R Lab 9. Regression Model Selection

```
> install.packages("leaps")
> library(leaps)
> setwd("C:\\Users\\baron\\Documents\\Teach\\615 Regression\\Data")

> HOMES = read.csv("HOME_SALES.csv")
> attach(HOMES)
```

1. Exhaustive Search

 # This command finds the best model for each p = number of independent variables. The best model is determined by the lowest SSErr. When the command is too long, it will be continued on the next line after "+"

```
> reg.models = regsubsets( SALES_PRICE ~ FINISHED_AREA + BEDROOMS +
+ BATHROOMS + GARAGE_SIZE + YEAR_BUILT + as.factor(STYLE) + LOT_SIZE
+ + AIR_CONDITIONER + POOL + QUALITY + HIGHWAY, data=HOMES )

> summary(reg.models)
         FINISHED_AREA BEDROOMS BATHROOMS GARAGE_SIZE YEAR_BUILT as.factor(STYLE)2 as.factor(STYLE)3 LOT_SIZE
1  ( 1 ) "*"           " "      " "       " "         " "        " "               " "               " "
2  ( 1 ) "*"           " "      " "       " "         "*"        " "               " "               " "
3  ( 1 ) "*"           " "      " "       " "         " "        " "               " "               " "
4  ( 1 ) "*"           " "      " "       " "         "*"        " "               " "               " "
5  ( 1 ) "*"           " "      " "       " "         "*"        " "               " "               "*"
6  ( 1 ) "*"           " "      " "       " "         "*"        " "               "*"               "*"
7  ( 1 ) "*"           " "      " "       " "         "*"        "*"               "*"               "*"
8  ( 1 ) "*"           " "      "*"       " "         "*"        "*"               "*"               "*"
         AIR_CONDITIONERYES POOLYES QUALITYLOW QUALITYMEDIUM HIGHWAYYES
1  ( 1 ) " "                " "     " "        " "           " "
2  ( 1 ) " "                " "     " "        " "           " "
3  ( 1 ) " "                " "     "*"        "*"           " "
4  ( 1 ) " "                " "     "*"        "*"           " "
5  ( 1 ) " "                " "     "*"        "*"           " "
6  ( 1 ) " "                " "     "*"        "*"           " "
7  ( 1 ) " "                " "     "*"        "*"           " "
8  ( 1 ) " "                " "     "*"        "*"           " "
```

 # Next, choose the best model rank p according to some criteria:

```
> summary(reg.models)$adjr2
[1] 0.6708995 0.7171632 0.7870865 0.7994151 0.8155416 0.8211868 0.8225431 0.8238597
> summary(reg.models)$cp
[1] 463.64547 326.02707 118.63868  82.85571  35.97259  20.23787  17.21319  14.32099
> summary(reg.models)$bic
[1] -568.6342 -642.4608 -785.4496 -811.3369 -849.8408 -860.8207 -859.5521 -858.1983
> which.max(summary(reg.models)$adjr2)
[1] 8
> which.min(summary(reg.models)$bic)
[1] 6
```

 # According to adjusted R^2 and Mallows Cp, the best model uses all 8 variables. By the BIC criterion, use 6 variables.

 # Recall that plain R^2 is not a fair measure of performance. It always increases with p:

```
> summary(reg.models)$rsq
[1] 0.6715312 0.7182489 0.7883125 0.8009551 0.8173119 0.8232461 0.8249274 0.8265644
```

 # To use all 11 X-variables available, change the mv option. I'm too lazy to count variables, so I entered a surely larger number (33)

```
> reg.models = regsubsets( SALES_PRICE ~ FINISHED_AREA + BEDROOMS +
+ + BATHROOMS + GARAGE_SIZE + YEAR_BUILT + as.factor(STYLE) + LOT_SIZE
+ + AIR_CONDITIONER + POOL + QUALITY + HIGHWAY, data=HOMES, nvmax=33 )
> which.max(summary(reg.models)$adjr2)
[1] 11
```

