Regression Analysis Using Statgraphics Centurion

Neil W. Polhemus, CTO, StatPoint Technologies, Inc.

Web site: www.statgraphics.com

STATPOINT
TECHNOLOGIES, INC.

Outline

- Regression Models
- Examples – Single X
  - Simple regression
  - Nonlinear models
  - Calibration
  - Comparison of regression lines
- Examples – Multiple X
  - Regression model selection (stepwise, all possible)
  - Logistic regression
  - Poisson regression

Regression Model Setup

- Dependent variable: Y
- Independent variable(s): $X_1, X_2, \ldots, X_k$
- Error term: ε

Model: $Y = f(X_1, X_2, \ldots, X_k) + \varepsilon$

Types of Regression Models (#1)

Procedure	Dependent variable	Independent variables
Simple Regression	continuous	1 continuous
Polynomial Regression	continuous	1 continuous
Box-Cox Transformations	continuous	1 continuous
Calibration Models	continuous	1 continuous
Comparison of Regression Lines	continuous	1 continuous and 1 categorical

Types of Regression Models (#2)

Procedure	Dependent variable	Independent variables
Multiple Regression	continuous	2+ continuous
Regression Model Selection	continuous	2+ continuous
Nonlinear Regression	continuous	1+ continuous
Ridge Regression	continuous	2+ continuous
Partial Least Squares	continuous	2+ continuous
General Linear Models	1+ continuous	2+ continuous or categorical variables

Types of Regression Models (#3)

Procedure	Dependent variable	Independent variables
Logistic Regression	proportions	1+ continuous or categorical
Probit Analysis	proportions	1+ continuous or categorical
Poisson Regression	counts	1+ continuous or categorical
Negative Binomial Regression	counts	1+ continuous or categorical
Life Data - Parametric Models	failure times	1+ continuous or categorical

Example 1: Stability study

C:\DocData16\nonlin.sgd

	weeks	chlorine
	weeks since production	percent available
1	8	0.49
2	8	0.49
3	10	0.48
4	10	0.47
5	10	0.48
6	10	0.47
7	12	0.46
8	12	0.46
9	12	0.45
10	12	0.43
11	14	0.45
12	14	0.43
13	14	0.43
14	16	0.44
15	16	0.43
16	16	0.43
17	18	0.46
18	18	0.45
19	20	0.42

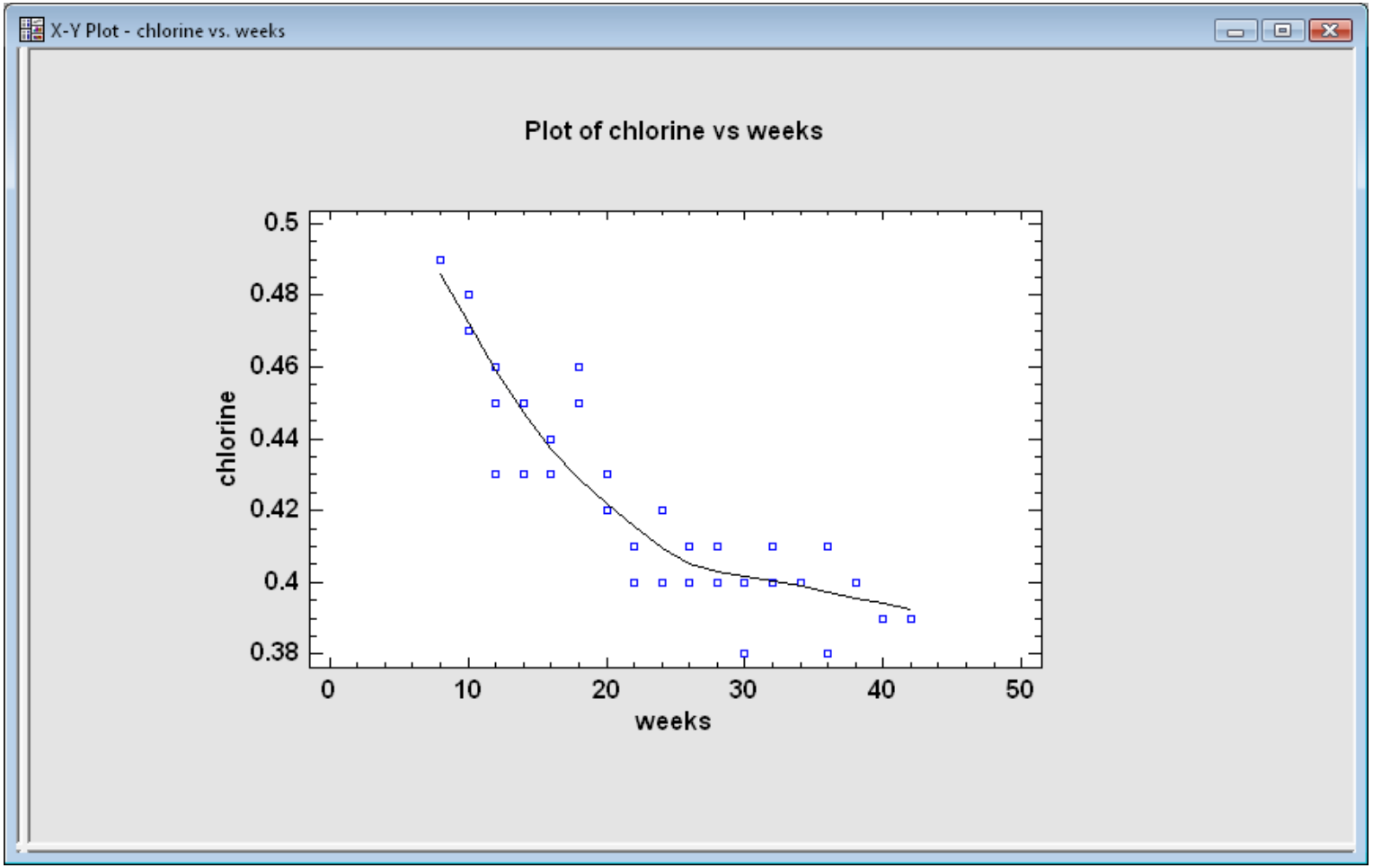
nonlin B C

Y: percent of available chlorine

X: number of weeks since production

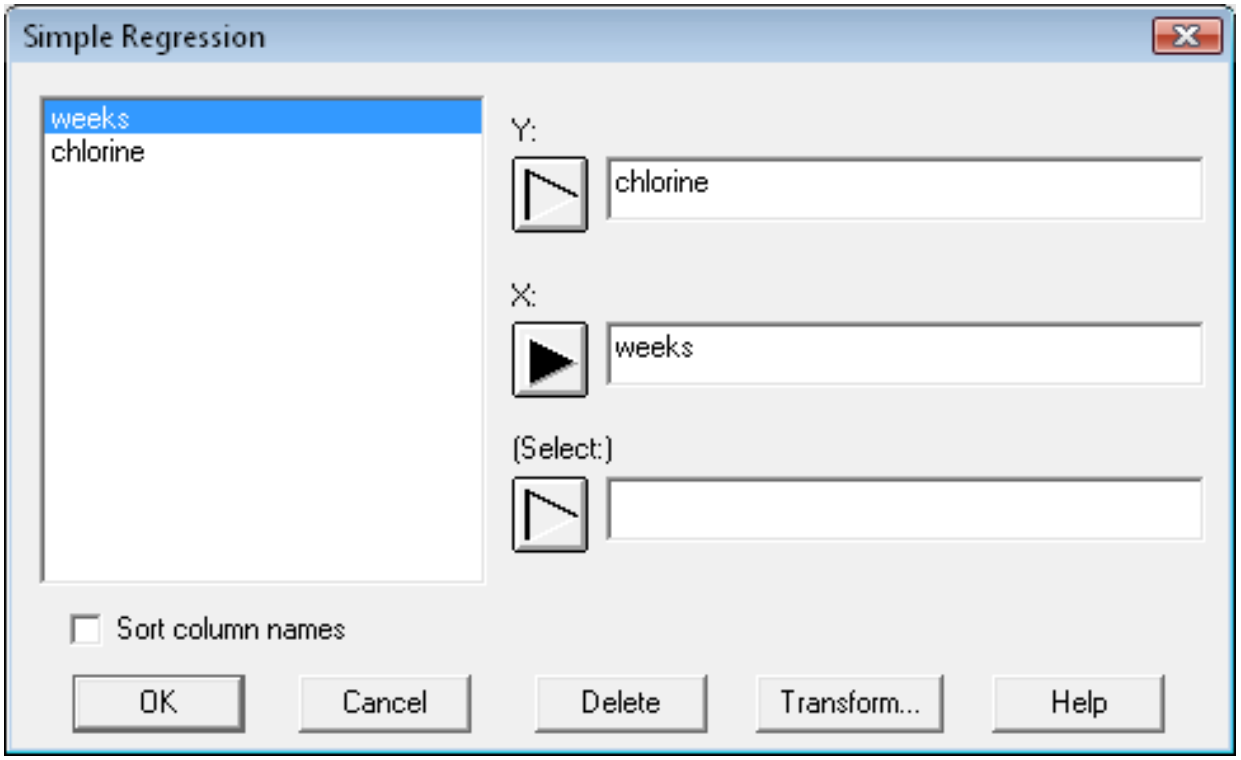
Lower acceptable limit for Y: 0.40

X-Y Scatterplot with Smooth



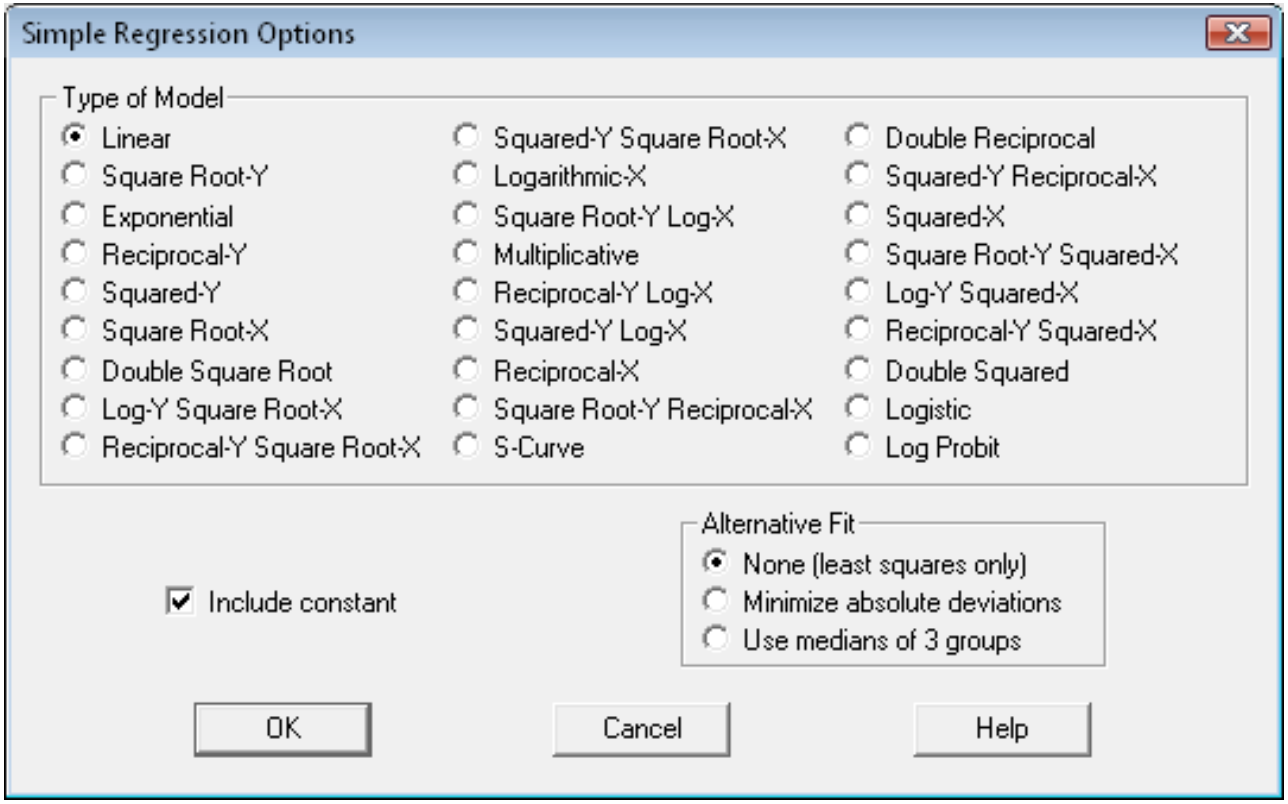
STATPOINT TECHNOLOGIES, INC.

Simple Regression



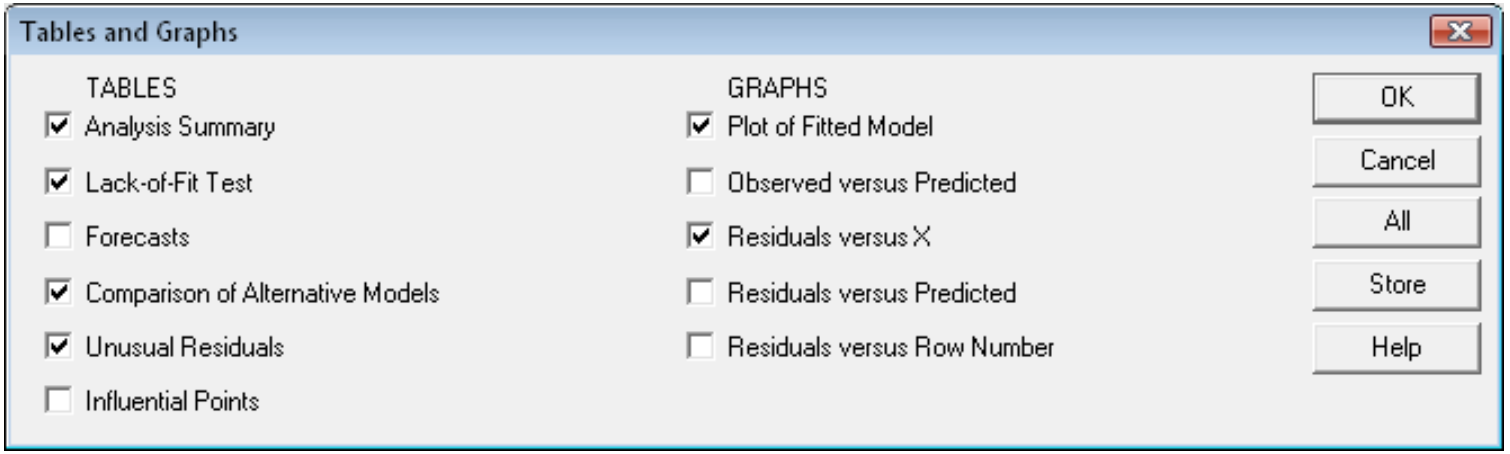
STATPOINT TECHNOLOGIES, INC.

