Automated Cost Estimating Integrated Tools

Pooled Regression and Dummy Variables in CO$TAT

Jeff McDowell

September 19, 2012

PRT 141

Outline

- Define Dummy Variable
- CER Example Using Linear Regression
- The Pattern
- CER Example Using Log-Linear Regression
- Define Pooled Regression
- Learning Curve Example
- Tips and Tricks

Define Dummy Variable

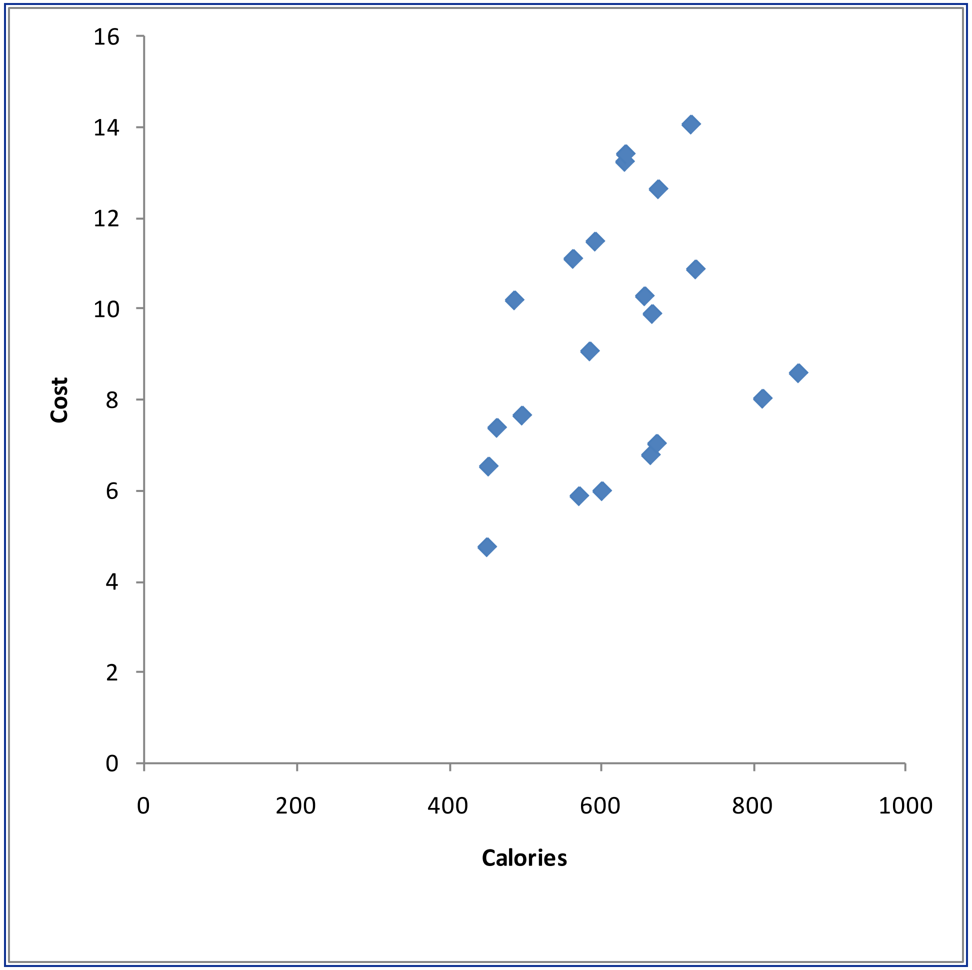
- **A dummy variable is a numeric proxy for a qualitative differentiating characteristic.**
  - Used to stratify data into subsets of distinct classes.
  - Classes are represented in regression using 'dummy' variables.

- **In contrast to other variables that can take values over some continuous range**

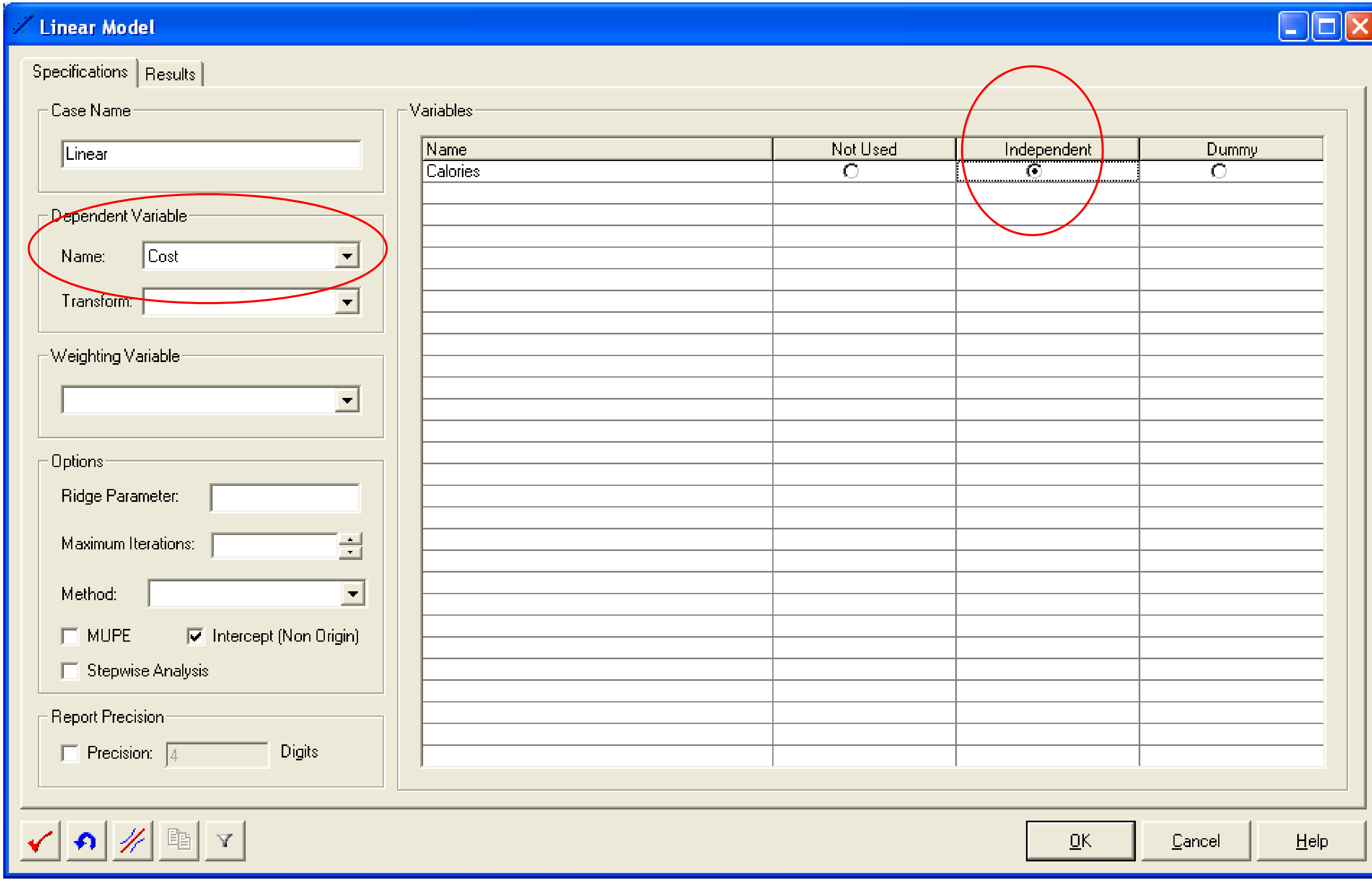
CER Example Using Linear Regression

Sample Dataset #1

Observations	Cost of Meal	Calories
Variable ID	Cost	Calories
Meal Number 1	6.78	665
Meal Number 2	4.76	450
Meal Number 3	8.02	812
Meal Number 4	5.99	601
Meal Number 5	8.58	859
Meal Number 6	7.03	673
Meal Number 7	5.88	571
Meal Number 8	10.27	657
Meal Number 9	10.86	724
Meal Number 10	9.06	585
Meal Number 11	7.38	463
Meal Number 12	6.53	452
Meal Number 13	9.88	667
Meal Number 14	7.65	496
Meal Number 15	11.09	563
Meal Number 16	13.22	631
Meal Number 17	10.18	486
Meal Number 18	11.47	592
Meal Number 19	13.39	632
Meal Number 20	14.04	718
Meal Number 21	12.62	675



CO$TAT Linear Model



Simple Example

Linear Analysis for Dataset Meals Dataset, Simple

Wednesday, 16 May 2012, 1:29 pm

I. Model Form and Equation Table

Model Form:	Unweighted Linear model
Number of Observations Used:	21
Equation in Unit Space:	Cost = 4.279 + 0.008081 * Calories

OK

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	4.2789	3.3219		1.2881	0.2131	0.7869
Calories	0.0081	0.0053	0.3304	1.5258	0.1434	0.8566

Weak

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
2.6434	10.92%	6.23%	0.3304

Weak

Analysis of Variance

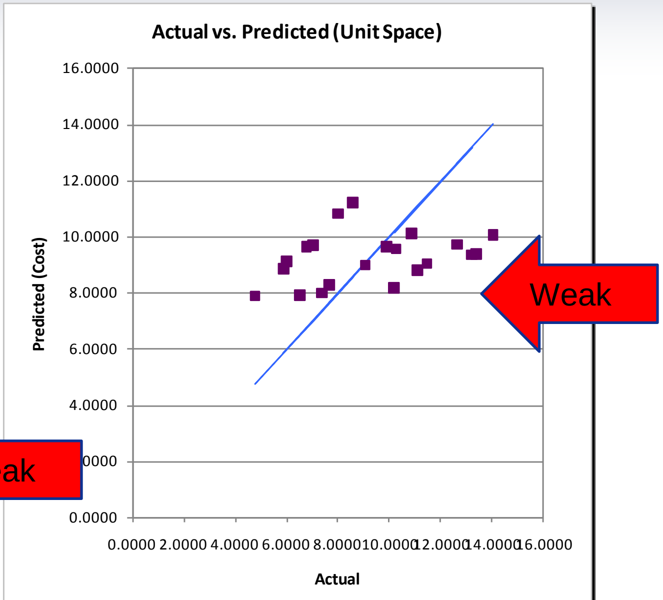
Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	1	16.2674	16.2674	2.3281	0.1434	0.8566
Residual (Error)	19	132.7607	6.9874			
Total	20	149.0281				

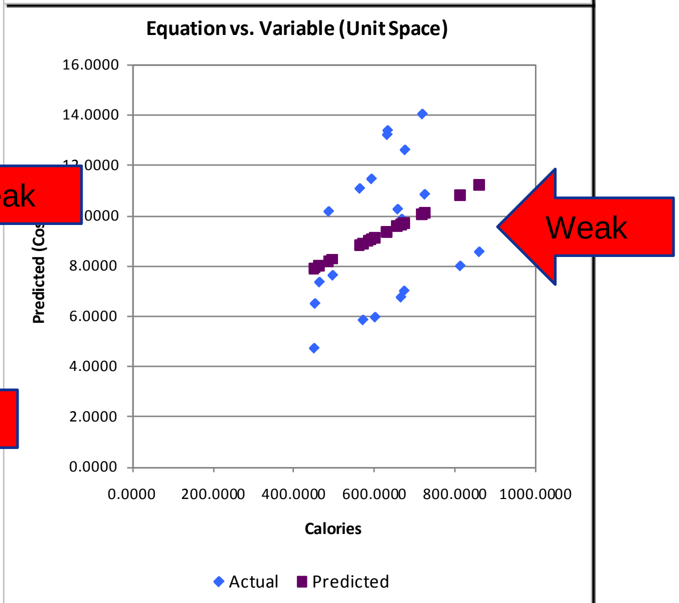
Weak

Summary of Predictive Measures

Average Actual (Avg Act)	9.2705
Standard Error (SE)	2.6434
Root Mean Square (RMS) of % Errors	31.02%
Mean Absolute Deviation (Mad) of % Errors	25.81%
Coef of Variation based on Std Error (SE/Avg Act)	28.51%
Coef of Variation based on MAD Res (MAD Res/Avg Act)	23.64%
Pearson's Correlation Coefficient between Act & Pred	33.04%
Adjusted R-Squared in Unit Space	6.23%

