SAND2007-5030P

Workshop: Analyzing the Data You Already Have – An Introduction to Model Fitting

Don Lifke

Corporate Lean / Six Sigma Black Belt

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Objectives

- Learn new ways to look at existing data
- Understand limitations with typical data analysis
- Understand what modeling is and why it is useful
- Learn to use statistical software and understand why Excel is not the best choice for efficient statistical analysis

Define “Statistics”

Define “statistics” in two words:

_____ _____

Exercise

You want to sell your house. It has the following features:

- 2000 square feet
- 0.2 acre lot
- 2 years old
- 3 bedrooms
- 3 full bathrooms

What should your asking price be?

Open *House Data for Summit Tutorial.xls*

What type of analyses would you do on this data?

-
-
-

Spend 5 minutes determining what you'd list your house for, using the Excel data.

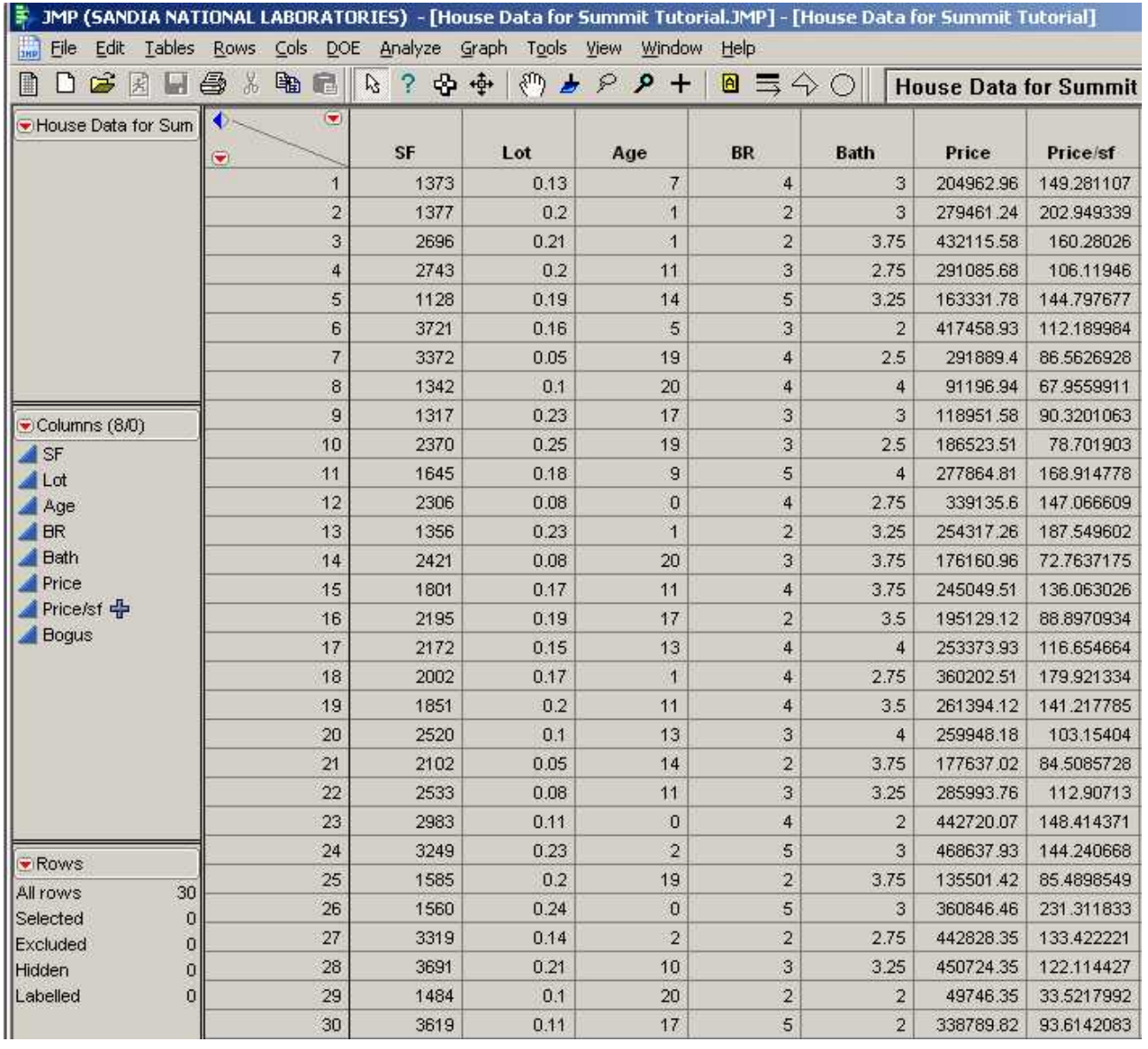
Exercise Time

5 minutes

List Class Prices

- Listing Prices:
  - $
  - $
  - $
  - $
  - $
  - $

Open *House Data for Summit Tutorial.jmp* Data File



	SF	Lot	Age	BR	Bath	Price	Price/sf
1	1373	0.13	7	4	3	204962.96	149.281107
2	1377	0.2	1	2	3	279461.24	202.949339
3	2696	0.21	1	2	3.75	432115.58	160.28026
4	2743	0.2	11	3	2.75	291085.68	106.11946
5	1128	0.19	14	5	3.25	163331.78	144.797677
6	3721	0.16	5	3	2	417458.93	112.189984
7	3372	0.05	19	4	2.5	291889.4	86.5626928
8	1342	0.1	20	4	4	91196.94	67.9559911
9	1317	0.23	17	3	3	118951.58	90.3201063
10	2370	0.25	19	3	2.5	186523.51	78.701903
11	1645	0.18	9	5	4	277864.81	168.914778
12	2306	0.08	0	4	2.75	339135.6	147.066609
13	1356	0.23	1	2	3.25	254317.26	187.549602
14	2421	0.08	20	3	3.75	176160.96	72.7637175
15	1801	0.17	11	4	3.75	245049.51	136.063026
16	2195	0.19	17	2	3.5	195129.12	88.8970934
17	2172	0.15	13	4	4	253373.93	116.654664
18	2002	0.17	1	4	2.75	360202.51	179.921334
19	1851	0.2	11	4	3.5	261394.12	141.217785
20	2520	0.1	13	3	4	259948.18	103.15404
21	2102	0.05	14	2	3.75	177637.02	84.5085728
22	2533	0.08	11	3	3.25	285993.76	112.90713
23	2983	0.11	0	4	2	442720.07	148.414371
24	3249	0.23	2	5	3	468637.93	144.240668
25	1585	0.2	19	2	3.75	135501.42	85.4898549
26	1560	0.24	0	5	3	360846.46	231.311833
27	3319	0.14	2	2	2.75	442828.35	133.422221
28	3691	0.21	10	3	3.25	450724.35	122.114427
29	1484	0.1	20	2	2	49746.35	33.5217992
30	3619	0.11	17	5	2	338789.82	93.6142083

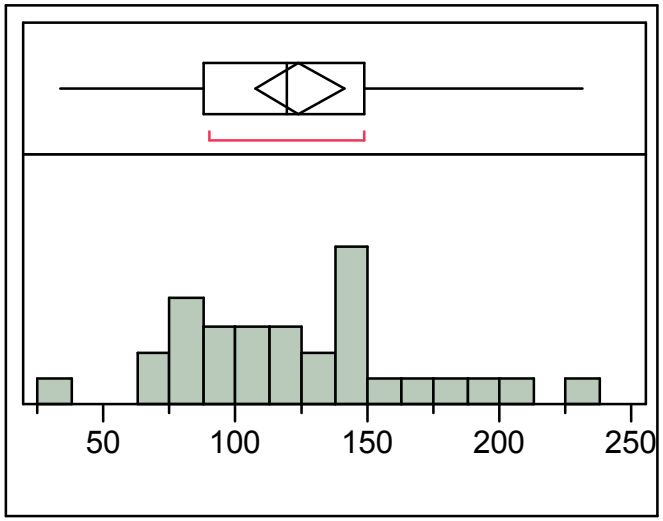
Exercise:
What Will Your Listing Price Be?

Analyze the distribution of Price/sf

Average = $124.36 per SF

Distributions

Price/sf



Quantiles

100.0%	maximum	231.31
99.5%		231.31
97.5%		231.31
90.0%		186.79
75.0%	quartile	148.63
50.0%	median	119.38
25.0%	quartile	88.31
10.0%		73.36
2.5%		33.52
0.5%		33.52
0.0%	minimum	33.52

Moments

Mean	124.36354
Std Dev	44.020787
Std Err Mean	8.0370594
upper 95% Mean	140.80117
lower 95% Mean	107.92591
N	30

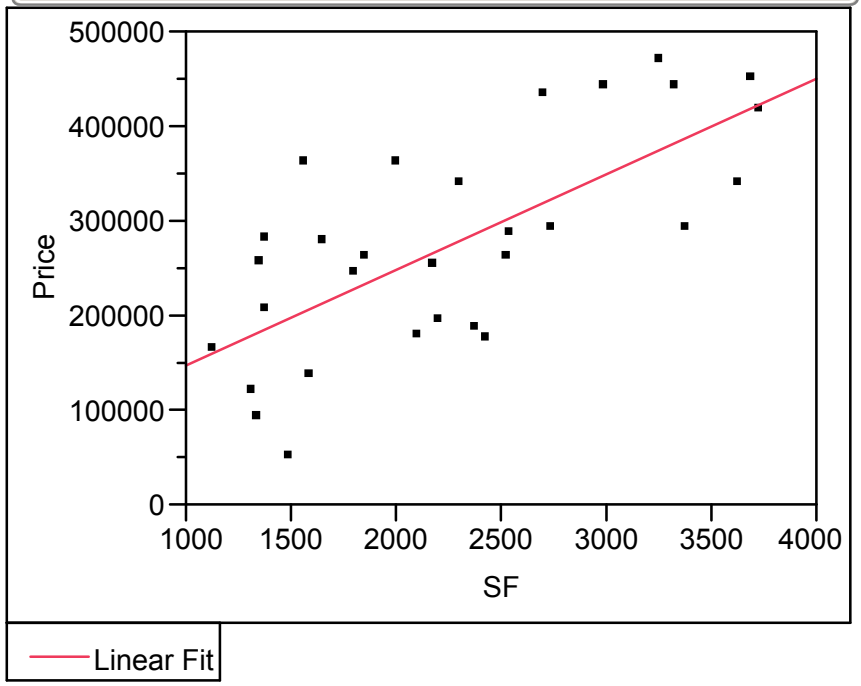
Sandia National Laboratories

Exercise:
What Will Your Listing Price Be?

- Perform a Fit Y by X for Price vs. SF
- Add a Line Fit

Price = $45,962 + $101.34*SF

Bivariate Fit of Price By SF



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1.8791e+11	1.879e+11	28.1954
Error	28	1.8661e+11	6.6647e+9	**Prob > F**
C. Total	29	3.7452e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	45962.927	45654	1.01	0.3227
SF	101.33845	19.0847	5.31	<.0001*

Sandia National Laboratories

Exercise Time

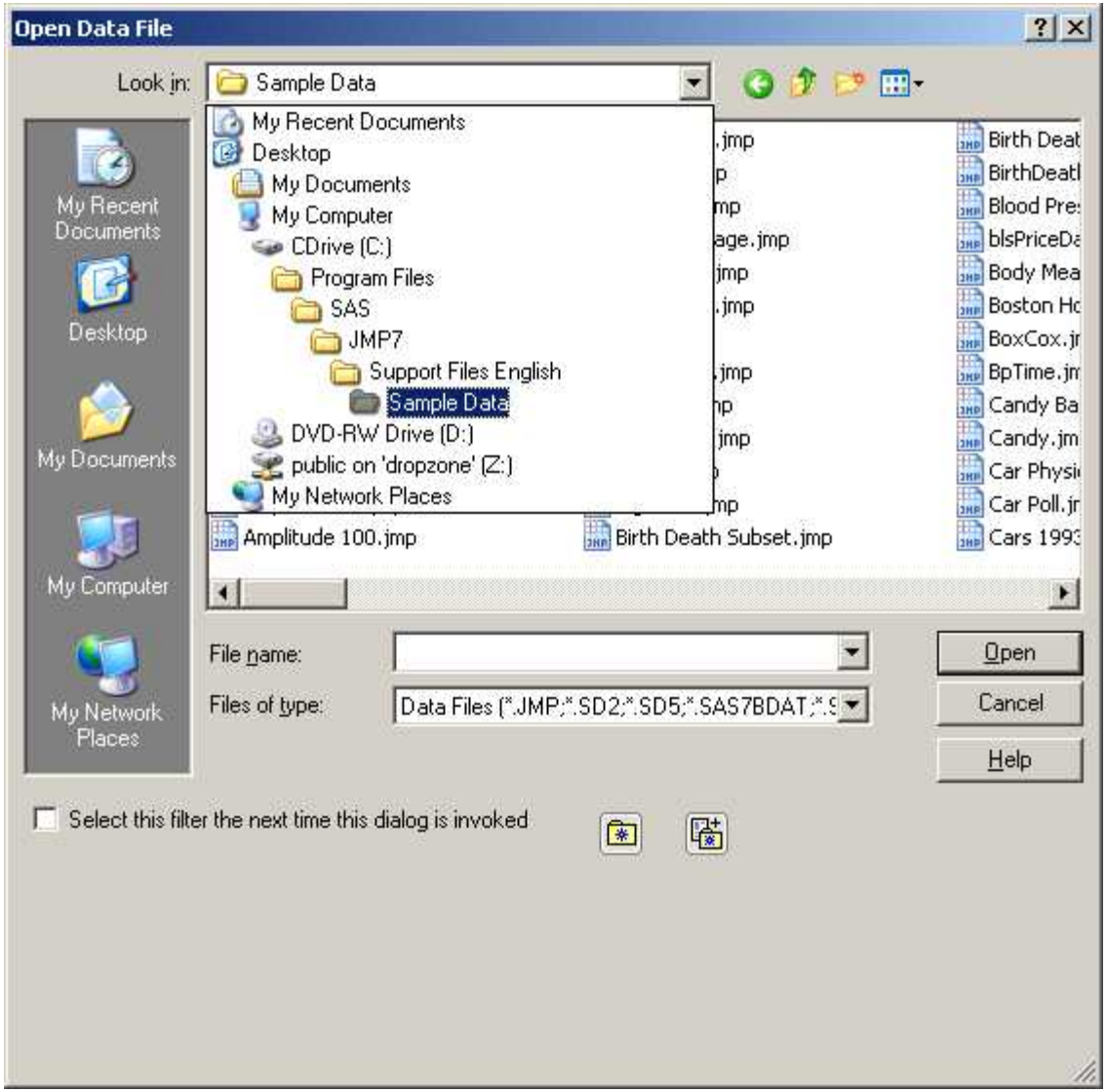
Using this new information, Spend 5 minutes determining what you'd list your house for, using the JMP data.

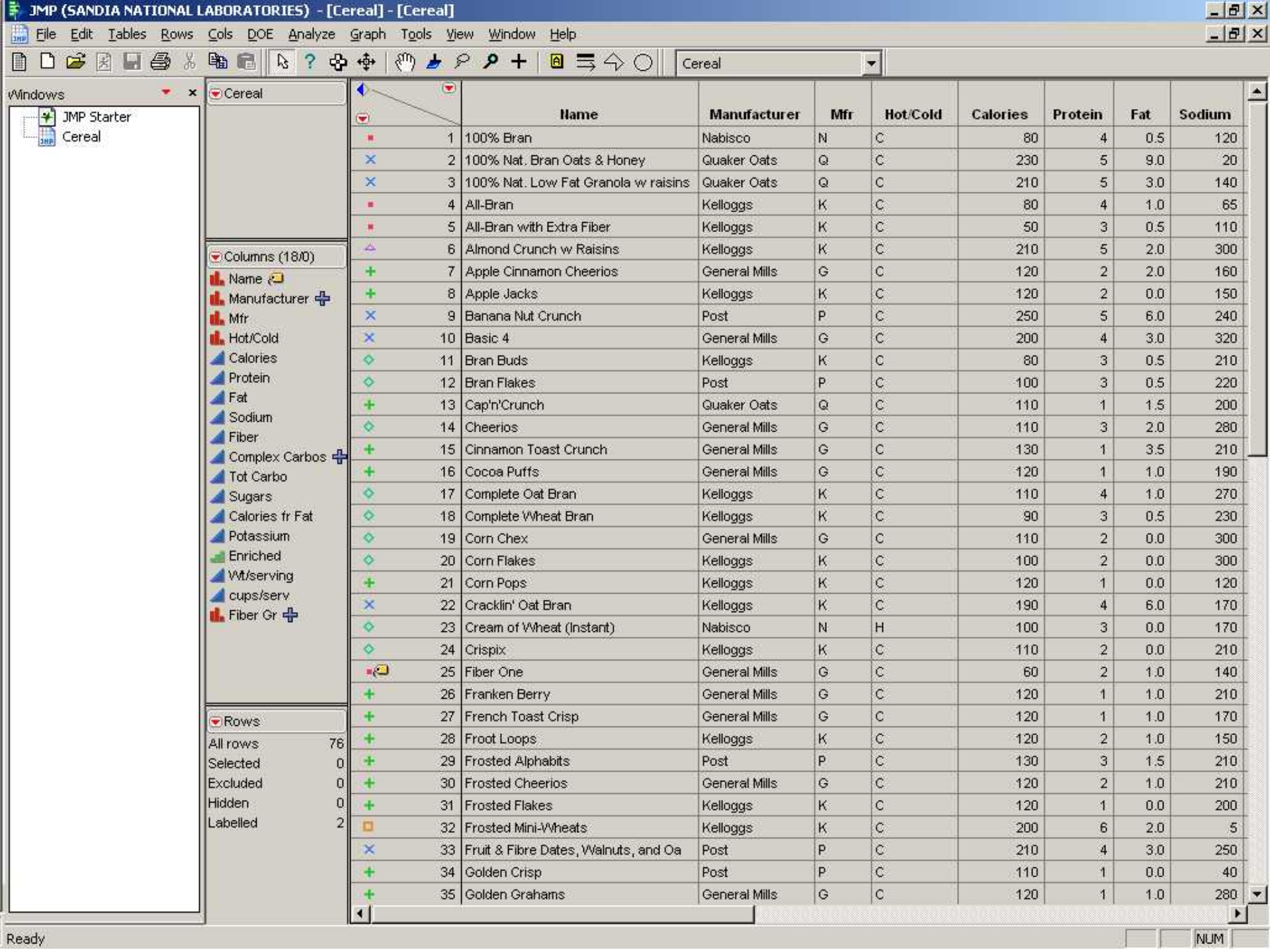
5 minutes

Exercise:
What Will Your Listing Price Be?

