Crush the Real Estate Market With Multiple Linear Regression

Scott C. Sterbenz, P.E., Ford Motor Company

ASQ World Conference, Session M19

ASQ
The Global Voice of Quality

Session Objectives

At the conclusion of this session, you will:

- Know when to use multiple linear regression
- Understand how to collect and analyze data using multiple linear regression
- Realize the dangers of multicollinearity and troublesome residuals—and how to handle them
- Recognize the power of multiple linear regression for problem solving through analysis of real estate data

Presentation Outline

- I. The Organization I Serve
  - Ford Motor Company
- II. Multiple Linear Regression
  - Background
  - Data Collection
  - Data Analysis & Interpretation
- III. Application
- IV. Questions

Ford Motor Company

Ford Motor Company:

- Manufactures cars and trucks globally under the Ford and Lincoln brands
- Established in 1903
- Visit us at http://www.ford.com

Go Further

Multiple Linear Regression Background

What is the goal?

- Explain the behavior of a dependent variable (Y) based on the behavior of multiple independent variables (Xs)
  - ✓ Cumulative effects
  - ✓ Interaction effects
  - ✓ Limited curvature
  - ✓ Attribute or variable predictors
  - ✓ Mathematical equation
  - ✓ Prediction and optimization abilities

Multiple Linear Regression Background

When do you use it?

- Empirical data
  - ✓ "Here's the data—tell me what it says"
- Inability to set up a designed experiment
  - ✓ You can't dictate specific combinations of the Xs to be run simultaneously

Multiple Linear Regression Background

Why not use Multiple Nonlinear Regression?

- Multiple linear regression can model curvature, but multiple nonlinear regression can do it better
  - ✓ Finding the “best” curvature model can be difficult
- Multiple linear regression is easier to setup and interpret
  - ✓ P-values for the coefficients
  - ✓ Confidence intervals around predictions

Multiple Linear Regression Background

Why not use Multiple Nonlinear Regression?

- Multiple linear regression equation examples

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + e$$

$$Y = b_0 + b_1 \ln(x_1) + b_2 x_2 + b_3 x_3 + b_4 (x_3)^2 + e$$

- Multiple nonlinear regression equation examples

$$Y = b_0 + \frac{x_1}{x_2} + e$$

$$Y = x_1 \cos(x_1) + x_2 \sin(x_3) + e$$

A regression equation is linear when it is linear in the parameters. You can transform predictors in ways that create curvature, but the equation must be linear in the parameters.

Multiple Linear Regression Data Collection

Guidelines for Data Collection:

- At least two more rows of data than the number of terms to be analyzed
  - ✓ Absolute minimum
  - ✓ Don’t forget to count curvature and interactions
- Ranges of predictor variables
  - ✓ Reasonably balanced (attribute)
  - ✓ Reasonably spread across range of interest (variable)

Multiple Linear Regression Data Collection

Guidelines for Data Collection:

- Additional considerations for attribute predictors:
    - ✓ Assign numbers instead of category names
- 2-4 categories:
    - ✓ If it makes sense, order the categories logically
- Large number of ordinal categories (Likert scale):
    - ✓ Treat the attribute predictor as variable

Attribute Predictor Raw Values	Attribute Predictors Numerical Values
Yes / No	1 / 0
Small / Medium / Large	1 / 2 / 3

Multiple Linear Regression Data Analysis

Procedure:

1. Check for potential curvature
2. Assess the magnitude of multicollinearity
3. Reduce the model
    a. Manually using ANOVA table
    b. Best Subsets
    c. Stepwise functions
4. Check the residuals
5. Validate the model
6. Predict and/or optimize

Multiple Linear Regression Data Analysis

To illustrate Multiple Linear Regression, consider the real estate market in Flat Rock, Michigan:

I am planning to flip a house and want to know:

- What buying price indicates a bargain?
- What features of the house should I spend money on?
- What price should I list at for a quick and profitable sale?

Data is from MLS real estate listings in Flat Rock, MI (ZIP Code 48134) —single, detached, non-foreclosed homes on one acre or less, as of September 28, 2017

Multiple Linear Regression Data Analysis

Response: List Price

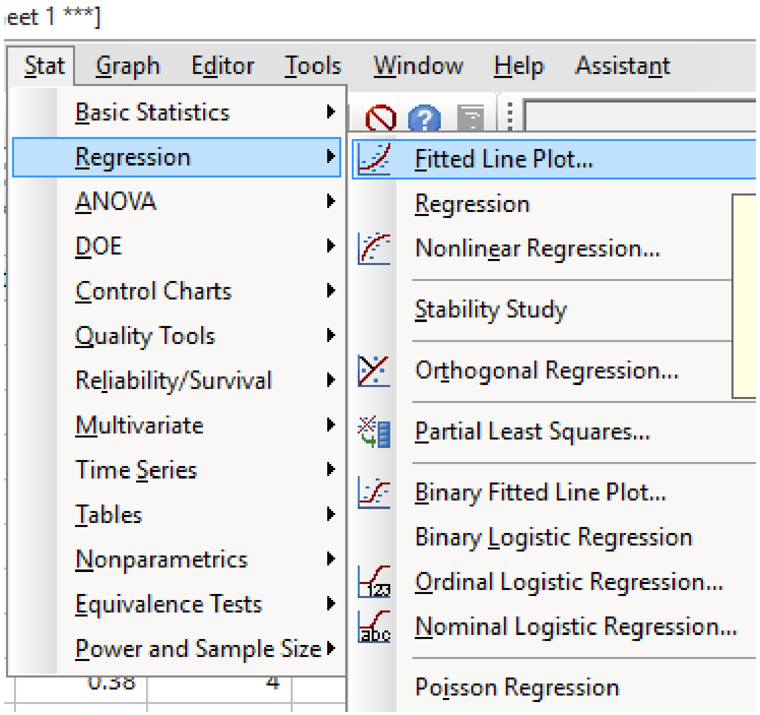
Variable Predictors	Attribute Predictors
Year Built	Number of Bedrooms
Lot Size (acres)	Number of Full Bathrooms
Square Footage	Number of Half Bathrooms
Quality of Landscaping	Garage Size
	Pool
	Finished Basement
	Number of Stories
	Construction Material

Data for this illustration is from MLS real estate listings in Flat Rock, MI—single, detached, non-foreclosed homes on one acre or less, as of September 28, 2017

Multiple Linear Regression Data Analysis

Step 1: Check for potential curvature

- Generate a fitted line plot for all Xs versus the Y
- No need to check attribute predictors with two categories



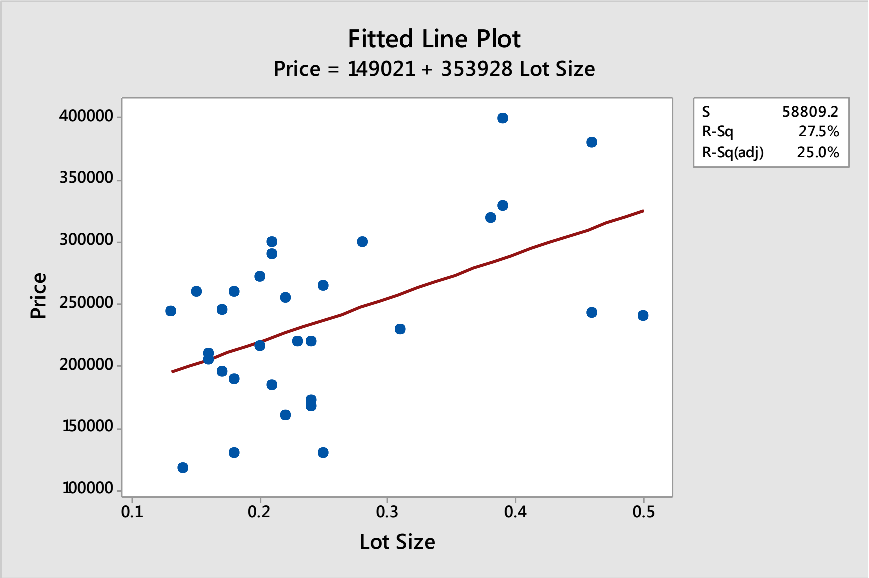
Multiple Linear Regression Data Analysis

Step 1: Check for potential curvature

- If there appears to be curvature:
  - ✓ Run quadratic or cubic models
  - ✓ Run simple transformations, like log or natural log
  - ✓ Compare the r-sq (adj.) values
    - ➢ If the r-sq (adj.) value increases by more than 10%, then curvature should be considered

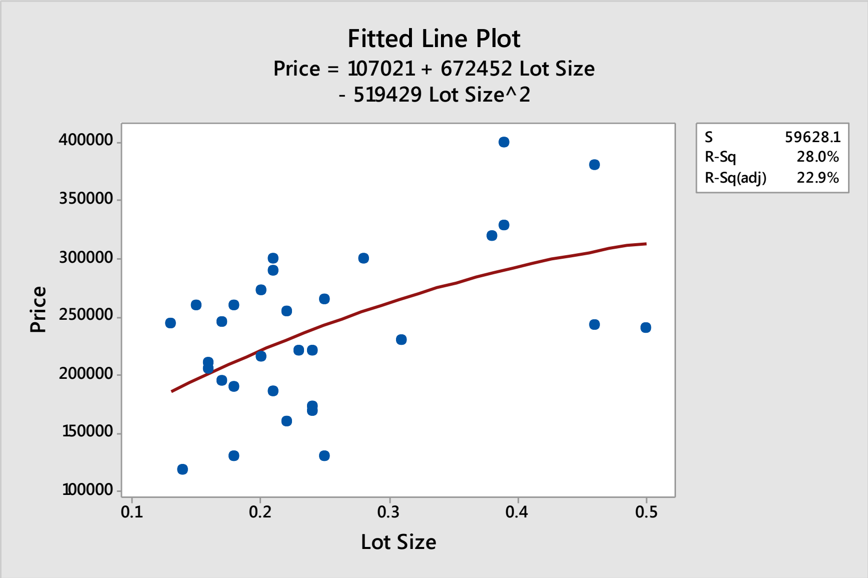
Multiple Linear Regression Data Analysis