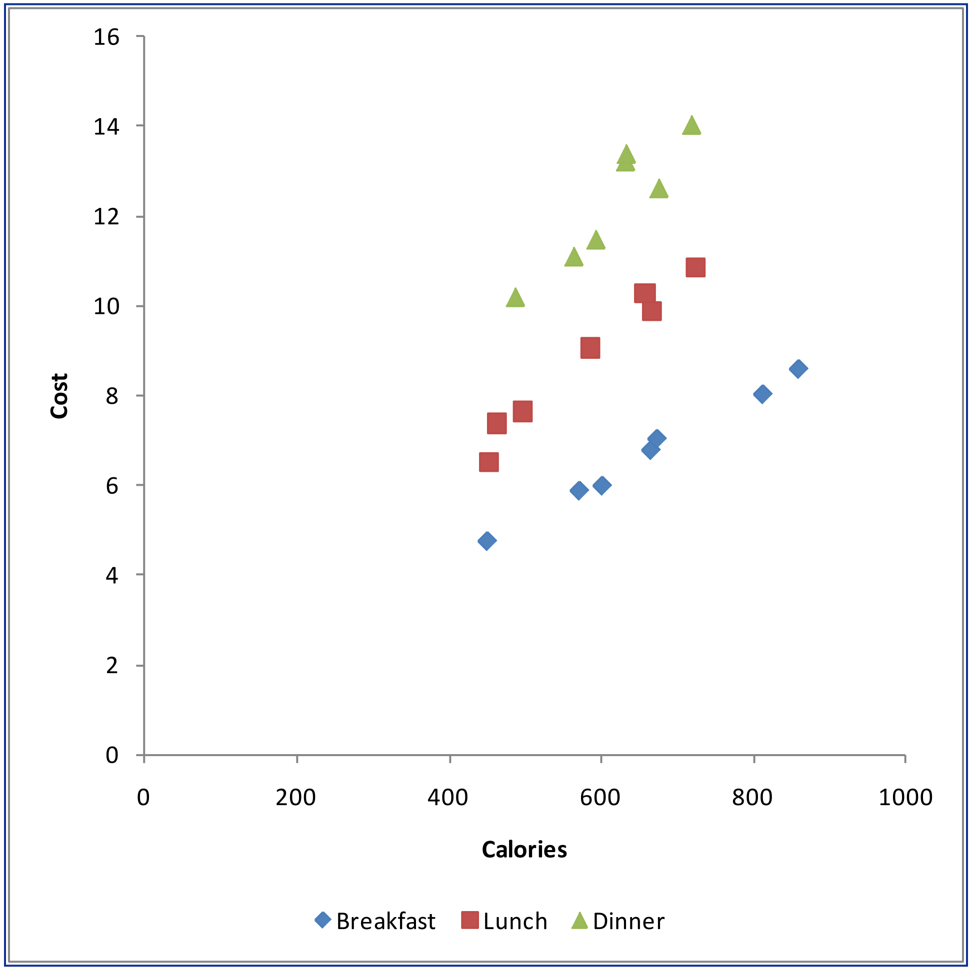
Weak





Stratify the Data into Classes

Observations	Cost of Meal	Calories	Type of Meal
Variable ID	Cost	Calories	
Meal Number 1	6.78	665	Breakfast
Meal Number 2	4.76	450	Breakfast
Meal Number 3	8.02	812	Breakfast
Meal Number 4	5.99	601	Breakfast
Meal Number 5	8.58	859	Breakfast
Meal Number 6	7.03	673	Breakfast
Meal Number 7	5.88	571	Breakfast
Meal Number 8	10.27	657	Lunch
Meal Number 9	10.86	724	Lunch
Meal Number 10	9.06	585	Lunch
Meal Number 11	7.38	463	Lunch
Meal Number 12	6.53	452	Lunch
Meal Number 13	9.88	667	Lunch
Meal Number 14	7.65	496	Lunch
Meal Number 15	11.09	563	Dinner
Meal Number 16	13.22	631	Dinner
Meal Number 17	10.18	486	Dinner
Meal Number 18	11.47	592	Dinner
Meal Number 19	13.39	632	Dinner
Meal Number 20	14.04	718	Dinner
Meal Number 21	12.62	675	Dinner



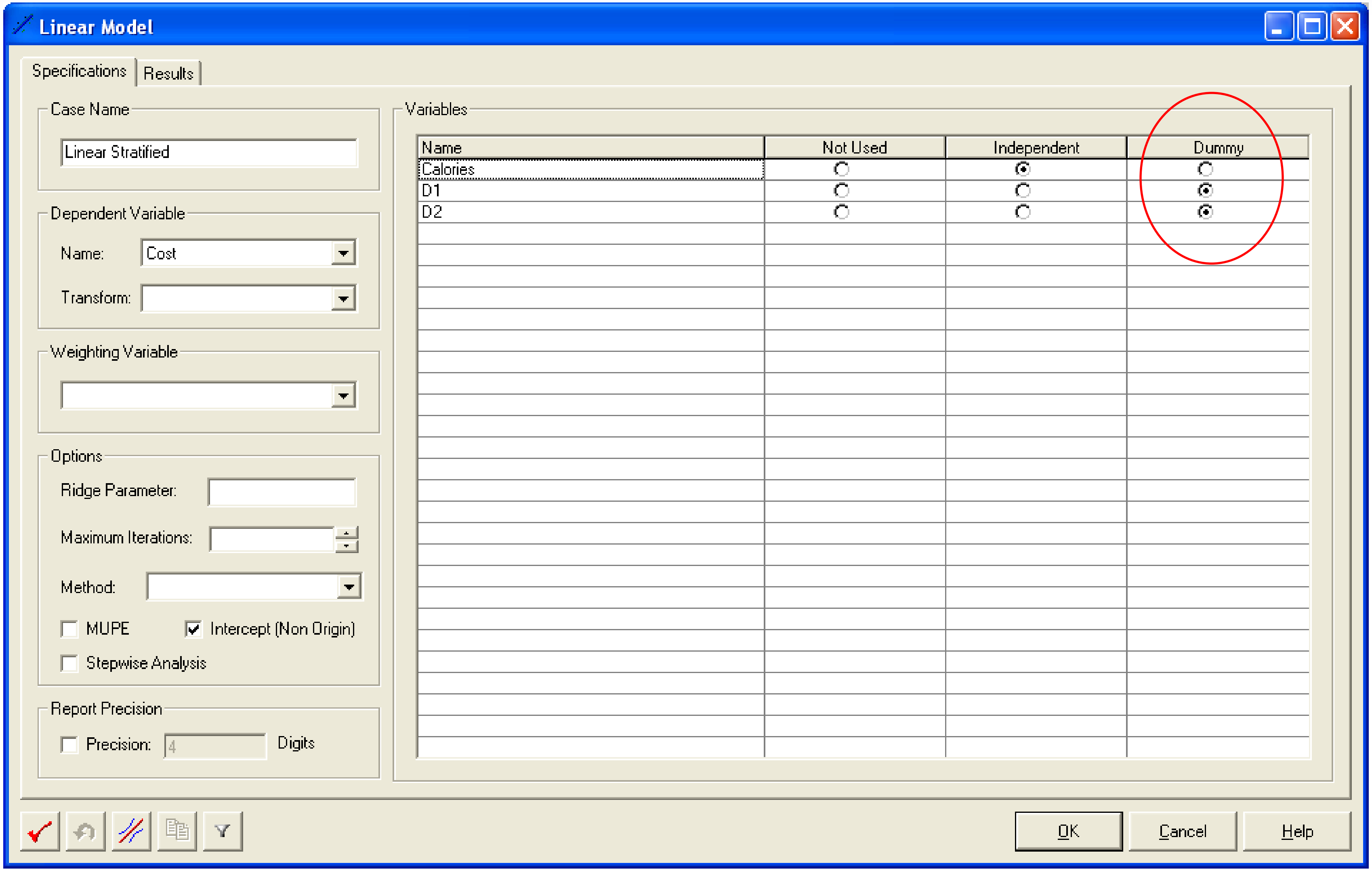
Create a Proxy

- **Treatment of *r* classes requires introduction of *r-1* dummy variables**
  - 3 classes
  - (3-1) = 2 dummy variables
- Let's call these D1 and D2 and denote the pair as {D1, D2} where:
  - {0,0} = Breakfast
  - {1,0} = Lunch
  - {0,1} = Dinner

Dataset Ready for CO$TAT

Observations	Cost of Meal	Calories	Type of Meal	Class of Meal Proxy 1 of 2	Class of Meal Proxy 2 of 2
Variable ID	Cost	Calories		D1	D2
Meal Number 1	6.78	665	Breakfast	0	0
Meal Number 2	4.76	450	Breakfast	0	0
Meal Number 3	8.02	812	Breakfast	0	0
Meal Number 4	5.99	601	Breakfast	0	0
Meal Number 5	8.58	859	Breakfast	0	0
Meal Number 6	7.03	673	Breakfast	0	0
Meal Number 7	5.88	571	Breakfast	0	0
Meal Number 8	10.27	657	Lunch	1	0
Meal Number 9	10.86	724	Lunch	1	0
Meal Number 10	9.06	585	Lunch	1	0
Meal Number 11	7.38	463	Lunch	1	0
Meal Number 12	6.53	452	Lunch	1	0
Meal Number 13	9.88	667	Lunch	1	0
Meal Number 14	7.65	496	Lunch	1	0
Meal Number 15	11.09	563	Dinner	0	1
Meal Number 16	13.22	631	Dinner	0	1
Meal Number 17	10.18	486	Dinner	0	1
Meal Number 18	11.47	592	Dinner	0	1
Meal Number 19	13.39	632	Dinner	0	1
Meal Number 20	14.04	718	Dinner	0	1
Meal Number 21	12.62	675	Dinner	0	1

CO$TAT Linear Model



Stratified Example

Linear Analysis for Dataset Meals Dataset, Linear Stratified
Wednesday, 16 May 2012, 2:09 pm

I. Model Form and Equation Table

Model Form:	Unweighted Linear model
Number of Observations Used:	21
Equation in Unit Space:	Cost = (-1.38) + 0.01224 * Calories + 3.111 * D1 + 6.151 * D2

OK

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	-1.3804	0.7556		-1.8269	0.0852	0.9148
Calories	0.0122	0.0011	0.5006	11.1063	0.0000	1.0000
D1	3.1110	0.2940	0.5505	10.5806	0.0000	1.0000
D2	6.1514	0.2840	1.0885	21.6571	0.0000	1.0000

OK

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
0.5222	96.89%	96.34%	0.9843

OK

Analysis of Variance

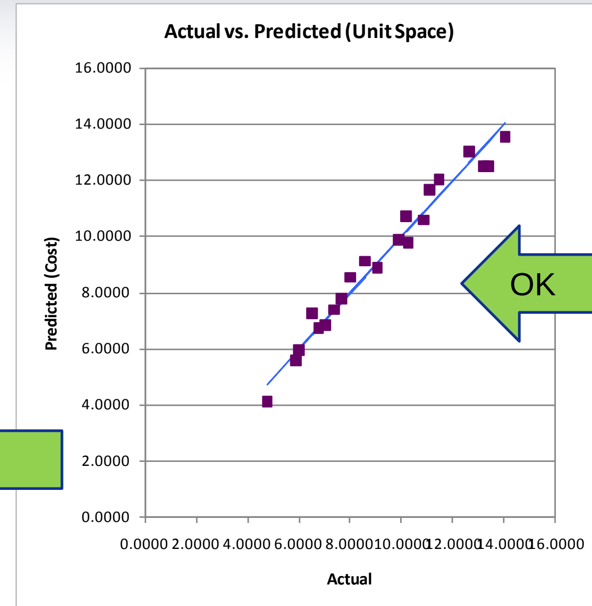
Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	3	144.3926	48.1309	176.5110	0.0000	1.0000
Residual (Error)	17	4.6355	0.2727			
Total	20	149.0281				