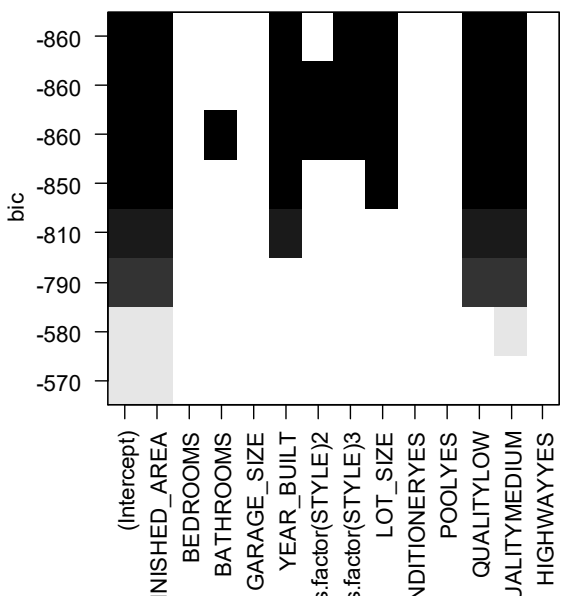
2. Sequential Search

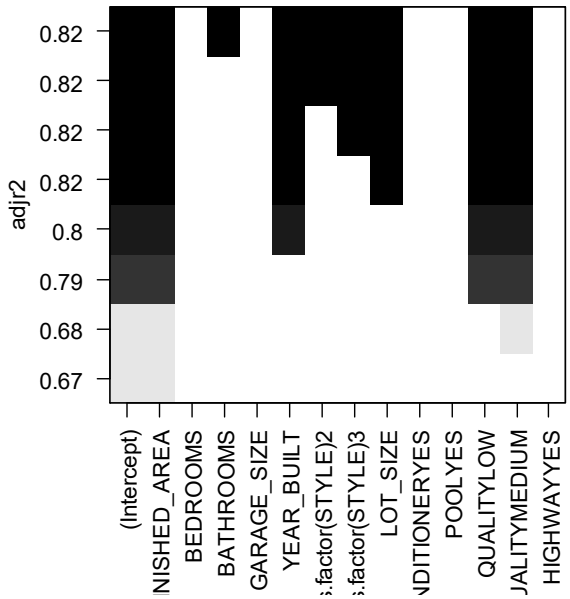
 # For stepwise or backward elimination variable selection, use method="forward" or method="backward".

```
> reg.backward = regsubsets( SALES_PRICE ~ FINISHED_AREA + BEDROOMS +
+ BATHROOMS + GARAGE_SIZE + YEAR_BUILT + as.factor(STYLE) + LOT_SIZE
+ + AIR_CONDITIONER + POOL + QUALITY + HIGHWAY, data=HOMES, method = "backward" )
```

 # There is a nice way to visualize results, ranking models by the chosen "scale". Black color means the variable is included into the model; white means that it is excluded.

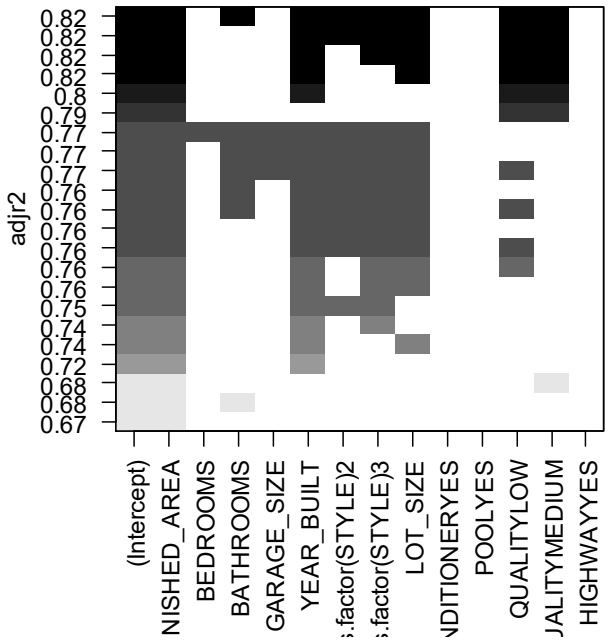
```
> plot(reg.backward)
> plot(reg.backward, scale = "adjr2" )
```