Step 1: Check for potential curvature



Fitted Line Plot

Price = 149021 + 353928 Lot Size



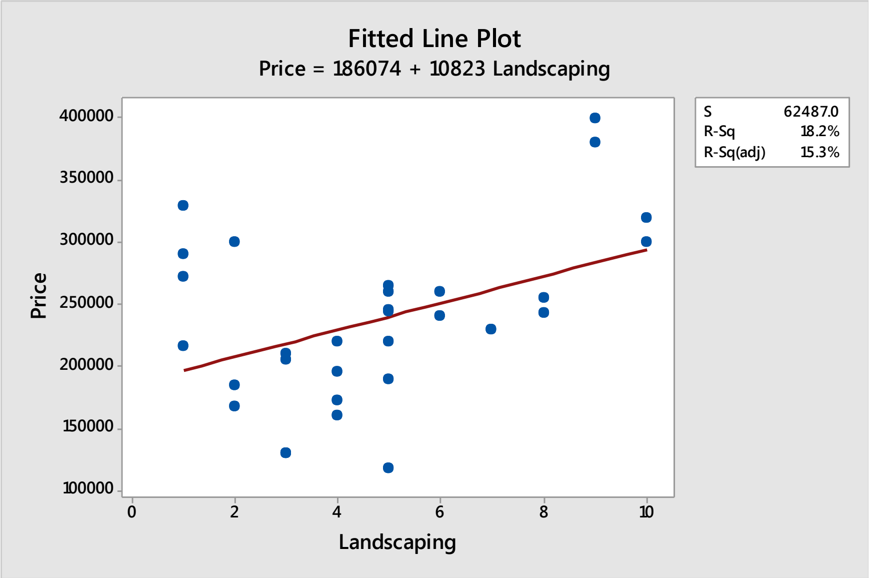
Fitted Line Plot

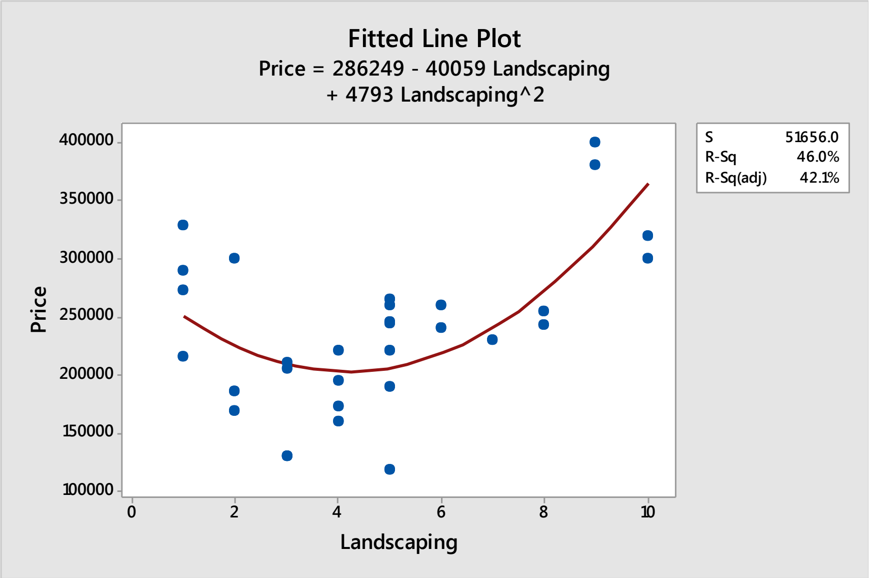
Price = 107021 + 672452 Lot Size
- 519429 Lot Size^2

The r-sq (adj.) value for the quadratic did not increase by more than 10%--curvature not considered for this X.

Multiple Linear Regression Data Analysis

Step 1: Check for potential curvature

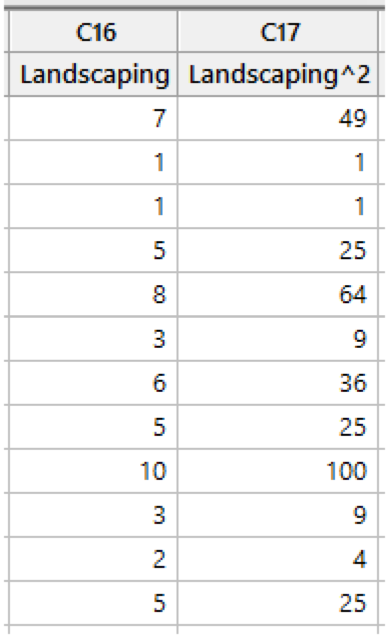




The r-sq (adj.) value for the quadratic increased by more than 10%--curvature considered for this X.

Multiple Linear Regression Data Analysis

Step 1: Check for potential curvature



C16	C17
Landscaping	Landscaping^2
7	49
1	1
1	1
5	25
8	64
3	9
6	36
5	25
10	100
3	9
2	4
5	25

Curvature	No Curvature
Quality of Landscaping	Year Built
	Square Footage
	Number of Bedrooms
	Number of Full Bathrooms
	Number of Half Bathrooms
	Garage Size
	Pool
	Finished Basement
	Number of Stories
	Construction Material

Add the curvature columns

Multiple Linear Regression Data Analysis

Step 2: Assess the magnitude of multicollinearity

- Multicollinearity is a measure of the correlation of the predictors
- Multicollinearity is measured by the Variance Inflation Factor
- High multicollinearity can lead to double-counting, inadvertent cancellation, and poor predictive models

$$VIF = \frac{1}{1 - R_i^2}$$

Guideline:

VIF < 5

Multiple Linear Regression Data Analysis

Step 2: Assess the magnitude of multicollinearity

- When there is high multicollinearity, you must remove redundant terms from the analysis:
  - ✓ The predictor that makes less sense
  - ✓ The predictor that is harder to measure
  - ✓ The predictor that is further away from the response

Multiple Linear Regression Data Analysis

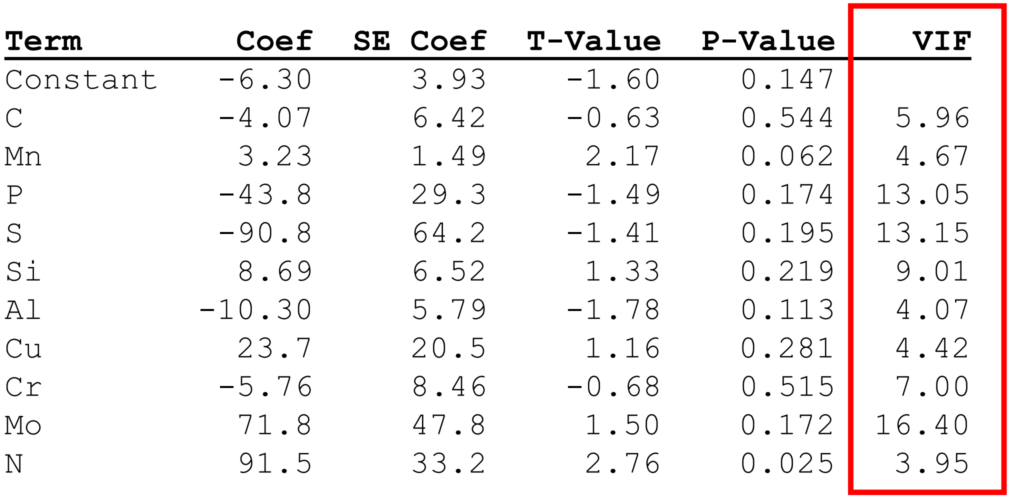
Step 2: Assess the magnitude of multicollinearity

- Examples of multicollinearity:
  - ✓ Predictors related by a mathematical equation
    - ➢ Drop the calculated value
    - ➢ Keep the predictor that is the cause, not the effect
    - ➢ Use trial and error to get the best r-sq values
  - ✓ Linear and non-linear terms of the same predictor
    - ➢ Only assess multicollinearity with one of the two

Multiple Linear Regression Data Analysis

Step 2: Assess the magnitude of multicollinearity

- Example of multicollinearity from a metal stamping study:
  - ➢ Molybdenum, Phosphorous, and Sulfur are the most multicollinear



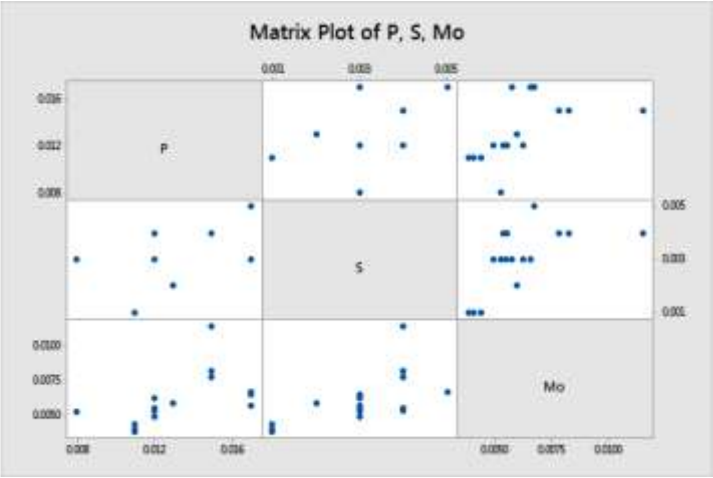
```
Term         Coef  SE Coef  T-Value  P-Value     VIF
Constant    -6.30     3.93    -1.60    0.147
C           -4.07     6.42    -0.63    0.544    5.96
Mn           3.23     1.49     2.17    0.062    4.67
P           -43.8     29.3    -1.49    0.174   13.05
S           -90.8     64.2    -1.41    0.195   13.15
Si           8.69     6.52     1.33    0.219    9.01
Al         -10.30     5.79    -1.78    0.113    4.07
Cu           23.7     20.5     1.16    0.281    4.42
Cr          -5.76     8.46    -0.68    0.515    7.00
Mo           71.8     47.8     1.50    0.172   16.40
N            91.5     33.2     2.76    0.025    3.95
```

Multiple Linear Regression Data Analysis

Step 2: Assess the magnitude of multicollinearity

- Remove Molybdenum
  - ➢ Mo is multicollinear to both S and P
  - ➢ S and P do not appear to be multicollinear

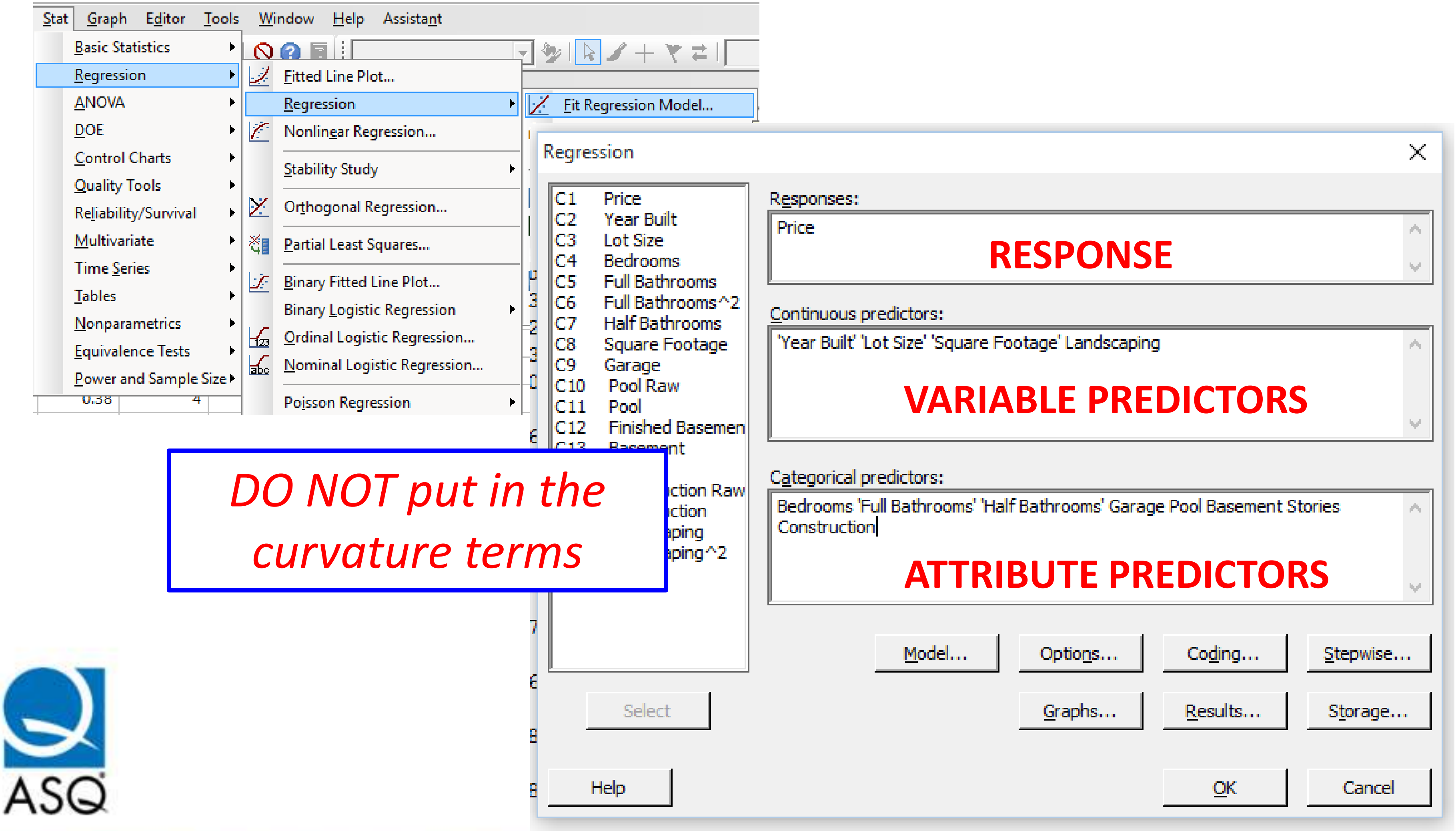
```
Term         Coef   SE Coef   T-Value   P-Value    VIF
Constant    -1.22      2.12     -0.57     0.580
C           -8.19      6.20     -1.32     0.219   4.87
Mn           1.62      1.10      1.48     0.174   2.23
P           -10.2      20.2     -0.50     0.627   3.41
S           -13.5      40.9     -0.33     0.750   4.68
Si           1.28      4.55      0.28     0.784   3.85
Al          -9.01      6.11     -1.47     0.174   3.98
Cu            1.2      14.9      0.08     0.939   2.05
Cr           4.17      5.62      0.74     0.477   2.72
N            69.3      31.7      2.18     0.057   3.17
```



