DSA/ISE 5103 Intelligent Data Analytics
Assignment #4

Name: Jiwon Jeon
Date: 10/31/2016

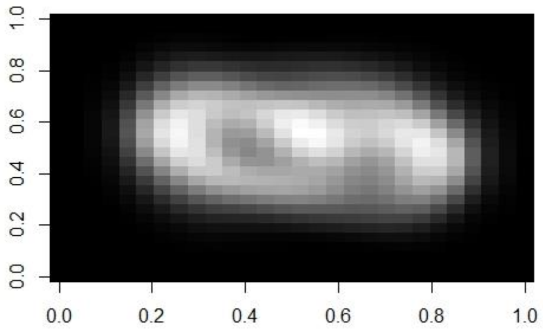
1 Digit Recognition

(a) The eigenvectors of the digit data set is the covariance matrix of its PCA (without scaling).

```
                 PC1           PC2           PC3           PC4           PC5           PC6           PC7
pixel0  2.219274e-20 -5.732181e-19  6.287447e-20 -1.759315e-19  2.794486e-19 -3.717744e-19  7.823462e-19
pixel1  2.081668e-17  1.110223e-16  2.081668e-17  8.326673e-17 -8.326673e-17  1.387779e-16 -5.551115e-16
pixel2 -1.942890e-16  0.000000e+00  4.857226e-17 -4.163336e-17  5.551115e-17  1.387779e-16 -3.053113e-16
pixel3 -1.387779e-16  1.110223e-16  4.336809e-17 -1.110223e-16  2.498002e-16 -1.942890e-16  7.216450e-16
pixel4  5.551115e-17  0.000000e+00 -1.387779e-17 -1.110223e-16 -5.551115e-17  1.665335e-16  1.942890e-16
pixel5  1.110223e-16  1.387779e-16  2.081668e-17 -1.422473e-16  1.110223e-16 -2.220446e-16  6.661338e-16
```

(b) JPEG file of mean digit

As PCA is run by setting center = TRUE, the mean digit is the value of centroids of PCA of training data. The image is constructed as below;



Code for better image (Extra credit)

```
jpeg("meanDigitImp.jpg", width = 2800, height = 2800, res = 500)
image(digitMatrix, col = grey(seq(0,1,length=256)))
```

(c) As the mean image is constructed in Problem 1.(a), all training digits can be estimated (reconstructed) by using the weight of the corresponding datum and the eigenvector at certain dimensions. PCA projections are the weights of each datum. Therefore the images are reconstructed by the following equation;

Reconstructed Image = mean Image + weight of Image · eigenvector of dimension k

The followings are the results;

