homework3

syh

May 22, 2017

```
# in this version, we use combine common methods and mice to deal with nas
# to manipulate data set
```

```
# 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
```

```
# we can find that there are lots of columns related with basement with nas.
# these missing value can be due to not being exist
all_data[is.na(all_data$BsmtCond),
              c("BsmtExposure","BsmtQual","BsmtFinType2", "BsmtFullBath", "BsmtFinType1","BsmtHalfBath"
```

```
##      BsmtExposure BsmtQual BsmtFinType2 BsmtFullBath BsmtFinType1
## 18           <NA>     <NA>         <NA>            0         <NA>
## 40           <NA>     <NA>         <NA>            0         <NA>
## 91           <NA>     <NA>         <NA>            0         <NA>
## 103          <NA>     <NA>         <NA>            0         <NA>
```

```
## 157          <NA>     <NA>         <NA>            0         <NA>
## 183          <NA>     <NA>         <NA>            0         <NA>
## 260          <NA>     <NA>         <NA>            0         <NA>
## 343          <NA>     <NA>         <NA>            0         <NA>
## 363          <NA>     <NA>         <NA>            0         <NA>
## 372          <NA>     <NA>         <NA>            0         <NA>
## 393          <NA>     <NA>         <NA>            0         <NA>
## 521          <NA>     <NA>         <NA>            0         <NA>
## 533          <NA>     <NA>         <NA>            0         <NA>
## 534          <NA>     <NA>         <NA>            0         <NA>
## 554          <NA>     <NA>         <NA>            0         <NA>
## 647          <NA>     <NA>         <NA>            0         <NA>
## 706          <NA>     <NA>         <NA>            0         <NA>
## 737          <NA>     <NA>         <NA>            0         <NA>
## 750          <NA>     <NA>         <NA>            0         <NA>
## 779          <NA>     <NA>         <NA>            0         <NA>
## 869          <NA>     <NA>         <NA>            0         <NA>
## 895          <NA>     <NA>         <NA>            0         <NA>
## 898          <NA>     <NA>         <NA>            0         <NA>
## 985          <NA>     <NA>         <NA>            0         <NA>
## 1001         <NA>     <NA>         <NA>            0         <NA>
## 1012         <NA>     <NA>         <NA>            0         <NA>
## 1036         <NA>     <NA>         <NA>            0         <NA>
## 1046         <NA>     <NA>         <NA>            0         <NA>
## 1049         <NA>     <NA>         <NA>            0         <NA>
## 1050         <NA>     <NA>         <NA>            0         <NA>
## 1091         <NA>     <NA>         <NA>            0         <NA>
## 1180         <NA>     <NA>         <NA>            0         <NA>
## 1217         <NA>     <NA>         <NA>            0         <NA>
## 1219         <NA>     <NA>         <NA>            0         <NA>
## 1233         <NA>     <NA>         <NA>            0         <NA>
## 1322         <NA>     <NA>         <NA>            0         <NA>
## 1413         <NA>     <NA>         <NA>            0         <NA>
## 1586         <NA>     <NA>         <NA>            0         <NA>
## 1594         <NA>     <NA>         <NA>            0         <NA>
## 1730         <NA>     <NA>         <NA>            0         <NA>
## 1779         <NA>     <NA>         <NA>            0         <NA>
## 1815         <NA>     <NA>         <NA>            0         <NA>
## 1848         <NA>     <NA>         <NA>            0         <NA>
## 1849         <NA>     <NA>         <NA>            0         <NA>
## 1857         <NA>     <NA>         <NA>            0         <NA>
## 1858         <NA>     <NA>         <NA>            0         <NA>
## 1859         <NA>     <NA>         <NA>            0         <NA>
## 1861         <NA>     <NA>         <NA>            0         <NA>
## 1916         <NA>     <NA>         <NA>            0         <NA>
## 2041           Mn       Gd          Rec            1          GLQ
## 2051         <NA>     <NA>         <NA>            0         <NA>
## 2067         <NA>     <NA>         <NA>            0         <NA>
## 2069         <NA>     <NA>         <NA>            0         <NA>
## 2121         <NA>     <NA>         <NA>           NA         <NA>
## 2123         <NA>     <NA>         <NA>            0         <NA>
## 2186           No       TA          Unf            0          BLQ
## 2189         <NA>     <NA>         <NA>           NA         <NA>
## 2190         <NA>     <NA>         <NA>            0         <NA>
```

```
## 2191         <NA>     <NA>         <NA>            0         <NA>
## 2194         <NA>     <NA>         <NA>            0         <NA>
## 2217         <NA>     <NA>         <NA>            0         <NA>
## 2225         <NA>     <NA>         <NA>            0         <NA>
## 2388         <NA>     <NA>         <NA>            0         <NA>
## 2436         <NA>     <NA>         <NA>            0         <NA>
## 2453         <NA>     <NA>         <NA>            0         <NA>
## 2454         <NA>     <NA>         <NA>            0         <NA>
## 2491         <NA>     <NA>         <NA>            0         <NA>
## 2499         <NA>     <NA>         <NA>            0         <NA>
## 2525           Av       TA          Unf            0          ALQ
## 2548         <NA>     <NA>         <NA>            0         <NA>
## 2553         <NA>     <NA>         <NA>            0         <NA>
## 2565         <NA>     <NA>         <NA>            0         <NA>
## 2579         <NA>     <NA>         <NA>            0         <NA>
## 2600         <NA>     <NA>         <NA>            0         <NA>
## 2703         <NA>     <NA>         <NA>            0         <NA>
## 2764         <NA>     <NA>         <NA>            0         <NA>
## 2767         <NA>     <NA>         <NA>            0         <NA>
## 2804         <NA>     <NA>         <NA>            0         <NA>
## 2805         <NA>     <NA>         <NA>            0         <NA>
## 2825         <NA>     <NA>         <NA>            0         <NA>
## 2892         <NA>     <NA>         <NA>            0         <NA>
## 2905         <NA>     <NA>         <NA>            0         <NA>
##      BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 18              0          0          0         0           0
## 40              0          0          0         0           0
## 91              0          0          0         0           0
## 103             0          0          0         0           0
## 157             0          0          0         0           0
## 183             0          0          0         0           0
## 260             0          0          0         0           0
## 343             0          0          0         0           0
## 363             0          0          0         0           0
## 372             0          0          0         0           0
## 393             0          0          0         0           0
## 521             0          0          0         0           0
## 533             0          0          0         0           0
## 534             0          0          0         0           0
## 554             0          0          0         0           0
## 647             0          0          0         0           0
## 706             0          0          0         0           0
## 737             0          0          0         0           0
## 750             0          0          0         0           0
## 779             0          0          0         0           0
## 869             0          0          0         0           0
## 895             0          0          0         0           0
## 898             0          0          0         0           0
## 985             0          0          0         0           0
## 1001            0          0          0         0           0
## 1012            0          0          0         0           0
## 1036            0          0          0         0           0
## 1046            0          0          0         0           0
## 1049            0          0          0         0           0
```

```
## 1050    0    0    0    0    0
## 1091    0    0    0    0    0
## 1180    0    0    0    0    0
## 1217    0    0    0    0    0
## 1219    0    0    0    0    0
## 1233    0    0    0    0    0
## 1322    0    0    0    0    0
## 1413    0    0    0    0    0
## 1586    0    0    0    0    0
## 1594    0    0    0    0    0
## 1730    0    0    0    0    0
## 1779    0    0    0    0    0
## 1815    0    0    0    0    0
## 1848    0    0    0    0    0
## 1849    0    0    0    0    0
## 1857    0    0    0    0    0
## 1858    0    0    0    0    0
## 1859    0    0    0    0    0
## 1861    0    0    0    0    0
## 1916    0    0    0    0    0
## 2041    0 1044  382    0 1426
## 2051    0    0    0    0    0
## 2067    0    0    0    0    0
## 2069    0    0    0    0    0
## 2121   NA   NA   NA   NA   NA
## 2123    0    0    0    0    0
## 2186    1 1033    0   94 1127
## 2189   NA    0    0    0    0
## 2190    0    0    0    0    0
## 2191    0    0    0    0    0
## 2194    0    0    0    0    0
## 2217    0    0    0    0    0
## 2225    0    0    0    0    0
## 2388    0    0    0    0    0
## 2436    0    0    0    0    0
## 2453    0    0    0    0    0
## 2454    0    0    0    0    0
## 2491    0    0    0    0    0
## 2499    0    0    0    0    0
## 2525    0  755    0  240  995
## 2548    0    0    0    0    0
## 2553    0    0    0    0    0
## 2565    0    0    0    0    0
## 2579    0    0    0    0    0
## 2600    0    0    0    0    0
## 2703    0    0    0    0    0
## 2764    0    0    0    0    0
## 2767    0    0    0    0    0
## 2804    0    0    0    0    0
## 2805    0    0    0    0    0
## 2825    0    0    0    0    0
## 2892    0    0    0    0    0
## 2905    0    0    0    0    0
```

```
# there is no basement for these houses
```

```
# we can create a another type value, "None" or 0 for these NAs.

all_data[is.na(all_data$BsmtCond), "BsmtCond"] <- "None"
all_data[is.na(all_data$BsmtExposure), "BsmtExposure"] <- "None"
all_data[is.na(all_data$BsmtQual), "BsmtQual"] <- "None"
all_data[is.na(all_data$BsmtFinType2), "BsmtFinType2"] <- "None"
all_data[is.na(all_data$BsmtFinType1), "BsmtFinType1"] <- "None"

all_data[is.na(all_data$BsmtHalfBath), "BsmtHalfBath"] <- 0
all_data[is.na(all_data$BsmtFinSF1), "BsmtFinSF1"] <- 0
all_data[is.na(all_data$BsmtFinSF2), "BsmtFinSF2"] <- 0
all_data[is.na(all_data$BsmtUnfSF), "BsmtUnfSF"] <- 0
all_data[is.na(all_data$TotalBsmtSF), "TotalBsmtSF"] <- 0
all_data[is.na(all_data$BsmtFullBath), "BsmtFullBath"] <- 0
```

```
# let's deal with MasVnrType, MasVnrArea
all_data[is.na(all_data$MasVnrType),"MasVnrArea"]
```

```
##  [1]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## [18]  NA  NA  NA 198  NA  NA  NA
```

```
# the same reason as basement
```

```
table(all_data$MasVnrType)
```

```
##
##  BrkCmn BrkFace    None   Stone
##      25     879    1742     249
```

```
all_data[is.na(all_data$MasVnrType), "MasVnrType"] <- "None"
all_data[is.na(all_data$MasVnrArea), "MasVnrArea"] <- 0
```

```
# GarageCars, GarageSize,
# there are many features about garage, but only one of them is missing
# it's due to data transportation, probably

# it's same reason for Kitchen
all_data[is.na(all_data$KitchenQual),"KitchenAbvGr"]
```

