```

(ii) Impute Missing Values with PMM

Make data set of `numeric` variables `hd.numericRaw`

Make data set of `factor` variables `hd.factorRaw`

For each column with missing data, impute missing values with `PMM`

1. Imputation completed with our created function called `imputeWithPMM()`
2. Applies function to columns with missing data via dynamic `dplyr` logic
3. Note `seeImputation()` function to visualize the imputation from prior homework 4, not shown for simplicity in viewing

```
## [1] "LotFrontage" "MasVnrArea"  "GarageYrBlt"
```

```
## [1] "For imputation results of LotFrontage, see OutputPMM/Imputation_With_PMM_LotFrontage.pdf"
```

```
## [1] "For imputation results of MasVnrArea, see OutputPMM/Imputation_With_PMM_MasVnrArea.pdf"
```

```
## [1] "For imputation results of GarageYrBlt, see OutputPMM/Imputation_With_PMM_GarageYrBlt.pdf"
```

(iii) Factor Level Collapsing

Overview: Create `Other` Bin for Columns over `4` Unique Values

- Applied to any `factor` column (previously `character`) with over 4 unique values
- Applies `fct_lump()` function to columns via dynamic `dplyr` logic

```
## [1] "Before cleaning, there are 14 factor columns with more than 4 unique values"
```

```
## [1] "After cleaning, there are 14 columns with more than 4 unique values (omitting NA's)"
```

(iv) Remove Outliers from Numeric Data

Overview: Using numeric data frame, remove *some* outliers from each column without dwindling the entire data set. See steps below to create data frame `hd.CleanedNoOutliers`.

- **Please note that NOT all models use this data set with removed outliers**.
  - Only models which are sensitive to outliers use this data frame without outliers.
  - For example, the linear model using this data frame.

Outlier removal steps below:

- Since there are so many outliers in each column, we are only going to remove some outliers
- If you count the number of outliers by column, the 75% of columns contain less than 50 outliers.
- However, some contain up to 200. Since remove ALL outliers would reduce the size of the data to less than 300 observations, we are removing up to 50 per numeric column.

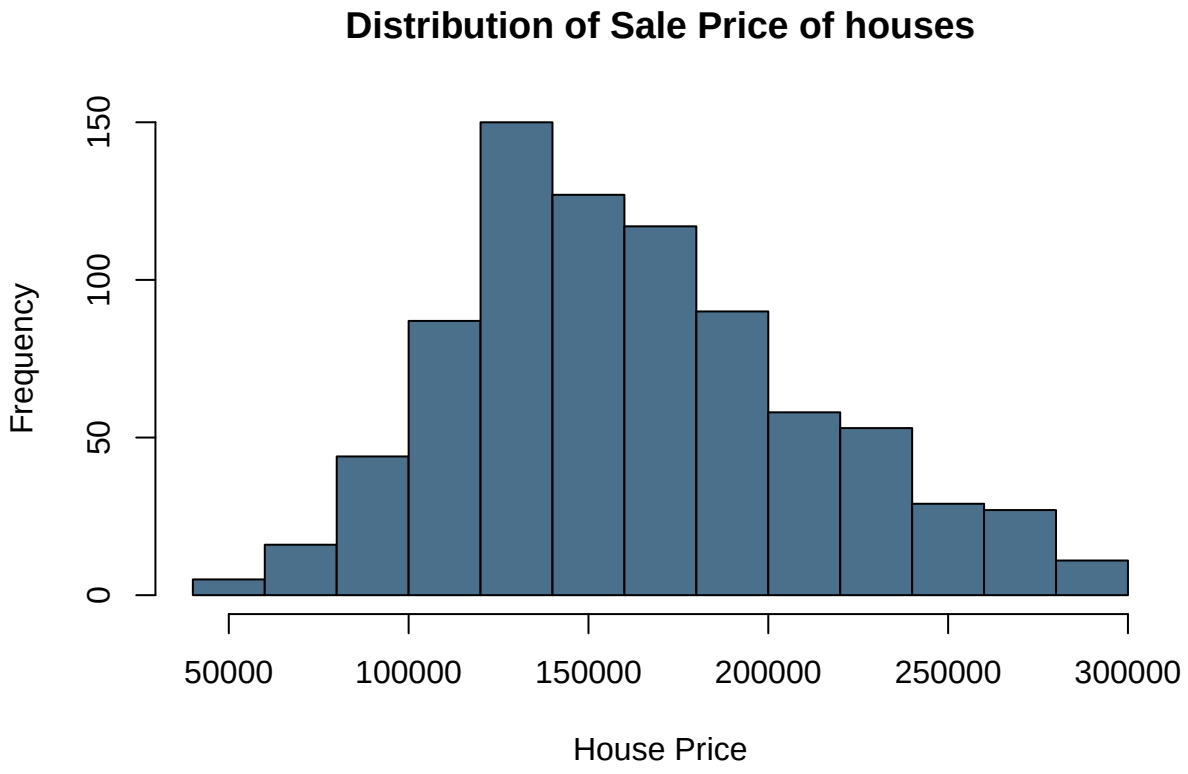
```
## [1] "Of the columns with outliers, removed up to 75th percentile of num. outliers."
```

```
## [1] "See that the 75th percentile of columns with outliers contain 51.75 outliers"
```

Exploratory Data Analysis

Checking the distribution of Sale Price of houses

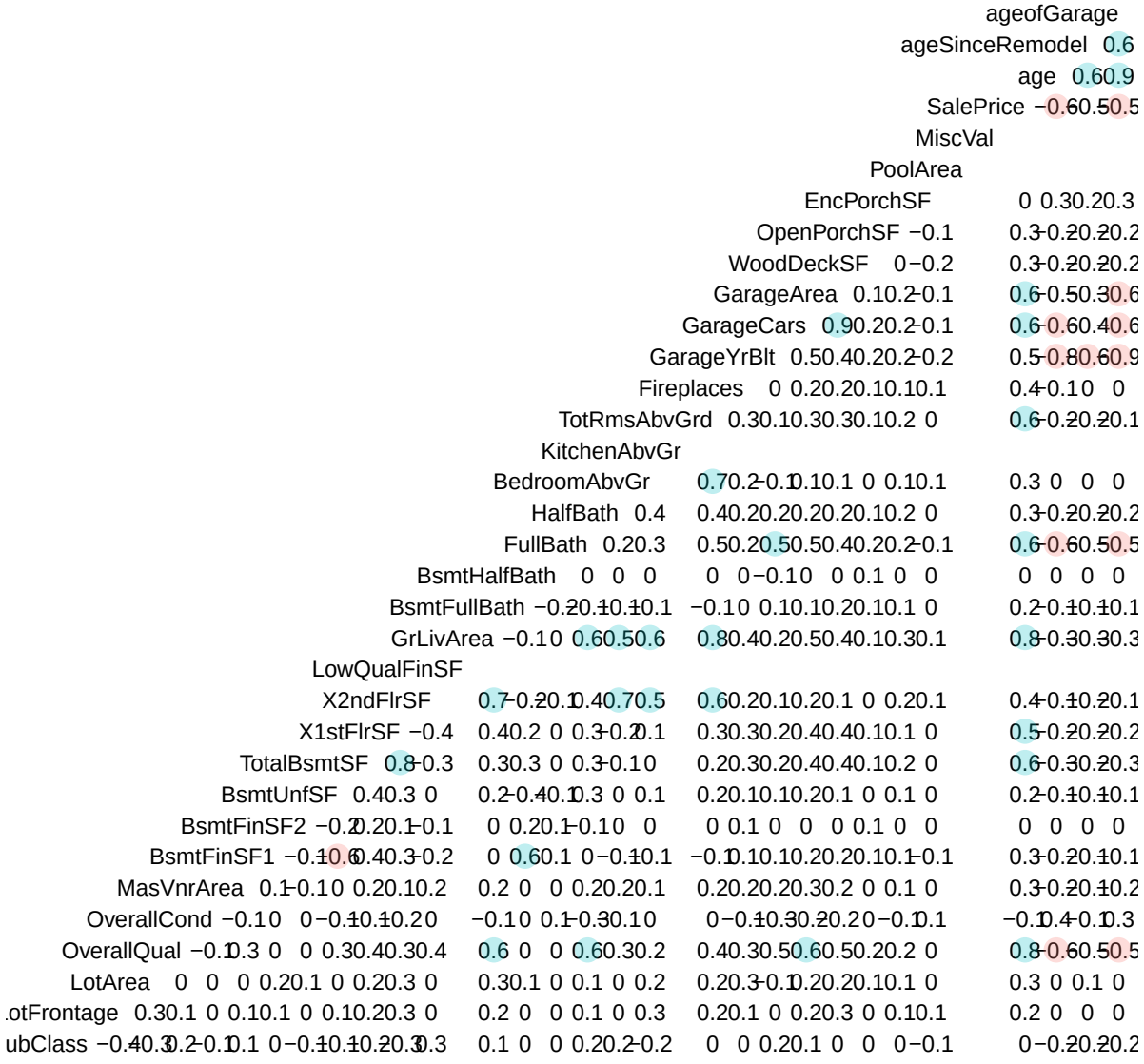
```
hist(hd.CleanedNoOutliers$SalePrice,
     col = 'skyblue4',
     main = 'Distribution of Sale Price of houses',
     xlab = 'House Price')
```



- After removing the desired outliers, we can see that the distribution of `log(Sale Price)` looks like a normal distribution with few outliers on the left tail.

Correlation between features in the dataset

