Predictive model building

Bao Doquang, Dhwanit Agarwal, Akksay Singh and Shristi Singh

April 19, 2020

Part 1.

This analysis attempts to find the important factors which determine the rent amount per square foot of a property. We are interested in estimating the percentage change in rent per square foot from 18 different factors individually as well as collectively for the entire model.

We got a sample of 7894 commercial rental properties and 18 independent variables which can affect percentage change in rental income per square foot. We selected log-lin model because this model gives percentage change in dependent variable which in our model is rent per square foot. Since it contains 18 independent variables, the model is very complex and for that we need a large sample size to avoid overfitting. We have 7894 observations containing 685 properties that have been awarded either LEED or EnergyStar certification, we conclude that our sample size is large enough to give generalized predictions. We used dummy variables for qualitative factors which are: renovated, class_a, class_b, green_rating, net and amenities. These qualitative factors increase the rental income and they are often significant in explaining the rental income variation. In actual model we collapsed LEED and EnergyStar in a single variable named "green_rating".

We fit the OLS linear regression model that includes these 18 variables as predictors of rental income. Our results are on page 8 below. Overall, the model is a good fit as Adjusted R^2 = 0.68 which indicates that our model explains 68% of the variation in rental income. The F Statistics =925.2 on 18 and 7801 degrees of freedom, is highly significant which means that at least one the coefficients in the model is significantly not equal to zero i.e. ??i =!0. The third point is to look at individual 't' scores for each coefficient or ??i. We find that only three variables: stories, renovated, and gas cost are not significant factors in explaining the movement in rental income while the remaining 15 variables are significant. We can conclude that these variables have ??i s = 0 or they do not impact rental income. We can drop these variables in future research. Most of the remaining variables are individually significant at 1 percent or less which is great.

The coefficients in a log-lin model tell how much percentage change will happen in rental income on an average when they are multiplied by 100 when the independent variable increases by 1 unit while keeping rest of the variables constant.

Part 2.

In our model, the average increase in rental income is 2.98% when the greenrating increases by 1 unit while all other things do not change. This variable is significant at 0.01 level.

From our analysis, we conclude that our log-lin OLS model does a good job in explaining the relationship between rental income and the other 18 factors determining the rental income. Our model's sign's of coefficients are in the expected directions such as increase in bad factors like age and electricity costs would be negatively related to rental income while increase in good factors like green rating, size, employment growth, amenities would positively increase the rental income.

```
library(rsample)  # data splitting
library(glmnet)   # implementing regularized regression approaches
```

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

```
## Loaded glmnet 3.0-2
```

```
library(dplyr)    # basic data manipulation procedures
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)  # plotting
library(DAAG)
```

```
## Loading required package: lattice
```