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

```
# same with features about exterior
all_data[is.na(all_data$Exterior1st), c("ExterCond","Exterior2nd","ExterQual")]
```

```
##      ExterCond Exterior2nd ExterQual
## 2152        TA        <NA>        TA
```

```
# so we plan to use mice to impute these values from other features
```

```
# first we should 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)
```

```
library(mice)
```

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

```
##
##  iter imp variable
##   1   1  MSZoning  Utilities  Exterior1st  Exterior2nd  Electrical  KitchenQual  Functional  GarageCa
##   2   1  MSZoning  Utilities  Exterior1st  Exterior2nd  Electrical  KitchenQual  Functional  GarageCa
##   3   1  MSZoning  Utilities  Exterior1st  Exterior2nd  Electrical  KitchenQual  Functional  GarageCa
##   4   1  MSZoning  Utilities  Exterior1st  Exterior2nd  Electrical  KitchenQual  Functional  GarageCa
##   5   1  MSZoning  Utilities  Exterior1st  Exterior2nd  Electrical  KitchenQual  Functional  GarageCa
```

```
real_all_data <- complete(im_all_data)
```

```
# 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
```

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

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

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

```
# 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

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

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

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

```
# 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
```

```
summary(simple_lm)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = real_all_data[c(1:1460), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68198 -0.04714  0.00157  0.05206  0.68198
##
## Coefficients: (7 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.072e+00  4.782e+00   1.688 0.091658 .
## MSSubClass30   -7.849e-02  2.171e-02  -3.615 0.000313 ***
## MSSubClass40   -6.110e-02  6.702e-02  -0.912 0.362112
## MSSubClass45   -2.403e-01  1.030e-01  -2.332 0.019867 *
## MSSubClass50   -1.496e-02  3.989e-02  -0.375 0.707638
## MSSubClass60   -2.891e-02  3.511e-02  -0.823 0.410410
## MSSubClass70    1.284e-02  3.815e-02   0.337 0.736500
## MSSubClass75   -6.461e-02  7.040e-02  -0.918 0.358907
## MSSubClass80   -5.363e-02  5.783e-02  -0.927 0.353922
## MSSubClass85   -9.106e-03  4.824e-02  -0.189 0.850294
## MSSubClass90   -2.246e-02  3.265e-02  -0.688 0.491715
## MSSubClass120  -4.050e-02  6.727e-02  -0.602 0.547266
## MSSubClass160  -1.325e-01  8.025e-02  -1.652 0.098885 .
## MSSubClass180  -8.283e-02  8.906e-02  -0.930 0.352525
## MSSubClass190   2.693e-02  1.283e-01   0.210 0.833729
## MSZoning2       4.815e-01  5.601e-02   8.597  < 2e-16 ***
## MSZoning3       4.325e-01  5.587e-02   7.741 2.07e-14 ***
## MSZoning4       4.458e-01  4.828e-02   9.233  < 2e-16 ***
## MSZoning5       3.927e-01  4.521e-02   8.685  < 2e-16 ***
## LotArea         2.895e-06  4.947e-07   5.852 6.27e-09 ***
## Street2         1.352e-01  5.705e-02   2.369 0.017991 *
## LotShape2       2.751e-02  1.897e-02   1.451 0.147106
## LotShape3       3.070e-02  3.974e-02   0.773 0.439911
## LotShape4       7.633e-03  7.361e-03   1.037 0.299974
```