Use the matrix plot to help isolate the multicollinearity

Multiple Linear Regression Data Analysis

Step 2: Assess the magnitude of multicollinearity



Responses: Price

RESPONSE

Continuous predictors: 'Year Built' 'Lot Size' 'Square Footage' Landscaping

VARIABLE PREDICTORS

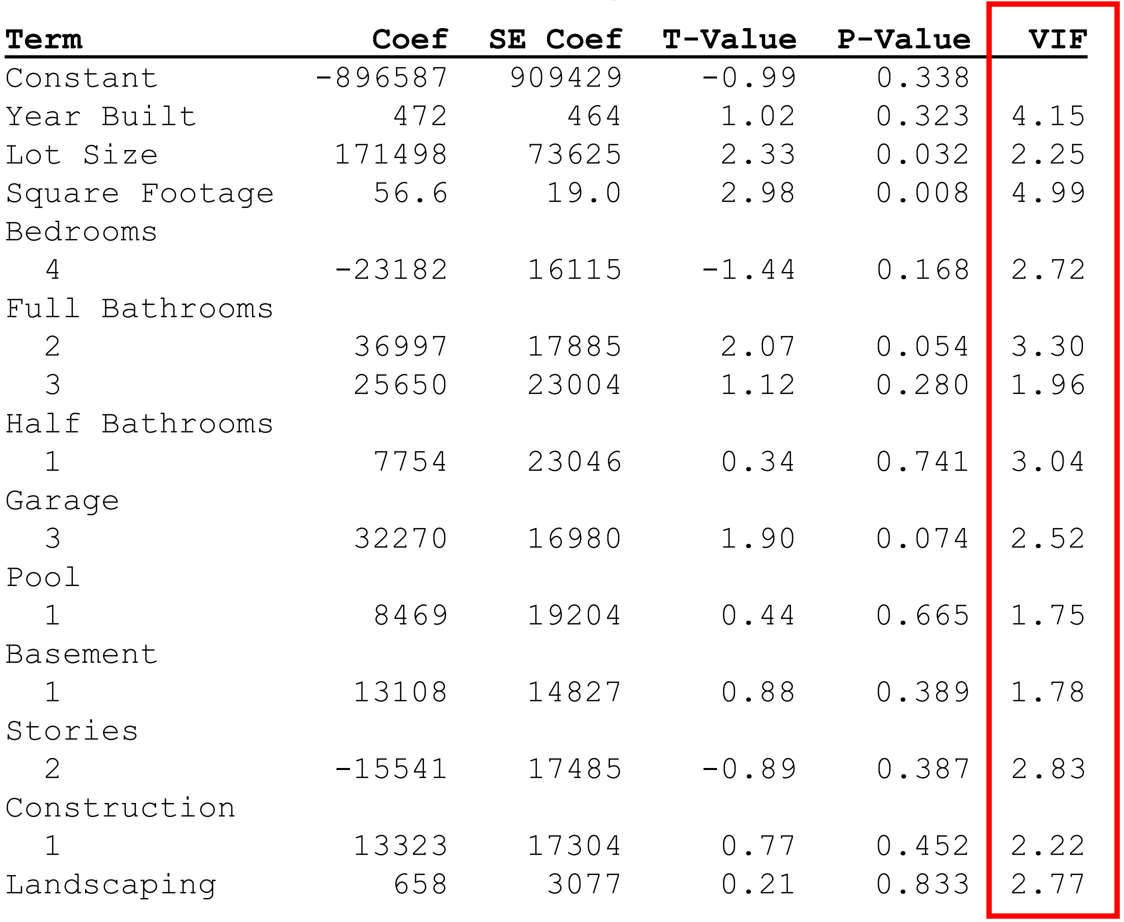
Categorical predictors: Bedrooms 'Full Bathrooms' 'Half Bathrooms' Garage Pool Basement Stories Construction

ATTRIBUTE PREDICTORS

DO NOT put in the curvature terms

Multiple Linear Regression Data Analysis

Step 2: Assess the magnitude of multicollinearity



```
Term                  Coef   SE Coef  T-Value  P-Value   VIF
Constant           -896587    909429    -0.99    0.338
Year Built             472       464     1.02    0.323  4.15
Lot Size            171498     73625     2.33    0.032  2.25
Square Footage        56.6      19.0     2.98    0.008  4.99
Bedrooms
  4                 -23182     16115    -1.44    0.168  2.72
Full Bathrooms
  2                  36997     17885     2.07    0.054  3.30
  3                  25650     23004     1.12    0.280  1.96
Half Bathrooms
  1                   7754     23046     0.34    0.741  3.04
Garage
  3                  32270     16980     1.90    0.074  2.52
Pool
  1                   8469     19204     0.44    0.665  1.75
Basement
  1                  13108     14827     0.88    0.389  1.78
Stories
  2                 -15541     17485    -0.89    0.387  2.83
Construction
  1                  13323     17304     0.77    0.452  2.22
Landscaping            658      3077     0.21    0.833  2.77
```

Guideline:
VIF < 5

There is no significant multicollinearity

Methodology Summary

Step 3: Reduce the model

Method	Advantages	Disadvantages
ANOVA Table	1. Total Control	1. Tedious 2. Can be tough to handle multicollinearity
Best Subsets	1. Handles large number of predictors 2. Handles mild multicollinearity	1. Does not always eliminate multicollinearity 2. Requires trial and error to pick the best model
Stepwise	1. Handles large number of predictors 2. Easy to analyze curvature and large numbers of interactions	1. Sometimes too aggressive with model reduction 2. Does not always eliminate multicollinearity

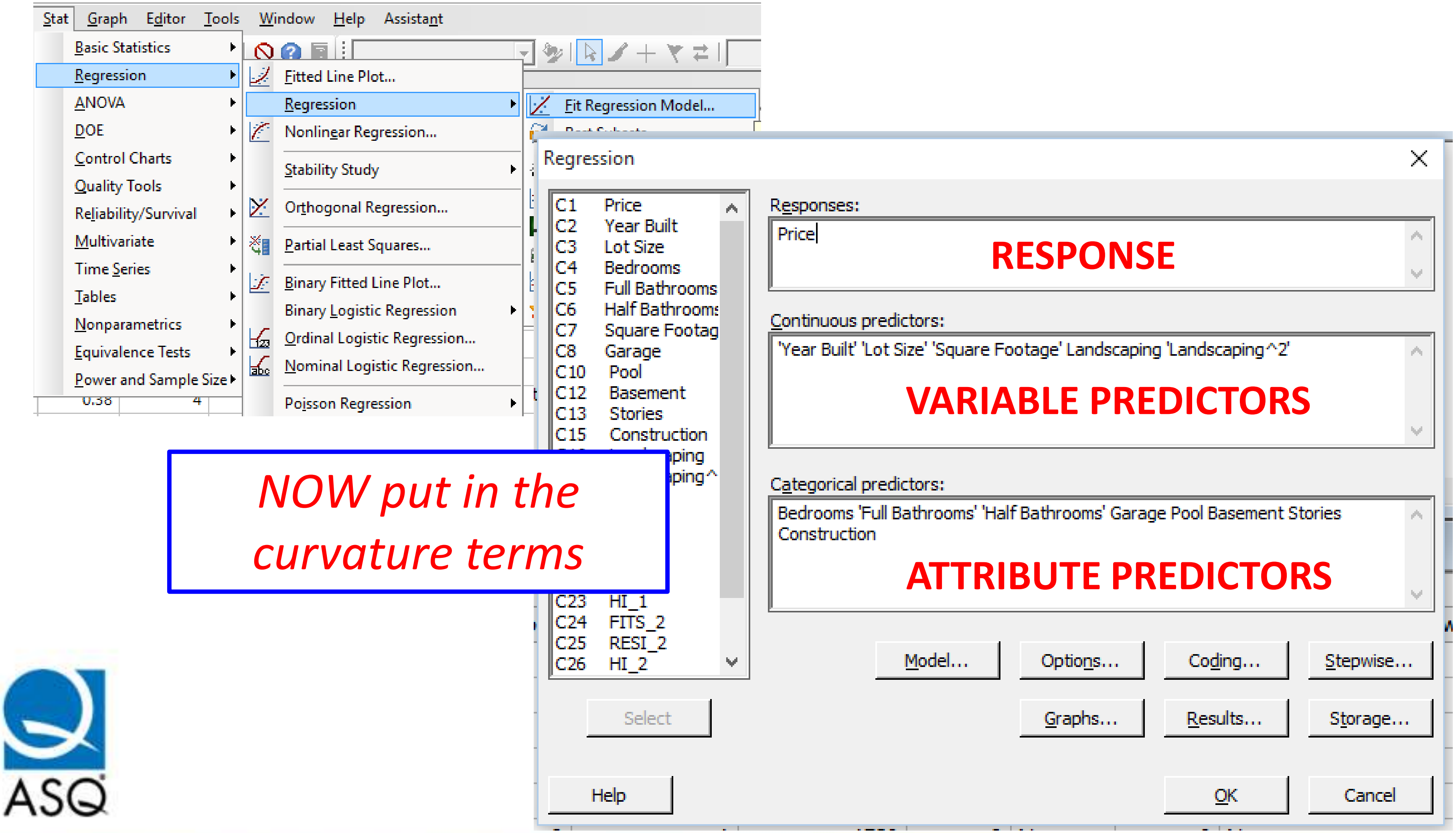
Multiple Linear Regression Data Analysis

Step 3: Reduce the model (ANOVA Table Method)

- Record the r-sq (adj.) and r-sq (pred.) for each iteration
- Remove one term at a time
  - ✓ Highest p-values first
  - ✓ Follow guidelines for hierarchy
- Continue as long as r-sq (adj.) and/or r-sq (pred.) increase
- Stop when:
  - ✓ No more terms have p-values greater than 0.05
  - ✓ The r-sq (adj.) and r-sq (pred.) start dropping

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (ANOVA Table Method)



Responses: Price

RESPONSE

Continuous predictors: 'Year Built' 'Lot Size' 'Square Footage' Landscaping 'Landscaping^2'

VARIABLE PREDICTORS

Categorical predictors: Bedrooms 'Full Bathrooms' 'Half Bathrooms' Garage Pool Basement Stories Construction