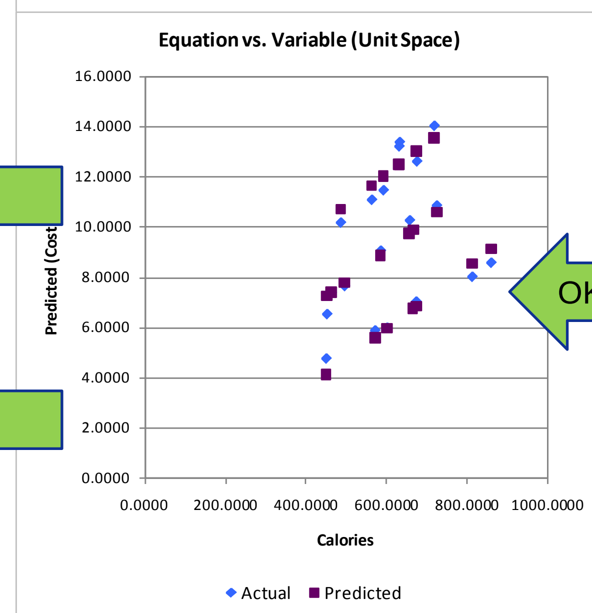
OK

Summary of Predictive Measures

Average Actual (Avg Act)	9.2705
Standard Error (SE)	0.5222
Root Mean Square (RMS) of % Errors	5.47%
Mean Absolute Deviation (Mad) of % Errors	4.32%
Coef of Variation based on Std Error (SE/Avg Act)	5.63%
Coef of Variation based on MAD Res (MAD Res/Avg Act)	4.22%
Pearson's Correlation Coefficient between Act & Pred	98.43%
Adjusted R-Squared in Unit Space	96.34%

OK





Final Equations

Cost = (-1.38) + 0.01224 * Calories + 3.111 * D1 + 6.151 * D2

When {D1,D2} = {0, 0}

Cost = (-1.38) + 0.01224 * Calories + 3.111 * 0 + 6.151 * 0

Cost = (-1.38) + 0.01224 * Calories + 0 + 0

Cost = (-1.38) + 0.01224 * Calories ← Breakfast CER

When {D1,D2} = {1, 0}

Cost = (-1.38) + 0.01224 * Calories + 3.111 * 1 + 6.151 * 0

Cost = (-1.38) + 0.01224 * Calories + 3.111 + 0

Cost = 1.731 + 0.01224 * Calories ← Lunch CER

When {D1,D2} = {0, 1}

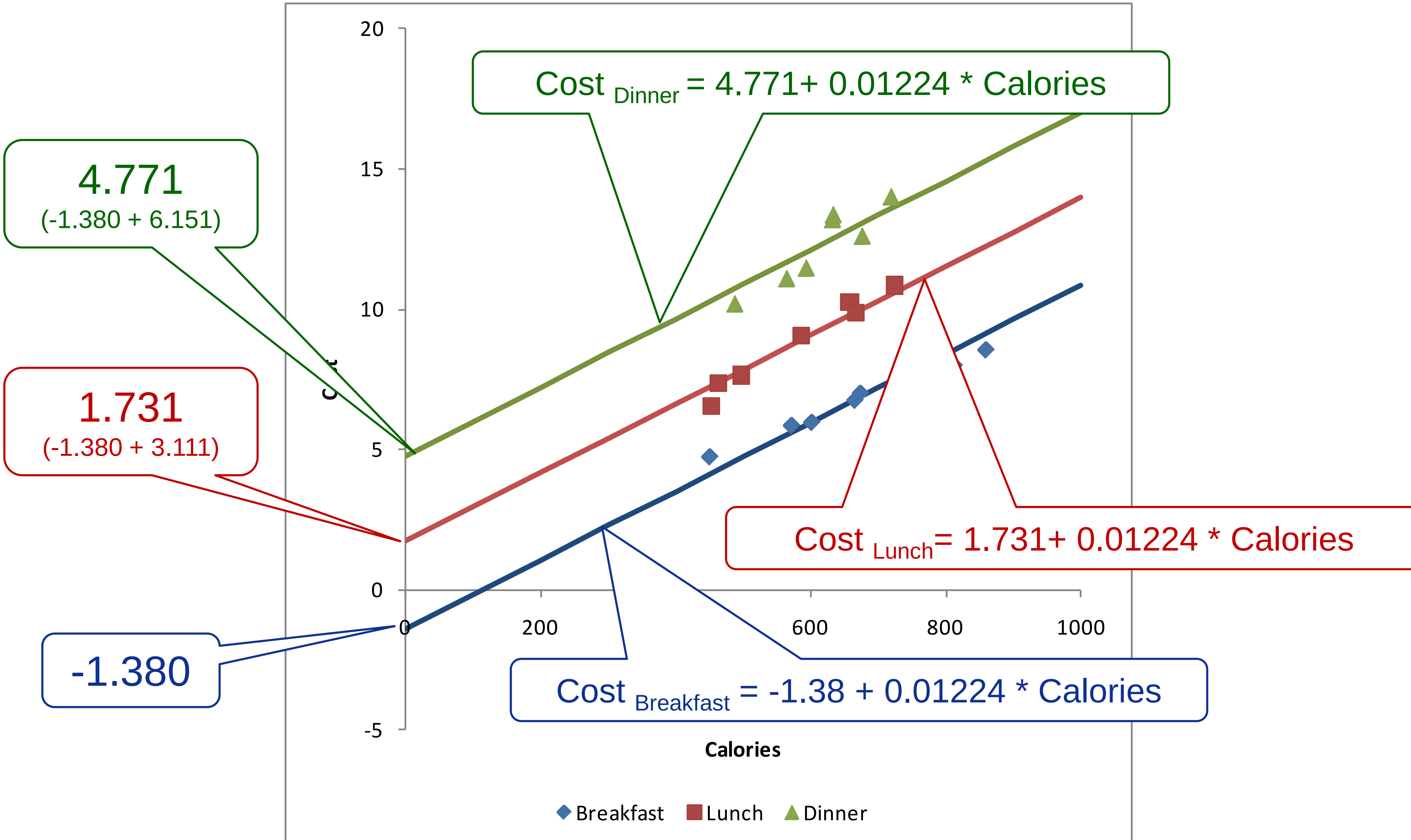
Cost = (-1.38) + 0.01224 * Calories + 3.111 * 0 + 6.151 * 1

Cost = (-1.38) + 0.01224 * Calories + 0 + 6.151

Cost = 4.71 + 0.01224 * Calories ← Dinner CER

Final Equations

- **Cost = (-1.38) + 0.01224 * Calories + 3.111 * D1 + 6.151 * D2**



Degrees-of-Freedom

- **Each dummy variable will consume one degree-of-freedom**

Cost = 4.279 + 0.008081 * Calories

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	1	16.2674	16.2674	2.3281	0.1434	0.8566
Residual (Error)	19	132.7607	6.9874			
Total	20	149.0281				

Cost = (-1.38) + 0.01224 * Calories + 3.111 * D1 + 6.151 * D2

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	3	144.3926	48.1309	176.5110	0.0000	1.0000
Residual (Error)	17	4.6355	0.2727			
Total	20	149.0281				

The Pattern in Creating and Assigning Dummy Variables

- Keep in mind, the dummy variables are to represent a unique proxy for each class …

  … in the simplest way possible.

- For one class, add no dummy variables
- For two classes, add one dummy variable
  - With two possible values 1 and 0
- For three classes, add two dummy variables
  - With three possible paired values {1,0}, {0,1}, {0,0}
- For four classes, add three dummy variables
  - With four possible value sets {1,0,0}, {0,1,0}, {0,0,1}, {0,0,0}

- In each instance there is a class where every dummy variable is set to zero

Formal Pattern Statement

" … we can deal with r levels by the introduction of (r-1) dummy variables. The basic allocation pattern is obtained by writing down an (r-1) x (r-1) **I** matrix and adding a row of (r-1) zeros."*Applied Regression Analysis*, Second Edition, Norman Draper and Harry Smith, 1981

Six (6) classes for example:

X_1	X_2	X_3	X_4	X_5
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1
0	0	0	0	0

5x5 Identity Matrix

A row of 5 zeros

Matrix Math Statement

Regression expressed in matrix terms:

$$b = (X'X)^{-1}X'Y$$

where **X** is a matrix of observed independent variables

(include a column of ones in **X** for the intercept)

where **Y** is a matrix of observed dependent variables

where **b** is a matrix of regression equation coefficients

Example Matrices

$$b = \begin{bmatrix} b0 \\ b1 \\ b2 \\ b3 \end{bmatrix}$$

$$X = \begin{array}{c} \begin{array}{cccc} \text{Intercept} & \text{Calories} & \text{D1} & \text{D2} \end{array} \\ \begin{bmatrix} 1 & 665 & 0 & 0 \\ 1 & 450 & 0 & 0 \\ 1 & 812 & 0 & 0 \\ 1 & 601 & 0 & 0 \\ 1 & 859 & 0 & 0 \\ 1 & 673 & 0 & 0 \\ 1 & 571 & 0 & 0 \\ 1 & 657 & 1 & 0 \\ 1 & 724 & 1 & 0 \\ 1 & 585 & 1 & 0 \\ 1 & 463 & 1 & 0 \\ 1 & 452 & 1 & 0 \\ 1 & 667 & 1 & 0 \\ 1 & 496 & 1 & 0 \\ 1 & 563 & 0 & 1 \\ 1 & 631 & 0 & 1 \\ 1 & 486 & 0 & 1 \\ 1 & 592 & 0 & 1 \\ 1 & 632 & 0 & 1 \\ 1 & 718 & 0 & 1 \\ 1 & 675 & 0 & 1 \end{bmatrix} \end{array}$$

$$Y = \begin{array}{c} \text{Cost} \\ \begin{bmatrix} 6.78 \\ 4.76 \\ 8.02 \\ 5.99 \\ 8.58 \\ 7.03 \\ 5.88 \\ 10.27 \\ 10.86 \\ 9.06 \\ 7.38 \\ 6.53 \\ 9.88 \\ 7.65 \\ 11.09 \\ 13.22 \\ 10.18 \\ 11.47 \\ 13.39 \\ 14.04 \\ 12.62 \end{bmatrix} \end{array}$$

Matrix Math Statement

Regression expressed in matrix terms:

$b = (X'X)^{-1}X'Y$

where **X** is a matrix of observed independent variables

(include a column of ones in **X** for the intercept)

where **Y** is a matrix of observed dependent variables

where **b** is a matrix of regression equation coefficients

For regression in matrix terms to work,

"$(X'X)^{-1}$ must exist"

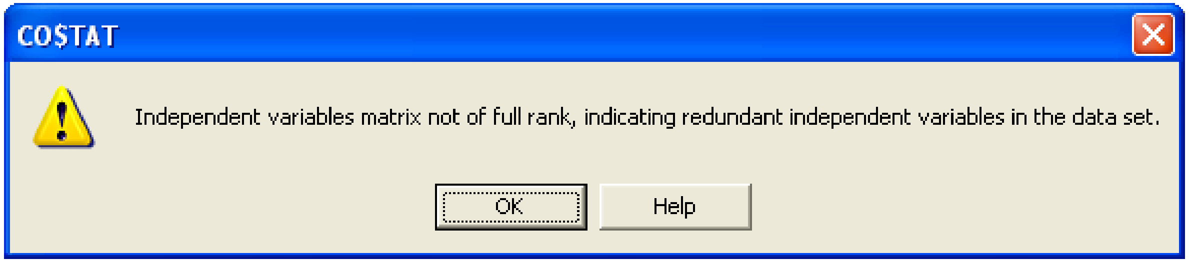
When does it not exist?

When any column of the matrix is not independent of the other columns.

Incorrect Example Matrix X

Intercept	Calories	D1	D2	D3
1	665	1	0	0
1	450	1	0	0
1	812	1	0	0
1	601	1	0	0
1	859	1	0	0
1	673	1	0	0
1	571	1	0	0
1	657	0	1	0
1	724	0	1	0
1	585	0	1	0
1	463	0	1	0
1	452	0	1	0
1	667	0	1	0
1	496	0	1	0
1	563	0	0	1
1	631	0	0	1
1	486	0	0	1
1	592	0	0	1
1	632	0	0	1
1	718	0	0	1
1	675	0	0	1

Incorrect!