Image 15-5

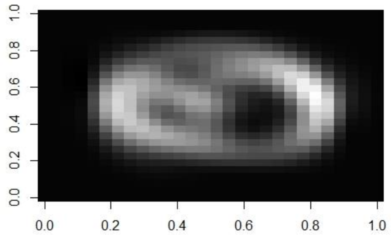


Image 15-20

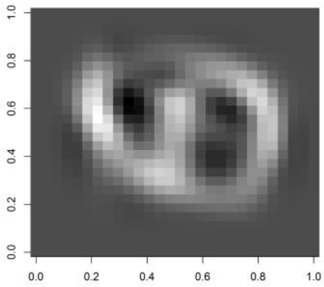


Image 15-100

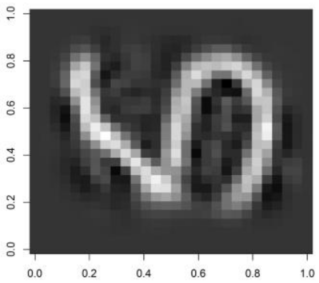


Image 100-5

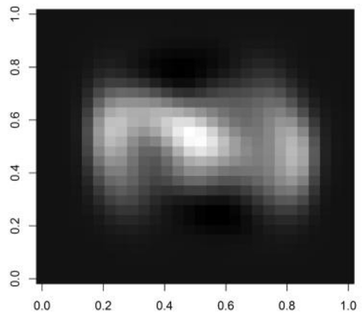


Image 100-20

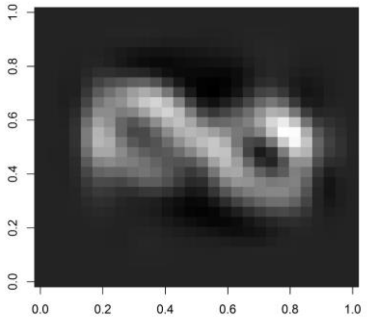
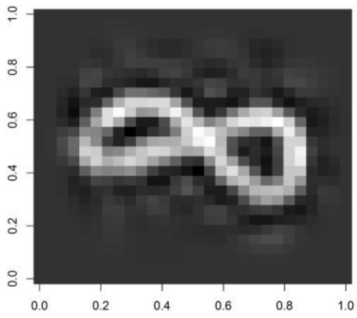


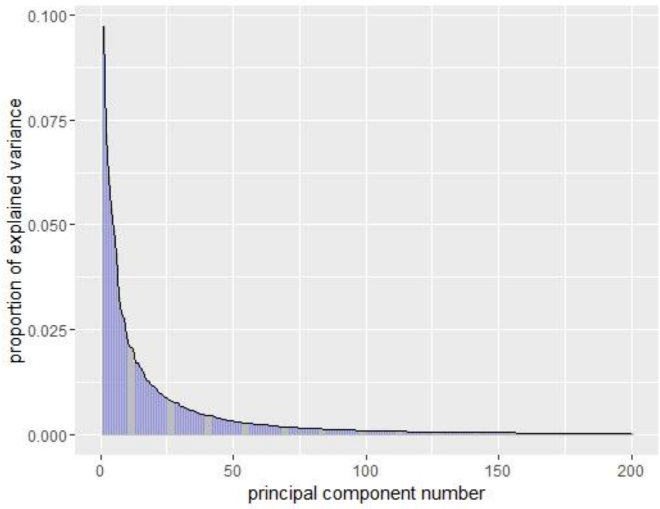
Image 100-100



As the dimension k increases, the image becomes clearer to recognize.

(d) From the PCA and screeplot, the dimension k is chosen to be 50 covering more than 80% of variance;

```
                             PC49      PC50      PC51      PC52      PC53     PC54     PC55     PC56
Standard deviation     107.47404 105.09350 103.91721 102.65543 100.14739 99.51852 98.31831 96.13662
Proportion of Variance   0.00336   0.00322   0.00314   0.00307   0.00292  0.00288  0.00281  0.00269
Cumulative Proportion    0.82247   0.82569   0.82883   0.83190   0.83482  0.83770  0.84052  0.84321
```



The average mahalanobis distances (D^2) from each of seven(7) test data to the transformed training data are calculated as below;

Test datum	1	2	3	4	5	6	7
D^2	90.41244	94.72214	104.41728	77.63699	122.62263	124.38858	151.43048

The mahalanobis distance is different from Eucleadian distance. Therefore, the table does not mean that test datum 4 is the physically closest one to the training digits. It means that datum 4 is aligned with the most dominant variance region.

(e) In order to find the least value of k matching the test digit image 4, 5, and 6 to the training digit images, the mahalanobis distance of each test image was calculated and the label was compared to the label of training digit images. Once the label is matched the least iteration (k) for mahalanobis calculated was determined. The results are;

Test datum	4	5	6
k	4	11	2

2 Predicting house prices

(a) Features with more than 20% of missing data are firstly removed from the housing data as those features significantly affect the analysis and prediction of SalePrice. Out of 73 given features, following six variables were deleted;