```
## LandContour2    3.250e-02  2.379e-02   1.366 0.172224
## LandContour3   -3.587e-04  2.991e-02  -0.012 0.990433
## LandContour4    2.849e-02  1.693e-02   1.683 0.092722 .
## Utilities2     -2.603e-01  1.282e-01  -2.030 0.042533 *
## LotConfig2      2.909e-02  1.455e-02   1.999 0.045821 *
## LotConfig3     -3.379e-02  1.818e-02  -1.858 0.063354 .
## LotConfig4     -8.176e-02  5.742e-02  -1.424 0.154746
## LotConfig5     -1.431e-02  7.996e-03  -1.789 0.073808 .
## LandSlope2      3.566e-02  1.836e-02   1.942 0.052364 .
## LandSlope3     -1.930e-01  5.184e-02  -3.722 0.000206 ***
## Neighborhood2   2.683e-02  8.967e-02   0.299 0.764842
## Neighborhood3  -1.936e-03  5.372e-02  -0.036 0.971267
## Neighborhood4   4.006e-02  4.367e-02   0.917 0.359130
## Neighborhood5   2.942e-02  4.264e-02   0.690 0.490383
## Neighborhood6  -8.113e-03  3.306e-02  -0.245 0.806197
## Neighborhood7   1.116e-01  3.975e-02   2.808 0.005062 **
## Neighborhood8  -6.701e-02  3.694e-02  -1.814 0.069923 .
## Neighborhood9   4.168e-03  3.540e-02   0.118 0.906306
## Neighborhood10 -2.261e-03  4.946e-02  -0.046 0.963548
## Neighborhood11 -1.220e-01  5.589e-02  -2.184 0.029179 *
## Neighborhood12 -5.213e-02  3.747e-02  -1.392 0.164322
## Neighborhood13 -2.895e-02  3.596e-02  -0.805 0.421043
## Neighborhood14  4.623e-02  3.881e-02   1.191 0.233803
## Neighborhood15  1.078e-03  6.356e-02   0.017 0.986466
## Neighborhood16  8.427e-02  3.391e-02   2.485 0.013088 *
## Neighborhood17 -2.832e-02  3.685e-02  -0.769 0.442316
## Neighborhood18 -2.879e-02  4.422e-02  -0.651 0.515150
## Neighborhood19 -1.519e-02  3.728e-02  -0.408 0.683688
## Neighborhood20  2.308e-03  3.579e-02   0.064 0.948584
## Neighborhood21  3.301e-02  4.145e-02   0.796 0.425989
## Neighborhood22  1.314e-01  3.854e-02   3.409 0.000674 ***
## Neighborhood23  9.870e-03  4.466e-02   0.221 0.825110
## Neighborhood24  1.599e-02  3.755e-02   0.426 0.670418
## Neighborhood25  4.749e-02  4.797e-02   0.990 0.322432
## Condition12     2.237e-02  2.267e-02   0.987 0.323910
## Condition13     7.572e-02  1.868e-02   4.054 5.35e-05 ***
## Condition14     5.769e-02  4.520e-02   1.276 0.202035
## Condition15     8.241e-02  3.375e-02   2.442 0.014754 *
## Condition16    -4.000e-02  4.113e-02  -0.973 0.330937
## Condition17     3.116e-02  3.092e-02   1.008 0.313642
## Condition18     6.417e-03  7.986e-02   0.080 0.935971
## Condition19     6.043e-02  5.779e-02   1.046 0.295908
## Condition22     1.992e-01  1.133e-01   1.758 0.078957 .
## Condition23     1.686e-01  1.005e-01   1.678 0.093528 .
## Condition24     3.258e-01  1.759e-01   1.853 0.064197 .
## Condition25    -6.849e-01  1.337e-01  -5.124 3.48e-07 ***
## Condition26    -4.493e-01  2.259e-01  -1.989 0.046956 *
## Condition27     7.833e-02  1.490e-01   0.526 0.599216
## Condition28     1.332e-01  1.306e-01   1.020 0.307922
## BldgType2      -5.225e-02  1.260e-01  -0.415 0.678442
## BldgType3              NA         NA      NA       NA
## BldgType4      -2.794e-02  7.160e-02  -0.390 0.696420
## BldgType5       3.642e-03  6.812e-02   0.053 0.957368
## HouseStyle2     2.051e-01  1.028e-01   1.995 0.046232 *
```

