Code & Outputs

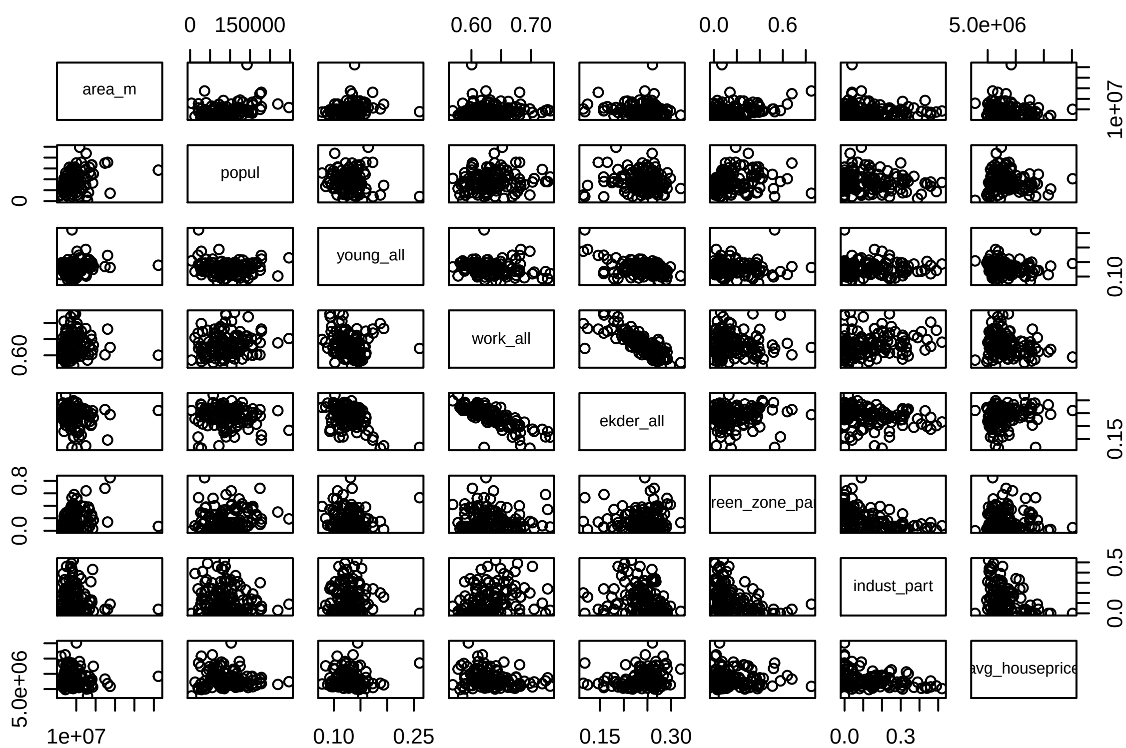
Liziyao, 1500017776

2018.1.15

```
library(stats)
library(MASS)
library(forecast)
library(lars)
```

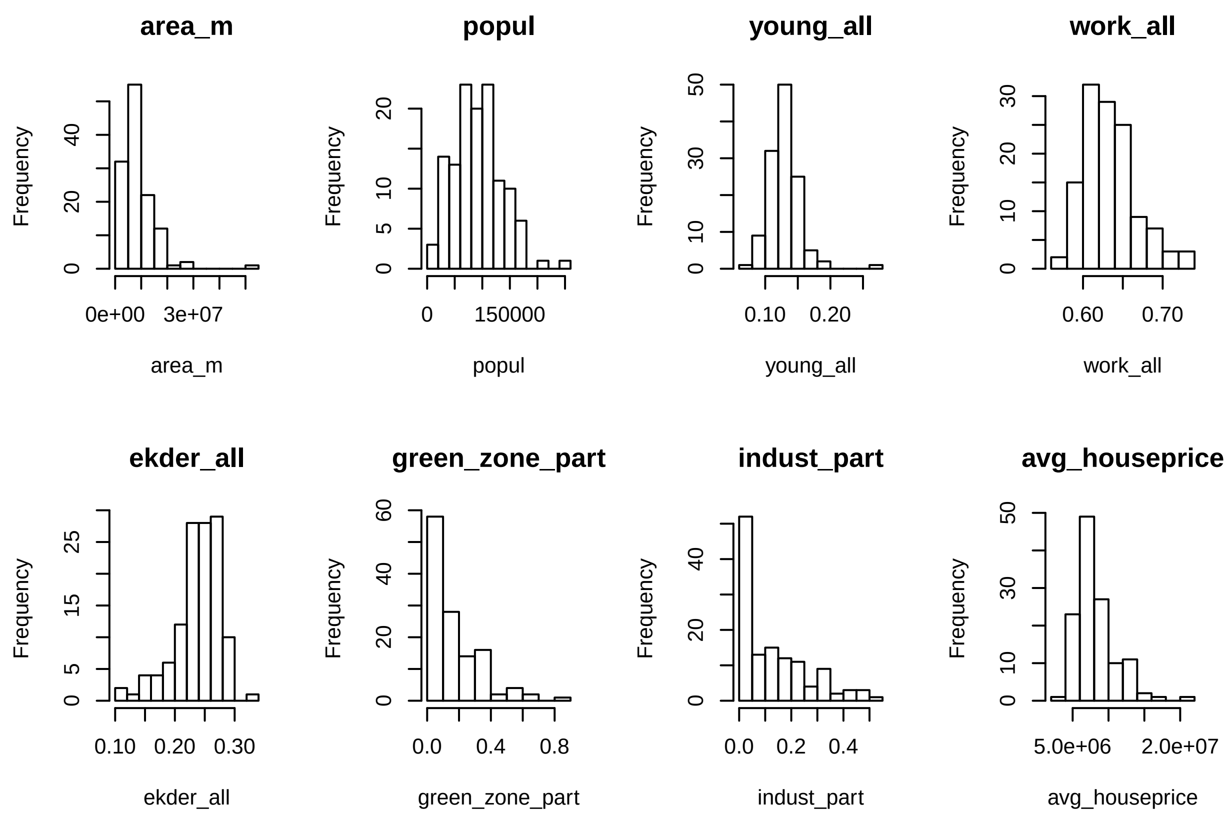
```
## Loaded lars 1.2
```

```
data_original=read.csv("moscow_districts.csv")
data=data_original
n=nrow(data);p=ncol(data)


###  first exploration  ###
# pairwise scatter plot
numeric_cols=c(2,3,4,5,6,7,8,18)
data_numeric=data[,numeric_cols]
pairs(data_numeric)
```



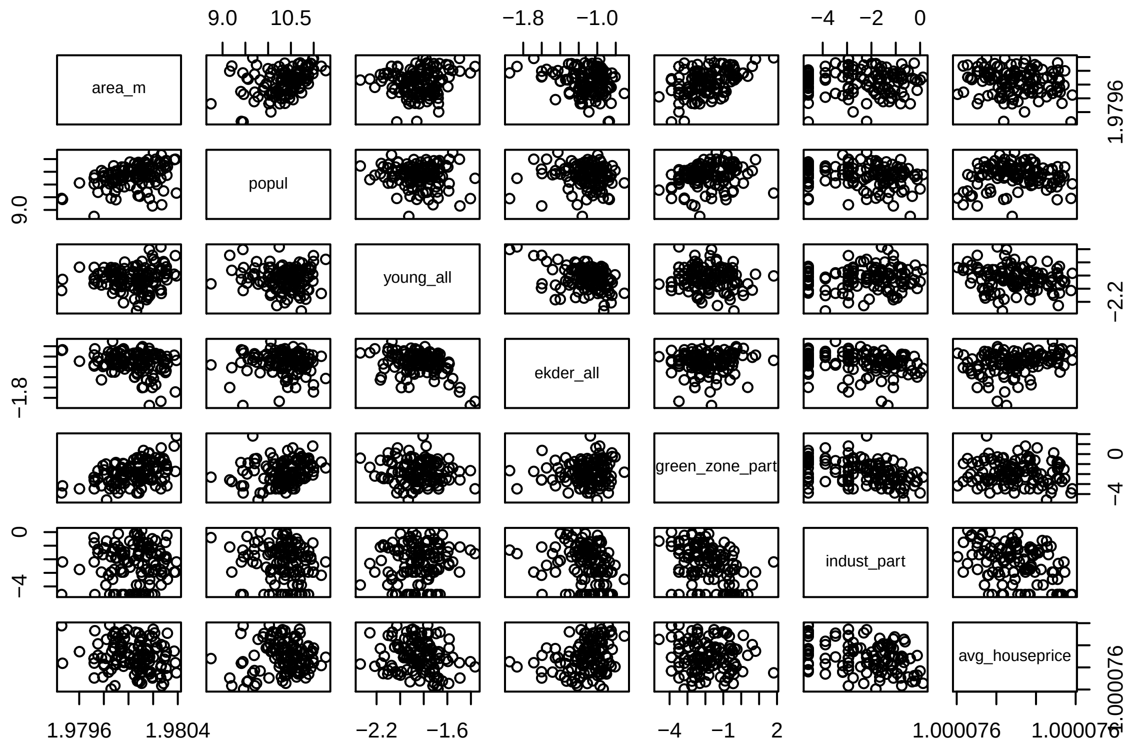