One of the columns is <u>not</u> independent of the other columns

Intercept = D1 + D2 + D3

CER Example Using Log Linear Regression

Log Linear Equations

Cost = 0.01623 * Calories ^ 0.9282 * 1.482 ^ D1 * 1.959 ^ D2

$\text{Cost} = 0.01623 * \text{Calories}^{0.9282} * 1.482^{D1} * 1.959^{D2}$

When {D1,D2} = {0, 0}

$\text{Cost} = 0.01623 * \text{Calories}^{0.9282} * 1.482^{0} * 1.959^{0}$

$\text{Cost} = 0.01623 * \text{Calories}^{0.9282} * 1 * 1$

$\text{Cost} = 0.01623 * \text{Calories}^{0.9282}$ ← Breakfast CER

When {D1,D2} = {1, 0}

$\text{Cost} = 0.01623 * \text{Calories}^{0.9282} * 1.482^{1} * 1.959^{0}$

$\text{Cost} = 0.01623 * \text{Calories}^{0.9282} * 1.482 * 1$

$\text{Cost} = 0.02405 * \text{Calories}^{0.9282}$ ← Lunch CER

When {D1,D2} = {0, 1}

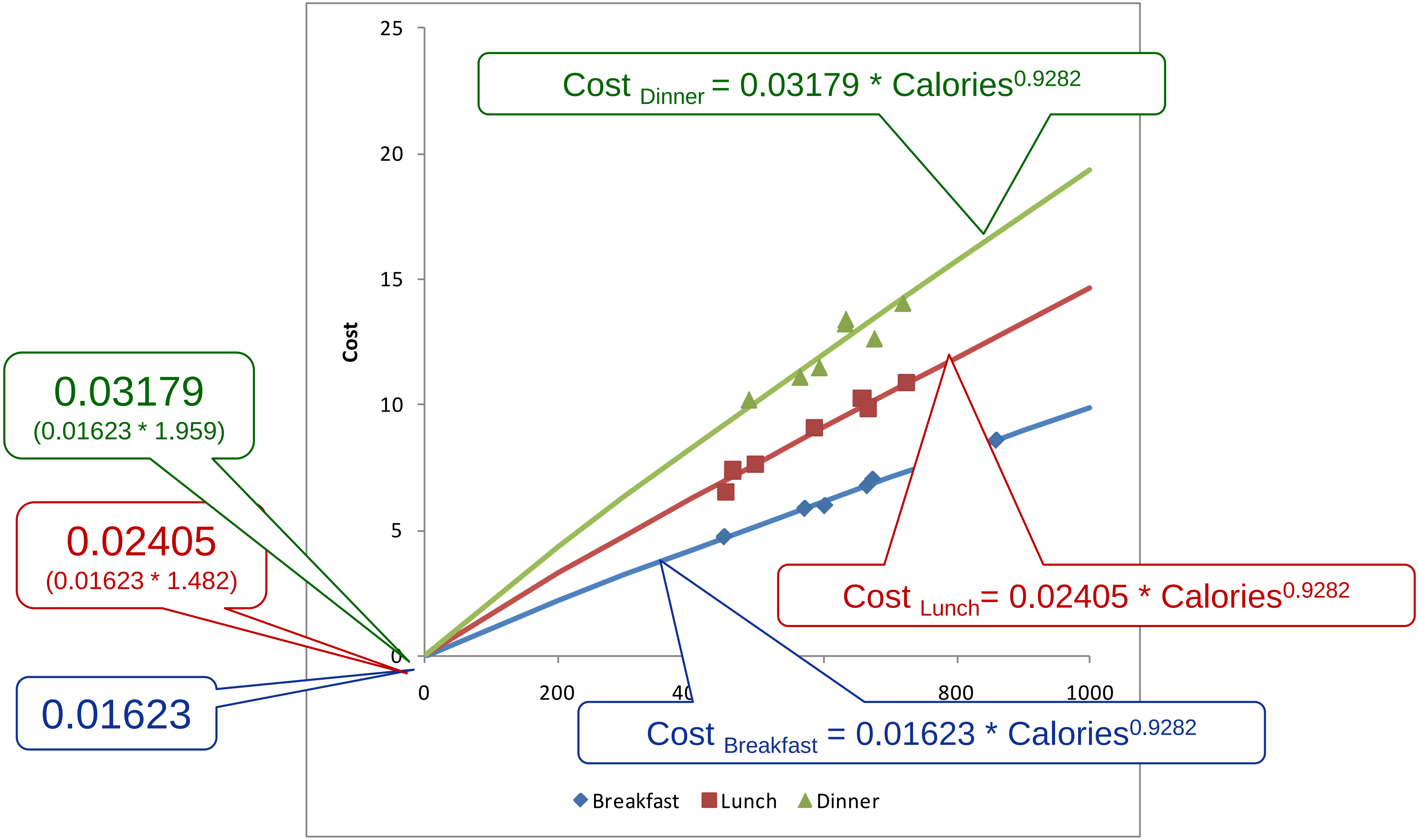
$\text{Cost} = 0.01623 * \text{Calories}^{0.9282} * 1.482^{0} * 1.959^{1}$

$\text{Cost} = 0.01623 * \text{Calories}^{0.9282} * 1 * 1.959$

$\text{Cost} = 0.03179 * \text{Calories}^{0.9282}$ ← Dinner CER

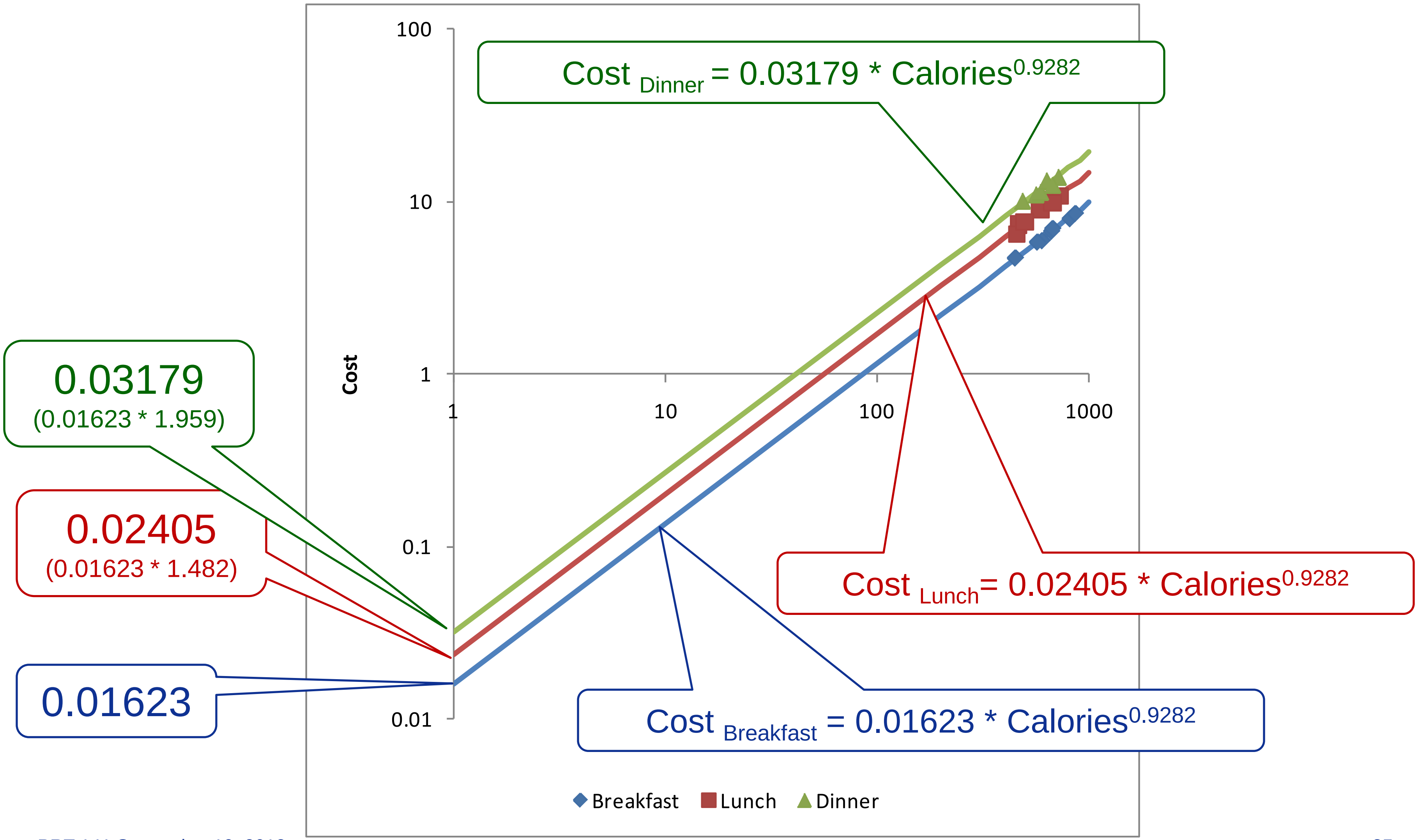
Final Log Linear Equations

$$Cost = 0.01623 * Calories^{0.9282} * 1.482^{D1} * 1.959^{D2}$$



Final Log Linear Equations (in Log Space)

$$Cost = 0.01623 * Calories^{0.9282} * 1.482^{D1} * 1.959^{D2}$$



Example Using Pooled Learning

Discerning a Common Slope Across Several Systems

- **Regressing slopes on each individual system and combining the results is problematic.**

System	Slope (%)
System #1	90
System #2	89
System #3	93
System #4	92
System #5	95

Mean = 91.8
Median = 92.0

System	Slope (%)
System #1	90
System #2	89
System #3	117
System #4	92
System #5	95

Mean = 96.6
Median = 92.0

System	Qty Slope (%)	Rate Slope (%)
System #1	90	90
System #2	68	103
System #3	93	95
System #4	107	62
System #5	95	91

Mean = 90.6; 88.2
Median = 93.0; 91.0

- **Pooled Regression is a Superior Technique**

Define Pooled

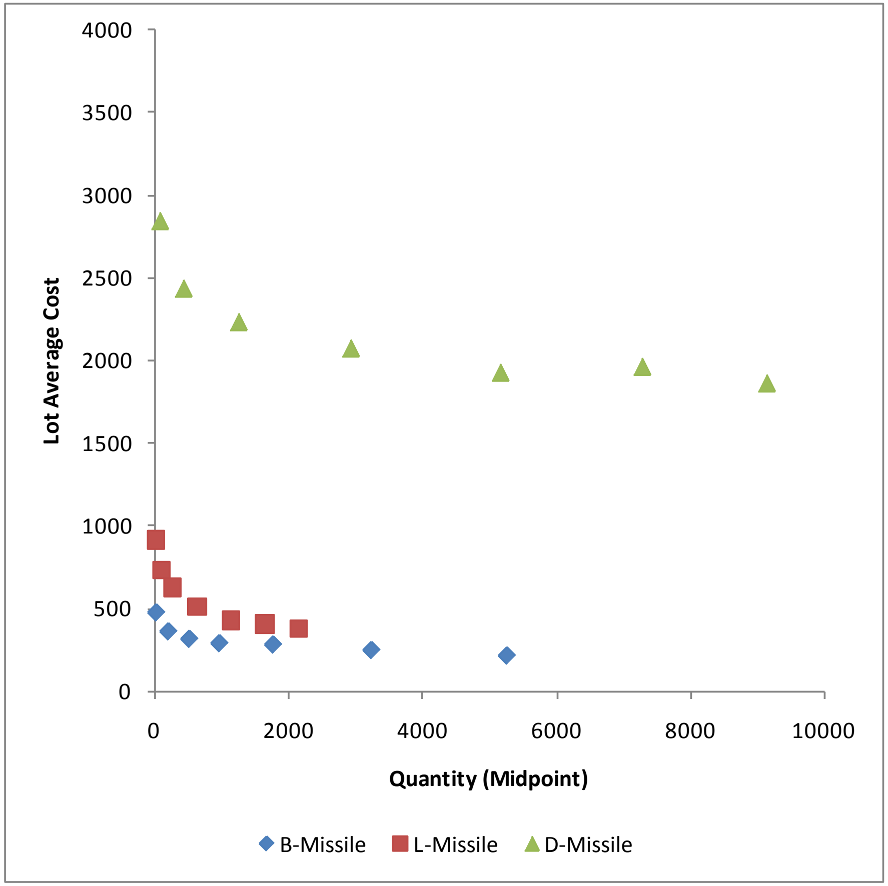
- **Treating each systems' lots as classes with dummy variables enables the analyst to assemble multiple systems into a single "pooled" learning regression.**