- Write Student's Listing Prices on the Board
- Put These in JMP for Future Reference
- We'll Return to this Problem Later

Analyzing the Data You Have: Open Cereal File

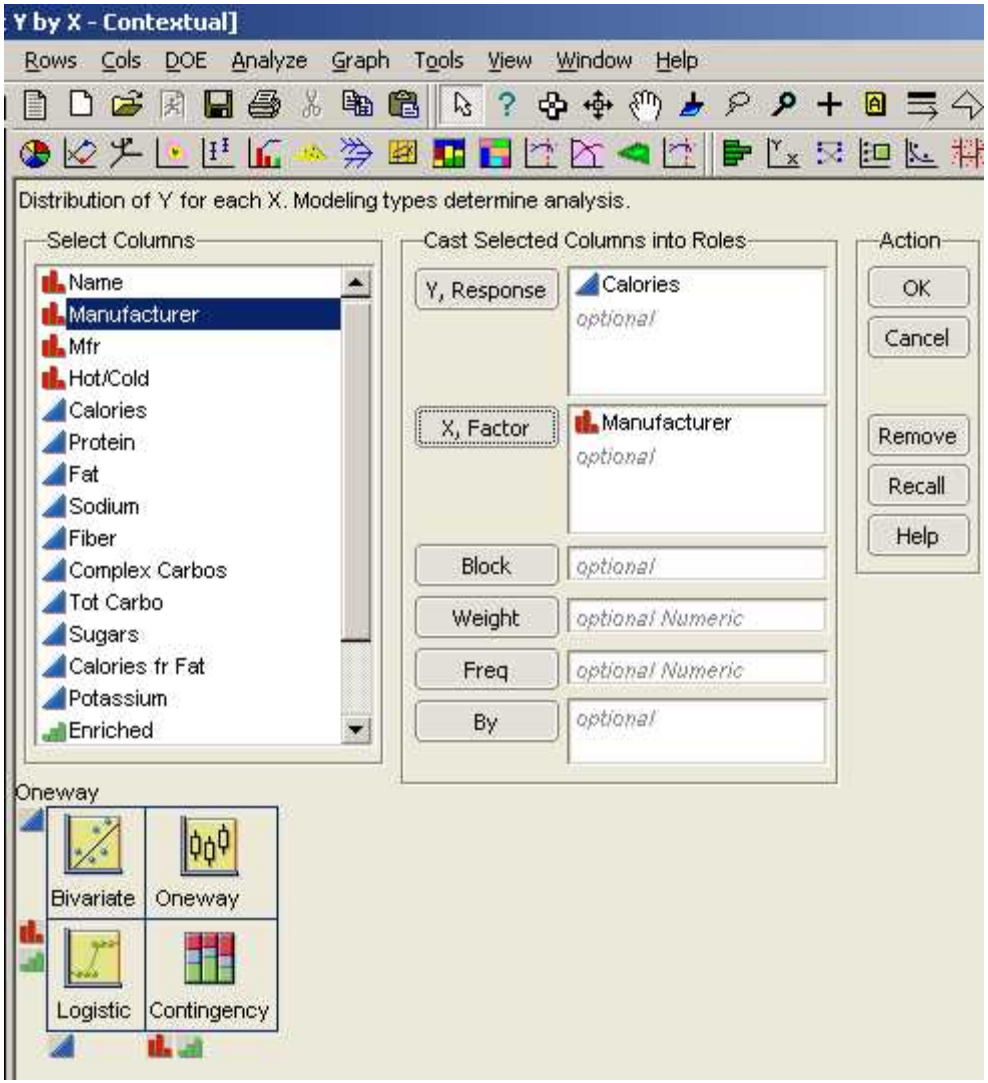



JMP (SANDIA NATIONAL LABORATORIES) - [Cereal] - [Cereal]
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Cereal
Windows
JMP Starter
Cereal
Cereal
Columns (18/0)
Name
Manufacturer
Mfr
Hot/Cold
Calories
Protein
Fat
Sodium
Fiber
Complex Carbos
Tot Carbo
Sugars
Calories fr Fat
Potassium
Enriched
Wt/serving
cups/serv
Fiber Gr
Rows
All rows 76
Selected 0
Excluded 0
Hidden 0
Labelled 2
Name Manufacturer Mfr Hot/Cold Calories Protein Fat Sodium
1 100% Bran Nabisco N C 80 4 0.5 120
2 100% Nat. Bran Oats & Honey Quaker Oats Q C 230 5 9.0 20
3 100% Nat. Low Fat Granola w raisins Quaker Oats Q C 210 5 3.0 140
4 All-Bran Kelloggs K C 80 4 1.0 65
5 All-Bran with Extra Fiber Kelloggs K C 50 3 0.5 110
6 Almond Crunch w Raisins Kelloggs K C 210 5 2.0 300
7 Apple Cinnamon Cheerios General Mills G C 120 2 2.0 160
8 Apple Jacks Kelloggs K C 120 2 0.0 150
9 Banana Nut Crunch Post P C 250 5 6.0 240
10 Basic 4 General Mills G C 200 4 3.0 320
11 Bran Buds Kelloggs K C 80 3 0.5 210
12 Bran Flakes Post P C 100 3 0.5 220
13 Cap'n'Crunch Quaker Oats Q C 110 1 1.5 200
14 Cheerios General Mills G C 110 3 2.0 280
15 Cinnamon Toast Crunch General Mills G C 130 1 3.5 210
16 Cocoa Puffs General Mills G C 120 1 1.0 190
17 Complete Oat Bran Kelloggs K C 110 4 1.0 270
18 Complete Wheat Bran Kelloggs K C 90 3 0.5 230
19 Corn Chex General Mills G C 110 2 0.0 300
20 Corn Flakes Kelloggs K C 100 2 0.0 300
21 Corn Pops Kelloggs K C 120 1 0.0 120
22 Cracklin' Oat Bran Kelloggs K C 190 4 6.0 170
23 Cream of Wheat (Instant) Nabisco N H 100 3 0.0 170
24 Crispix Kelloggs K C 110 2 0.0 210
25 Fiber One General Mills G C 60 2 1.0 140
26 Franken Berry General Mills G C 120 1 1.0 210
27 French Toast Crisp General Mills G C 120 1 1.0 170
28 Froot Loops Kelloggs K C 120 2 1.0 150
29 Frosted Alphabits Post P C 130 3 1.5 210
30 Frosted Cheerios General Mills G C 120 2 1.0 210
31 Frosted Flakes Kelloggs K C 120 1 0.0 200
32 Frosted Mini-Wheats Kelloggs K C 200 6 2.0 5
33 Fruit & Fibre Dates, Walnuts, and Oa Post P C 210 4 3.0 250
34 Golden Crisp Post P C 110 1 0.0 40
35 Golden Grahams General Mills G C 120 1 1.0 280
Ready
NUM

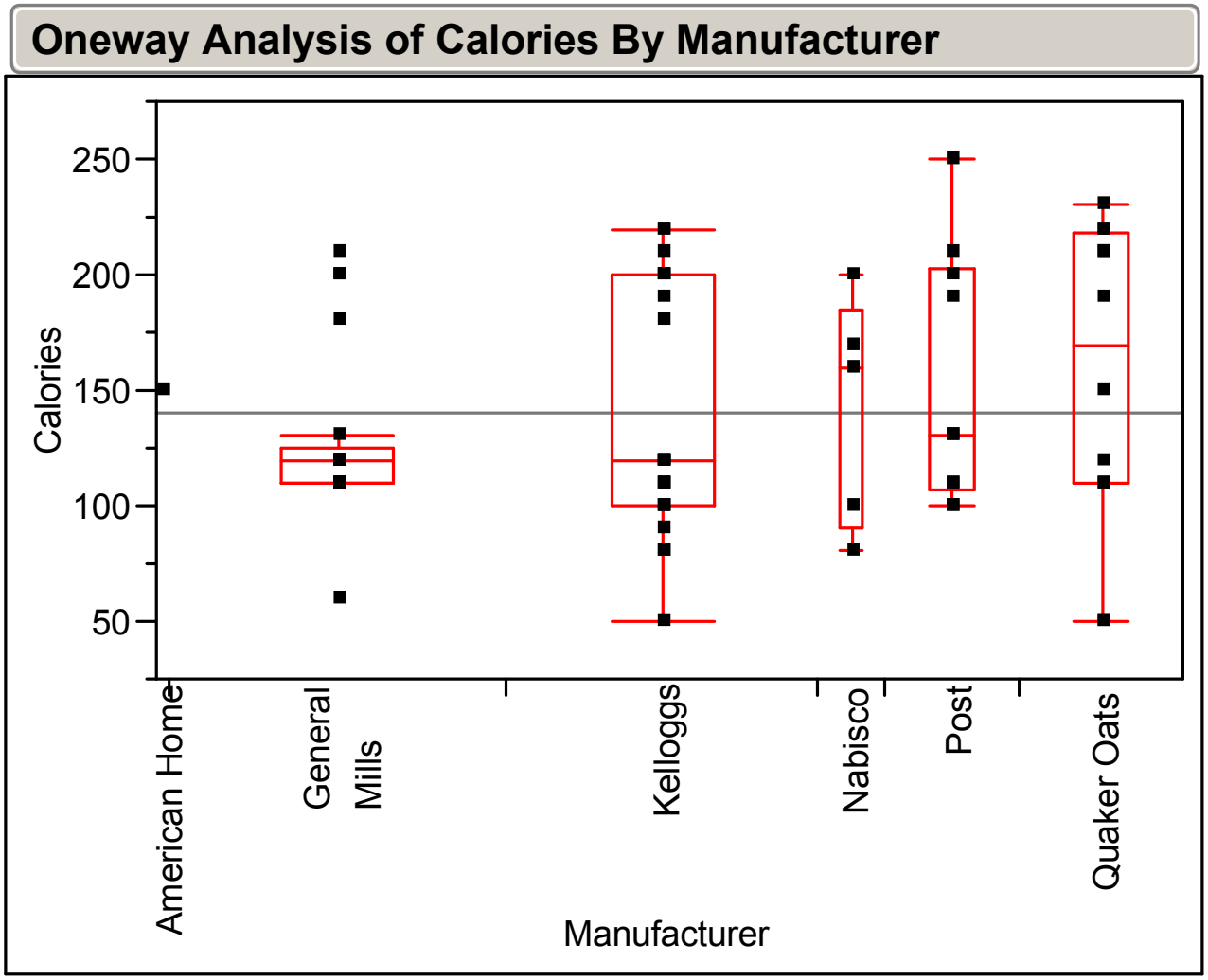
Before We Proceed

- Rows → Clear Row States
- File → Preferences
  - Click Reports
    - Change Graph Marker Size to medium.
  - Click Platforms
    - Select Distribution. Under Options, select Stack.
  - Click OK.

Does Calorie Content Depend on the Manufacturer? Analyze → Fit Y by X



Right-click Title Bar and Select Quantiles



Student's t-test

Background:

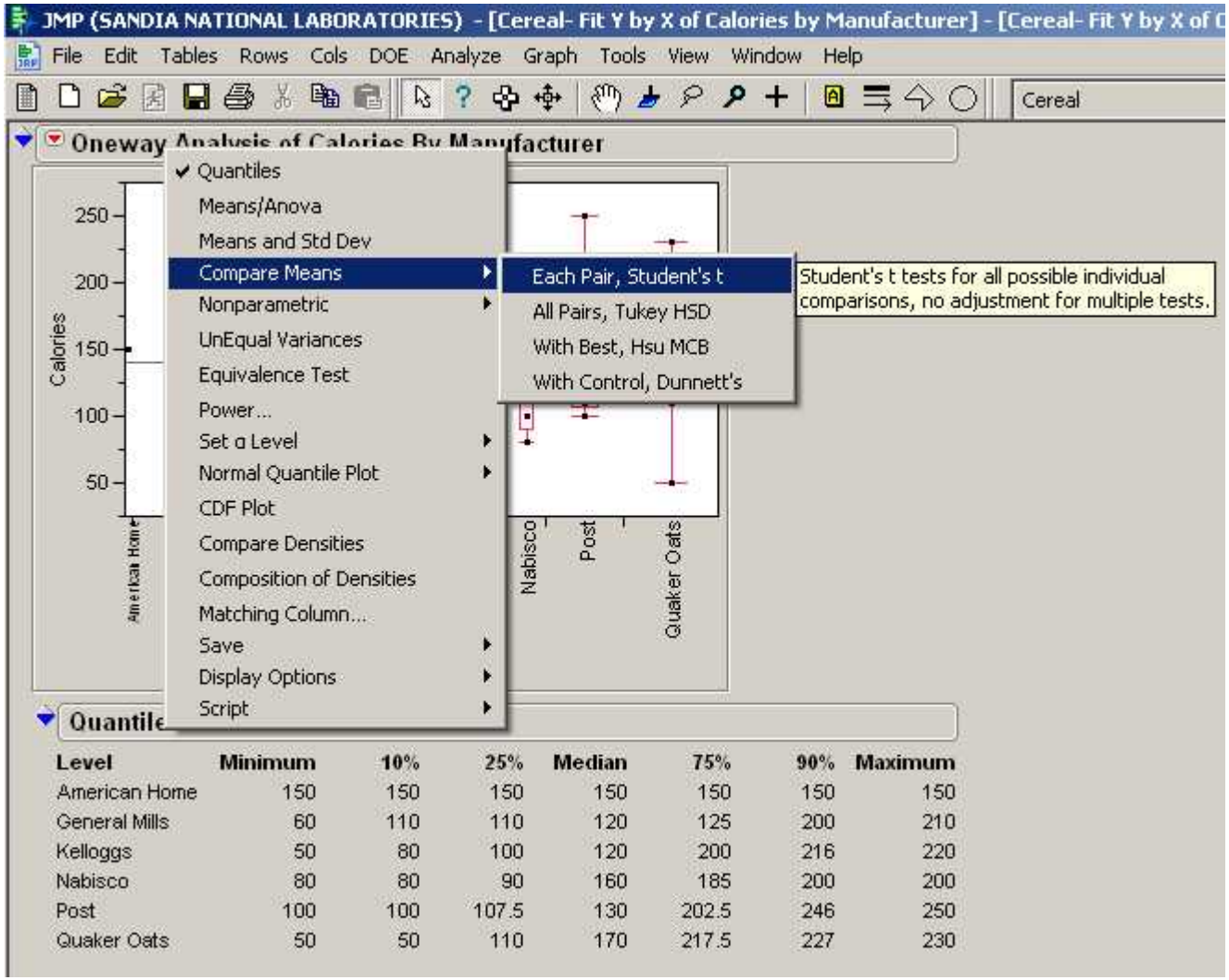
This test has evolved over the years.

- The original test was called a z-test, which uses a normal distribution as a reference distribution. However, the z-test requires knowing the true population variances - not usually the case.
- A statistician who called himself 'Student' improved the test by basing it on the t distribution, which uses variance estimates from samples. Thus the name Student's t-test.
- The Student's t-test was adapted to work if variances in the two sample groups were different. Sometimes this approach is called the Aspin-Welch Student t-test.
- Then, F.E. Satterthwaite developed a better approximation for degrees of freedom.

