homework3

syh

May 22, 2017

```
# in this version, we use mice to imute missing value
```

```
# get data
raw_train <- read.csv(file = "H:/kaggle/houseprice/data/train.csv",
                      stringsAsFactors = FALSE)

raw_test <- read.csv(file = "H:/kaggle/houseprice/data/test.csv", stringsAsFactors = F)

raw_test$SalePrice <- rep(0,dim(raw_test)[1])

all_data <- rbind(raw_train, raw_test)
```

```
# deal with NA value
# first have a look which columns have NAs
na_sort <- sapply(all_data, function(x){
  sum(is.na(x))
})

na_sort
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             4           486             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0          2721             0             0             2
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             0             0             0             1             1
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##            24            23             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##            81            82            82            79             1
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##            80             1             1             1             0
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             1             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             2             2             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             1             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             2             0          1420           157           159
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##           159             1             1           159           159
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
```

```
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0          2909          2348          2814
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             1             0
##     SalePrice
##             0
```

```
# at first we remove columns with na in excess of 5% of all
keep_col <- which(na_sort < dim(all_data)[1] * 0.05)
all_data <- all_data[keep_col]
```

```
# check other columns with NAs
sort(sapply(all_data, function(x){
  sum(is.na(x))
}), decreasing = TRUE)
```

```
##      BsmtCond  BsmtExposure      BsmtQual  BsmtFinType2  BsmtFinType1
##            82            82            81            80            79
##    MasVnrType    MasVnrArea      MSZoning     Utilities  BsmtFullBath
##            24            23             4             2             2
##  BsmtHalfBath    Functional   Exterior1st   Exterior2nd    BsmtFinSF1
##             2             2             1             1             1
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF    Electrical   KitchenQual
##             1             1             1             1             1
##    GarageCars    GarageArea      SaleType            Id    MSSubClass
##             1             1             1             0             0
##       LotArea        Street      LotShape   LandContour     LotConfig
##             0             0             0             0             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl     ExterQual     ExterCond    Foundation
##             0             0             0             0             0
##       Heating     HeatingQC    CentralAir     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea      FullBath      HalfBath  BedroomAbvGr
##             0             0             0             0             0
##  KitchenAbvGr  TotRmsAbvGrd    Fireplaces    PavedDrive    WoodDeckSF
##             0             0             0             0             0
##   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold SaleCondition     SalePrice
##             0             0             0             0             0
```

```
# 2. Missingness is caused by that it doesn't exsit
# by mice, to impute them.
library(mice)
# md.pattern(raw_train)
```

```
# convert character to factor
cha_col <- c("MSSubClass", "MSZoning", "Street", "LotShape", "LandContour",
             "Utilities", "LotConfig", "LandSlope", "Neighborhood", "Condition1", "Condition2","BldgType
"PavedDrive", "MoSold", "SaleType", "SaleCondition")
all_data[cha_col] <- lapply(all_data[cha_col], as.factor)
```

```r
# str(all_data)
# summary(all_data)
```

```r
# impute nas by mice
im_all_data <- mice(data = all_data, m = 1, method = "cart",printFlag = F)
real_all_data <- complete(im_all_data)
```

```r
# check again if there is no missing value
sort(sapply(real_all_data, function(x){sum(is.na(x))}), decreasing = TRUE)
```

```
##            Id    MSSubClass      MSZoning       LotArea        Street
##             0             0             0             0             0
##      LotShape   LandContour     Utilities     LotConfig     LandSlope
##             0             0             0             0             0
##  Neighborhood    Condition1    Condition2      BldgType    HouseStyle
##             0             0             0             0             0
##   OverallQual   OverallCond     YearBuilt  YearRemodAdd     RoofStyle
##             0             0             0             0             0
##      RoofMatl   Exterior1st   Exterior2nd    MasVnrType    MasVnrArea
##             0             0             0             0             0
##     ExterQual     ExterCond    Foundation      BsmtQual      BsmtCond
##             0             0             0             0             0
##  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2    BsmtFinSF2
##             0             0             0             0             0
##     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC    CentralAir
##             0             0             0             0             0
##    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF     GrLivArea
##             0             0             0             0             0
##  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath  BedroomAbvGr
##             0             0             0             0             0
##  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional    Fireplaces
##             0             0             0             0             0
##    GarageCars    GarageArea    PavedDrive    WoodDeckSF   OpenPorchSF
##             0             0             0             0             0
## EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea       MiscVal
##             0             0             0             0             0
##        MoSold        YrSold      SaleType SaleCondition     SalePrice
##             0             0             0             0             0
```

```r
# there isn't missing value any more.
```

```r
# record real_all_data data set
write.csv(file = "H:/kaggle/houseprice/data/real_all_data_mice.csv", x = real_all_data)
```

```r
#create real_all_data without id
real_all_data <- subset(real_all_data, select = -Id)
```

```r
# feature engineering
# how many years are these houses
# train_no_miss$Age <- 2017 - train_no_miss[,"YearBuilt"]

# total Floor square feet + basement
# train_no_miss$tot_Flo_area <- train_no_miss$X1stFlrSF
                # + train_no_miss$X2ndFlrSF
                # + train_no_miss$TotalBsmtSF
```

```r
# how many years house last since repairing
# train_no_miss$rep_yea <- 2017 - train_no_miss$YearRemodAdd
```

```r
# transform sale price to more normal,
# in order to subject to assumptin of linear regression
real_all_data$SalePrice <- log(real_all_data$SalePrice)
```

```r
# plot(density(whole_train$SalePrice))
```

```r
# train a simple linear regression model first
simple_lm <- lm(SalePrice ~ ., data = real_all_data[c(1:1460),])
```

```
## Warning: contrasts dropped from factor MSSubClass due to missing levels
```

```r
summary(simple_lm)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = real_all_data[c(1:1460), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68269 -0.04610  0.00138  0.05188  0.68269