```
# marginal distr.
par(mfrow=c(2,4))
```

```r
for (i in 1:8){
  hist(data_numeric[,i], 10,
       main=colnames(data_numeric)[i],
       xlab=colnames(data_numeric)[i])
}
```

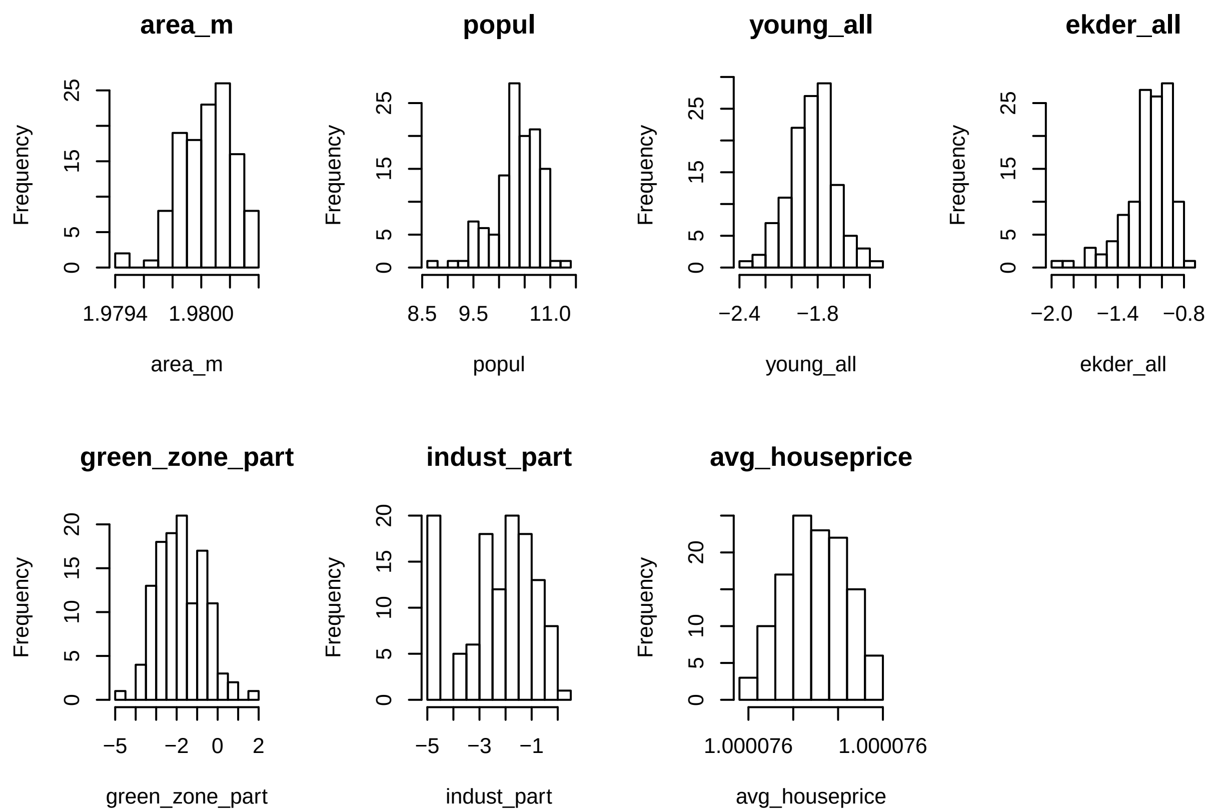



```r
par(mfrow=c(1,1))


###  remove outliers and bad features  ###
outliers=c(27,46,47,61)
bad_features=c(5)
data=data[-outliers,-bad_features]
n=nrow(data);p=ncol(data)