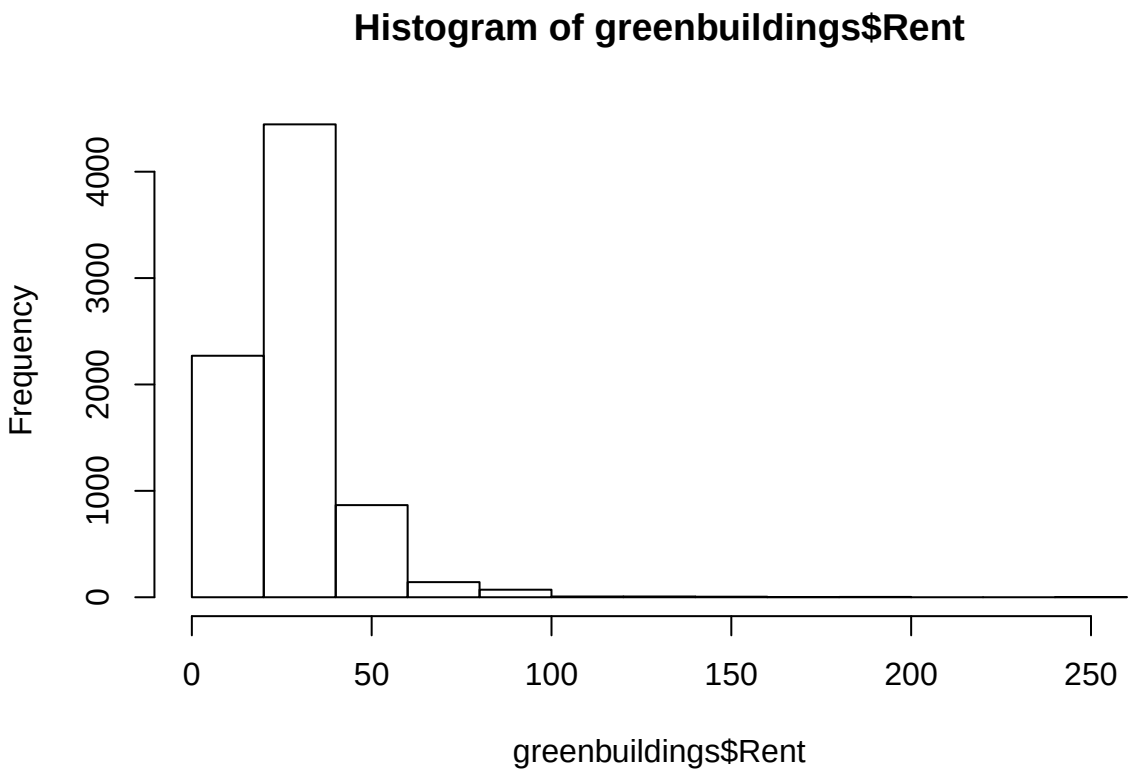
```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:DAAG':
##
##     hills
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

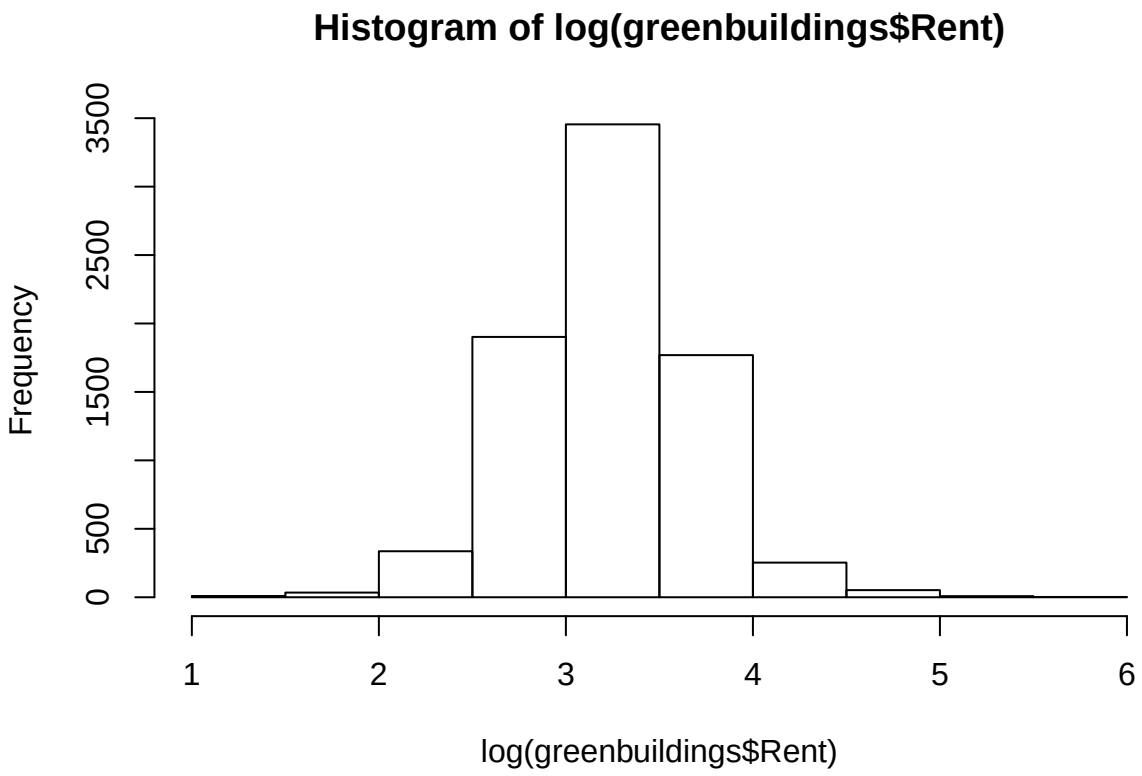
```
# import data and examine it

greenbuildings <- read.csv("greenbuildings.csv")
#View(greenbuildings)
ok <- complete.cases(greenbuildings)
greenbuildings <- greenbuildings[ok,]


