# بنام خدا

*Hossein Vatani*

*May 31, 2016*

# Variable Selection

# انتخاب متغیرهای تشریح

تابدینجای بحث، در تمامی مثالها فرض براین بوده که برای ساخت نمونه پسرفت(Regression Model) تمامی متغیرهای تشریح موجود، در ساخت نمونه مورد استفاده هستند. لیکن در غالب موارد موجود در دنیای حقیقی، اینگونه نیست و انتخاب متغیرهای تشریح مربوط و موثر، خود یکی از مراحل تشکیل نمونه پسرفت می باشد.

دراین زمینه دو روش کلی وجود دارد:۱-تمامی حالات ممکن و ۲-دستیابی خودکار

# روش بررسی تمامی حالات ممکن

در روش اول: تمامی حالتهای نمونه پسرفت از زیر مجموعه های متغیرهای تشریح ساخته می شوند و با توجه به عواملی مشخص(مانند $Adjusted\ R^2, AIC, BIC$) به نمونه امتیازی تعلق می گیرد که در نهایت به ما در انتخاب نمونه مناسب کمک می کند.

دستور ()regsubsets از بسته leaps امکان انجام این کار را مهیا می سازد که خروجی آن بصورت تصویری و بر اساس عامل درخواستی در دستور می باشد. در تصویر خروجی بهترین نمونه، نمونه ای می باشد که کمترین مقدار را دارد.

مثال- اطلاعاتی در مورد قیمت خانه هایی با مشخصاتی مانند تعداد اتاق،گاراژ،حمام، قسمت تفریحی و... جمع آوری ده است. می خواهیم بدانیم که کدامیک از عوامل جمع آوری شده بر قیمت خانه تاثیر گذار می باشند.

```
#price,lotsize,bedrooms,bathrms,stories,driveway,recroom,fullbase,gashw,airco
,garagepl,prefarea
```

دراینجا ما متغیر قیمت را متغیر پاسخ درنظر میگیریم و سایر متغیرها را بعنوان متغیرهای تشریح .

```
House=read.csv("./File/Housing.txt",header = TRUE,sep = ",")
str(House)
```

```
## 'data.frame':    101 obs. of  12 variables:
##  $ price   : num  42000 38500 49500 60500 61000 66000 66000 69000 83800 88
500 ...
##  $ lotsize : num  5850 4000 3060 6650 6360 4160 3880 4160 4800 5500 ...
##  $ bedrooms: num  3 2 3 3 2 3 3 3 3 3 ...
##  $ bathrms : num  1 1 1 1 1 1 2 1 1 2 ...
##  $ stories : num  2 1 1 2 1 1 2 3 1 4 ...
##  $ driveway: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ recroom : num  0 0 0 1 0 1 0 0 1 1 ...
##  $ fullbase: num  1 0 0 0 0 1 1 0 1 0 ...
##  $ gashw   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ airco   : num  0 0 0 0 0 1 0 0 0 1 ...
##  $ garagepl: num  1 0 0 0 0 0 2 0 0 1 ...
##  $ prefarea: num  0 0 0 0 0 0 0 0 0 0 ...
```

دستور مورد نظر regsubsets می باشد که پس از اجرای آن باید با استفاده خاصی از دستور plot زیر مجموعه های ساخته شده را مشاهده کرد.

توجه شود که در نمودارها هرکدام که امتیاز بالاتری دارد(نمره منفی کمتری دارد) نمونه پسرفت مناسب تر لذا آن متغیرهای تشریح اثرگذار تر می باشند.