```
ggcorr(hd.CleanedNoOutliers, geom='blank', label=T, label_size=3, hjust=1,
       size=3, layout.exp=2) +
  geom_point(size = 4, aes(color = coefficient > 0, alpha = abs(coefficient) >= 0.5)) +
  scale_alpha_manual(values = c("TRUE" = 0.25, "FALSE" = 0)) +
  guides(color = F, alpha = F)
```



- We can see that `SalePrice` has strong correlations with `GarageArea`, `GarageCars`, `TotRmsAbvGrd`, `FullBath`, `GrLivArea`, `X1stFlrSF`, `TotalBsmtSF`, `OverallQual`.

1 (a) - OLS Model

i.

Hold-out validation set

- Since, we have deleted some of the outlier values during data pre-processing, using 10% of the data as test and remaining 90% as train

```
idx <- sample(nrow(hd.CleanedNoOutliers), nrow(hd.CleanedNoOutliers)*0.1)
test <- hd.CleanedNoOutliers[idx,]
train <- hd.CleanedNoOutliers[-idx,]
```

Fit the OLS Model

`Model 1:`

- Linear model containing:
  - *Independent variables:* `GarageArea + GarageCars + TotRmsAbvGrd + FullBath + GrLivArea + X1stFlrSF + TotalBsmtSF + OverallQual`
  - *Predicted variable:* `SalePrice`

```
ols.mdl1 <- lm(log(SalePrice) ~ GarageArea + GarageCars + TotRmsAbvGrd
               + FullBath + GrLivArea + X1stFlrSF + TotalBsmtSF + OverallQual, data=train)
```

- **For Model 1**: Adjusted R-squared is `0.8138`, AIC is `-847.5004` and BIC is `-801.5289` and RMSE is `171456.2`.
- Still trying to improve the existing model.
- No multicollinearity detected.

`Model 2:`

- This model created is based on `Principal Component Analysis`.
  - Uses `numeric` data for Principal Component Analysis
  - Then appends the `factor` data to the data *without `NULL` values*
  - Finally, uses `stepAIC()` to best model data

Now we choose number of PC's that explain 75% of the variation

- Note this threshold is just a judgement call. No significance behind 75%

```
## [1] "There are 9 principal components that explain up to 75% of the variation in the data"
```

Fit the Model

- Linear model containing:
  - Principal components explaining 75% of variation in `numeric` data
  - Non-null `factor` data
  - *Predicted variable:* `SalePrice`
- Then use `stepAIC()` to identify which variables are actually important for model

```
# Fit data using PC's, non-null factors
fit.ols <- lm(log(SalePrice) ~ ., data = df.ols)

# Reduce to only important variables
ols.mdl2 <- stepAIC(fit.ols, direction="both")
```

- Reporting `all the variables` of the best model (`Model 2`):

Coefficient estimates:

```
##                         Estimate  Std. Error     t value      Pr(>|t|)
## (Intercept)         11.750812989 0.070572721 166.5064467  0.000000e+00
## PC1                  0.085308630 0.002596806  32.8513650 1.228936e-142
## PC2                 -0.010561408 0.002377986  -4.4413250  1.042851e-05
## PC3                 -0.039634708 0.003254975 -12.1766554  5.279649e-31
## PC4                 -0.024003798 0.002822715  -8.5037985  1.156974e-16
## PC5                  0.005798362 0.003618046   1.6026227  1.094810e-01
## PC6                 -0.034961360 0.003888694  -8.9905143  2.397987e-18
## PC7                 -0.035478855 0.004030116  -8.8034334  1.085165e-17
## PC8                  0.023691189 0.003914467   6.0522137  2.357286e-09
## PC9                  0.012900529 0.003831303   3.3671385  8.021788e-04
## MSZoningRH          -0.155281324 0.048405796  -3.2079077  1.399594e-03
## MSZoningRL          -0.076195541 0.023870196  -3.1920785  1.477397e-03
## MSZoningRM          -0.112040672 0.025882001  -4.3289029  1.723238e-05
## LandContourHLS       0.068338692 0.031208673   2.1897340  2.888158e-02
## LandContourLow       0.027166217 0.035101247   0.7739388  4.392352e-01
## LandContourLvl      -0.013398247 0.020677198  -0.6479721  5.172211e-01
## LotConfigCulDSac     0.037781966 0.016887184   2.2373159  2.558795e-02
## LotConfigInside      0.002762348 0.010406987   0.2654321  7.907568e-01
## LotConfigother      -0.034009501 0.021333271  -1.5942000  1.113545e-01
## NeighborhoodNAmes   -0.028770723 0.017004357  -1.6919619  9.111004e-02
## NeighborhoodOldTown -0.075577450 0.023044310  -3.2796577  1.092170e-03
## Neighborhoodother   -0.046228499 0.020067169  -2.3036881  2.153998e-02
## NeighborhoodOther   -0.017954301 0.013433591  -1.3365228  1.818242e-01
## Condition1Feedr      0.025370534 0.029429948   0.8620652  3.889547e-01
## Condition1Norm       0.067072946 0.024612122   2.7251996  6.591070e-03
## Condition1RR         0.033490932 0.032848006   1.0195727  3.082927e-01
## Condition1Other      0.039642473 0.040902315   0.9691988  3.327896e-01
## RoofStyleHip         0.004294792 0.010706586   0.4011355  6.884460e-01
## RoofStyleother       0.116084017 0.029992467   3.8704390  1.190733e-04
## Exterior1stMetalSd   0.031902089 0.014203201   2.2461197  2.501564e-02
## Exterior1stVinylSd   0.034807711 0.012779963   2.7236160  6.622406e-03
## Exterior1stWd Sdng  -0.001866158 0.015509886  -0.1203206  9.042646e-01
```

```
## Exterior1stOther     0.029850129 0.013134462   2.2726571  2.335720e-02
## ExterCondAvg         0.024465959 0.013104165   1.8670369  6.232526e-02
## ExterCondBelowAvg   -0.028206229 0.037612797  -0.7499104  4.535673e-01
## FoundationCBlock    -0.005642087 0.016235609  -0.3475131  7.283131e-01
## Foundationother      0.016893468 0.034969237   0.4830951  6.291832e-01
## FoundationPConc      0.039165182 0.018822234   2.0807935  3.782544e-02
## Heatingother         0.056545563 0.031911300   1.7719605  7.684755e-02
## CentralAirY          0.082040972 0.023994006   3.4192277  6.653749e-04
## ElectricalFuseF     -0.106173289 0.035538965  -2.9875178  2.913481e-03
## ElectricalSBrkr     -0.029760756 0.016370115  -1.8179931  6.950378e-02
## KitchenQualAvg      -0.029183361 0.011003699  -2.6521409  8.184269e-03
## KitchenQualBelowAvg -0.029761432 0.028444256  -1.0463073  2.957899e-01
## FunctionalMaj2      -0.237697554 0.071641338  -3.3178827  9.551995e-04
## FunctionalMin1       0.062883101 0.049139606   1.2796827  2.010919e-01
## FunctionalMin2       0.080433431 0.047272156   1.7014970  8.930549e-02
## FunctionalMod        0.043346203 0.066547424   0.6513581  5.150348e-01
## FunctionalTyp        0.134738856 0.040649447   3.3146541  9.661205e-04
## PavedDriveP         -0.004099131 0.028215727  -0.1452782  8.845341e-01
## PavedDriveY          0.036519383 0.018330534   1.9922705  4.673961e-02
```

p-values:

```
##         value
## 1.929377e-309
```

Adjusted R-squared:

```
## [1] 0.895541
```

AIC:

```
## [1] -1235.314
```

BIC:

```
## [1] -996.2623
```

VIF:

```
##                 GVIF Df GVIF^(1/(2*Df))
## PC1          3.349133  1        1.830064
## PC2          1.405989  1        1.185744
## PC3          2.200222  1        1.483315
## PC4          1.152293  1        1.073449
## PC5          1.291353  1        1.136377
## PC6          1.375471  1        1.172805
## PC7          1.302623  1        1.141325
## PC8          1.192701  1        1.092109
## PC9          1.101191  1        1.049377
## MSZoning     2.968290  3        1.198812
## LandContour  1.454248  3        1.064404
```

```
## LotConfig    1.357456  3        1.052255
## Neighborhood 4.781496  4        1.216033
## Condition1   1.677851  4        1.066827
## RoofStyle    1.310911  2        1.070024
## Exterior1st  3.351989  4        1.163223
## ExterCond    1.406588  2        1.089035
## Foundation   4.765501  3        1.297233
## Heating      1.670439  1        1.292455
## CentralAir   1.946796  1        1.395276
## Electrical   1.670954  2        1.136949
## KitchenQual  2.771675  2        1.290285
## Functional   1.655958  5        1.051732
## PavedDrive   1.582856  2        1.121658
```

RMSE:

```
## [1] 0.09705064
```

- So, we can say that using PCA followed by stepAIC the OLS regression model is better as compared to the other OLS model built based on their `adjusted R-squared` value.
- There is also no multicollinearity found in the model as the VIF values are less than 10.

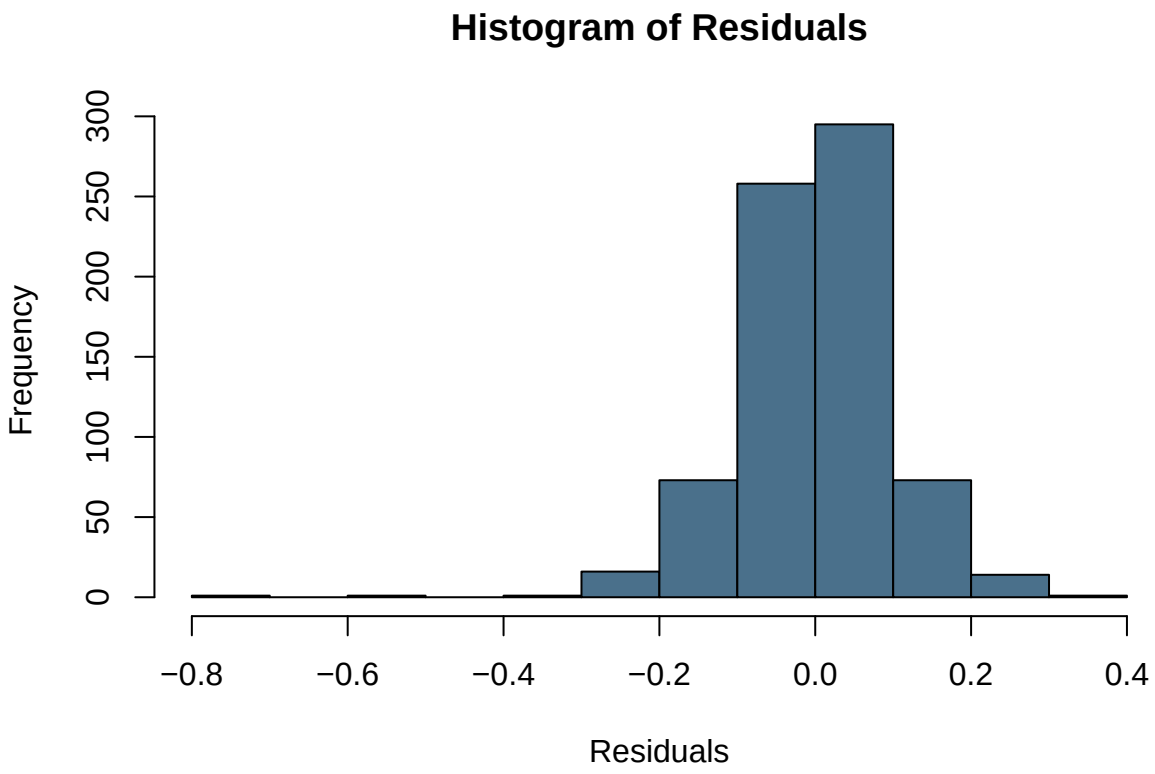
Model	Notes	Hyperparameters	RMSE	Rsquared
OLS	lm + 2-way interactions	N/A	0.0970506	0.895541

ii. Complete analysis of the residuals

A linear regression model is considered fit if the below assumptions are met:

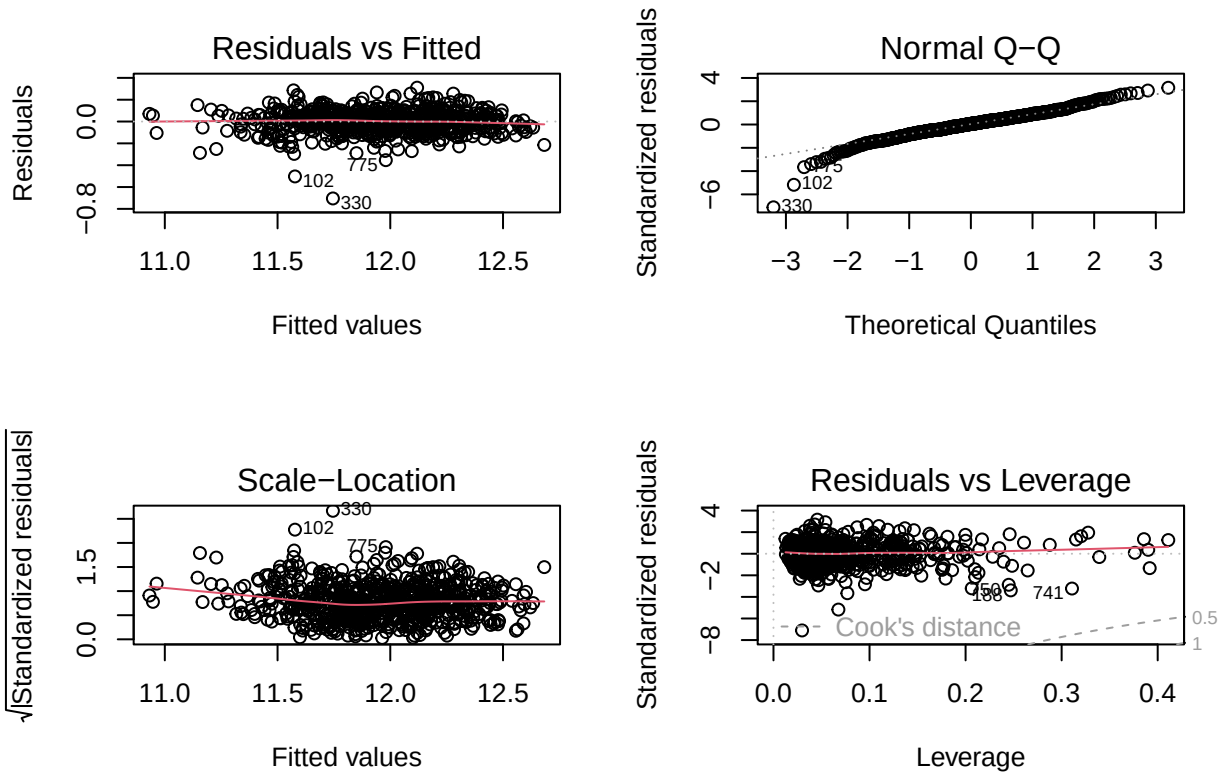
- **Residuals should follow normal distribution**
- **There should be no heteroscedasticity**
- **There should be no multicollinearity**

```
hist(ols.mdl2$residuals,
     col = 'skyblue4',
     main = 'Histogram of Residuals',
     xlab = 'Residuals')
```



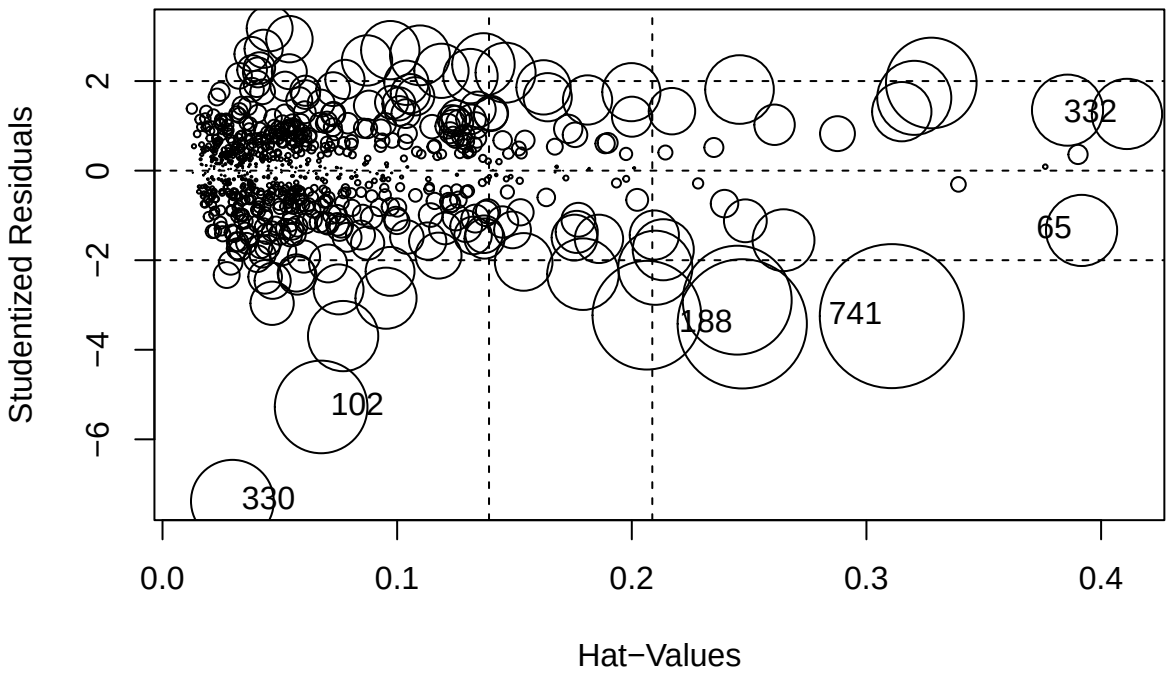
We can see that the residuals are `normally distributed` with a little longer left tail, maybe due to presence of outliers.