# note that shares is hugely skewed
# probably want a log transformation here
hist(greenbuildings$Rent)
```




```
summary(greenbuildings$Rent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   19.50   25.20   28.42   34.18  250.00
```

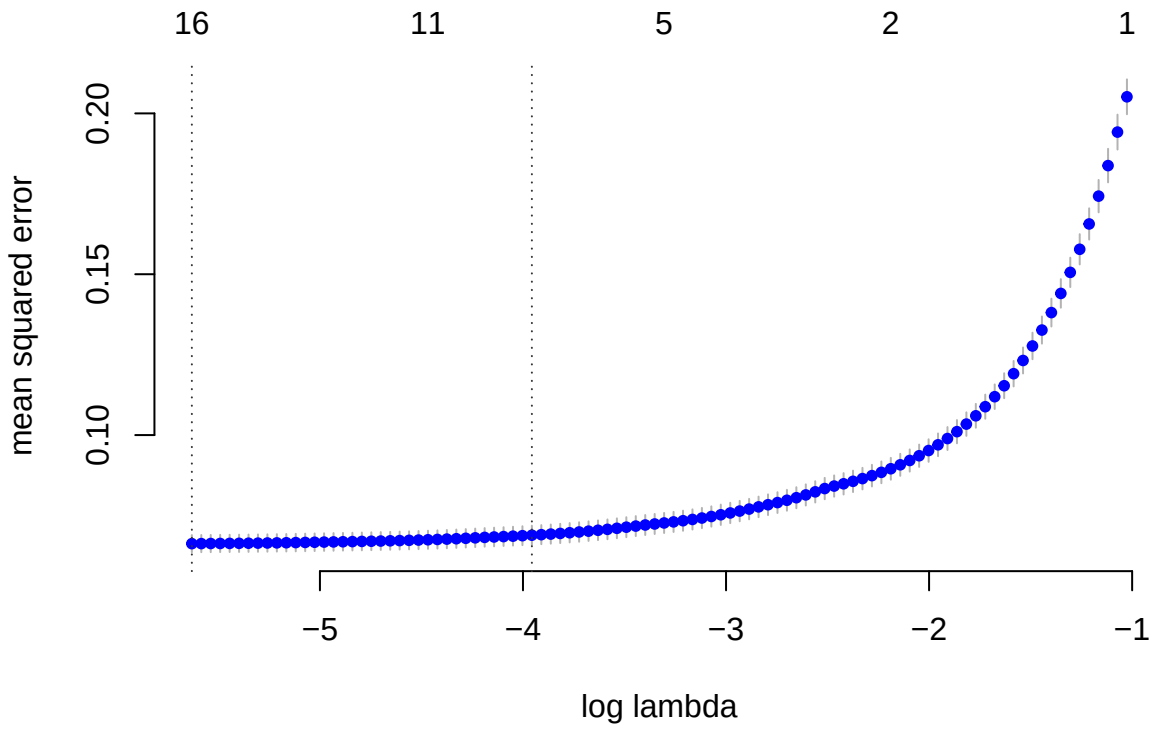
```
# much nicer :-)
hist(log(greenbuildings$Rent))
```




```
#### lasso (glmnet does L1-L2, gamlr does L0-L1)
# I want to fit a lasso regression and do cross validation of K=10 folds
# inorder to automate finiding independent variables and training & testing my data multiple times.
# cv.gamlr command in the gamlr does it for me.
# download gamlr library
library(gamlr)

# i create a matrix of all my independent varaibles except for url from online_news data to make it eas
# the sparse.model.matrix function.
x = sparse.model.matrix( log(Rent) ~  . - CS_PropertyID - LEED -Energystar  , data=greenbuildings, stan

y = log(greenbuildings$Rent) # pull out `y' too just for convenience and do log(shares)- dependent vari



# Here I fit my lasso regression to the data and do my cross validation of k=10 n folds
# the cv.gamlr command does both things at once.
#(verb just prints progress)
cvl = cv.gamlr(x, y, nfold=10, verb=TRUE)
```