- LotFrontage - Alley
- FireplaceQu - PoolQC
- Fence - MiscFeature

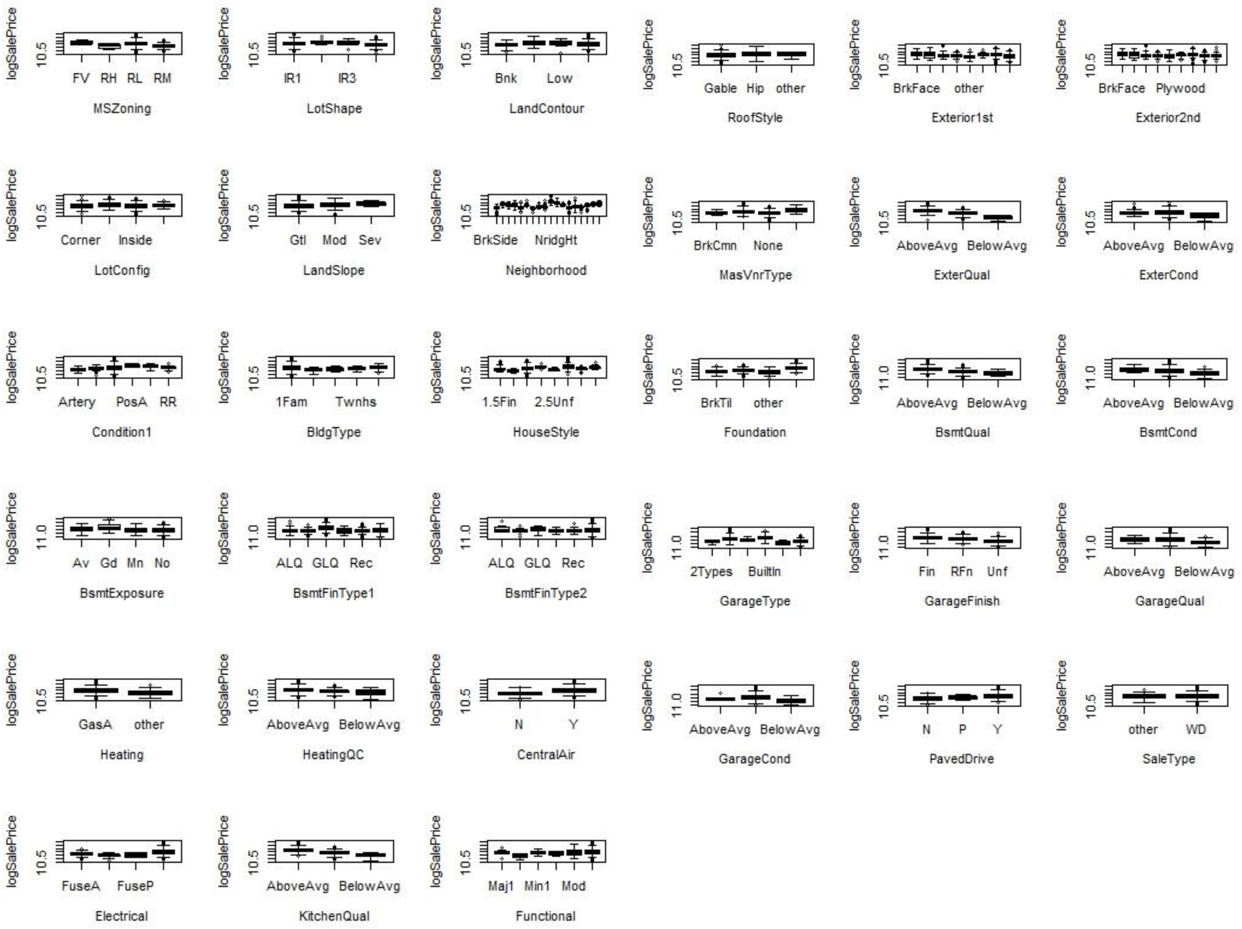
The remaining features were separated into numeric type and factor type and examined for the relevance to SalePrice using correlation matrix and boxplots. 34 variables are numeric type and 33 variables are factor type;

i) 11 numeric features were selected based on the correlation and corrplot;

```
              EncPorchSF     PoolArea      MiscVal       MoSold       YrSold  logSalePrice
MSSubClass  -0.050728962 -0.0095558814  0.01314841 -0.005618783 -0.052303557  -0.04326468
LotArea      0.005689478 -0.0018316646  0.03137070  0.008233596  0.001499983   0.28947856
OverallQual -0.013056195  0.0007167063 -0.06373901  0.046070110  0.010031486   0.80807663
OverallCond  0.109092239  0.0145337866  0.04873526  0.018117255  0.028431476  -0.02991029
YearBuilt   -0.298706401  0.0085615095 -0.07138404 -0.028235404  0.029917418   0.56225432
YearRemodAdd -0.150132016  0.0100888326 -0.03663186 -0.011621848  0.095877573   0.51289597
```