 # To see more models, use option "nbest", which is the number of models of each size p to be compared.

```
> plot(reg.backward, scale = "adjr2" )
> reg.backward = regsubsets( SALES_PRICE ~ FINISHED_AREA + BEDROOMS +
+ BATHROOMS + GARAGE_SIZE + YEAR_BUILT + as.factor(STYLE) + LOT_SIZE
+ + AIR_CONDITIONER + POOL + QUALITY + HIGHWAY, data=HOMES, method = "backward", nbest=3)
```



 # We can also choose the best model by means of a stepwise procedure, starting with one model and ending with another. At each step, the algorithm compares contributions by each variable.

```
> reg.null = lm( SALES_PRICE ~ 1, data=HOMES )
> reg.full = lm( SALES_PRICE ~ . - ID - STYLE + as.factor(STYLE), data=HOMES )
```

 # This is another way of writing the same regression model. Use all variables (.) except those with "- "

 # Here is the forward variable selection:

```
> step( reg.null, scope=list( lower=reg.null, upper=reg.full ), direction="forward" )
Start:  AIC=5144.47
SALES_PRICE ~ 1

                   Df Sum of Sq     RSS    AIC
+ FINISHED_AREA     1   6655486 3255426 4565.3
+ QUALITY           2   6541783 3369129 4585.2
+ BATHROOMS         1   4632615 5278297 4817.6
+ GARAGE_SIZE       1   3308629 6602283 4934.4
+ YEAR_BUILT        1   3058493 6852419 4953.8
+ BEDROOMS          1   1693147 8217765 5048.7
+ as.factor(STYLE)  2   1308922 8601990 5074.5
+ AIR_CONDITIONER   1    825458 9085454 5101.1
+ LOT_SIZE          1    498038 9412873 5119.6
+ POOL              1    213035 9697877 5135.1
<none>                          9910912 5144.5
+ HIGHWAY           1     25746 9885166 5145.1

Step:  AIC=4565.32
SALES_PRICE ~ FINISHED_AREA

                   Df Sum of Sq     RSS    AIC
+ QUALITY           2   1157409 2098016 4340.0
+ YEAR_BUILT        1    463016 2792410 4487.2
+ as.factor(STYLE)  2    386713 2868713 4503.3
+ GARAGE_SIZE       1    273127 2982298 4521.6
+ BATHROOMS         1     96767 3158659 4551.6
+ LOT_SIZE          1     91880 3163546 4552.4
+ AIR_CONDITIONER   1     50865 3204561 4559.1
+ BEDROOMS          1     27613 3227813 4562.9
<none>                          3255426 4565.3
```

```
+ POOL              1      1864 3253561 4567.0
+ HIGHWAY           1        16 3255409 4567.3

Step:  AIC=4339.99
SALES_PRICE ~ FINISHED_AREA + QUALITY

                   Df Sum of Sq     RSS    AIC
+ YEAR_BUILT        1    125299 1972717 4309.8
+ as.factor(STYLE)  2    113542 1984474 4314.9
+ LOT_SIZE          1     97690 2000326 4317.1
+ GARAGE_SIZE       1     63943 2034074 4325.8
+ BATHROOMS         1     37966 2060050 4332.5
<none>                         2098016 4340.0
+ AIR_CONDITIONER   1      7086 2090930 4340.2
+ POOL              1      1261 2096755 4341.7
+ HIGHWAY           1      1233 2096784 4341.7
+ BEDROOMS          1       328 2097688 4341.9

Step:  AIC=4309.85
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT

                   Df Sum of Sq     RSS    AIC
+ LOT_SIZE          1    162111 1810606 4267.1
+ as.factor(STYLE)  2    111543 1861174 4283.5
+ GARAGE_SIZE       1     35423 1937294 4302.4
+ BATHROOMS         1     18673 1954044 4306.9