```
## fold 1,2,3,4,5,6,7,8,9,10,done.
```

```
# plot the out-of-sample deviance as a function of log lambda
plot(cvl, bty="n")
```




```
min(cvl$cvm)       # minimum MSE
```

```
## [1] 0.06624338
```

```
## [1] 0.06615445
cvl$lambda.min     # lambda for this min MSE
```

```
## [1] 0.003585894
```

```
## [1] 0.003585894


cvl$cvm[cvl$lambda == cvl$lambda.1se]  # 1 st.error of min MSE
```

```
## numeric(0)
```

```
## [1] 0.06908108
cvl$lambda.1se  # lambda for this MSE
```

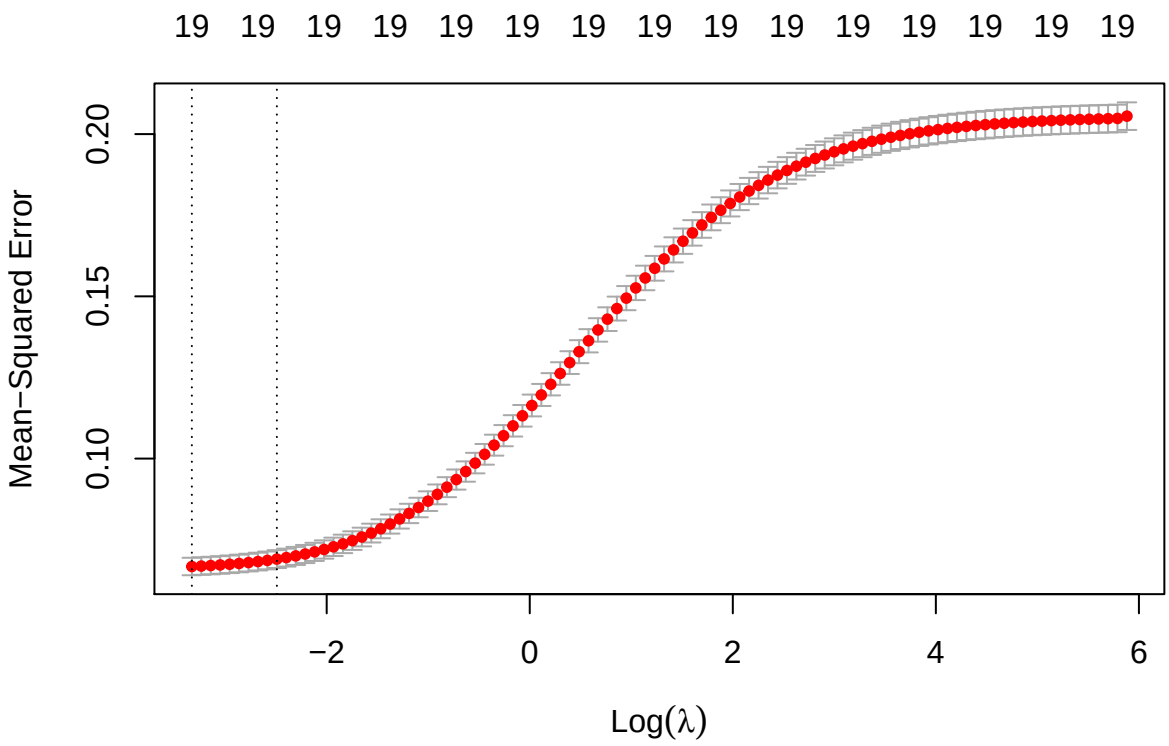
```
## [1] 0.01913684
```

