> bonus <- read.csv(file.choose(), header=TRUE)

> attach(bonus)

> model1 = lm(Value ~ Size + Age + Garage + Rooms + Baths)

> summary(model1)

Call:

lm(formula = Value ~ Size + Age + Garage + Rooms + Baths)

Residuals:

      1       2       3       4       5       6       7       8       9

  5.257 -13.719  11.236 -66.261  -3.276   7.706   2.882  50.270   5.904

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 110.1807   154.2452   0.714   0.5266

Size          0.2162     0.0800   2.703   0.0736 .

Age          -2.1893     1.0160  -2.155   0.1201

Garage      -38.4491    57.7353  -0.666   0.5531

Rooms        -0.7231    33.4941  -0.022   0.9841

Baths         1.1132    45.3529   0.025   0.9820

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.57 on 3 degrees of freedom

Multiple R-squared:  0.9599,           Adjusted R-squared:  0.8931

F-statistic: 14.36 on 5 and 3 DF,  p-value: 0.02628

> model2 = lm(Value ~ Garage)

> summary(model2)

Call:

lm(formula = Value ~ Garage)

Residuals:

    Min      1Q  Median      3Q     Max

-136.25  -75.48  -38.95   20.16  310.22

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)   353.95      75.96   4.659  0.00232 \*\*

Garage         85.54      65.79   1.300  0.23470

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 145.5 on 7 degrees of freedom

Multiple R-squared:  0.1945,           Adjusted R-squared:  0.07946

F-statistic: 1.691 on 1 and 7 DF,  p-value: 0.2347

> model3 = lm(Value ~ Rooms)

> summary(model3)

Call:

lm(formula = Value ~ Rooms)

Residuals:

    Min      1Q  Median      3Q     Max

-125.19  -66.09  -34.29   25.81  172.59

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  -203.05     201.66  -1.007   0.3475

Rooms          91.89      28.84   3.186   0.0154 \*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 103.5 on 7 degrees of freedom

Multiple R-squared:  0.5919,           Adjusted R-squared:  0.5336

F-statistic: 10.15 on 1 and 7 DF,  p-value: 0.01536

> # We see that the p-values of the variables individually are much better than before since in the original model, each variable is only the marginal effect that variable provides over the other variables and many are measuring the same type of thing in a 'big' house.

Vars = data.frame(Size, Age, Garage, Rooms, Baths, Value)

pairs(Vars, upper.panel=NULL)

> # We see somewhat of a curved pattern in Rooms and Size versus Value, let's try some squared or quadriatic terms and see if they are significant

>

> Rooms2 = Rooms\*Rooms

> Size2 = Size\*Size

> model4 = lm(Value ~ Size + Size2)

> summary(model4)

Call:

lm(formula = Value ~ Size + Size2)

Residuals:

    Min      1Q  Median      3Q     Max

-44.174 -25.528  -0.731  33.452  43.128

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  3.054e+02  1.354e+02   2.256   0.0649 .

Size        -1.612e-01  1.392e-01  -1.158   0.2908

Size2        9.873e-05  3.427e-05   2.881   0.0280 \*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.4 on 6 degrees of freedom

Multiple R-squared:  0.9493,           Adjusted R-squared:  0.9324

F-statistic: 56.21 on 2 and 6 DF,  p-value: 0.00013

> model5 = lm(Value ~ Size2)

> summary(model5)

Call:

lm(formula = Value ~ Size2)

Residuals:

    Min      1Q  Median      3Q     Max

-43.393 -38.246   4.828  20.622  64.337

Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.523e+02  3.015e+01   5.053  0.00147 \*\*

Size2       5.958e-05  5.789e-06  10.292 1.77e-05 \*\*\*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.35 on 7 degrees of freedom

Multiple R-squared:  0.938, Adjusted R-squared:  0.9292

F-statistic: 105.9 on 1 and 7 DF,  p-value: 1.769e-05

> model6 = lm(Value ~ Rooms + Rooms2)

> summary(model6)

Call:

lm(formula = Value ~ Rooms + Rooms2)

Residuals:

    Min      1Q  Median      3Q     Max

-104.71  -23.25   -5.63   18.65  106.59

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  1486.98     663.07   2.243   0.0661 .