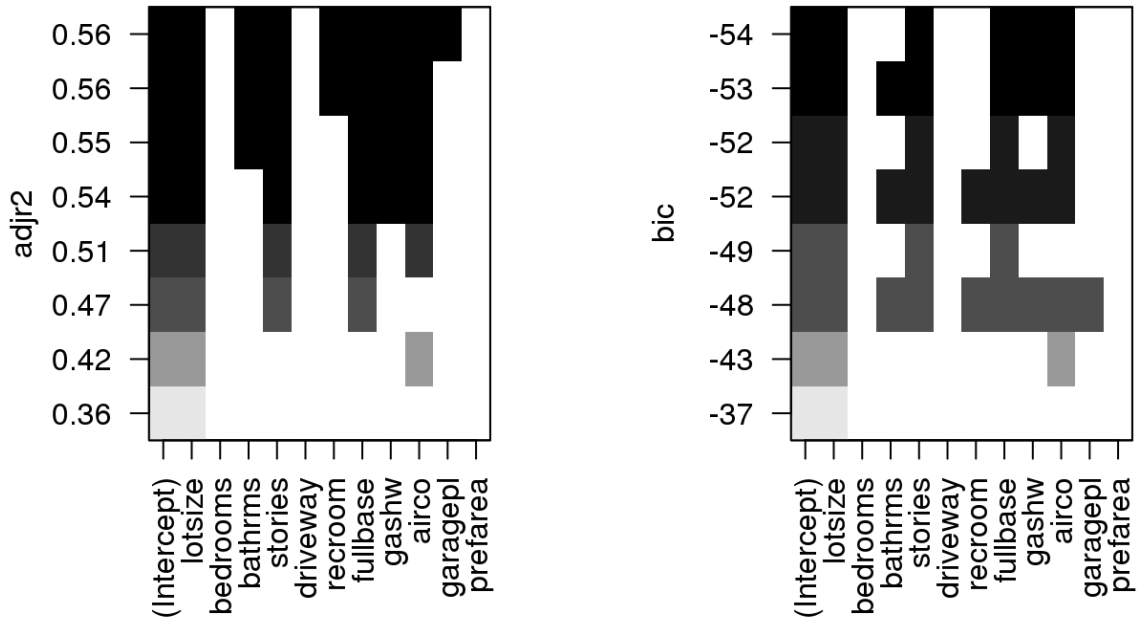
```
require(leaps)
```

```
## Loading required package: leaps
```

```
HouseLeaps=regsubsets(price~lotsize+bedrooms+bathrms+stories+driveway+recroom
+fullbase+gashw+airco+garagepl+prefarea,data=House)
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 1 linear dependencies found
```

```
par(mfrow=c(1,2))
plot(HouseLeaps,scale="adjr2")
plot(HouseLeaps,scale="bic")
```




همانطور که از نتایج بر می آید، در حالتی که adjusted R-square creteria بصورت کمترین حالت در نظر گرفته شود، عامل های موثر lotsize,bathrooms,stories,recroom,fullbase,gashw,airco,garagepl می باشد و همچنین در حالتی که مقدار BIC بصورت کمترین درنظر گرفته شود عاملهای lotsize,fullbase,gashw,airco عاملهای موثر می باشند.

# روش خودکار

در این روش ما به سه طریق می توانیم عمل کنیم : ۱- روبجلو ۲-روبعقب ۳-هردوطریق

بدین صورت که در ابتدای امر
دو نمونه پسرفت ایجاد می کنیم که در یکی هیچ متغیر تشریحی وجود ندارد و در دیگری تمامی متغیرهای تشریح حاضرند.(نمونه صفر و نمونه کامل) سپس بطریق یکی از سه روش مذکور نمونه را بصورت مرحله به مرحله ادامه می دهیم تا نتیجه خوبی حاصل شود.