Selected features: OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, X1stFlrSF, GrLivArea, FullBath, TotRmsAbvGrd, Fireplaces, GarageCars, and GarageArea

i) 11 factor features were selected based on the boxplot and their positive/negative relationship with SalePrice;



Selected features: LandSlope, Neighborhood, Condition1, ExterQual, BsmtQual, Heating, HeatingQC, CentralAir, KitchenQual, GarageFinish, and SaleType

(b) OLS Model

i. lm function

Using the selected 22 numeric and factor type variables explained above, OLS model was built by lm function. The followings are the statistical results;

- Adjusted R-squared value: 0.8835
- RMSE: 0.1171
- p-value: <2.2e-16
- AIC: -1148.745
- BIC: -912.976
- vif:

```
                   GVIF Df    GVIF^(1/(2*Df))
OverallQual   3.110298  1       1.763604
YearBuilt    6.254332  1       2.500866
YearRemodAdd  2.372056  1       1.540148
TotalBsmtSF   5.598857  1       2.366190
X1stFlrSF    6.270644  1       2.504125
GrLivArea    6.414034  1       2.532594
FullBath     2.731896  1       1.652845
TotRmsAbvGrd  3.978103  1       1.994518
Fireplaces    1.601176  1       1.265376
GarageCars    4.273168  1       2.067164
GarageArea    3.751542  1       1.936890
LandSlope    1.465700  2       1.100300
Neighborhood 37.025093 17       1.112070
Condition1    1.717971  5       1.055605
ExterQual    3.194361  2       1.336891
BsmtQual     3.629805  2       1.380292
Heating     1.268952  1       1.126478
HeatingQC    1.759281  2       1.151685
CentralAir    1.417728  1       1.190684
KitchenQual   2.922062  2       1.307442
GarageFinish  2.190224  2       1.216528
SaleType     1.052156  1       1.025746
```

- coefficient estimates:

```
     (Intercept)         OverallQual           YearBuilt        YearRemodAdd         TotalBsmtSF 
    6.251482e+00        5.901031e-02        1.108664e-03        1.251991e-03        1.950091e-04 
       X1stFlrSF           GrLivArea            FullBath        TotRmsAbvGrd          Fireplaces 
   -4.236167e-05        2.604333e-04       -3.091330e-02       -5.077716e-04        4.598519e-02 
      GarageCars          GarageArea        LandSlopeMod        LandSlopeSev NeighborhoodClearCr 
    1.276575e-02        2.023510e-04        7.042397e-02        1.368088e-01        5.630396e-02 
NeighborhoodCollgCr NeighborhoodCrawfor NeighborhoodEdwards NeighborhoodGilbert  NeighborhoodIDOTRR 
   -2.807765e-02        1.141078e-01       -7.461123e-02       -1.672405e-02       -4.841873e-02 
NeighborhoodMitchel   NeighborhoodNAmes NeighborhoodNoRidge NeighborhoodNridgHt  NeighborhoodNWAmes 
   -2.843747e-02       -2.876504e-02        1.364760e-02        9.684472e-03       -1.014808e-02 
NeighborhoodOldTown   Neighborhoodother  NeighborhoodSawyer NeighborhoodSawyerW NeighborhoodSomerst 
   -9.837482e-02       -7.249834e-02       -1.766534e-02       -4.020414e-02       -1.308602e-02 
 NeighborhoodTimber      Condition1Feedr       Condition1Norm       Condition1PosA       Condition1PosN 
    1.514507e-02        2.487536e-02        5.916969e-02        6.760042e-02        5.870856e-02 
    Condition1RR         ExterQualAvg   ExterQualBelowAvg         BsmtQualAvg    BsmtQualBelowAvg 
    3.659562e-02        2.494952e-03       -7.338532e-02       -7.647184e-03       -1.030050e-02 
    Heatingother         HeatingQCAvg   HeatingQCBelowAvg          CentralAirY      KitchenQualAvg 
    7.848956e-02       -2.516366e-02       -3.162247e-02        1.549176e-01       -4.824306e-02 
KitchenQualBelowAvg      GarageFinishRFn      GarageFinishUnf           SaleTypeWD 
   -5.984341e-02       -1.575791e-02       -4.264204e-02       -1.615152e-02 
```