Rooms        -422.44     197.79  -2.136   0.0766 .

Rooms2         37.90      14.49   2.616   0.0398 \*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.45 on 6 degrees of freedom

Multiple R-squared:  0.8093,           Adjusted R-squared:  0.7457

F-statistic: 12.73 on 2 and 6 DF,  p-value: 0.006935

> # now decision depends on the alpha we pick. if we pick alpha is .1 we get both variables are significant. R2 and R2adjusted are down quite a bit, that makes us unhappy.

> model7 = lm(Value ~ Size + Size2 + Rooms + Rooms2  + Garage + Baths + Age)

> summary(model7)

Call:

lm(formula = Value ~ Size + Size2 + Rooms + Rooms2 + Garage +

    Baths + Age)

Residuals:

      1       2       3       4       5       6       7       8       9

  7.213 -25.932   3.933  -7.797   8.514   2.239  13.814   2.924  -4.908

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.240e+03  4.967e+02   2.496    0.243

Size         2.641e-01  2.320e-01   1.139    0.459

Size2       -4.240e-05  6.264e-05  -0.677    0.621

Rooms       -4.267e+02  2.013e+02  -2.120    0.281

Rooms2       3.586e+01  1.663e+01   2.156    0.276

Garage       2.617e+01  4.858e+01   0.539    0.685

Baths        2.645e+00  3.040e+01   0.087    0.945

Age         -1.158e-01  1.170e+00  -0.099    0.937

Residual standard error: 33.19 on 1 degrees of freedom

Multiple R-squared:  0.994, Adjusted R-squared:  0.9521

F-statistic:  23.7 on 7 and 1 DF,  p-value: 0.1569

> # interaction terms are when you think there is a synergy that exists between 2 variables.

AB = Age\*Baths

model8 = lm(Value ~ Size + Rooms + Garage + Baths + Age + AB)

summary(model8)

Call:

lm(formula = Value ~ Size + Rooms + Garage + Baths + Age + AB)

Residuals:

      1       2       3       4       5       6       7       8       9

 8.5387 -2.5531 -6.1391 -0.7524 -4.3931 -0.9758 -5.2871  2.8995  8.6624

 #This whole part is not necessary

#in this case if the sample size is really small,then it will appear

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.089e+03  4.246e+02  -7.274   0.0184 \*

Size        -2.791e-01  6.796e-02  -4.107   0.0545 .

Rooms        2.173e+02  2.981e+01   7.289   0.0183 \*

Garage       2.496e+02  4.027e+01   6.199   0.0250 \*

Baths        1.001e+03  1.327e+02   7.545   0.0171 \*

Age          5.141e+01  7.095e+00   7.247   0.0185 \*

AB          -2.273e+01  3.007e+00  -7.559   0.0171 \*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.17 on 2 degrees of freedom

Multiple R-squared:  0.9986,           Adjusted R-squared:  0.9946

F-statistic: 245.5 on 6 and 2 DF,  p-value: 0.004062

> # it got WAY better. Not only is the interaction term (AB) significant, all the other variables are waaaay better. Also r2 is really high and close to r2 adjusted. This is great

> # This interaction term is significant at alpha=0.05, as are almost all the other variables now. Both R-squareds are also extremely high and close which is great. THe only downside about interaction terms is that there isn't a graph to help you know one should exist. you just need to know something about the data!

> # You can include a squared term without the original X if the squared term is significant and the original X is not. However, if the interaction term is significant, you also must keep the original 2 terms that make it in your model, even if their p-values are high and not significant.

>

> #Model: Value=-3089-0.2791(Size) + 217.3(Rooms) + 249.6(GaErage) + 1001(baths) + 51.41(Age) -22.73(Age\*Baths)> bonus <- read.csv(file.choose(), header=TRUE)