```
## HouseStyle3    -2.226e-02  4.015e-02  -0.554 0.579376
## HouseStyle4     3.004e-03  7.792e-02   0.039 0.969253
## HouseStyle5     9.885e-02  7.435e-02   1.329 0.183937
## HouseStyle6    -1.004e-02  3.675e-02  -0.273 0.784695
## HouseStyle7    -2.611e-02  5.360e-02  -0.487 0.626187
## HouseStyle8     3.342e-02  6.318e-02   0.529 0.596919
## OverallQual2    4.768e-01  1.429e-01   3.336 0.000875 ***
## OverallQual3    5.629e-01  1.312e-01   4.290 1.93e-05 ***
## OverallQual4    6.030e-01  1.298e-01   4.645 3.77e-06 ***
## OverallQual5    6.455e-01  1.305e-01   4.946 8.64e-07 ***
## OverallQual6    6.772e-01  1.309e-01   5.175 2.67e-07 ***
## OverallQual7    7.131e-01  1.309e-01   5.447 6.22e-08 ***
## OverallQual8    7.674e-01  1.316e-01   5.832 7.02e-09 ***
## OverallQual9    8.379e-01  1.340e-01   6.255 5.52e-10 ***
## OverallQual10   8.571e-01  1.374e-01   6.238 6.10e-10 ***
## OverallCond2   -5.203e-01  2.131e-01  -2.441 0.014783 *
## OverallCond3   -6.556e-01  2.251e-01  -2.913 0.003649 **
## OverallCond4   -5.845e-01  2.266e-01  -2.580 0.009998 **
## OverallCond5   -5.387e-01  2.263e-01  -2.381 0.017427 *
## OverallCond6   -5.006e-01  2.263e-01  -2.212 0.027144 *
## OverallCond7   -4.658e-01  2.263e-01  -2.058 0.039764 *
## OverallCond8   -4.547e-01  2.263e-01  -2.009 0.044759 *
## OverallCond9   -4.009e-01  2.282e-01  -1.757 0.079231 .
## YearBuilt       1.668e-03  3.775e-04   4.420 1.08e-05 ***
## YearRemodAdd    8.202e-04  2.519e-04   3.256 0.001162 **
## RoofStyle2     -1.485e-02  8.312e-02  -0.179 0.858233
## RoofStyle3     -2.428e-02  9.095e-02  -0.267 0.789558
## RoofStyle4     -1.457e-02  8.329e-02  -0.175 0.861192
## RoofStyle5      4.879e-02  9.685e-02   0.504 0.614465
## RoofStyle6      5.093e-01  1.698e-01   3.000 0.002755 **
## RoofMatl2       2.560e+00  1.507e-01  16.985  < 2e-16 ***
## RoofMatl3       2.967e+00  2.173e-01  13.654  < 2e-16 ***
## RoofMatl4       2.804e+00  2.126e-01  13.184  < 2e-16 ***
## RoofMatl5       2.550e+00  1.905e-01  13.387  < 2e-16 ***
## RoofMatl6       2.585e+00  1.725e-01  14.984  < 2e-16 ***
## RoofMatl7       2.480e+00  1.665e-01  14.896  < 2e-16 ***
## RoofMatl8       2.649e+00  1.550e-01  17.097  < 2e-16 ***
## Exterior1st2    6.045e-02  1.504e-01   0.402 0.687788
## Exterior1st3   -2.011e-01  1.302e-01  -1.545 0.122721
## Exterior1st4    1.292e-01  5.838e-02   2.212 0.027120 *
## Exterior1st5   -1.120e-01  1.255e-01  -0.892 0.372342
## Exterior1st6   -2.771e-02  8.777e-02  -0.316 0.752269
## Exterior1st7    4.469e-02  5.916e-02   0.755 0.450126
## Exterior1st8    1.801e-02  1.256e-01   0.143 0.885971
## Exterior1st9    8.364e-02  6.752e-02   1.239 0.215702
## Exterior1st10   4.388e-02  5.828e-02   0.753 0.451637
## Exterior1st11   1.270e-01  1.106e-01   1.148 0.251057
## Exterior1st12   8.130e-02  6.577e-02   1.236 0.216644
## Exterior1st13   5.188e-02  6.153e-02   0.843 0.399264
## Exterior1st14   1.779e-02  5.679e-02   0.313 0.754111
## Exterior1st15   6.764e-02  6.122e-02   1.105 0.269430
## Exterior2nd2   -3.247e-02  1.011e-01  -0.321 0.748149
## Exterior2nd3    3.143e-02  9.353e-02   0.336 0.736918
## Exterior2nd4   -7.679e-02  6.021e-02  -1.275 0.202420
```

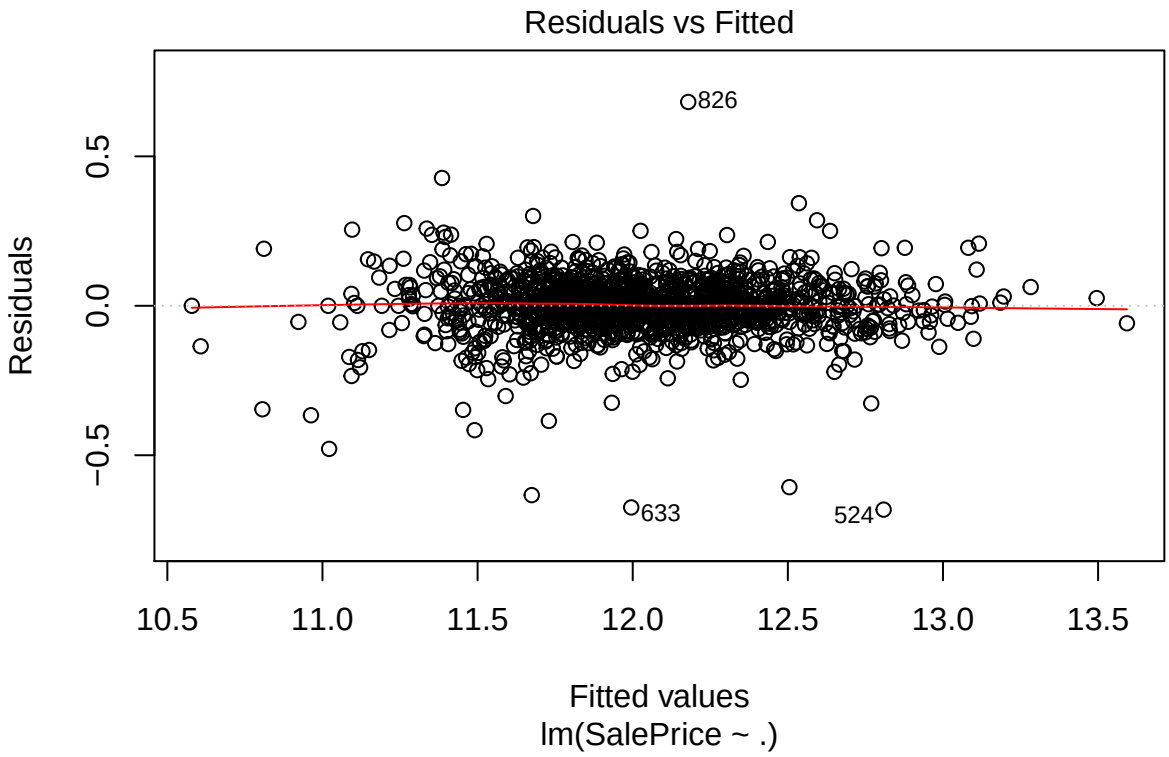
```
## Exterior2nd5           NA         NA      NA       NA
## Exterior2nd6    6.853e-02  8.575e-02   0.799 0.424349
## Exterior2nd7   -3.464e-02  5.680e-02  -0.610 0.542138
## Exterior2nd8   -1.989e-02  6.505e-02  -0.306 0.759881
## Exterior2nd9   -4.720e-02  6.566e-02  -0.719 0.472427
## Exterior2nd10  -1.223e-01  1.242e-01  -0.985 0.324925
## Exterior2nd11  -3.083e-02  5.515e-02  -0.559 0.576265
## Exterior2nd12  -1.006e-01  7.841e-02  -1.283 0.199770
## Exterior2nd13  -4.806e-02  6.243e-02  -0.770 0.441590
## Exterior2nd14  -2.110e-02  5.929e-02  -0.356 0.721950
## Exterior2nd15   1.505e-04  5.475e-02   0.003 0.997807
## Exterior2nd16  -5.420e-02  5.679e-02  -0.954 0.340079
## MasVnrType2     4.621e-02  3.084e-02   1.498 0.134332
## MasVnrType3     3.783e-02  3.112e-02   1.216 0.224290
## MasVnrType4     5.537e-02  3.266e-02   1.695 0.090259 .
## MasVnrArea      2.128e-05  2.635e-05   0.808 0.419468
## ExterQual2      7.032e-02  5.618e-02   1.252 0.210907
## ExterQual3      7.919e-03  2.357e-02   0.336 0.736924
## ExterQual4      4.407e-03  2.560e-02   0.172 0.863365
## ExterCond2     -7.048e-02  8.536e-02  -0.826 0.409172
## ExterCond3     -4.549e-02  8.153e-02  -0.558 0.576945
## ExterCond4     -1.847e-01  1.660e-01  -1.113 0.266088
## ExterCond5     -3.395e-02  8.174e-02  -0.415 0.678006
## Foundation2     9.318e-03  1.461e-02   0.638 0.523836
## Foundation3     2.804e-02  1.566e-02   1.791 0.073619 .
## Foundation4    -1.106e-02  4.565e-02  -0.242 0.808651
## Foundation5     1.105e-01  4.958e-02   2.230 0.025956 *
## Foundation6    -1.179e-01  6.616e-02  -1.781 0.075116 .
## BsmtQual2      -5.199e-03  2.951e-02  -0.176 0.860172
## BsmtQual3      -1.966e-02  1.547e-02  -1.271 0.203918
## BsmtQual4       1.225e-01  1.683e-01   0.728 0.466972
## BsmtQual5      -2.205e-02  1.897e-02  -1.163 0.245190
## BsmtCond2       2.874e-02  2.411e-02   1.192 0.233453
## BsmtCond3              NA         NA      NA       NA
## BsmtCond4      -1.277e-02  1.481e-01  -0.086 0.931314
## BsmtCond5       3.216e-02  1.944e-02   1.654 0.098333 .
## BsmtExposure2   3.028e-02  1.391e-02   2.177 0.029650 *
## BsmtExposure3  -1.103e-03  1.383e-02  -0.080 0.936434
## BsmtExposure4  -8.758e-03  1.001e-02  -0.875 0.381759
## BsmtExposure5  -3.596e-02  1.056e-01  -0.340 0.733588
## BsmtFinType12  -4.356e-03  1.256e-02  -0.347 0.728787
## BsmtFinType13   9.386e-03  1.151e-02   0.816 0.414819
## BsmtFinType14  -2.766e-02  1.696e-02  -1.631 0.103094
## BsmtFinType15          NA         NA      NA       NA
## BsmtFinType16  -2.807e-03  1.357e-02  -0.207 0.836144
## BsmtFinType17  -1.617e-02  1.337e-02  -1.210 0.226681
## BsmtFinSF1      1.362e-04  2.431e-05   5.600 2.65e-08 ***
## BsmtFinType22  -7.647e-02  3.436e-02  -2.226 0.026226 *
## BsmtFinType23  -1.725e-02  4.263e-02  -0.405 0.685803
## BsmtFinType24  -3.639e-02  3.371e-02  -1.079 0.280626
## BsmtFinType25  -1.193e-01  1.150e-01  -1.038 0.299547
## BsmtFinType26  -2.957e-02  3.228e-02  -0.916 0.359796
## BsmtFinType27  -2.161e-02  3.439e-02  -0.628 0.529853
## BsmtFinSF2      1.306e-04  4.136e-05   3.156 0.001636 **
```

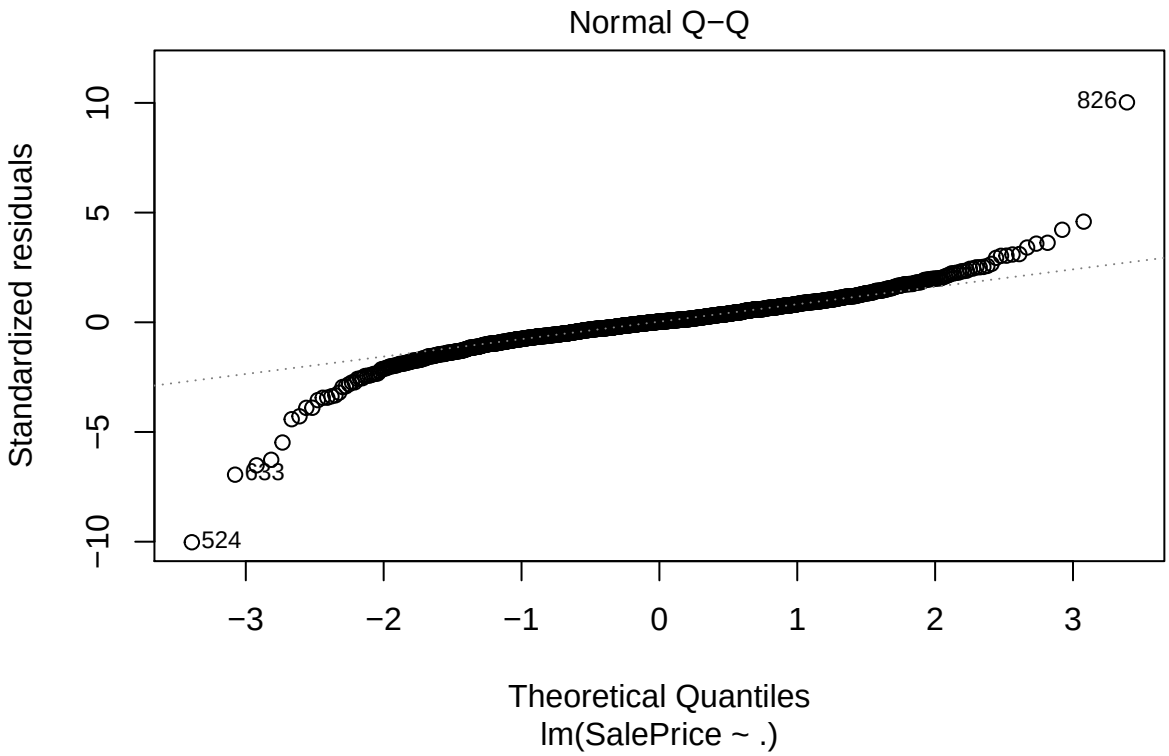
```
## BsmtUnfSF       8.164e-05  2.210e-05   3.694 0.000231 ***
## TotalBsmtSF            NA         NA      NA       NA
## Heating2        1.774e-01  1.176e-01   1.508 0.131696
## Heating3        2.151e-01  1.210e-01   1.778 0.075667 .
## Heating4       -1.367e-02  1.268e-01  -0.108 0.914173
## Heating5        1.300e-01  1.444e-01   0.900 0.368118
## Heating6        2.357e-01  1.345e-01   1.753 0.079888 .
## HeatingQC2     -2.432e-02  2.170e-02  -1.120 0.262773
## HeatingQC3     -2.028e-02  9.443e-03  -2.147 0.031975 *
## HeatingQC4     -1.086e-01  1.223e-01  -0.888 0.374739
## HeatingQC5     -3.200e-02  9.388e-03  -3.409 0.000674 ***
## CentralAir2     5.676e-02  1.778e-02   3.192 0.001448 **
## Electrical2     1.283e-03  2.717e-02   0.047 0.962340
## Electrical3    -5.654e-02  7.842e-02  -0.721 0.471053
## Electrical4            NA         NA      NA       NA
## Electrical5    -2.714e-02  1.367e-02  -1.985 0.047366 *
## X1stFlrSF       2.265e-04  2.540e-05   8.920  < 2e-16 ***
## X2ndFlrSF       2.174e-04  2.391e-05   9.089  < 2e-16 ***
## LowQualFinSF    1.658e-04  8.468e-05   1.958 0.050441 .
## GrLivArea              NA         NA      NA       NA
## BsmtFullBath    2.316e-02  9.008e-03   2.571 0.010265 *
## BsmtHalfBath    4.938e-03  1.381e-02   0.358 0.720634
## FullBath        3.008e-02  1.006e-02   2.991 0.002839 **
## HalfBath        2.711e-02  9.522e-03   2.847 0.004494 **
## BedroomAbvGr    1.407e-03  6.350e-03   0.221 0.824745
## KitchenAbvGr   -6.049e-02  2.754e-02  -2.197 0.028246 *
## KitchenQual2   -4.083e-02  2.888e-02  -1.414 0.157725
## KitchenQual3   -5.540e-02  1.623e-02  -3.413 0.000664 ***
## KitchenQual4   -5.846e-02  1.814e-02  -3.223 0.001303 **
## TotRmsAbvGrd    2.884e-03  4.338e-03   0.665 0.506303
## Functional2    -2.338e-01  6.697e-02  -3.492 0.000498 ***
## Functional3     1.415e-03  3.970e-02   0.036 0.971569
## Functional4    -1.474e-02  4.024e-02  -0.366 0.714298
## Functional5    -9.870e-02  4.839e-02  -2.040 0.041613 *
## Functional6    -2.924e-01  1.258e-01  -2.325 0.020254 *
## Functional7     3.008e-02  3.501e-02   0.859 0.390473
## Fireplaces      2.552e-02  6.069e-03   4.206 2.80e-05 ***
## GarageCars      2.388e-02  9.886e-03   2.416 0.015837 *
## GarageArea      1.211e-04  3.412e-05   3.549 0.000401 ***
## PavedDrive2     2.890e-03  2.486e-02   0.116 0.907487
## PavedDrive3     2.089e-02  1.556e-02   1.343 0.179491
## WoodDeckSF      8.815e-05  2.636e-05   3.344 0.000850 ***
## OpenPorchSF     7.013e-05  5.265e-05   1.332 0.183089
## EnclosedPorch   1.196e-04  5.693e-05   2.101 0.035804 *
## X3SsnPorch      1.702e-04  1.013e-04   1.681 0.092961 .
## ScreenPorch     2.684e-04  5.510e-05   4.872 1.25e-06 ***
## PoolArea        1.858e-04  8.302e-05   2.238 0.025392 *
## MiscVal         5.635e-07  6.436e-06   0.088 0.930252
## MoSold2        -1.704e-03  2.167e-02  -0.079 0.937340
## MoSold3        -4.070e-03  1.886e-02  -0.216 0.829206
## MoSold4         6.941e-03  1.805e-02   0.385 0.700611
## MoSold5         9.901e-03  1.723e-02   0.575 0.565589
## MoSold6         1.629e-02  1.694e-02   0.962 0.336388
## MoSold7         6.741e-03  1.716e-02   0.393 0.694469
```