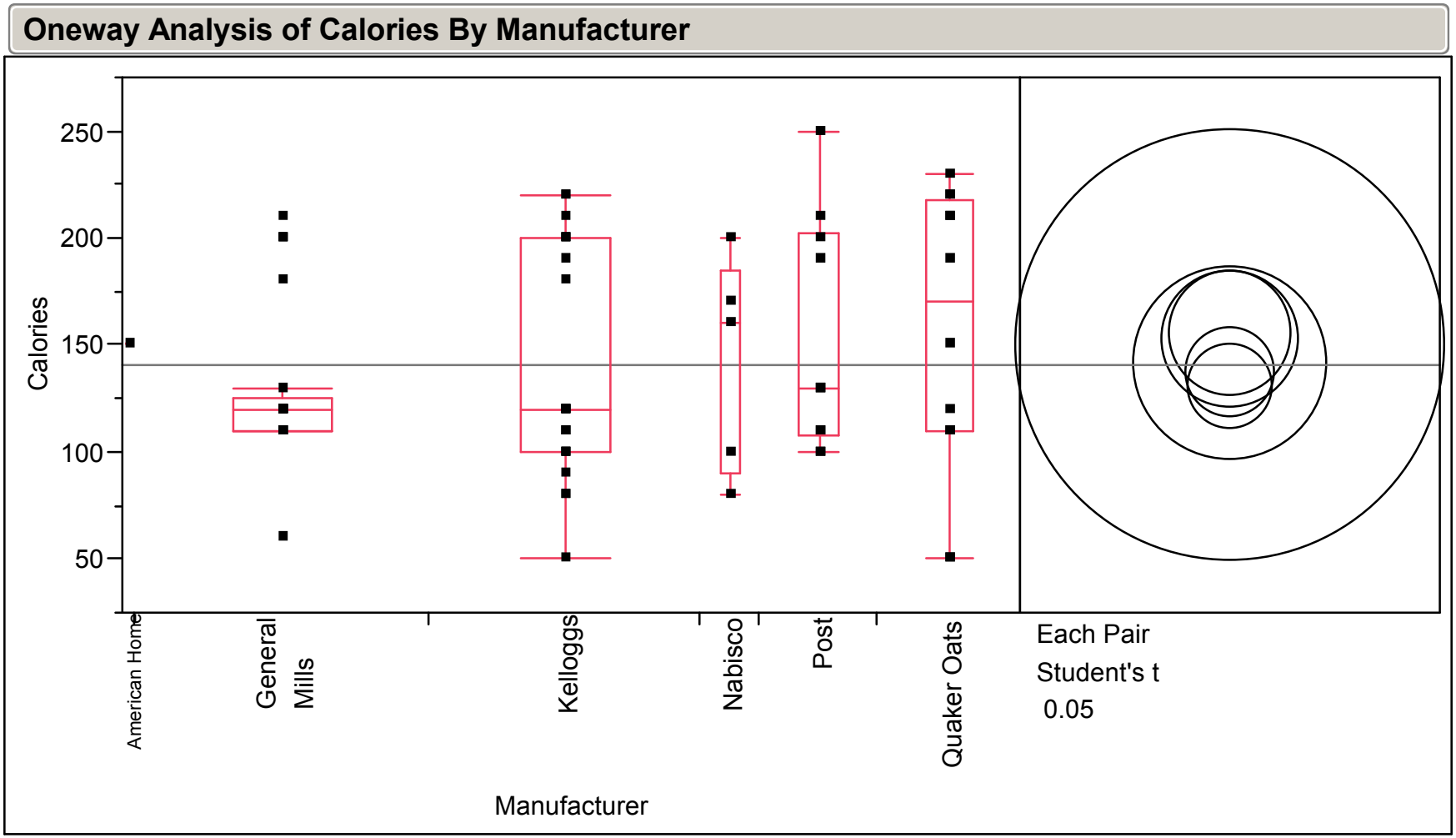
Thus the full name of this improved test is the:

Aspin-Welch-Satterthwaite-Student's t-test.
(commonly called the t-test for short).

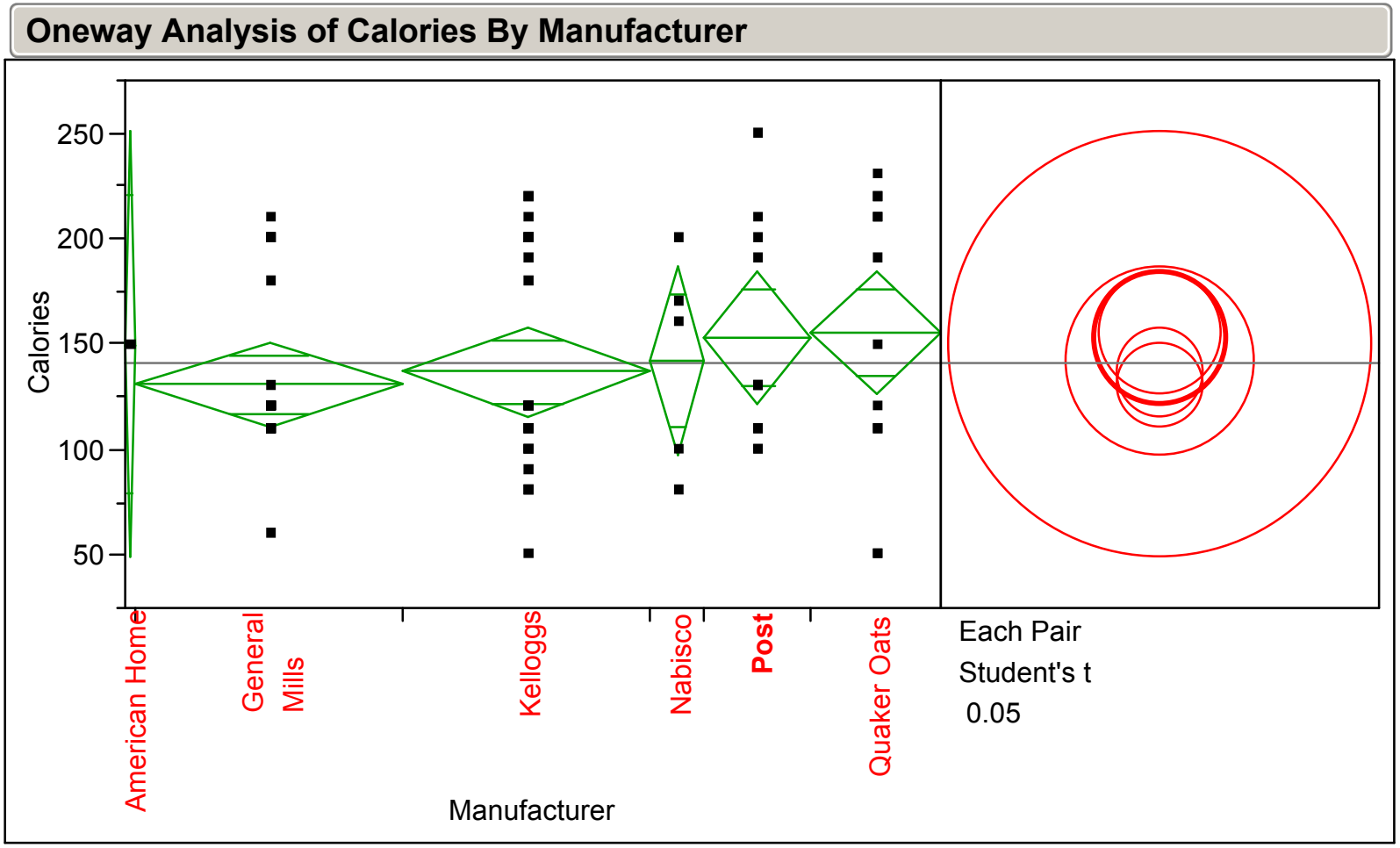
Compare Means → Each Pair, Student's t



Click Any Circle – All Others Turn Red



Right-click Title Bar – Click Means/ANOVA, Uncheck Quantiles

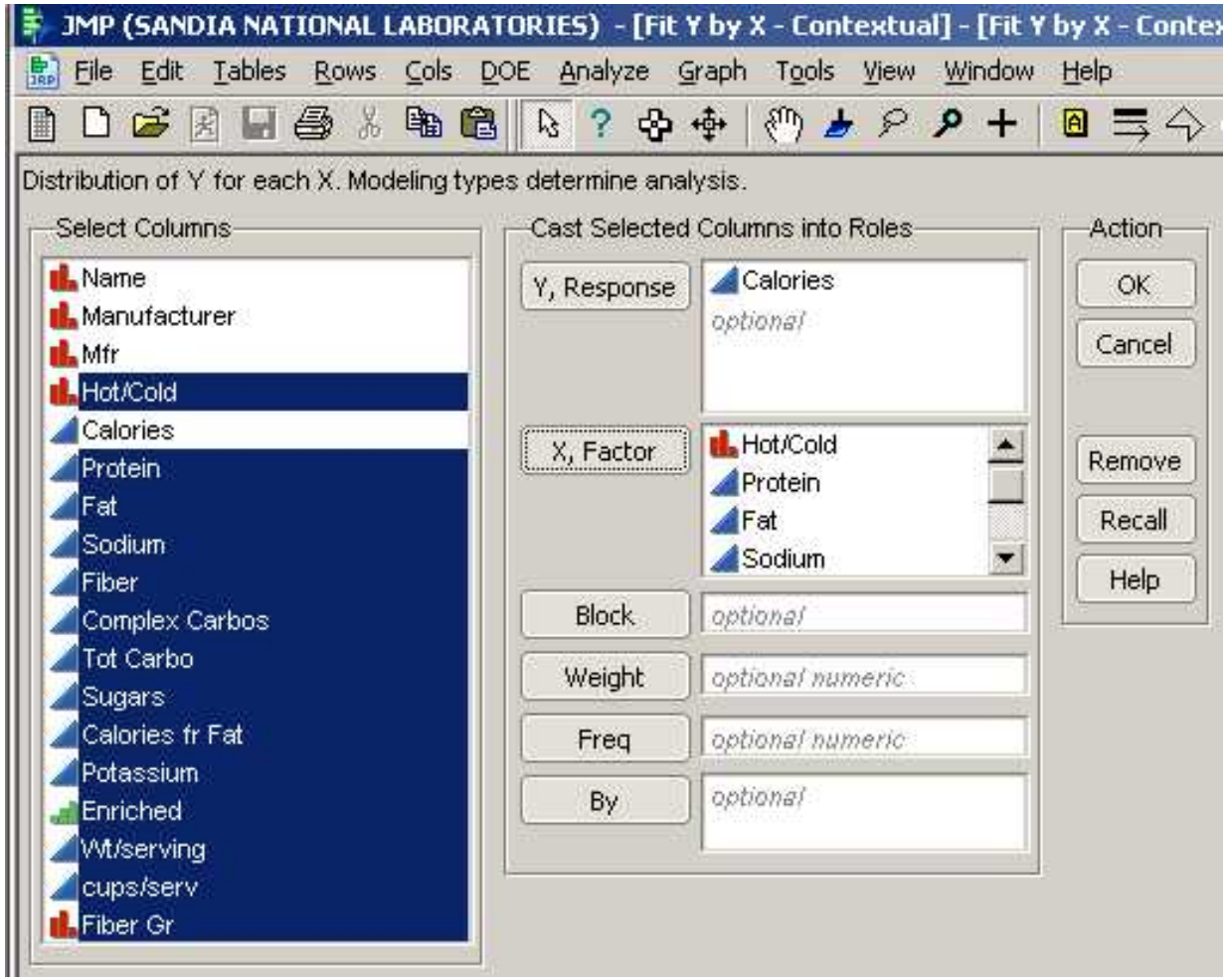


None of these are different, statistically

What Else Affects Calories?

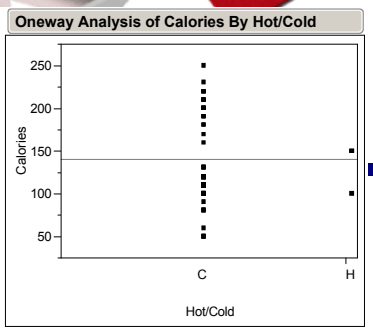
Analyze → Fit Y by X

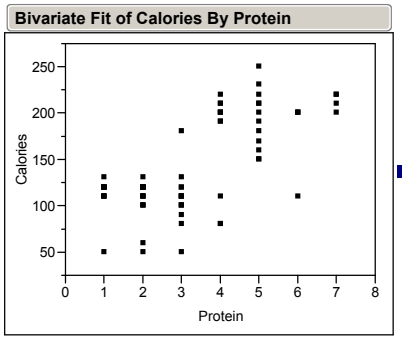
Select Calories for Y, those below for X

