Multivariate Regression

 # We'll be predicting the home sales price based on various characteristics of the home.
 # For most of our analysis, we can use the same commands as in the Univariate Regression, but notice
 # that the interpretation may be different.

```
> A = read.csv("HOME_SALES.csv")
> names(A)
 [1] "ID"              "SALES_PRICE"     "FINISHED_AREA"   "BEDROOMS"
 [5] "BATHROOMS"       "GARAGE_SIZE"     "YEAR_BUILT"      "STYLE"
 [9] "LOT_SIZE"        "AIR_CONDITIONER" "POOL"            "QUALITY"
[13] "HIGHWAY"
> attach(A)
> reg = lm(SALES_PRICE ~ FINISHED_AREA + BEDROOMS + BATHROOMS +
GARAGE_SIZE + YEAR_BUILT )
> summary(reg)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.962e+03  4.171e+02  -7.101 4.13e-12 ***
FINISHED_AREA  1.276e-01  7.166e-03  17.806  < 2e-16 ***
BEDROOMS      -1.255e+01  3.894e+00  -3.223  0.00135 **
BATHROOMS      1.042e+01  4.945e+00   2.107  0.03561 *
GARAGE_SIZE    2.724e+01  5.930e+00   4.593 5.49e-06 ***
YEAR_BUILT     1.480e+00  2.153e-01   6.872 1.83e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.26 on 516 degrees of freedom
Multiple R-squared:  0.7356,    Adjusted R-squared:  0.7331
F-statistic: 287.1 on 5 and 516 DF,  p-value: < 2.2e-16
```

 # What? A negative coefficient for the Bathrooms? A house with more bathrooms is cheaper?
 # Answer: yes, as long as the area of the house remains constant.

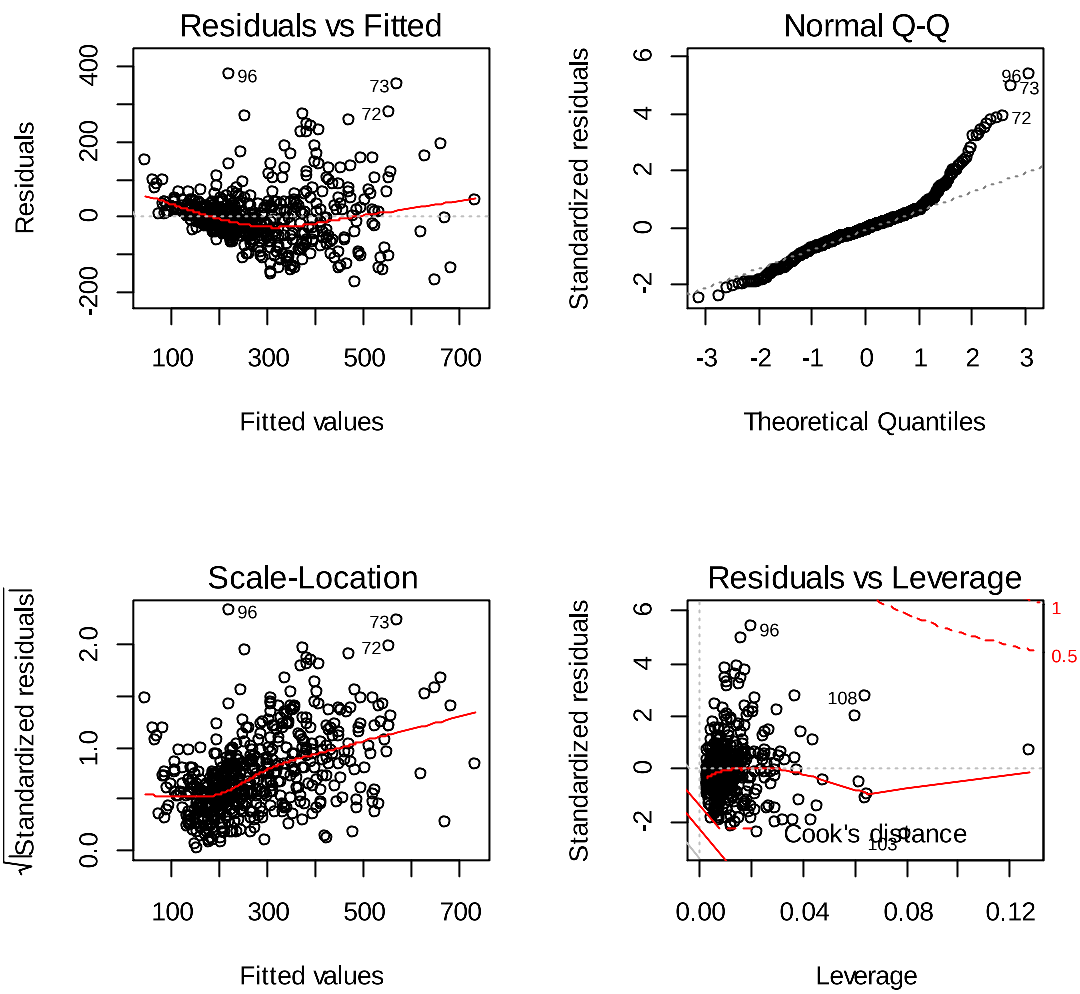
```
> anova(reg)
Analysis of Variance Table

Response: SALES_PRICE
               Df  Sum Sq Mean Sq   F value    Pr(>F)
FINISHED_AREA   1 6655486 6655486 1310.6215 < 2.2e-16 ***
BEDROOMS        1   27613   27613    5.4376   0.02009 *
BATHROOMS       1  142710  142710   28.1030 1.708e-07 ***
GARAGE_SIZE     1  224987  224987   44.3053 7.197e-11 ***
YEAR_BUILT      1  239808  239808   47.2239 1.832e-11 ***
Residuals     516 2620307    5078
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

 # FINISHED_AREA <u>alone</u> explains `6655486`. BEDROOMS explains an <u>additional</u> amount of 27613. Etc.

Residual plots