```
## MoSold8        -3.377e-03  1.815e-02  -0.186 0.852435
## MoSold9        -4.771e-03  2.067e-02  -0.231 0.817496
## MoSold10       -1.420e-02  1.967e-02  -0.722 0.470493
## MoSold11       -5.661e-03  1.983e-02  -0.285 0.775330
## MoSold12       -1.163e-03  2.140e-02  -0.054 0.956670
## YrSold         -2.682e-03  2.347e-03  -1.143 0.253298
## SaleType2       7.304e-02  7.995e-02   0.913 0.361162
## SaleType3       1.202e-01  4.559e-02   2.636 0.008494 **
## SaleType4      -2.775e-02  5.233e-02  -0.530 0.595971
## SaleType5       1.444e-02  5.663e-02   0.255 0.798793
## SaleType6       5.819e-02  5.912e-02   0.984 0.325179
## SaleType7       8.779e-02  7.089e-02   1.238 0.215799
## SaleType8       7.422e-02  6.634e-02   1.119 0.263427
## SaleType9      -1.775e-02  1.915e-02  -0.927 0.353938
## SaleCondition2  9.489e-02  6.592e-02   1.439 0.150272
## SaleCondition3  6.876e-02  3.963e-02   1.735 0.082988 .
## SaleCondition4  1.724e-02  2.775e-02   0.622 0.534369
## SaleCondition5  6.493e-02  1.314e-02   4.941 8.87e-07 ***
## SaleCondition6  6.126e-03  6.809e-02   0.090 0.928330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1031 on 1208 degrees of freedom
## Multiple R-squared:  0.9448, Adjusted R-squared:  0.9334
## F-statistic: 82.41 on 251 and 1208 DF,  p-value: < 2.2e-16
```

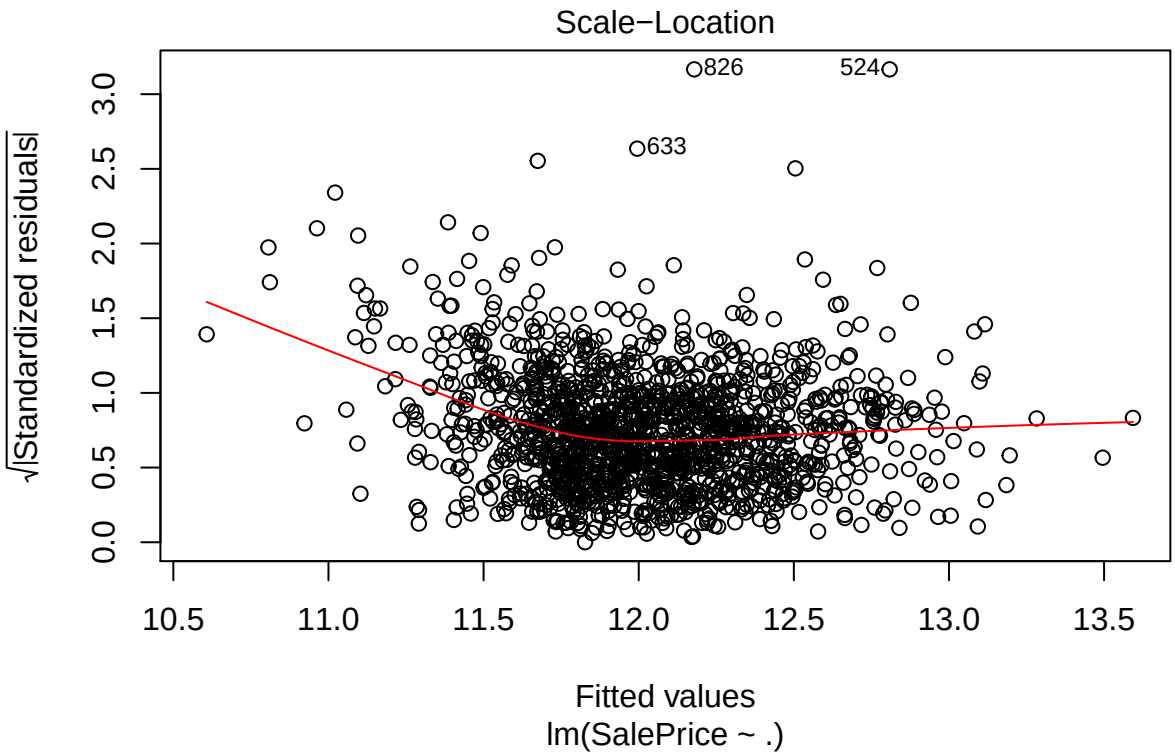
```
# residual plot
# 1st: residual is unbais and homoscedastic,
# 2nd: basically, residual is subject to normal distribution which means variacne of error is constant.
# 4th: we know some outlier: 89, 826, 524 high leaverage and high residual
plot(simple_lm)
```