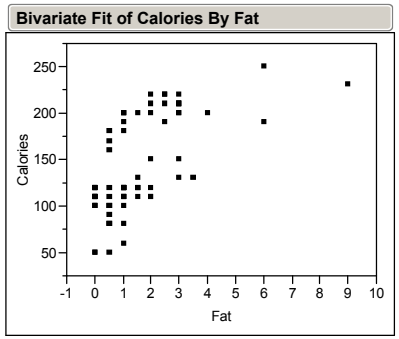


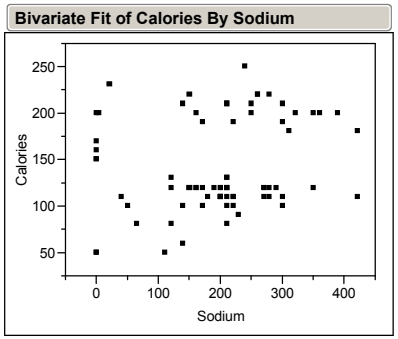
Sandia National Laboratories

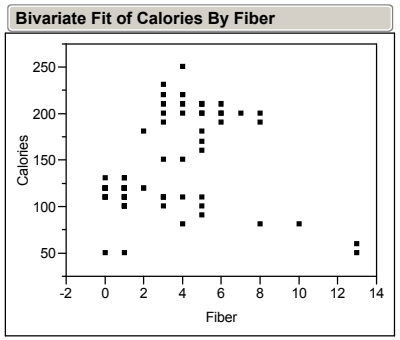
Fit Y by X Plots

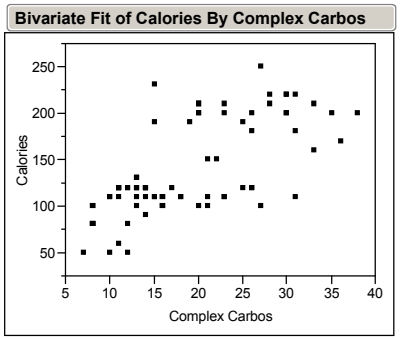


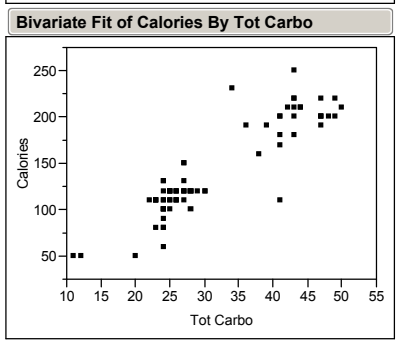


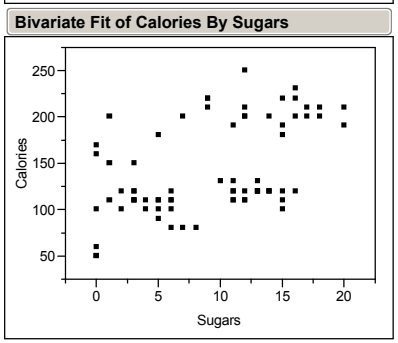


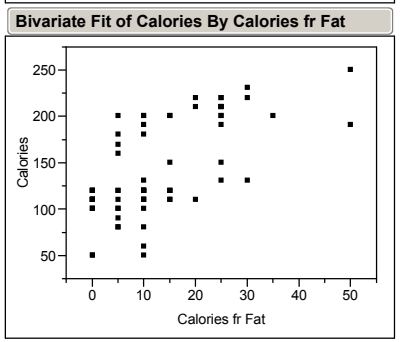


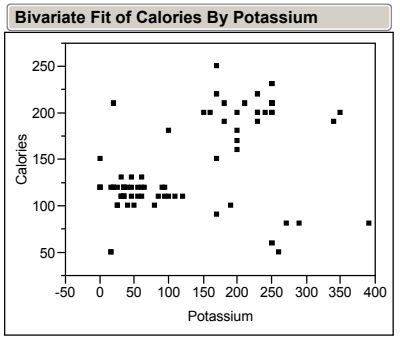


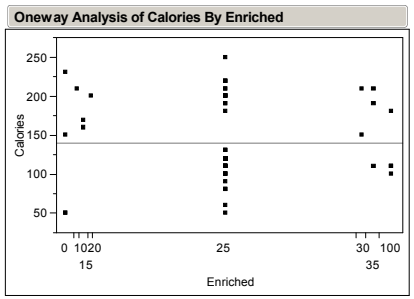


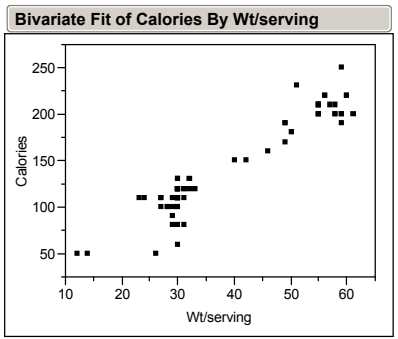


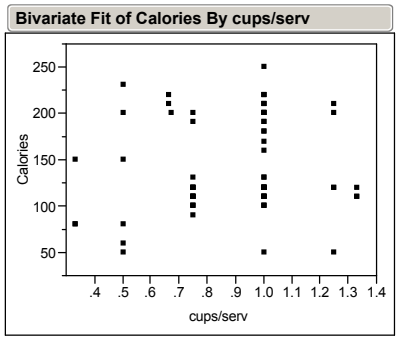


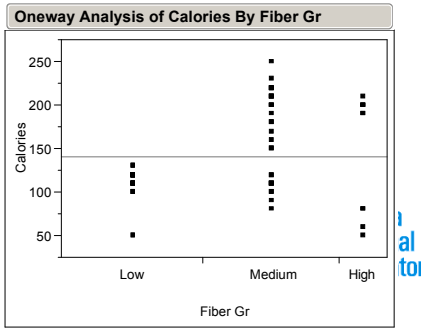






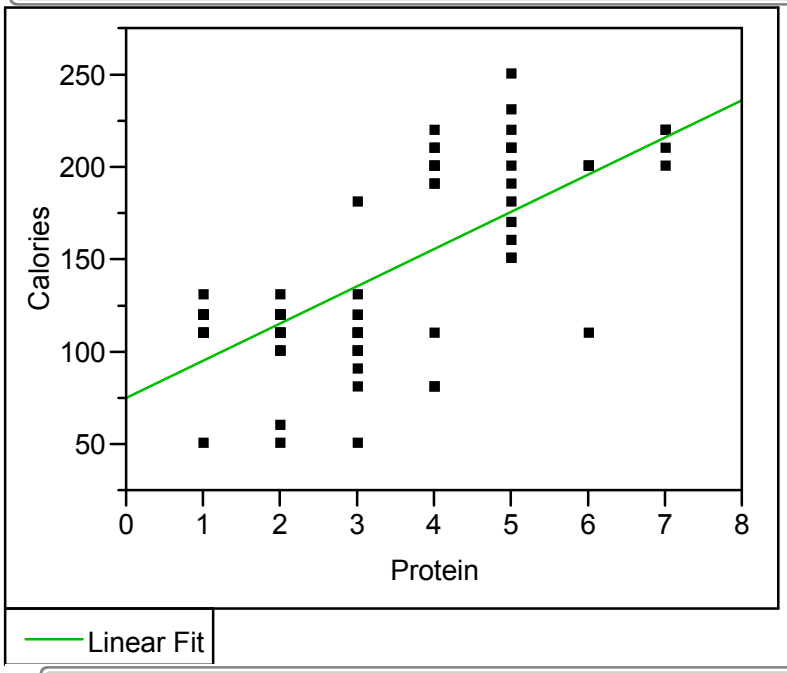






Look Closer at Protein
Right-click title bar – Fit Line

Bivariate Fit of Calories By Protein



Linear Fit

Calories = 74.874142 + 20.200669 Protein

Summary of Fit

RSquare	0.495772
RSquare Adj	0.488958
Root Mean Square Error	35.46409
Mean of Response	140.5263
Observations (or Sum Wgts)	76

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	91509.03	91509.0	72.7589
Error	74	93069.92	1257.7	Prob > F
C. Total	75	184578.95		<.0001*

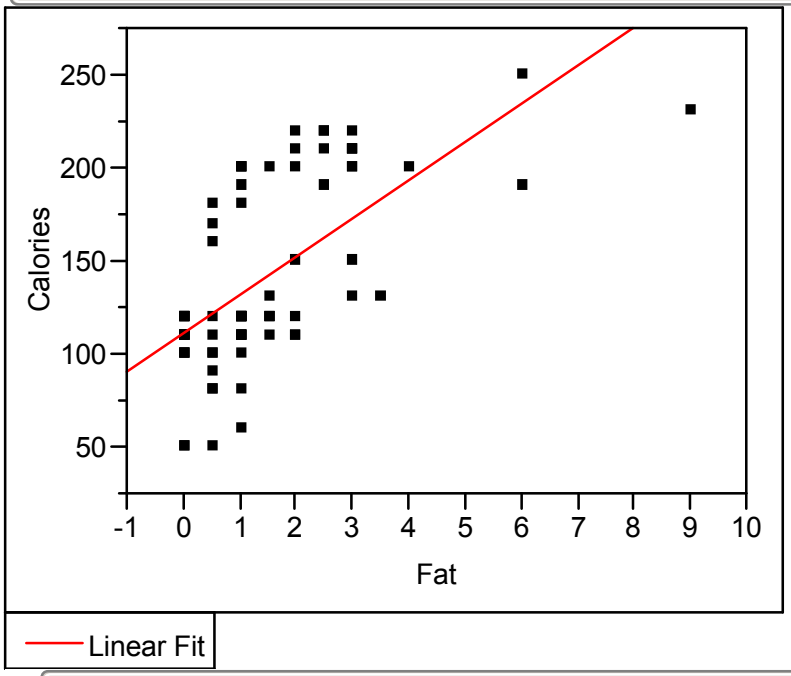
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	74.874142	8.705646	8.60	<.0001*
Protein	20.200669	2.368223	8.53	<.0001*

It appears that each g of protein adds 20.2 calories

Look Closer at Fat
Right-click title bar – Fit Line

Bivariate Fit of Calories By Fat



Linear Fit

Calories = 110.77451 + 20.555796 Fat

Summary of Fit

RSquare	0.4173
RSquare Adj	0.409425
Root Mean Square Error	38.12395
Mean of Response	140.5263
Observations (or Sum Wgts)	76

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	77024.73	77024.7	52.9949
Error	74	107554.22	1453.4	Prob > F
C. Total	75	184578.95		<.0001*

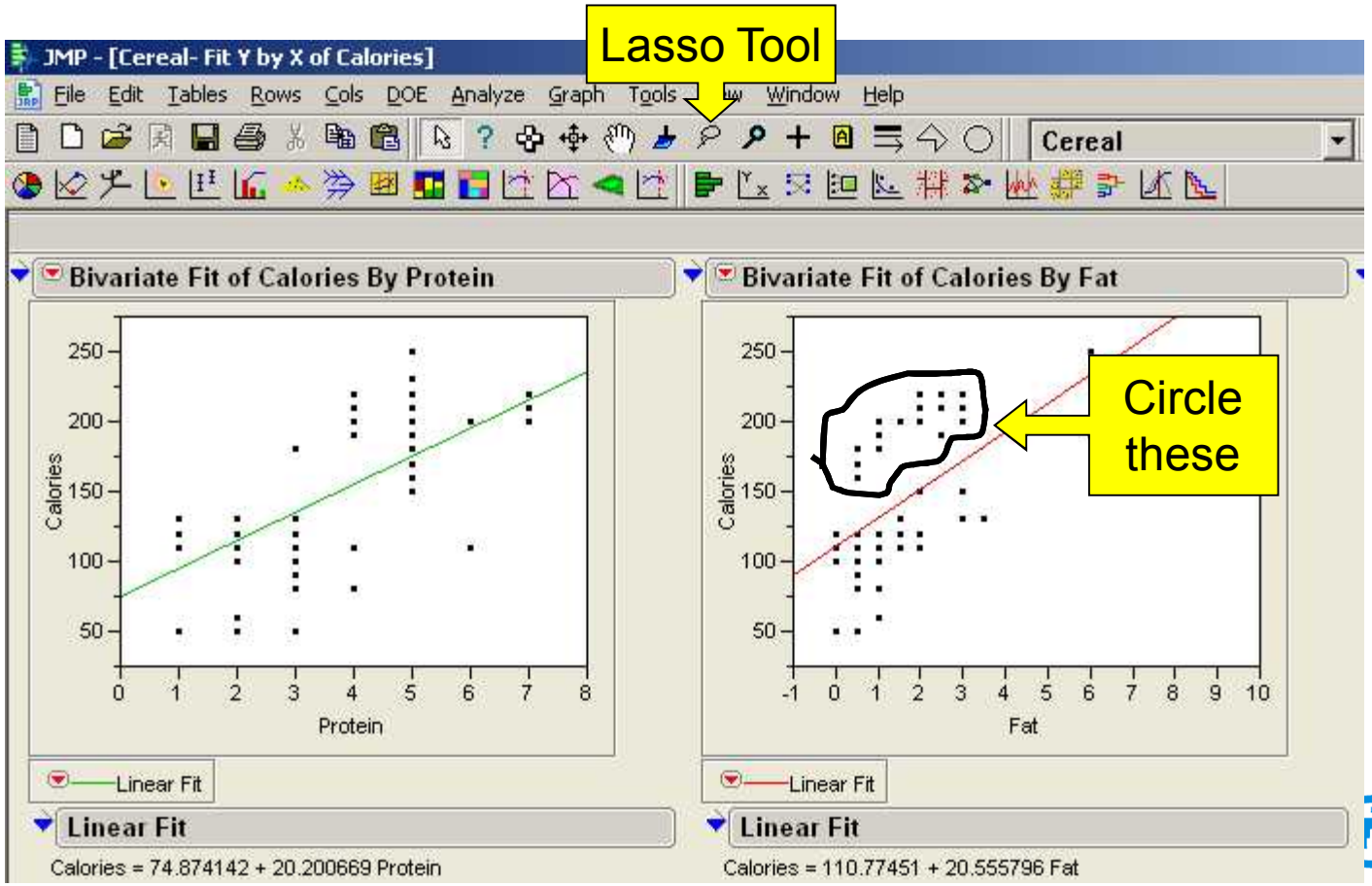
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	110.77451	5.985571	18.51	<.0001*
Fat	20.555796	2.82369	7.28	<.0001*

It appears that each g of fat adds 20.6 calories

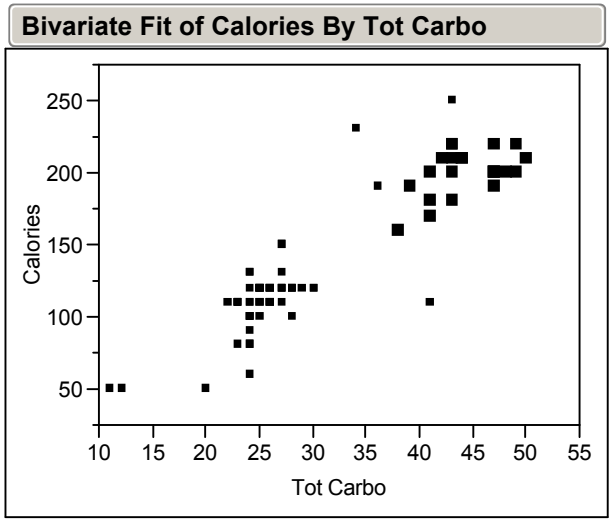
Fat Looks Bimodal

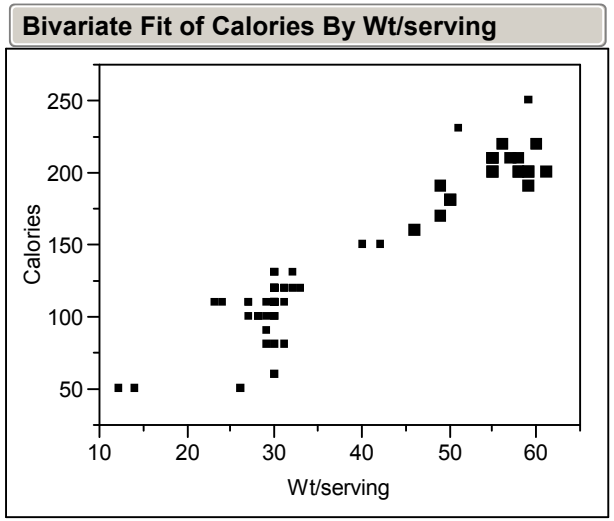
Let's investigate a little further.
Click the Lasso tool and circle the points shown.



Now Look at the Fit Y by X Plots

What do you observe?

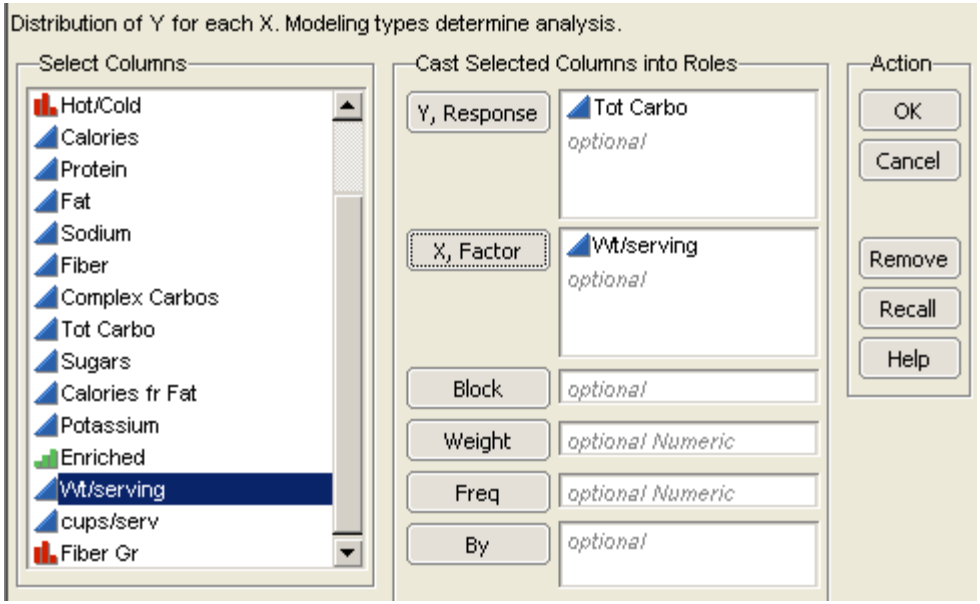


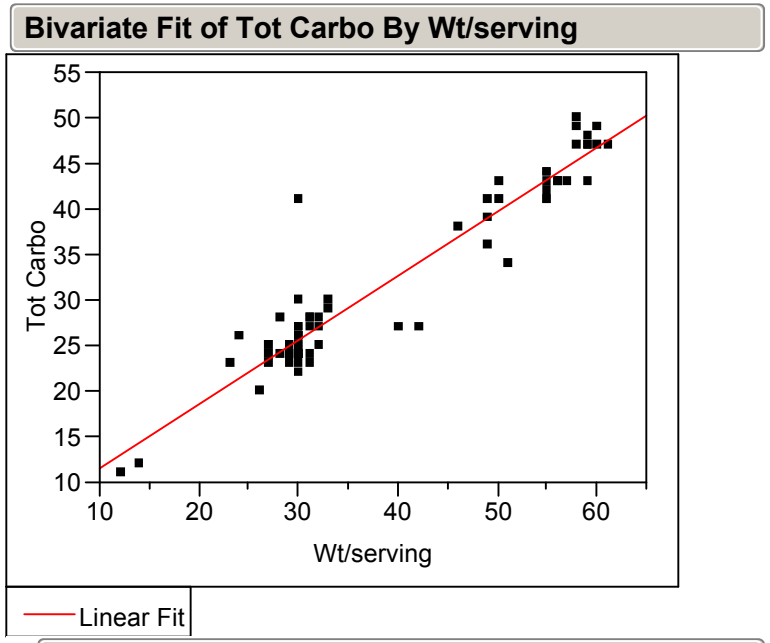


It appears these are the cereals with higher Total Carbs or higher Wt/serving.