Analysis Options



STATPOINT TECHNOLOGIES, INC.

Tables and Graphs



STATPOINT TECHNOLOGIES, INC.

Analysis Window

Simple Regression - chlorine vs. weeks

Simple Regression - chlorine vs. weeks

Dependent variable: chlorine (percent available)
Independent variable: weeks (weeks since production)
Linear model: Y = a + b*X

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*

Analysis of Variance with Lack-of-Fit

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	0.0295587	1	0.0295587	124.88	0.0000
Residual	0.00994133	42	0.000236698		
Lack-of-Fit	0.00757467	16	0.000473417	5.20	0.0001
Pure Error	0.00236667	26	0.0000910256		
Total (Corr.)	0.0395	43			

Comparison of Alternative Models

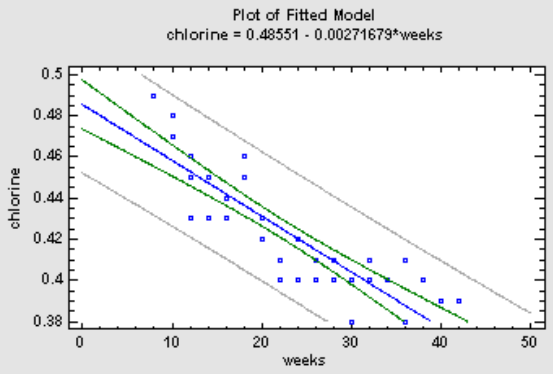
Model	*Correlation*	*R-Squared*
Squared-Y reciprocal-X	0.9367	87.75%
Reciprocal-X	0.9333	87.11%
Square root-Y reciprocal-X	0.9312	86.71%
S-curve model	0.9288	86.27%
Double reciprocal	-0.9233	85.25%
Reciprocal-Y logarithmic-X	0.9219	84.99%

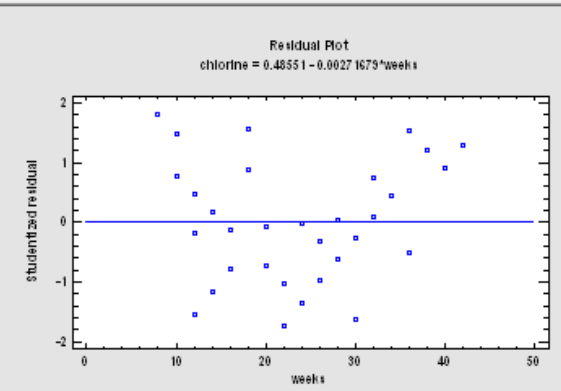
Unusual Residuals

			Predicted		*Studentized*
Row	*X*	*Y*	*Y*	*Residual*	*Residual*

The StatAdvisor

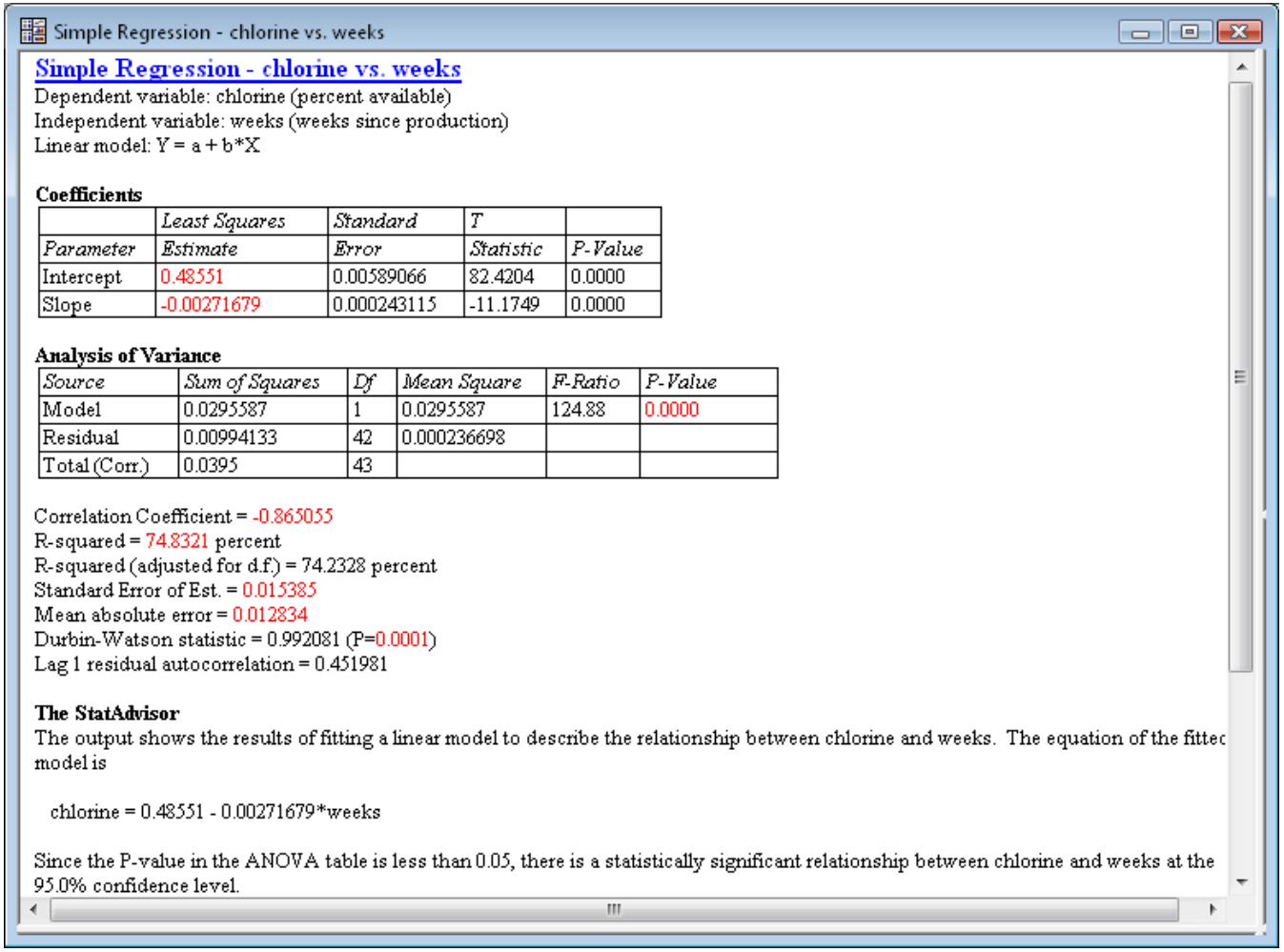
The table of unusual residuals lists all observations which have Studentized residuals greater than 2 residuals measure how many standard deviations each observed value of chlorine deviates from a m that observation. In this case, there are no Studentized residuals greater than 2.





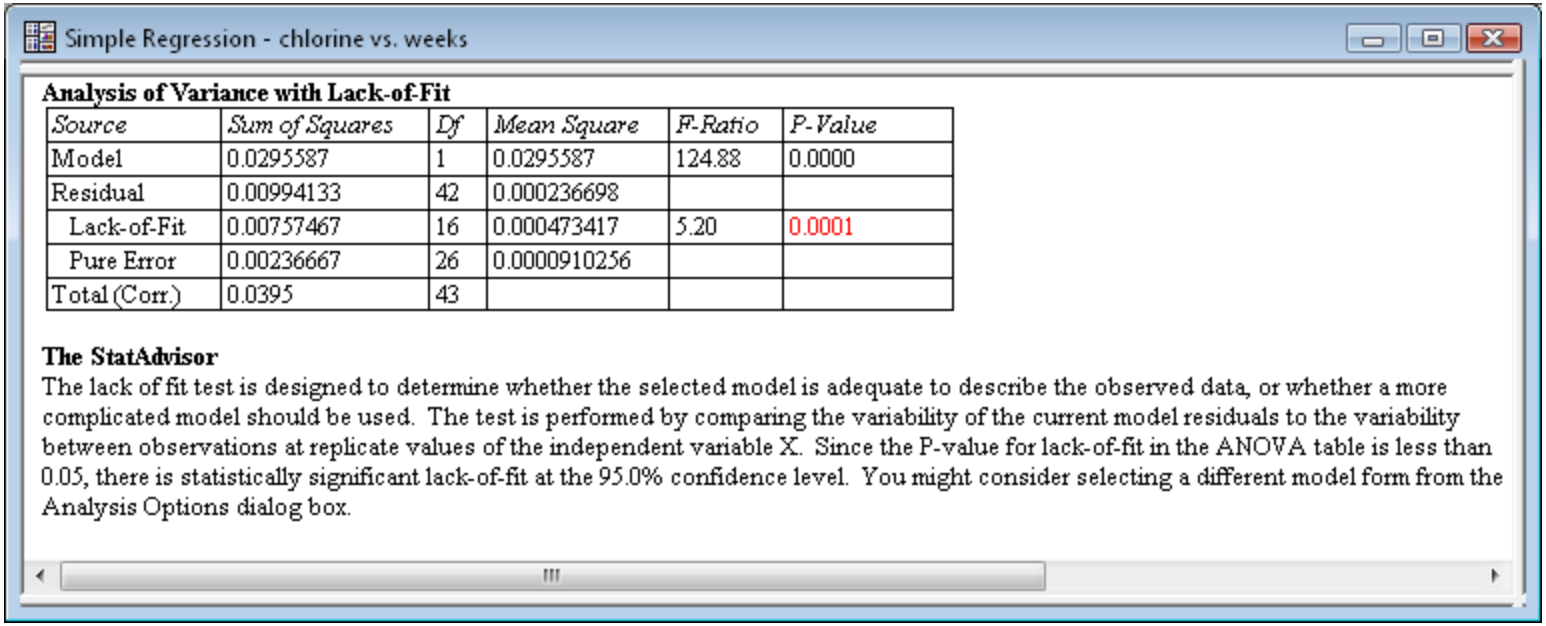
STATPOINT TECHNOLOGIES, INC.

Analysis Summary



STATPOINT TECHNOLOGIES, INC.

Lack-of-Fit Test


Simple Regression - chlorine vs. weeks

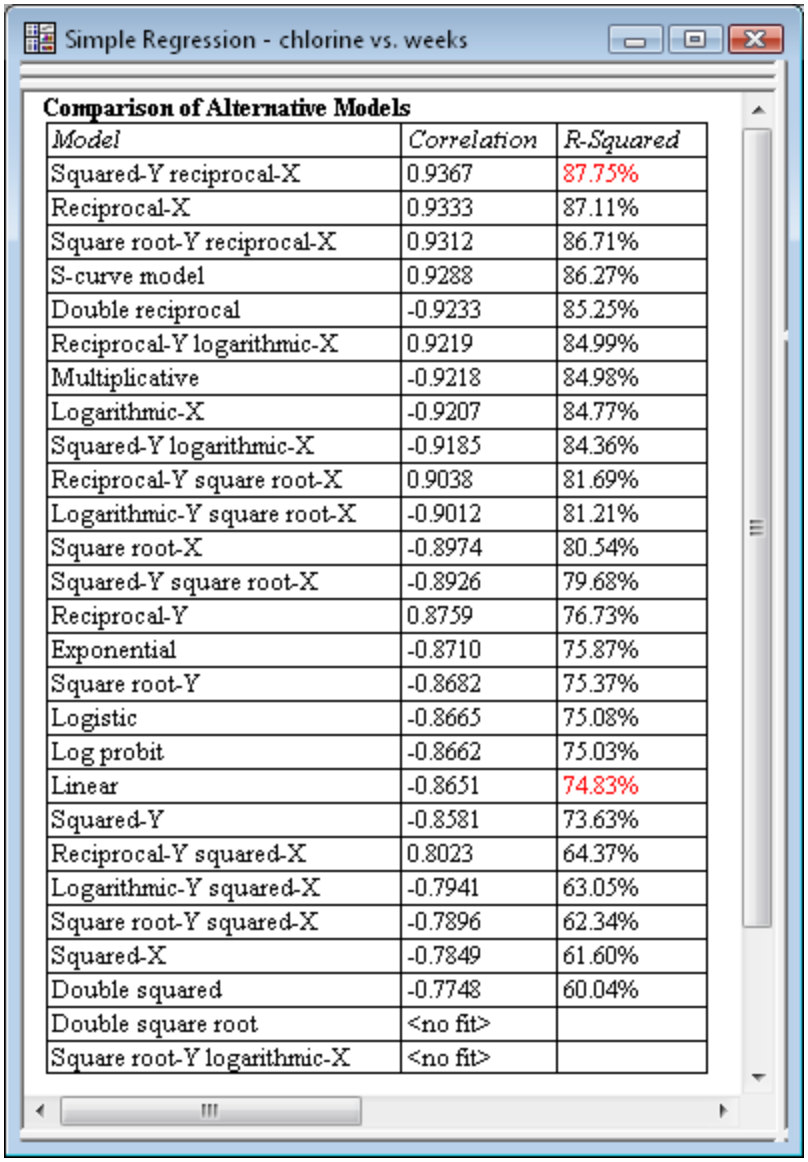
Analysis of Variance with Lack-of-Fit

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	0.0295587	1	0.0295587	124.88	0.0000
Residual	0.00994133	42	0.000236698		
Lack-of-Fit	0.00757467	16	0.000473417	5.20	0.0001
Pure Error	0.00236667	26	0.0000910256		
Total (Corr.)	0.0395	43			

The StatAdvisor

The lack of fit test is designed to determine whether the selected model is adequate to describe the observed data, or whether a more complicated model should be used. The test is performed by comparing the variability of the current model residuals to the variability between observations at replicate values of the independent variable X. Since the P-value for lack-of-fit in the ANOVA table is less than 0.05, there is statistically significant lack-of-fit at the 95.0% confidence level. You might consider selecting a different model form from the Analysis Options dialog box.