Sample Dataset #2

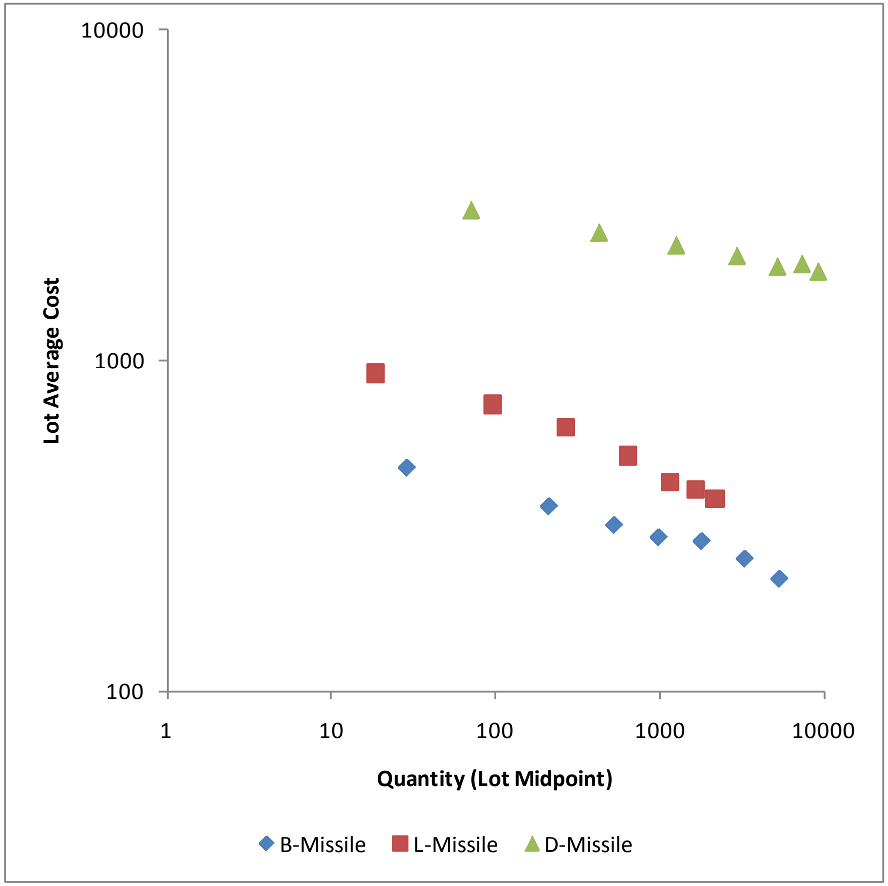
Observations	Lot Total Cost	Quantity	First Unit of Lot	Last Unit of Lot	Missile System	Class of Missile Proxy 1 of 2	Class of Missile Proxy 2 of 2
Variable ID	LTC	Quantity	First_Unit	Last_Unit		D1	D2
B-Missile Lot 1	38,094	80	1	80	B-Missile	0	0
B-Missile Lot 2	108,843	300	81	380	B-Missile	0	0
B-Missile Lot 3	95,421	300	381	680	B-Missile	0	0
B-Missile Lot 4	181,016	620	681	1,300	B-Missile	0	0
B-Missile Lot 5	284,094	1,000	1,301	2,300	B-Missile	0	0
B-Missile Lot 6	502,042	2,000	2,301	4,300	B-Missile	0	0
B-Missile Lot 7	435,818	2,000	4,301	6,300	B-Missile	0	0
L-Missile Lot 1	45,838	50	1	50	L-Missile	1	0
L-Missile Lot 2	73,711	100	51	150	L-Missile	1	0
L-Missile Lot 3	157,126	250	151	400	L-Missile	1	0
L-Missile Lot 4	257,549	500	401	900	L-Missile	1	0
L-Missile Lot 5	215,310	500	901	1,400	L-Missile	1	0
L-Missile Lot 6	203,333	500	1,401	1,900	L-Missile	1	0
L-Missile Lot 7	191,125	500	1,901	2,400	L-Missile	1	0
D-Missile Lot 1	568,734	200	1	200	D-Missile	0	1
D-Missile Lot 2	1,217,274	500	201	700	D-Missile	0	1
D-Missile Lot 3	2,677,707	1,200	701	1,900	D-Missile	0	1
D-Missile Lot 4	4,557,980	2,200	1,901	4,100	D-Missile	0	1
D-Missile Lot 5	4,235,452	2,200	4,101	6,300	D-Missile	0	1
D-Missile Lot 6	3,921,465	2,000	6,301	8,300	D-Missile	0	1
D-Missile Lot 7	3,162,458	1,700	8,301	10,000	D-Missile	0	1

Fake Data

Sample Dataset #2

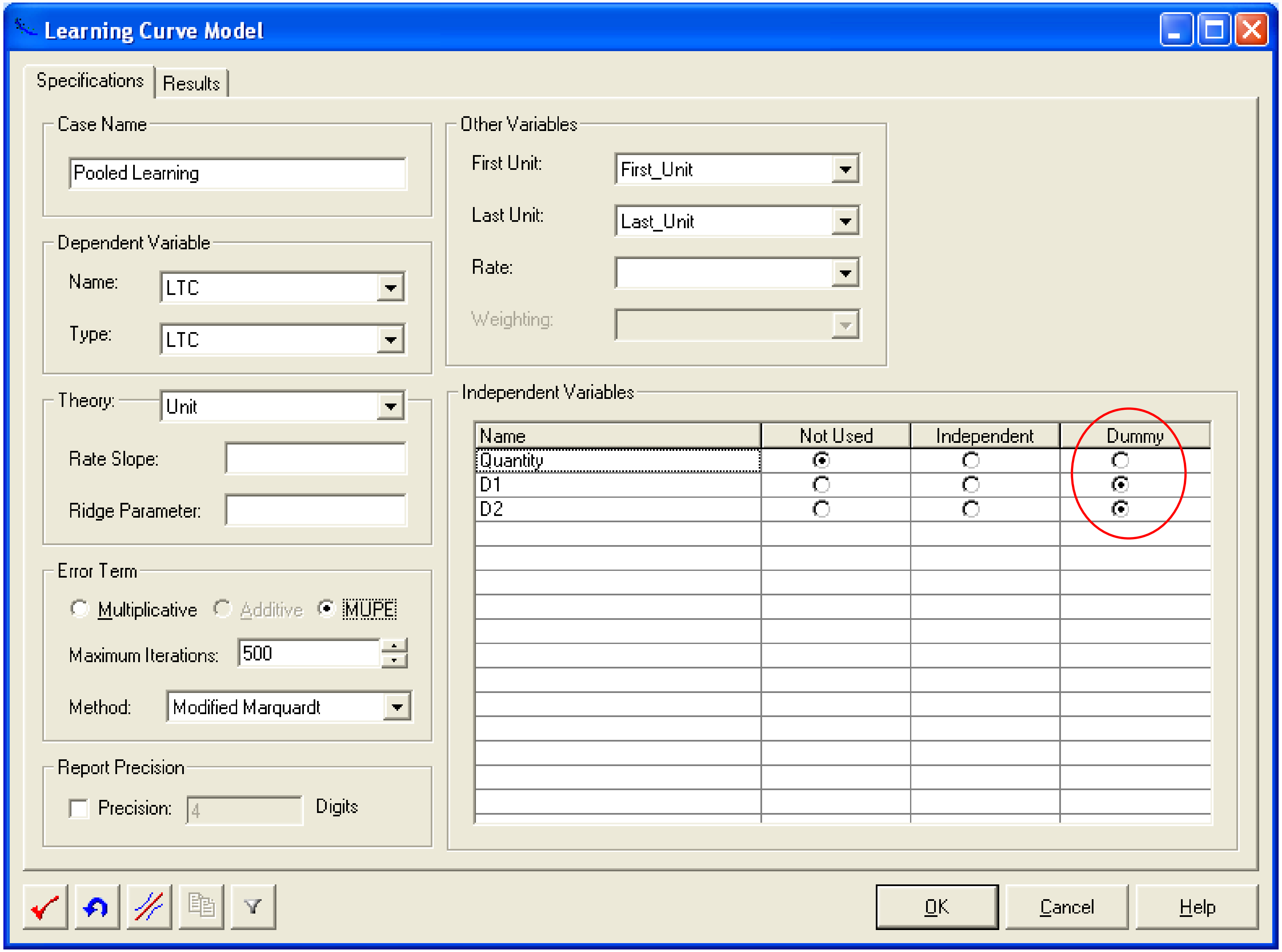


Unit Space



Log Space

CO$TAT Learning Curve Model



CO$TAT Results

I. Model Form and Equation Table

Model Form:	Unweighted Learning Curve (Unit Theory)
Number of Observations Used:	21
Equation in Unit Space:	UNIT_COST = 756.4 * UNIT_NUM ^ (-0.138) * 1.642 ^ D1 * 7.979 ^ D2
T1:	756.4175
Quantity Slope:	90.88%
Dependent Variable:	LTC
Quantity Variable:	First Unit and Last Unit Variables

OK

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	6.6286	0.0805		82.3717	0.0000	1.0000
CUM_QTY	-0.1380	0.0113	-0.2802	-12.2526	0.0000	1.0000
EXP_D1	0.4958	0.0454	0.2734	10.9165	0.0000	1.0000
EXP_D2	2.0768	0.0459	1.1453	45.2499	0.0000	1.0000

OK

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
0.0840	99.22%	99.08%	0.9961

OK

Analysis of Variance

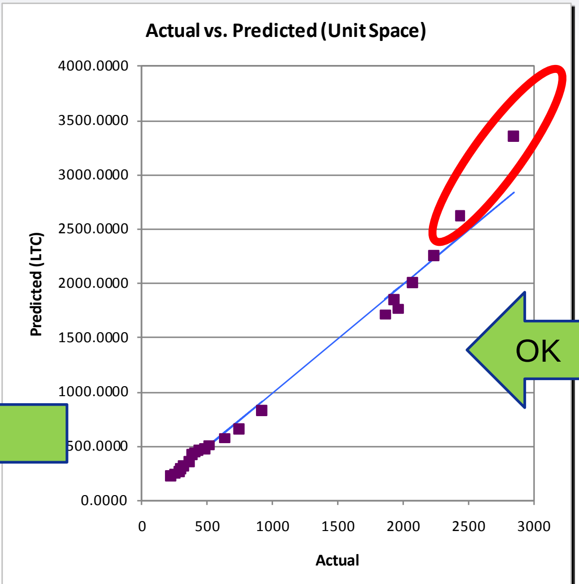
Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	3	15.2251	5.0750	720.0595	0.0000	1.0000
Residual (Error)	17	0.1198	0.0070			
Total	20	15.3450				

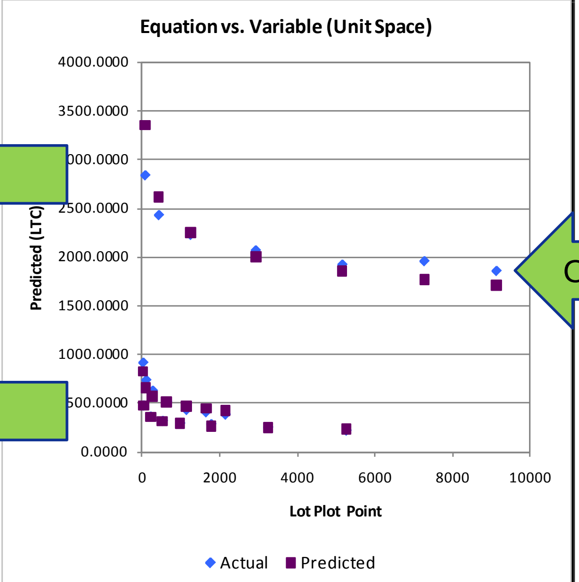
OK

Summary of Predictive Measures

Average Actual (Avg Act)	1026.0338
Standard Error (SE)	151.2369
Root Mean Square (RMS) of % Errors	7.70%
Mean Absolute Deviation (Mad) of % Errors	5.99%
Coef of Variation based on Std Error (SE/Avg Act)	14.74%
Coef of Variation based on MAD Res (MAD Res/Avg Act)	7.33%
Pearson's Correlation Coefficient between Act & Pred	98.99%
Adjusted R-Squared in Unit Space	97.04%