<none>                         1972717 4309.8
+ HIGHWAY           1      2819 1969897 4311.1
+ POOL              1      1977 1970739 4311.3
+ BEDROOMS          1       947 1971769 4311.6
+ AIR_CONDITIONER   1         1 1972715 4311.8

Step:  AIC=4267.09
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE

                   Df Sum of Sq     RSS    AIC
+ as.factor(STYLE)  2     75477 1735129 4248.9
+ GARAGE_SIZE       1     18407 1792199 4263.8
+ BATHROOMS         1     10996 1799610 4265.9
+ HIGHWAY           1      9076 1801530 4266.5
<none>                         1810606 4267.1
+ POOL              1      5428 1805178 4267.5
+ AIR_CONDITIONER   1      2881 1807725 4268.3
+ BEDROOMS          1      2718 1807888 4268.3

Step:  AIC=4248.86
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE +
    as.factor(STYLE)

                  Df Sum of Sq     RSS    AIC
+ BATHROOMS        1   16224.0 1718905 4246.0
+ HIGHWAY          1   14868.0 1720261 4246.4
+ GARAGE_SIZE      1   13414.0 1721715 4246.8
<none>                         1735129 4248.9
+ POOL             1    4427.5 1730702 4249.5
+ BEDROOMS         1    1195.7 1733933 4250.5
+ AIR_CONDITIONER  1    1139.3 1733990 4250.5

Step:  AIC=4245.95
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE +
    as.factor(STYLE) + BATHROOMS

                  Df Sum of Sq     RSS    AIC
```

```
+ HIGHWAY          1   13979.1 1704926 4243.7
+ GARAGE_SIZE      1   12070.4 1706835 4244.3
<none>                         1718905 4246.0
+ BEDROOMS         1    5561.3 1713344 4246.3
+ POOL             1    2753.1 1716152 4247.1
+ AIR_CONDITIONER  1     854.6 1718051 4247.7

Step:  AIC=4243.69
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE +
    as.factor(STYLE) + BATHROOMS + HIGHWAY

                  Df Sum of Sq     RSS    AIC
+ GARAGE_SIZE      1   12047.4 1692879 4242.0
<none>                         1704926 4243.7
+ BEDROOMS         1    5208.8 1699717 4244.1
+ POOL             1    2520.3 1702406 4244.9
+ AIR_CONDITIONER  1     593.4 1704333 4245.5

Step:  AIC=4241.99
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE +
    as.factor(STYLE) + BATHROOMS + HIGHWAY + GARAGE_SIZE

                  Df Sum of Sq     RSS    AIC
<none>                         1692879 4242.0
+ BEDROOMS         1    5648.2 1687230 4242.2
+ POOL             1    2181.1 1690698 4243.3
+ AIR_CONDITIONER  1     120.0 1692759 4244.0

Call:
lm(formula = SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT +
    LOT_SIZE + as.factor(STYLE) + BATHROOMS + HIGHWAY + GARAGE_SIZE,
    data = HOMES)

Coefficients:
      (Intercept)      FINISHED_AREA         QUALITYLOW      QUALITYMEDIUM         YEAR_BUILT           LOT_SIZE
       -2.346e+03          1.006e-01         -1.409e+02         -1.326e+02          1.249e+00          1.348e-03
as.factor(STYLE)2  as.factor(STYLE)3          BATHROOMS         HIGHWAYYES        GARAGE_SIZE
       -1.592e+01         -3.806e+01          8.281e+00         -3.653e+01          9.397e+00
```

 # The final model contains the following variables: FINISHED_AREA, QUALITY, YEAR_BUILT, LOT_SIZE, STYLE, BATHROOMS, HIGHWAY, and GARAGE_SIZE.