```
par(mfrow=c(2,2)) #combining multiple plots together
plot(ols.mdl2)
```



- From the *Residuals vs Fitted* plot, we can see there are points above and below the 0 line.
- There is also a pattern seen like a `very slight curvature pattern` towards the end which indicates that there maybe a systematic lack of fit.
- The mean of residuals is almost zero which implies there is no biasing involved.
- From the *Normal Q-Q* plot, we can see that most of the points are `very close to the dotted line`, indicating that the residuals follow a normal distribution, except some points which might be outliers which maybe affecting the regression line fit of data.
- Here the *Scale-Location* plot suggests that the red line is roughly horizontal across the plot and the spread of magnitude looks unequal, at some fitted values there are more residuals as compared to other like the ones in between 11.5 and and 12.5, indicating some heteroskedasticity.
- From the *Residuals vs Leverage* plot, we can see that there are no influential points close to the Cook's distance line in our regression model. We need to check `influencePlot` to see if we are missing any leverage.

```
influencePlot(ols.mdl2)
```




```
##       StudRes        Hat      CookD
## 65  -1.335443 0.39168503 0.02249004
## 102 -5.276791 0.06759709 0.03808267
## 188 -3.421388 0.24706397 0.07415174
## 330 -7.382532 0.02977788 0.03041331
## 332  1.266114 0.41102149 0.02191578
## 741 -3.249003 0.31080036 0.09204992
```

- We can now see some high influential points for the fitted values.

```
#ncv Test
ncvTest(ols.mdl2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 12.50368, Df = 1, p = 0.00040615
```

Since `p-value` is less than significance level (α) of `0.05`, that means we **`reject the null hypothesis`** of constant error variance which indicates heteroscedasticity.

```
VIF(ols.mdl2)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## PC1          3.349133  1        1.830064
```

```
## PC2          1.405989  1        1.185744
## PC3          2.200222  1        1.483315
## PC4          1.152293  1        1.073449
## PC5          1.291353  1        1.136377
## PC6          1.375471  1        1.172805
## PC7          1.302623  1        1.141325
## PC8          1.192701  1        1.092109
## PC9          1.101191  1        1.049377
## MSZoning     2.968290  3        1.198812
## LandContour  1.454248  3        1.064404
## LotConfig    1.357456  3        1.052255
## Neighborhood 4.781496  4        1.216033
## Condition1   1.677851  4        1.066827
## RoofStyle    1.310911  2        1.070024
## Exterior1st  3.351989  4        1.163223
## ExterCond    1.406588  2        1.089035
## Foundation   4.765501  3        1.297233
## Heating      1.670439  1        1.292455
## CentralAir   1.946796  1        1.395276
## Electrical   1.670954  2        1.136949
## KitchenQual  2.771675  2        1.290285
## Functional   1.655958  5        1.051732
## PavedDrive   1.582856  2        1.121658
```

Generally, VIF values which are greater than 5 or 7 are the cause of multicollinearity which we do not see in our model.

Improving the current model:

- To improve our model, we need to remove some influential observations from our model and then fit the regression model to the data.
- We can re-build the model with new predictors.
- We can also perform variable transformation such as Box-Cox or use better evolved models like SVM, PCR etc., and see how it works.

1 (b) - PLS Model

Model Setup

- Using the whole data set after PMM imputation and factor level collapsing without omitting any outliers
- Using the predictors - `GarageArea`, `GarageCars`, `TotRmsAbvGrd`, `FullBath`, `GrLivArea`, `X1stFlrSF`, `TotalBsmtSF`, `OverallQual` which has strong correlations with response variable - `SalePrice`

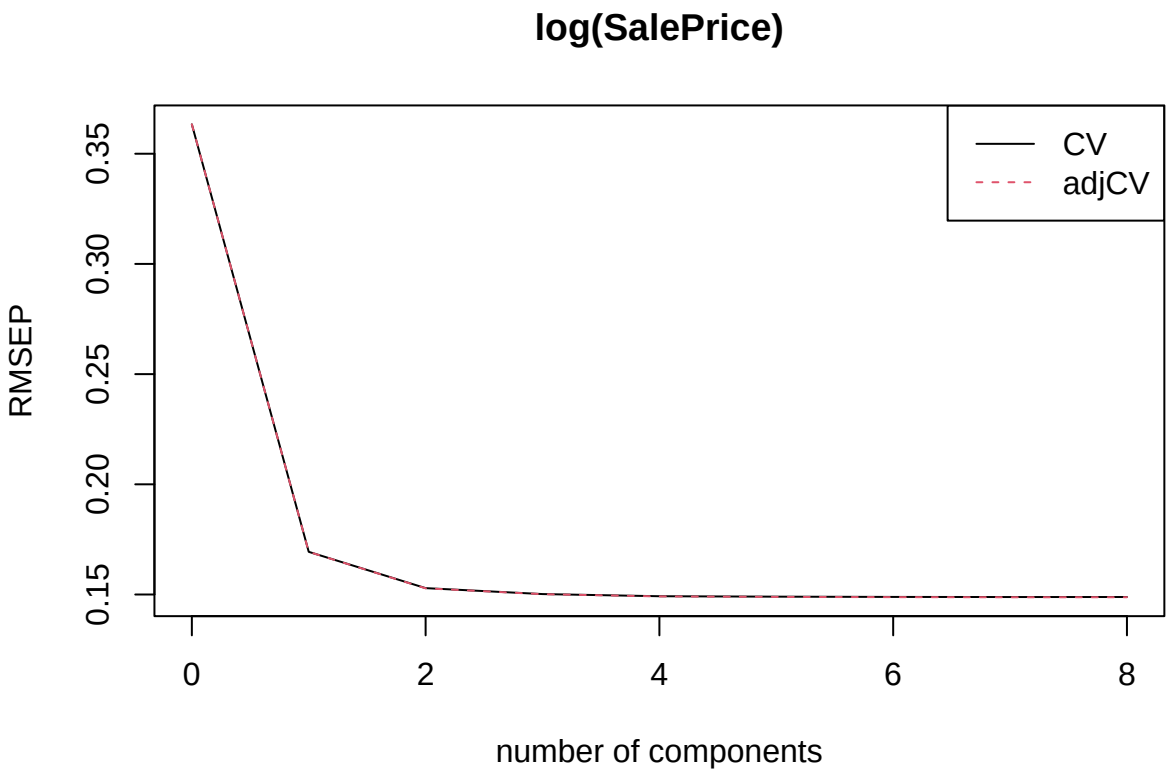
```
#creating a PLS model to predict the log of the sale price
#using 5-fold CV

pls.model <- plsr(log(SalePrice) ~ GarageArea + GarageCars + TotRmsAbvGrd
              + FullBath + GrLivArea + X1stFlrSF + TotalBsmtSF + OverallQual,
              data=hd.Cleaned, scale=TRUE, validation='CV', k=5)
```

- Hyperparameter tuning to determine the number of PLS components with RMSE as the error metric

```
#report chart
summary(pls.model)
```

```
## Data:    X dimension: 1000 8
##  Y dimension: 1000 1
## Fit method: kernelpls
## Number of components considered: 8
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV          0.3633   0.1694   0.1529   0.1502   0.1492   0.1490   0.1489
## adjCV       0.3633   0.1693   0.1528   0.1501   0.1491   0.1489   0.1488
##        7 comps  8 comps
## CV      0.1489   0.1489
## adjCV   0.1488   0.1488
##
## TRAINING: % variance explained
##                 1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X                 54.34    62.66    74.93    79.61    83.32    95.82    97.73
## log(SalePrice)    78.36    82.58    83.18    83.49    83.60    83.60    83.60
##                 8 comps
## X                 100.0
## log(SalePrice)     83.6
```

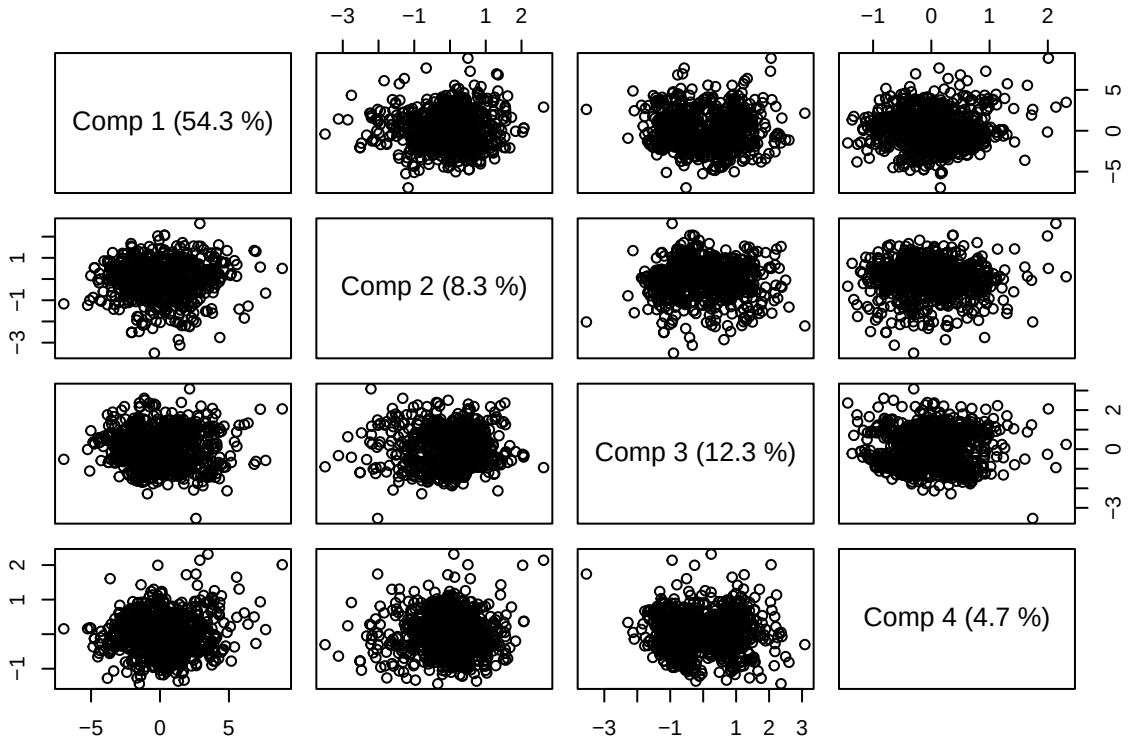
```
plot(RMSEP(pls.model),legendpos="topright")
```



- From the table, we can see that if we use `6` PLS components only in our model, the RMSE drops to `0.1486` and after that even if we keep adding components the RMSE still is the same.
- Though we are eyeballing the CV component, but from the plot we can see that fitting `4` PLS components is enough because even if we are adding 2 more components there is not much difference in the CV component.
- Using the final model with `four PLS components` to make predictions

Fit the Model