```
## Warning: not plotting observations with leverage one:
##   121, 251, 272, 326, 333, 376, 399, 534, 584, 596, 667, 822, 945, 949, 1012, 1188, 1231, 1271, 1276
```

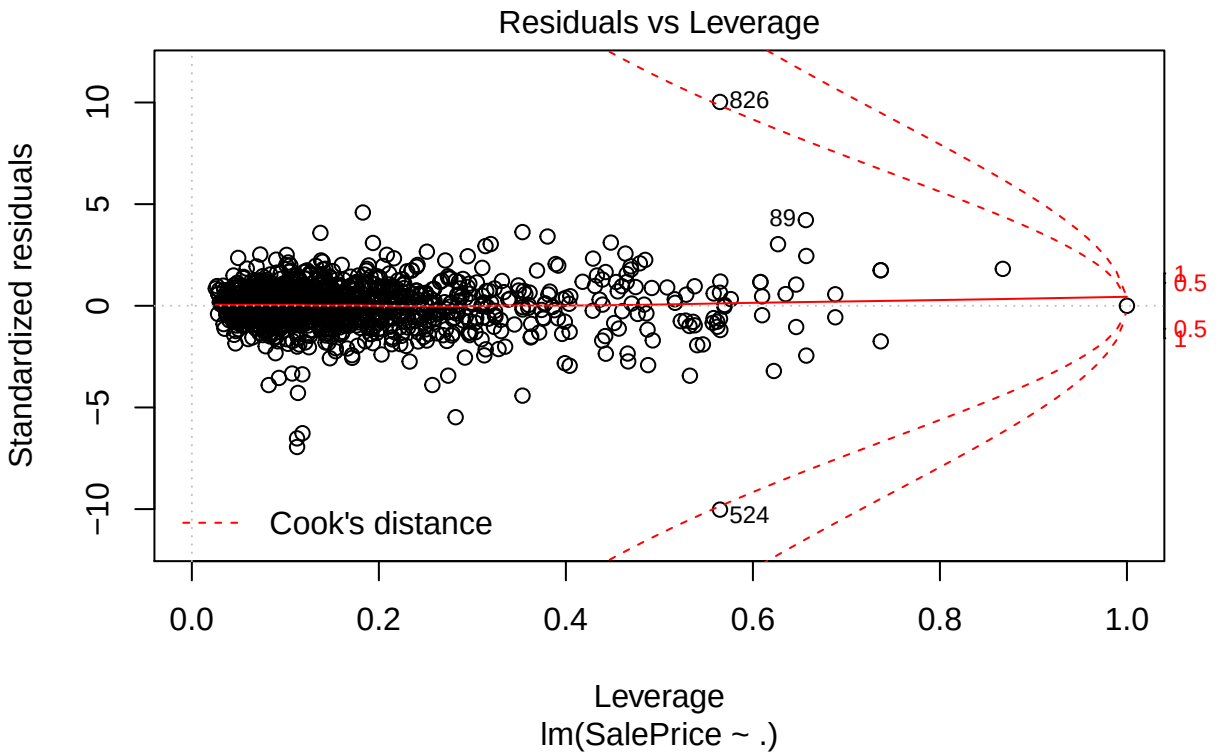

Residuals vs Fitted
826
633
524
Residuals
0.5
0.0
−0.5
10.5
11.0
11.5
12.0
12.5
13.0
13.5
Fitted values
lm(SalePrice ~ .)