 # Similarly, we can conduct variable selection using backward elimination …

```
> step( reg.full, scope=list( lower=reg.null, upper=reg.full ), direction="backward" )
Start:  AIC=4245.46
SALES_PRICE ~ (ID + FINISHED_AREA + BEDROOMS + BATHROOMS + GARAGE_SIZE +
    YEAR_BUILT + STYLE + LOT_SIZE + AIR_CONDITIONER + POOL +
    QUALITY + HIGHWAY) - ID - STYLE + as.factor(STYLE)

                   Df Sum of Sq     RSS    AIC
- AIR_CONDITIONER   1       220 1684897 4243.5
- POOL              1      2268 1686946 4244.2
- BEDROOMS          1      5941 1690618 4245.3
<none>                         1684678 4245.5
- GARAGE_SIZE       1     11577 1696254 4247.0
- HIGHWAY           1     13194 1697872 4247.5
- BATHROOMS         1     16784 1701462 4248.6
- as.factor(STYLE)  2     76668 1761346 4264.7
- LOT_SIZE          1    115690 1800367 4278.1
- YEAR_BUILT        1    129778 1814456 4282.2
- QUALITY           2    532390 2217067 4384.8
- FINISHED_AREA     1    620617 2305294 4407.2

Step:  AIC=4243.52
SALES_PRICE ~ FINISHED_AREA + BEDROOMS + BATHROOMS + GARAGE_SIZE +
```

```
    YEAR_BUILT + LOT_SIZE + POOL + QUALITY + HIGHWAY + as.factor(STYLE)

                   Df Sum of Sq     RSS    AIC
- POOL              1      2333 1687230 4242.2
- BEDROOMS          1      5800 1690698 4243.3
<none>                         1684897 4243.5
- GARAGE_SIZE       1     12136 1697034 4245.3
- HIGHWAY           1     13363 1698261 4245.6
- BATHROOMS         1     16787 1701684 4246.7
- as.factor(STYLE)  2     77691 1762589 4263.1
- LOT_SIZE          1    116770 1801668 4276.5
- YEAR_BUILT        1    135320 1820218 4281.8
- QUALITY           2    532765 2217662 4382.9
- FINISHED_AREA     1    620687 2305584 4405.2

Step:  AIC=4242.25
SALES_PRICE ~ FINISHED_AREA + BEDROOMS + BATHROOMS + GARAGE_SIZE +
    YEAR_BUILT + LOT_SIZE + QUALITY + HIGHWAY + as.factor(STYLE)

                   Df Sum of Sq     RSS    AIC
- BEDROOMS          1      5648 1692879 4242.0
<none>                         1687230 4242.2
- GARAGE_SIZE       1     12487 1699717 4244.1
- HIGHWAY           1     13589 1700820 4244.4
- BATHROOMS         1     18250 1705480 4245.9
- as.factor(STYLE)  2     78845 1766076 4262.1
- LOT_SIZE          1    114823 1802054 4274.6
- YEAR_BUILT        1    133452 1820683 4280.0
- QUALITY           2    533252 2220482 4381.6
- FINISHED_AREA     1    626691 2313922 4405.1

Step:  AIC=4241.99
SALES_PRICE ~ FINISHED_AREA + BATHROOMS + GARAGE_SIZE + YEAR_BUILT +
    LOT_SIZE + QUALITY + HIGHWAY + as.factor(STYLE)

                   Df Sum of Sq     RSS    AIC
<none>                         1692879 4242.0
- GARAGE_SIZE       1     12047 1704926 4243.7
- HIGHWAY           1     13956 1706835 4244.3
- BATHROOMS         1     14033 1706912 4244.3
- as.factor(STYLE)  2     80912 1773791 4262.4
- LOT_SIZE          1    113074 1805952 4273.7
- YEAR_BUILT        1    134752 1827631 4280.0
- QUALITY           2    563738 2256616 4388.0
- FINISHED_AREA     1    635028 2327907 4406.3

Call:
lm(formula = SALES_PRICE ~ FINISHED_AREA + BATHROOMS + GARAGE_SIZE +
    YEAR_BUILT + LOT_SIZE + QUALITY + HIGHWAY + as.factor(STYLE),
    data = HOMES)

Coefficients:
      (Intercept)      FINISHED_AREA          BATHROOMS        GARAGE_SIZE
       -2.346e+03          1.006e-01          8.281e+00          9.397e+00
       YEAR_BUILT           LOT_SIZE        QUALITYLOW      QUALITYMEDIUM
        1.249e+00          1.348e-03         -1.409e+02         -1.326e+02
       HIGHWAYYES  as.factor(STYLE)2  as.factor(STYLE)3
       -3.653e+01         -1.592e+01         -3.806e+01
```