```
> par(mfrow=c(2,2))
> plot(reg)
```



Confidence intervals for the slopes.

```
> confint(reg, level=0.90)
                        5 %          95 %
(Intercept)    -3649.4960341 -2274.7356492
FINISHED_AREA      0.1157872     0.1394038
BEDROOMS         -18.9655253    -6.1333354
BATHROOMS          2.2702046    18.5680623
GARAGE_SIZE       17.4654181    37.0078645
YEAR_BUILT         1.1248812     1.8345057
```

 # Confidence intervals for the slopes with Bonferroni adjustment (just 5 slopes; suppose we are not interested in the interval for the intercept).

```
> confint(reg, level = 1 - 0.10/5)
                        1 %          99 %
(Intercept)   -3935.5690391 -1988.6626442
FINISHED_AREA     0.1108729     0.1443181
BEDROOMS        -21.6357675    -3.4630932
BATHROOMS        -1.1212061    21.9594731
GARAGE_SIZE      13.3988430    41.0744396
YEAR_BUILT        0.9772158     1.9821710
```

 # Testing several slopes in one hypothesis.
 # H_0: $\beta_4 = 0$ and $\beta_5 = 0$ vs H_1: either $\beta_4 \neq 0$ or $\beta_5 \neq 0$
 # Consider a reduced model without these variables. Compare two models
 # via a partial F-test.

```
> reg.reduced = lm(SALES_PRICE ~ FINISHED_AREA + BEDROOMS + BATHROOMS )
> anova(reg.reduced, reg)
Analysis of Variance Table

Model 1: SALES_PRICE ~ FINISHED_AREA + BEDROOMS + BATHROOMS
Model 2: SALES_PRICE ~ FINISHED_AREA + BEDROOMS + BATHROOMS + GARAGE_SIZE
+
    YEAR_BUILT
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    518 3085103
2    516 2620307  2    464796 45.765 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```