The adjusted R squared value is high and RMSE shows fairly low values. The AIC and BIC also have very low (negative) values indicating this linear model with 22 features represents the predictivity reasonably. The individual vif is not that high and remains in the bearable range (<10), however, the average vif is 2.835265 which is greater than 1 and therefore, reconstruction of variables is needed.

In order to find the best OLS model, stepwise regression was used and statistical results were obtained. The stepwise regression leaves much smaller AIC value of -3503.540 with the final model as below (15 variables);

```
logSalePrice ~ OverallQual + YearBuilt + YearRemodAdd + TotalBsmtSF +
    GrLivArea + FullBath + Fireplaces + GarageArea + LandSlope +
    Neighborhood + Heating + HeatingQC + CentralAir + KitchenQual +
    GarageFinish
```

```
           Step Df     Deviance Resid. Df Resid. Dev       AIC
1                                     776   10.63208 -3491.994
2     - BsmtQual  2 0.0041508335      778   10.63623 -3495.672
3    - ExterQual  2 0.0169653597      780   10.65320 -3498.357
4 - TotRmsAbvGrd  1 0.0001811555      781   10.65338 -3500.343
5     - SaleType  1 0.0060496022      782   10.65943 -3501.874
6   - GarageCars  1 0.0134706007      783   10.67290 -3502.833
7   - Condition1  5 0.1227217181      788   10.79562 -3503.400
8    - X1stFlrSF  1 0.0243741461      789   10.81999 -3503.540
```

- Adjusted R-squared value: 0.8834
- RMSE: 0.1171
- p-value: <2.2e-16
- vif:

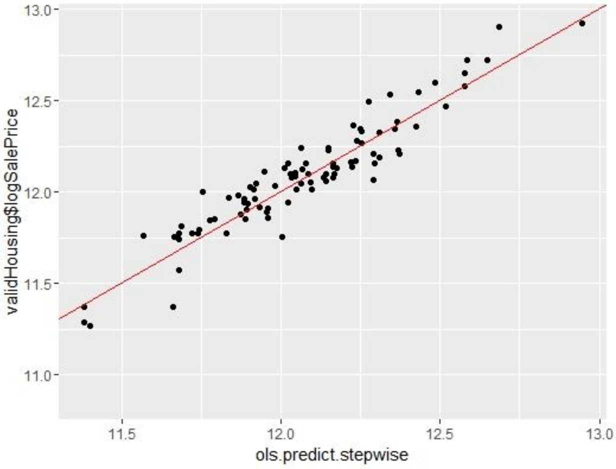
```
                  GVIF Df GVIF^(1/(2*Df))
 OverallQual   2.932478  1        1.712448
 YearBuilt     5.317470  1        2.305964
 YearRemodAdd  2.239944  1        1.496644
 TotalBsmtSF   1.656396  1        1.287011
 GrLivArea     2.954582  1        1.718890
 FullBath      2.595910  1        1.611183
 Fireplaces    1.541053  1        1.241392
 GarageArea    1.727771  1        1.314447
 LandSlope     1.395798  2        1.086940
 Neighborhood 17.002437 17        1.086905
 Heating       1.220629  1        1.104821
 HeatingQC     1.670208  2        1.136823
 CentralAir    1.355203  1        1.164132
 KitchenQual   2.551143  2        1.263816
 GarageFinish  2.078387  2        1.200692
```

- coefficient estimates:

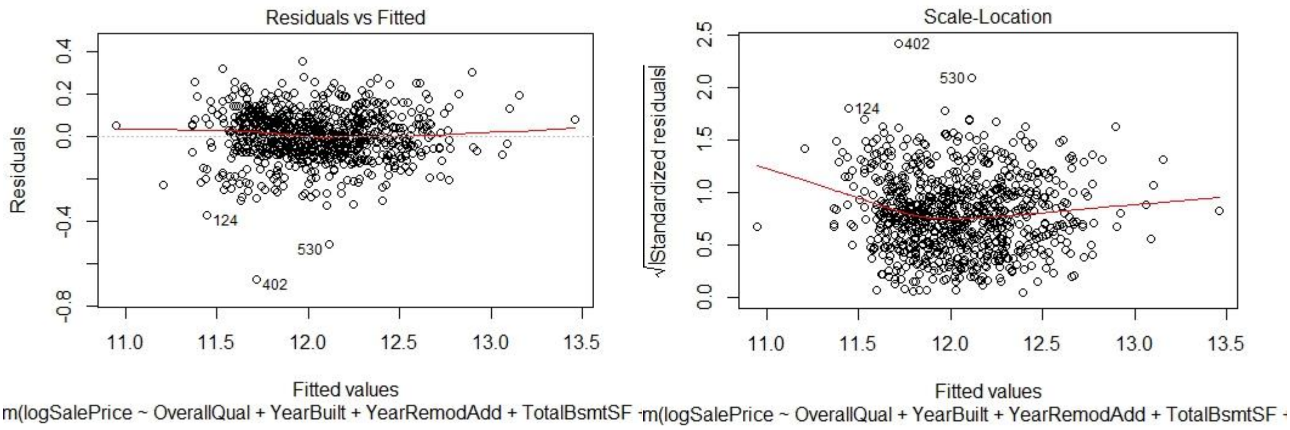
```
        (Intercept)         OverallQual           YearBuilt        YearRemodAdd         TotalBsmtSF 
       5.7797846697        0.0600623175        0.0013101163        0.0013092745        0.0001646231 
          GrLivArea            FullBath          Fireplaces          GarageArea        LandSlopeMod 
       0.0002532115       -0.0285954494        0.0441821759        0.0002225915        0.0642191316 
        LandSlopeSev NeighborhoodClearCr NeighborhoodCollgCr NeighborhoodCrawfor NeighborhoodEdwards 
       0.1353821022        0.0624884548       -0.0295981760        0.1173246442       -0.0779823872 
NeighborhoodGilbert  NeighborhoodIDOTRR  NeighborhoodMitchel   NeighborhoodNAmes NeighborhoodNoRidge 
      -0.0160594068       -0.0453667286       -0.0301712604       -0.0386578970        0.0174405969 
NeighborhoodNridgHt  NeighborhoodNWAmes  NeighborhoodOldTown   Neighborhoodother  NeighborhoodSawyer 
       0.0078515565       -0.0167116387       -0.1089095730       -0.0737480023       -0.0308358435 
NeighborhoodSawyerW NeighborhoodSomerst  NeighborhoodTimber        Heatingother        HeatingQCAvg 
      -0.0471396463       -0.0145227797        0.0122957443        0.0852685399       -0.0240176943 
  HeatingQCBelowAvg         CentralAirY      KitchenQualAvg KitchenQualBelowAvg      GarageFinishRFn 
      -0.0311719894        0.1520843465       -0.0477590105       -0.0611035185       -0.0150096260 
     GarageFinishUnf 
      -0.0420380825 
```

The statistical results were not drastically changed, however, the average vif decreased to 2.310478, which is slightly less value than that of previous model.