Comparison of Alternative Models

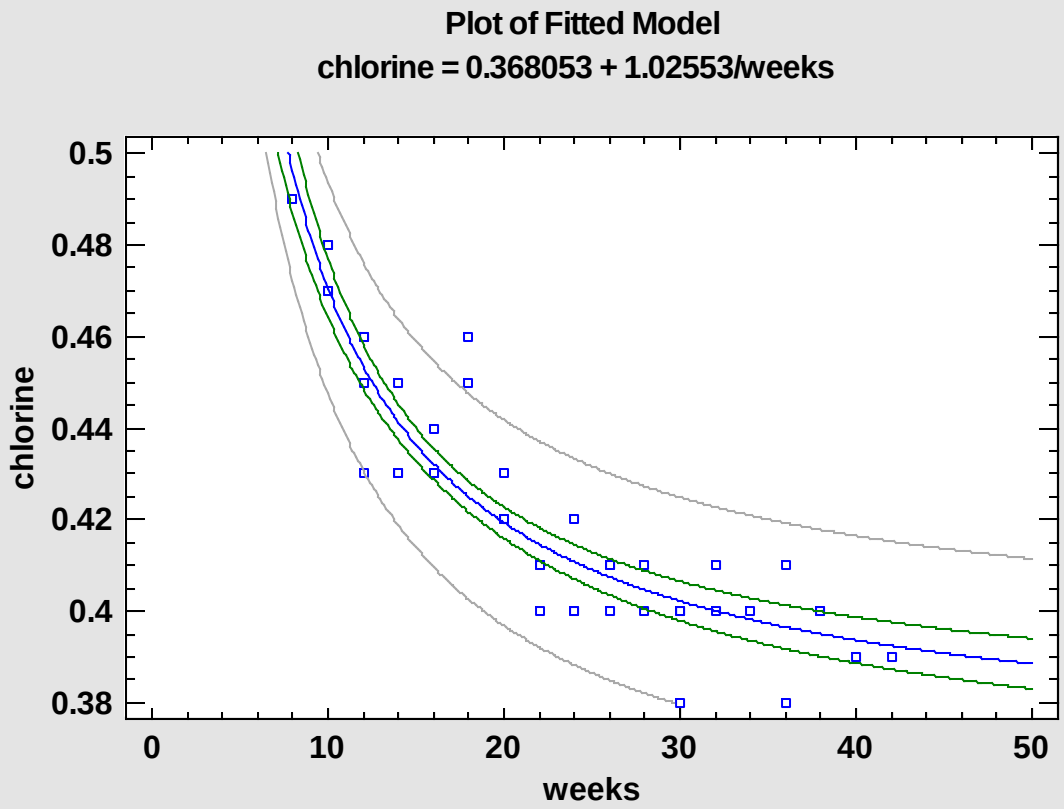


Comparison of Alternative Models

Model	*Correlation*	*R-Squared*
Squared-Y reciprocal-X	0.9367	87.75%
Reciprocal-X	0.9333	87.11%
Square root-Y reciprocal-X	0.9312	86.71%
S-curve model	0.9288	86.27%
Double reciprocal	-0.9233	85.25%
Reciprocal-Y logarithmic-X	0.9219	84.99%
Multiplicative	-0.9218	84.98%
Logarithmic-X	-0.9207	84.77%
Squared-Y logarithmic-X	-0.9185	84.36%
Reciprocal-Y square root-X	0.9038	81.69%
Logarithmic-Y square root-X	-0.9012	81.21%
Square root-X	-0.8974	80.54%
Squared-Y square root-X	-0.8926	79.68%
Reciprocal-Y	0.8759	76.73%
Exponential	-0.8710	75.87%
Square root-Y	-0.8682	75.37%
Logistic	-0.8665	75.08%
Log probit	-0.8662	75.03%
Linear	-0.8651	74.83%
Squared-Y	-0.8581	73.63%
Reciprocal-Y squared-X	0.8023	64.37%
Logarithmic-Y squared-X	-0.7941	63.05%
Square root-Y squared-X	-0.7896	62.34%
Squared-X	-0.7849	61.60%
Double squared	-0.7748	60.04%
Double square root	<no fit>	
Square root-Y logarithmic-X	<no fit>	

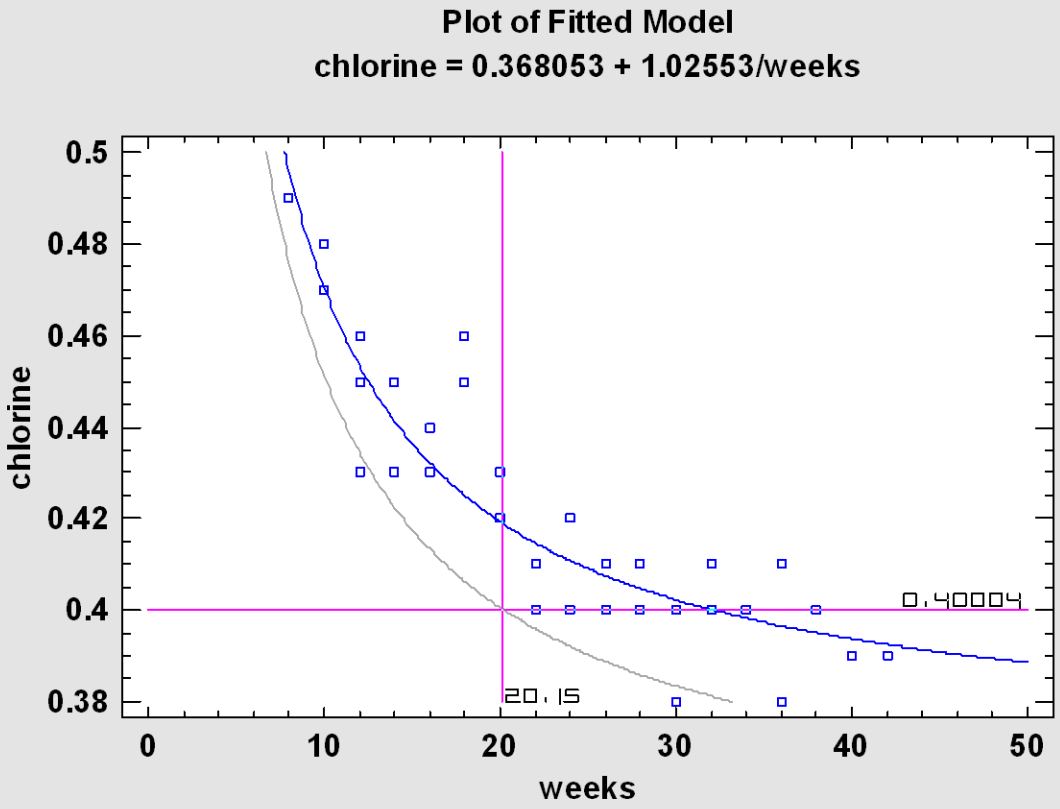
STATPOINT TECHNOLOGIES, INC.

Fitted Reciprocal-X Model



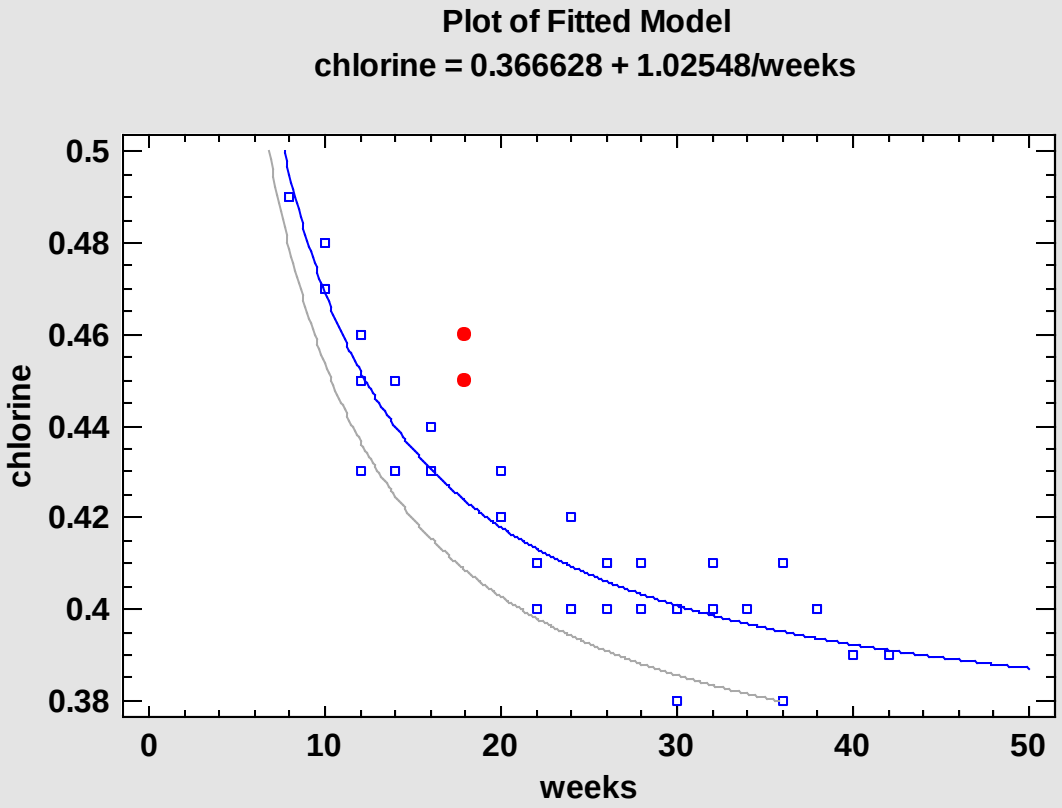
STATPOINT TECHNOLOGIES, INC.

Lower 95% Prediction Limit



STATPOINT TECHNOLOGIES, INC.

Outlier Removal



STATPOINT TECHNOLOGIES, INC.

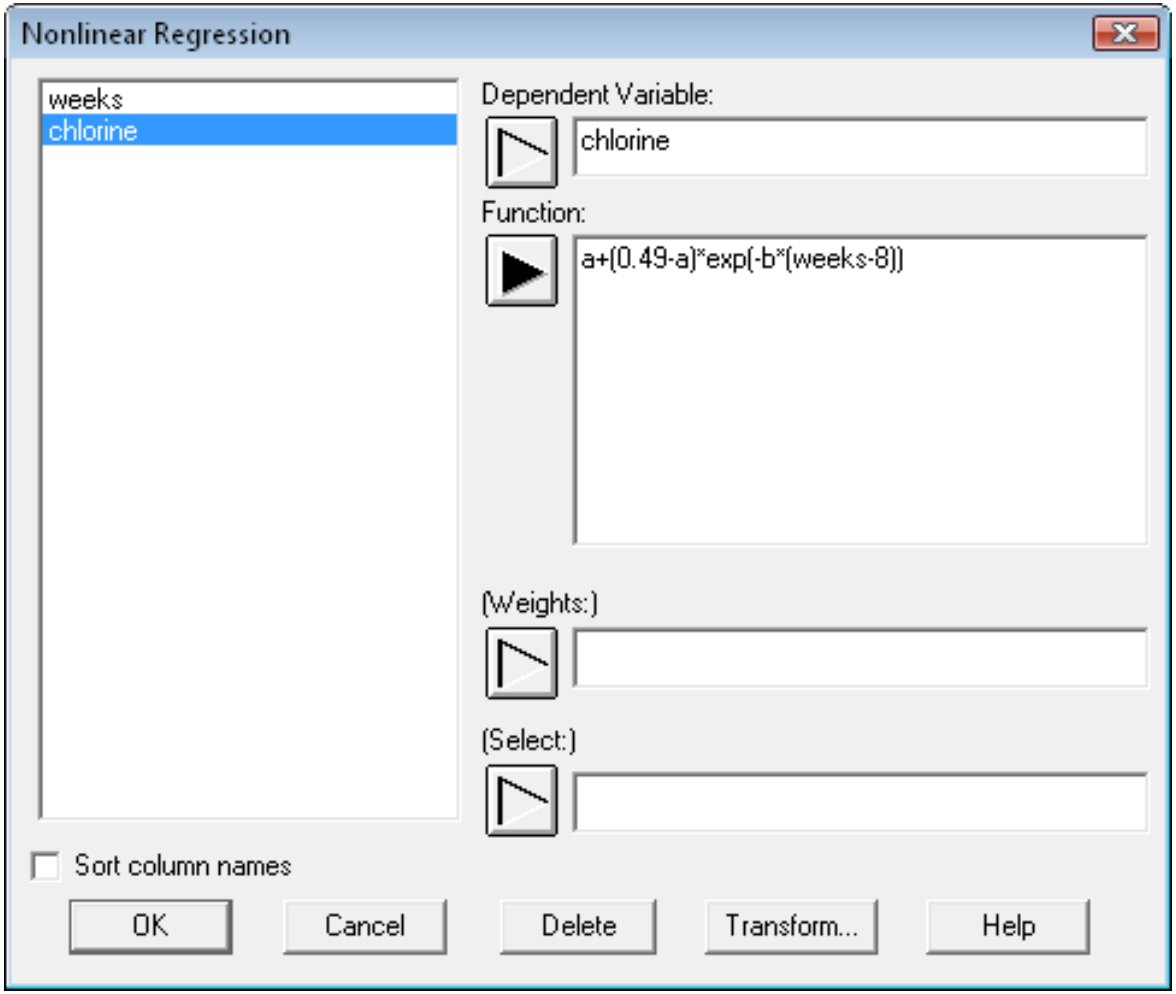
Example 2: Nonlinear Regression

Draper and Smith in Applied Regression Analysis suggest fitting a model of the form

$$Y = a + (0.49-a)\exp[-b(x-8)]$$

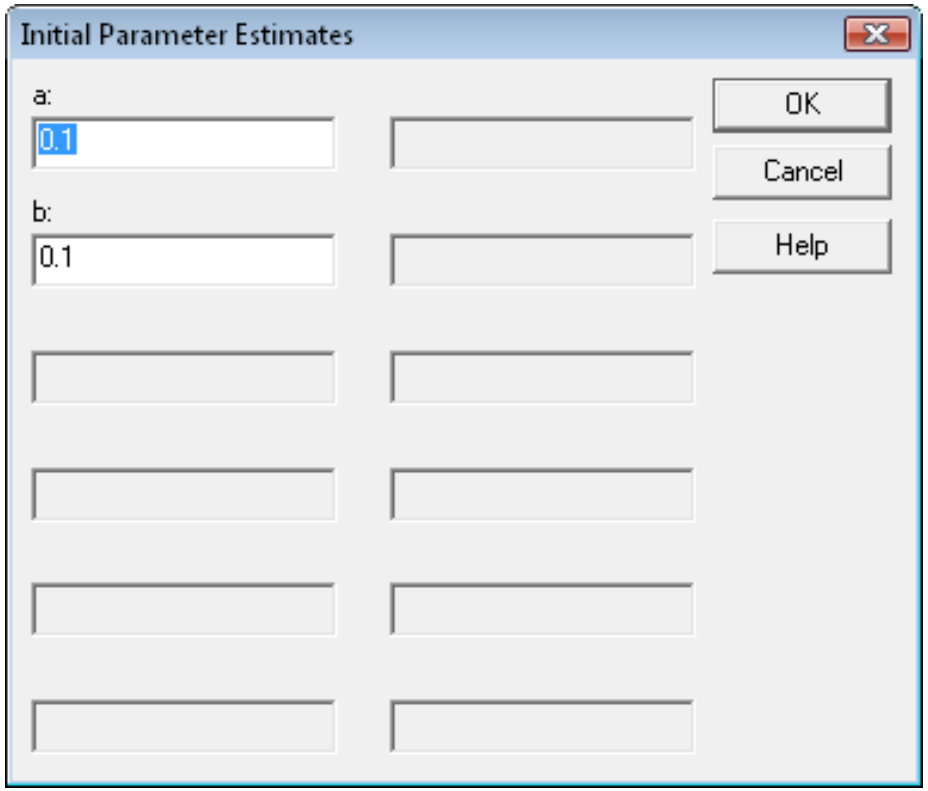
Since the model is nonlinear in the parameters, it requires a search procedure to find the best solution.