 # … and stepwise regression, where the algorithm considers either adding or removing variables at each step:

```
> step( reg.null, scope=list( lower=reg.null, upper=reg.full ), direction="both" )
Start:  AIC=5144.47
SALES_PRICE ~ 1

                   Df Sum of Sq     RSS    AIC
+ FINISHED_AREA     1   6655486 3255426 4565.3
+ QUALITY           2   6541783 3369129 4585.2
+ BATHROOMS         1   4632615 5278297 4817.6
+ GARAGE_SIZE       1   3308629 6602283 4934.4
+ YEAR_BUILT        1   3058493 6852419 4953.8
+ BEDROOMS          1   1693147 8217765 5048.7
+ as.factor(STYLE)  2   1308922 8601990 5074.5
+ AIR_CONDITIONER   1    825458 9085454 5101.1
+ LOT_SIZE          1    498038 9412873 5119.6
+ POOL              1    213035 9697877 5135.1
```

```
<none>                          9910912 5144.5
+ HIGHWAY           1     25746 9885166 5145.1

Step:  AIC=4565.32
SALES_PRICE ~ FINISHED_AREA

                   Df Sum of Sq     RSS    AIC
+ QUALITY           2   1157409 2098016 4340.0
+ YEAR_BUILT        1    463016 2792410 4487.2
+ as.factor(STYLE)  2    386713 2868713 4503.3
+ GARAGE_SIZE       1    273127 2982298 4521.6
+ BATHROOMS         1     96767 3158659 4551.6
+ LOT_SIZE          1     91880 3163546 4552.4
+ AIR_CONDITIONER   1     50865 3204561 4559.1
+ BEDROOMS          1     27613 3227813 4562.9
<none>                          3255426 4565.3
+ POOL              1      1864 3253561 4567.0
+ HIGHWAY           1        16 3255409 4567.3
- FINISHED_AREA     1   6655486 9910912 5144.5

Step:  AIC=4339.99
SALES_PRICE ~ FINISHED_AREA + QUALITY

                   Df Sum of Sq     RSS    AIC
+ YEAR_BUILT        1    125299 1972717 4309.8
+ as.factor(STYLE)  2    113542 1984474 4314.9
+ LOT_SIZE          1     97690 2000326 4317.1
+ GARAGE_SIZE       1     63943 2034074 4325.8
+ BATHROOMS         1     37966 2060050 4332.5
<none>                          2098016 4340.0
+ AIR_CONDITIONER   1      7086 2090930 4340.2
+ POOL              1      1261 2096755 4341.7
+ HIGHWAY           1      1233 2096784 4341.7
+ BEDROOMS          1       328 2097688 4341.9
- QUALITY           2   1157409 3255426 4565.3
- FINISHED_AREA     1   1271112 3369129 4585.2

Step:  AIC=4309.85
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT

                   Df Sum of Sq     RSS    AIC
+ LOT_SIZE          1    162111 1810606 4267.1
+ as.factor(STYLE)  2    111543 1861174 4283.5
+ GARAGE_SIZE       1     35423 1937294 4302.4
+ BATHROOMS         1     18673 1954044 4306.9
<none>                          1972717 4309.8
+ HIGHWAY           1      2819 1969897 4311.1
+ POOL              1      1977 1970739 4311.3
+ BEDROOMS          1       947 1971769 4311.6
+ AIR_CONDITIONER   1         1 1972715 4311.8
- YEAR_BUILT        1    125299 2098016 4340.0
- QUALITY           2    819693 2792410 4487.2
- FINISHED_AREA     1   1247408 3220125 4563.6

Step:  AIC=4267.09
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE

                   Df Sum of Sq     RSS    AIC
+ as.factor(STYLE)  2     75477 1735129 4248.9
+ GARAGE_SIZE       1     18407 1792199 4263.8
+ BATHROOMS         1     10996 1799610 4265.9
+ HIGHWAY           1      9076 1801530 4266.5
<none>                          1810606 4267.1
+ POOL              1      5428 1805178 4267.5
+ AIR_CONDITIONER   1      2881 1807725 4268.3
+ BEDROOMS          1      2718 1807888 4268.3
- LOT_SIZE          1    162111 1972717 4309.8
- YEAR_BUILT        1    189720 2000326 4317.1
- QUALITY           2    786607 2597213 4451.4
- FINISHED_AREA     1   1125143 2935749 4517.4