The validation data set was applied to the stepwise regression model, confirming RMSE = 0.1047958 and R^2 = 0.8877451. The validation value of SalePrice and the predicted value from the stepwise regression model is plotted below;



ii. Residuals



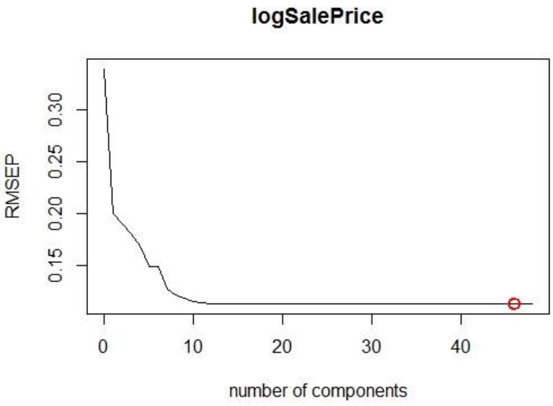
The residual patterns of stepwise OLS regression are displayed above. The residuals show a good alignment and symmetricity about residual 0 axis, however, the standard residuals line indicates that the OLS model needs a support from non-linear terms in order to increase the fitness. Moreover, we can detect several outliers from the plots which can highly affect the regression of the data. Therefore, we need to consider adding or changing variables, introducing non-linear formula or evaluating impact of outliers. The other interesting point is that the residuals show the following ncvTest results with $p < 0.05$:

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 4.832866    Df = 1    p = 0.02792218
```

Therefore, the variance of residuals should be accepted as not constant, and OLS may not be a best model for this data set.

(c) PLS model using hyperparameter tuning with cross-validation method, the number of component is recommended as 46 with RMSE of 0.1124. (see the chart and plot below)

```
(Intercept)     1 comps     2 comps     3 comps     4 comps     5 comps     6 comps     7 comps
    0.3394      0.1994      0.1904      0.1801      0.1676      0.1490      0.1485      0.1269
   8 comps     9 comps    10 comps    11 comps    12 comps    13 comps    14 comps    15 comps
    0.1212      0.1180      0.1148      0.1136      0.1131      0.1129      0.1127      0.1126
  16 comps    17 comps    18 comps    19 comps    20 comps    21 comps    22 comps    23 comps
    0.1125      0.1125      0.1125      0.1124      0.1124      0.1124      0.1124      0.1124
  24 comps    25 comps    26 comps    27 comps    28 comps    29 comps    30 comps    31 comps
    0.1124      0.1124      0.1124      0.1124      0.1124      0.1124      0.1124      0.1124
  32 comps    33 comps    34 comps    35 comps    36 comps    37 comps    38 comps    39 comps
    0.1124      0.1124      0.1124      0.1124      0.1124      0.1124      0.1124      0.1124
  40 comps    41 comps    42 comps    43 comps    44 comps    45 comps    46 comps    47 comps
    0.1124      0.1124      0.1124      0.1124      0.1124      0.1124      0.1124      0.1124
```



Using the recommended number of component, pls model provides the following coefficients;

```
        OverallQual           YearBuilt         YearRemodAdd          TotalBsmtSF             X1stFlrSF
       6.200526e-02        1.072015e-03         1.427450e-03         1.808820e-04         -3.227195e-05
          GrLivArea            FullBath         TotRmsAbvGrd           Fireplaces            GarageCars
       2.657846e-04       -3.472379e-02        -1.889469e-03         4.494253e-02          1.122061e-02
         GarageArea        LandSlopeMod         LandSlopeSev NeighborhoodClearCr NeighborhoodCollgCr
       2.036161e-04        6.647231e-02         1.219577e-01         6.648541e-02         -2.886288e-02
NeighborhoodCrawfor NeighborhoodEdwards  NeighborhoodGilbert  NeighborhoodIDOTRR NeighborhoodMitchel
       1.220569e-01       -6.884880e-02        -1.805327e-02        -4.991052e-02         -2.932592e-02
  NeighborhoodNAmes NeighborhoodNoRidge  NeighborhoodNridgHt  NeighborhoodNWAmes NeighborhoodOldTown
      -2.032884e-02        1.758078e-02         6.385443e-03        -4.628257e-03         -9.774068e-02
  Neighborhoodother  NeighborhoodSawyer NeighborhoodSawyerW NeighborhoodSomerst  NeighborhoodTimber
      -6.915921e-02       -1.364660e-02        -3.825544e-02        -1.274993e-02          1.792155e-02
    Condition1Feedr      Condition1Norm       Condition1PosA       Condition1PosN          Condition1RR
       4.271067e-02        6.488131e-02         7.947263e-02         5.286246e-02          4.225894e-02
       ExterQualAvg   ExterQualBelowAvg          BsmtQualAvg    BsmtQualBelowAvg           Heatingother
       3.331361e-03       -7.443845e-02        -1.009613e-02         4.688324e-03          7.007799e-02
       HeatingQCAvg   HeatingQCBelowAvg          CentralAirY      KitchenQualAvg KitchenQualBelowAvg
      -1.869659e-02       -4.842971e-02         1.494895e-01        -4.466112e-02         -4.478581e-02
    GarageFinishRFn     GarageFinishUnf           SaleTypeWD
      -1.834999e-02       -4.553599e-02        -3.222139e-03
```

Caret package in R provides more intuitive and better results. The function determines the number of components as 3 with the following RMSE and R squared values;