```
final.pls <- plsr(log(SalePrice) ~ GarageArea + GarageCars + TotRmsAbvGrd
              + FullBath + GrLivArea + X1stFlrSF + TotalBsmtSF + OverallQual,4,
              data=hd.Cleaned, scale=TRUE, validation='CV', k=5)

plot(final.pls, plottype = "scores", comps = 1:4)
```



- From the above plot, we can see that by using only four PLS components we can describe about 80% of the variation in the response variable.
- Metric Calculations:

Model	Notes	Hyperparameters	RMSE	Rsquared
PLS	pls	ncomp = 4	0.1474771	0.0218368

- If we now compare between our preferred OLS model and PLS model on basis of RMSE values, we can see that PLS model's efficiency is much higher.
- RMSE for chosen OLS model was `ols.mdl2.rsme` whereas for PLS model is `0.1475`.
- But we see that the adjusted R-squared value for PLS model has significantly reduced to about `2%`.
- We know that adjusted R-squared identifies the percentage of variance in the response that is explained by the predictors which PCA handles in a better way as PCA finds the composite variables of predictors that maximally explain the variability of the data, whereas PLS finds the composite variables of predictors that are most predictive of the response variable. So maybe that's why we have a less adjusted R-squared whereas a better RMSE value.

1 (c) - LASSO Model

Model Setup

- We first setup our cross-validation strategy
- Then create a dataframe with PMM imputed values, and only whole columns without NA. Does not omit outliers
- Then we train the model using `glmnet` which actually fits the elastic net

```
ctrl <- trainControl(method  = "repeatedcv",
                     number  = 5, # 5 fold cross validation
                     repeats = 2  # 2 repeats
                     )

# The data (PMM imputed values, and only whole columns without NA. Does not omit outliers)
df.lasso <- cbind(SalePrice = hd.numericClean$SalePrice,
                  hd.numericClean, hd.factorClean)
```

Fit the Model

```
# Train and tune the SVM
fit.lasso <- train(data = df.lasso,
                   log(SalePrice) ~ .,
                   method     = "glmnet",          # Elastic net
                   preProc    = c("center","scale"), # Center and scale data
                   tuneLength = 10,  #10 values of alpha and 10 lamda values for each
                   trControl  = ctrl)
```

- The variables with non-zero coefficients of the final model:

```
lasso.coeff <- drop(coef(fit.lasso$finalModel, fit.lasso$bestTune$lambda))

lasso.coeff[lasso.coeff != 0]
```

```
##         (Intercept)          MSSubClass         LotFrontage             LotArea
##        1.200247e+01       -4.622764e-04        3.690484e-03        2.268817e-02
##         OverallQual         OverallCond          MasVnrArea          BsmtFinSF1
##        8.366135e-02        4.663499e-02        4.635555e-04        3.098434e-02
##          BsmtFinSF2         TotalBsmtSF        LowQualFinSF           GrLivArea
##        2.158533e-03        4.325055e-02       -1.526556e-04        1.290654e-01
##        BsmtFullBath            HalfBath        BedroomAbvGr        KitchenAbvGr
##        9.563906e-03        1.821680e-07       -9.628194e-04       -8.480775e-03
##          Fireplaces          GarageCars          GarageArea          WoodDeckSF
##        2.402993e-02        2.426696e-02        2.038675e-02        5.272743e-03
##         OpenPorchSF          EncPorchSF                 age     ageSinceRemodel
##        6.835049e-03        1.074429e-02       -4.919920e-02       -1.225833e-02
##          MSZoningRH          MSZoningRM         LotShapeIR3         LotShapeReg
##       -2.192053e-03       -2.541142e-02       -1.475741e-04       -1.279656e-03
```

```
##      LandContourHLS    LotConfigCulDSac        LandSlopeMod NeighborhoodOldTown
##        3.540319e-03        4.275002e-03        2.338920e-03       -4.495761e-03
##   Neighborhoodother   NeighborhoodOther      Condition1Norm      BldgTypeDuplex
##       -3.233973e-03        1.412591e-03        1.393124e-02       -3.611239e-04
##       BldgTypeTwnhs     HouseStyleOther      RoofStyleother  Exterior1stWd Sdng
##       -7.962517e-03       -2.673045e-03        8.598469e-03       -7.649646e-04
##    Exterior1stOther  Exterior2ndVinylSd        ExterQualAvg   ExterQualBelowAvg
##        1.504308e-03        2.483326e-03       -4.265986e-03       -4.827393e-03
##        ExterCondAvg   ExterCondBelowAvg     FoundationPConc        HeatingQCAvg
##        8.808943e-04       -1.141435e-03        1.716070e-02       -5.822882e-03
##   HeatingQCBelowAvg         CentralAirY      KitchenQualAvg KitchenQualBelowAvg
##       -1.085355e-04        9.798714e-03       -6.964044e-03       -2.545173e-03
##      FunctionalMaj2       FunctionalTyp         PavedDriveY
##       -1.091277e-02        1.196188e-02        5.845931e-03
```

Model	Notes	Hyperparameters	RMSE	Rsquared
Lasso	caret and elasticnet	Alpha = 0.9 , Lambda = 0.00385954380548551	0.1008889	0.923388

1 (d) - Model Variants

1 (d, i) - PCR Model

Model Setup

- Uses `numeric` data for Principal Component Analysis
  - Data includes outliers
  - Chose number of PC's that explain 75% of the variation. This is just a general judgement call to keep the number of principal components low.
- Then appends the `factor` columns *without* ***`NULL`*** *values* and `SalePrice` to the data
- Finally, uses `stepAIC()` to best model data
- See interpretation at end

```
## [1] "There are 12 principal components that explain up to 75% of the variation in the data"
```

Join on the `factor` data and `SalePrice`

```
df.pcr <- cbind(SalePrice = hd.numericClean$SalePrice, chosenPCs, hd.factorClean)
```

Fit the Model

- Linear model containing:
  - Principal components explaining 75% of variation in `numeric` data
  - Non-null `factor` data
  - *Predicted variable:* `log(SalePrice)`
- Then use `stepAIC()` to identify which variables are actually important for model

```
# Fit data using PC's, non-null factors
fit.pcr <- lm(log(SalePrice) ~ ., data = df.pcr)

# Reduce to only important variables
fit.pcrReduced <- stepAIC(fit.pcr, direction="both")
```

Model	Notes	Hyperparameters	RMSE	Rsquared
PCR	lm, prcomp, and stepAIC	N/A	0.1014045	0.917838

View results of step `AIC` model

```
summary(fit.pcrReduced)
```

```
##
## Call:
## lm(formula = log(SalePrice) ~ PC1 + PC3 + PC4 + PC5 + PC7 + PC8 +
##     PC9 + PC11 + PC12 + MSZoning + LandContour + LotConfig +
```