تابع step() برای این موضوع قابل استفاده می باشد.

```
Null=lm(price~1,data=House)
Full=lm(price~.,data=House)
step(Null,scope = list(lower=Null,upper=Full),direction = "forward")
```

```
## Start:  AIC=1994.04
## price ~ 1
##
##            Df  Sum of Sq        RSS    AIC
## + lotsize   1 1.6481e+10 2.8324e+10 1950.2
## + garagepl  1 8.3693e+09 3.6435e+10 1975.4
## + airco     1 7.5501e+09 3.7255e+10 1977.6
## + bedrooms  1 5.2636e+09 3.9541e+10 1983.5
## + fullbase  1 4.5330e+09 4.0272e+10 1985.4
## + bathrms   1 3.7721e+09 4.1033e+10 1987.2
## + stories   1 3.6259e+09 4.1179e+10 1987.6
## + recroom   1 3.2740e+09 4.1531e+10 1988.5
## + driveway  1 2.7830e+09 4.2022e+10 1989.6
## + gashw     1 1.6857e+09 4.3119e+10 1992.2
## <none>                   4.4805e+10 1994.0
##
## Step:  AIC=1950.18
## price ~ lotsize
##
##            Df  Sum of Sq        RSS    AIC
## + airco     1 2911573123 2.5412e+10 1941.3
## + fullbase  1 2330699770 2.5993e+10 1943.6
## + stories   1 2189590547 2.6134e+10 1944.1
## + gashw     1 1943587513 2.6380e+10 1945.1
## + bedrooms  1 1919194619 2.6404e+10 1945.2
## + recroom   1 1339851600 2.6984e+10 1947.3
## + bathrms   1 1046105948 2.7278e+10 1948.4
## + garagepl  1  725592513 2.7598e+10 1949.6
## <none>                   2.8324e+10 1950.2
## + driveway  1  124771002 2.8199e+10 1951.7
##
## Step:  AIC=1941.33
## price ~ lotsize + airco
##
##            Df  Sum of Sq        RSS    AIC
## + gashw     1 2248267198 2.3164e+10 1934.1
## + fullbase  1 1927017274 2.3485e+10 1935.5
## + bedrooms  1 1719023252 2.3693e+10 1936.3
## + stories   1 1602036506 2.3810e+10 1936.8
## + bathrms   1 1287074904 2.4125e+10 1938.1
## + recroom   1 1184825446 2.4227e+10 1938.6
## + garagepl  1  503712543 2.4908e+10 1941.3
## <none>                   2.5412e+10 1941.3
## + driveway  1   80588517 2.5331e+10 1943.0
##
## Step:  AIC=1934.07
```

```
## price ~ lotsize + airco + gashw
##
##            Df  Sum of Sq        RSS    AIC
## + bedrooms  1 1551252668 2.1613e+10 1929.1
## + bathrms   1 1427801587 2.1736e+10 1929.7
## + fullbase  1 1426812568 2.1737e+10 1929.7
## + stories   1 1350969630 2.1813e+10 1930.1
## + recroom   1 1332137092 2.1832e+10 1930.2
## <none>                   2.3164e+10 1934.1
## + garagepl  1  333563704 2.2830e+10 1934.6
## + driveway  