##
## Coefficients: (5 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.911e+00  4.770e+00   1.658 0.097490 .
## MSSubClass30   -7.819e-02  2.155e-02  -3.627 0.000298 ***
## MSSubClass40   -5.850e-02  6.695e-02  -0.874 0.382403
## MSSubClass45   -2.435e-01  1.029e-01  -2.367 0.018084 *
## MSSubClass50   -1.618e-02  3.978e-02  -0.407 0.684264
## MSSubClass60   -3.218e-02  3.509e-02  -0.917 0.359310
## MSSubClass70    1.001e-02  3.807e-02   0.263 0.792712
## MSSubClass75   -6.244e-02  7.028e-02  -0.888 0.374471
## MSSubClass80   -5.354e-02  5.775e-02  -0.927 0.354047
## MSSubClass85   -1.107e-02  4.817e-02  -0.230 0.818240
## MSSubClass90   -2.276e-02  3.262e-02  -0.698 0.485469
## MSSubClass120  -3.984e-02  6.717e-02  -0.593 0.553228
## MSSubClass160  -1.359e-01  7.999e-02  -1.699 0.089502 .
## MSSubClass180  -8.109e-02  8.890e-02  -0.912 0.361890
## MSSubClass190   3.610e-02  1.278e-01   0.282 0.777623
## MSZoning2       4.797e-01  5.591e-02   8.581  < 2e-16 ***
## MSZoning3       4.298e-01  5.585e-02   7.696 2.91e-14 ***
## MSZoning4       4.451e-01  4.821e-02   9.233  < 2e-16 ***
## MSZoning5       3.914e-01  4.515e-02   8.670  < 2e-16 ***
## LotArea         2.898e-06  4.942e-07   5.864 5.83e-09 ***
## Street2         1.352e-01  5.699e-02   2.373 0.017796 *
## LotShape2       2.819e-02  1.893e-02   1.489 0.136739
## LotShape3       3.125e-02  3.971e-02   0.787 0.431496
## LotShape4       7.873e-03  7.330e-03   1.074 0.283047
## LandContour2    3.266e-02  2.375e-02   1.375 0.169264
## LandContour3   -1.474e-03  2.974e-02  -0.050 0.960480
## LandContour4    2.841e-02  1.687e-02   1.683 0.092558 .
## Utilities2     -2.616e-01  1.281e-01  -2.042 0.041351 *
## LotConfig2      2.902e-02  1.454e-02   1.996 0.046148 *
```

```
## LotConfig3     -3.406e-02  1.815e-02  -1.876 0.060853 .
## LotConfig4     -8.205e-02  5.736e-02  -1.430 0.152871
## LotConfig5     -1.442e-02  7.980e-03  -1.807 0.070956 .
## LandSlope2      3.647e-02  1.827e-02   1.996 0.046169 *
## LandSlope3     -1.933e-01  5.177e-02  -3.734 0.000197 ***
## Neighborhood2   3.182e-02  8.938e-02   0.356 0.721875
## Neighborhood3  -2.893e-03  5.365e-02  -0.054 0.957005
## Neighborhood4   4.007e-02  4.358e-02   0.919 0.358052
## Neighborhood5   2.967e-02  4.252e-02   0.698 0.485396
## Neighborhood6  -7.257e-03  3.301e-02  -0.220 0.826033
## Neighborhood7   1.113e-01  3.971e-02   2.804 0.005130 **
## Neighborhood8  -6.854e-02  3.687e-02  -1.859 0.063259 .
## Neighborhood9   4.496e-03  3.537e-02   0.127 0.898860
## Neighborhood10 -1.747e-03  4.938e-02  -0.035 0.971788
## Neighborhood11 -1.208e-01  5.582e-02  -2.164 0.030627 *
## Neighborhood12 -5.236e-02  3.742e-02  -1.399 0.161983
## Neighborhood13 -2.938e-02  3.591e-02  -0.818 0.413357
## Neighborhood14  4.729e-02  3.875e-02   1.220 0.222520
## Neighborhood15  1.582e-03  6.349e-02   0.025 0.980131
## Neighborhood16  8.273e-02  3.386e-02   2.443 0.014709 *
## Neighborhood17 -2.799e-02  3.680e-02  -0.761 0.447028
## Neighborhood18 -2.947e-02  4.411e-02  -0.668 0.504125
## Neighborhood19 -1.531e-02  3.722e-02  -0.411 0.681018
## Neighborhood20  3.048e-03  3.574e-02   0.085 0.932055
## Neighborhood21  3.358e-02  4.140e-02   0.811 0.417454
## Neighborhood22  1.316e-01  3.850e-02   3.418 0.000652 ***
## Neighborhood23  9.213e-03  4.460e-02   0.207 0.836382
## Neighborhood24  1.611e-02  3.752e-02   0.429 0.667759
## Neighborhood25  4.777e-02  4.793e-02   0.997 0.319104
## Condition12     2.282e-02  2.262e-02   1.009 0.313214
## Condition13     7.647e-02  1.864e-02   4.101 4.38e-05 ***
## Condition14     5.793e-02  4.513e-02   1.284 0.199538
## Condition15     8.263e-02  3.371e-02   2.451 0.014398 *
## Condition16    -3.889e-02  4.110e-02  -0.946 0.344192
## Condition17     3.248e-02  3.087e-02   1.052 0.292978
## Condition18     6.774e-03  7.979e-02   0.085 0.932361
## Condition19     6.072e-02  5.771e-02   1.052 0.292986
## Condition22     2.020e-01  1.132e-01   1.785 0.074582 .
## Condition23     1.739e-01  1.003e-01   1.734 0.083183 .
## Condition24     3.314e-01  1.757e-01   1.886 0.059471 .
## Condition25    -6.776e-01  1.335e-01  -5.075 4.47e-07 ***
## Condition26    -4.406e-01  2.255e-01  -1.954 0.050973 .
## Condition27     8.478e-02  1.489e-01   0.569 0.569328
## Condition28     1.404e-01  1.304e-01   1.077 0.281699
## BldgType2      -6.062e-02  1.256e-01  -0.483 0.629339
## BldgType3              NA         NA      NA       NA
## BldgType4      -2.697e-02  7.150e-02  -0.377 0.706093
## BldgType5       4.102e-03  6.806e-02   0.060 0.951948
## HouseStyle2     2.055e-01  1.026e-01   2.002 0.045484 *
## HouseStyle3    -2.473e-02  4.004e-02  -0.618 0.536955
## HouseStyle4     1.153e-03  7.789e-02   0.015 0.988193
## HouseStyle5     9.597e-02  7.432e-02   1.291 0.196839
## HouseStyle6    -8.116e-03  3.667e-02  -0.221 0.824897
## HouseStyle7    -2.600e-02  5.332e-02  -0.488 0.625865
```