Let's Take a Closer Look

Analyze → Fit Y by X





Linear Fit

Tot Carbo = 4.3400284 + 0.7073036 Wt/serving

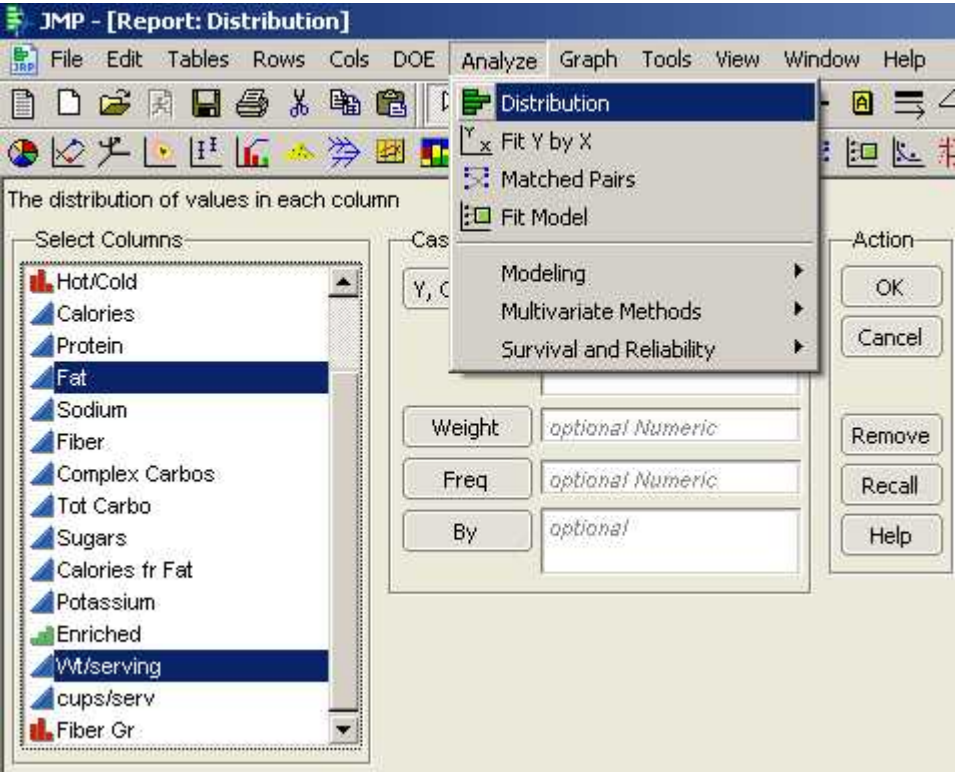
Summary of Fit

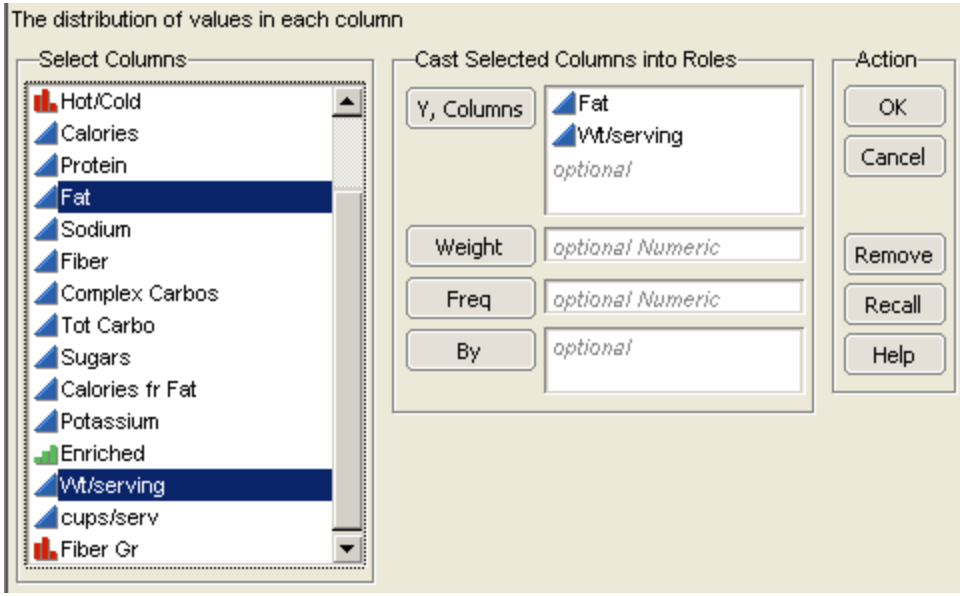
RSquare	0.909686
RSquare Adj	0.908449
Root Mean Square Error	2.937065
Mean of Response	31.45333
Observations (or Sum Wgts)	75

Total Carbs and Wt/serving are closely correlated.

Let's Look at the Distributions of Fat and Wt/serving

Analyze → Distribution
Click Fat, Hold [Ctrl] and click Wt/serving (you may need to scroll down)
Click [Y, Columns], then [OK]

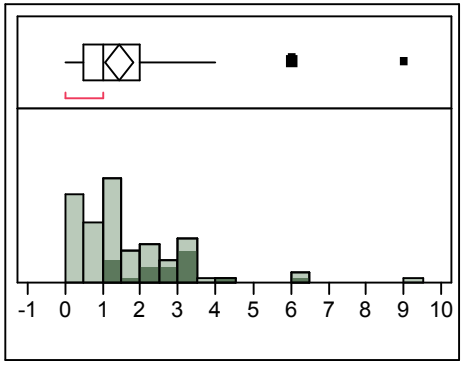




Distributions

Distributions

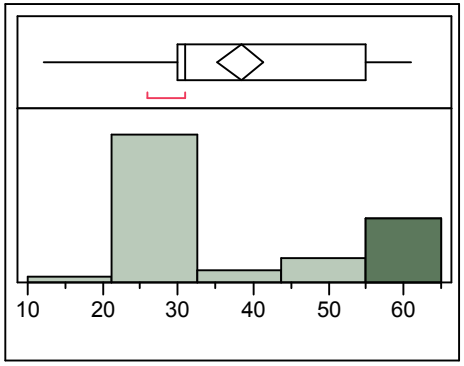
Fat



Quantiles		
100.0%	maximum	9.0000
99.5%		9.0000
97.5%		6.2250
90.0%		3.0000
75.0%	quartile	2.0000
50.0%	median	1.0000
25.0%	quartile	0.5000
10.0%		0.0000
2.5%		0.0000
0.5%		0.0000
0.0%	minimum	0.0000

Moments	
Mean	1.4473684
Std Dev	1.5590145
Std Err Mean	0.1788312
upper 95% Mean	1.8036185
lower 95% Mean	1.0911183
N	76

Wt/serving



Quantiles		
100.0%	maximum	61.000
99.5%		61.000
97.5%		60.100
90.0%		58.400
75.0%	quartile	55.000
50.0%	median	31.000
25.0%	quartile	30.000
10.0%		27.000
2.5%		13.800
0.5%		12.000
0.0%	minimum	12.000

Moments	
Mean	38.333333
Std Dev	13.089436
Std Err Mean	1.5114379
upper 95% Mean	41.344939
lower 95% Mean	35.321728
N	75

- Click the Grabber (hand) and grab the top of a bar and drag it downward.
- Click the Arrow and then click the right bar on the Wt/serving histogram. Look at the Fat histogram.

Exercise

Your doctor has told you that your triglycerides are high, and that you need to eat foods high in fiber and low in sugar. Which cereal should you choose?

Hint: Start by looking at the distributions.

(It's probably not Cocoa Puffs.)

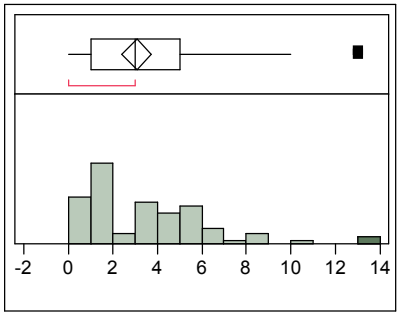
5 minutes

Exercise – One Possible Solution

- Click Analyze → Distribution
  - Choose Fiber and Sugars.
  - Click on bars until you find a good choice. The obvious choice is the highest fiber bar. Those happen to appear in the low sugar bar.

Distributions

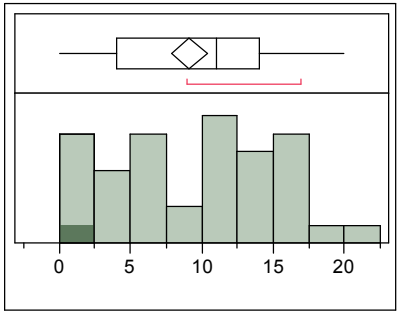
Fiber



Quantiles		
100.0%	maximum	13.000
99.5%		13.000
97.5%		13.000
90.0%		6.300
75.0%	quartile	5.000
50.0%	median	3.000
25.0%	quartile	1.000
10.0%		0.000
2.5%		0.000
0.5%		0.000
0.0%	minimum	0.000

Moments	
Mean	3.0657895
Std Dev	2.8720981
Std Err Mean	0.3294522
upper 95% Mean	3.722092
lower 95% Mean	2.4094869
N	76

Sugars



Quantiles		
100.0%	maximum	20.000
99.5%		20.000
97.5%		20.000
90.0%		16.000
75.0%	quartile	14.000
50.0%	median	11.000
25.0%	quartile	4.000
10.0%		0.700
2.5%		0.000
0.5%		0.000
0.0%	minimum	0.000

Moments	
Mean	9.1447368
Std Dev	5.7543119
Std Err Mean	0.6600648
upper 95% Mean	10.459653
lower 95% Mean	7.8298202
N	76

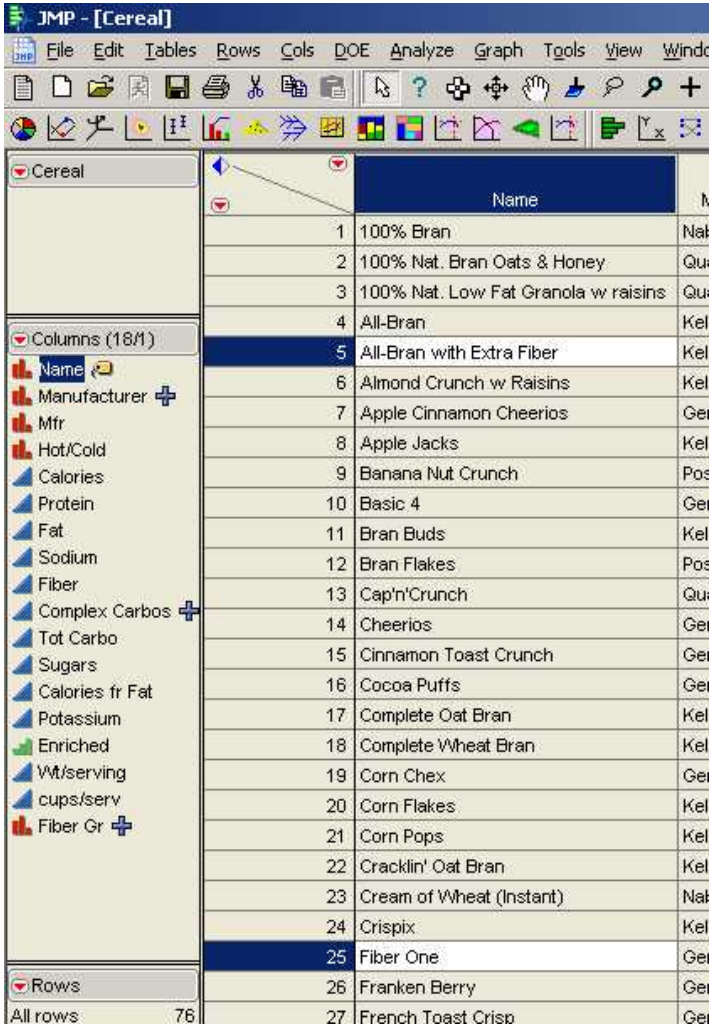
Exercise – One Possible Solution

Go back to the data (Window → Cereal) to see that you have two choices:

- All-Bran with Extra Fiber
- Fiber One

Both have 13g fiber and 0g sugar

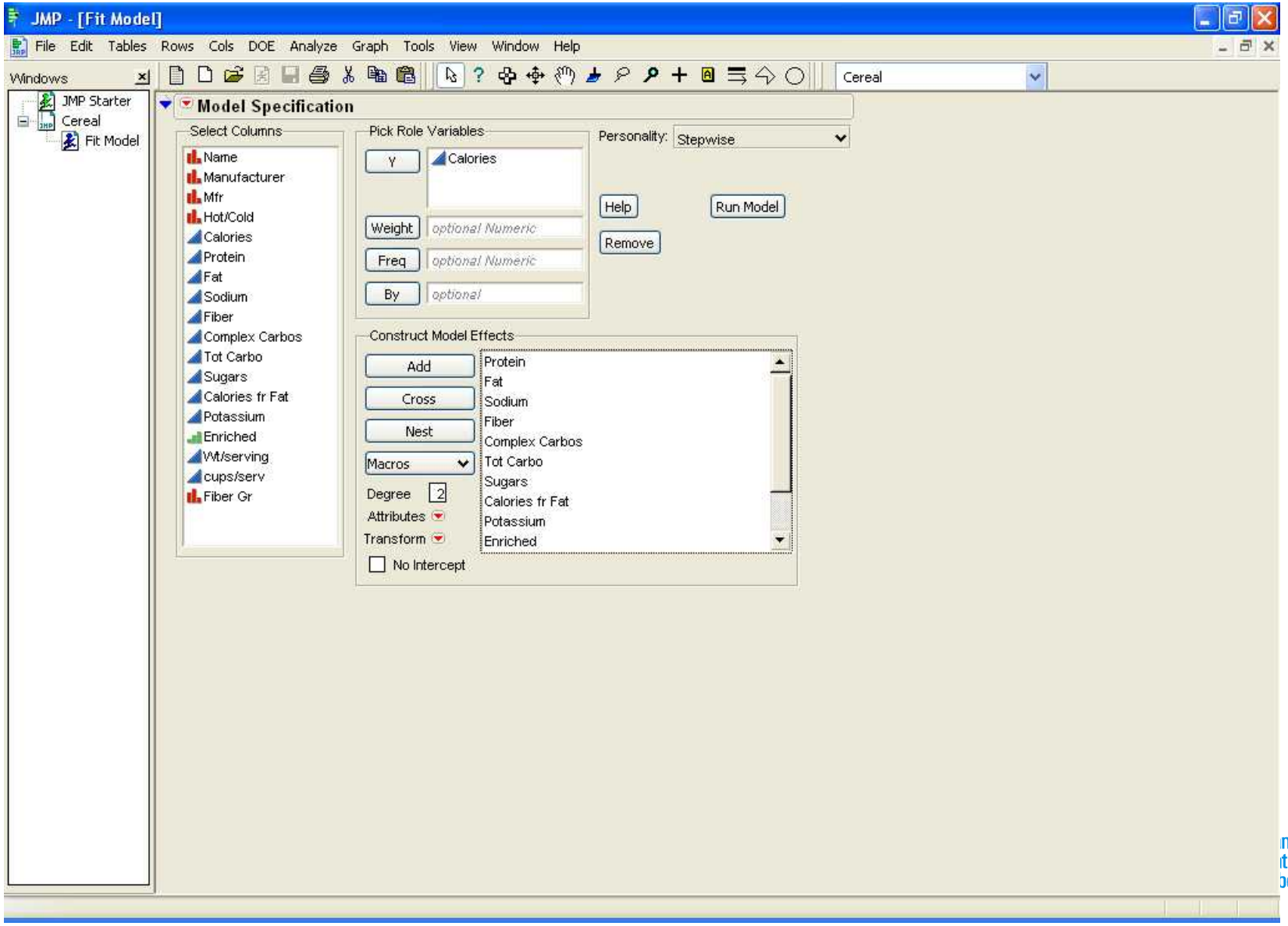
Third option:
Eat the box. 48g fiber, 0g sugar



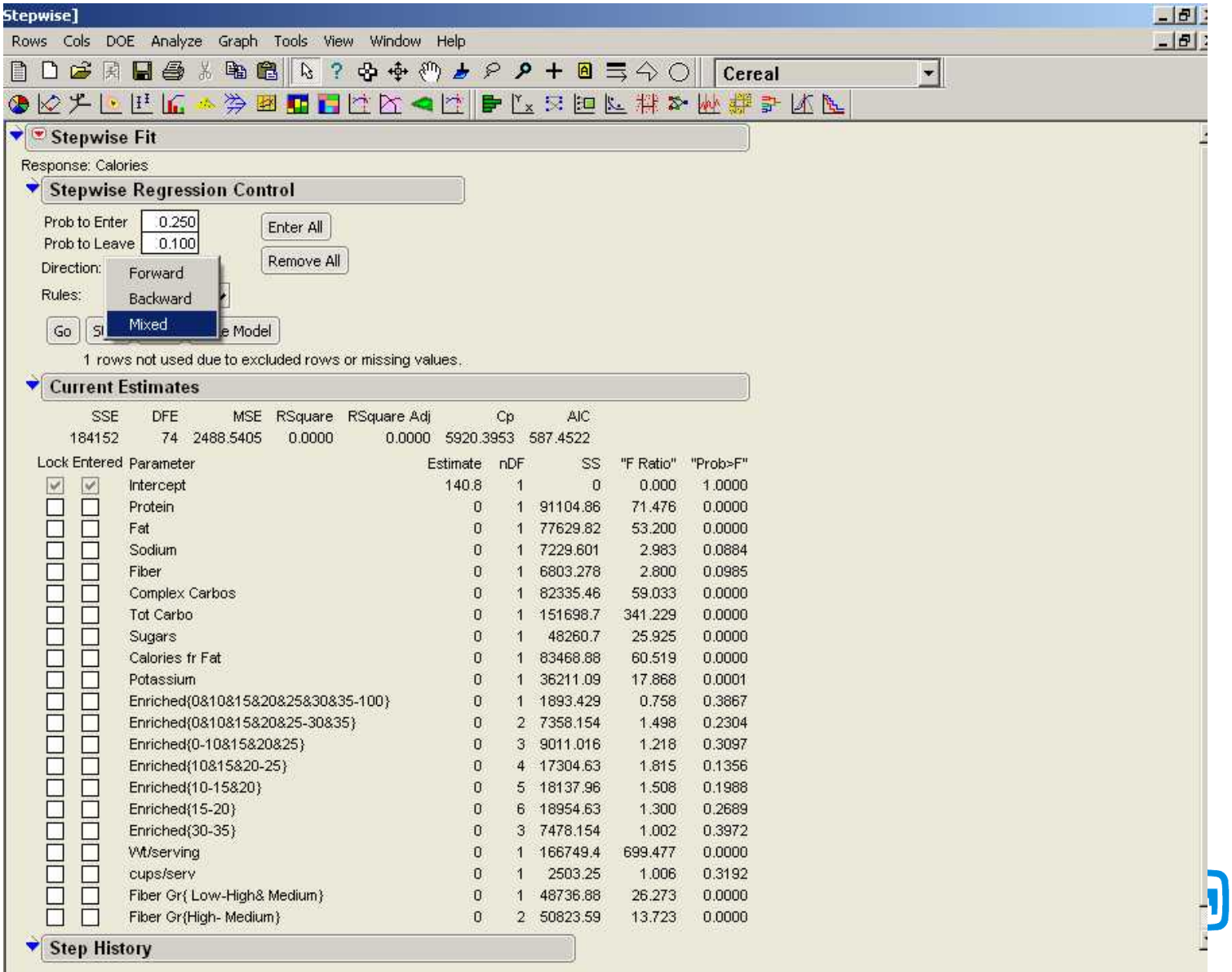
Which Would Be Your Worst Choice?