Data Input Dialog Box



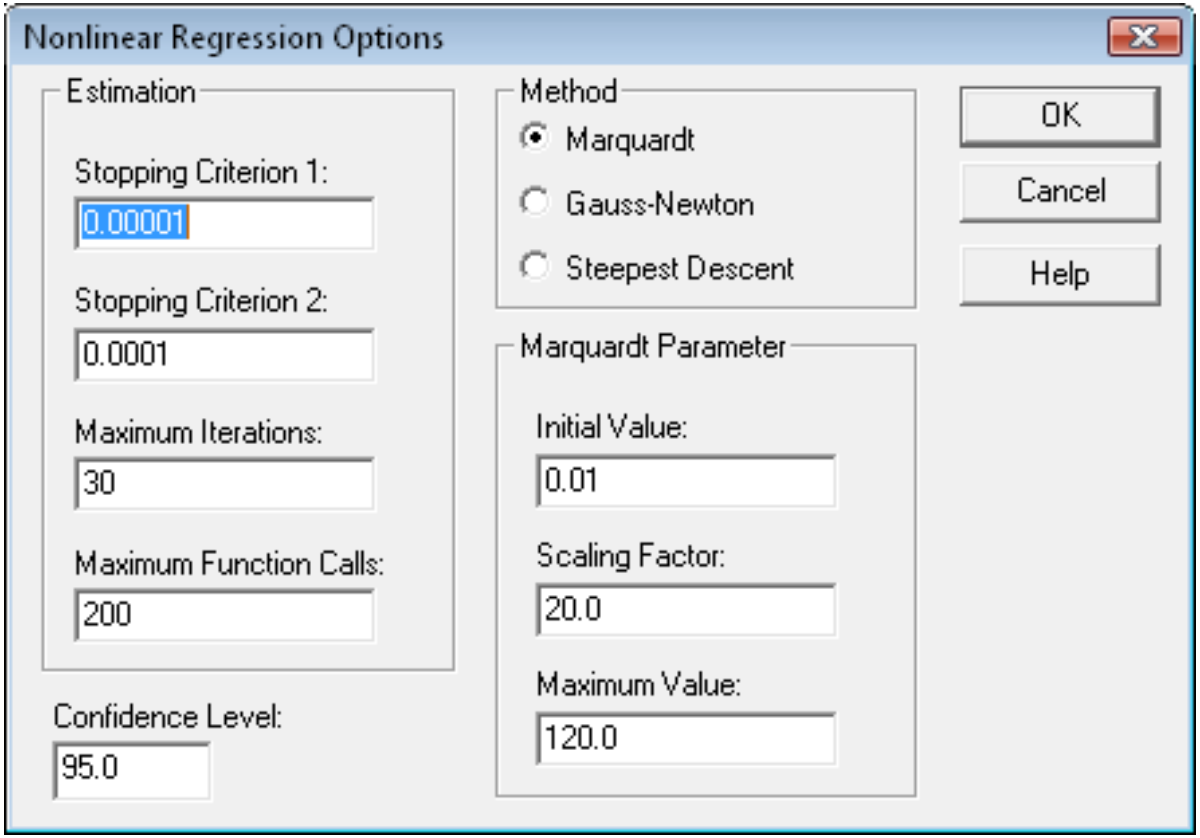
STATPOINT TECHNOLOGIES, INC.

Initial Parameter Estimates



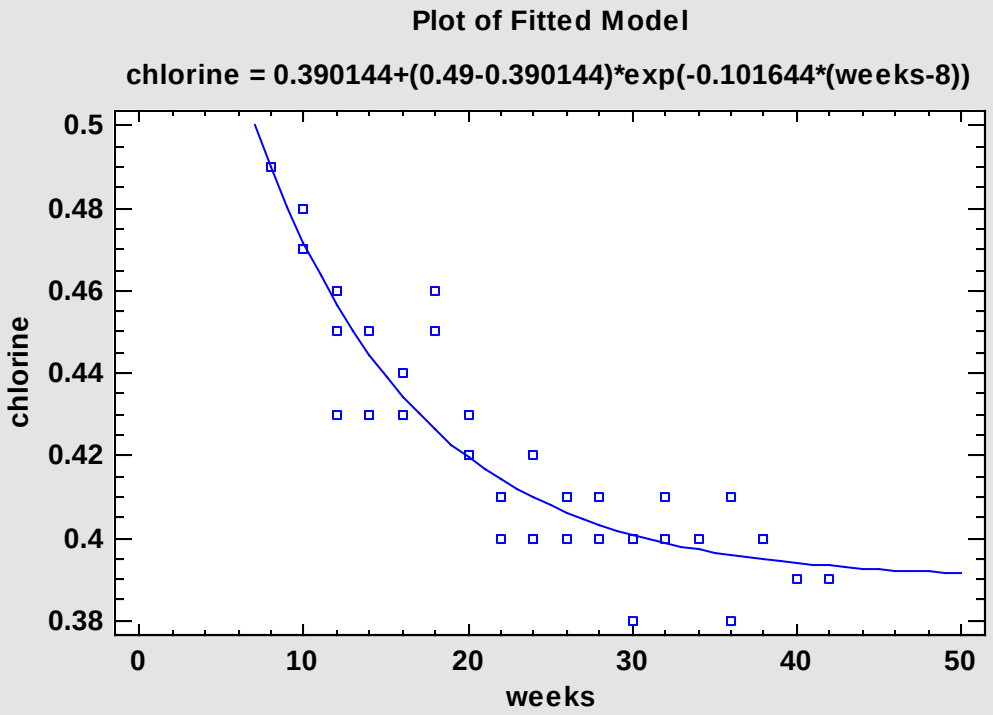
STATPOINT TECHNOLOGIES, INC.

Analysis Options



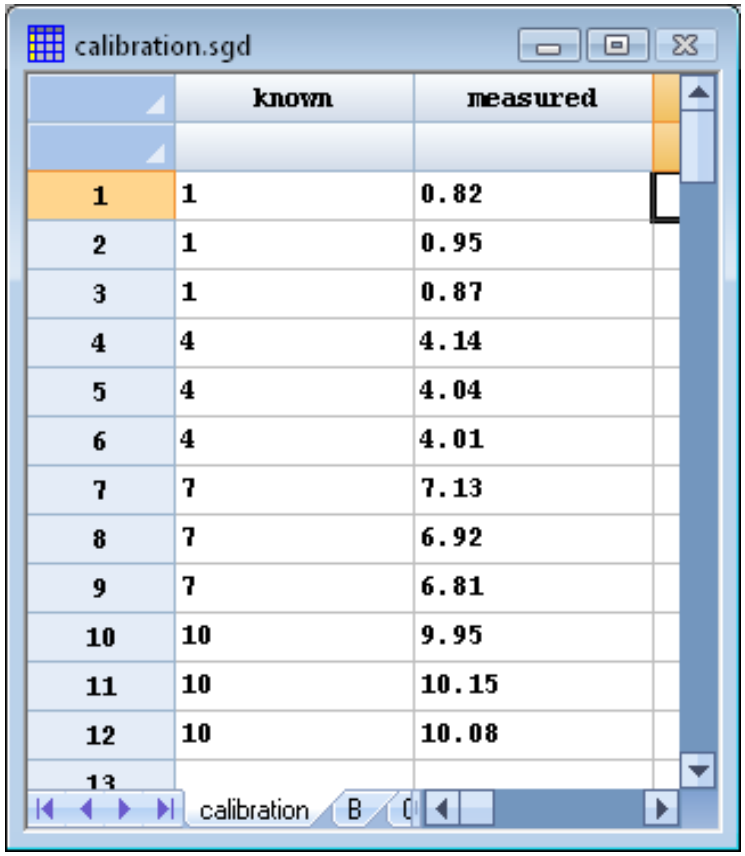
STATPOINT TECHNOLOGIES, INC.

Plot of Fitted Model



STATPOINT TECHNOLOGIES, INC.

Example 3: Calibration



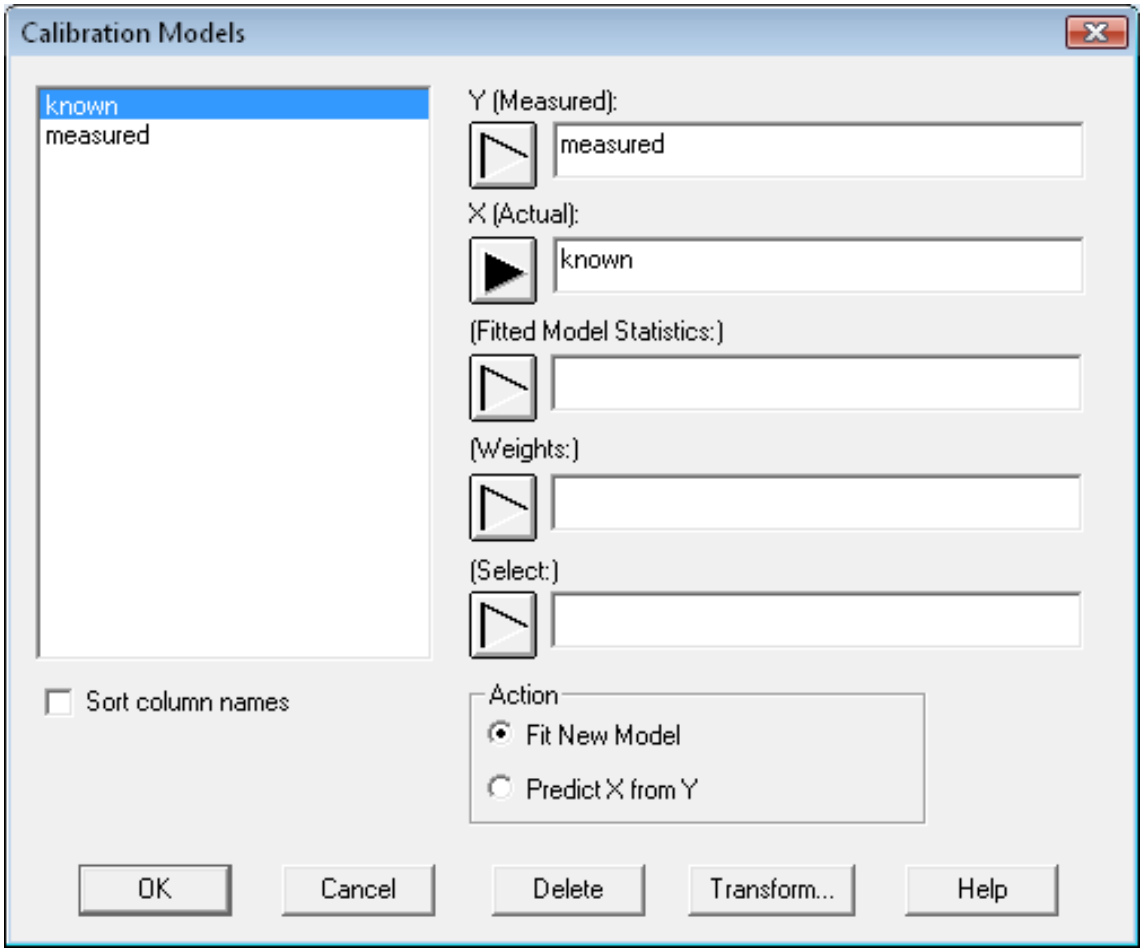
The general calibration problem is that of determining the likely value of X given an observed value of Y.

Typically: X = item characteristic, Y = measured value

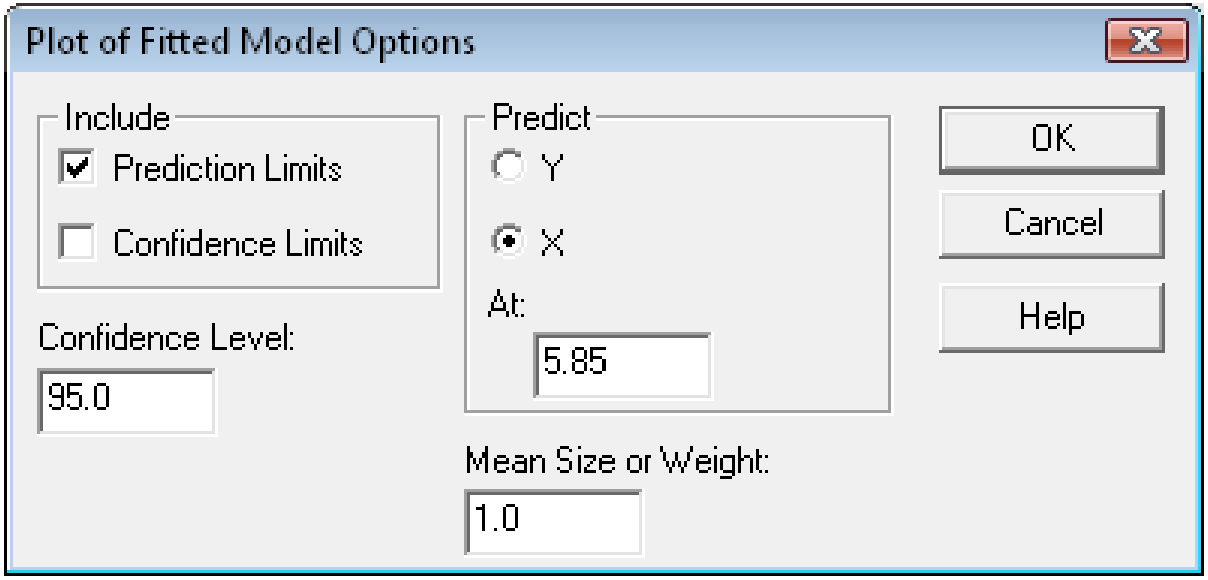
Step 1: Build a regression model using samples with known values of X ("golden samples").