```
## HouseStyle8     3.162e-02  6.287e-02   0.503 0.615082
## OverallQual2    4.700e-01  1.434e-01   3.278 0.001075 **
## OverallQual3    5.570e-01  1.320e-01   4.221 2.62e-05 ***
## OverallQual4    5.963e-01  1.307e-01   4.564 5.52e-06 ***
## OverallQual5    6.397e-01  1.313e-01   4.872 1.25e-06 ***
## OverallQual6    6.717e-01  1.317e-01   5.102 3.90e-07 ***
## OverallQual7    7.075e-01  1.317e-01   5.373 9.28e-08 ***
## OverallQual8    7.617e-01  1.322e-01   5.761 1.06e-08 ***
## OverallQual9    8.328e-01  1.347e-01   6.184 8.53e-10 ***
## OverallQual10   8.526e-01  1.380e-01   6.178 8.84e-10 ***
## OverallCond2   -5.104e-01  2.134e-01  -2.391 0.016940 *
## OverallCond3   -6.468e-01  2.251e-01  -2.873 0.004139 **
## OverallCond4   -5.740e-01  2.268e-01  -2.531 0.011498 *
## OverallCond5   -5.278e-01  2.264e-01  -2.331 0.019926 *
## OverallCond6   -4.897e-01  2.265e-01  -2.162 0.030829 *
## OverallCond7   -4.549e-01  2.265e-01  -2.009 0.044803 *
## OverallCond8   -4.441e-01  2.266e-01  -1.960 0.050255 .
## OverallCond9   -3.891e-01  2.286e-01  -1.702 0.089005 .
## YearBuilt       1.712e-03  3.773e-04   4.537 6.28e-06 ***
## YearRemodAdd    8.180e-04  2.519e-04   3.247 0.001197 **
## RoofStyle2     -1.494e-02  8.306e-02  -0.180 0.857270
## RoofStyle3     -2.417e-02  9.089e-02  -0.266 0.790311
## RoofStyle4     -1.443e-02  8.322e-02  -0.173 0.862352
## RoofStyle5      4.768e-02  9.678e-02   0.493 0.622363
## RoofStyle6      5.084e-01  1.696e-01   2.997 0.002784 **
## RoofMatl2       2.557e+00  1.488e-01  17.179  < 2e-16 ***
## RoofMatl3       2.972e+00  2.157e-01  13.778  < 2e-16 ***
## RoofMatl4       2.800e+00  2.111e-01  13.263  < 2e-16 ***
## RoofMatl5       2.545e+00  1.885e-01  13.501  < 2e-16 ***
## RoofMatl6       2.580e+00  1.704e-01  15.145  < 2e-16 ***
## RoofMatl7       2.475e+00  1.645e-01  15.050  < 2e-16 ***
## RoofMatl8       2.645e+00  1.531e-01  17.277  < 2e-16 ***
## Exterior1st2    5.895e-02  1.493e-01   0.395 0.692936
## Exterior1st3   -1.943e-01  1.301e-01  -1.493 0.135689
## Exterior1st4    1.290e-01  5.820e-02   2.217 0.026811 *
## Exterior1st5   -1.112e-01  1.254e-01  -0.887 0.375396
## Exterior1st6   -3.074e-02  8.776e-02  -0.350 0.726163
## Exterior1st7    4.415e-02  5.901e-02   0.748 0.454468
## Exterior1st8    1.757e-02  1.254e-01   0.140 0.888559
## Exterior1st9    8.197e-02  6.742e-02   1.216 0.224266
## Exterior1st10   4.403e-02  5.814e-02   0.757 0.449040
## Exterior1st11   1.310e-01  1.095e-01   1.197 0.231719
## Exterior1st12   7.763e-02  6.582e-02   1.179 0.238458
## Exterior1st13   5.171e-02  6.147e-02   0.841 0.400428
## Exterior1st14   1.695e-02  5.661e-02   0.299 0.764648
## Exterior1st15   6.757e-02  6.117e-02   1.105 0.269576
## Exterior2nd2   -3.224e-02  1.009e-01  -0.319 0.749422
## Exterior2nd3    2.970e-02  9.331e-02   0.318 0.750293
## Exterior2nd4   -7.729e-02  5.976e-02  -1.293 0.196112
## Exterior2nd5           NA         NA      NA       NA
## Exterior2nd6    7.016e-02  8.571e-02   0.818 0.413235
## Exterior2nd7   -3.474e-02  5.659e-02  -0.614 0.539457
## Exterior2nd8   -2.096e-02  6.494e-02  -0.323 0.746881
## Exterior2nd9   -4.578e-02  6.548e-02  -0.699 0.484599
```