OK





Final Equations

$$UNIT\_COST = 756.4 * UNIT\_NUM \wedge (-0.138) * 1.642 \wedge D1 * 7.979 \wedge D2$$

$$UNIT\_COST = 756.4 * UNIT\_NUM^{(-0.138)} * 1.642^{D1} * 7.979^{D2}$$

When {D1,D2} = {0, 0}

$$UNIT\_COST = 756.4 * UNIT\_NUM^{(-0.138)} * 1.642^{0} * 7.979^{0}$$

$$UNIT\_COST = 756.4 * UNIT\_NUM^{(-0.138)} * 1 * 1$$

$$UNIT\_COST = 756.4 * UNIT\_NUM^{(-0.138)}$$ ← B-Missile Learning Curve

When {D1,D2} = {1, 0}

$$UNIT\_COST = 756.4 * UNIT\_NUM^{(-0.138)} * 1.642^{1} * 7.979^{0}$$

$$UNIT\_COST = 756.4 * UNIT\_NUM^{(-0.138)} * 1.642 * 1$$

$$UNIT\_COST = 1242. * UNIT\_NUM^{(-0.138)}$$ ← L-Missile Learning Curve

When {D1,D2} = {0, 1}

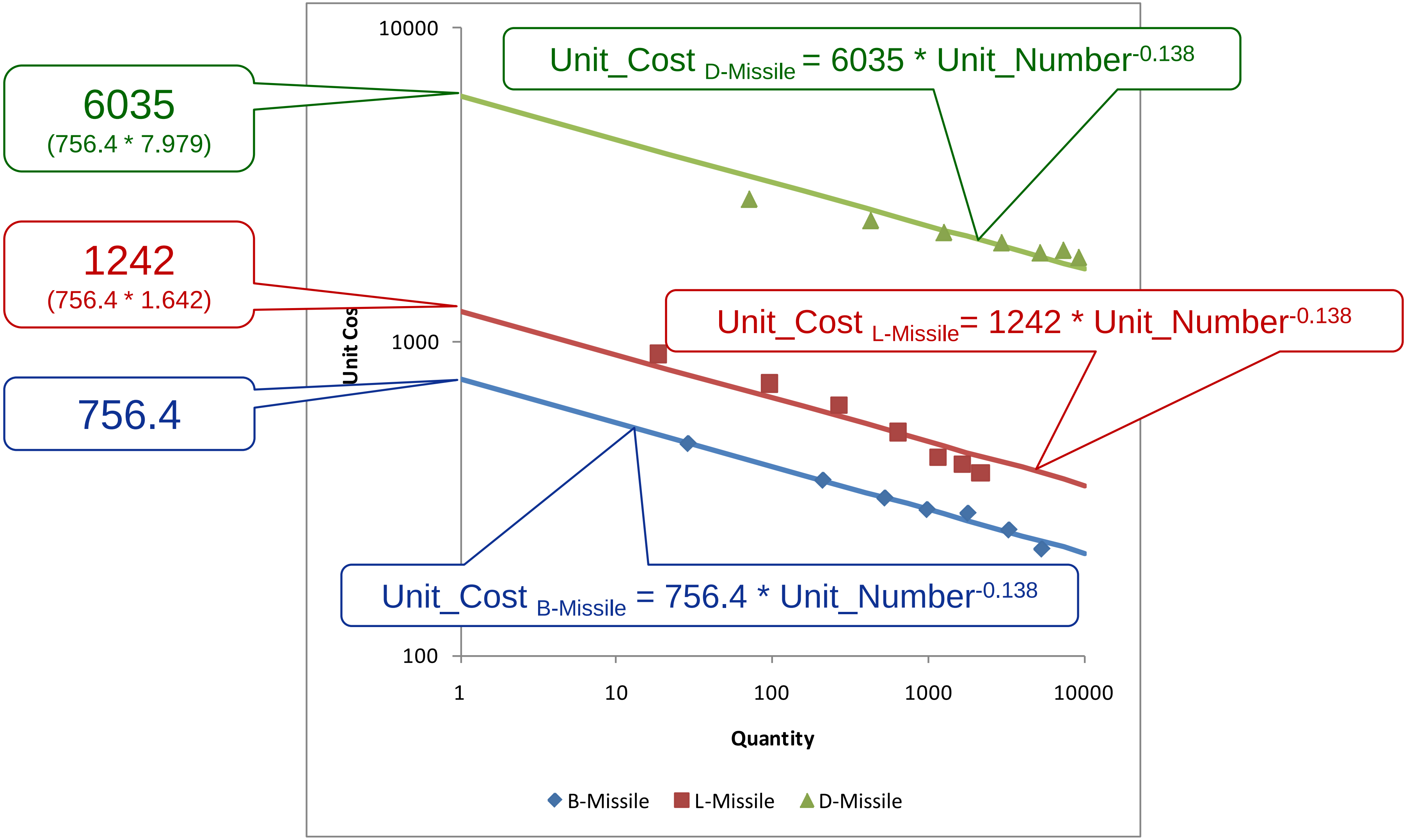
$$UNIT\_COST = 756.4 * UNIT\_NUM^{(-0.138)} * 1.642^{0} * 7.979^{1}$$

$$UNIT\_COST = 756.4 * UNIT\_NUM^{(-0.138)} * 1 * 7.979$$

$$UNIT\_COST = 6035. * UNIT\_NUM^{(-0.138)}$$ ← D-Missile Learning Curve

Final Equations (Log Space)

UNIT_COST = 756.4 * UNIT_NUM ^ (-0.138) * 1.642 ^ D1 * 7.979 ^ D2



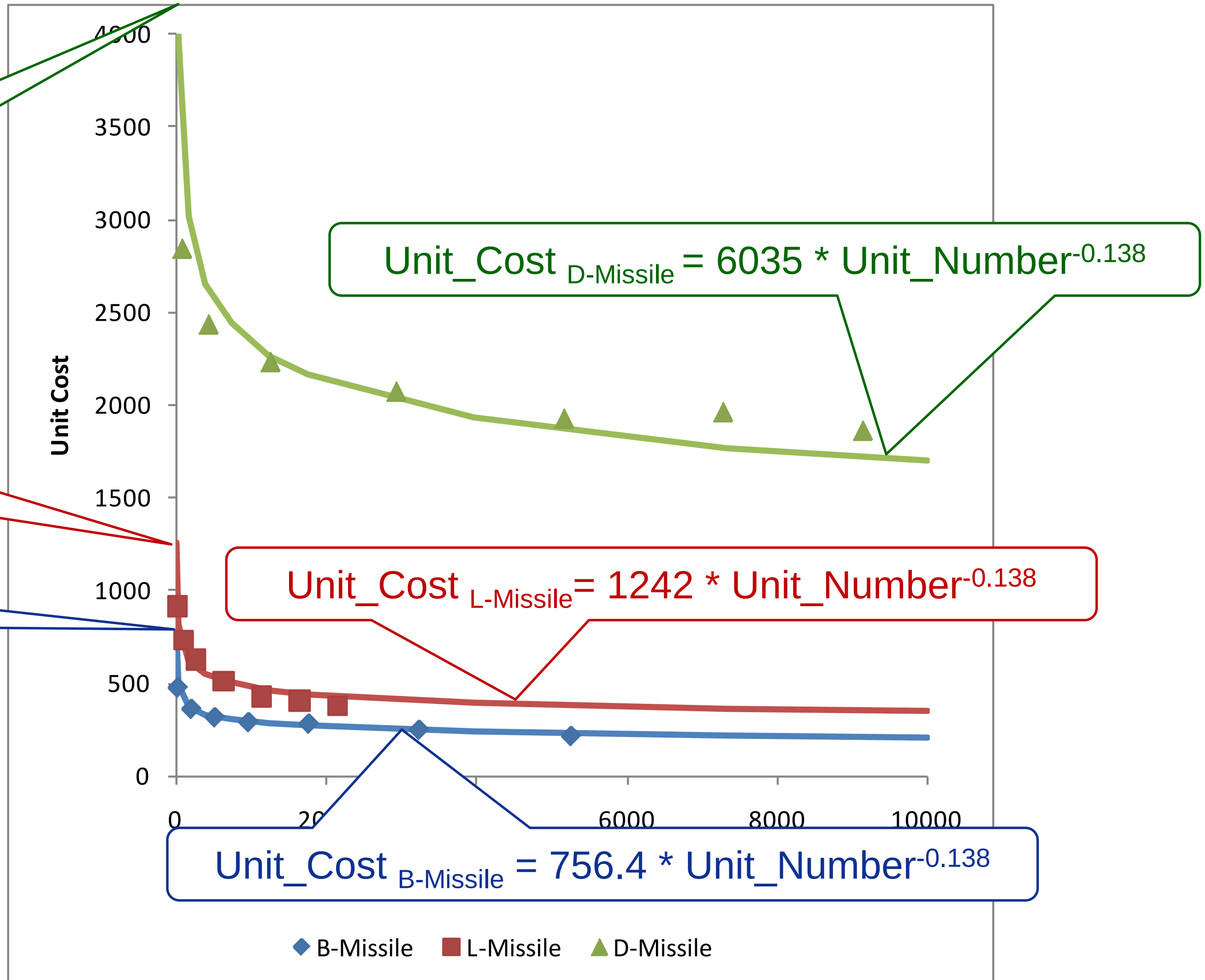
Final Equations (Unit Space)

UNIT_COST = 756.4 * UNIT_NUM ^ (-0.138) * 1.642 ^ D1 * 7.979 ^ D2

6035
(756.4 * 7.979)

1242
(756.4 * 1.642)

756.4



Tricks and Tips
Using Filters for Data Subsets

Analyze a Single System

- **No need to create separate sheets for each system.**
- **In CO$TAT 7.1 an option was added which allows the user to filter out particular data points from the dataset.**
  - A Filter button is located at the bottom of each analysis specification form which launches a Select Filter dialog. It allows the user to include/exclude specific data points for the report.
  - Previous versions required use of a weighting column to include/exclude observations.