ATTRIBUTE PREDICTORS

NOW put in the curvature terms

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (ANOVA Table Method)

```
Source             DF      Adj SS       Adj MS  F-Value  P-Value
Regression         14  1.26172E+11  9012298901    11.84    0.000
  Year Built        1    638002524   638002524     0.84    0.373
  Lot Size          1   3018800891  3018800891     3.97    0.064
  Square Footage    1   4045862273  4045862273     5.32    0.035
  Bedrooms          1   1112111781  1112111781     1.46    0.244
  Garage            1   2696517002  2696517002     3.54    0.078
  Full Bathrooms    2   3176697774  1588348887     2.09    0.156
  Landscaping       1    192561581   192561581     0.25    0.622
  Landscaping^2     1    272320710   272320710     0.36    0.558
  Half Bathrooms    1    219085571   219085571     0.29    0.599
  Pool              1    114669784   114669784     0.15    0.703
  Basement          1    628990496   628990496     0.83    0.377
  Stories           1    646848565   646848565     0.85    0.370
  Construction      1    457387476   457387476     0.60    0.449
Error              16  12174017488   760876093
Total              30  1.38346E+11

Model Summary

      S     R-sq  R-sq(adj)  R-sq(pred)
27584.0   91.20%     83.50%      60.71%
```

POOL is the first term removed from the model

Remember to record the R-sq (adj.) and R-sq (pred.) values

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (ANOVA Table Method)

```
Source               DF       Adj SS       Adj MS  F-Value  P-Value
Regression           13  1.26058E+11   9696731910    13.41    0.000
  Year Built          1    728192520    728192520     1.01    0.330
  Lot Size            1   2946771677   2946771677     4.08    0.060
  Square Footage      1   4179568329   4179568329     5.78    0.028
  Bedrooms            1   1009333288   1009333288     1.40    0.254
  Garage              1   2594213198   2594213198     3.59    0.075
  Full Bathrooms      2   3068342357   1534171178     2.12    0.150
  Landscaping         1    172166503    172166503     0.24    0.632
  Landscaping^2       1    300047549    300047549     0.42    0.528
  Half Bathrooms      1    199972802    199972802     0.28    0.606
  Basement            1    657100099    657100099     0.91    0.354
  Stories             1    590027691    590027691     0.82    0.379
  Construction        1    428753995    428753995     0.59    0.452
Error                17  12288687272    722863957
Total                30  1.38346E+11

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
26886.1  91.12%     84.32%      68.08%
```

HALF BATHROOMS is the next term removed from the model

Notice the R-sq (adj.) and R-sq (pred.) values both increased

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (ANOVA Table Method)

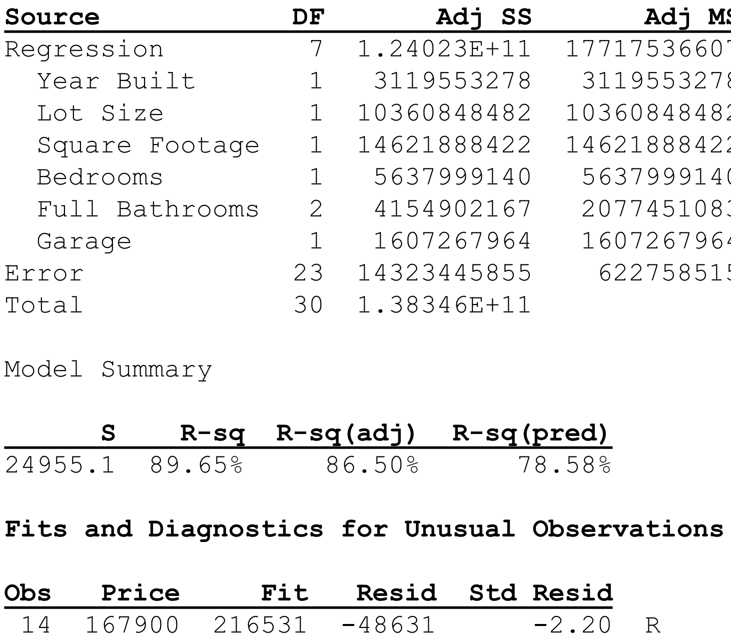
Term Removed	P-Value	R-sq (adj.)	R-sq (pred.)
		83.50	60.71
Pool	0.703	84.32	68.08
Half Bathrooms	0.606	84.95	68.77
Landscaping^2	0.634	85.56	72.62
Landscaping	0.547	86.01	76.11
Construction Mat'l	0.414	86.21	77.54
Stories	0.377	86.33	77.37
Basement	0.405	86.50	78.58
Garage Size	0.122	85.61	76.92

Maximized Result

Even though Garage Size has a p-value greater than 0.05, it still adds value to the model

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (ANOVA Table Method)



```
Source                DF       Adj SS       Adj MS  F-Value  P-Value
Regression             7  1.24023E+11  17717536607    28.45    0.000
  Year Built           1   3119553278   3119553278     5.01    0.035
  Lot Size             1  10360848482  10360848482    16.64    0.000
  Square Footage       1  14621888422  14621888422    23.48    0.000
  Bedrooms             1   5637999140   5637999140     9.05    0.006
  Full Bathrooms       2   4154902167   2077451083     3.34    0.053
  Garage               1   1607267964   1607267964     2.58    0.122
Error                 23  14323445855    622758515
Total                 30  1.38346E+11

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
24955.1  89.65%     86.50%      78.58%

Fits and Diagnostics for Unusual Observations

Obs   Price     Fit   Resid  Std Resid
 14  167900  216531  -48631      -2.20  R
```

Summary of the Final Model

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (ANOVA Table Method)

- Things to check on the final model:
  - ✓ Do the direction of the coefficients mostly make sense?
  - ✓ Are there a lot of unusual observations?
  - ✓ Is the r-sq (adj.) really low?
  - ✓ Is there a large difference between the r-sq (adj.) and r-sq (pred.) values?

All these things are indicators of an unstable model, or a model with poor predictive ability

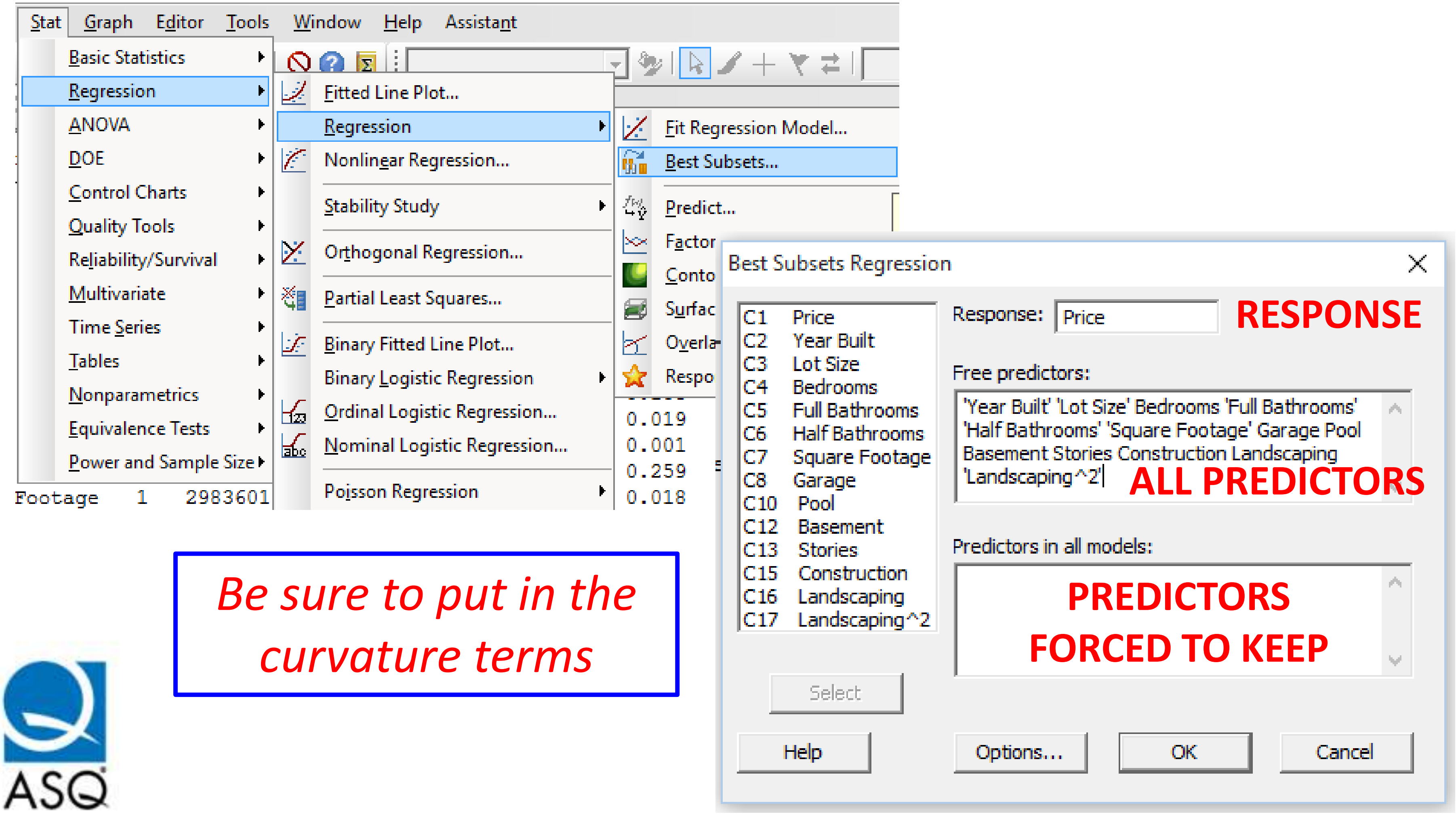
Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Best Subsets)

- Theory behind best subsets:
  - ✓ Methodically increases the number of predictors considered, starting with one, by increments of one
  - ✓ Reports the best two results for each
- Useful when there is a lot of multicollinearity or a lot of predictors

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Best Subsets)



C1 Price
C2 Year Built
C3 Lot Size
C4 Bedrooms
C5 Full Bathrooms
C6 Half Bathrooms
C7 Square Footage
C8 Garage
C10 Pool
C12 Basement
C13 Stories
C15 Construction
C16 Landscaping
C17 Landscaping^2

Response: Price

RESPONSE

Free predictors: 'Year Built' 'Lot Size' Bedrooms 'Full Bathrooms' 'Half Bathrooms' 'Square Footage' Garage Pool Basement Stories Construction Landscaping 'Landscaping^2'