```
## Warning: not plotting observations with leverage one:
##   121, 251, 272, 326, 333, 376, 399, 534, 584, 596, 667, 822, 945, 949, 1012, 1188, 1231, 1271, 1276
```




```
## 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:  0.9334, it seems it's pretty good.
# F-statistic: 82.41, 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
```

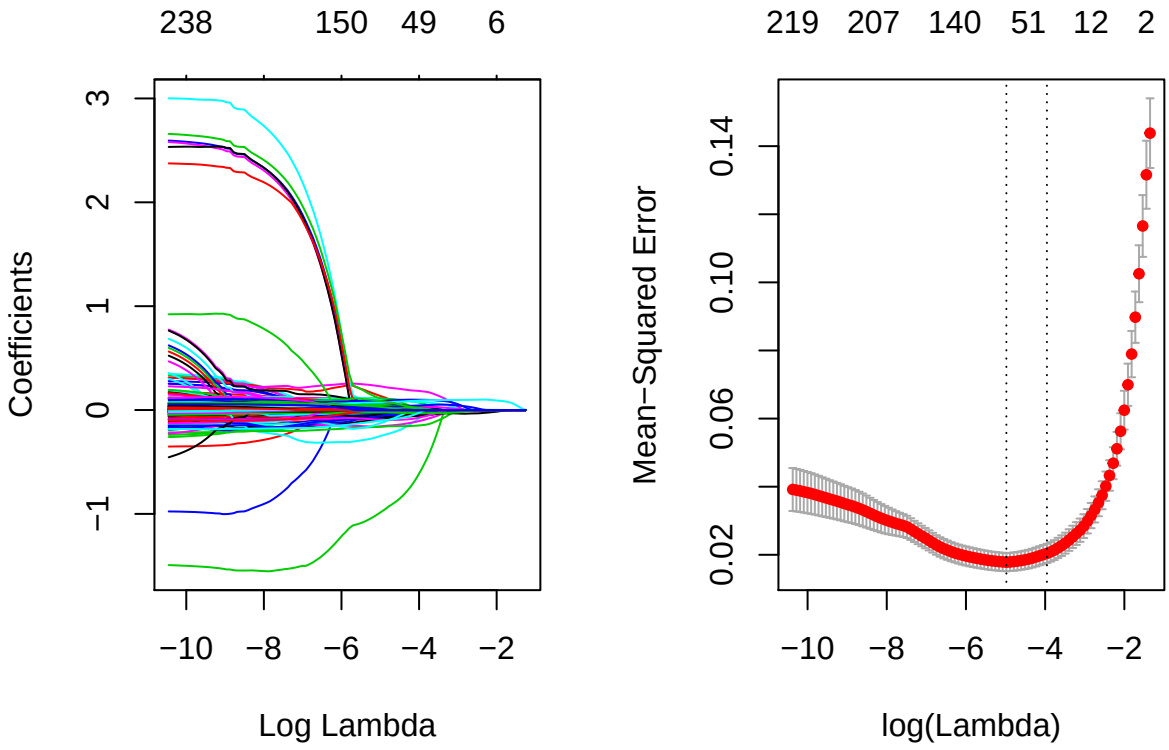
```
## Loading required package: foreach
```