1   33605346 2.3130e+10 1935.9
##
## Step:  AIC=1929.14
## price ~ lotsize + airco + gashw + bedrooms
##
##            Df  Sum of Sq        RSS    AIC
## + recroom   1 1350395461 2.0262e+10 1924.7
## + fullbase  1 1301681354 2.0311e+10 1924.9
## + bathrms   1  746726575 2.0866e+10 1927.6
## <none>                   2.1613e+10 1929.1
## + stories   1  388293618 2.1224e+10 1929.3
## + garagepl  1  332426659 2.1280e+10 1929.6
## + driveway  1  107524018 2.1505e+10 1930.6
##
## Step:  AIC=1924.68
## price ~ lotsize + airco + gashw + bedrooms + recroom
##
##            Df Sum of Sq        RSS    AIC
## + fullbase  1 635677235 1.9626e+10 1923.5
## + stories   1 503993901 1.9758e+10 1924.2
## + bathrms   1 497586346 1.9765e+10 1924.2
## <none>                  2.0262e+10 1924.7
## + garagepl  1 362208726 1.9900e+10 1924.9
## + driveway  1 145898498 2.0116e+10 1926.0
##
## Step:  AIC=1923.5
## price ~ lotsize + airco + gashw + bedrooms + recroom + fullbase
##
##            Df Sum of Sq        RSS    AIC
## + stories   1 857308148 1.8769e+10 1921.0
## + bathrms   1 454768006 1.9172e+10 1923.2
## <none>                  1.9626e+10 1923.5
## + garagepl  1 271231258 1.9355e+10 1924.1
## + driveway  1 196008938 1.9430e+10 1924.5
##
## Step:  AIC=1921.03
```

```
## price ~ lotsize + airco + gashw + bedrooms + recroom + fullbase +
##     stories
##
##            Df Sum of Sq        RSS    AIC
## + garagepl  1 421593247 1.8348e+10 1920.8
## + bathrms   1 409766032 1.8359e+10 1920.8
## <none>                  1.8769e+10 1921.0
## + driveway  1 109256718 1.8660e+10 1922.5
##
## Step:  AIC=1920.76
## price ~ lotsize + airco + gashw + bedrooms + recroom + fullbase +
##     stories + garagepl
##
##            Df Sum of Sq        RSS    AIC
## <none>                  1.8348e+10 1920.8
## + bathrms   1 216431185 1.8131e+10 1921.6
## + driveway  1 102185716 1.8245e+10 1922.2
```

```
##
## Call:
## lm(formula = price ~ lotsize + airco + gashw + bedrooms + recroom +
##     fullbase + stories + garagepl, data = House)
##
## Coefficients:
## (Intercept)      lotsize        airco        gashw     bedrooms
##    6081.372        5.201    10216.861    38965.014     2638.114
##     recroom     fullbase      stories     garagepl
##    7411.125     6890.120     5593.040     2950.457
```