ALL PREDICTORS

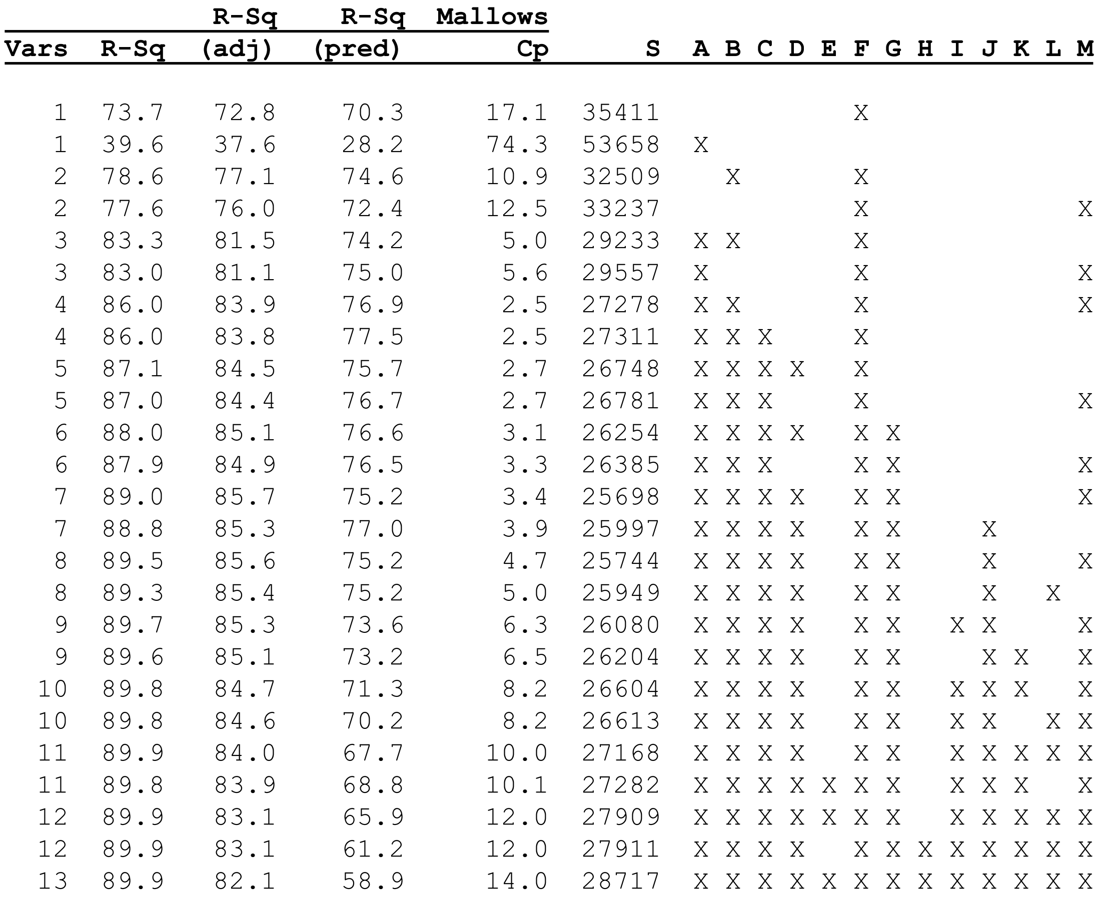
Predictors in all models:

PREDICTORS FORCED TO KEEP

Be sure to put in the curvature terms

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Best Subsets)



```
             R-Sq    R-Sq   Mallows
Vars  R-Sq  (adj)  (pred)        Cp      S  A B C D E F G H I J K L M

   1  73.7   72.8    70.3      17.1  35411            X
   1  39.6   37.6    28.2      74.3  53658  X
   2  78.6   77.1    74.6      10.9  32509    X       X
   2  77.6   76.0    72.4      12.5  33237            X             X
   3  83.3   81.5    74.2       5.0  29233  X X       X
   3  83.0   81.1    75.0       5.6  29557  X         X             X
   4  86.0   83.9    76.9       2.5  27278  X X       X             X
   4  86.0   83.8    77.5       2.5  27311  X X X     X
   5  87.1   84.5    75.7       2.7  26748  X X X X   X
   5  87.0   84.4    76.7       2.7  26781  X X X     X             X
   6  88.0   85.1    76.6       3.1  26254  X X X X   X X
   6  87.9   84.9    76.5       3.3  26385  X X X     X X           X
   7  89.0   85.7    75.2       3.4  25698  X X X X   X X           X
   7  88.8   85.3    77.0       3.9  25997  X X X X   X X     X
   8  89.5   85.6    75.2       4.7  25744  X X X X   X X     X     X
   8  89.3   85.4    75.2       5.0  25949  X X X X   X X     X   X
   9  89.7   85.3    73.6       6.3  26080  X X X X   X X   X X     X
   9  89.6   85.1    73.2       6.5  26204  X X X X   X X     X X   X
  10  89.8   84.7    71.3       8.2  26604  X X X X   X X   X X X   X
  10  89.8   84.6    70.2       8.2  26613  X X X X   X X   X X   X X
  11  89.9   84.0    67.7      10.0  27168  X X X X   X X   X X X X X
  11  89.8   83.9    68.8      10.1  27282  X X X X X X X   X X X   X
  12  89.9   83.1    65.9      12.0  27909  X X X X X X X   X X X X X
  12  89.9   83.1    61.2      12.0  27911  X X X X   X X X X X X X X
  13  89.9   82.1    58.9      14.0  28717  X X X X X X X X X X X X X
```

PREDICTORS

X INDICATES A KEEPER

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Best Subsets)

- Make the best selection:

✓ Choose 4 or 5 “finalists” with the best R-sq (adj.) and R-sq (pred.) values

✓ Mallows’ Cp should be less than or equal to the number of terms in the model

✓ S (Standard deviation of the residuals) should be as small as possible

Mallows’ Cp statistic helps guard against overfit, by estimating the mean squared prediction error (MSPE)

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Best Subsets)

```
             R-Sq    R-Sq  Mallows
Vars  R-Sq  (adj)  (pred)       Cp      S  A B C D E F G H I J K L M
   1  73.7   72.8    70.3     17.1  35411            X
   1  39.6   37.6    28.2     74.3  53658  X
   2  78.6   77.1    74.6     10.9  32509    X       X
   2  77.6   76.0    72.4     12.5  33237            X             X
   3  83.3   81.5    74.2      5.0  29233  X X       X
   3  83.0   81.1    75.0      5.6  29557  X         X             X
   4  86.0   83.9    76.9      2.5  27278  X X       X             X
   4  86.0   83.8    77.5      2.5  27311  X X X     X
   5  87.1   84.5    75.7      2.7  26748  X X X X   X
   5  87.0   84.4    76.7      2.7  26781  X X X     X             X
   6  88.0   85.1    76.6      3.1  26254  X X X X   X X
   6  87.9   84.9    76.5      3.3  26385  X X X     X X           X
   7  89.0   85.7    75.2      3.4  25698  X X X X   X X           X
   7  88.8   85.3    77.0      3.9  25997  X X X X   X X     X
   8  89.5   85.6    75.2      4.7  25744  X X X X   X X     X     X
   8  89.3   85.4    75.2      5.0  25949  X X X X   X X     X   X
   9  89.7   85.3    73.6      6.3  26080  X X X X   X X   X X     X
   9  89.6   85.1    73.2      6.5  26204  X X X X   X X     X X   X
  10  89.8   84.7    71.3      8.2  26604  X X X X   X X   X X X   X
  10  89.8   84.6    70.2      8.2  26613  X X X X   X X   X X   X X
  11  89.9   84.0    67.7     10.0  27168  X X X X   X X   X X X X X
  11  89.8   83.9    68.8     10.1  27282  X X X X X X X   X X X   X
  12  89.9   83.1    65.9     12.0  27909  X X X X X X X   X X X X X
  12  89.9   83.1    61.2     12.0  27911  X X X X   X X X X X X X X
  13  89.9   82.1    58.9     14.0  28717  X X X X X X X X X X X X X
```

Best Selection

Many of these models are OK to investigate further

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Best Subsets)

- Check your work:
  - ✓ Is the multicollinearity acceptable
  - ✓ Do the direction of the coefficients mostly make sense?
  - ✓ Are there a lot of unusual observations?
  - ✓ Is there a large difference between the r-sq (adj.) and r-sq (pred.) values?
- The Best Subsets method and the ANOVA Table method may not give the same result

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Best Subsets)

```
Source              DF       Adj SS       Adj MS  F-Value  P-Value
Regression           8  1.24523E+11  15565348315    24.77    0.000
  Year Built         1   3574815058   3574815058     5.69    0.026
  Lot Size           1   5294459291   5294459291     8.43    0.008
  Square Footage     1   9799990327   9799990327    15.60    0.001
  Landscaping^2      1    500030269    500030269     0.80    0.382
  Bedrooms           1   2994519757   2994519757     4.77    0.040
  Full Bathrooms     2   2884737024   1442368512     2.30    0.124
  Garage             1   1737703697   1737703697     2.77    0.110
Error               22  13823415586    628337072
Total               30  1.38346E+11

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
25066.7  90.01%     86.37%      75.84%

Fits and Diagnostics for Unusual Observations

Obs   Price     Fit   Resid  Std Resid
 14  167900  208912  -41012      -2.00  R
 29  299900  342674  -42774      -2.18  R
```

Summary of the Final Model

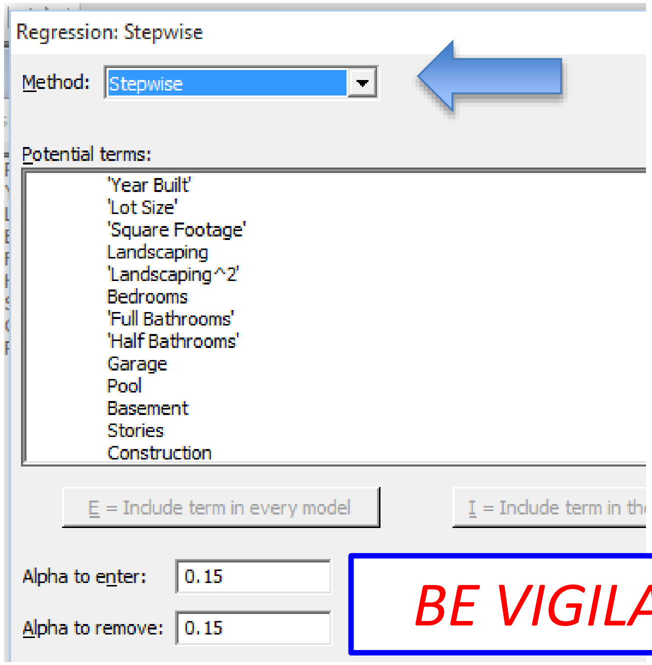
Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Stepwise)

- Theory behind Stepwise:
  - ✓ Starts with the strongest single predictor
  - ✓ Incrementally adds to the model until the best result is achieved
- Useful when:
  - ✓ There is a lot of multicollinearity
  - ✓ There are a lot of predictors
  - ✓ You want to study interactions

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Stepwise)



Regression: Stepwise

Method: Stepwise

Potential terms:

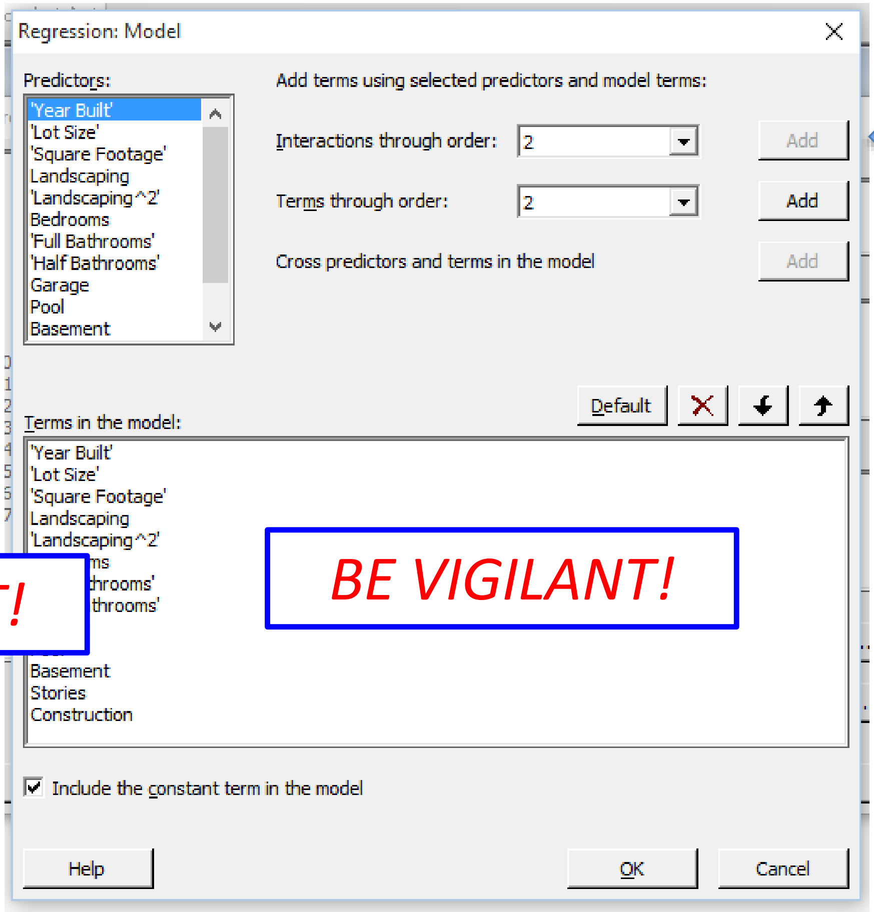
'Year Built'
'Lot Size'
'Square Footage'
Landscaping
'Landscaping^2'
Bedrooms
'Full Bathrooms'
'Half Bathrooms'
Garage
Pool
Basement
Stories
Construction

Alpha to enter: 0.15

Alpha to remove: 0.15

BE VIGILANT!