Step:  AIC=4248.86
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE +
    as.factor(STYLE)

                   Df Sum of Sq     RSS    AIC
+ BATHROOMS         1     16224 1718905 4246.0
+ HIGHWAY           1     14868 1720261 4246.4
+ GARAGE_SIZE       1     13414 1721715 4246.8
```

```
<none>                          1735129 4248.9
+ POOL              1      4427 1730702 4249.5
+ BEDROOMS          1      1196 1733933 4250.5
+ AIR_CONDITIONER   1      1139 1733990 4250.5
- as.factor(STYLE)  2     75477 1810606 4267.1
- LOT_SIZE          1    126045 1861174 4283.5
- YEAR_BUILT        1    177680 1912809 4297.7
- QUALITY           2    599006 2334136 4399.7
- FINISHED_AREA     1    955248 2690378 4475.8

Step:  AIC=4245.95
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE +
    as.factor(STYLE) + BATHROOMS

                   Df Sum of Sq     RSS    AIC
+ HIGHWAY           1     13979 1704926 4243.7
+ GARAGE_SIZE       1     12070 1706835 4244.3
<none>                          1718905 4246.0
+ BEDROOMS          1      5561 1713344 4246.3
+ POOL              1      2753 1716152 4247.1
+ AIR_CONDITIONER   1       855 1718051 4247.7
- BATHROOMS         1     16224 1735129 4248.9
- as.factor(STYLE)  2     80705 1799610 4265.9
- LOT_SIZE          1    117122 1836027 4278.4
- YEAR_BUILT        1    151782 1870687 4288.1
- QUALITY           2    601044 2319949 4398.5
- FINISHED_AREA     1    688877 2407782 4419.9

Step:  AIC=4243.69
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE +
    as.factor(STYLE) + BATHROOMS + HIGHWAY

                   Df Sum of Sq     RSS    AIC
+ GARAGE_SIZE       1     12047 1692879 4242.0
<none>                          1704926 4243.7
+ BEDROOMS          1      5209 1699717 4244.1
+ POOL              1      2520 1702406 4244.9
+ AIR_CONDITIONER   1       593 1704333 4245.5
- HIGHWAY           1     13979 1718905 4246.0
- BATHROOMS         1     15335 1720261 4246.4
- as.factor(STYLE)  2     86288 1791214 4265.5
- LOT_SIZE          1    123460 1828386 4278.2
- YEAR_BUILT        1    158009 1862935 4288.0
- QUALITY           2    596493 2301419 4396.3
- FINISHED_AREA     1    687244 2392170 4418.5

Step:  AIC=4241.99
SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT + LOT_SIZE +
    as.factor(STYLE) + BATHROOMS + HIGHWAY + GARAGE_SIZE

                   Df Sum of Sq     RSS    AIC
<none>                          1692879 4242.0
+ BEDROOMS          1      5648 1687230 4242.2
+ POOL              1      2181 1690698 4243.3
- GARAGE_SIZE       1     12047 1704926 4243.7
+ AIR_CONDITIONER   1       120 1692759 4244.0
- HIGHWAY           1     13956 1706835 4244.3
- BATHROOMS         1     14033 1706912 4244.3
- as.factor(STYLE)  2     80912 1773791 4262.4
- LOT_SIZE          1    113074 1805952 4273.7
- YEAR_BUILT        1    134752 1827631 4280.0
- QUALITY           2    563738 2256616 4388.0
- FINISHED_AREA     1    635028 2327907 4406.3

Call:
lm(formula = SALES_PRICE ~ FINISHED_AREA + QUALITY + YEAR_BUILT +
    LOT_SIZE + as.factor(STYLE) + BATHROOMS + HIGHWAY + GARAGE_SIZE,
    data = HOMES)

Coefficients:
      (Intercept)      FINISHED_AREA         QUALITYLOW      QUALITYMEDIUM
       -2.346e+03          1.006e-01         -1.409e+02         -1.326e+02
       YEAR_BUILT           LOT_SIZE  as.factor(STYLE)2  as.factor(STYLE)3
        1.249e+00          1.348e-03         -1.592e+01         -3.806e+01
        BATHROOMS         HIGHWAYYES        GARAGE_SIZE
        8.281e+00         -3.653e+01          9.397e+00
```