با توجه به خروجی نهایی می توان متغیرهای
lotsize,airco,gashw,bedrooms,recroom,fullbase,stories,garagepl بعنوان متغیرهای تشریح مربوط در نظر گرفت. روش فوق دو حالت دیگر نیز دارد که در ادامه می آیند.

```
step(Full,scope = list(lower=Null,upper=Full),direction = "backward")
```

```
## Start:  AIC=1922.74
## price ~ lotsize + bedrooms + bathrms + stories + driveway + recroom +
##     fullbase + gashw + airco + garagepl + prefarea
##
##
## Step:  AIC=1922.74
## price ~ lotsize + bedrooms + bathrms + stories + driveway + recroom +
##     fullbase + gashw + airco + garagepl
##
##             Df  Sum of Sq        RSS    AIC
## - driveway   1  150804630 1.8131e+10 1921.6
## - bedrooms   1  190179360 1.8170e+10 1921.8
## - garagepl   1  206221763 1.8187e+10 1921.9
## - bathrms    1  265050099 1.8245e+10 1922.2
## <none>                   1.7980e+10 1922.7
## - recroom    1  600623376 1.8581e+10 1924.0
## - stories    1  811177823 1.8791e+10 1925.2
## - fullbase   1  898258632 1.8879e+10 1925.6
## - gashw      1 1460359833 1.9441e+10 1928.5
## - airco      1 2013194739 1.9994e+10 1931.3
## - lotsize    1 3815537999 2.1796e+10 1940.0
##
## Step:  AIC=1921.57
## price ~ lotsize + bedrooms + bathrms + stories + recroom + fullbase +
##     gashw + airco + garagepl
##
##             Df  Sum of Sq        RSS    AIC
## - bedrooms   1  152463324 1.8284e+10 1920.4
## - bathrms    1  216431185 1.8348e+10 1920.8
## - garagepl   1  228258400 1.8359e+10 1920.8
## <none>                   1.8131e+10 1921.6
## - recroom    1  601352384 1.8732e+10 1922.8
## - fullbase   1  869577272 1.9001e+10 1924.3
## - stories    1  927185676 1.9058e+10 1924.6
## - gashw      1 1519534437 1.9651e+10 1927.6
## - airco      1 2033627009 2.0165e+10 1930.2
## - lotsize    1 4666704391 2.2798e+10 1942.5
##
## Step:  AIC=1920.41
## price ~ lotsize + bathrms + stories + recroom + fullbase + gashw +
##     airco + garagepl
##
##             Df  Sum of Sq        RSS    AIC
## - garagepl   1  224584598 1.8508e+10 1919.6
## - bathrms    1  326899778 1.8610e+10 1920.2
## <none>                   1.8284e+10 1920.4
```

```
## - recroom   1  570825346 1.8854e+10 1921.5
## - fullbase  1 1002332508 1.9286e+10 1923.8
## - gashw     1 1518112377 1.9802e+10 1926.4
## - stories   1 1742275988 2.0026e+10 1927.5
## - airco     1 2014085962 2.0298e+10 1928.9
## - lotsize   1 4970001189 2.3254e+10 1942.5
##
## Step:  AIC=1919.63
## price ~ lotsize + bathrms + stories + recroom + fullbase + gashw +
##     airco
##
##            Df  Sum of Sq        RSS    AIC
## <none>                   1.8508e+10 1919.6
## - recroom   1  509346546 1.9018e+10 1920.3
## - bathrms   1  570543776 1.9079e+10 1920.7
## - fullbase  1 1063279568 1.9571e+10 1923.2
## - stories   1 1576783001 2.0085e+10 1925.8
## - gashw     1 1663725738 2.0172e+10 1926.2
## - airco     1 2246772052 2.0755e+10 1929.1
## - lotsize   1 7819998209 2.6328e+10 1952.9
```

```
##
## Call:
## lm(formula = price ~ lotsize + bathrms + stories + recroom +
##     fullbase + gashw + airco, data = House)
##
## Coefficients:
## (Intercept)      lotsize      bathrms      stories      recroom
##    3898.408        5.874     5953.176     6033.802     6239.980
##    fullbase        gashw        airco
##    7422.860    41928.529    11121.234
```

```
step(Null,scope = list(lower=Null,upper=Full),direction = "both")