###  marginal transformations  ###
# logit trans. for precentage variables
logit=function(x) {
  y=x+.01
    # considering plausible 0s and percentages all small(<0.8).
  log(y/(1-y))
}
percentage_cols=c("young_all","ekder_all",
                  "green_zone_part","indust_part")
for (i in percentage_cols) {
```

```
  data[,i]=logit(data[,i])
}
# boxcox trans. for area,popul and avg_houseprice
boxcox_features=c("area_m","popul","avg_houseprice")
boxcox_lambdas=c(0,0,0)
names(boxcox_lambdas)=boxcox_features
for(i in boxcox_features){
  i
  boxcox_lambdas[i]=BoxCox.lambda(data[,i])
  data[,i]=BoxCox(data[,i],boxcox_lambdas[i])
}
# transformed scatter & hist plots
numeric_cols=c(2,3,4,5,6,7,17)
data_numeric=data[,numeric_cols]
pairs(data_numeric)
```



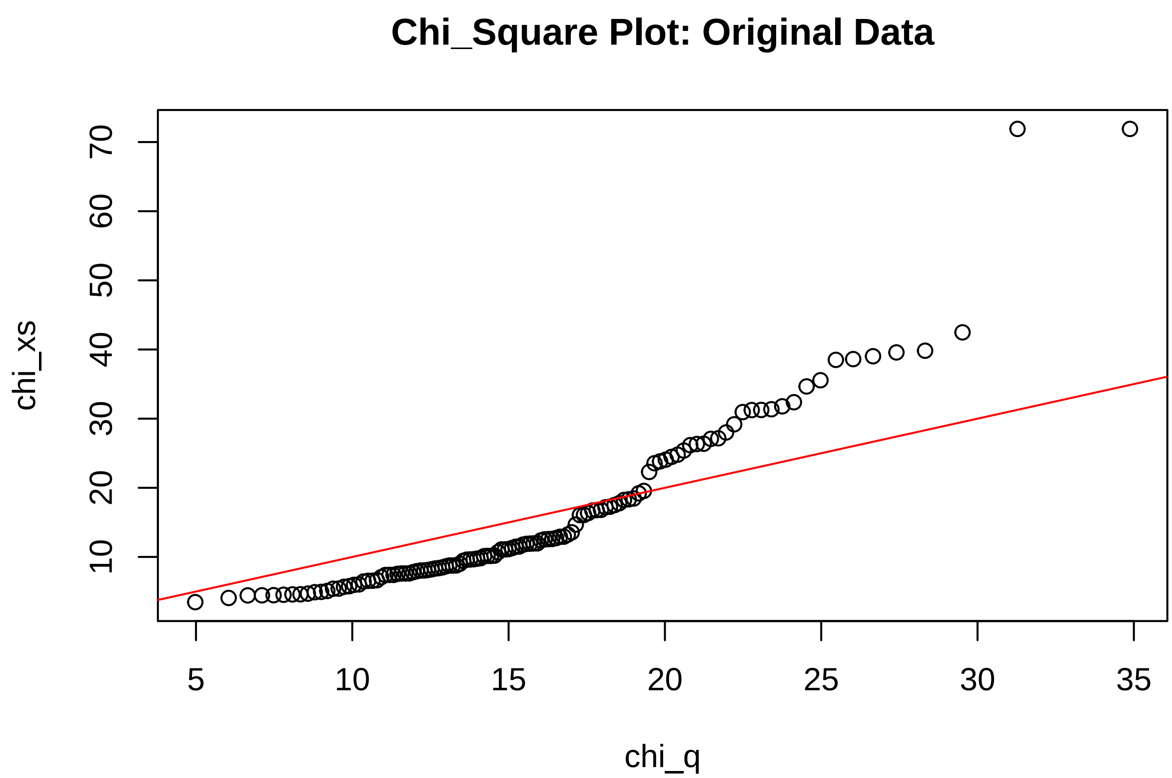

```
par(mfrow=c(2,4))
for (i in 1:7){
  hist(data_numeric[,i], 10,
       main=colnames(data_numeric)[i],
       xlab=colnames(data_numeric)[i])
}
par(mfrow=c(1,1))
```



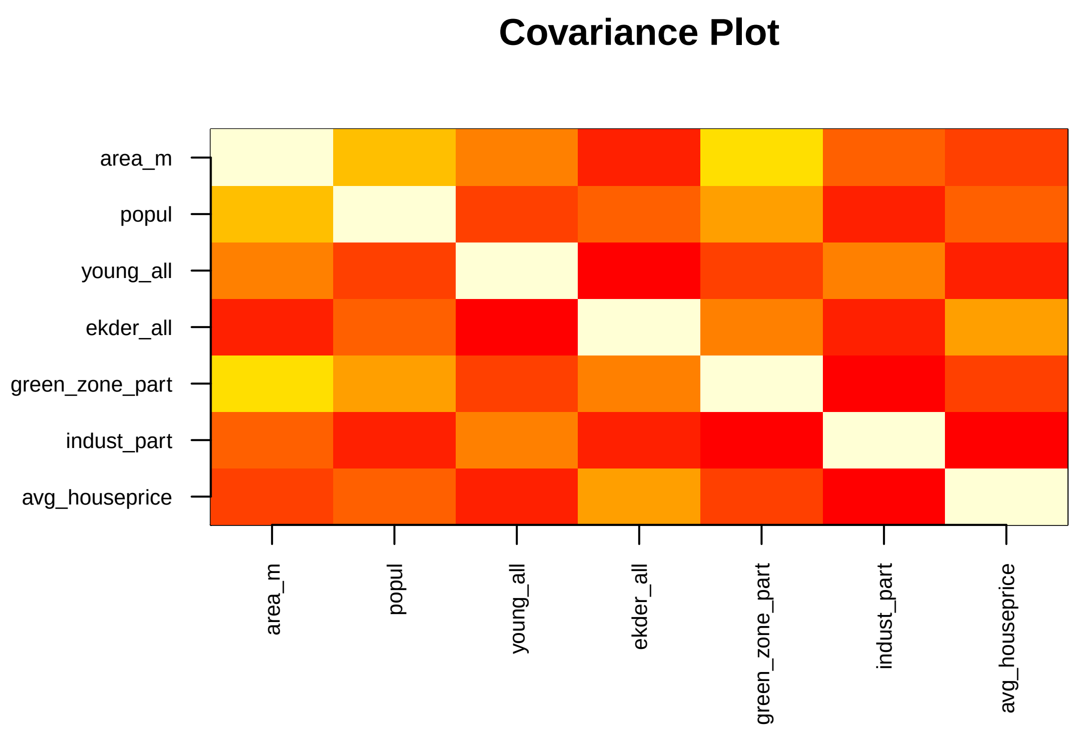

```r
###  statistical description: multivariate normal? ###
# all variables
x=as.matrix(data[,-1])
z=scale(x)
S=cov(z)
lambda=eigen(S)$values
lambda  # check for condition number: alright (10.21)
```

```
##  [1] 2.4859117 2.1465400 1.7222787 1.5529539 1.2409005 1.0538084 0.9143542
##  [8] 0.8272683 0.6955410 0.6542747 0.6126054 0.5852945 0.4726848 0.4192168
## [15] 0.3800029 0.2363641
```

```r
chi_x=diag(z%*%solve(S)%*%t(z))
chi_xs=sort(chi_x)
chi_q=qchisq(p=((1:n)-.5)/n,df=ncol(x))
plot(chi_q,chi_xs,main="Chi_Square Plot: Original Data")
lines(x=c(-1,50),y=c(-1,50),col="red")  # can't say normal
```



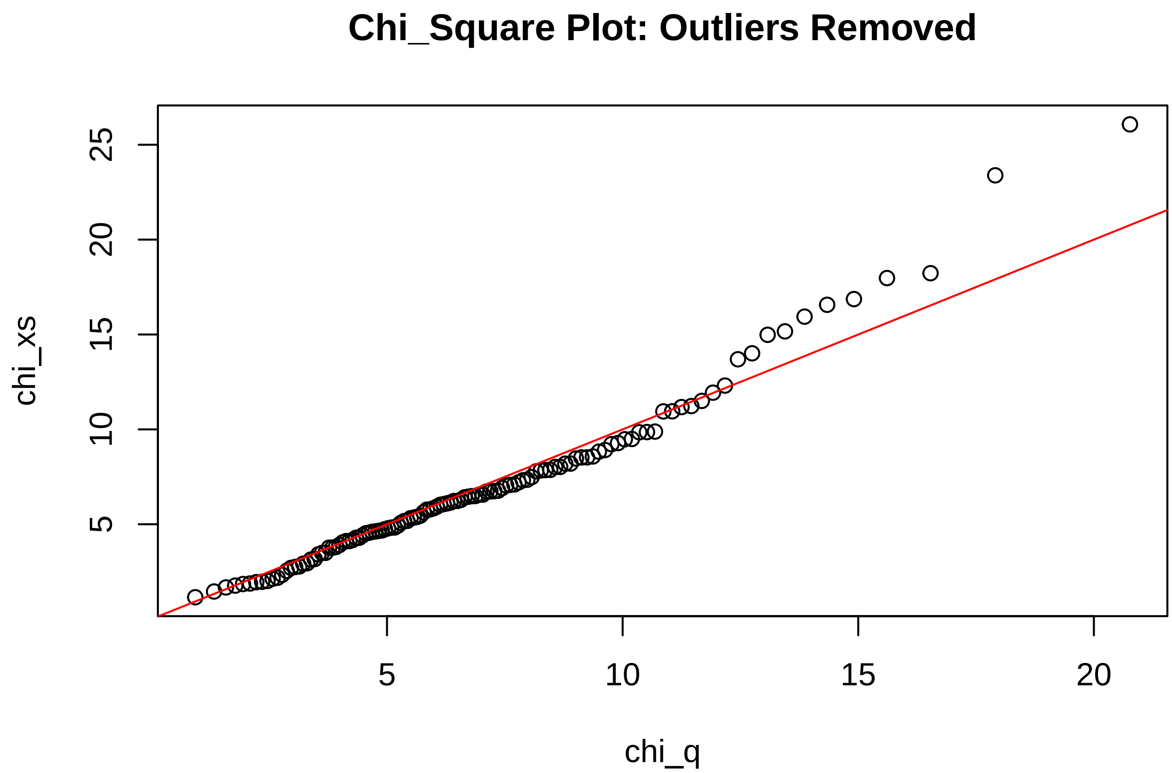

```
# numerical variables
x=as.matrix(data_numeric)
n=nrow(x);p=ncol(x)
z=scale(x)
S=cov(z)
# covariance plot
par(mar=c(7,6,5,4)+.1)
Splot=S[,7:1]
image(Splot,xaxt = 'n', yaxt='n', main="Covariance Plot")
axis(2,labels=colnames(Splot),at=(0:6)/6,las=1,cex.axis=.7)
axis(1,labels=rownames(Splot),at=(0:6)/6,las=3,cex.axis=.7)
```



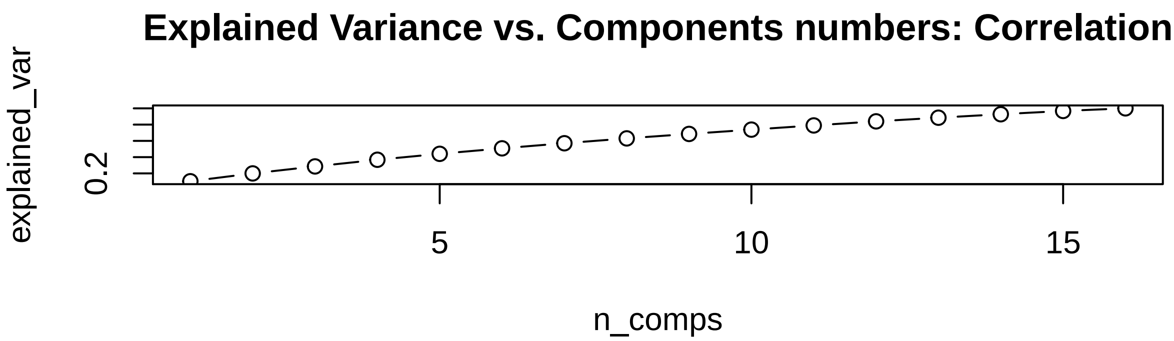

```
par(mar=c(5,4,4,2)+.1)

lambda=eigen(S)$values
lambda  # check for condition number: better (7.11)
```