```
ncomp      RMSE  Rsquared     RMSESD RsquaredSD
1     1 0.1995513 0.6545181 0.01225490 0.04991349
2     2 0.1905452 0.6823291 0.01262157 0.05599725
3     3 0.1801621 0.7150044 0.01372067 0.05851881
```

(d) Using Caret package, LASSO model was built with hyperparameter tuning (CV=5). The model provides the lambda and RMSE results as below;

```
  alpha       lambda      RMSE   Rsquared      RMSESD  RsquaredSD
1  0.10 0.0005362793 0.1206977 0.8745925 0.008614193 0.01924217
2  0.10 0.0053627935 0.1203710 0.8752999 0.008727701 0.01939775
3  0.10 0.0536279349 0.1230244 0.8725163 0.009540844 0.02049030
4  0.55 0.0005362793 0.1205558 0.8749039 0.008639999 0.01924979
5  0.55 0.0053627935 0.1203111 0.8756501 0.009022786 0.01980501
6  0.55 0.0536279349 0.1369725 0.8576269 0.010702220 0.01573734
7  1.00 0.0005362793 0.1204271 0.8752018 0.008662886 0.01925244
8  1.00 0.0053627935 0.1205703 0.8753845 0.009261739 0.01977306
9  1.00 0.0536279349 0.1558459 0.8320109 0.012681615 0.01955133
```

The coefficients are;

```
(Intercept)          5.9539854694
OverallQual          0.0640504212
YearBuilt            0.0009882522
YearRemodAdd         0.0014961687
TotalBsmtSF          0.0001550254
GrLivArea            0.0002416047
FullBath            -0.0068595922
Fireplaces           0.0457212738
GarageCars           0.0146060900
GarageArea           0.0001928034
LandSlopeMod         0.0463747422
LandSlopeSev         0.1080960521
NeighborhoodClearCr  0.0826975387
NeighborhoodCrawfor  0.1290723022
NeighborhoodEdwards -0.0360704934
NeighborhoodIDOTRR  -0.0102684555
NeighborhoodNoRidge  0.0335223612
NeighborhoodNridgHt  0.0131988387
NeighborhoodOldTown -0.0779098877
Neighborhoodother   -0.0395386383
NeighborhoodSawyerW -0.0067770057
NeighborhoodTimber   0.0223606746
Condition1Norm       0.0188089951
ExterQualBelowAvg   -0.0270374581
BsmtQualAvg         -0.0025141674
Heatingother         0.0354963150
HeatingQCAvg        -0.0131849215
HeatingQCBelowAvg   -0.0211734467
CentralAirY          0.1403312561
KitchenQualAvg      -0.0357057179
KitchenQualBelowAvg -0.0134441476
GarageFinishRFn     -0.0029578870
GarageFinishUnf     -0.0331000380
```

(e) Based on the RMSE of each model from (b) to (d), OLS is selected for prediction of SalePrice of test data. The prediction result is attached to this homework submission. (testHousing_result_Jeon-HW4.csv)

head(test)

```
Id SalePrice
1 197409.5
2 148443.9
3 277744.8
4 137379.5
5 135310.6
6 100343.7
```