Step 2: For another sample with unknown X, predict X from Y.

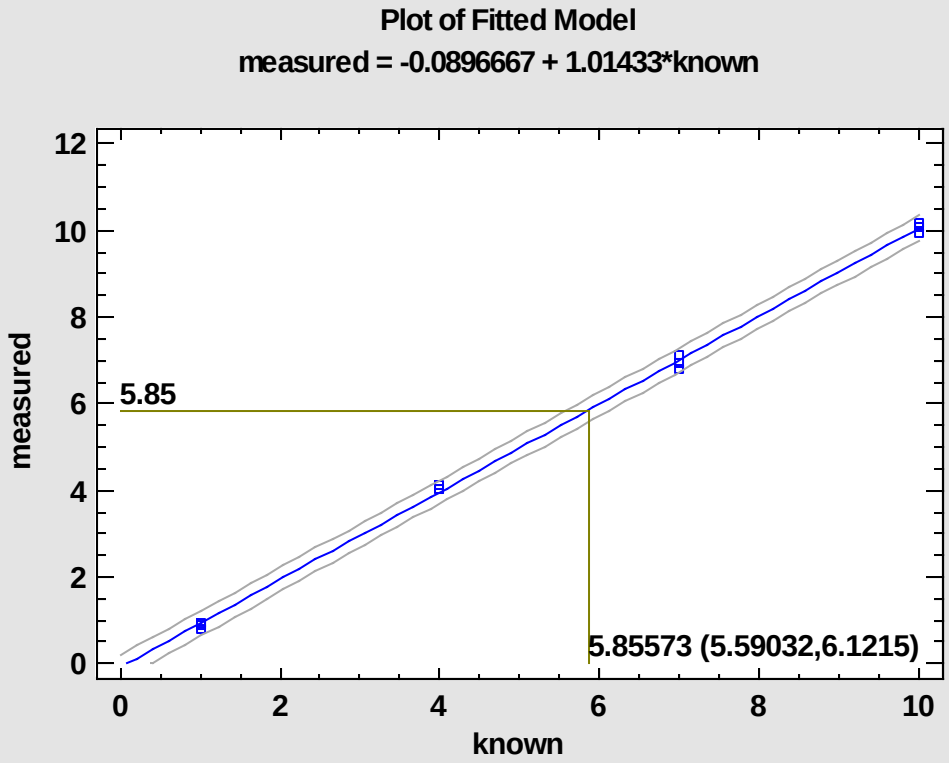
Data Input Dialog Box



Reverse Prediction



Plot of Fitted Model



Example 4: Comparison of Regression Lines

C:\DocData16\soap.sgd

	Line	Scrap	Speed
1	1	218	100
2	1	248	125
3	1	360	220
4	1	351	205
5	1	470	300
6	1	394	255
7	1	332	225
8	1	321	175
9	1	410	270
10	1	260	170
11	1	241	155
12	1	331	190
13	1	275	140
14	1	425	290
15	1	367	265
16	2	140	105
17	2	277	215