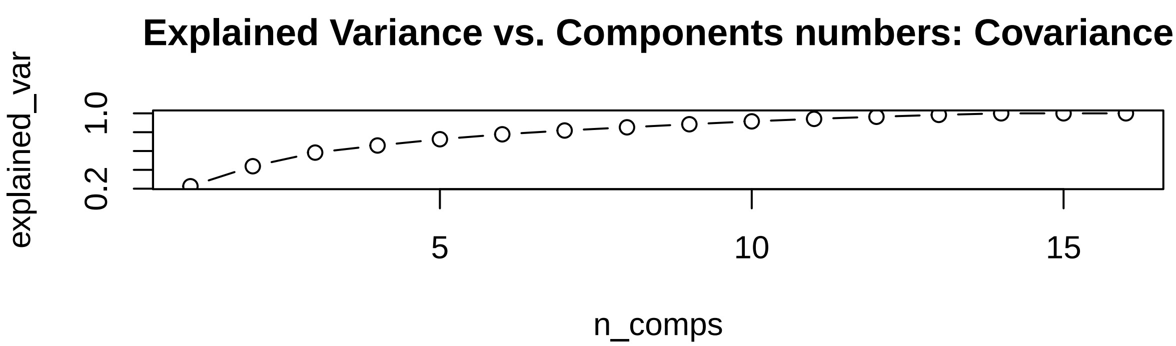
```
## [1] 1.9220102 1.8152041 1.0019369 0.8742839 0.5618169 0.5536257 0.2711223
```

```
chi_x=diag(z%*%solve(S)%*%t(z))
chi_xs=sort(chi_x)
chi_q=qchisq(p=((1:n)-.5)/n,df=ncol(x))
plot(chi_q,chi_xs,main="Chi_Square Plot: Outliers Removed")
lines(x=c(-1,50),y=c(-1,50),col="red")  # can say normal now
```




```
###  PCA & Factor Analysis  ###
x=as.matrix(data[,c(-1)])
par(mfrow=c(2,1))
x.pca=princomp(x,cor=T)
cumsdev=cumsum(x.pca$sdev)/sum(x.pca$sdev)
plot(cumsdev,type='b',
     xlab="n_comps",ylab="explained_var",
     main="Explained Variance vs. Components numbers: Correlation")
    # no dominant components; not very compressable
x.pca_cov=princomp(x,cor=F)
cumsdev_cov=cumsum(x.pca_cov$sdev)/sum(x.pca_cov$sdev)
plot(cumsdev_cov,type='b',
     xlab="n_comps",ylab="explained_var",
     main="Explained Variance vs. Components numbers: Covariance")
```






```r
    # no dominant components; more compressable than correl
par(mfrow=c(1,1))
x.fact=factanal(x,5,scores="Bartlett",rotation="varimax")
x.fact$loadings
```

```
##
## Loadings:
##                     Factor1 Factor2 Factor3 Factor4 Factor5
## area_m               0.891           0.355   0.138
## popul                0.494
## young_all           -0.101           0.659  -0.177
## ekder_all                           -0.586
## green_zone_part      0.585  -0.374  -0.180  -0.326
## indust_part         -0.238  -0.175   0.332   0.821
## healthcare_centers   0.261   0.317  -0.129   0.234
## university_top_20            0.699          -0.166
## thermal_power_plant  0.139                   0.353   0.154
## incineration                         0.427
## oil_chemistry                                0.151   0.984
## radiation            0.339   0.157           0.124
## railroad_terminal            0.541
## big_market                                           0.282
## nuclear_reactor                     -0.129   0.317
## avg_houseprice               0.556  -0.326  -0.200  -0.154
##
##                Factor1 Factor2 Factor3 Factor4 Factor5
```

```
## SS loadings      1.668   1.416   1.383   1.229   1.132
## Proportion Var   0.104   0.088   0.086   0.077   0.071
## Cumulative Var   0.104   0.193   0.279   0.356   0.427
```