```
## [1] 0.01516562

#fitted coefficients at minimum MSE
coef(cvl, select="min")
```

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                              seg100
## intercept          2.405860e+00
```

```
## cluster            2.888478e-05
## size               7.390862e-08
## empl_gr            2.507732e-03
## leasing_rate       4.649535e-04
## stories            2.305963e-04
## age               -9.899540e-04
## renovated          .
## class_a            9.182501e-02
## class_b            2.579294e-02
## green_rating       1.910632e-02
## net               -3.024764e-02
## amenities          3.902586e-02
## cd_total_07       -3.056906e-05
## hd_total07         .
## total_dd_07       -1.821279e-05
## Precipitation      4.260804e-05
## Gas_Costs          .
## Electricity_Costs  .
## cluster_rent       3.087431e-02
```

```r
# Apply CV Ridge regression to data
cvr <- cv.glmnet(
  x ,
  y ,
  alpha = 0
)

# plot MSE as a function of log(lambda)
plot(cvr)
```




```
min(cvr$cvm)       # minimum MSE
```

```
## [1] 0.06674327
```

```
## [1] 0.06679016  #value observed
cvr$lambda.min     # lambda for this min MSE
```

```
## [1] 0.03585894
```

```
## [1] 0.03585894

cvr$cvm[cvr$lambda == cvr$lambda.1se]  # 1 st.error of min MSE
```

```
## [1] 0.06902818
```

```
## [1] 0.06908108
cvr$lambda.1se  # lambda for this MSE
```

```
## [1] 0.0828388
```

```
## [1] 0.0828388

#fitted coefficients at minimum MSE
coef(cvr, select="min")
```

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                               1
## (Intercept)        2.441321e+00
## cluster            4.932116e-05
```

```
## size                5.929577e-08
## empl_gr             3.695284e-03
## leasing_rate        9.056597e-04
## stories             7.218613e-04
## age                -9.316485e-04
## renovated          -9.610125e-03
## class_a             8.716012e-02
## class_b             2.226612e-02
## green_rating        2.177658e-02
## net                -5.986778e-02
## amenities           3.835981e-02
## cd_total_07        -4.131382e-05
## hd_total07         -4.554181e-06
## total_dd_07        -1.738643e-05
## Precipitation       2.215165e-03
## Gas_Costs          -4.588401e+00
## Electricity_Costs   3.232539e+00
## cluster_rent        2.449920e-02
```

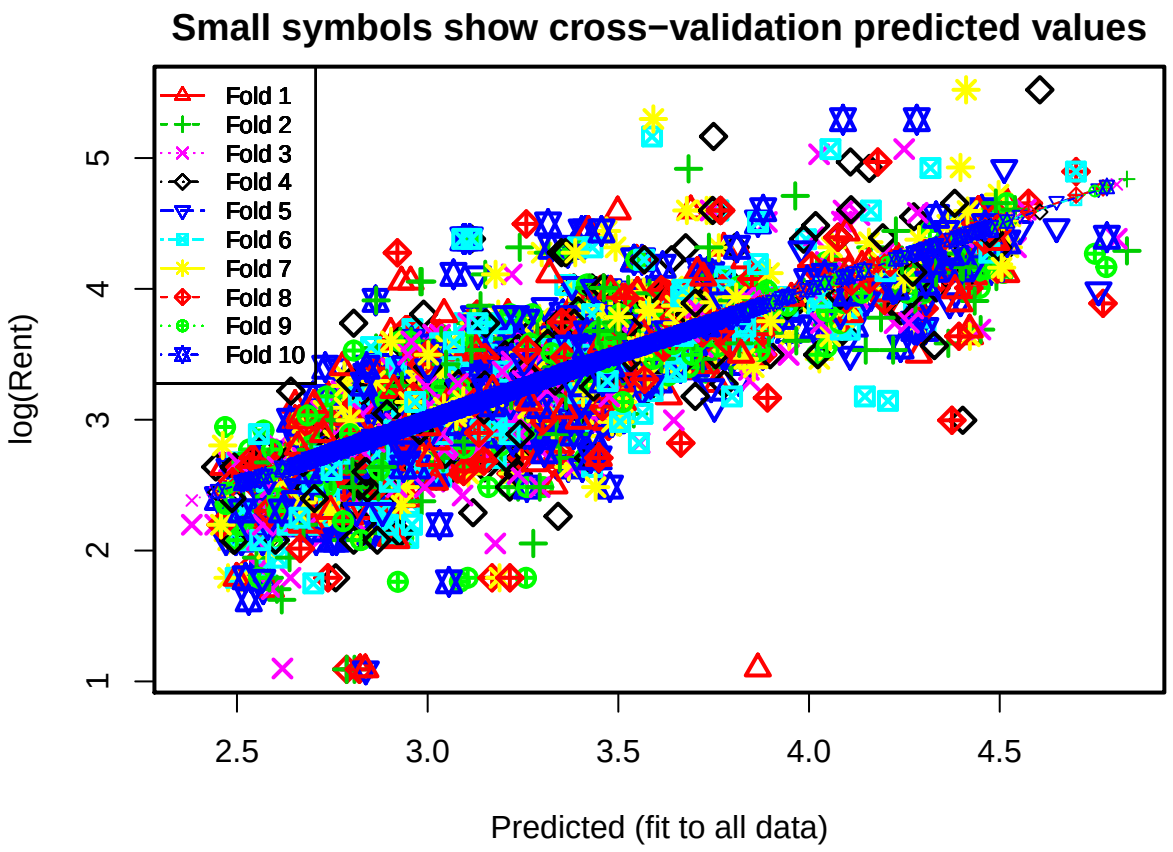
```
#Apply OLS to data
linear_fit = lm(log(Rent) ~ . - CS_PropertyID - LEED -Energystar , data = greenbuildings) #no scaling
summary(linear_fit)
```