```
##     Condition1 + HouseStyle + RoofStyle + Exterior1st + ExterQual +
##     ExterCond + Foundation + Heating + CentralAir + KitchenQual +
##     Functional + PavedDrive, data = df.pcr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68636 -0.05897  0.00307  0.06656  0.30565
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          11.774996   0.054302 216.844  < 2e-16 ***
## PC1                   0.098491   0.002390  41.203  < 2e-16 ***
## PC3                  -0.053292   0.003105 -17.166  < 2e-16 ***
## PC4                  -0.022304   0.002873  -7.763 2.15e-14 ***
## PC5                  -0.039254   0.003331 -11.783  < 2e-16 ***
## PC7                   0.032241   0.003489   9.241  < 2e-16 ***
## PC8                  -0.006508   0.003291  -1.977 0.048292 *
## PC9                   0.010279   0.003409   3.015 0.002635 **
## PC11                 -0.006684   0.003518  -1.900 0.057751 .
## PC12                  0.019116   0.003626   5.273 1.67e-07 ***
## MSZoningRH           -0.069956   0.039876  -1.754 0.079697 .
## MSZoningRL           -0.032701   0.020118  -1.625 0.104392
## MSZoningRM           -0.113193   0.021847  -5.181 2.69e-07 ***
## LandContourHLS        0.070145   0.026706   2.627 0.008764 **
## LandContourLow       -0.001503   0.028375  -0.053 0.957763
## LandContourLvl       -0.018397   0.018162  -1.013 0.311365
## LotConfigCulDSac      0.043649   0.015168   2.878 0.004096 **
## LotConfigInside       0.002382   0.009120   0.261 0.794041
## LotConfigother       -0.009613   0.019221  -0.500 0.617106
## Condition1Feedr       0.049685   0.024749   2.008 0.044971 *
## Condition1Norm        0.092921   0.020406   4.554 5.96e-06 ***
## Condition1RR          0.052281   0.029225   1.789 0.073950 .
## Condition1Other       0.025893   0.031177   0.831 0.406463
## HouseStyle1Story     -0.077291   0.012992  -5.949 3.79e-09 ***
## HouseStyle2Story     -0.008212   0.014619  -0.562 0.574420
## HouseStyleSLvl       -0.033715   0.019857  -1.698 0.089856 .
## HouseStyleOther      -0.057367   0.019180  -2.991 0.002853 **
## RoofStyleHip          0.015028   0.009350   1.607 0.108338
## RoofStyleother        0.096958   0.024652   3.933 9.00e-05 ***
## Exterior1stMetalSd    0.029047   0.012581   2.309 0.021166 *
## Exterior1stVinylSd    0.023898   0.011334   2.108 0.035250 *
## Exterior1stWd Sdng   -0.003500   0.013312  -0.263 0.792662
## Exterior1stOther      0.032190   0.011555   2.786 0.005448 **
## ExterQualAvg         -0.037847   0.011573  -3.270 0.001113 **
## ExterQualBelowAvg    -0.097741   0.046764  -2.090 0.036876 *
## ExterCondAvg          0.023658   0.011646   2.031 0.042491 *
## ExterCondBelowAvg     0.009905   0.032326   0.306 0.759367
## FoundationCBlock      0.005254   0.014254   0.369 0.712526
## Foundationother       0.029435   0.025491   1.155 0.248483
## FoundationPConc       0.053354   0.016451   3.243 0.001223 **
## Heatingother          0.039708   0.025314   1.569 0.117073
## CentralAirY           0.064069   0.018388   3.484 0.000516 ***
## KitchenQualAvg       -0.018541   0.010437  -1.776 0.075974 .
## KitchenQualBelowAvg  -0.037692   0.025772  -1.463 0.143929
```

```
## FunctionalMaj2      -0.198090   0.061648  -3.213 0.001357 **
## FunctionalMin1       0.038981   0.038864   1.003 0.316106
## FunctionalMin2       0.034726   0.038020   0.913 0.361294
## FunctionalMod        0.018186   0.044960   0.405 0.685936
## FunctionalTyp        0.107790   0.031839   3.385 0.000740 ***
## PavedDriveP         -0.002384   0.025255  -0.094 0.924801
## PavedDriveY          0.050592   0.016186   3.126 0.001828 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1041 on 949 degrees of freedom
## Multiple R-squared:  0.922,  Adjusted R-squared:  0.9178
## F-statistic: 224.2 on 50 and 949 DF,  p-value: < 2.2e-16
```

View and Interpret Results

Please note all interpretations below are approximate, given the `stepAIC()` uses stochastic modeling.

Model performance evaluation:

- See that around 28 of the variables cannot be explained by random chance, with a probability of 90% or more (see significance codes above)
- Standard errors range from ± 1-5%, with average around 2%. Larger values may indicate higher uncertainty of the estimated coefficients.
- This model explains around 92% of the variation in the `log(SalePrice)`. See Adjusted R-Squared for reference.
- Note this model may exhibit selection bias, since the data excludes factor data with null values in the variable.
- This model would likely doe well for prediction of `log(SalePrice)`, given the small range of standard errors, high adjusted R squared, and number of significant variables. This model would obviously not do well for inference, given we are using principal components that mask the numeric data.

Practical significance evaluation:

- The principal components contribute positively about 20% of the sale price of the home
- Residential Medium Density (`MSZoningRM`) reduces the home price by around 12%, with a standard error of around 2%.
- If the exterior quality is below average (`ExterQualBelowAvg`), it reduces the home price by around 12%, with a standard error of around 5%.
- If the functionality of the home has 2 major deductions (`FunctionalMaj2`), it reduces the home price by around 20%, with a standard error of around 6%. While having typical functionality (`FunctionalTyp`) increases the home sale price by nearly 10%, with a standard error of 3%.
- See other coefficients of the data for other variables.

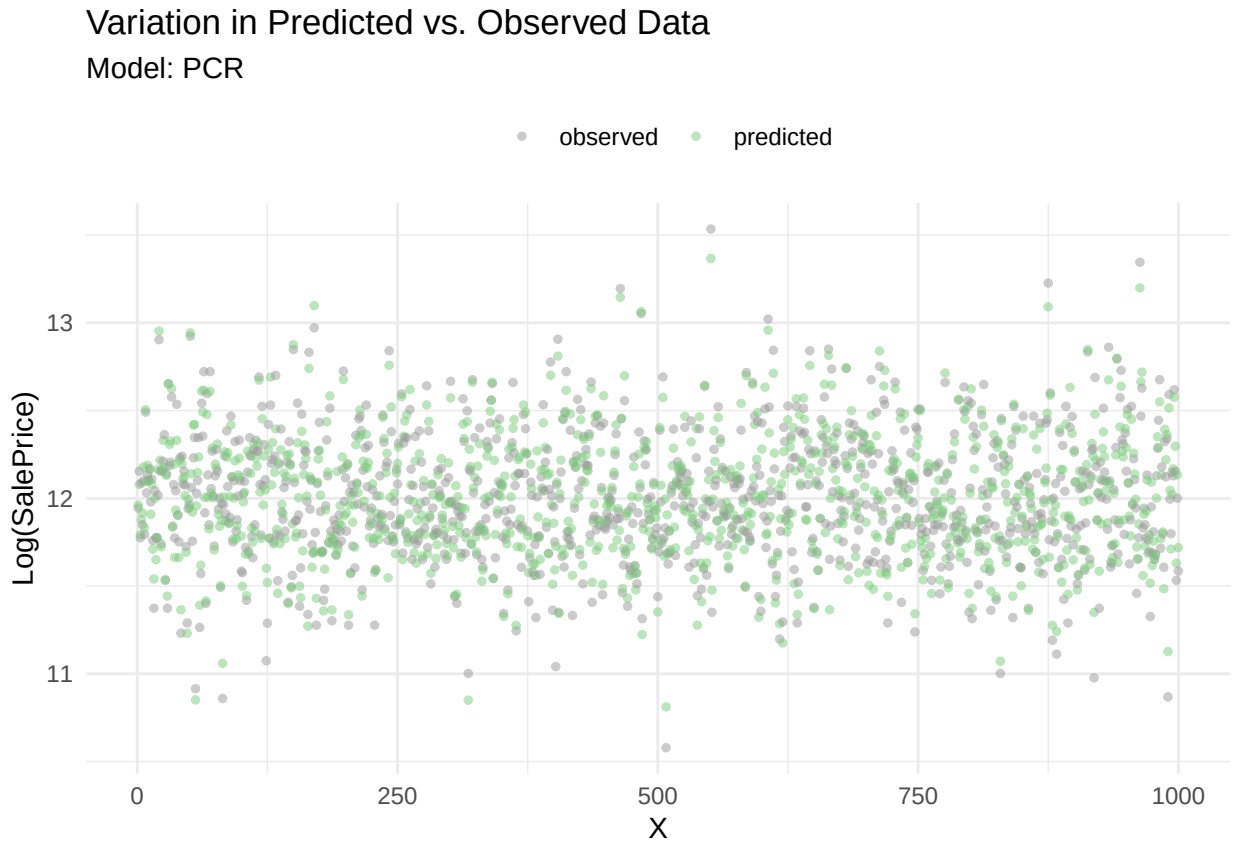
View Predicted vs. Actuals

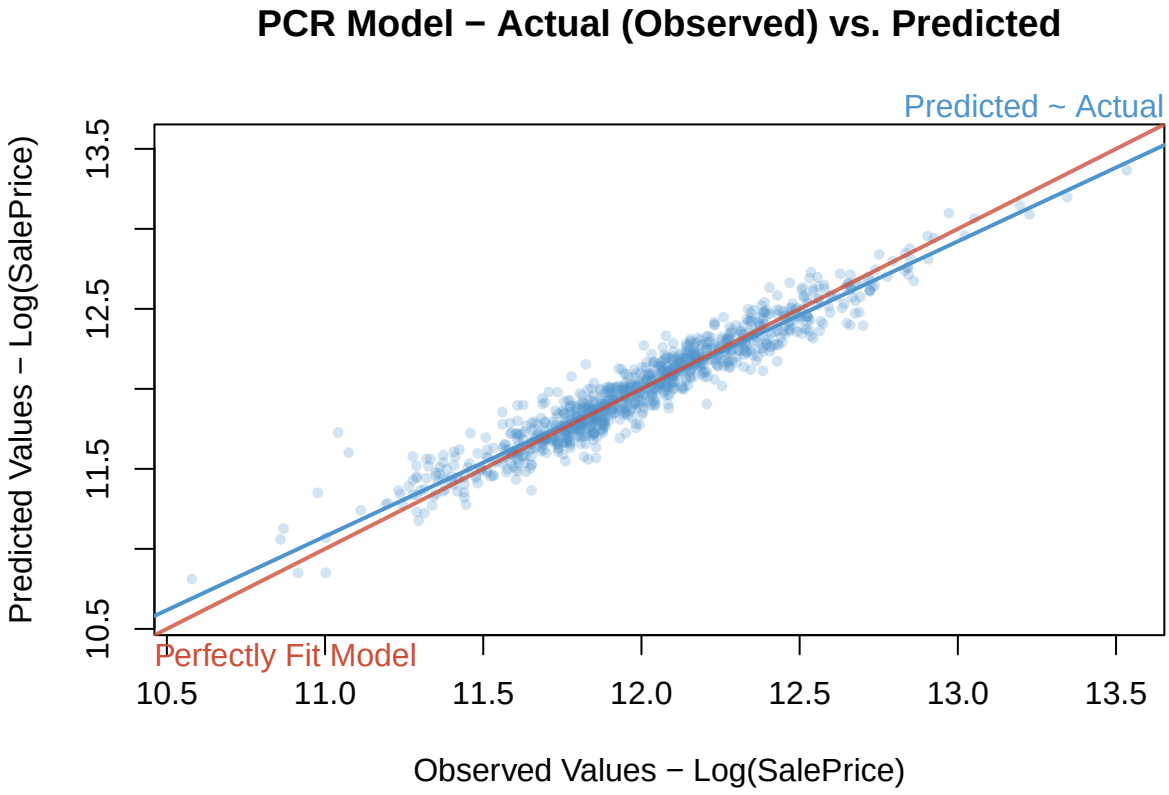
Note that the Function `predictedVsObserved()` created to compare predicted vs. observed values from the model. Uses `ggplot2` and model output to display the following. *See interpretation below.*

View results of the PCR Model

- See that the variation in the data is very closely resembled actual by changes in independent variables
- Implication? This model fits its own data well, but what is not know if it can predict out of sample data.
- Note that it the data (blue) deviates slightly from perfect line model (red), indicating that the model is slightly skewed from predicted and actual data.