Select "Stepwise"



Regression: Model

Predictors:

'Year Built'
'Lot Size'
'Square Footage'
Landscaping
'Landscaping^2'
Bedrooms
'Full Bathrooms'
'Half Bathrooms'
Garage
Pool
Basement

Add terms using selected predictors and model terms:

Interactions through order: 2

Terms through order: 2

Cross predictors and terms in the model

Terms in the model:

'Year Built'
'Lot Size'
'Square Footage'
Landscaping
'Landscaping^2'
ms
throoms'
throoms'
Basement
Stories
Construction

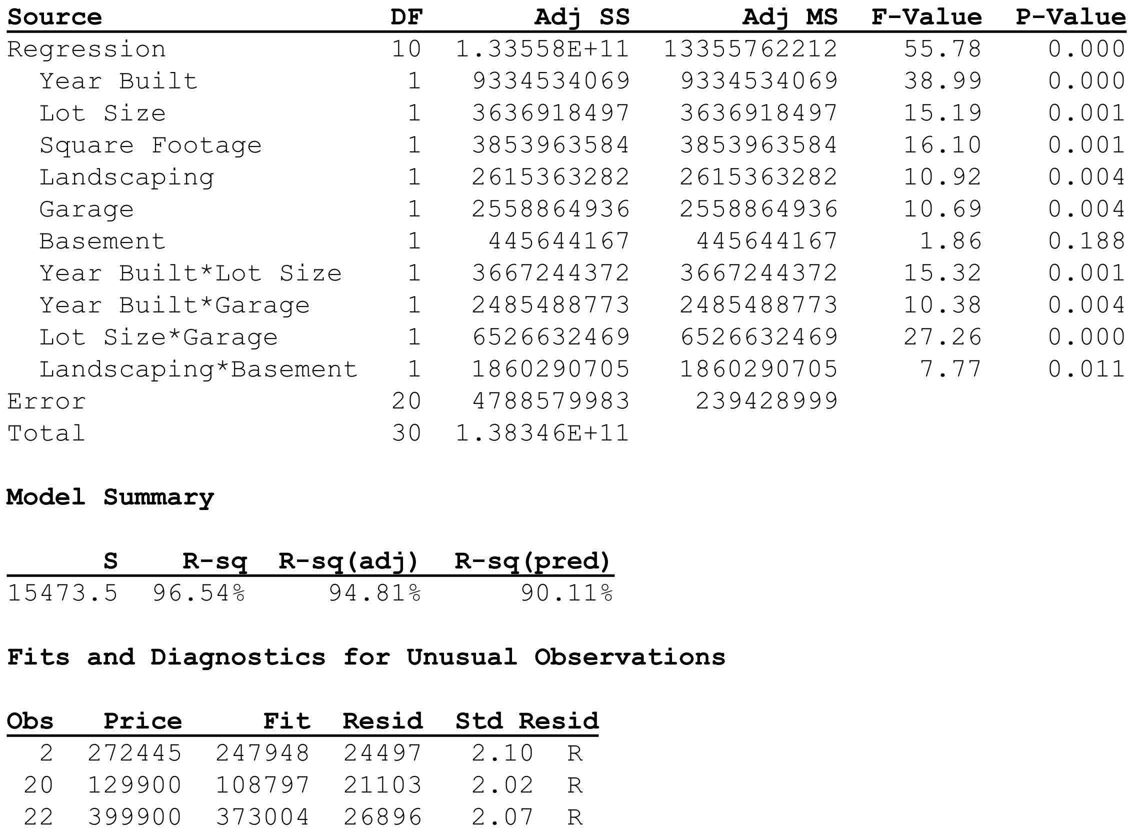
☑ Include the constant term in the model

BE VIGILANT!

Choose terms for interactions, select "Add"

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Stepwise)



```
Source                     DF       Adj SS       Adj MS  F-Value  P-Value
Regression                 10  1.33558E+11  13355762212    55.78    0.000
  Year Built                1   9334534069   9334534069    38.99    0.000
  Lot Size                  1   3636918497   3636918497    15.19    0.001
  Square Footage            1   3853963584   3853963584    16.10    0.001
  Landscaping               1   2615363282   2615363282    10.92    0.004
  Garage                    1   2558864936   2558864936    10.69    0.004
  Basement                  1    445644167    445644167     1.86    0.188
  Year Built*Lot Size       1   3667244372   3667244372    15.32    0.001
  Year Built*Garage         1   2485488773   2485488773    10.38    0.004
  Lot Size*Garage           1   6526632469   6526632469    27.26    0.000
  Landscaping*Basement      1   1860290705   1860290705     7.77    0.011
Error                      20   4788579983    239428999
Total                      30  1.38346E+11

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
15473.5  96.54%     94.81%      90.11%

Fits and Diagnostics for Unusual Observations

Obs   Price     Fit  Resid  Std Resid
  2  272445  247948  24497       2.10  R
 20  129900  108797  21103       2.02  R
 22  399900  373004  26896       2.07  R
```

Summary of the Final Model

Multiple Linear Regression Data Analysis

Step 3: Reduce the model (Stepwise)

- Check your work:
  - ✓ Is the multicollinearity acceptable
  - ✓ Are there a lot of unusual observations?
  - ✓ Is there a large difference between the r-sq (adj.) and r-sq (pred.) values?
- The Stepwise method, Best Subsets method, and the ANOVA Table method may not give the same result
- Inclusion of interaction effects can make coefficients seem illogical

Multiple Linear Regression Data Analysis

Summary of Three Methods

Measure	ANOVA Table	Best Subsets	Stepwise
Number of Terms	6	7	10
R-sq	89.65	90.01	96.54
R-sq (adj.)	86.50	86.37	94.81
R-sq (pred.)	78.58	75.84	90.11
Number of Unusual Observations	1	2	3

Which is the best model?

Multiple Linear Regression Data Analysis

Step 4: Check the residuals

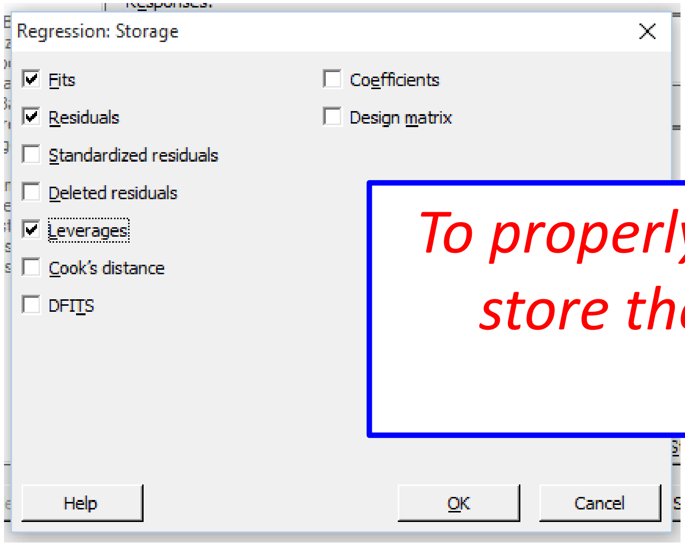
- Residuals must be:
  - ✓ Normally distributed
  - ✓ Independent
  - ✓ Homoscedastic (of equal variance) across the model

Violations of the ANOVA assumptions result in an unstable model or a model that has poor predictive ability

Multiple Linear Regression Data Analysis

Step 4: Check the residuals

- Residuals must be:
  - ✓ Normally distributed
  - ✓ Independent
  - ✓ Homoscedastic (of equal variance) across the model



Regression: Storage

☑ Fits
☑ Residuals
☐ Standardized residuals
☐ Deleted residuals
☑ Leverages
☐ Cook's distance
☐ DFITS
☐ Coefficients
☐ Design matrix

To properly analyze the residuals, store the Fits, Residuals, and Leverages

Multiple Linear Regression Data Analysis

Step 4: Check the residuals

- Generate the following:
  - ✓ Normality Plot of the Residuals
  - ✓ I-MR Control Chart of the Residuals
  - ✓ Scatter Plot of the Residuals vs. Fits

Multiple Linear Regression Data Analysis

Step 4: Check the residuals

Normality Plot:

- Outliers / non-normality
  - ✓ Check leverage

Scatter Plot:

- Smiles or frowns
  - ✓ Missed curvature
- Shark's mouth
  - ✓ Unstable model
- Significant outliers
  - ✓ Check leverage

Individuals Control Chart:

- Violate the control limits
  - ✓ Check leverage
- Patterns
  - ✓ Check noise factors

$$Max\ Leverage = \frac{2p}{n}$$

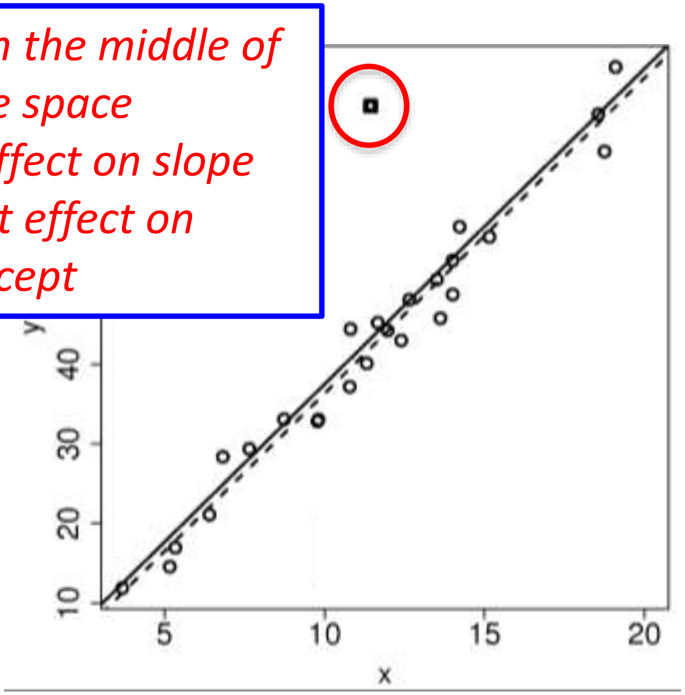
p = terms in the model + constant

n = number of rows of data

Multiple Linear Regression Data Analysis

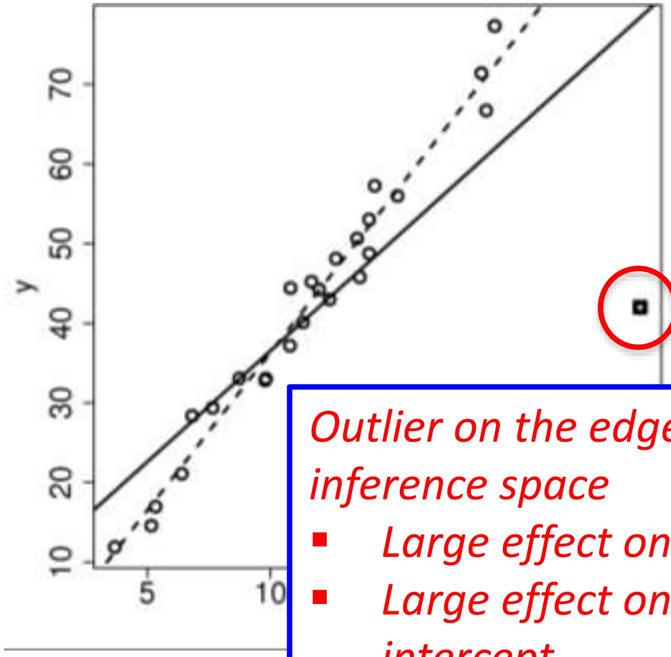
Step 4: Check the residuals

- Why worry about outliers and leverage:
  - ✓ Outliers can significantly impact the slope and intercept of the regression line