3. Model Validation

 # In the absence of another dataset, we'll split the data into the training and testing subsets.

```
> n = length(SALES_PRICE)
> n
[1] 522
```

 # Randomly choose indices for the testing (validation) data

```
> testing = sample(n,100)
> testing
  [1] 461 439 174 273 484 214   4 200  17 114  93 145 272 423 177
 [16] 199 426 239  97 100  81 203 311 278 108  91  27 349 433 309
 [31]  77 429 274 305 226 299 329 499 173  11 320 211 477 503 353
 [46]  16 240 236 153  63 248 404 310 187 205 370  40  78 209 427
 [61]   1  22 512 229 492 232 166  52 488 262 237 405 343 358 186
 [76] 411 419 365 480 354  92 231 500 415 453 283 441 475 223 104
 [91] 507 220  36 241 348 401  54 277 339  58
```

 # The remaining indices will be training

```
> training = -testing
> reg = lm( SALES_PRICE ~ . - ID - STYLE + as.factor(STYLE), data=HOMES, subset=training)
> Yhat = predict(reg,HOMES)
> length(Yhat)
[1] 522
```

 # Calculating the mean squared prediction error

```
> MSPE = mean((SALES_PRICE[testing] - Yhat[testing])^2)
> MSPE
[1] 3884.3
```

 # Let's compare with a reduced model that excludes the architectural style.

```
> reg1 = lm( SALES_PRICE ~ . - ID - STYLE, data=HOMES, subset=training )
> Yhat = predict(reg1,HOMES)
> MSPE = mean((SALES_PRICE[testing] - Yhat[testing])^2)
> MSPE
[1] 3865.107
```

 # The reduced model produces a lower MSPE, so it is better for prediction!

4. Visualization – scatterplot matrix

 # Scatterplot matrix – a way to visualize relations between the response and predictor variables.
 # It is used to show (1) whether there is a relation between Y and each X_j, (2) whether this relation is linear
 # or nonlinear, (3) whether there may be strong multicollinearity.

```
> par(mfrow=c(6,6))
> plot(HOMES[c(2,3,4,5,7,9)])
```