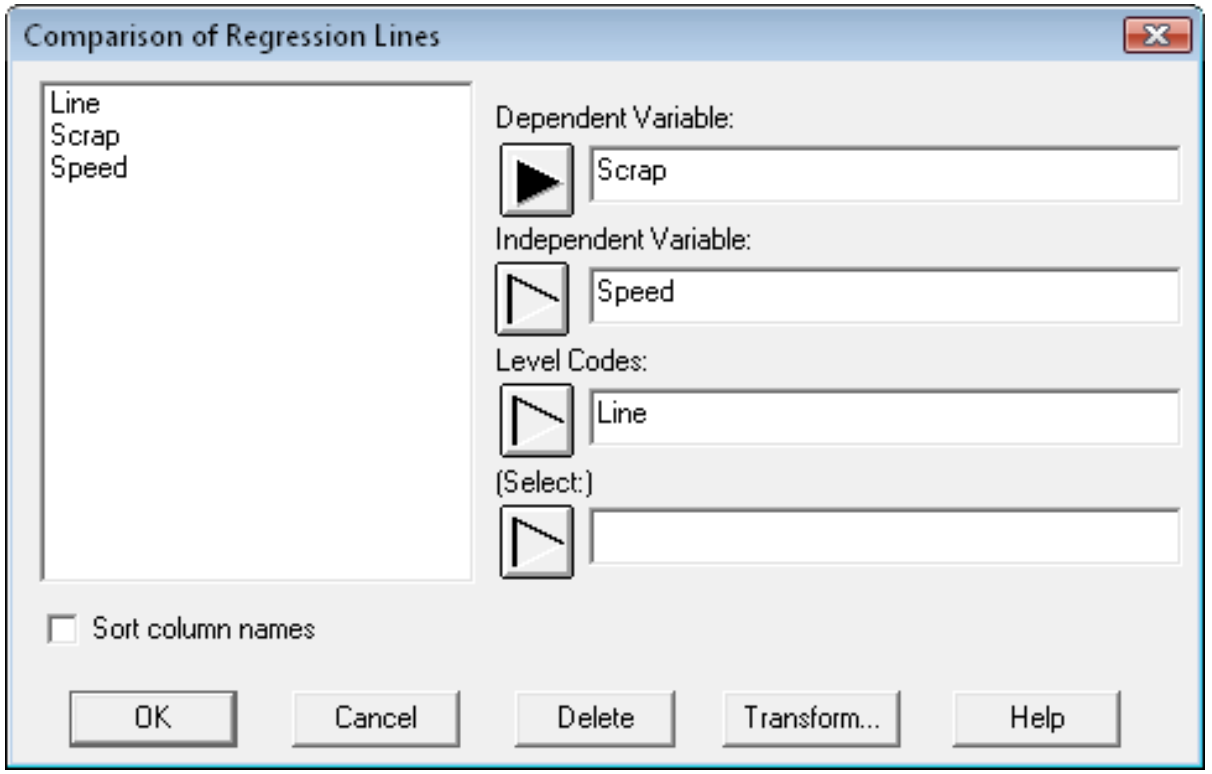
soap B C

Y: amount of scrap produced

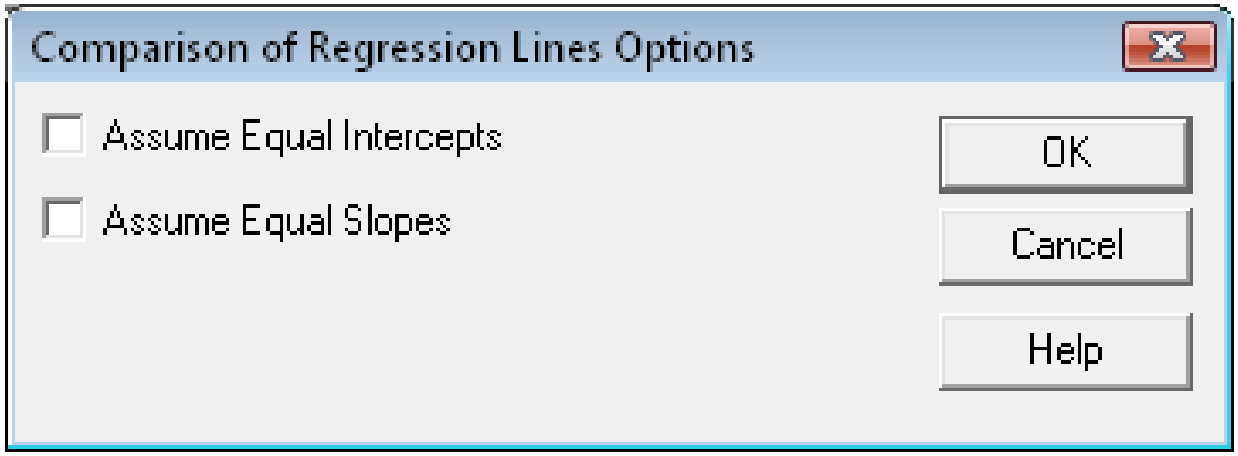
X: production line speed

Levels: line number

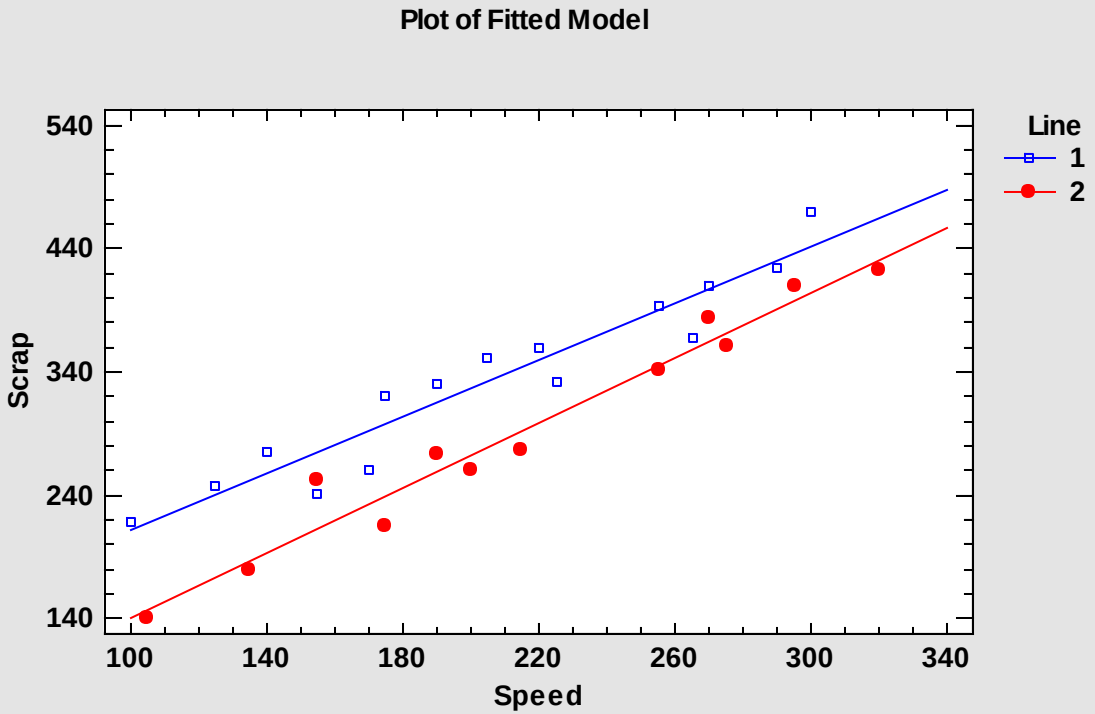
Data Input Dialog Box



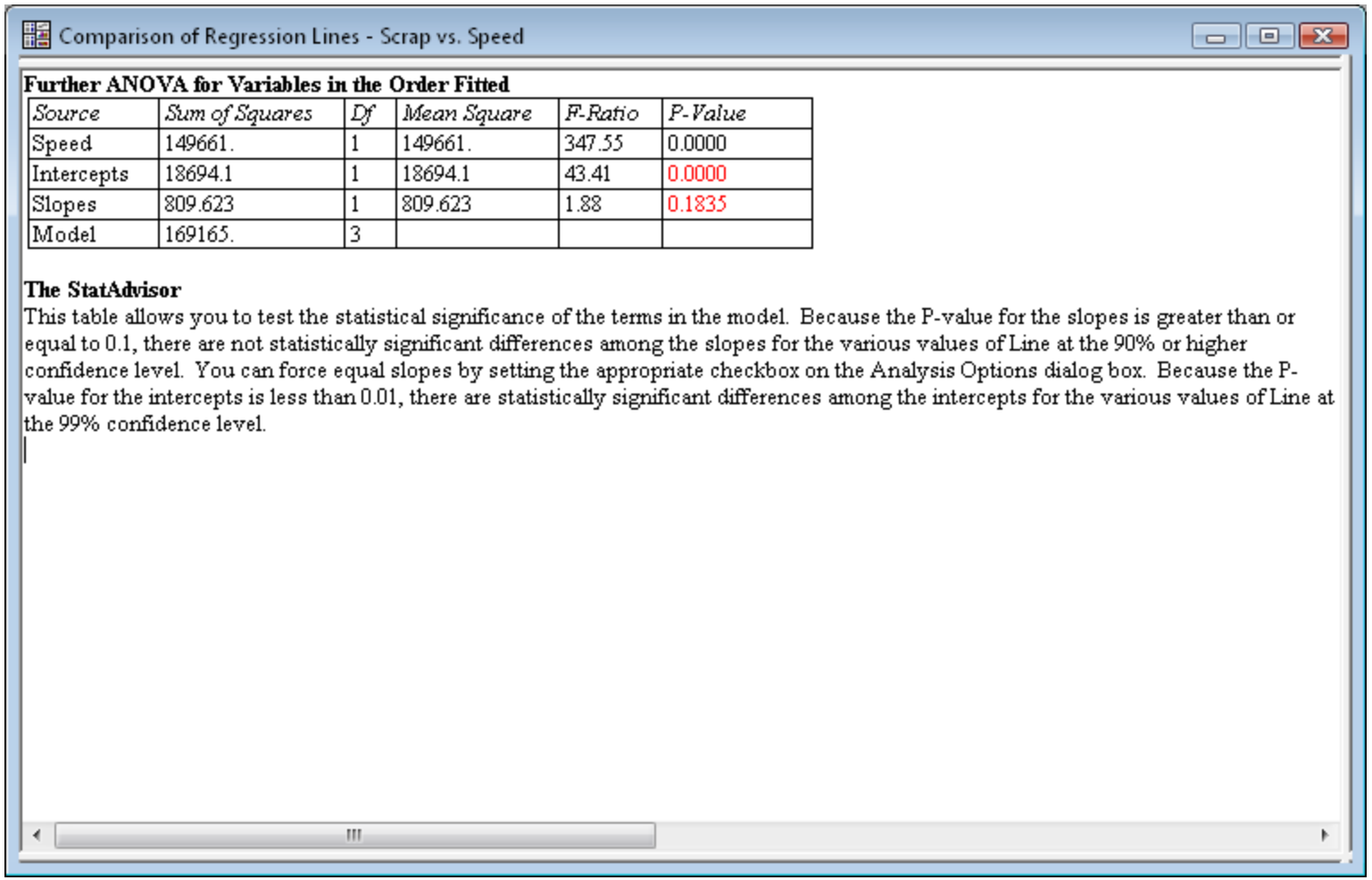
Analysis Options



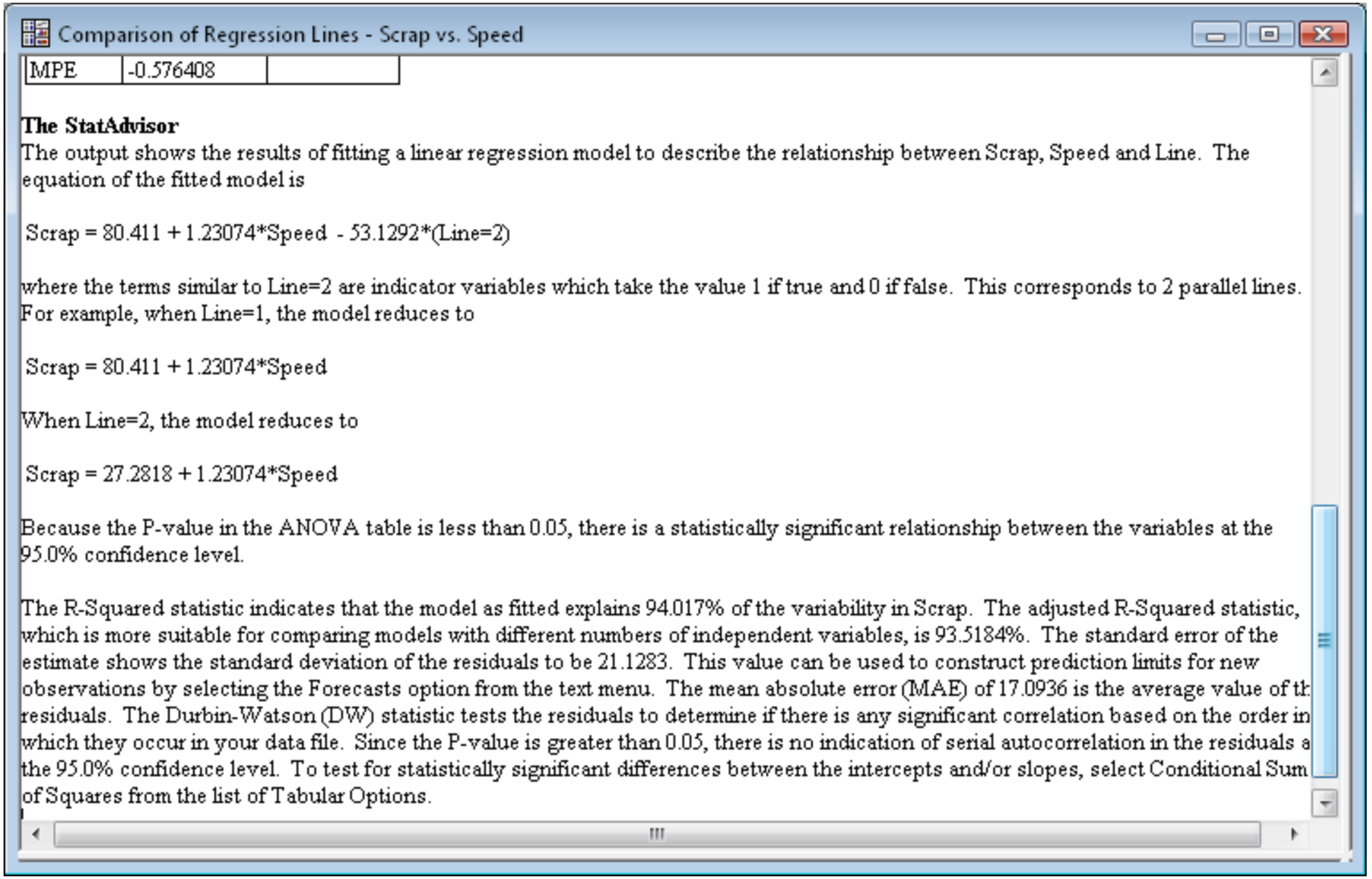
Plot of Fitted Model



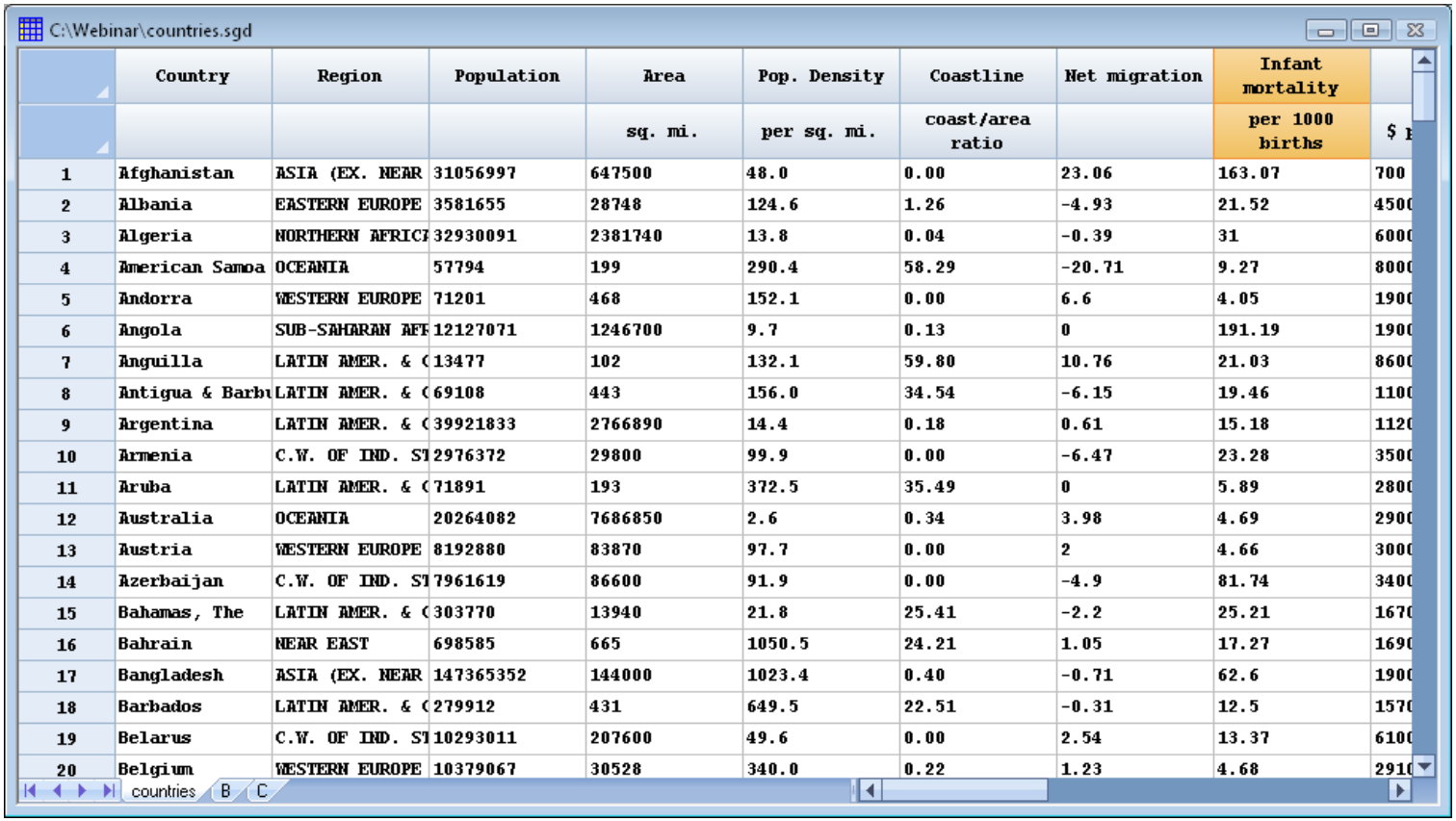
Significance Tests



Parallel Slope Model



Example 5: Multiple Regression



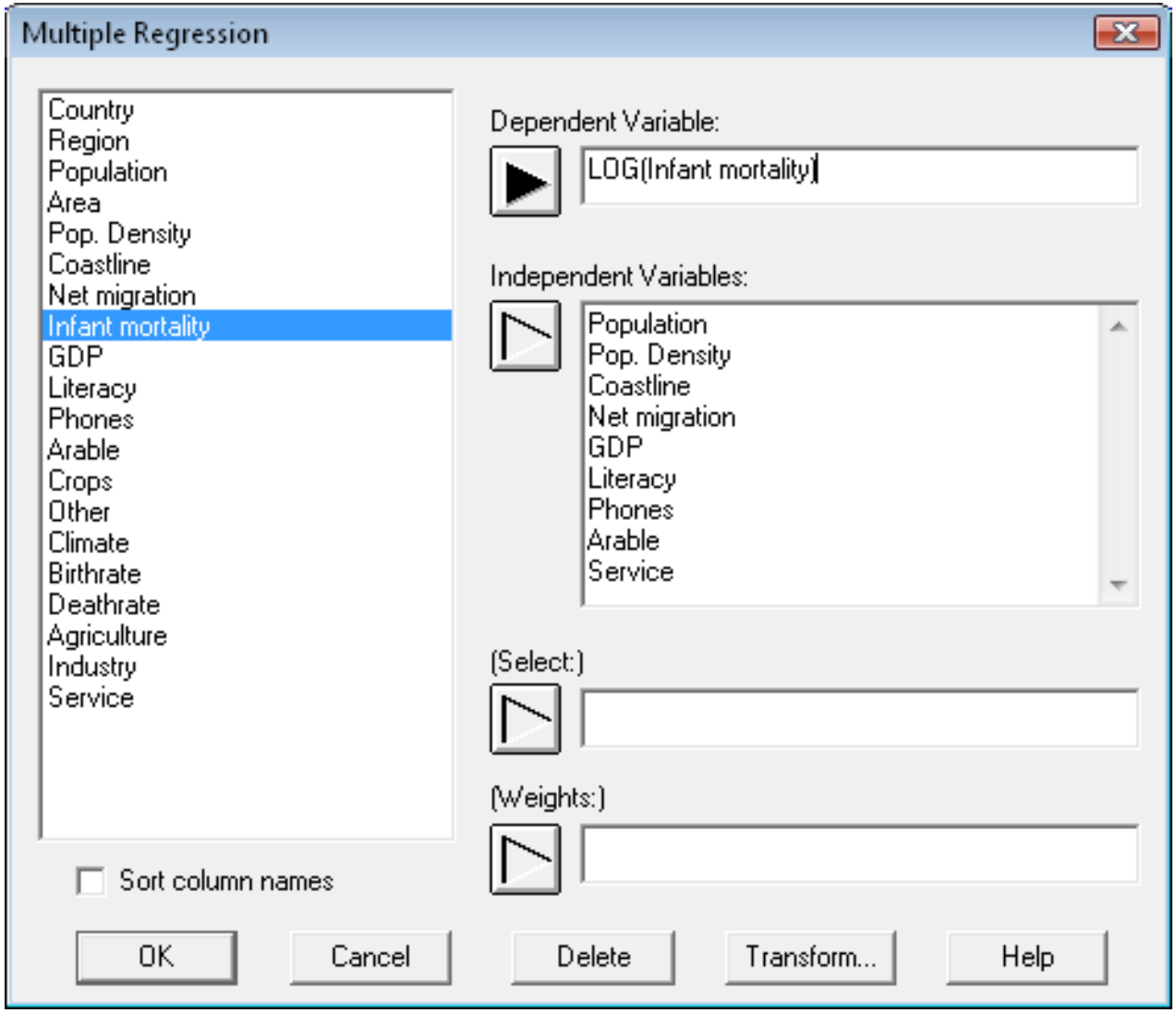
C:\Webinar\countries.sgd

	Country	Region	Population	Area	Pop. Density	Coastline	Net migration	Infant mortality	
				sq. mi.	per sq. mi.	coast/area ratio		per 1000 births	$ I
1	Afghanistan	ASIA (EX. NEAR	31056997	647500	48.0	0.00	23.06	163.07	700
2	Albania	EASTERN EUROPE	3581655	28748	124.6	1.26	-4.93	21.52	4500
3	Algeria	NORTHERN AFRIC	32930091	2381740	13.8	0.04	-0.39	31	6000
4	American Samoa	OCEANIA	57794	199	290.4	58.29	-20.71	9.27	8000
5	Andorra	WESTERN EUROPE	71201	468	152.1	0.00	6.6	4.05	1900
6	Angola	SUB-SAHARAN AF	12127071	1246700	9.7	0.13	0	191.19	1900
7	Anguilla	LATIN AMER. & (	13477	102	132.1	59.80	10.76	21.03	8600
8	Antigua & Barbu	LATIN AMER. & (	69108	443	156.0	34.54	-6.15	19.46	1100
9	Argentina	LATIN AMER. & (	39921833	2766890	14.4	0.18	0.61	15.18	1120
10	Armenia	C.W. OF IND. ST	2976372	29800	99.9	0.00	-6.47	23.28	3500
11	Aruba	LATIN AMER. & (	71891	193	372.5	35.49	0	5.89	2800
12	Australia	OCEANIA	20264082	7686850	2.6	0.34	3.98	4.69	2900
13	Austria	WESTERN EUROPE	8192880	83870	97.7	0.00	2	4.66	3000
14	Azerbaijan	C.W. OF IND. ST	7961619	86600	91.9	0.00	-4.9	81.74	3400
15	Bahamas, The	LATIN AMER. & (	303770	13940	21.8	25.41	-2.2	25.21	1670
16	Bahrain	NEAR EAST	698585	665	1050.5	24.21	1.05	17.27	1690
17	Bangladesh	ASIA (EX. NEAR	147365352	144000	1023.4	0.40	-0.71	62.6	1900
18	Barbados	LATIN AMER. & (	279912	431	649.5	22.51	-0.31	12.5	1570
19	Belarus	C.W. OF IND. ST	10293011	207600	49.6	0.00	2.54	13.37	6100
20	Belgium	WESTERN EUROPE	10379067	30528	340.0	0.22	1.23	4.68	2910