```
###  OLS Regression   ###
data_reg=data[,-1]
boxcox_model=lm(avg_houseprice~.,data=data_reg)
summary(boxcox_model)
```

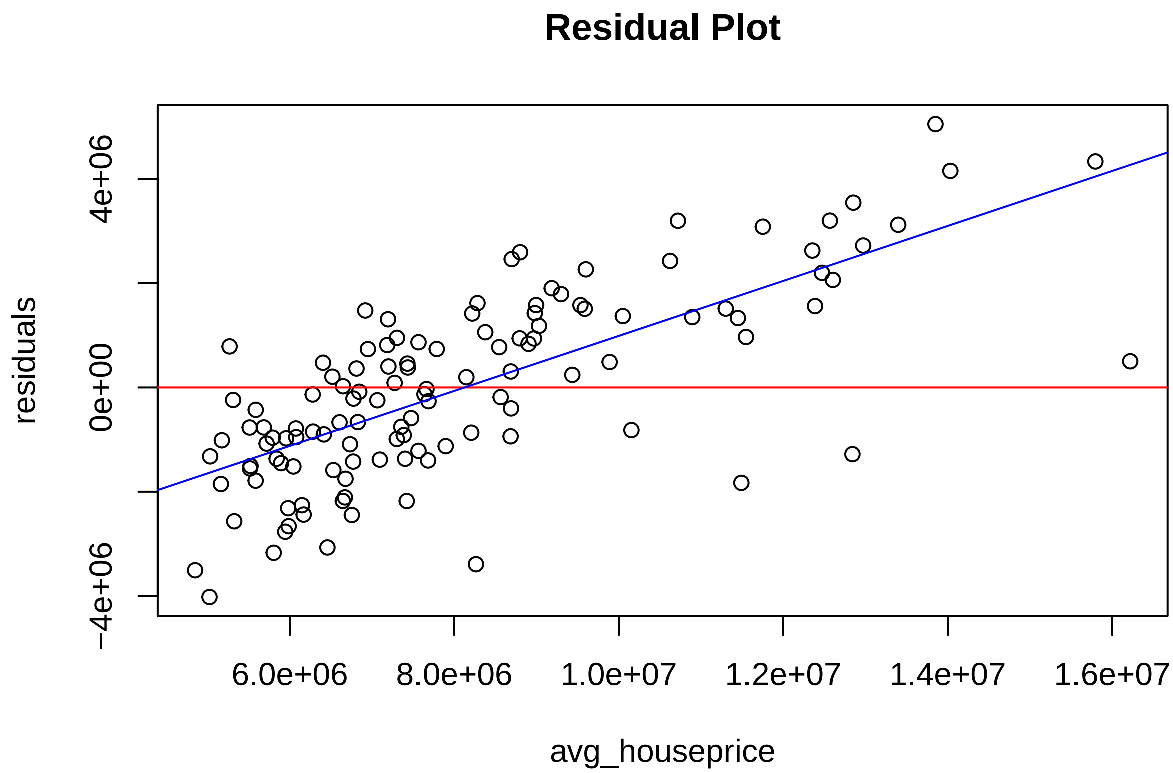
```
##
## Call:
## lm(formula = avg_houseprice ~ ., data = data_reg)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -6.845e-08 -1.894e-08  1.480e-10  2.282e-08  5.225e-08
##
## Coefficients:
##                      Estimate Std. Error   t value Pr(>|t|)
## (Intercept)          1.000e+00  4.028e-05 24830.893  < 2e-16 ***
## area_m              -1.116e-05  2.035e-05    -0.549 0.584455
## popul               -2.163e-11  7.044e-09    -0.003 0.997556
## young_all           -1.163e-08  1.803e-08    -0.645 0.520329
## ekder_all            1.984e-08  1.483e-08     1.338 0.183901
## green_zone_part     -4.376e-09  3.258e-09    -1.343 0.182153
## indust_part         -8.994e-09  2.527e-09    -3.560 0.000559 ***
## healthcare_centers   4.535e-09  2.064e-09     2.197 0.030209 *
## university_top_20    1.422e-08  6.463e-09     2.200 0.029979 *
## thermal_power_plant  1.311e-08  1.076e-08     1.218 0.225908
## incineration        -7.638e-09  1.632e-08    -0.468 0.640681
## oil_chemistry       -3.323e-08  2.233e-08    -1.488 0.139815
## radiation           -2.320e-09  5.979e-09    -0.388 0.698762
## railroad_terminal    1.712e-08  1.278e-08     1.339 0.183404
## big_market          -9.587e-10  1.277e-08    -0.075 0.940291
## nuclear_reactor      8.304e-10  1.421e-08     0.058 0.953508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.86e-08 on 105 degrees of freedom
## Multiple R-squared:  0.3931, Adjusted R-squared:  0.3064
## F-statistic: 4.535 on 15 and 105 DF,  p-value: 1.465e-06
```

```
data_reg[,16]=data_original[-outliers,18]
original_model=lm(avg_houseprice~.,data=data_reg)
summary(original_model)
```

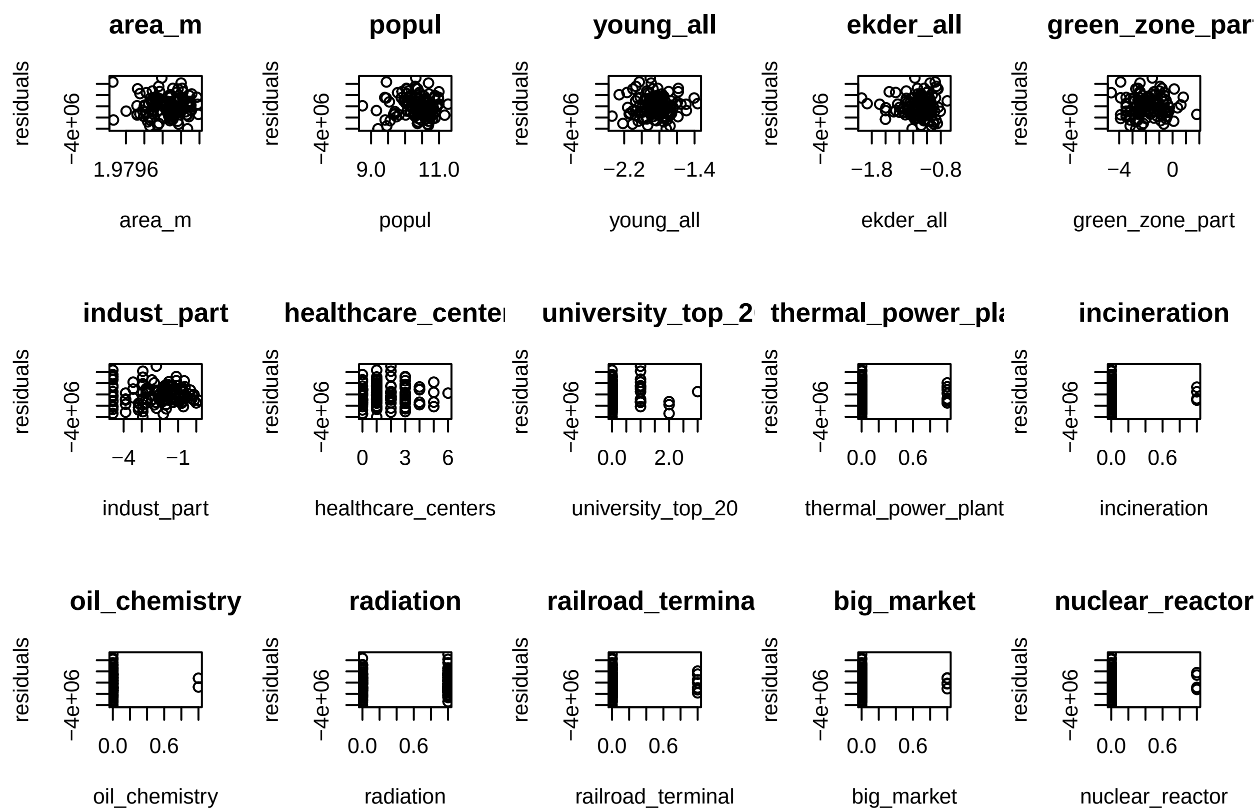
```
##
## Call:
## lm(formula = avg_houseprice ~ ., data = data_reg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4020441 -1279666  -134517  1307693  5052714
##
## Coefficients:
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            60271364 2685111487   0.022  0.98213