Outlier in the middle of inference space

- *No effect on slope*
- *Slight effect on intercept*

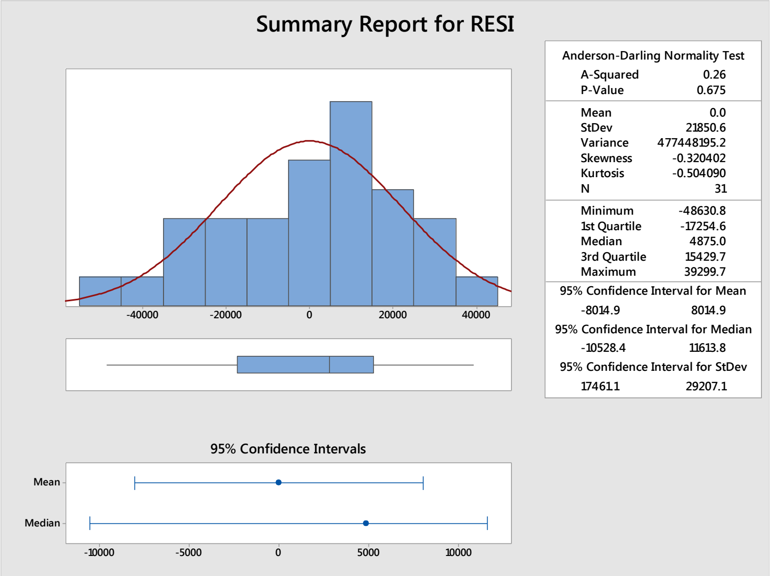


Outlier on the edges of inference space

- *Large effect on slope*
- *Large effect on intercept*

Multiple Linear Regression Data Analysis

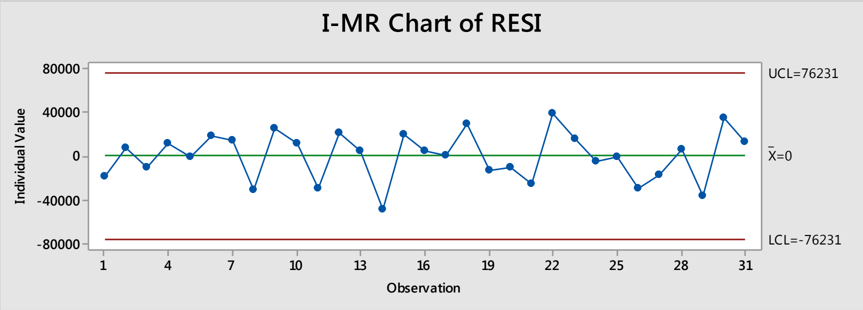
Step 4: Check the residuals (ANOVA Table Model)

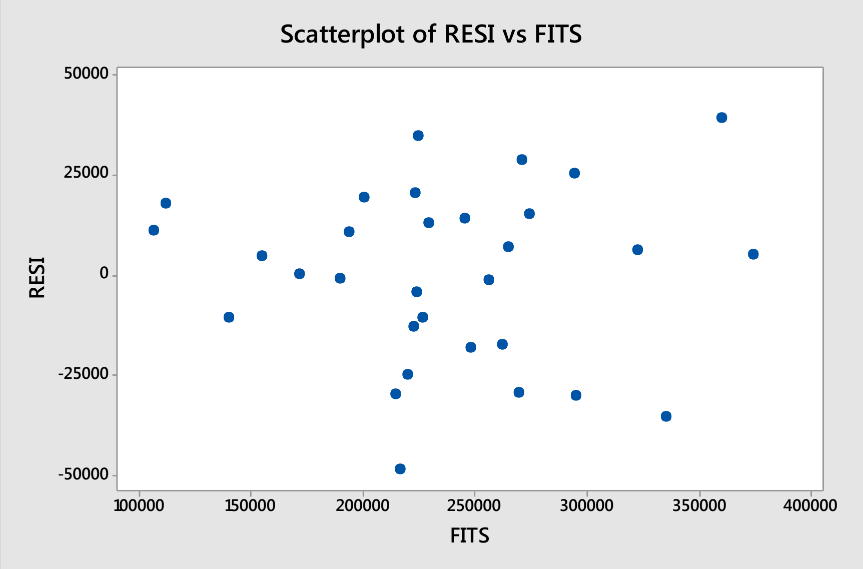


```
Anderson-Darling Normality Test
   A-Squared              0.26
   P-Value               0.675

   Mean                    0.0
   StDev               21850.6
   Variance        477448195.2
   Skewness          -0.320402
   Kurtosis          -0.504090
   N                        31

   Minimum            -48630.8
   1st Quartile       -17254.6
   Median               4875.0
   3rd Quartile        15429.7
   Maximum             39299.7
95% Confidence Interval for Mean
   -8014.9              8014.9
95% Confidence Interval for Median
   -10528.4            11613.8
95% Confidence Interval for StDev
   17461.1             29207.1
```

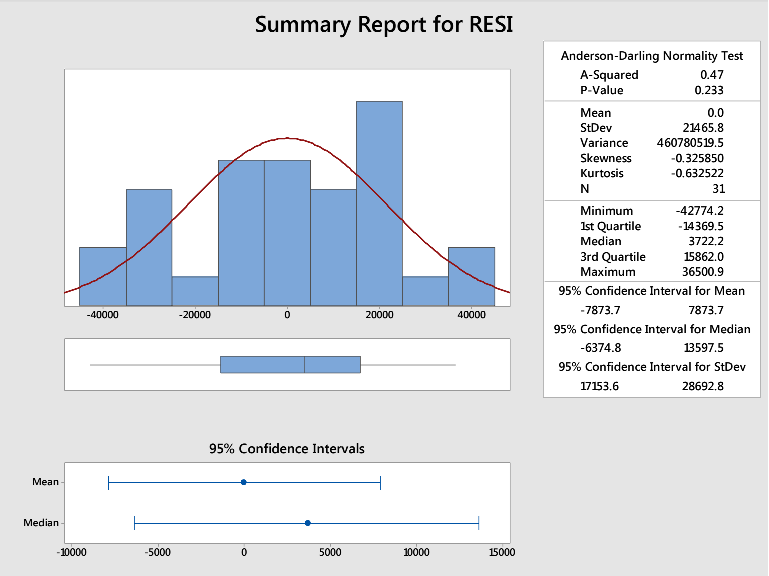




This is a stable and usable model

Multiple Linear Regression Data Analysis

Step 4: Check the residuals (Best Subsets Model)



Summary Report for RESI

Anderson-Darling Normality Test

A-Squared	0.47
P-Value	0.233
Mean	0.0
StDev	21465.8
Variance	460780519.5
Skewness	-0.325850
Kurtosis	-0.632522
N	31
Minimum	-42774.2
1st Quartile	-14369.5
Median	3722.2
3rd Quartile	15862.0
Maximum	36500.9

95% Confidence Interval for Mean

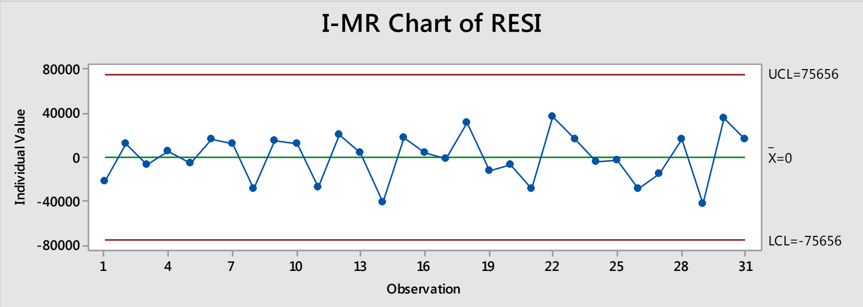
-7873.7 7873.7

95% Confidence Interval for Median

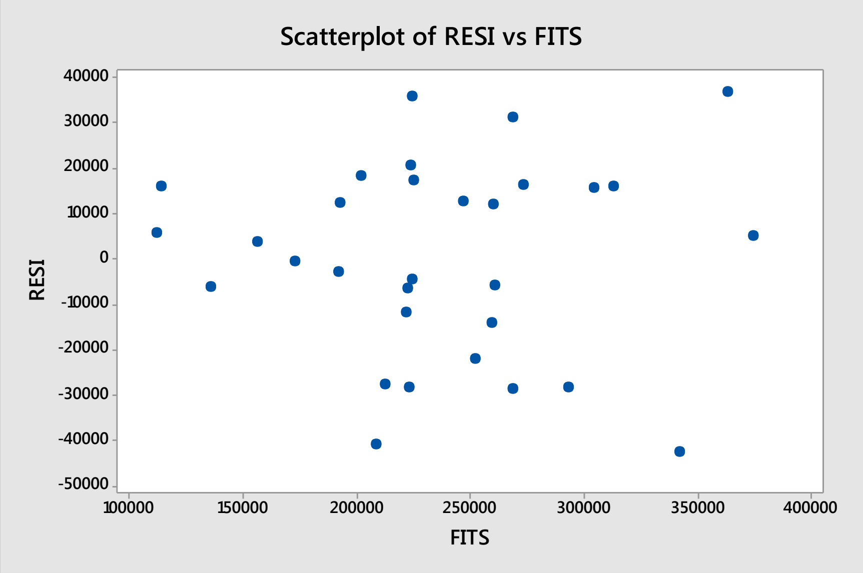
-6374.8 13597.5

95% Confidence Interval for StDev

17153.6 28692.8



I-MR Chart of RESI

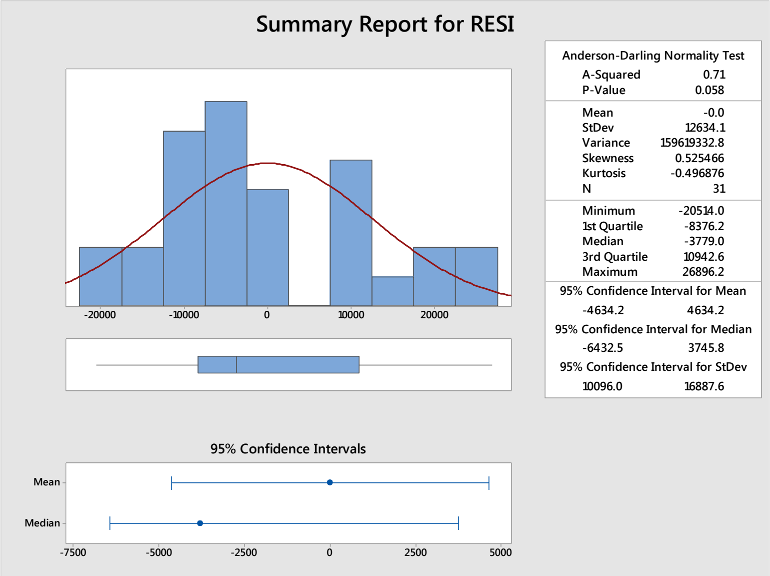


Scatterplot of RESI vs FITS

This is a stable and usable model

Multiple Linear Regression Data Analysis

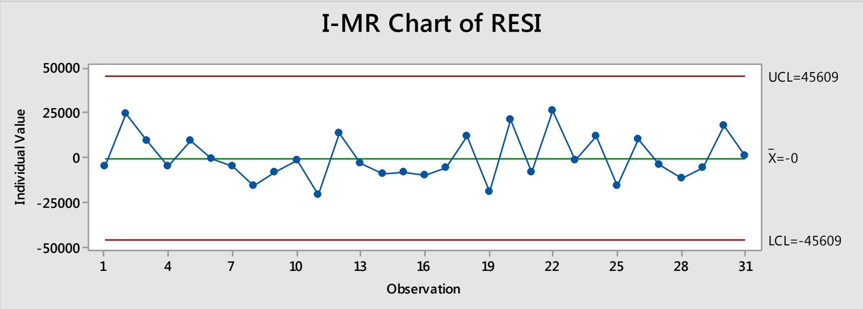
Step 4: Check the residuals (Stepwise Model)