```
# How do the predicted vs. Actuals Compare?
predictedVsObserved(observed  = log(df.pcr$SalePrice),
                    predicted = predict(fit.pcrReduced),
                    modelName = 'PCR')
```




PCR Model − Actual (Observed) vs. Predicted
Predicted ~ Actual
Perfectly Fit Model
Predicted Values − Log(SalePrice)
Observed Values − Log(SalePrice)
10.5
11.5
12.5
13.5
10.5
11.0
11.5
12.0
12.5
13.0
13.5

1 (d, ii) - SVM Model

Model Setup

```
ctrl <- trainControl(method  = "repeatedcv",
                     number  = 5, # 5 fold cross validation
                     repeats = 2  # 2 repeats
                     )

# The data (PMM imputed values, and only whole columns without NA. Does not omit outliers)
df.svm <- cbind(SalePrice = hd.numericClean$SalePrice,
                hd.numericClean, hd.factorClean)
```

Fit the Model

- *Predicted variable:* `log(SalePrice)`
- *Dependent variables*: non-null factor data (collapsed if over 4 unique values), and all numeric data (`pmm` imputed if needed). Includes outliers

```
# Train and tune the SVM
fit.svm <- train(data = df.svm,
                 log(SalePrice) ~ .,
                 method     = "svmRadial",         # Radial kernel
                 tuneLength = 9,                   # 9 values of the cost function
                 preProc    = c("center","scale"), # Center and scale data
                 trControl  = ctrl)
```

View and Interpret Results

- Note all numbers mentioned below are approximate
- See that the R Squared of the model is around 0.86, and RMSE is 0.14
- See that the model predicts the data well.
- Also, note that the model predicts the data with less error than the linear model. See this from the RMSE or scatter plot of predicted values.

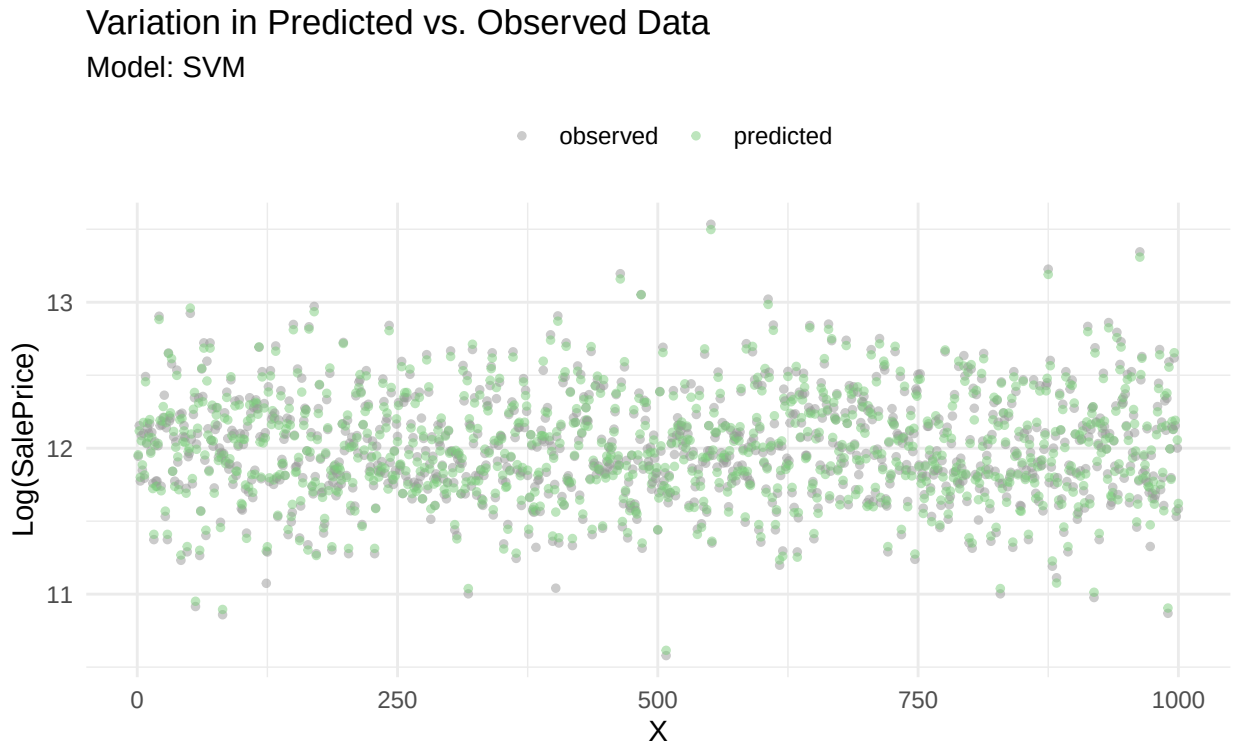
Model	Notes	Hyperparameters	RMSE	Rsquared
SVM	caret and svmRadial	C = 4 , Epsilon = 0.1	0.1364775	0.8622847

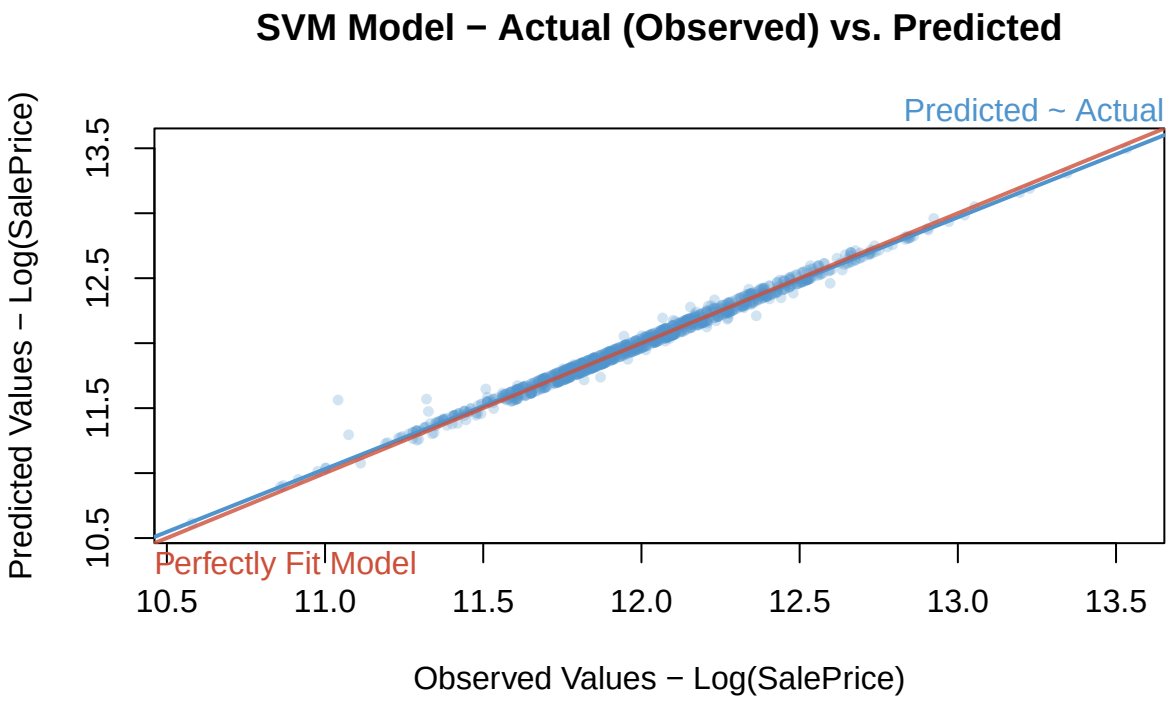
```
# Final model?
fit.svm$finalModel
```

```
## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 4
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.00708163906604968
##
```