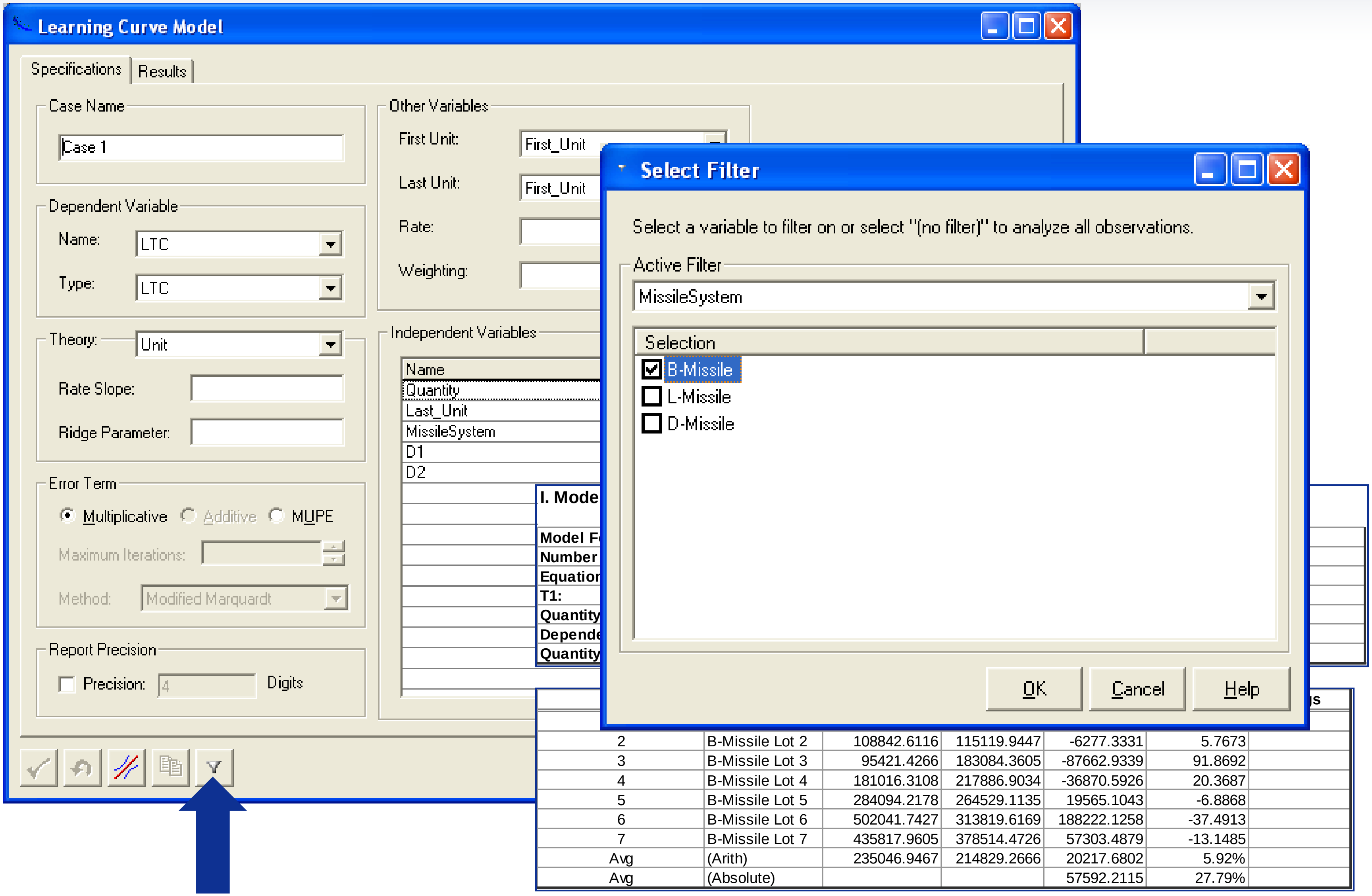
CO$TAT Sheet Ready for Filter

Observations	Lot Total Cost	Quantity	First Unit of Lot	Last Unit of Lot	Missile System	Class of Missile Proxy 1 of 2	Class of Missile Proxy 2 of 2
Variable ID	LTC	Quantity	First_Unit	Last_Unit	MissileSystem	D1	D2
B-Missile Lot 1	38,094	80	1	80	B-Missile	0	0
B-Missile Lot 2	108,843	300	81	380	B-Missile	0	0
B-Missile Lot 3	95,421	300	381	680	B-Missile	0	0
B-Missile Lot 4	181,016	620	681	1,300	B-Missile	0	0
B-Missile Lot 5	284,094	1,000	1,301	2,300	B-Missile	0	0
B-Missile Lot 6	502,042	2,000	2,301	4,300	B-Missile	0	0
B-Missile Lot 7	435,818	2,000	4,301	6,300	B-Missile	0	0
L-Missile Lot 1	45,838	50	1	50	L-Missile	1	0
L-Missile Lot 2	73,711	100	51	150	L-Missile	1	0
L-Missile Lot 3	157,126	250	151	400	L-Missile	1	0
L-Missile Lot 4	257,549	500	401	900	L-Missile	1	0
L-Missile Lot 5	215,310	500	901	1,400	L-Missile	1	0
L-Missile Lot 6	203,333	500	1,401	1,900	L-Missile	1	0
L-Missile Lot 7	191,125	500	1,901	2,400	L-Missile	1	0
D-Missile Lot 1	568,734	200	1	200	D-Missile	0	1
D-Missile Lot 2	1,217,274	500	201	700	D-Missile	0	1
D-Missile Lot 3	2,677,707	1,200	701	1,900	D-Missile	0	1
D-Missile Lot 4	4,557,980	2,200	1,901	4,100	D-Missile	0	1
D-Missile Lot 5	4,235,452	2,200	4,101	6,300	D-Missile	0	1
D-Missile Lot 6	3,921,465	2,000	6,301	8,300	D-Missile	0	1
D-Missile Lot 7	3,162,458	1,700	8,301	10,000	D-Missile	0	1

Fake Data

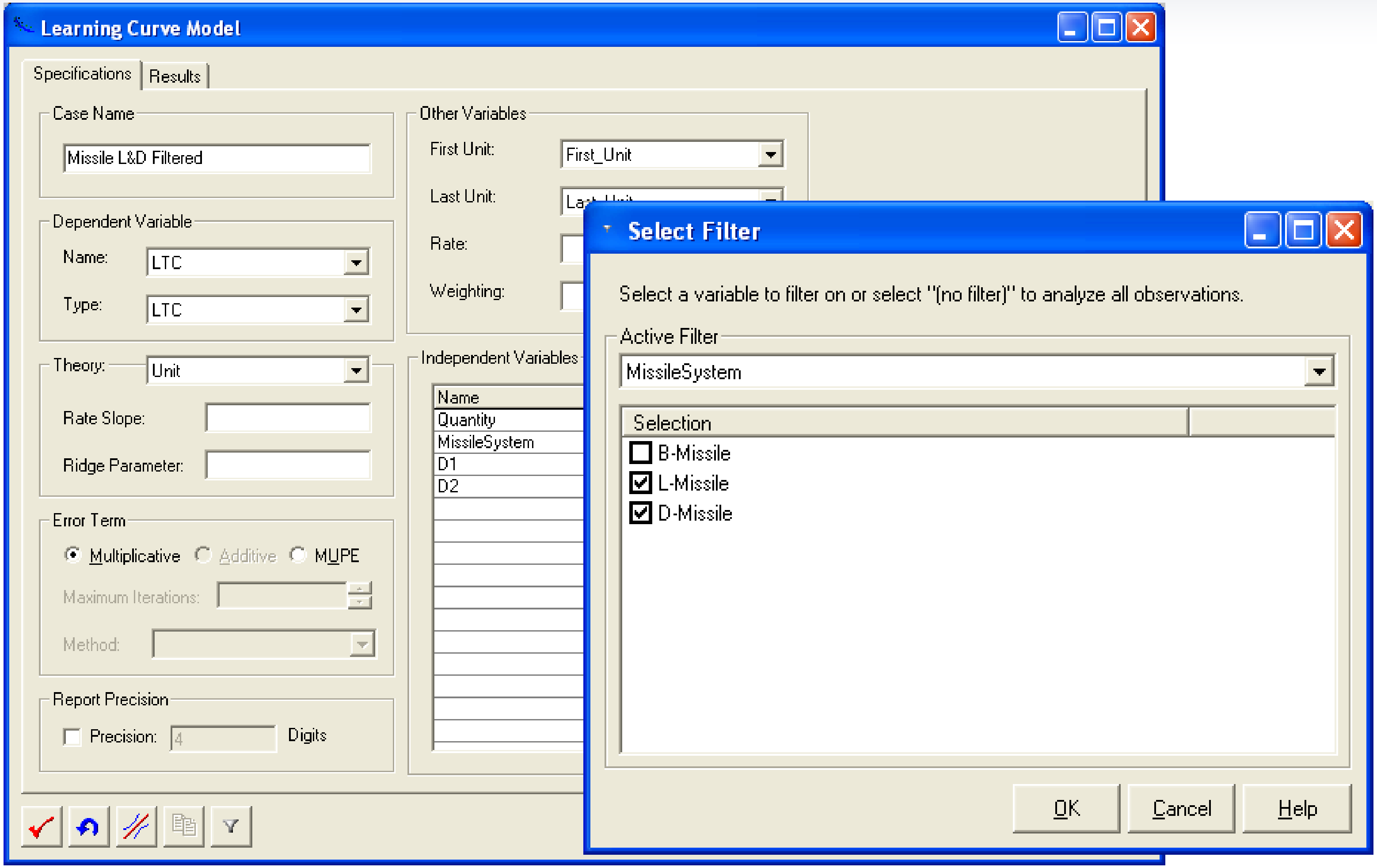
Give the Missile System column a CO$TAT Variable ID

Filter



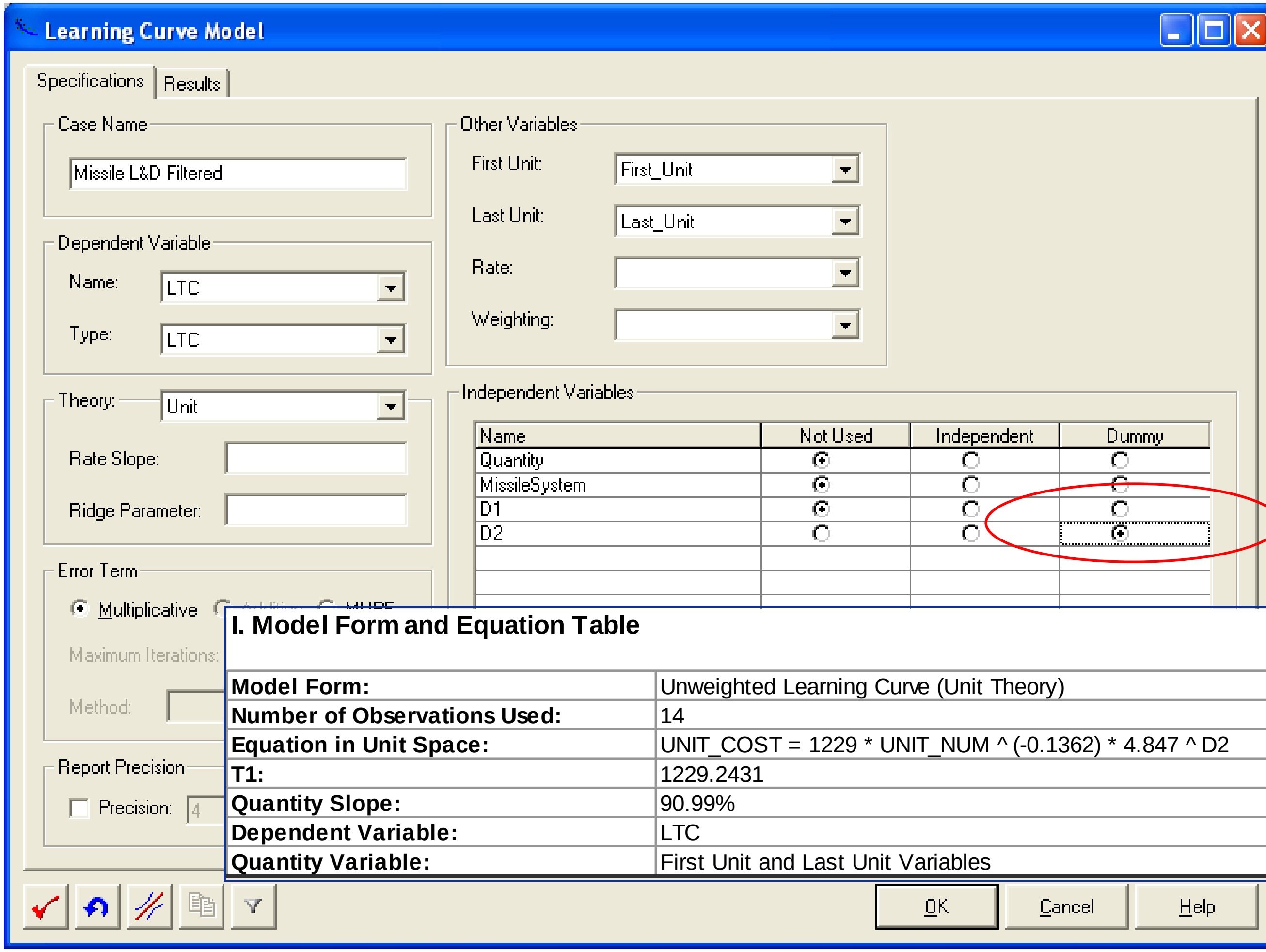
Learning Curve Model
Specifications
Results
Case Name
Case 1
Dependent Variable
Name: LTC
Type: LTC
Theory: Unit
Rate Slope:
Ridge Parameter:
Error Term
Multiplicative
Additive
MUPE
Maximum Iterations:
Method: Modified Marquardt
Report Precision
Precision: 4 Digits
Other Variables
First Unit: First_Unit
Last Unit: First_Unit
Rate:
Weighting:
Independent Variables
Name
Quantity
Last_Unit
MissileSystem
D1
D2
I. Mode
Model F
Number
Equation
T1:
Quantity
Depende
Quantity
Select Filter
Select a variable to filter on or select "(no filter)" to analyze all observations.
Active Filter
MissileSystem
Selection
B-Missile
L-Missile
D-Missile
OK
Cancel
Help
2 B-Missile Lot 2 108842.6116 115119.9447 -6277.3331 5.7673
3 B-Missile Lot 3 95421.4266 183084.3605 -87662.9339 91.8692
4 B-Missile Lot 4 181016.3108 217886.9034 -36870.5926 20.3687
5 B-Missile Lot 5 284094.2178 264529.1135 19565.1043 -6.8868
6 B-Missile Lot 6 502041.7427 313819.6169 188222.1258 -37.4913
7 B-Missile Lot 7 435817.9605 378514.4726 57303.4879 -13.1485
Avg (Arith) 235046.9467 214829.2666 20217.6802 5.92%
Avg (Absolute) 57592.2115 27.79%

Analyze a Subset of Systems



When treating a subset of systems, use the Filter as shown. But when running the analysis …

Analyze a Subset of Systems



I. Model Form and Equation Table

Model Form:	Unweighted Learning Curve (Unit Theory)
Number of Observations Used:	14
Equation in Unit Space:	UNIT_COST = 1229 * UNIT_NUM ^ (-0.1362) * 4.847 ^ D2
T1:	1229.2431
Quantity Slope:	90.99%
Dependent Variable:	LTC
Quantity Variable:	First Unit and Last Unit Variables

… remember to select one fewer dummy variables so that you still have a zero class.