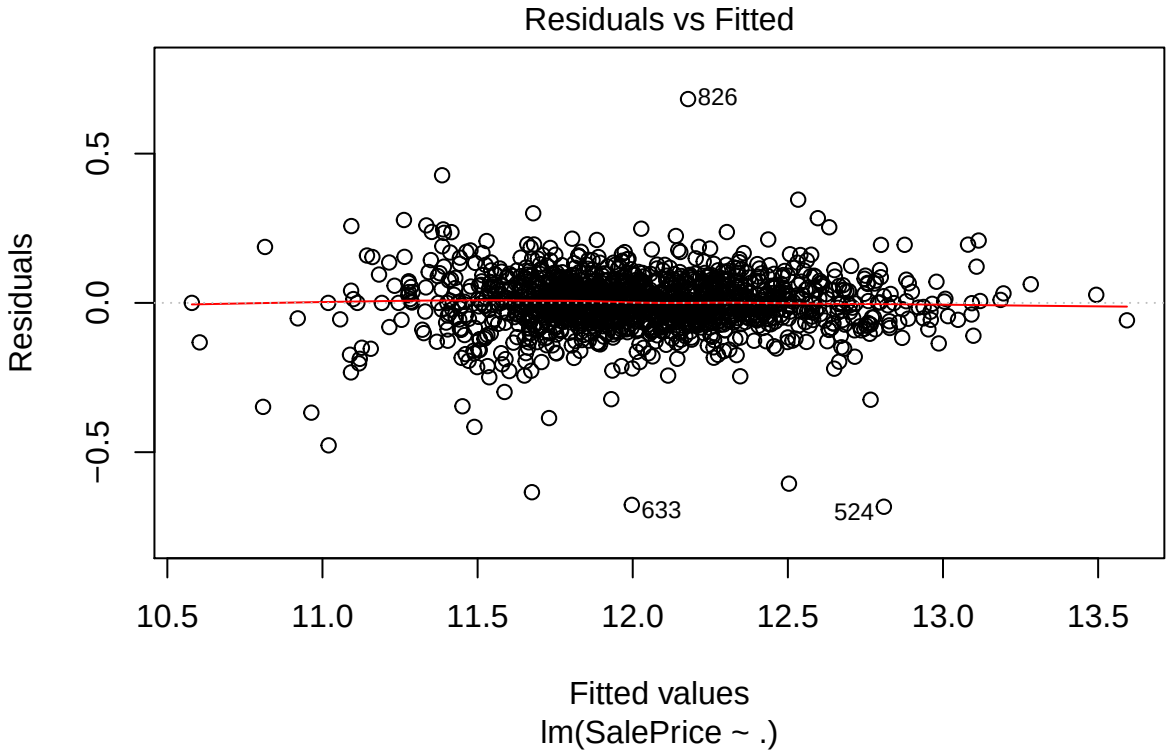
```
## Exterior2nd10  -1.226e-01  1.241e-01  -0.988 0.323224
## Exterior2nd11  -3.126e-02  5.495e-02  -0.569 0.569522
## Exterior2nd12  -1.047e-01  7.776e-02  -1.346 0.178479
## Exterior2nd13  -4.581e-02  6.232e-02  -0.735 0.462409
## Exterior2nd14  -2.159e-02  5.920e-02  -0.365 0.715392
## Exterior2nd15   3.448e-04  5.453e-02   0.006 0.994957
## Exterior2nd16  -5.380e-02  5.673e-02  -0.948 0.343148
## MasVnrType2     4.602e-02  3.081e-02   1.493 0.135589
## MasVnrType3     3.815e-02  3.107e-02   1.228 0.219674
## MasVnrType4     5.572e-02  3.258e-02   1.710 0.087468 .
## MasVnrArea      2.161e-05  2.628e-05   0.822 0.411193
## ExterQual2      7.330e-02  5.614e-02   1.306 0.191932
## ExterQual3      8.332e-03  2.355e-02   0.354 0.723552
## ExterQual4      4.641e-03  2.558e-02   0.181 0.856041
## ExterCond2     -7.127e-02  8.532e-02  -0.835 0.403709
## ExterCond3     -4.673e-02  8.147e-02  -0.574 0.566327
## ExterCond4     -1.867e-01  1.657e-01  -1.127 0.260116
## ExterCond5     -3.462e-02  8.168e-02  -0.424 0.671760
## Foundation2     8.719e-03  1.462e-02   0.596 0.551032
## Foundation3     2.776e-02  1.567e-02   1.771 0.076743 .
## Foundation4    -1.973e-02  3.587e-02  -0.550 0.582309
## Foundation5     1.102e-01  4.949e-02   2.226 0.026180 *
## Foundation6    -1.182e-01  6.608e-02  -1.788 0.074009 .
## BsmtQual2       3.158e-03  2.816e-02   0.112 0.910730
## BsmtQual3      -2.063e-02  1.544e-02  -1.337 0.181623
## BsmtQual4      -2.089e-02  1.882e-02  -1.110 0.267333
## BsmtCond2       2.693e-02  2.383e-02   1.130 0.258612
## BsmtCond3      -1.862e-02  1.478e-01  -0.126 0.899785
## BsmtCond4       2.916e-02  1.916e-02   1.522 0.128154
## BsmtExposure2   3.115e-02  1.387e-02   2.245 0.024923 *
## BsmtExposure3  -7.736e-04  1.364e-02  -0.057 0.954766
## BsmtExposure4  -7.409e-03  9.928e-03  -0.746 0.455642
## BsmtFinType12  -4.088e-03  1.252e-02  -0.326 0.744106
## BsmtFinType13   8.818e-03  1.146e-02   0.770 0.441707
## BsmtFinType14  -2.787e-02  1.694e-02  -1.645 0.100130
## BsmtFinType15  -3.500e-03  1.345e-02  -0.260 0.794756
## BsmtFinType16  -1.629e-02  1.321e-02  -1.233 0.217683
## BsmtFinSF1      1.332e-04  2.127e-05   6.262 5.28e-10 ***
## BsmtFinType22  -7.684e-02  3.433e-02  -2.238 0.025378 *
## BsmtFinType23  -2.629e-02  4.135e-02  -0.636 0.524997
## BsmtFinType24  -3.745e-02  3.365e-02  -1.113 0.265950
## BsmtFinType25  -3.045e-02  3.223e-02  -0.945 0.344915
## BsmtFinType26  -2.167e-02  3.435e-02  -0.631 0.528147
## BsmtFinSF2      1.285e-04  3.997e-05   3.215 0.001339 **
## BsmtUnfSF       7.840e-05  1.832e-05   4.281 2.01e-05 ***
## TotalBsmtSF            NA         NA      NA       NA
## Heating2        2.002e-01  1.156e-01   1.732 0.083557 .
## Heating3        2.395e-01  1.195e-01   2.004 0.045287 *
## Heating4        9.108e-03  1.248e-01   0.073 0.941835
## Heating5        1.531e-01  1.431e-01   1.070 0.284818
## Heating6        2.701e-01  1.323e-01   2.041 0.041428 *
## HeatingQC2     -2.420e-02  2.169e-02  -1.115 0.264940
## HeatingQC3     -2.019e-02  9.435e-03  -2.140 0.032529 *
## HeatingQC4     -1.084e-01  1.221e-01  -0.887 0.375022
```