- Clear Row States
- Go back to distributions
- Select the lowest Fiber bar in the histogram
- Hold [Ctrl] and click the sugar bars less than 15.
- Go back to the Data and see that your worst choice is …
- Golden Crisp (0g fiber, 15g sugar)

Analyze → Fit Model
Calories in Y, all below that in X



Select Personality → Stepwise, Run Model
Change Direction to Mixed



Stepwise Regression Control Panel

The Stepwise Regression Control Panel (Control Panel for short) has editable areas, buttons and a popup menu. You use these dialog features to limit regressor effect probabilities, determine the method of selecting effects, begin or stop the selection process, and create a model.

You use the Control Panel as follows:

Prob to Enter

is the significance probability that must be attributed to a regressor term for it to be considered as a forward step and entered into the model. Click the field to enter a value.

Prob to Leave

is the significance probability that must be attributed to a regressor term in order for it to be considered as a backward step and removed from the model. Click the field to enter a value.

Direction

accesses the popup menu shown here, which lets you choose how you want variables to enter the regression equation.

Stepwise Regression Control Panel

Forward brings in the regressor that most improves the fit, given that term is significant at the level specified by **Prob to Enter**.

Backward removes the regressor that affects the fit the least, given that term is not significant at the level specified in **Prob to Leave**.

Mixed alternates the forward and backward steps. It includes the most significant term that satisfies **Prob to Enter** and removes the least significant term satisfying **Prob to Leave**. It continues removing terms until the remaining terms are significant and then it changes to the forward direction.

Stepwise Regression Control Panel

Buttons on the controls panel let you control the stepwise processing:

Go
starts the selection process. The process continues to run in the background until the model is finished.

Stop
stops the background selection process.

Step
stops after each step of the stepwise process

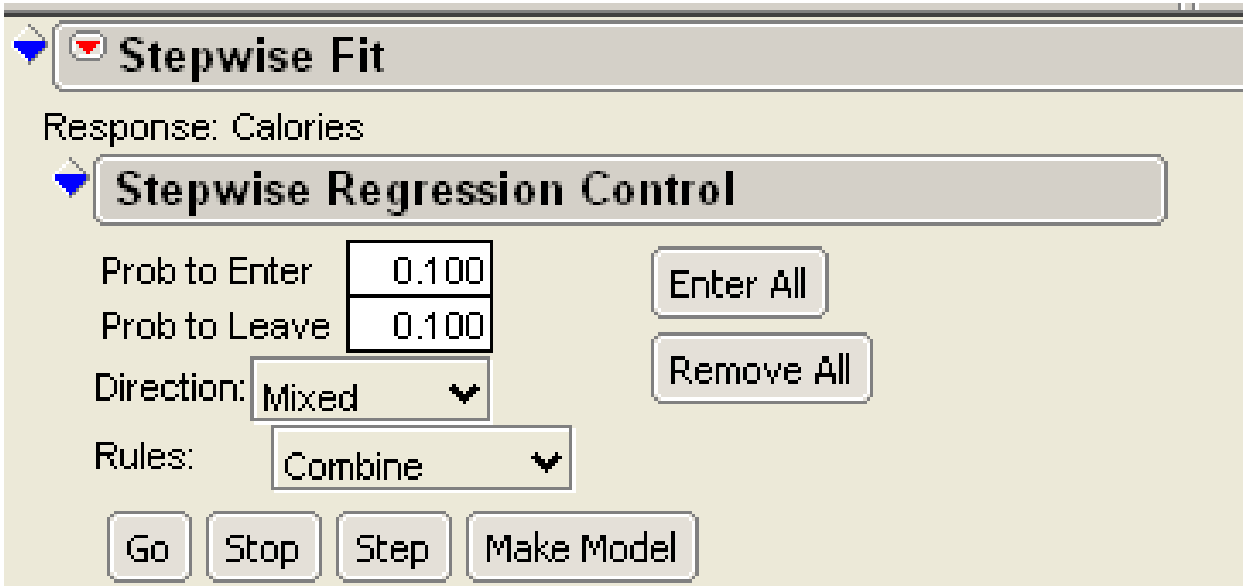
Enter All
enters all unlocked terms into the model.

Remove All
removes all terms from the model.

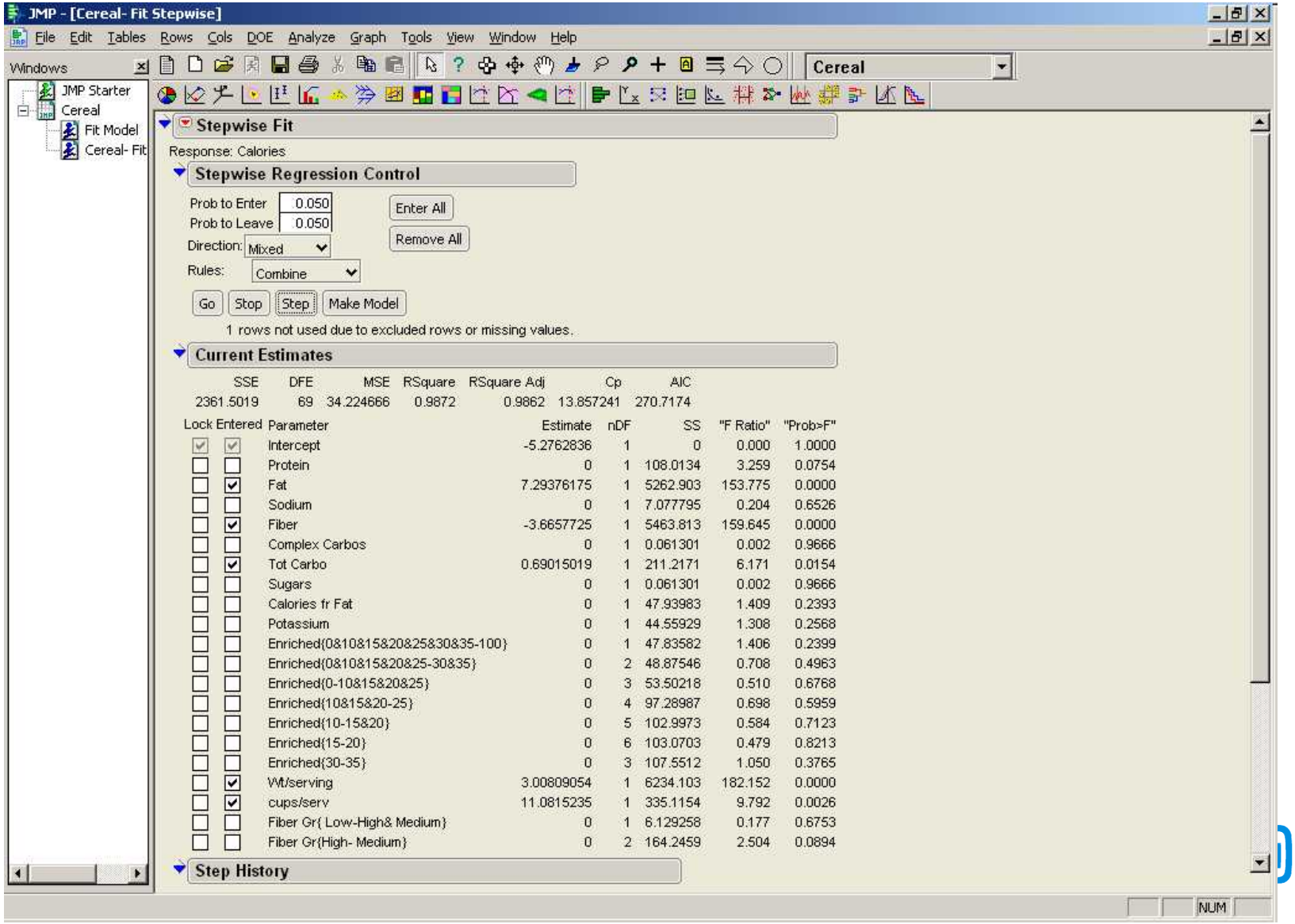
Make Model
forms a model for the Model Specification Dialog from the model currently showing in the Current Estimates table. In cases where there are nominal or ordinal terms, Make Model can create new data table columns to contain terms that are needed for the model.

A Few Changes First

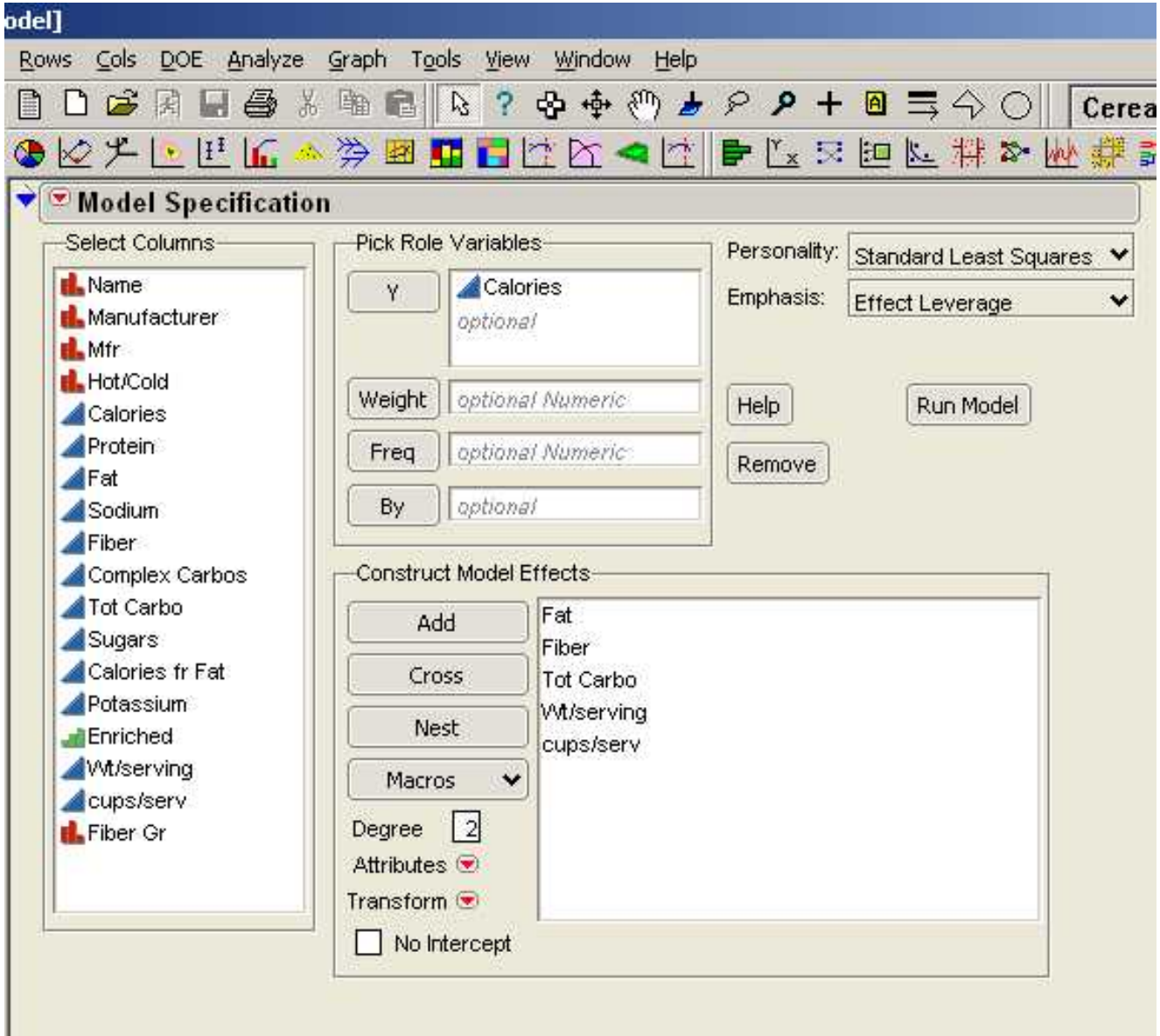
- You already changed direction to Mixed.
- Change “Prob to Enter” and “Prob to Leave” to 0.100



Click Step and Watch Factors Get Added to the Model



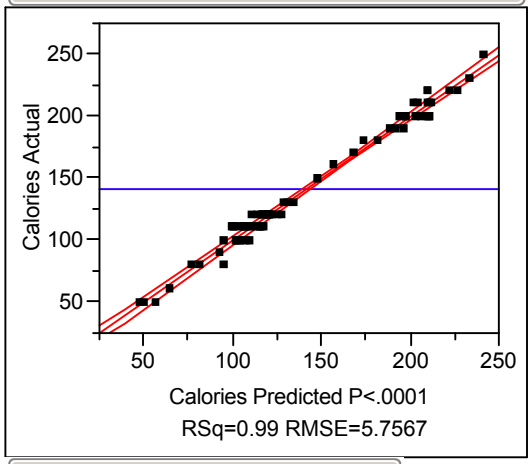
Click Make Model



Click Run Model

Whole Model

Actual by Predicted Plot



Summary of Fit

RSquare	0.987763
RSquare Adj	0.986683
Root Mean Square Error	5.756695
Mean of Response	140.8
Observations (or Sum Wgts)	75

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	181898.51	30316.4	914.8112
Error	68	2253.49	33.1	Prob > F
C. Total	74	184152.00		<.0001*

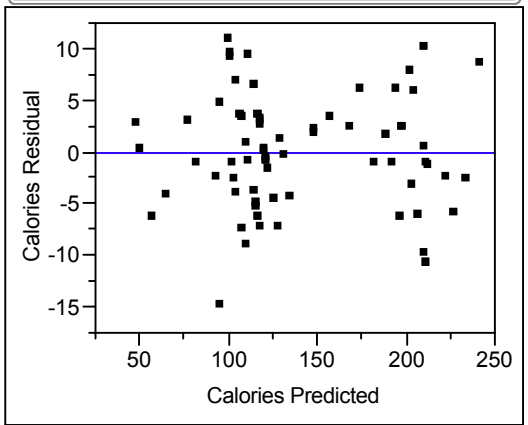
Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	65	2253.4886	34.6691	.
Pure Error	3	0.0000	0.0000	Prob > F
Total Error	68	2253.4886		.
				Max RSq
				1.0000

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-4.230658	3.478264	-1.22	0.2281
Protein	1.3019557	0.721158	1.81	0.0754
Fat	7.414865	0.582652	12.73	<.0001*
Fiber	-3.874463	0.308004	-12.58	<.0001*
Tot Carbo	0.8724868	0.291431	2.99	0.0038*
Wt/serving	2.755463	0.260157	10.59	<.0001*
cups/serv	10.097625	3.527139	2.86	0.0056*

Residual by Predicted Plot



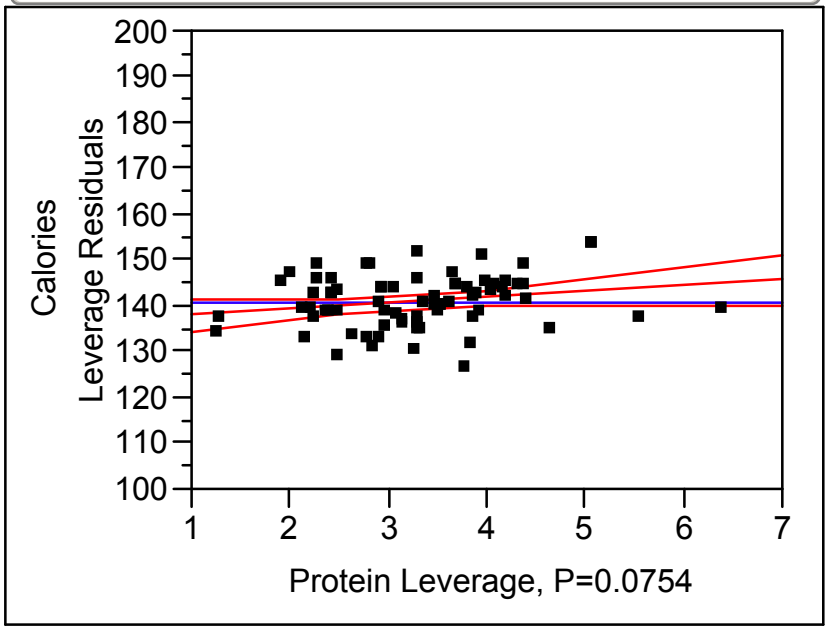
What is the Effect of Protein Now?