## area_m                -24261414 1356708098  -0.018  0.98577
## popul                   -674298     469610  -1.436  0.15401
## young_all               -756948    1202096  -0.630  0.53027
## ekder_all               1657048     988779   1.676  0.09674 .
## green_zone_part         -429736     217207  -1.978  0.05050 .
## indust_part             -778316     168436  -4.621 1.09e-05 ***
## healthcare_centers       259744     137611   1.888  0.06185 .
## university_top_20       1244742     430838   2.889  0.00469 **
## thermal_power_plant     1121356     717653   1.563  0.12117
## incineration            -218729    1087778  -0.201  0.84103
## oil_chemistry          -1628006    1488967  -1.093  0.27673
## radiation                -99023     398619  -0.248  0.80430
## railroad_terminal       1350374     852266   1.584  0.11610
## big_market              -139867     851214  -0.164  0.86980
## nuclear_reactor          129519     947243   0.137  0.89150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1907000 on 105 degrees of freedom
## Multiple R-squared:  0.4724, Adjusted R-squared:  0.3971
## F-statistic: 6.268 on 15 and 105 DF,  p-value: 2.929e-09
```

```r
    # R2 of original model is a lot better than a boxcox model
    # use original house price below.
residuals=original_model$residuals
plot(data_reg[,16],residuals,
     xlab="avg_houseprice",ylab="residuals",
     main="Residual Plot")
lines(x=c(0,2e7),y=c(0,0),col="red",type='c')
abline(lm(residuals~data_reg[,16]),col='blue')
```




```
par(mfrow=c(3,5))
for (i in 1:15) {
  plot(data_reg[,i],residuals,
       xlab=colnames(data_reg)[i],ylab="residuals",
       main=colnames(data_reg)[i])
}
```




```
par(mfrow=c(1,1))
data_onehot=read.csv("onehot.csv")
onehot_model=lm(avg_houseprice~.,data=data_onehot)
summary(onehot_model)  #  lower R2, but not significant
```

```
##
## Call:
## lm(formula = avg_houseprice ~ ., data = data_onehot)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3995456 -1072231        0  1006852  4858796
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -106143814 2873961497  -0.037   0.9706
## area_m                59997014 1452271592   0.041   0.9671
## popul                  -626765     507523  -1.235   0.2199
## young_all              -313543    1280829  -0.245   0.8071
## ekder_all              1759964    1035134   1.700   0.0924 .
## green_zone_part        -429222     228115  -1.882   0.0630 .
## indust_part            -775043     178851  -4.333 3.65e-05 ***
## healthcare_centers      257534     145116   1.775   0.0792 .
## university_top_20      1488904     465995   3.195   0.0019 **
## X1                     1364647    1243277   1.098   0.2751
## X2                    -1024027    2134526  -0.480   0.6325
```

```
## X8                     -78028     456786  -0.171   0.8647
## X9                     355166    1455298   0.244   0.8077
## X10                    515279    1486226   0.347   0.7296
## X11                   -243558    2030520  -0.120   0.9048
## X13                  -1280542    2043488  -0.627   0.5324
## X16                   1492146    1087727   1.372   0.1734
## X24                   -802875    1676084  -0.479   0.6330
## X25                   4635103    2049578   2.261   0.0260 *
## X32                    273396    1447711   0.189   0.8506
## X36                   -875518    2052712  -0.427   0.6707
## X40                   -702115    1195202  -0.587   0.5583
## X64                   -622982    2028044  -0.307   0.7594
## X65                    291100    2050951   0.142   0.8874
## X72                    264446    1465007   0.181   0.8571
## X73                   2582041    2021311   1.277   0.2046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1948000 on 95 degrees of freedom
## Multiple R-squared:  0.5019, Adjusted R-squared:  0.3709
## F-statistic:  3.83 on 25 and 95 DF,  p-value: 1.108e-06
```

```
###  different regression models  ###
rmse=function(true,pred){
  mean((true-pred)^2)^.5
}
train=sample(1:n,size=0.8*n)
test_y=data_reg$avg_houseprice[-train]
ols=lm(avg_houseprice~.,data=data_reg[train,])
ols_pred=predict(ols,data_reg[-train,])
ols_rmse=rmse(test_y,ols_pred)

step_ols=step(ols)
```

```
## Start:  AIC=2804.45
## avg_houseprice ~ area_m + popul + young_all + ekder_all + green_zone_part +
##     indust_part + healthcare_centers + university_top_20 + thermal_power_plant +
##     incineration + oil_chemistry + radiation + railroad_terminal +
##     big_market + nuclear_reactor
##
##                       Df  Sum of Sq        RSS    AIC
## - radiation            1 7.7244e+09 3.3463e+14 2802.4
## - big_market           1 2.2087e+11 3.3484e+14 2802.5
## - area_m               1 3.7925e+11 3.3500e+14 2802.6
## - incineration         1 4.8956e+11 3.3511e+14 2802.6
## - young_all            1 5.0845e+11 3.3513e+14 2802.6
## - nuclear_reactor      1 2.8452e+12 3.3746e+14 2803.3
## - healthcare_centers   1 3.7666e+12 3.3839e+14 2803.5
## - oil_chemistry        1 4.5661e+12 3.3919e+14 2803.8
## - popul                1 6.1388e+12 3.4076e+14 2804.2
## <none>                             3.3462e+14 2804.4
## - ekder_all            1 7.2008e+12 3.4182e+14 2804.5
## - thermal_power_plant  1 8.9841e+12 3.4360e+14 2805.0
## - green_zone_part      1 1.1323e+13 3.4594e+14 2805.6
## - railroad_terminal    1 1.1930e+13 3.4655e+14 2805.8
```