```
## 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"))
}
```

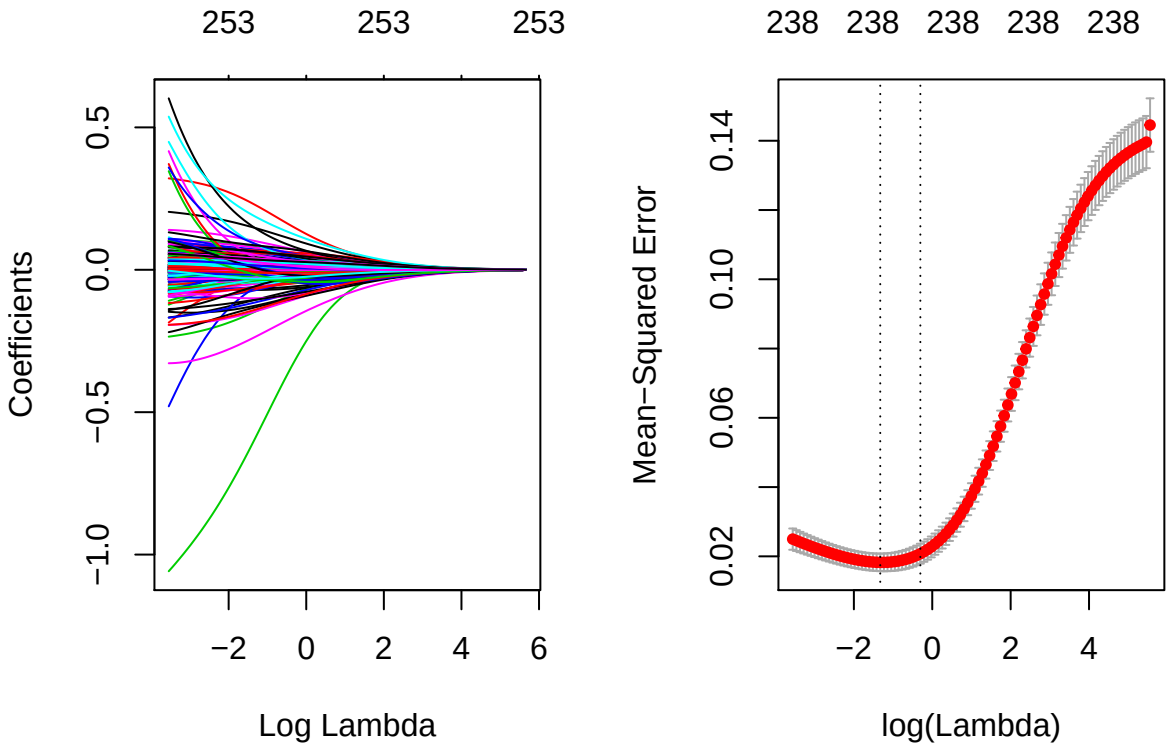
```r
par(mfrow=c(1,2))
plot(lasso_lm, xvar = "lambda", label = T)
plot(cvglm10)
```




```
# 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] 0.02164823
```

```
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] 0.01963155
```

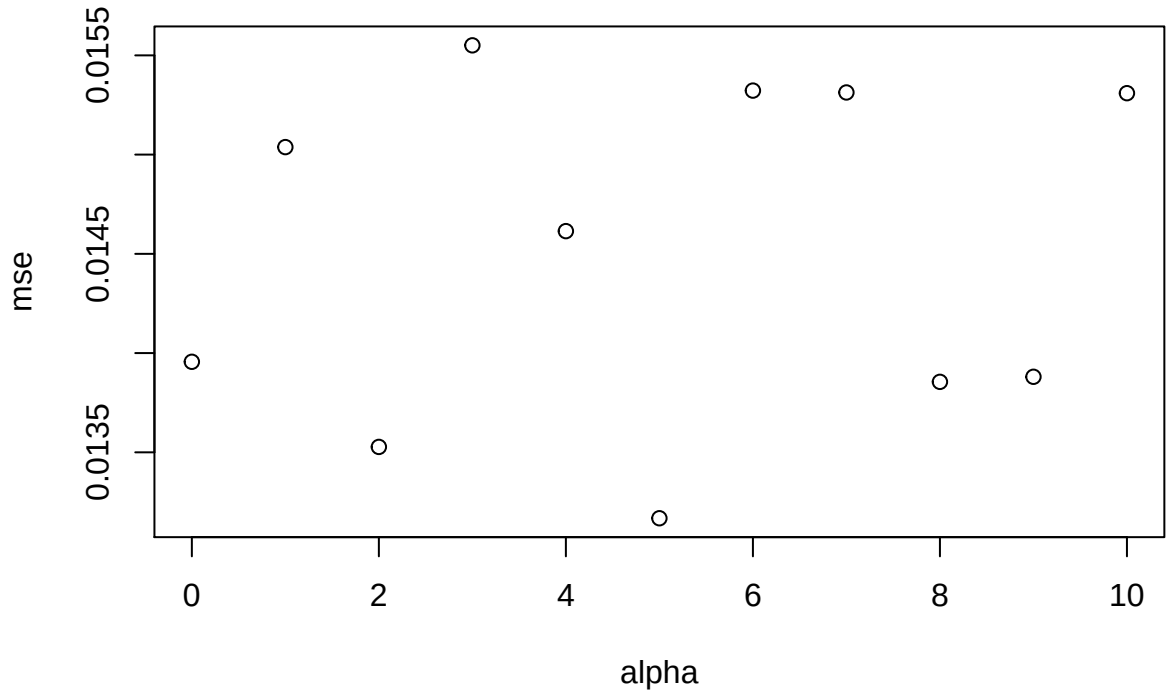
```
elnet_y <- predict.glmnet(object = elnet_lm, newx = x_test, s = cvglm5$lambda.1se)
mean((y_test - elnet_y)^2)
```

```
## [1] 0.02001308
```

```
# it seems that performance of elastic net is best
```

```
# 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] 0.01316788
```

```
# 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_1.csv",
          row.names = FALSE)
```