Tricks and Tips
Using Filters for Outliers

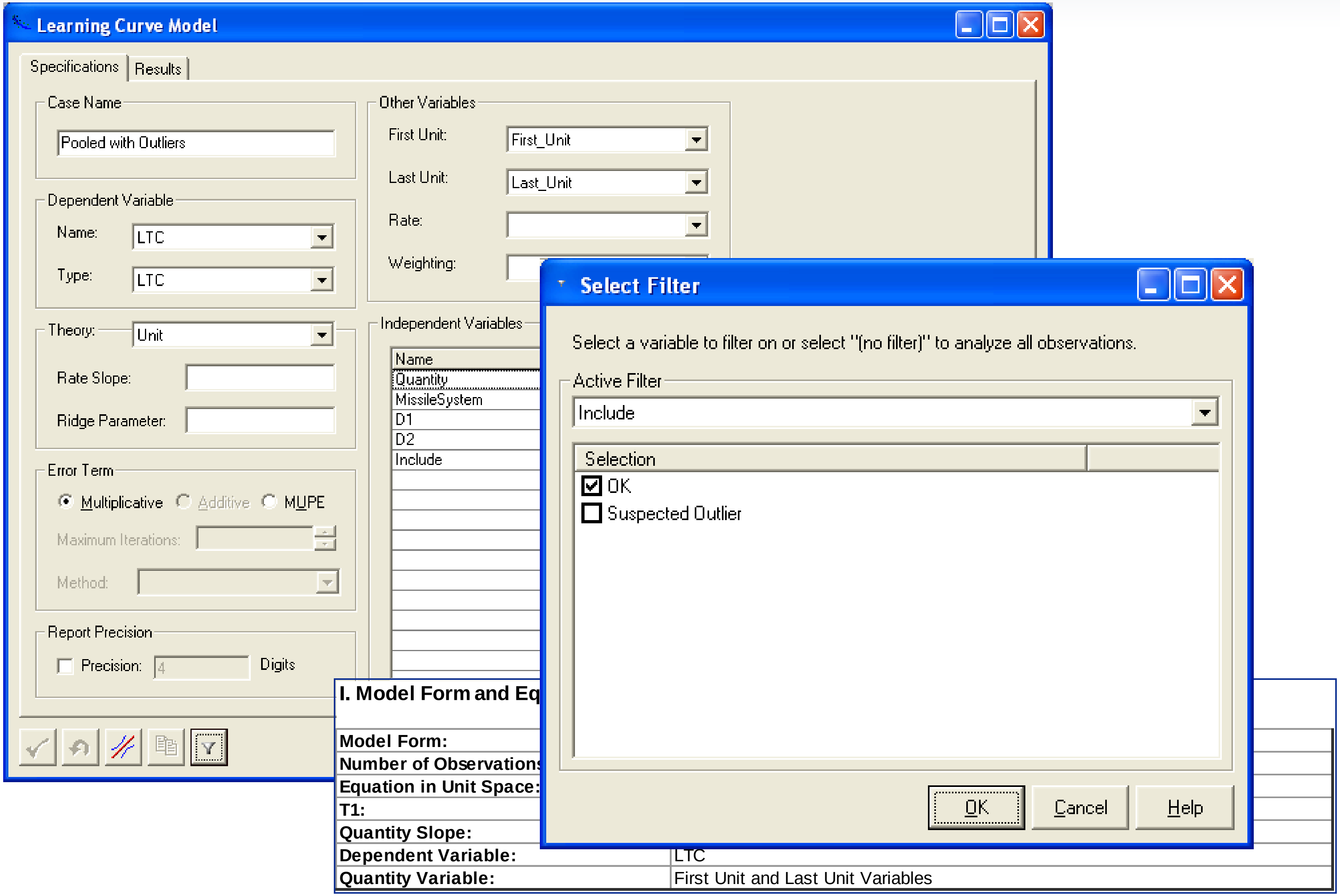
Outliers

Observations	Lot Total Cost	Quantity	First Unit of Lot	Last Unit of Lot	Missile System	Class of Missile Proxy 1 of 2	Class of Missile Proxy 2 of 2	Include
Variable ID	LTC	Quantity	First_Unit	Last_Unit	MissileSystem	D1	D2	Include
B-Missile Lot 1	38,094	80	1	80	B-Missile	0	0	OK
B-Missile Lot 2	108,843	300	81	380	B-Missile	0	0	OK
B-Missile Lot 3	95,421	300	381	680	B-Missile	0	0	OK
B-Missile Lot 4	181,016	620	681	1,300	B-Missile	0	0	OK
B-Missile Lot 5	284,094	1,000	1,301	2,300	B-Missile	0	0	OK
B-Missile Lot 6	502,042	2,000	2,301	4,300	B-Missile	0	0	OK
B-Missile Lot 7	435,818	2,000	4,301	6,300	B-Missile	0	0	OK
L-Missile Lot 1	45,838	50	1	50	L-Missile	1	0	OK
L-Missile Lot 2	73,711	100	51	150	L-Missile	1	0	OK
L-Missile Lot 3	157,126	250	151	400	L-Missile	1	0	OK
L-Missile Lot 4	257,549	500	401	900	L-Missile	1	0	OK
L-Missile Lot 5	215,310	500	901	1,400	L-Missile	1	0	OK
L-Missile Lot 6	203,333	500	1,401	1,900	L-Missile	1	0	OK
L-Missile Lot 7	191,125	500	1,901	2,400	L-Missile	1	0	OK
D-Missile Lot 1	568,734	200	1	200	D-Missile	0	1	Suspected Outlier
D-Missile Lot 2	1,217,274	500	201	700	D-Missile	0	1	Suspected Outlier
D-Missile Lot 3	2,677,707	1,200	701	1,900	D-Missile	0	1	OK
D-Missile Lot 4	4,557,980	2,200	1,901	4,100	D-Missile	0	1	OK
D-Missile Lot 5	4,235,452	2,200	4,101	6,300	D-Missile	0	1	OK
D-Missile Lot 6	3,921,465	2,000	6,301	8,300	D-Missile	0	1	OK
D-Missile Lot 7	3,162,458	1,700	8,301	10,000	D-Missile	0	1	OK

Fake Data

No need to delete rows to exclude observations from your regression. Add an “Include” column and assign meaningful labels to observations you may want to include/exclude during the course of analysis.

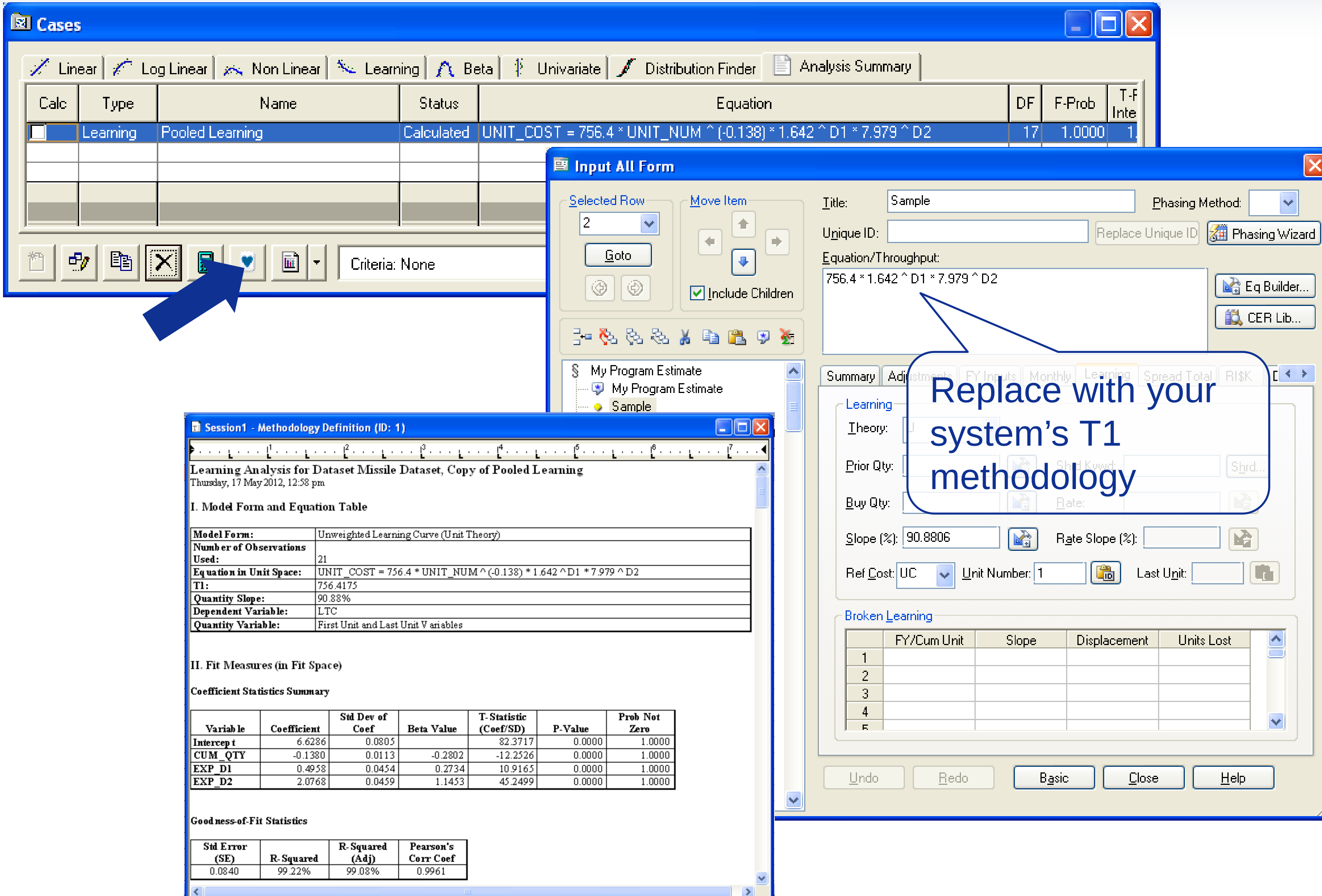
Outliers



Learning Curve Model
Specifications
Results
Case Name
Pooled with Outliers
Dependent Variable
Name:
LTC
Type:
LTC
Theory:
Unit
Rate Slope:
Ridge Parameter:
Error Term
Multiplicative
Additive
MUPE
Maximum Iterations:
Method:
Report Precision
Precision:
4
Digits
Other Variables
First Unit:
First_Unit
Last Unit:
Last_Unit
Rate:
Weighting:
Independent Variables
Name
Quantity
MissileSystem
D1
D2
Include
I. Model Form and Eq
Model Form:
Number of Observations
Equation in Unit Space:
T1:
Quantity Slope:
Dependent Variable:
LTC
Quantity Variable:
First Unit and Last Unit Variables
Select Filter
Select a variable to filter on or select "(no filter)" to analyze all observations.
Active Filter
Include
Selection
OK
Suspected Outlier
OK
Cancel
Help

Tricks and Tips
Results in ACEIT

Results In ACE



Disclaimers

Disclaimer

- **You are not limited to values of zero and one but there is no compelling reason to do otherwise**
  - The goal is only to represent the classes in the most expedient manner
  - Reducing final equations is unnecessarily complicated
  - Explaining why you chose values can be difficult to explain
  - Finally, if you have a measured or subjective value for the classes then consider using a straightforward independent variable
- **Non-intercept regression will run without a zero class**
  - But choosing no intercept changes your model considerably
  - The learning curve model treated by pooled regression, by definition, has an intercept

Summary

- **Regression: Use dummy variables to stratify CER data into subsets of distinct classes**
- **Pooled Regression: Use dummy variables to stratify learning curve data into distinct classes by system**
- ***r* classes require *r-1* dummy variables**
- **Keep in mind each dummy variable consumes a degree-of-freedom**
- **Remember to include a “zero class”**