```
## Number of Support Vectors : 666
##
## Objective Function Value : -164.9783
## Training error : 0.012595
```

```
# How do the predicted vs. Actuals Compare?
predictedVsObserved(observed  = log(df.svm$SalePrice),
                    predicted = predict(fit.svm, df.svm),
                    modelName = 'SVM')
```




SVM Model − Actual (Observed) vs. Predicted
Predicted ~ Actual
Perfectly Fit Model
Predicted Values − Log(SalePrice)
Observed Values − Log(SalePrice)
10.5
11.5
12.5
13.5
10.5
11.0
11.5
12.0
12.5
13.0
13.5

1 (d, iii) - MARS Model

Fit the Model

- *Predicted variable:* `log(SalePrice)`
- *Dependent variables*: non-null factor data (collapsed if over 4 unique values), and all numeric data (`pmm` imputed if needed). Includes outliers

```
# Train and tune the MARS model
fit.mars <- train(data = df.svm, # note this is fine since data is the same for this model
                  log(SalePrice) ~ .,
                  method     = "earth",             # Radial kernel
                  tuneLength = 9,                   # 9 values of the cost function
                  preProc    = c("center","scale"), # Center and scale data
                  trControl  = ctrl
                 )
```

```
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =
## 10, : These variables have zero variances: PoolArea
```

Model	Notes	Hyperparameters	RMSE	Rsquared
MARS	caret and earth	Degree = 1 , nprune = 17	0.1077457	0.9113543

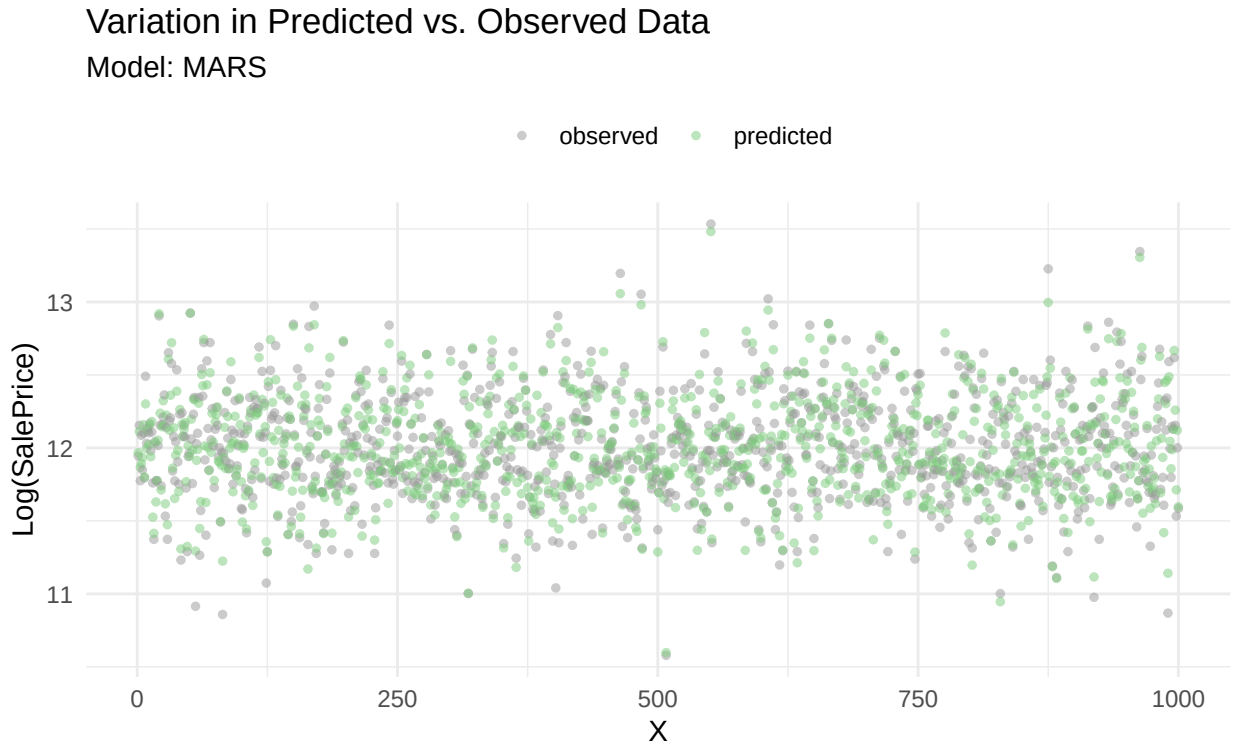
View and Interpret Results

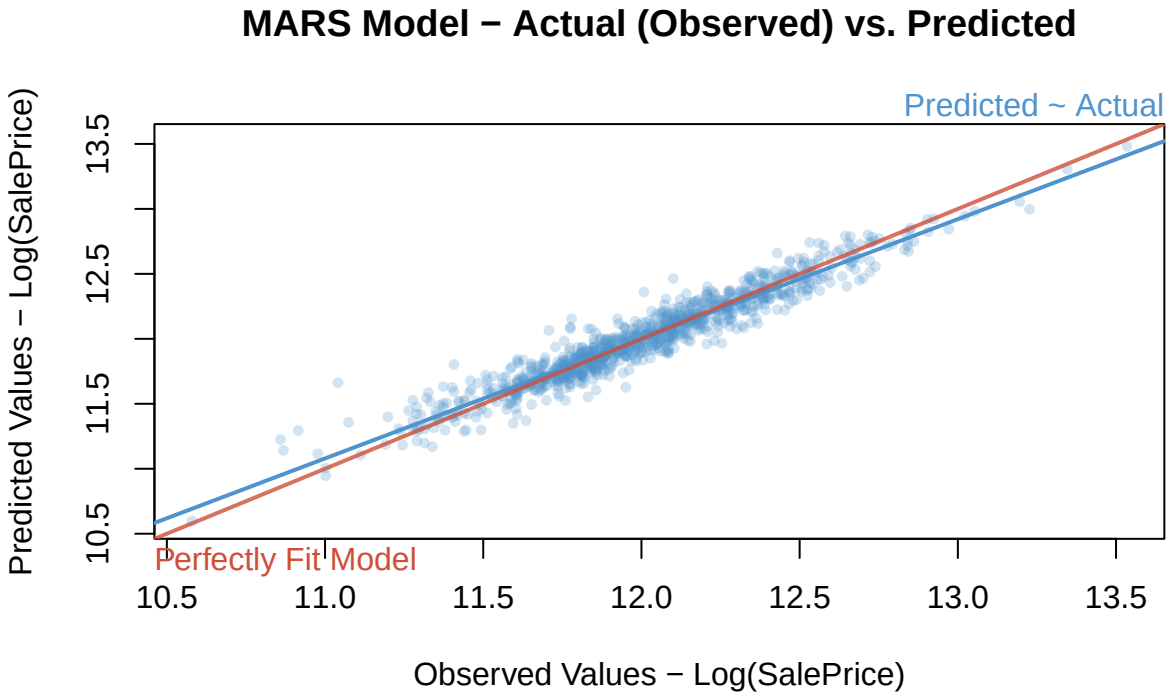
- See that the model overall performs very well, and in fact performs similarly to the PCR model (in terms of RMSE and Adjusted R Squared).
- Again, unsure if the model would do well for prediction of out of sample data, but fits this data extremely well.

```
# Final model?
fit.mars$finalModel
```

```
## Selected 17 of 21 terms, and 10 of 94 predictors (nprune=17)
## Termination condition: RSq changed by less than 0.001 at 21 terms
## Importance: GrLivArea, age, OverallQual, TotalBsmtSF, OverallCond, LotArea, ...
## Number of terms at each degree of interaction: 1 16 (additive model)
## GCV 0.011145    RSS 10.42157    GRSq 0.9155756    RSq 0.9208976
```

```
# How do the predicted vs. Actuals Compare?
predicted.mars = fit.mars[["finalModel"]][["fitted.values"]]
colnames(predicted.mars) <- 'predicted'

predictedVsObserved(observed  = log(df.svm$SalePrice),
                    predicted = predicted.mars,
                    modelName = 'MARS')
```


Variation in Predicted vs. Observed Data
Model: MARS
observed
predicted
Log(SalePrice)
13
12
11
0
250
500
750
1000
X


MARS Model − Actual (Observed) vs. Predicted
Predicted ~ Actual
Perfectly Fit Model
Predicted Values − Log(SalePrice)
10.5
11.5
12.5
13.5
10.5
11.0
11.5
12.0
12.5
13.0
13.5
Observed Values − Log(SalePrice)

Summary Table of Model Performance

Model	Notes	Hyperparameters	RMSE	Rsquared
OLS	lm	N/A	0.1312	0.8202
OLS	lm + 2-way interactions	N/A	0.0971	0.8955
PLS	pls	ncomp = 4	0.1475	0.0218
Lasso	caret and elasticnet	Alpha = 0.9 , Lambda = 0.00385954380548551	0.1009	0.9234
PCR	lm, prcomp, and stepAIC	N/A	0.1014	0.9178
SVM	caret and svmRadial	C = 4 , Epsilon = 0.1	0.1365	0.8623
MARS	caret and earth	Degree = 1 , nprune = 17	0.1077	0.9114

References

1. https://rpubs.com/staneaurelius/house_price_prediction
2. https://www.statology.org/partial-least-squares-in-r/
3. https://daviddalpiaz.github.io/r4sl/elastic-net.html