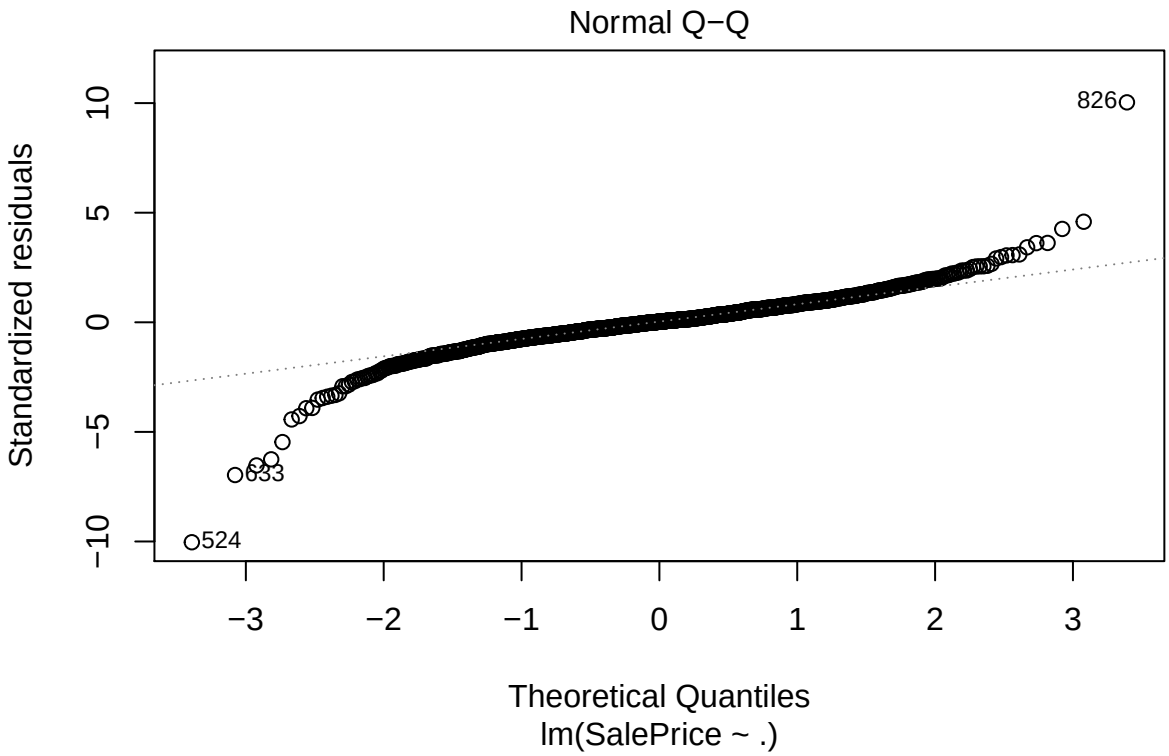
```
## HeatingQC5     -3.172e-02  9.395e-03  -3.376 0.000757 ***
## CentralAir2     5.646e-02  1.778e-02   3.176 0.001531 **
## Electrical2     1.429e-03  2.719e-02   0.053 0.958091
## Electrical3    -6.278e-02  7.775e-02  -0.807 0.419606
## Electrical4            NA         NA      NA       NA
## Electrical5    -2.702e-02  1.367e-02  -1.977 0.048324 *
## X1stFlrSF       2.300e-04  2.382e-05   9.655  < 2e-16 ***
## X2ndFlrSF       2.165e-04  2.386e-05   9.073  < 2e-16 ***
## LowQualFinSF    1.649e-04  8.460e-05   1.949 0.051468 .
## GrLivArea              NA         NA      NA       NA
## BsmtFullBath    2.346e-02  8.969e-03   2.616 0.009006 **
## BsmtHalfBath    5.135e-03  1.378e-02   0.373 0.709543
## FullBath        2.947e-02  1.005e-02   2.932 0.003432 **
## HalfBath        2.674e-02  9.497e-03   2.816 0.004938 **
## BedroomAbvGr    1.610e-03  6.338e-03   0.254 0.799562
## KitchenAbvGr   -6.077e-02  2.753e-02  -2.208 0.027453 *
## KitchenQual2   -4.163e-02  2.884e-02  -1.443 0.149152
## KitchenQual3   -5.588e-02  1.620e-02  -3.450 0.000580 ***
## KitchenQual4   -5.883e-02  1.811e-02  -3.248 0.001193 **
## TotRmsAbvGrd    2.803e-03  4.331e-03   0.647 0.517664
## Functional2    -2.375e-01  6.690e-02  -3.550 0.000400 ***
## Functional3     9.164e-04  3.969e-02   0.023 0.981584
## Functional4    -1.572e-02  4.020e-02  -0.391 0.695847
## Functional5    -9.812e-02  4.826e-02  -2.033 0.042226 *
## Functional6    -2.913e-01  1.257e-01  -2.318 0.020632 *
## Functional7     2.946e-02  3.497e-02   0.842 0.399675
## Fireplaces      2.538e-02  6.054e-03   4.192 2.96e-05 ***
## GarageCars      2.398e-02  9.872e-03   2.429 0.015274 *
## GarageArea      1.191e-04  3.402e-05   3.502 0.000479 ***
## PavedDrive2     3.003e-03  2.478e-02   0.121 0.903544
## PavedDrive3     2.091e-02  1.552e-02   1.347 0.178229
## WoodDeckSF      9.002e-05  2.634e-05   3.418 0.000652 ***
## OpenPorchSF     7.433e-05  5.237e-05   1.419 0.156084
## EnclosedPorch   1.222e-04  5.685e-05   2.150 0.031765 *
## X3SsnPorch      1.725e-04  1.010e-04   1.707 0.088044 .
## ScreenPorch     2.712e-04  5.505e-05   4.927 9.53e-07 ***
## PoolArea        1.848e-04  8.293e-05   2.229 0.026008 *
## MiscVal         6.090e-07  6.430e-06   0.095 0.924559
## MoSold2        -1.778e-03  2.154e-02  -0.083 0.934221
## MoSold3        -3.811e-03  1.885e-02  -0.202 0.839832
## MoSold4         7.555e-03  1.804e-02   0.419 0.675370
## MoSold5         1.006e-02  1.721e-02   0.584 0.559034
## MoSold6         1.660e-02  1.693e-02   0.981 0.326915
## MoSold7         7.242e-03  1.714e-02   0.423 0.672686
## MoSold8        -2.834e-03  1.815e-02  -0.156 0.875897
## MoSold9        -4.230e-03  2.065e-02  -0.205 0.837760
## MoSold10       -1.512e-02  1.956e-02  -0.773 0.439684
## MoSold11       -4.973e-03  1.982e-02  -0.251 0.801985
## MoSold12       -6.136e-04  2.138e-02  -0.029 0.977107
## YrSold         -2.654e-03  2.339e-03  -1.135 0.256721
## SaleType2       7.429e-02  7.988e-02   0.930 0.352574
## SaleType3       1.162e-01  4.551e-02   2.554 0.010768 *
## SaleType4      -2.745e-02  5.229e-02  -0.525 0.599648
## SaleType5       1.374e-02  5.656e-02   0.243 0.808128
```