```
##
## Call:
## lm(formula = log(Rent) ~ . - CS_PropertyID - LEED - Energystar,
##     data = greenbuildings)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76714 -0.12071  0.00929  0.13152  1.70644
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.406e+00  2.771e-02  86.819  < 2e-16 ***
## cluster            3.940e-05  7.720e-06   5.103 3.42e-07 ***
## size               7.892e-08  1.785e-08   4.421 9.98e-06 ***
## empl_gr            3.504e-03  4.608e-04   7.605 3.19e-14 ***
## leasing_rate       4.850e-04  1.448e-04   3.350 0.000812 ***
## stories            4.224e-04  4.400e-04   0.960 0.337089
## age               -1.049e-03  1.284e-04  -8.173 3.47e-16 ***
## renovated          5.566e-03  6.984e-03   0.797 0.425474
## class_a            1.126e-01  1.192e-02   9.452  < 2e-16 ***
## class_b            5.285e-02  9.329e-03   5.665 1.52e-08 ***
## green_rating       2.978e-02  1.083e-02   2.749 0.005997 **
## net               -4.694e-02  1.614e-02  -2.908 0.003651 **
## amenities          3.966e-02  6.807e-03   5.827 5.86e-09 ***
## cd_total_07       -6.156e-05  3.984e-06 -15.452  < 2e-16 ***
## hd_total07        -2.514e-05  2.437e-06 -10.316  < 2e-16 ***
## total_dd_07              NA         NA      NA       NA
## Precipitation      6.703e-04  4.347e-04   1.542 0.123137
## Gas_Costs          2.219e+00  2.081e+00   1.066 0.286442
## Electricity_Costs -1.569e+00  6.788e-01  -2.312 0.020828 *
## cluster_rent       3.114e-02  3.832e-04  81.271  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2563 on 7801 degrees of freedom
## Multiple R-squared:  0.681,  Adjusted R-squared:  0.6803
## F-statistic: 925.2 on 18 and 7801 DF,  p-value: < 2.2e-16
```

```
cvlm = cv.lm(data = greenbuildings, linear_fit, m=10, plotit = TRUE, printit = FALSE)
```

```
## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading
## Warning in cv.lm(data = greenbuildings, linear_fit, m = 10, plotit = TRUE, :
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```




```
print(linear_fit)
```

```
##
## Call:
## lm(formula = log(Rent) ~ . - CS_PropertyID - LEED - Energystar,
##     data = greenbuildings)
##
## Coefficients:
##       (Intercept)            cluster               size            empl_gr
##         2.406e+00          3.940e-05          7.892e-08          3.504e-03
##      leasing_rate              stories                age          renovated
##         4.850e-04          4.224e-04         -1.049e-03          5.566e-03
##           class_a            class_b       green_rating                net
##         1.126e-01          5.285e-02          2.978e-02         -4.694e-02
##         amenities        cd_total_07        hd_total07       total_dd_07
##         3.966e-02         -6.156e-05         -2.514e-05                NA
##     Precipitation          Gas_Costs  Electricity_Costs       cluster_rent
##         6.703e-04          2.219e+00         -1.569e+00          3.114e-02
```

```
#MSE for OLS = 0.0659
```