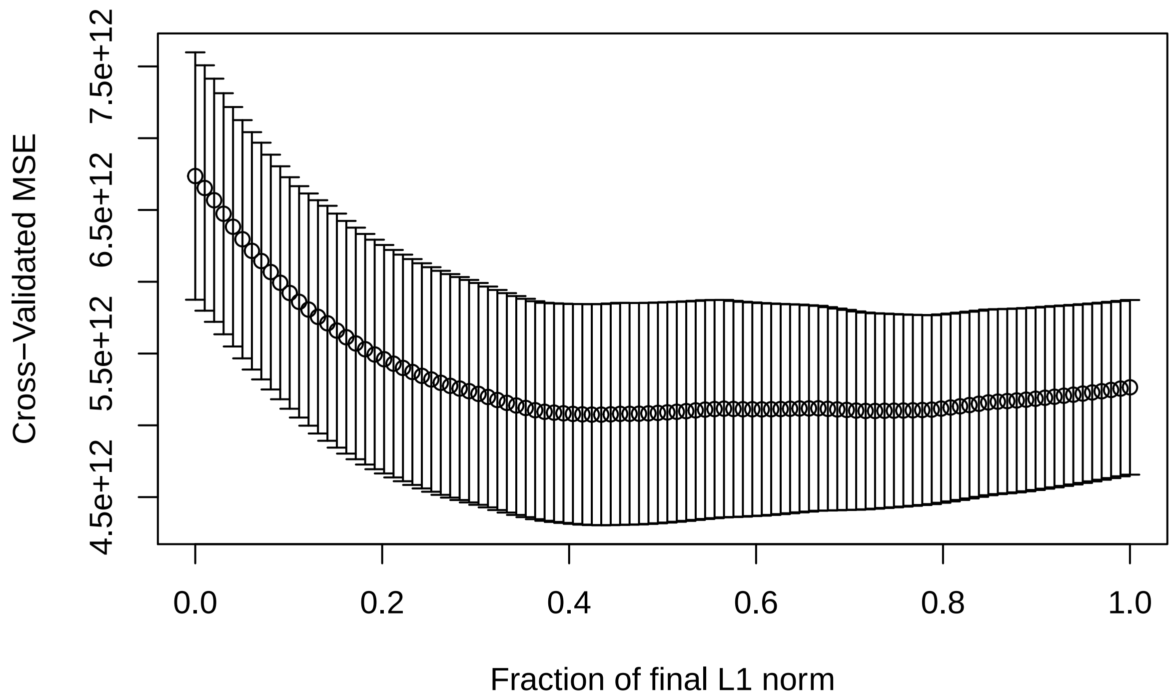
```
## - university_top_20    1 2.5220e+13 3.5984e+14 2809.4
## - indust_part          1 7.1056e+13 4.0567e+14 2820.9
##
## Step:  AIC=2802.45
## avg_houseprice ~ area_m + popul + young_all + ekder_all + green_zone_part +
##     indust_part + healthcare_centers + university_top_20 + thermal_power_plant +
##     incineration + oil_chemistry + railroad_terminal + big_market +
##     nuclear_reactor
##
##                       Df  Sum of Sq        RSS    AIC
## - big_market           1 2.1881e+11 3.3485e+14 2800.5
## - area_m               1 3.9416e+11 3.3502e+14 2800.6
## - young_all            1 5.0502e+11 3.3513e+14 2800.6
## - incineration         1 5.1119e+11 3.3514e+14 2800.6
## - nuclear_reactor      1 2.8424e+12 3.3747e+14 2801.3
## - healthcare_centers   1 3.7720e+12 3.3840e+14 2801.5
## - oil_chemistry        1 4.5645e+12 3.3919e+14 2801.8
## - popul                1 6.7284e+12 3.4136e+14 2802.4
## <none>                              3.3463e+14 2802.4
## - ekder_all            1 7.2911e+12 3.4192e+14 2802.5
## - thermal_power_plant  1 9.0202e+12 3.4365e+14 2803.0
## - green_zone_part      1 1.1318e+13 3.4594e+14 2803.6
## - railroad_terminal    1 1.2066e+13 3.4669e+14 2803.8
## - university_top_20    1 2.5758e+13 3.6038e+14 2807.6
## - indust_part          1 7.1422e+13 4.0605e+14 2819.0
##
## Step:  AIC=2800.51
## avg_houseprice ~ area_m + popul + young_all + ekder_all + green_zone_part +
##     indust_part + healthcare_centers + university_top_20 + thermal_power_plant +
##     incineration + oil_chemistry + railroad_terminal + nuclear_reactor
##
##                       Df  Sum of Sq        RSS    AIC
## - area_m               1 3.7803e+11 3.3522e+14 2798.6
## - young_all            1 4.5768e+11 3.3530e+14 2798.6
## - incineration         1 5.1657e+11 3.3536e+14 2798.7
## - nuclear_reactor      1 2.8160e+12 3.3766e+14 2799.3
## - healthcare_centers   1 3.8717e+12 3.3872e+14 2799.6
## - oil_chemistry        1 4.4460e+12 3.3929e+14 2799.8
## - popul                1 6.7221e+12 3.4157e+14 2800.4
## <none>                              3.3485e+14 2800.5
## - ekder_all            1 7.4847e+12 3.4233e+14 2800.6
## - thermal_power_plant  1 8.8017e+12 3.4365e+14 2801.0
## - green_zone_part      1 1.1175e+13 3.4602e+14 2801.7
## - railroad_terminal    1 1.2140e+13 3.4699e+14 2801.9
## - university_top_20    1 2.5561e+13 3.6041e+14 2805.6
## - indust_part          1 7.1241e+13 4.0609e+14 2817.0
##
## Step:  AIC=2798.62
## avg_houseprice ~ popul + young_all + ekder_all + green_zone_part +
##     indust_part + healthcare_centers + university_top_20 + thermal_power_plant +
##     incineration + oil_chemistry + railroad_terminal + nuclear_reactor
##
##                       Df  Sum of Sq        RSS    AIC
## - young_all            1 6.0972e+11 3.3583e+14 2796.8
```

```
## - incineration         1 6.2920e+11 3.3585e+14 2796.8
## - nuclear_reactor      1 2.8604e+12 3.3808e+14 2797.4
## - healthcare_centers   1 3.5712e+12 3.3879e+14 2797.6
## - oil_chemistry        1 4.6888e+12 3.3991e+14 2797.9
## <none>                              3.3522e+14 2798.6
## - thermal_power_plant  1 8.4431e+12 3.4367e+14 2799.0
## - ekder_all            1 8.4580e+12 3.4368e+14 2799.0
## - popul                1 8.8819e+12 3.4411e+14 2799.1
## - railroad_terminal    1 1.2021e+13 3.4724e+14 2800.0
## - green_zone_part      1 1.9147e+13 3.5437e+14 2801.9
## - university_top_20    1 2.5185e+13 3.6041e+14 2803.6
## - indust_part          1 7.5837e+13 4.1106e+14 2816.2
##
## Step:  AIC=2796.8
## avg_houseprice ~ popul + ekder_all + green_zone_part + indust_part +
##     healthcare_centers + university_top_20 + thermal_power_plant +
##     incineration + oil_chemistry + railroad_terminal + nuclear_reactor
##
##                       Df  Sum of Sq        RSS    AIC
## - incineration         1 7.1521e+11 3.3655e+14 2795.0
## - nuclear_reactor      1 3.0460e+12 3.3888e+14 2795.7
## - healthcare_centers   1 3.6083e+12 3.3944e+14 2795.8
## - oil_chemistry        1 4.2092e+12 3.4004e+14 2796.0
## <none>                              3.3583e+14 2796.8
## - popul                1 8.5217e+12 3.4436e+14 2797.2
## - thermal_power_plant  1 8.7115e+12 3.4454e+14 2797.2
## - railroad_terminal    1 1.2639e+13 3.4847e+14 2798.3
## - ekder_all            1 1.2770e+13 3.4860e+14 2798.4
## - green_zone_part      1 1.9119e+13 3.5495e+14 2800.1
## - university_top_20    1 2.4956e+13 3.6079e+14 2801.7
## - indust_part          1 7.7229e+13 4.1306e+14 2814.7
##
## Step:  AIC=2795
## avg_houseprice ~ popul + ekder_all + green_zone_part + indust_part +
##     healthcare_centers + university_top_20 + thermal_power_plant +
##     oil_chemistry + railroad_terminal + nuclear_reactor
##
##                       Df  Sum of Sq        RSS    AIC
## - nuclear_reactor      1 3.1774e+12 3.3973e+14 2793.9
## - oil_chemistry        1 3.9097e+12 3.4046e+14 2794.1
## - healthcare_centers   1 4.3677e+12 3.4092e+14 2794.2
## <none>                              3.3655e+14 2795.0
## - thermal_power_plant  1 8.2730e+12 3.4482e+14 2795.3
## - popul                1 8.7109e+12 3.4526e+14 2795.4
## - railroad_terminal    1 1.3069e+13 3.4962e+14 2796.7
## - ekder_all            1 1.3889e+13 3.5044e+14 2796.9
## - green_zone_part      1 1.8994e+13 3.5554e+14 2798.3
## - university_top_20    1 2.4527e+13 3.6108e+14 2799.8
## - indust_part          1 8.2683e+13 4.1923e+14 2814.1
##
## Step:  AIC=2793.9
## avg_houseprice ~ popul + ekder_all + green_zone_part + indust_part +
##     healthcare_centers + university_top_20 + thermal_power_plant +
##     oil_chemistry + railroad_terminal
```