```

```
## Start:  AIC=1994.04
## price ~ 1
##
##            Df  Sum of Sq        RSS    AIC
## + lotsize   1 1.6481e+10 2.8324e+10 1950.2
## + garagepl  1 8.3693e+09 3.6435e+10 1975.4
## + airco     1 7.5501e+09 3.7255e+10 1977.6
## + bedrooms  1 5.2636e+09 3.9541e+10 1983.5
## + fullbase  1 4.5330e+09 4.0272e+10 1985.4
## + bathrms   1 3.7721e+09 4.1033e+10 1987.2
## + stories   1 3.6259e+09 4.1179e+10 1987.6
## + recroom   1 3.2740e+09 4.1531e+10 1988.5
## + driveway  1 2.7830e+09 4.2022e+10 1989.6
## + gashw     1 1.6857e+09 4.3119e+10 1992.2
## <none>                   4.4805e+10 1994.0
##
## Step:  AIC=1950.18
## price ~ lotsize
##
##            Df  Sum of Sq        RSS    AIC
## + airco     1 2.9116e+09 2.5412e+10 1941.3
## + fullbase  1 2.3307e+09 2.5993e+10 1943.6
## + stories   1 2.1896e+09 2.6134e+10 1944.1
## + gashw     1 1.9436e+09 2.6380e+10 1945.1
## + bedrooms  1 1.9192e+09 2.6404e+10 1945.2
## + recroom   1 1.3399e+09 2.6984e+10 1947.3
## + bathrms   1 1.0461e+09 2.7278e+10 1948.4
## + garagepl  1 7.2559e+08 2.7598e+10 1949.6
## <none>                   2.8324e+10 1950.2
## + driveway  1 1.2477e+08 2.8199e+10 1951.7
## - lotsize   1 1.6481e+10 4.4805e+10 1994.0
##
## Step:  AIC=1941.33
## price ~ lotsize + airco
##
##            Df  Sum of Sq        RSS    AIC
## + gashw     1 2.2483e+09 2.3164e+10 1934.1
## + fullbase  1 1.9270e+09 2.3485e+10 1935.5
## + bedrooms  1 1.7190e+09 2.3693e+10 1936.3
## + stories   1 1.6020e+09 2.3810e+10 1936.8
## + bathrms   1 1.2871e+09 2.4125e+10 1938.1
## + recroom   1 1.1848e+09 2.4227e+10 1938.6
## + garagepl  1 5.0371e+08 2.4908e+10 1941.3
## <none>                   2.5412e+10 1941.3
## + driveway  1 8.0589e+07 2.5331e+10 1943.0
## - airco     1 2.9116e+09 2.8324e+10 1950.2
```

```
## - lotsize   1 1.1843e+10 3.7255e+10 1977.6
##
## Step:  AIC=1934.07
## price ~ lotsize + airco + gashw
##
##            Df  Sum of Sq        RSS    AIC
## + bedrooms  1 1.5513e+09 2.1613e+10 1929.1
## + bathrms   1 1.4278e+09 2.1736e+10 1929.7
## + fullbase  1 1.4268e+09 2.1737e+10 1929.7
## + stories   1 1.3510e+09 2.1813e+10 1930.1
## + recroom   1 1.3321e+09 2.1832e+10 1930.2
## <none>                   2.3164e+10 1934.1
## + garagepl  1 3.3356e+08 2.2830e+10 1934.6
## + driveway  1 3.3605e+07 2.3130e+10 1935.9
## - gashw     1 2.2483e+09 2.5412e+10 1941.3
## - airco     1 3.2163e+09 2.6380e+10 1945.1
## - lotsize   1 1.1907e+10 3.5070e+10 1973.5
##
## Step:  AIC=1929.14
## price ~ lotsize + airco + gashw + bedrooms
##
##            Df  Sum of Sq        RSS    AIC
## + recroom   1 1350395461 2.0262e+10 1924.7
## + fullbase  1 1301681354 2.0311e+10 1924.9
## + bathrms   1  746726575 2.0866e+10 1927.6
## <none>                   2.1613e+10 1929.1
## + stories   1  388293618 2.1224e+10 1929.3
## + garagepl  1  332426659 2.1280e+10 1929.6
## + driveway  1  107524018 2.1505e+10 1930.6
## - bedrooms  1 1551252668 2.3164e+10 1934.1
## - gashw     1 2080496614 2.3693e+10 1936.3
## - airco     1 3003315742 2.4616e+10 1940.2
## - lotsize   1 9630759680 3.1243e+10 1964.0
##
## Step:  AIC=1924.68
## price ~ lotsize + airco + gashw + bedrooms + recroom
##
##            Df  Sum of Sq        RSS    AIC
## + fullbase  1  635677235 1.9626e+10 1923.5
## + stories   1  503993901 1.9758e+10 1924.2
## + bathrms   1  497586346 1.9765e+10 1924.2
## <none>                   2.0262e+10 1924.7
## + garagepl  1  362208726 1.9900e+10 1924.9
## + driveway  1  145898498 2.0116e+10 1926.0
## - recroom   1 1350395461 2.1613e+10 1929.1
## - bedrooms  1 1569511036 2.1832e+10 1930.2
```