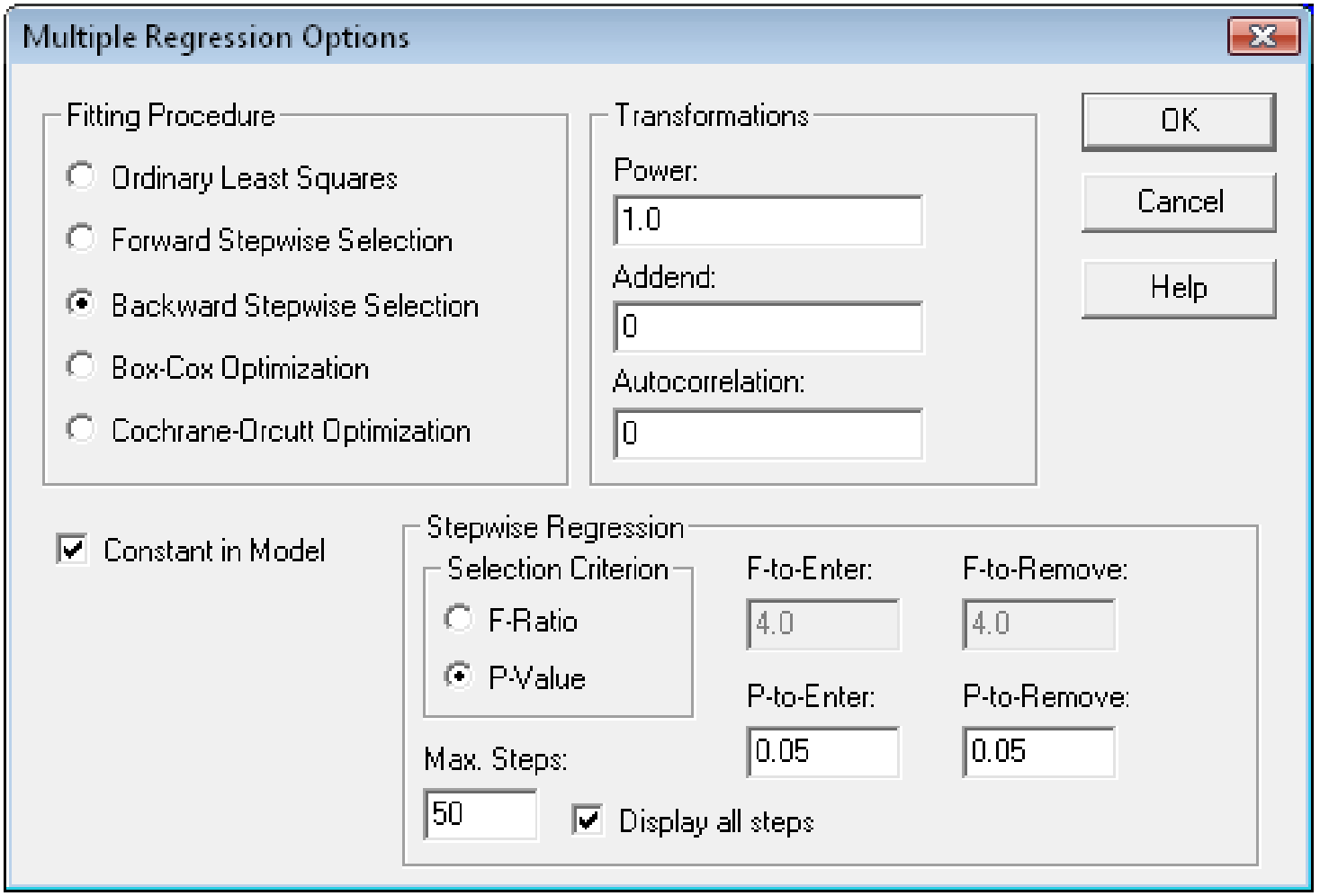
countries | B | C

STATPOINT TECHNOLOGIES, INC.

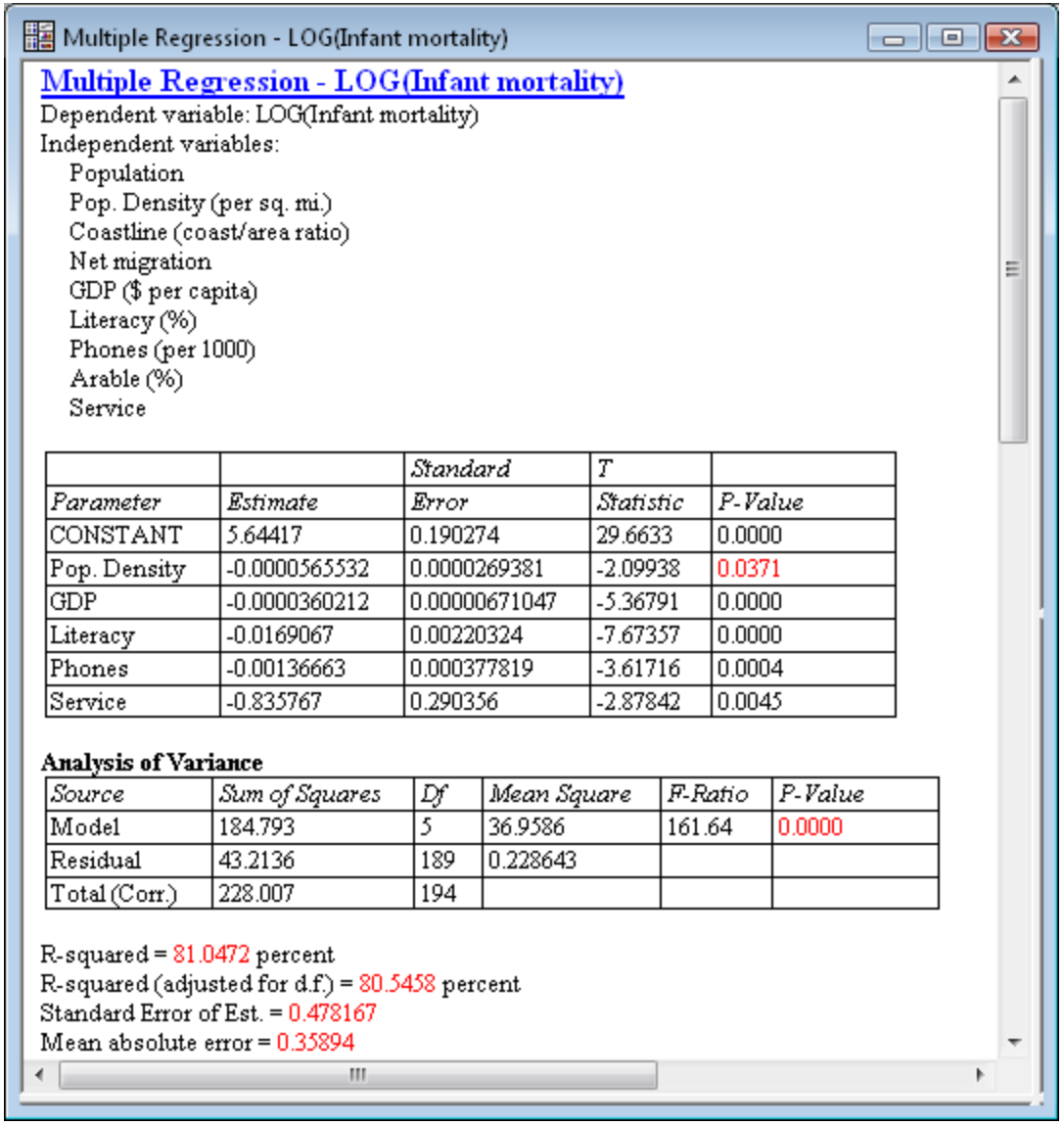
Stepwise Regression



Analysis Options



Selected Variables



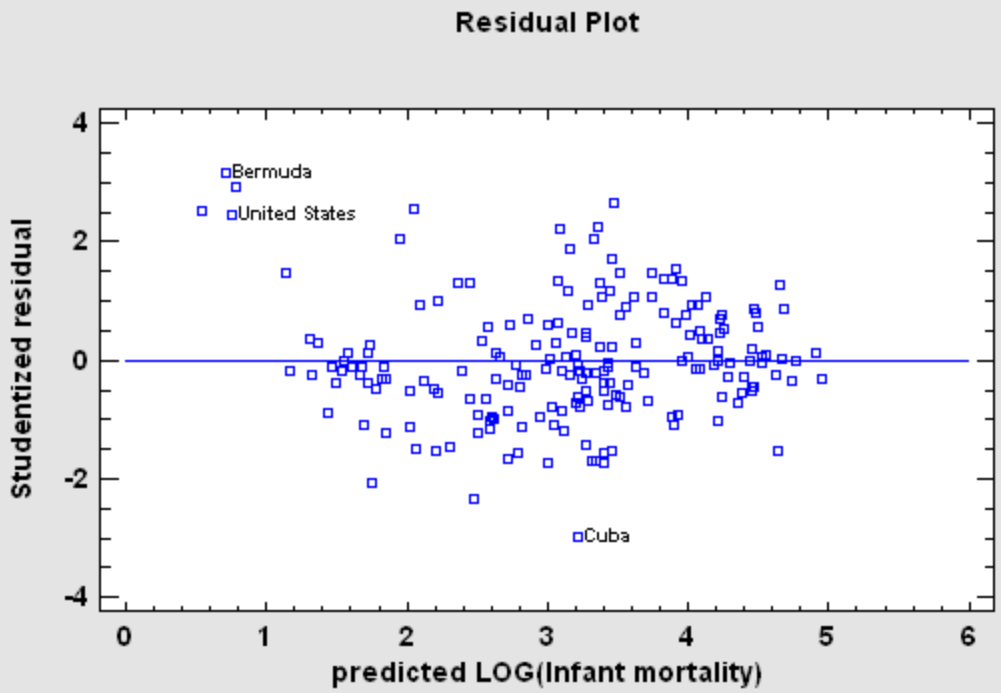
		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	5.64417	0.190274	29.6633	0.0000
Pop. Density	-0.0000565532	0.0000269381	-2.09938	0.0371
GDP	-0.0000360212	0.00000671047	-5.36791	0.0000
Literacy	-0.0169067	0.00220324	-7.67357	0.0000
Phones	-0.00136663	0.000377819	-3.61716	0.0004
Service	-0.835767	0.290356	-2.87842	0.0045

Analysis of Variance

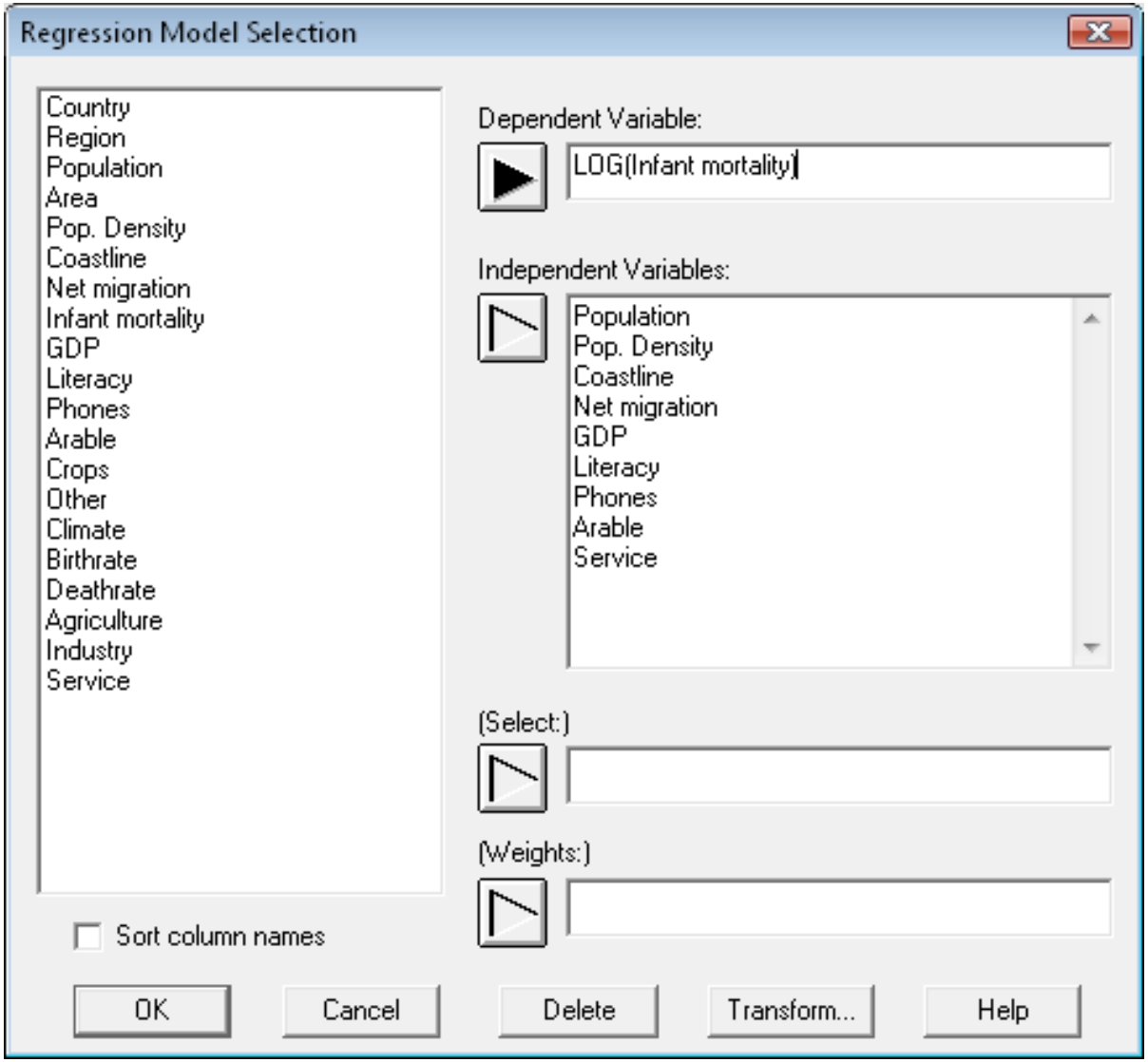
Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	184.793	5	36.9586	161.64	0.0000
Residual	43.2136	189	0.228643		
Total (Corr.)	228.007	194			

R-squared = 81.0472 percent
R-squared (adjusted for d.f.) = 80.5458 percent
Standard Error of Est. = 0.478167
Mean absolute error = 0.35894