```
## SaleType6       5.865e-02  5.909e-02   0.993 0.321088
## SaleType7       8.578e-02  7.088e-02   1.210 0.226444
## SaleType8       7.651e-02  6.625e-02   1.155 0.248377
## SaleType9      -1.749e-02  1.913e-02  -0.914 0.360833
## SaleCondition2  9.701e-02  6.587e-02   1.473 0.141104
## SaleCondition3  7.032e-02  3.951e-02   1.780 0.075305 .
## SaleCondition4  1.809e-02  2.772e-02   0.653 0.514009
## SaleCondition5  6.522e-02  1.312e-02   4.970 7.67e-07 ***
## SaleCondition6  9.141e-03  6.809e-02   0.134 0.893231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1031 on 1211 degrees of freedom
## Multiple R-squared:  0.9448, Adjusted R-squared:  0.9334
## F-statistic:  83.5 on 248 and 1211 DF,  p-value: < 2.2e-16
```

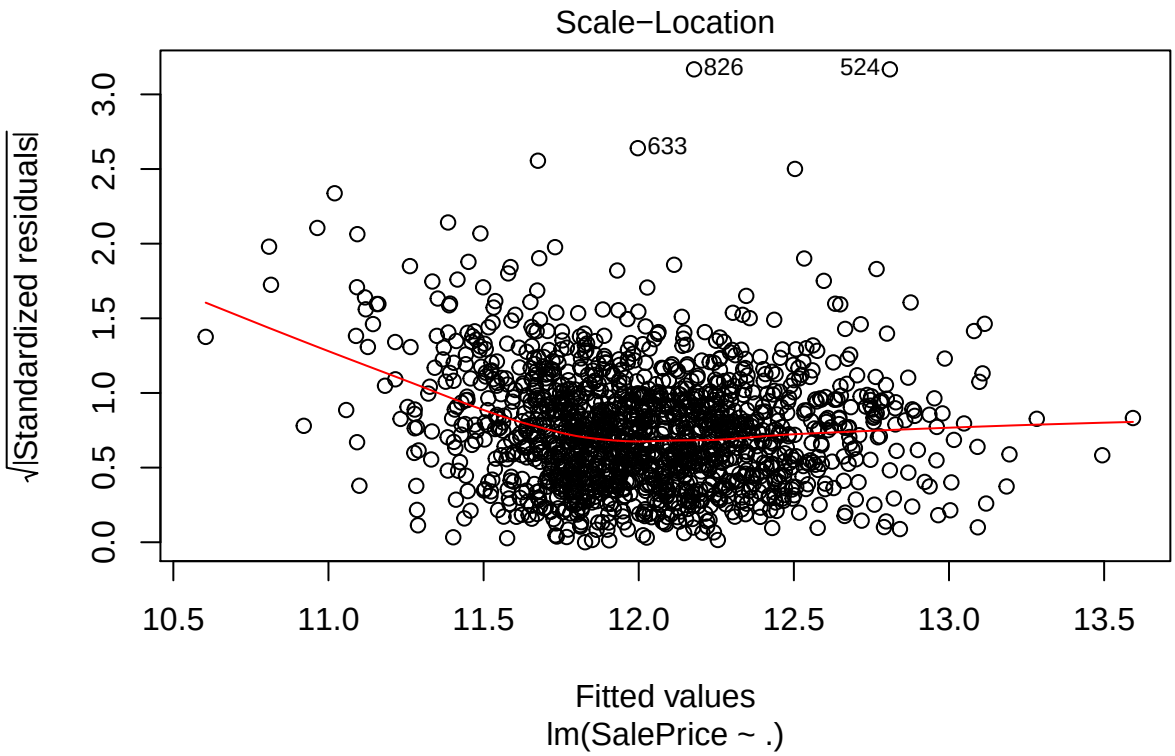
```r
# residual plot
# 1st: residual is unbais and homoscedastic,
# 2nd: basically, residual is subject to normal distribution which means variacne of error is constant.
# 4th: we know some outlier: 89, 826, 524 high leaverage and high residual
plot(simple_lm)
```

```
## Warning: not plotting observations with leverage one:
##   121, 251, 272, 326, 376, 399, 534, 584, 596, 667, 822, 945, 1012, 1188, 1231, 1271, 1276, 1299, 13
```

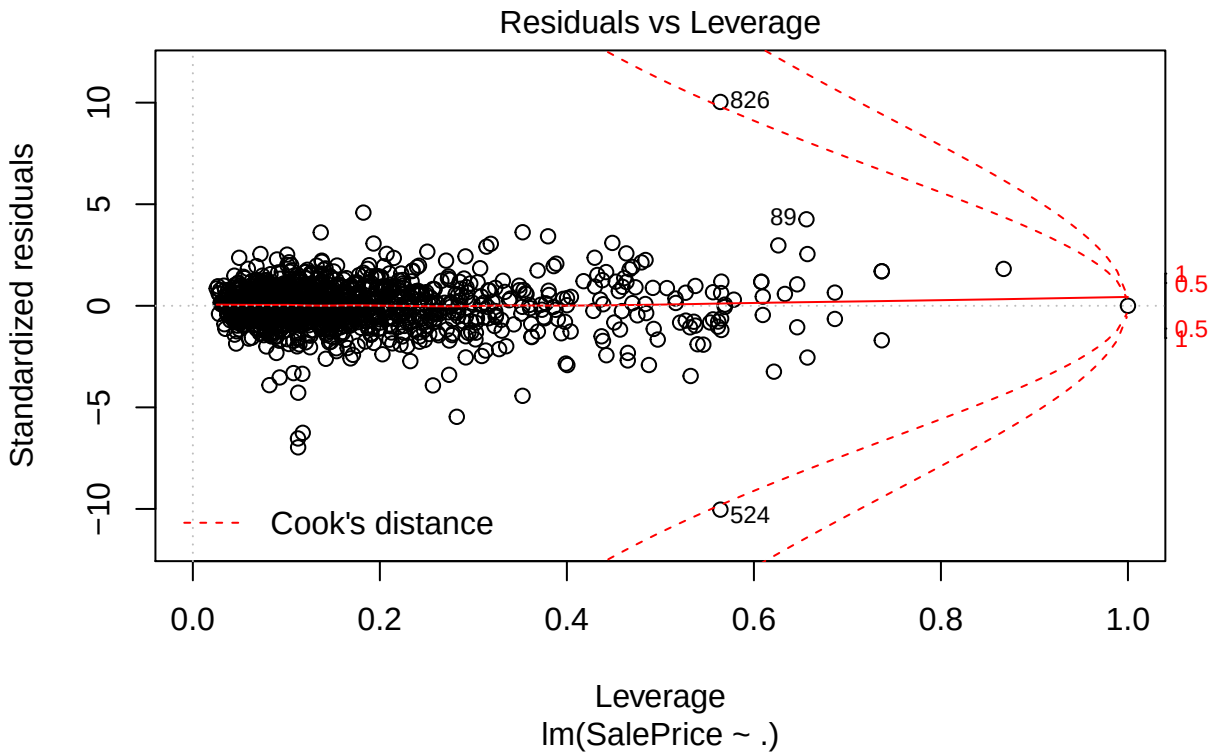





```
## Warning: not plotting observations with leverage one:
##   121, 251, 272, 326, 376, 399, 534, 584, 596, 667, 822, 945, 1012, 1188, 1231, 1271, 1276, 1299, 13
```




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```