```
## - gashw     1 2222407032 2.2485e+10 1933.1
## - airco     1 2845424245 2.3108e+10 1935.8
## - lotsize   1 8416553209 2.8679e+10 1957.4
##
## Step:  AIC=1923.5
## price ~ lotsize + airco + gashw + bedrooms + recroom + fullbase
##
##            Df  Sum of Sq        RSS    AIC
## + stories   1  857308148 1.8769e+10 1921.0
## + bathrms   1  454768006 1.9172e+10 1923.2
## <none>                   1.9626e+10 1923.5
## + garagepl  1  271231258 1.9355e+10 1924.1
## + driveway  1  196008938 1.9430e+10 1924.5
## - fullbase  1  635677235 2.0262e+10 1924.7
## - recroom   1  684391342 2.0311e+10 1924.9
## - bedrooms  1 1469841779 2.1096e+10 1928.7
## - gashw     1 1813476015 2.1440e+10 1930.3
## - airco     1 2608608664 2.2235e+10 1934.0
## - lotsize   1 8055846359 2.7682e+10 1955.9
##
## Step:  AIC=1921.03
## price ~ lotsize + airco + gashw + bedrooms + recroom + fullbase +
##     stories
##
##            Df  Sum of Sq        RSS    AIC
## - bedrooms  1  309567266 1.9079e+10 1920.7
## + garagepl  1  421593247 1.8348e+10 1920.8
## + bathrms   1  409766032 1.8359e+10 1920.8
## <none>                   1.8769e+10 1921.0
## + driveway  1  109256718 1.8660e+10 1922.5
## - recroom   1  667529450 1.9437e+10 1922.5
## - stories   1  857308148 1.9626e+10 1923.5
## - fullbase  1  988991483 1.9758e+10 1924.2
## - gashw     1 1576297672 2.0345e+10 1927.1
## - airco     1 2104077762 2.0873e+10 1929.7
## - lotsize   1 8154844895 2.6924e+10 1955.1
##
## Step:  AIC=1920.67
## price ~ lotsize + airco + gashw + recroom + fullbase + stories
##
##            Df  Sum of Sq        RSS    AIC
## + bathrms   1  570543776 1.8508e+10 1919.6
## + garagepl  1  468228596 1.8610e+10 1920.2
## <none>                   1.9079e+10 1920.7
## + bedrooms  1  309567266 1.8769e+10 1921.0
## - recroom   1  645571393 1.9724e+10 1922.0
```

```
## + driveway  1   56505730 1.9022e+10 1922.4
## - fullbase  1 1211054944 2.0290e+10 1924.8
## - gashw     1 1552818736 2.0632e+10 1926.5
## - stories   1 2017582661 2.1096e+10 1928.7
## - airco     1 2036365524 2.1115e+10 1928.8
## - lotsize   1 9088158821 2.8167e+10 1957.6
##
## Step:  AIC=1919.63
## price ~ lotsize + airco + gashw + recroom + fullbase + stories +
##     bathrms
##
##            Df  Sum of Sq        RSS    AIC
## <none>                   1.8508e+10 1919.6
## - recroom   1  509346546 1.9018e+10 1920.3
## + garagepl  1  224584598 1.8284e+10 1920.4
## - bathrms   1  570543776 1.9079e+10 1920.7
## + bedrooms  1  148789522 1.8359e+10 1920.8
## + driveway  1  132430836 1.8376e+10 1920.9
## - fullbase  1 1063279568 1.9571e+10 1923.2
## - stories   1 1576783001 2.0085e+10 1925.8
## - gashw     1 1663725738 2.0172e+10 1926.2
## - airco     1 2246772052 2.0755e+10 1929.1
## - lotsize   1 7819998209 2.6328e+10 1952.9
```

```
##
## Call:
## lm(formula = price ~ lotsize + airco + gashw + recroom + fullbase +
##     stories + bathrms, data = House)
##
## Coefficients:
## (Intercept)      lotsize        airco        gashw      recroom
##    3898.408        5.874    11121.234    41928.529     6239.980
##    fullbase      stories      bathrms
##    7422.860     6033.802     5953.176
```