```
##
##                      Df  Sum of Sq        RSS    AIC
## - oil_chemistry        1 4.4260e+12 3.4415e+14 2793.1
## - healthcare_centers   1 5.6955e+12 3.4542e+14 2793.5
## <none>                             3.3973e+14 2793.9
## - popul                1 8.3785e+12 3.4810e+14 2794.2
## - thermal_power_plant  1 9.9759e+12 3.4970e+14 2794.7
## - railroad_terminal    1 1.2306e+13 3.5203e+14 2795.3
## - ekder_all            1 1.5248e+13 3.5497e+14 2796.1
## - green_zone_part      1 1.9542e+13 3.5927e+14 2797.3
## - university_top_20    1 2.3968e+13 3.6369e+14 2798.4
## - indust_part          1 8.0691e+13 4.2042e+14 2812.4
##
## Step:  AIC=2793.14
## avg_houseprice ~ popul + ekder_all + green_zone_part + indust_part +
##     healthcare_centers + university_top_20 + thermal_power_plant +
##     railroad_terminal
##
##                      Df  Sum of Sq        RSS    AIC
## - healthcare_centers   1 5.7233e+12 3.4988e+14 2792.7
## <none>                             3.4415e+14 2793.1
## - popul                1 7.9105e+12 3.5206e+14 2793.3
## - thermal_power_plant  1 8.0500e+12 3.5220e+14 2793.4
## - railroad_terminal    1 1.3248e+13 3.5740e+14 2794.8
## - ekder_all            1 1.5620e+13 3.5977e+14 2795.4
## - green_zone_part      1 1.9586e+13 3.6374e+14 2796.5
## - university_top_20    1 2.3461e+13 3.6761e+14 2797.5
## - indust_part          1 8.5210e+13 4.2936e+14 2812.4
##
## Step:  AIC=2792.73
## avg_houseprice ~ popul + ekder_all + green_zone_part + indust_part +
##     university_top_20 + thermal_power_plant + railroad_terminal
##
##                      Df  Sum of Sq        RSS    AIC
## - popul                1 4.2889e+12 3.5416e+14 2791.9
## - thermal_power_plant  1 6.7260e+12 3.5660e+14 2792.6
## <none>                             3.4988e+14 2792.7
## - railroad_terminal    1 1.3740e+13 3.6362e+14 2794.4
## - ekder_all            1 1.6400e+13 3.6628e+14 2795.1
## - green_zone_part      1 2.0943e+13 3.7082e+14 2796.3
## - university_top_20    1 2.8615e+13 3.7849e+14 2798.3
## - indust_part          1 8.0253e+13 4.3013e+14 2810.6
##
## Step:  AIC=2791.9
## avg_houseprice ~ ekder_all + green_zone_part + indust_part +
##     university_top_20 + thermal_power_plant + railroad_terminal
##
##                      Df  Sum of Sq        RSS    AIC
## - thermal_power_plant  1 6.0659e+12 3.6023e+14 2791.5
## <none>                             3.5416e+14 2791.9
## - railroad_terminal    1 1.4350e+13 3.6851e+14 2793.7
## - ekder_all            1 1.6208e+13 3.7037e+14 2794.2
## - green_zone_part      1 2.4959e+13 3.7912e+14 2796.4
## - university_top_20    1 2.8460e+13 3.8262e+14 2797.3
```

```
## - indust_part          1 7.7790e+13 4.3195e+14 2809.0
##
## Step:  AIC=2791.53
## avg_houseprice ~ ekder_all + green_zone_part + indust_part +
##     university_top_20 + railroad_terminal
##
##                     Df  Sum of Sq        RSS    AIC
## <none>                             3.6023e+14 2791.5
## - railroad_terminal  1 1.6954e+13 3.7718e+14 2793.9
## - ekder_all          1 1.7205e+13 3.7744e+14 2794.0
## - green_zone_part    1 2.2557e+13 3.8279e+14 2795.4
## - university_top_20  1 2.7154e+13 3.8738e+14 2796.5
## - indust_part        1 7.1758e+13 4.3199e+14 2807.0
```

```r
step_ols_pred=predict(step_ols,data_reg[-train,])
step_ols_rmse=rmse(test_y,step_ols_pred)

trainx=as.matrix(data_reg[train,-16])
trainy=data_reg[train,16]
testx=as.matrix(data_reg[-train,-16])

ridge=lm.ridge(avg_houseprice~.,data=data_reg[train,],
               lambda=seq(0,100,length=10001))
ridge_k=which.min(ridge$GCV)
ridge_coef=coef(ridge)[ridge_k,]
ridge_pred=cbind(1,testx)%*%ridge_coef
ridge_rmse=rmse(test_y,ridge_pred)

lasso=lars(trainx,trainy)
lasso_cv=cv.lars(trainx,trainy,K=5)
```




```
lasso_k=lasso_cv$index[which.min(lasso_cv$cv)]
lasso_coef=coef(lasso,s=lasso_k,mode="fraction")
lasso_interc=predict(lasso,s=lasso_k,mode="fraction",
                     newx=t(numeric(15)))$fit
lasso_pred=lasso_interc+testx%*%lasso_coef
lasso_rmse=rmse(test_y,lasso_pred)

totalx=data_reg[,-16]
pcax=predict(princomp(totalx))[,1:10]
totalx=as.data.frame(cbind(pcax,data_reg[,16]))
colnames(totalx)=c(1:10,"hp")
pca=lm(hp~.,data=totalx[train,])
pca_pred=predict(pca,totalx[-train,])
pca_rmse=rmse(test_y,pca_pred)

c(ols_rmse,step_ols_rmse,ridge_rmse,lasso_rmse,pca_rmse)
```

```
## [1] 1511196 1524160 1417529 1413124 1364878
```