```
# we can draw conclusion that basically underlying relations is linear
# Adjusted R-squared:  0.9358, it seems it's pretty good.
# F-statistic: 83.59, it's not very high, though.
```

```
# scale numerical features
# for purpose of comparing predection power of different features
# num_col <- setdiff(colnames(whole_train),cha_col)
# whole_train[setdiff(num_col,"SalePrice")] <- sapply(whole_train[setdiff(num_col,"SalePrice")], scale)
```

```
# get ind and dep data
dep_data <- real_all_data$SalePrice
# x must be a matrix for glmnet
ind_data <- model.matrix(~., subset(real_all_data, select = -SalePrice))


# split all data into for training model and for prediction
x_model <- ind_data[c(1:dim(raw_train)[1]),]
y_model <- dep_data[c(1:dim(raw_train)[1])]


x_pre <- ind_data[-c(1:dim(raw_train)[1]),]

# split training data into train and test
set.seed(1000)
train_ind <- sample(x = 1:dim(x_model)[1], size = dim(x_model)[1] * 0.7)

x_train <- x_model[train_ind,]
```

```r
y_train <- y_model[train_ind]

x_test <- x_model[-train_ind,]
y_test <- y_model[-train_ind]
```

```r
# resolve multicolinearity and select lamda
# we choose different penality methods
library(glmnet)
```

```
## Loading required package: Matrix
```

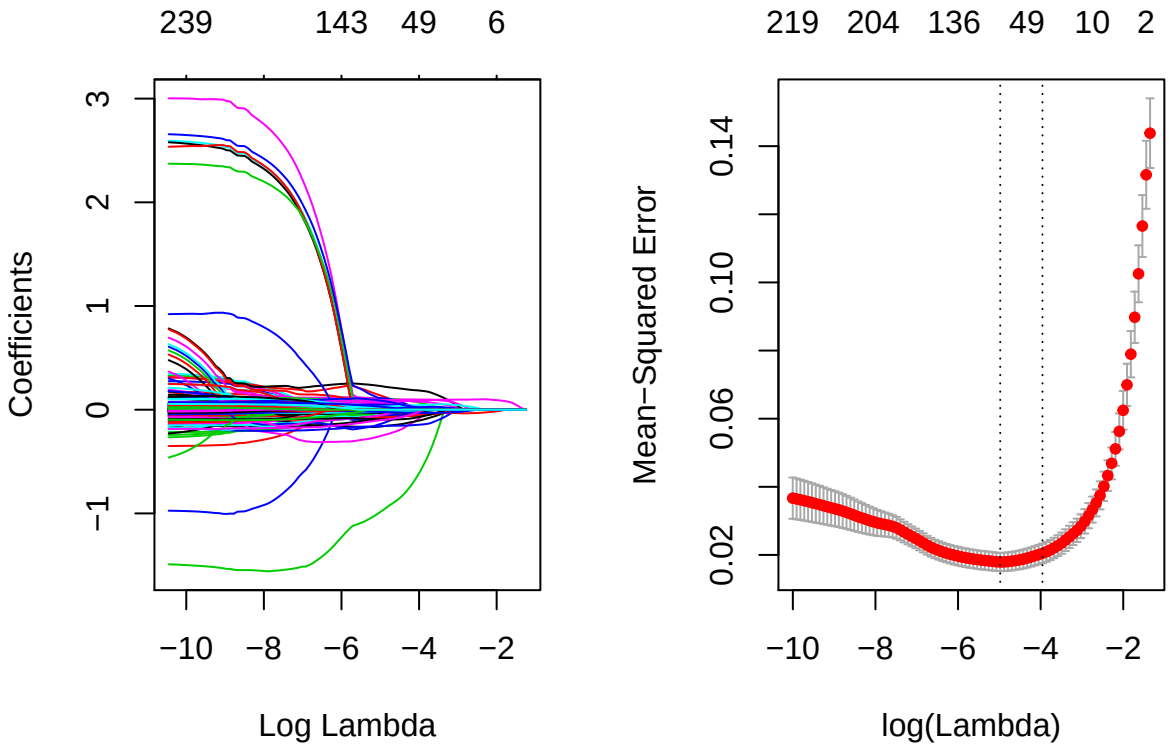
```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-10
```

```r
lasso_lm <- glmnet(x = x_train, y = y_train, alpha = 1)
ridge_lm <- glmnet(x = x_train, y = y_train, alpha = 0)
elnet_lm <- glmnet(x = x_train, y = y_train, alpha = 0.5)
```

```r
# train 11 models with different alpha, different penalty, ranging from 0 to 1
# by cross validation,default folders are 10
for(i in c(0:10)){
  assign(paste("cvglm", i,sep = ""),
         cv.glmnet(x = x_test, y = y_test, alpha = i/10,
                   type.measure = "mse", family = "gaussian"))
}
```

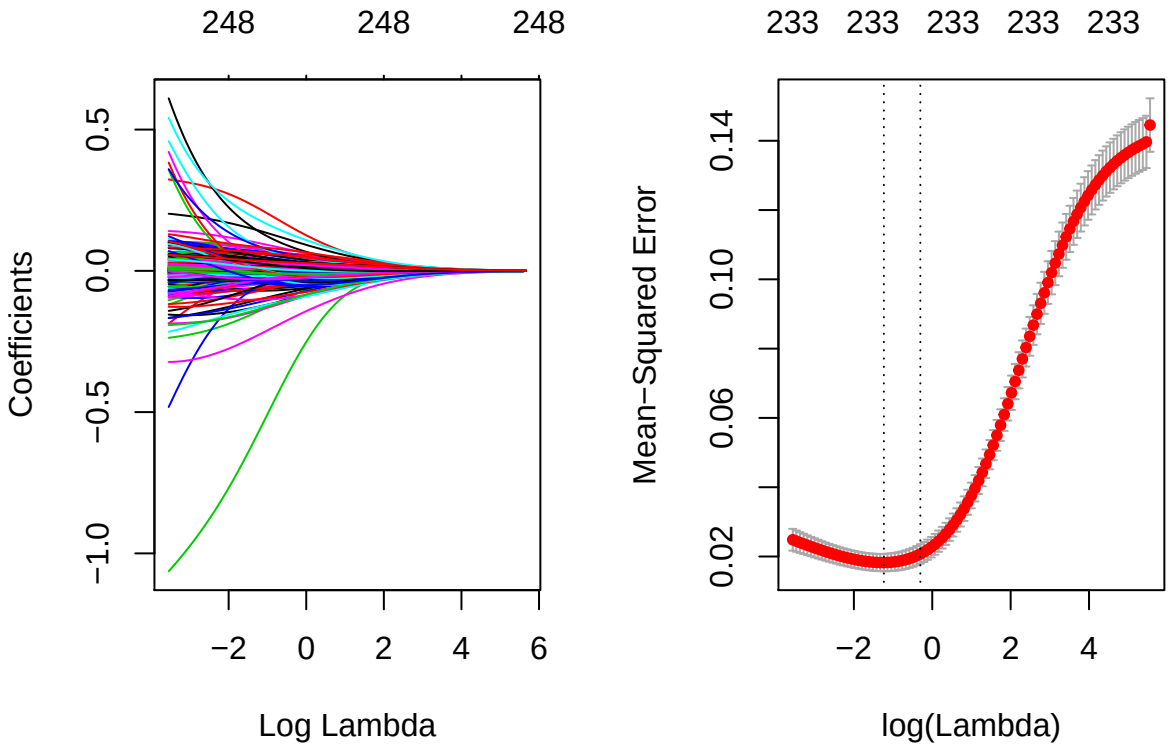
```r