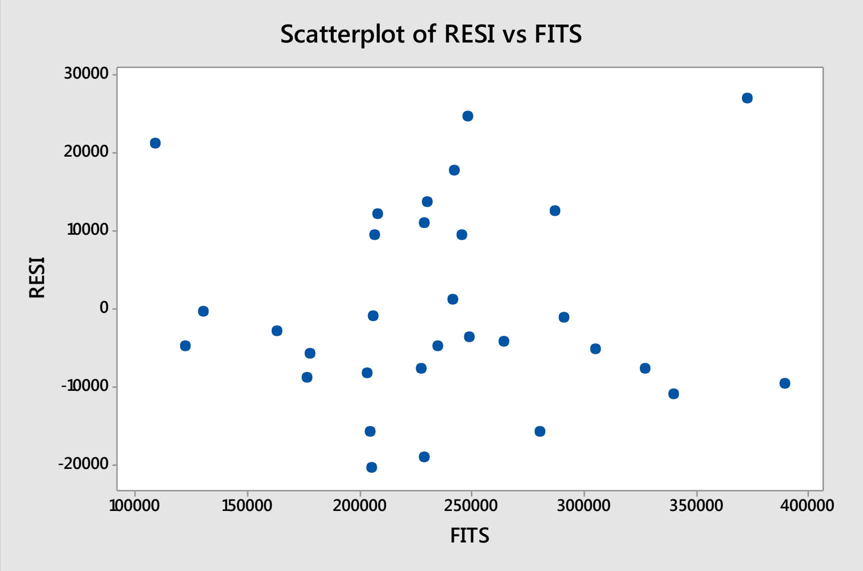

Summary Report for RESI

```
Anderson-Darling Normality Test
   A-Squared              0.71
   P-Value               0.058

   Mean                   -0.0
   StDev               12634.1
   Variance        159619332.8
   Skewness           0.525466
   Kurtosis          -0.496876
   N                        31

   Minimum            -20514.0
   1st Quartile        -8376.2
   Median              -3779.0
   3rd Quartile        10942.6
   Maximum             26896.2
95% Confidence Interval for Mean
   -4634.2              4634.2
95% Confidence Interval for Median
   -6432.5              3745.8
95% Confidence Interval for StDev
   10096.0             16887.6
```



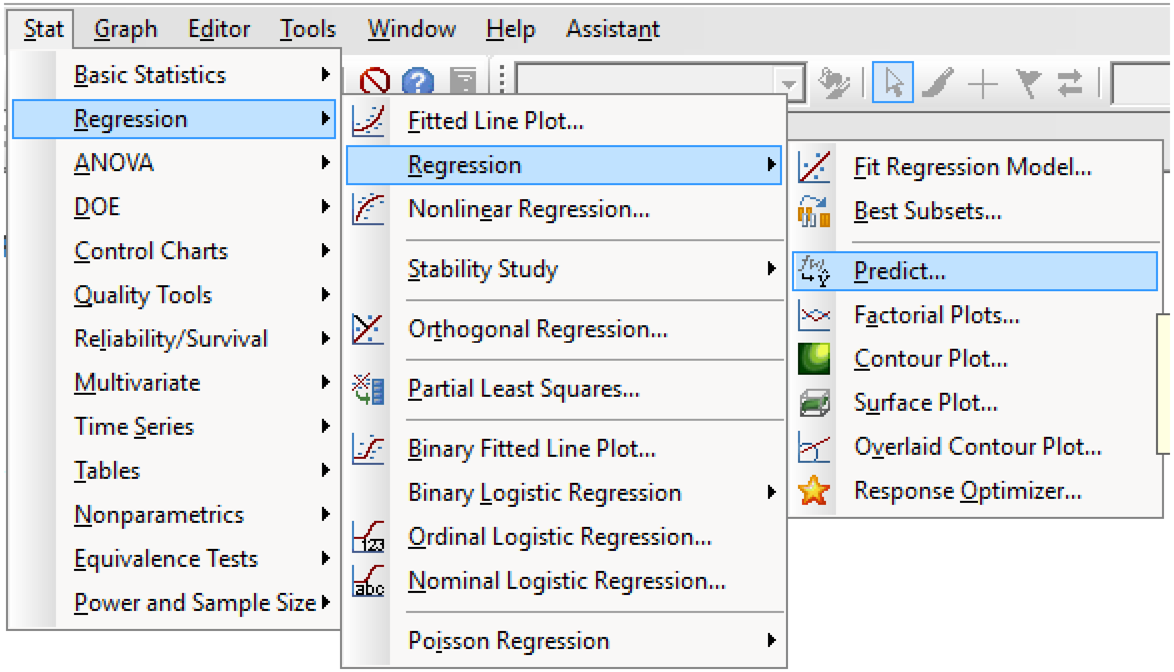


This is a stable and usable model

Multiple Linear Regression Data Analysis

Step 5: Validate the model

- Collect a data point that was not used to generate the model
- Use the regression model's prediction interval



Multiple Linear Regression Data Analysis

Step 5: Validate the model

28384 Hunter Ct,
Flat Rock, MI 48134

● FOR SALE
$255,000

Predictor	Value
Year Built	2001
Lot Size (acres)	0.28
Square Footage	2,440
Quality of Landscaping	5
Number of Bedrooms	4
Number of Full Bathrooms	2
Number of Half Bathrooms	1
Garage Size	3
Pool	No
Finished Basement	No
Number of Stories	2
Construction Material	Brick

Multiple Linear Regression Data Analysis

Step 5: Validate the model

Model	Model Exact Fit	Prediction Interval
ANOVA Table	$283,677	($228,508, $338,845)
Best Subsets	$282,056	($226,374, $337,738)
Stepwise	$291,265	($251,568, $330,963)

28384 Hunter Ct,
Flat Rock, MI 48134

FOR SALE
$255,000

***MODEL VALIDATED**—List price included in the prediction interval*

*The **TYPICAL HOUSE** with these features should be listed at the **EXACT FIT** of the model.*

Other features not gleaned from the listing move the price within the prediction interval

Multiple Linear Regression Data Analysis

28384 Hunter Ct,
Flat Rock, MI 48134

FOR SALE
$255,000

Using the Prediction Interval:

- Commands a higher price:
  - ✓ Cul-de-sac lot
  - ✓ Privacy fence
  - ✓ Hardwood / new floors
- Commands a lower price:
  - ✓ Busy street
  - ✓ Older or basic appliances
  - ✓ Non-neutral colors

Perhaps the model can be enhanced by adding these variables, although they are harder to assess from the listing

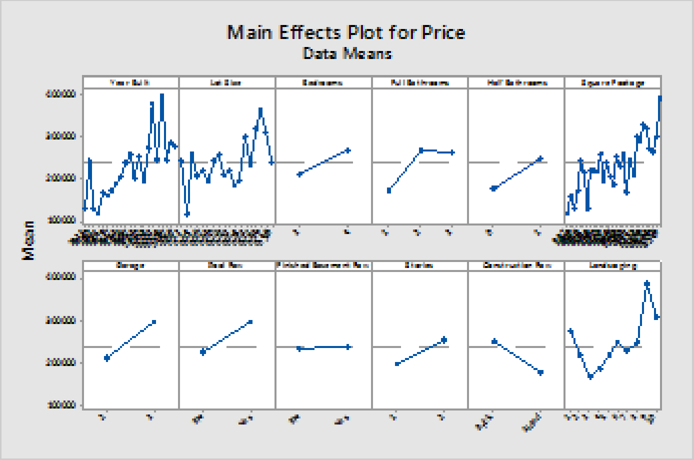
Application

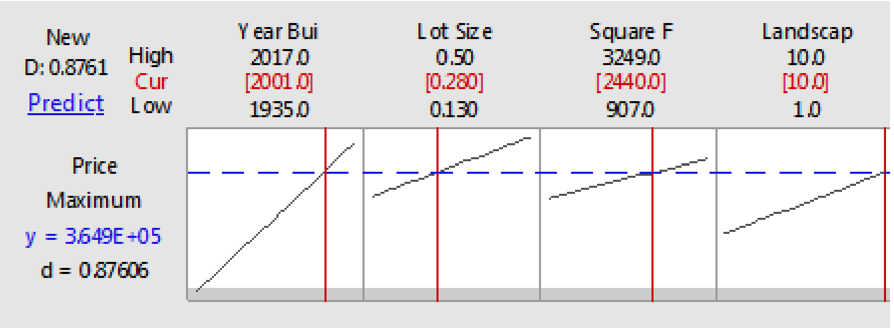
Think as an INVESTOR, not a STATISTICIAN:

- Look for a price well-below the TYPICAL price, preferably outside of the prediction interval
- Inspect the house to determine what you can change or improve
- Prioritize based on return on investment

Buy LOW

Sell HIGH





Factorial Plots / Response Modeler

Application

Think as an INVESTOR, not a STATISTICIAN:

House Feature	Changeable	Rank	Make the Most Profit
Square Footage	Yes	1	Addition / remodel unused space
Year Built	No	2	N/A
Lot Size	No	3	N/A
Landscaping	Yes	4	Professionally done; add deck or patio
Bedrooms	Yes	5	Convert den or loft to bedroom
Full Bathrooms	Yes	6	Convert half bath to full bath
Garage	No	7	N/A
Basement	Yes	8	Finish the basement with usable area

Application

Step 6: Predict

Flip Budget

- Convert half bath to full bath ($4,000)
- Convert attic above garage to bedroom ($16,000)
- Landscape with deck and/or patio ($11,000)
- Replace flooring ($7,000)
- Upgrade kitchen counters and appliances ($6,000)

```
Fit      SE Fit         95% PI
323392   27859.4   (265456, 381329)
```

Purchase Price: $255,000

Investment: $44,000

Selling Price: $364,000

Profit: $65,000

Review Session Objectives

Now that the session has concluded, do you:

- Know when to use multiple linear regression
- Understand how to collect and analyze data using multiple linear regression
- Realize the dangers of multicollinearity and troublesome residuals—and how to handle them
- Recognize the power of multiple linear regression for problem solving through analysis of real estate data

ASQ®

The Global Voice of Quality®

Questions and Discussion

Ford

Go Further