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-4.230658	3.478264	-1.22	0.2281
Protein	1.3019557	0.721158	1.81	0.0754
Fat	7.414865	0.582652	12.73	<.0001*
Fiber	-3.874463	0.308004	-12.58	<.0001*
Tot Carbo	0.8724868	0.291431	2.99	0.0038*
Wt/serving	2.755463	0.260157	10.59	<.0001*
cups/serv	10.097625	3.527139	2.86	0.0056*

Protein

Leverage Plot



Recall that it was 20.2 calories per g when we just looked at calories vs. protein. Nutritionists tell us that the real number is 4 calories per g of fat.

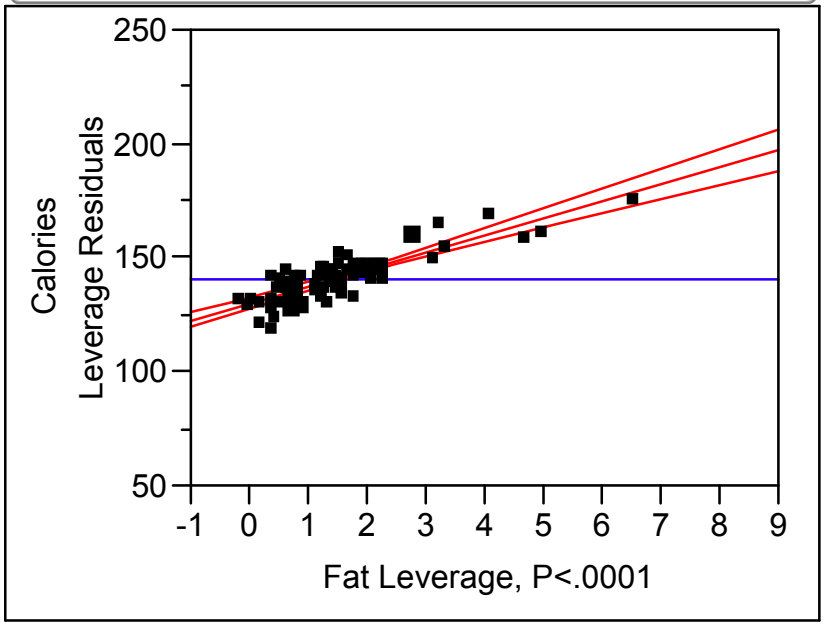
What is the Effect of Fat Now?

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-4.230658	3.478264	-1.22	0.2281
Protein	1.3019557	0.721158	1.81	0.0754
Fat	7.414865	0.582652	12.73	<.0001*
Fiber	-3.874463	0.308004	-12.58	<.0001*
Tot Carbo	0.8724868	0.291431	2.99	0.0038*
Wt/serving	2.755463	0.260157	10.59	<.0001*
cups/serv	10.097625	3.527139	2.86	0.0056*

Fat

Leverage Plot

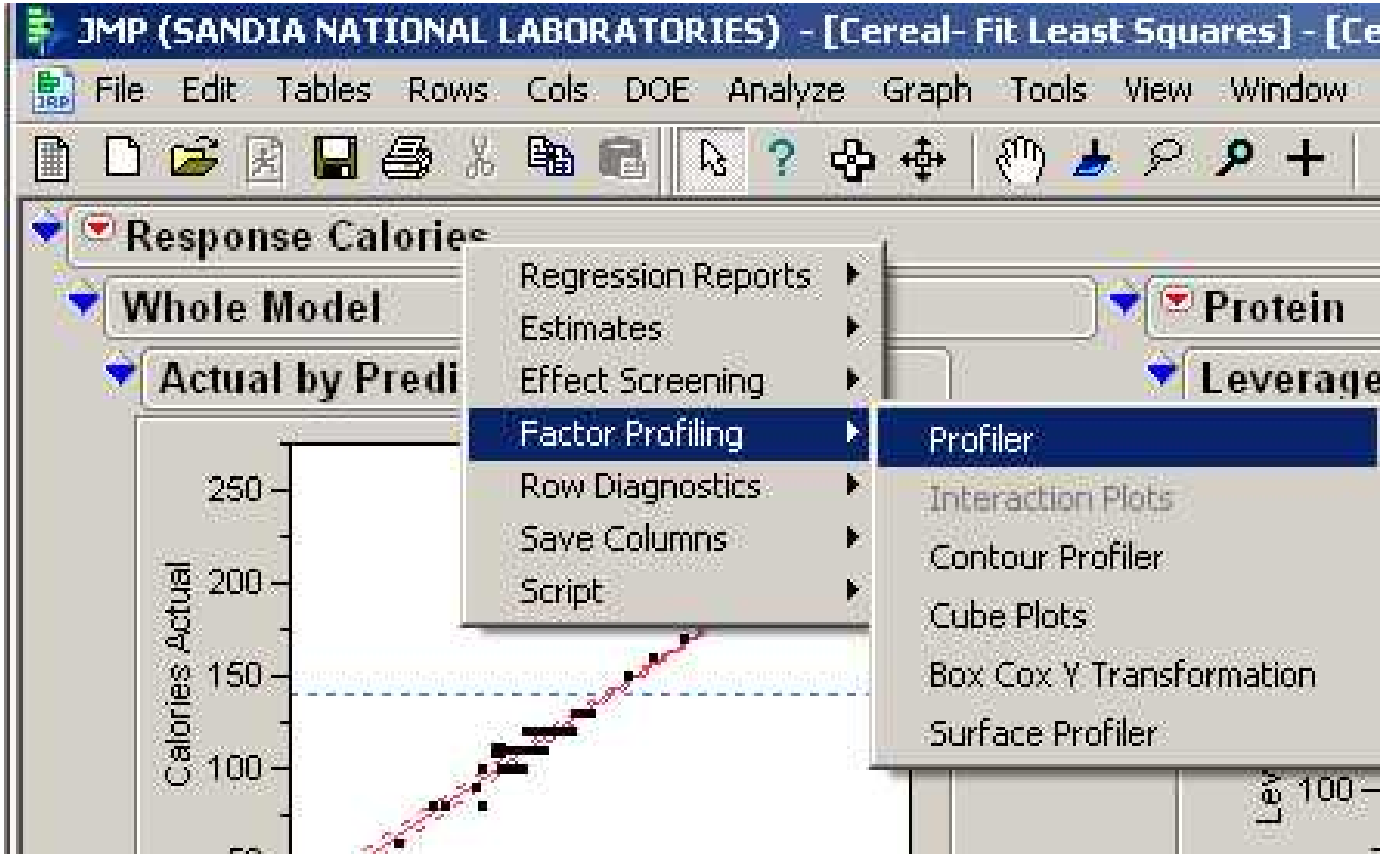


Recall that it was 20.6 calories per g when we just looked at calories vs. fat. Nutritionists tell us that the real number is 9 calories per g of fat.

You Need to Look at the Model!

- When we just looked at calories vs. protein, we concluded that each gram of protein adds 20.2 calories.
- When we looked at the entire model, we discovered that each gram of protein really only adds 1.3 calories!
- When we just looked at calories vs. protein, we concluded that each gram of fat adds 20.6 calories.
- When we looked at the entire model, we discovered that each gram of fat really only adds 7.4 calories!
- Looking at the entire model captured all contributing factors, and gave us coefficients closer to what we've been told by nutritionists.

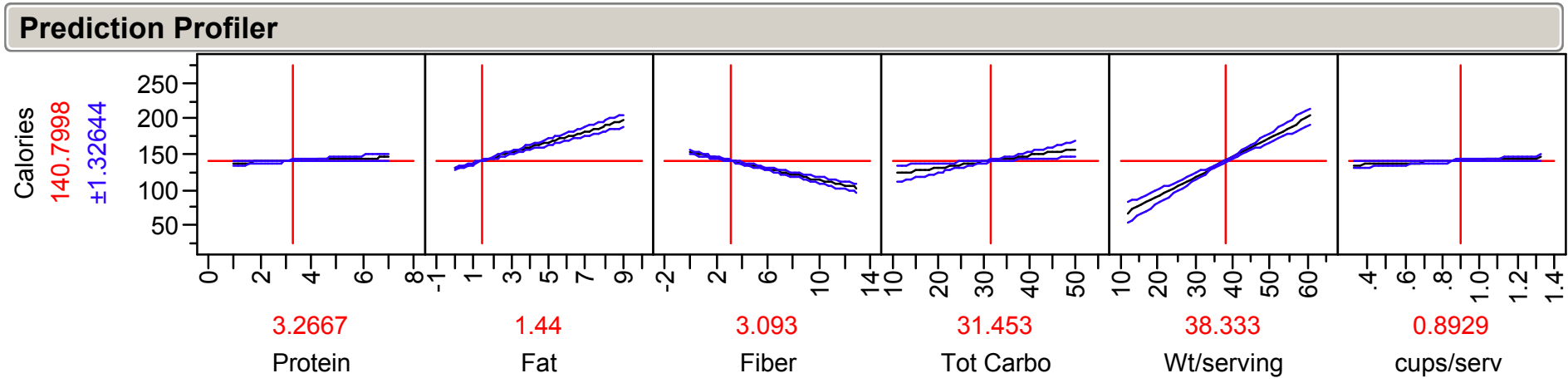
Right-click Response Calories title bar
Select Factor Profiling → Profiler



Scroll Down to Prediction Profiler

Try moving the vertical lines.

Which factors could you change in order to reduce calories?

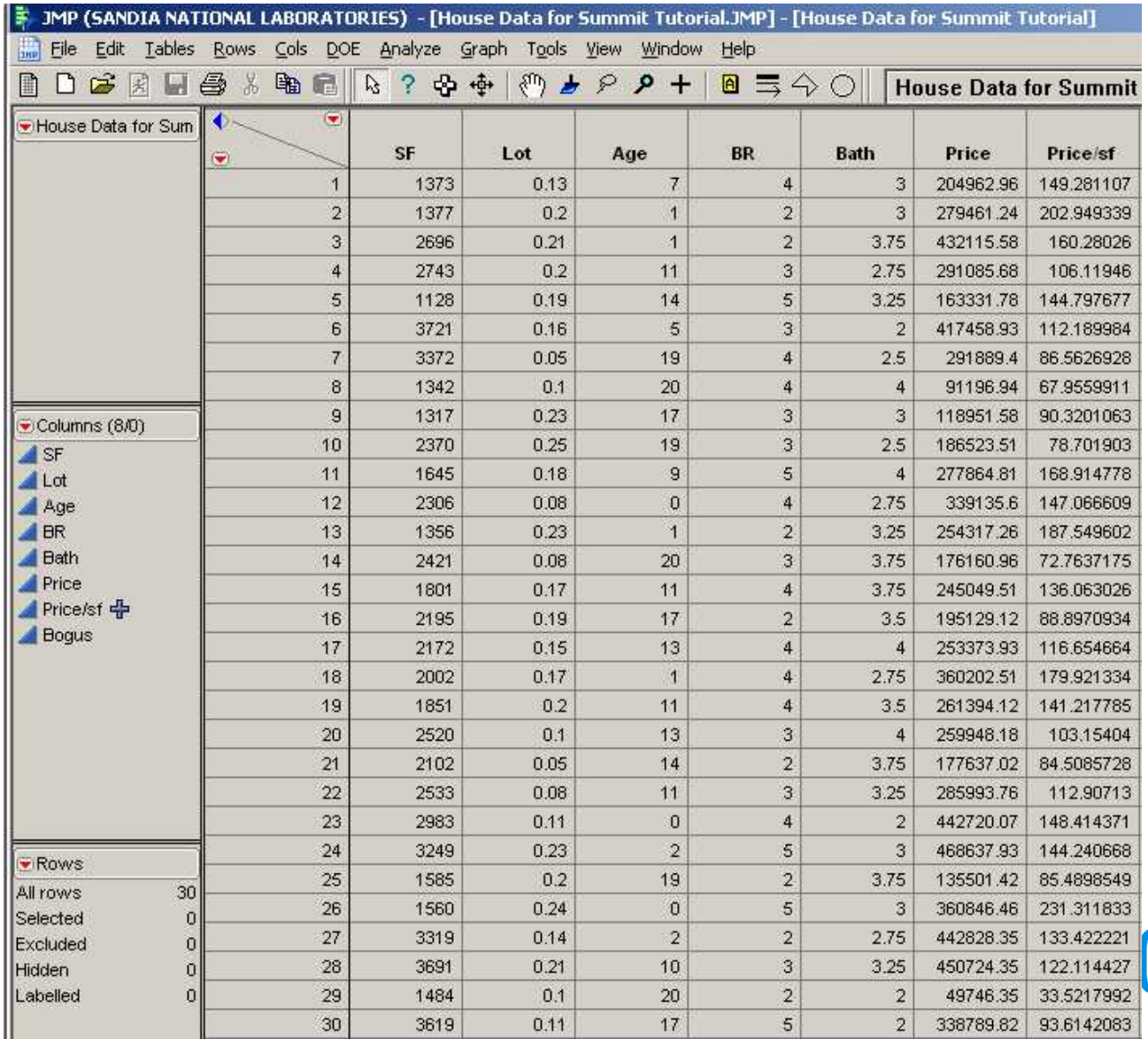


Back to the Exercise

You want to sell your house. It has the following features:

- 2000 square feet
- 0.2 acre lot
- 2 years old
- 3 bedrooms
- 3 full bathrooms

Load *House Data for Summit Tutorial.jmp* Data File



	SF	Lot	Age	BR	Bath	Price	Price/sf
1	1373	0.13	7	4	3	204962.96	149.281107
2	1377	0.2	1	2	3	279461.24	202.949339
3	2696	0.21	1	2	3.75	432115.58	160.28026
4	2743	0.2	11	3	2.75	291085.68	106.11946
5	1128	0.19	14	5	3.25	163331.78	144.797677
6	3721	0.16	5	3	2	417458.93	112.189984
7	3372	0.05	19	4	2.5	291889.4	86.5626928
8	1342	0.1	20	4	4	91196.94	67.9559911
9	1317	0.23	17	3	3	118951.58	90.3201063
10	2370	0.25	19	3	2.5	186523.51	78.701903
11	1645	0.18	9	5	4	277864.81	168.914778
12	2306	0.08	0	4	2.75	339135.6	147.066609
13	1356	0.23	1	2	3.25	254317.26	187.549602
14	2421	0.08	20	3	3.75	176160.96	72.7637175
15	1801	0.17	11	4	3.75	245049.51	136.063026
16	2195	0.19	17	2	3.5	195129.12	88.8970934
17	2172	0.15	13	4	4	253373.93	116.654664
18	2002	0.17	1	4	2.75	360202.51	179.921334
19	1851	0.2	11	4	3.5	261394.12	141.217785
20	2520	0.1	13	3	4	259948.18	103.15404
21	2102	0.05	14	2	3.75	177637.02	84.5085728
22	2533	0.08	11	3	3.25	285993.76	112.90713
23	2983	0.11	0	4	2	442720.07	148.414371
24	3249	0.23	2	5	3	468637.93	144.240668
25	1585	0.2	19	2	3.75	135501.42	85.4898549
26	1560	0.24	0	5	3	360846.46	231.311833
27	3319	0.14	2	2	2.75	442828.35	133.422221
28	3691	0.21	10	3	3.25	450724.35	122.114427
29	1484	0.1	20	2	2	49746.35	33.5217992
30	3619	0.11	17	5	2	338789.82	93.6142083

Sandia National Laboratories

Exercise:
What Will Your Listing Price Be?

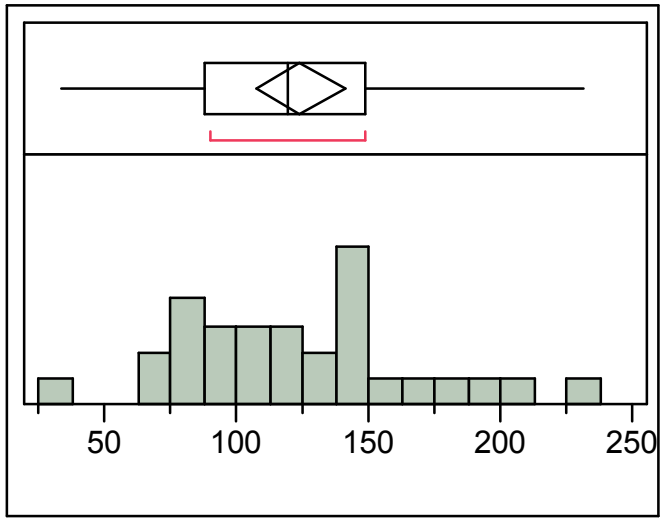
You Analyzed the distribution of Price/sf

Average = $124.36 per SF

Therefore, $124.36 / sf x 2,000 sf = $248,720

Distributions

Price/sf



Quantiles

100.0%	maximum	231.31
99.5%		231.31
97.5%		231.31
90.0%		186.79
75.0%	quartile	148.63
50.0%	median	119.38
25.0%	quartile	88.31
10.0%		73.36
2.5%		33.52
0.5%		33.52
0.0%	minimum	33.52

Moments

Mean	124.36354
Std Dev	44.020787
Std Err Mean	8.0370594
upper 95% Mean	140.80117
lower 95% Mean	107.92591
N	30

Exercise:
What Will Your Listing Price Be?