par(mfrow=c(1,2))
plot(lasso_lm, xvar = "lambda", label = T)
plot(cvglm10)
```




```
# let calculate the mse of these three models
lasso_y <- predict.glmnet(object = lasso_lm, newx = x_test, s = cvglm10$lambda.1se)
mean((y_test - lasso_y)^2)
```

```
## [1] 0.02154153
```

```
par(mfrow=c(1,2))
plot(ridge_lm, xvar = "lambda", label = T)
plot(cvglm0)
```




```
ridge_y <- predict.glmnet(object = ridge_lm, newx = x_test, s = cvglm0$lambda.1se)
mean((y_test - ridge_y)^2)
```

```
## [1] 0.01963535
```

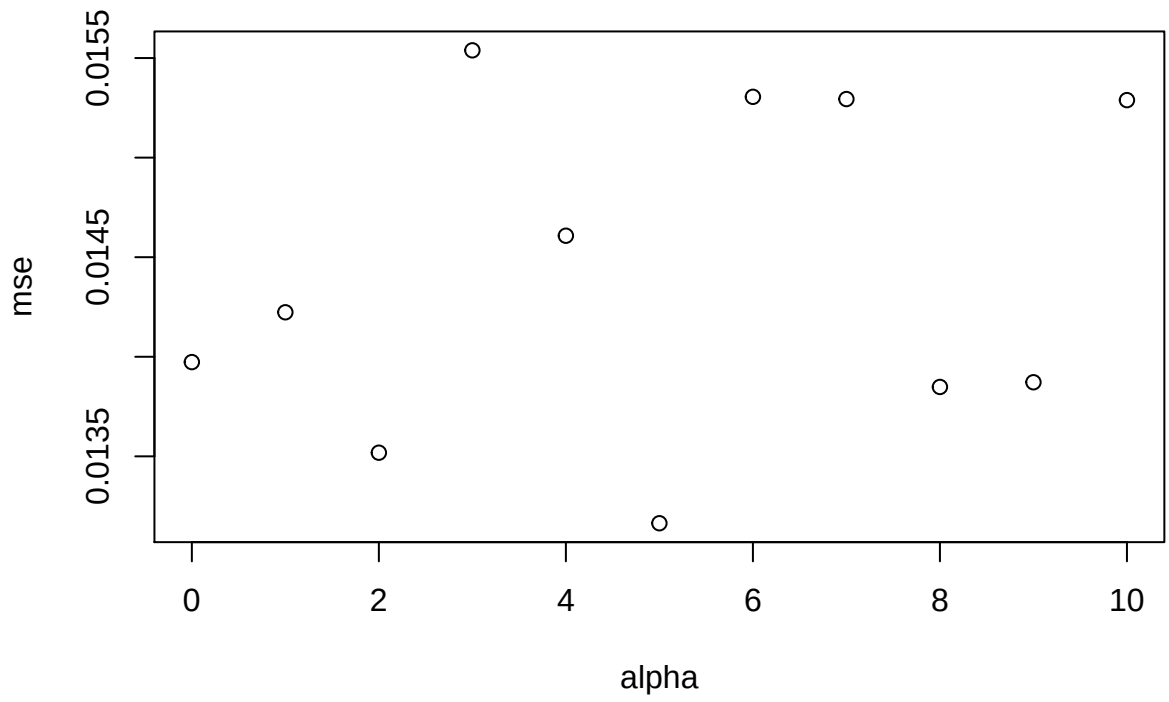
```
elnet_y <- predict.glmnet(object = elnet_lm, newx = x_test, s = cvglm5$lambda.1se)
mean((y_test - elnet_y)^2)
```

```
## [1] 0.01991286
```

```
# it seems that performance of ridge is best
# let's check if there is other elastic net that out-performs it.
```

```
# let choose the optimal lambda from elastic models
cvglm <- list(cvglm0,cvglm1,cvglm2,cvglm3,cvglm4,cvglm5,
           cvglm6,cvglm7,cvglm8,cvglm9,cvglm10
           )
mse <- NULL
for(i in c(1:11)){
  y_pre <- predict.cv.glmnet(object = cvglm[[i]],newx = x_test,
                             s = cvglm[[i]]$lambda.1se)
  mse <- c(mse, mean((y_test - y_pre)^2))
}
```

```
plot(x = c(0:10), y = mse, xlab = "alpha", ylab = "mse")
```




```
# print the model's index with optimal lambda
which.min(mse) - 1
```

```
## [1] 5
```

```
min(mse)
```

```
## [1] 0.0131641
```

```
# we got different model from hoemwork3_1
# use this one on test data
```

```
# use prediction on test data using our optimal model
pre_pri <- predict.cv.glmnet(object = cvglm5, newx = x_pre,
                             s = cvglm5$lambda.1se)
```

```
result <- data.frame(Id = raw_test$Id, SalePrice = exp(pre_pri))
names(result) <- c("Id", "SalePrice")
```

```
# write result back to csv
write.csv(x = result, file = "H:/kaggle/houseprice/data/submission_2.csv",
          row.names = FALSE)
```