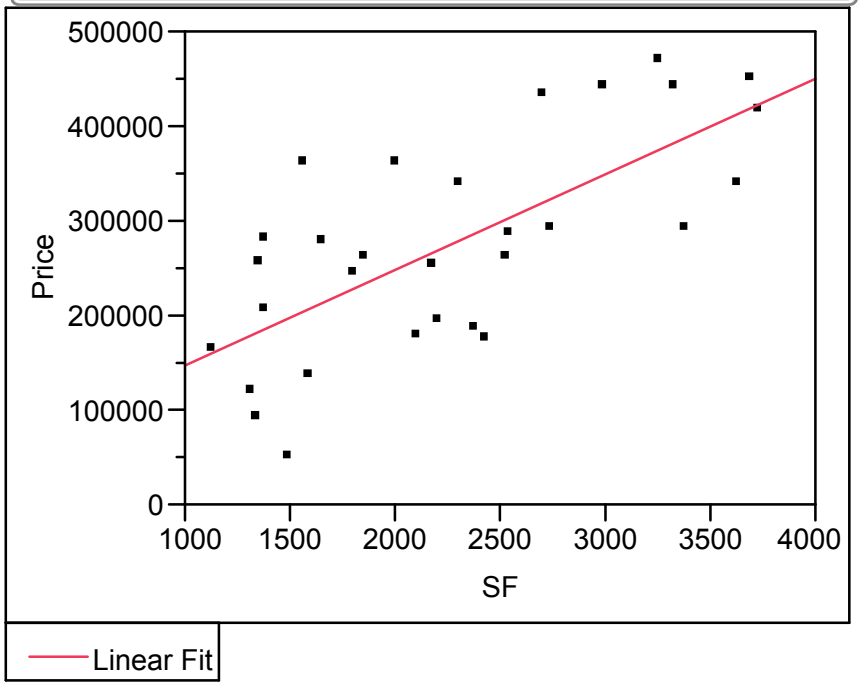
You performed a Fit Y by X for Price vs. SF

You also added a Line Fit

Price = $45,962 + $101.34*SF

Therefore, Price = $248,642

Bivariate Fit of Price By SF



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1.8791e+11	1.879e+11	28.1954
Error	28	1.8661e+11	6.6647e+9	**Prob > F**
C. Total	29	3.7452e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	45962.927	45654	1.01	0.3227
SF	101.33845	19.0847	5.31	<.0001*

Exercise:
What Will Your Listing Price Be?

- Review Student's Listing Prices previously captured on the Board

You Want to Sell Your House

- Your realtor pulls up the set of data for recent home sales in your zip code, and tells you the average selling price was $124.36 per square foot. (Data file provided.)
- Your realtor breaks out the calculator and tells you your home is worth $124.36/ft^2 x 2,000 ft^2 = $248,720.
- Your realtor tells you to list your house for $260,000. “That leaves a little room for negotiating,” they explain.
- You’re just about to sign the listing paperwork, but you remember the Workshop from the Black Belt Summit.

Should You Listen to Your Realtor? Exercise

- Create a model for home price, including only significant factors.
- Determine the value of your home based on the model.
- Capture the students' listing prices on the board.
- Are these much different than what your Realtor recommended?

Exercise Time

15 minutes

Solutions

Create a model for home price, including only significant factors.

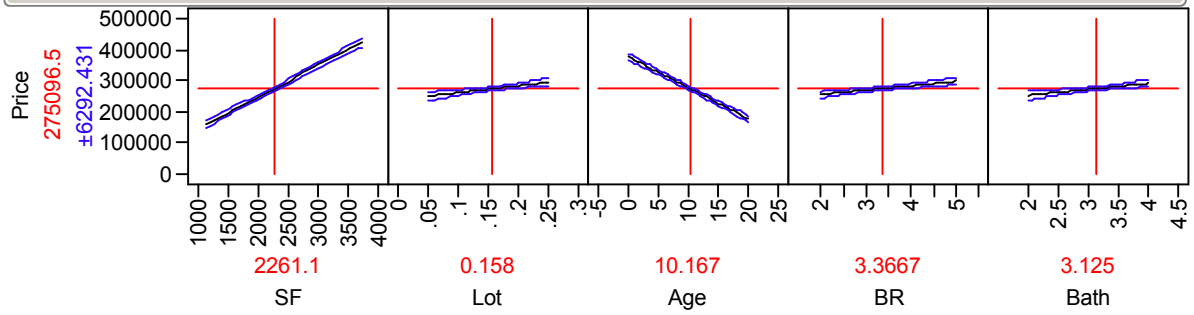
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3242.1849	28037.38	0.12	0.9089
SF	100.26837	4.377378	22.91	<.0001*
Lot	228519.2	55577.05	4.11	0.0004*
Age	-9954.605	456.565	-21.80	<.0001*
BR	14362.019	2925.965	4.91	<.0001*
Bath	19803.935	5364.676	3.69	0.0011*

Prediction Expression

+14362.0191583007*BR
+19803.9349350972*Bath

Prediction Profiler



Sandia National Laboratories

Solutions

Determine the value of your home based on the model.

$$3242.18 + 100.27 * (2000) + 228519.20 * (0.2) - 9954.60 * (2) + 14362.02 * (3) + 19803.93 * (3) =$$

$332,075

Should you listen to your Realtor and list your house for $260,000?

What Does the Model Tell You?

- Which factors are statistically significant?
- What are the coefficients for these factors?
- In particular, what is the coefficient for $/square foot?

A Word of Caution

- Three types of variables
  - Continuous
    - Time
    - Distance
  - Ordinal
    - Character data with an order (poor, fair, good, better, best)
    - Numerical data with unequal spacing (4 = strongly agree, 3 = agree, 2 = disagree, 1 = strongly disagree)
  - Nominal
    - Character data with no specific order (green, blue, yellow)
    - Numerical data with no specific order (NASCAR car #)
- Should BR and Bath be treated as continuous variables?
- What if we had treated them as Ordinal Variables?

Treating BR and Bath as Ordinal

- If Time Permits, change BR and Bath to Ordinal and redo the analysis

House Data for Sum

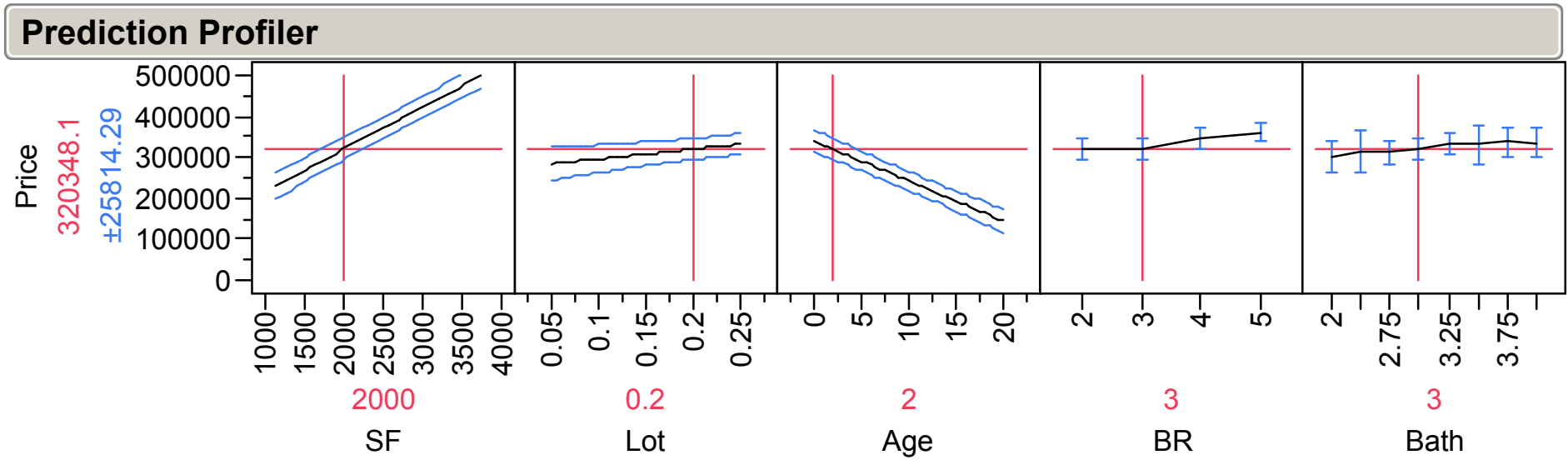
	SF	Lot	Age	BR
1	1373	0.13	7	4
2	1377	0.2	1	2
3	2696	0.21	1	2
4	2743	0.2	11	3
5	1128	0.19	14	5
6	3721	0.16	5	3
7	3372	0.05	19	4
8	1342	0.1	20	4
9	1317	0.23	17	3
10	2370	0.25	19	3
11	1645	0.18	9	5
12	2306	0.08	0	4
13	1356	0.23	1	2
14	2421	0.08	20	3
15	1801	0.17	11	4
16	2195	0.19	17	2

Columns (7/1)
- SF
- Lot
- Age
- BR

✔ Continuous
Ordinal
Nominal

Columns (7/0)
- SF
- Lot
- Age
- BR
- Bath
- Price
- Price/sf

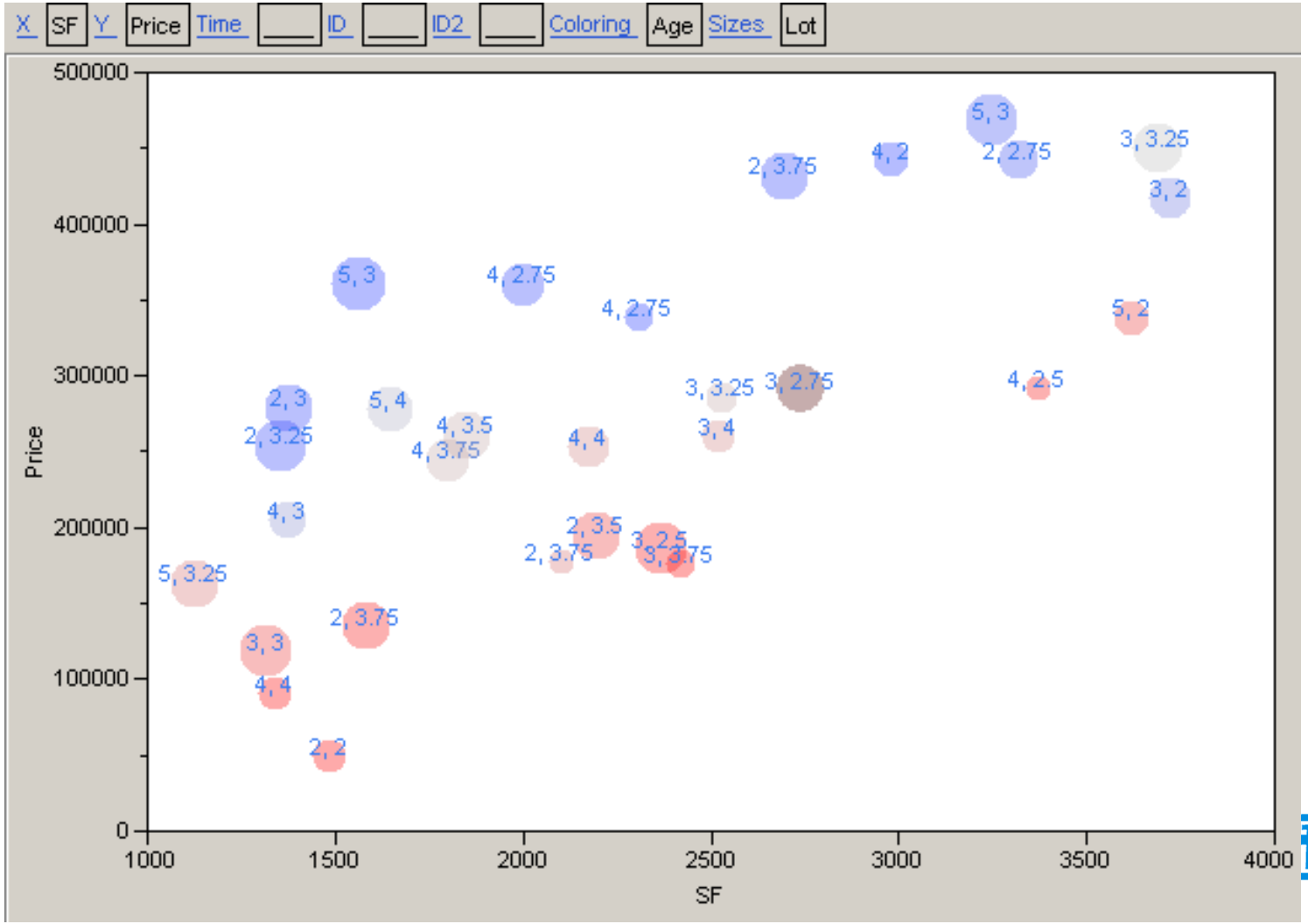
Use Prediction Profiler



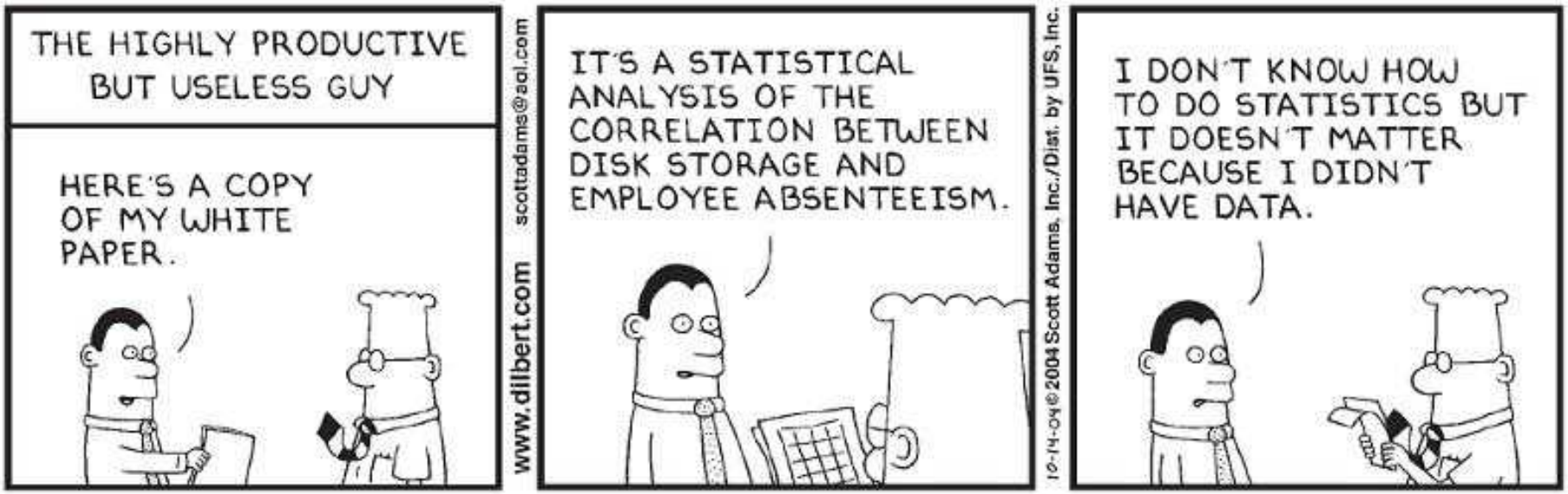
What is the predicted price now?

One Last Tip: Visually Display your Data!

How many dimensions are shown in this single graph?



Questions?


Dilbert
THE HIGHLY PRODUCTIVE BUT USELESS GUY
HERE'S A COPY OF MY WHITE PAPER.
www.dilbert.com scottadams@aol.com
IT'S A STATISTICAL ANALYSIS OF THE CORRELATION BETWEEN DISK STORAGE AND EMPLOYEE ABSENTEEISM.
10-14-04 © 2004 Scott Adams, Inc./Dist. by UFS, Inc.
I DON'T KNOW HOW TO DO STATISTICS BUT IT DOESN'T MATTER BECAUSE I DIDN'T HAVE DATA.

"… all models are wrong; the practical question is how wrong do they have to be to not be useful …"

George Box and Norman Draper, Empirical Model Building and Response Surfaces, John Wiley, 1987, pg. 74