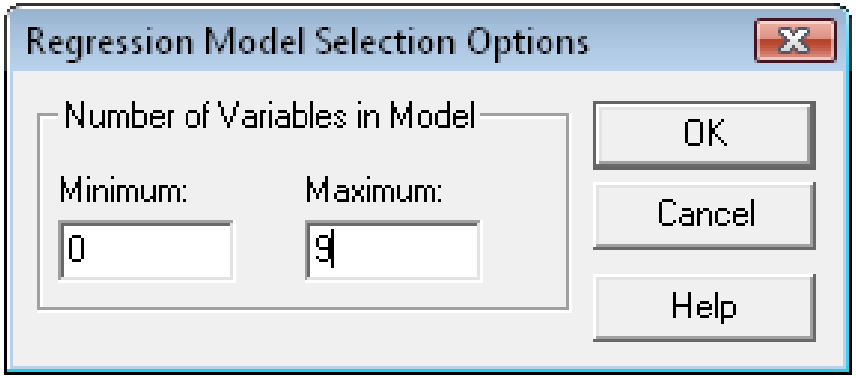
Residual Plot



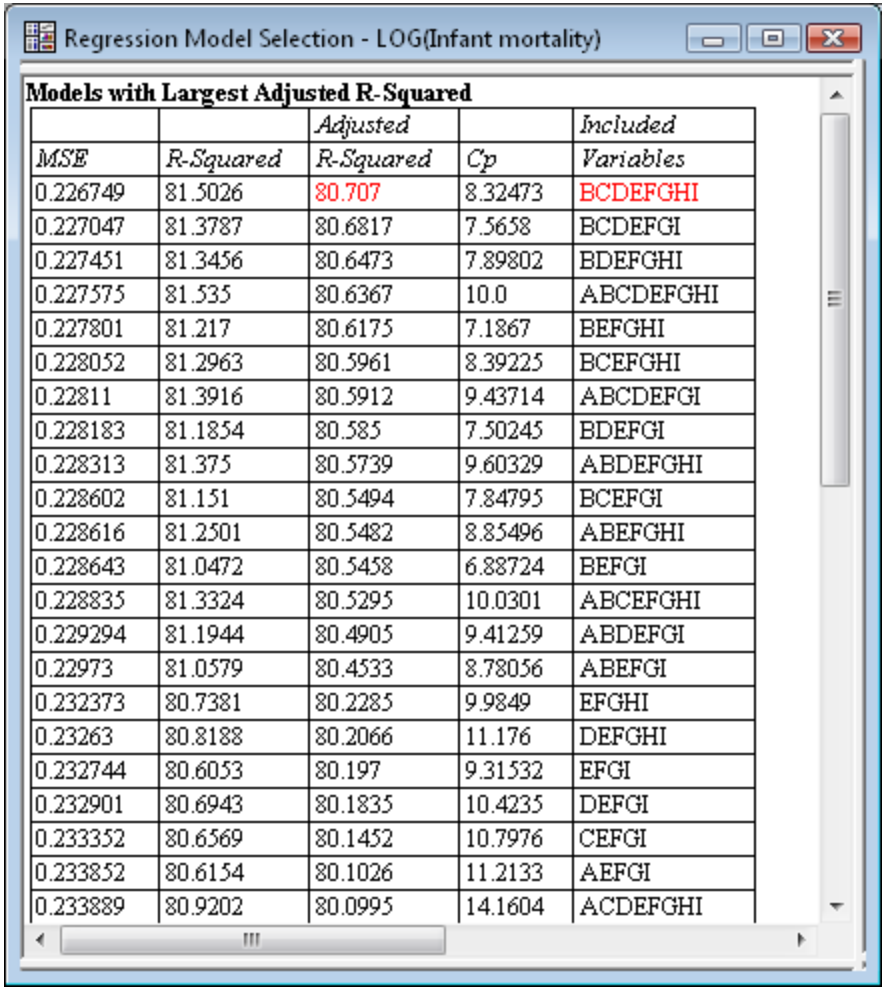
All Possible Regressions



Analysis Options



Best Adjusted R-Squared Models



Models with Largest Adjusted R-Squared

MSE	*R-Squared*	*Adjusted R-Squared*	*Cp*	*Included Variables*
0.226749	81.5026	80.707	8.32473	BCDEFGHI
0.227047	81.3787	80.6817	7.5658	BCDEFGI
0.227451	81.3456	80.6473	7.89802	BDEFGHI
0.227575	81.535	80.6367	10.0	ABCDEFGHI
0.227801	81.217	80.6175	7.1867	BEFGHI
0.228052	81.2963	80.5961	8.39225	BCEFGHI
0.22811	81.3916	80.5912	9.43714	ABCDEFGI
0.228183	81.1854	80.585	7.50245	BDEFGI
0.228313	81.375	80.5739	9.60329	ABDEFGHI
0.228602	81.151	80.5494	7.84795	BCEFGI
0.228616	81.2501	80.5482	8.85496	ABEFGHI
0.228643	81.0472	80.5458	6.88724	BEFGI
0.228835	81.3324	80.5295	10.0301	ABCEFGHI
0.229294	81.1944	80.4905	9.41259	ABDEFGI
0.22973	81.0579	80.4533	8.78056	ABEFGI
0.232373	80.7381	80.2285	9.9849	EFGHI
0.23263	80.8188	80.2066	11.176	DEFGHI
0.232744	80.6053	80.197	9.31532	EFGI
0.232901	80.6943	80.1835	10.4235	DEFGI
0.233352	80.6569	80.1452	10.7976	CEFGI
0.233852	80.6154	80.1026	11.2133	AEFGI
0.233889	80.9202	80.0995	14.1604	ACDEFGHI

Example 6: Logistic Regression

Response variable may be in the form of proportions or binary (0/1).

failures.sgd

	Load	Specimens	Failures
1	5	600	13
2	35	500	95
3	70	600	189
4	80	300	95
5	90	300	130
6			
7			
8			
9			
10			
11			
12			

failures B C

birthweight.sgd

	Low birth weight 0=No, 1=Yes	Age	Weight	Smoker 1=Yes	Doctor visits first trimester
121	0	20	158	0	1
122	0	26	160	0	0
123	0	21	115	0	1
124	0	22	129	0	0
125	0	25	130	0	2
126	0	31	120	0	2
127	0	35	170	0	1
128	0	19	120	1	0
129	0	24	116	0	1
130	0	45	123	0	1
131	1	28	120	1	0
132	1	29	130	0	2
133	1	34	187	1	0
134	1	25	105	0	0
135	1	25	85	0	0
136	1	27	150	0	0
137	1	23	97	0	1
138	1	24	128	0	1
139	1	24	132	0	0
140	1	21	165	1	1

birthweight B C

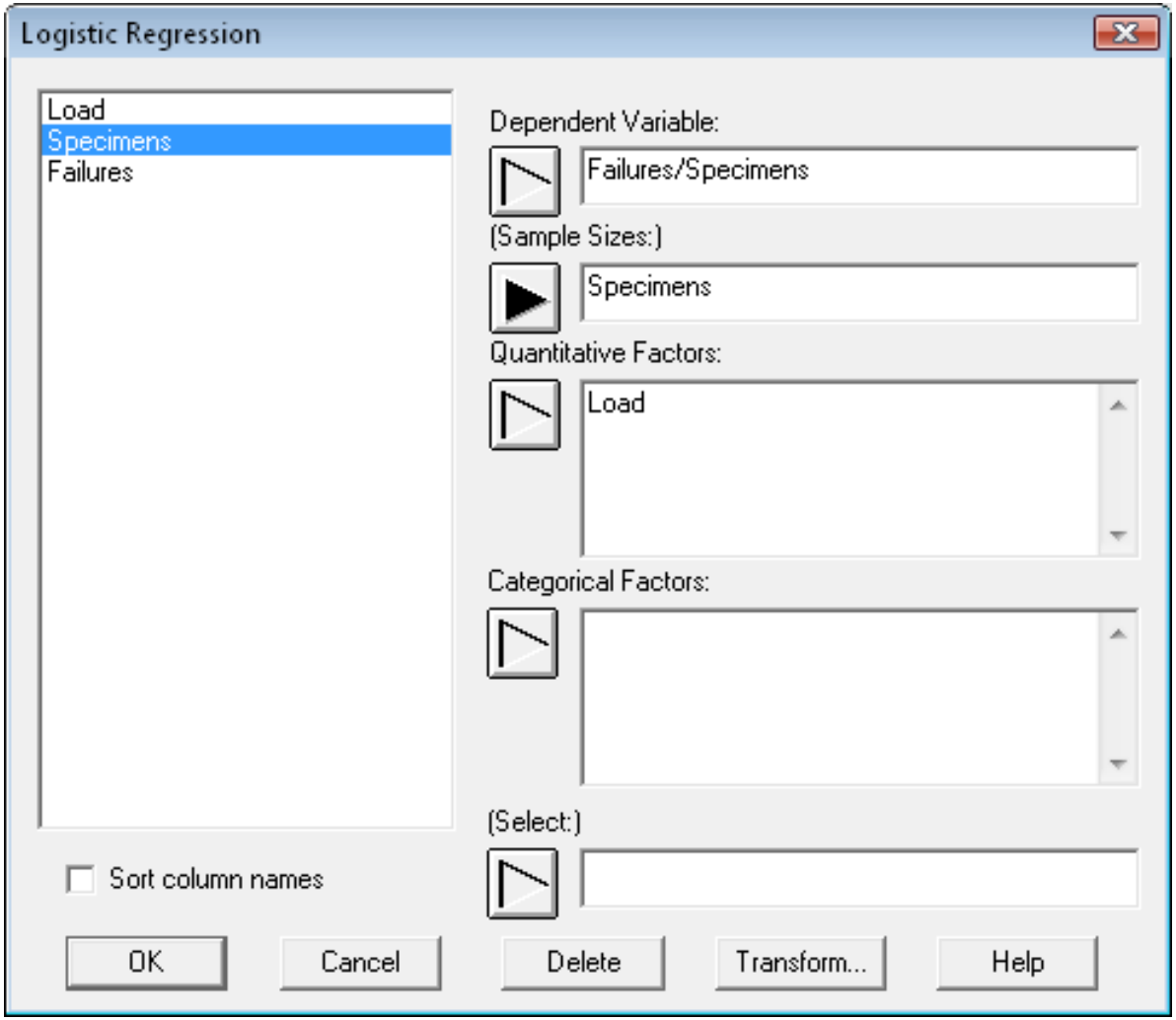
Logistic Model

Let P(Event) be the probability an event occurs at specified values of the independent variables X.

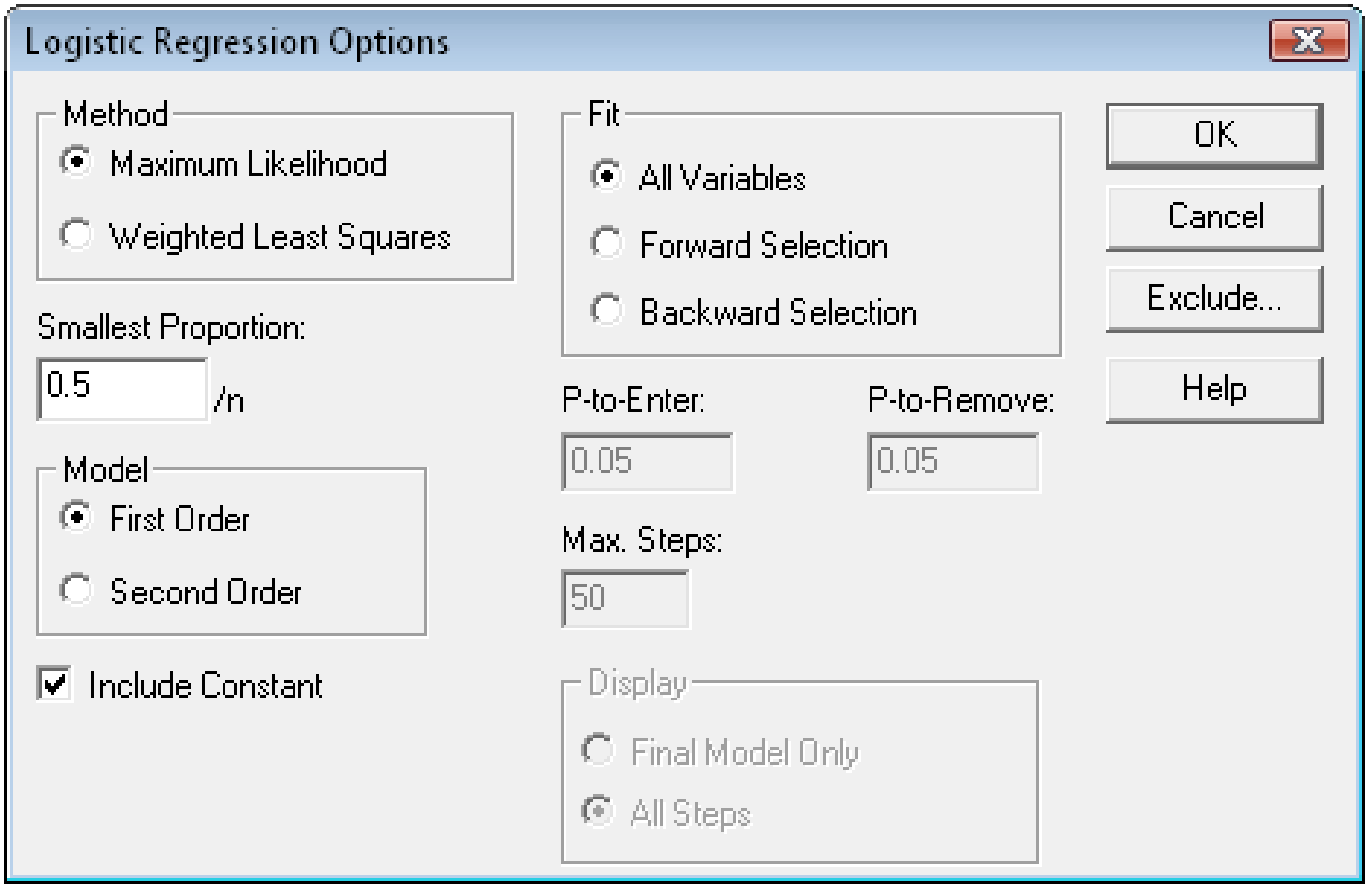
$$P(Event) = \frac{1}{1+\exp\left[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)\right]} \quad (1)$$

$$\log\left(\frac{P(Event)}{1-P(Event)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \quad (2)$$

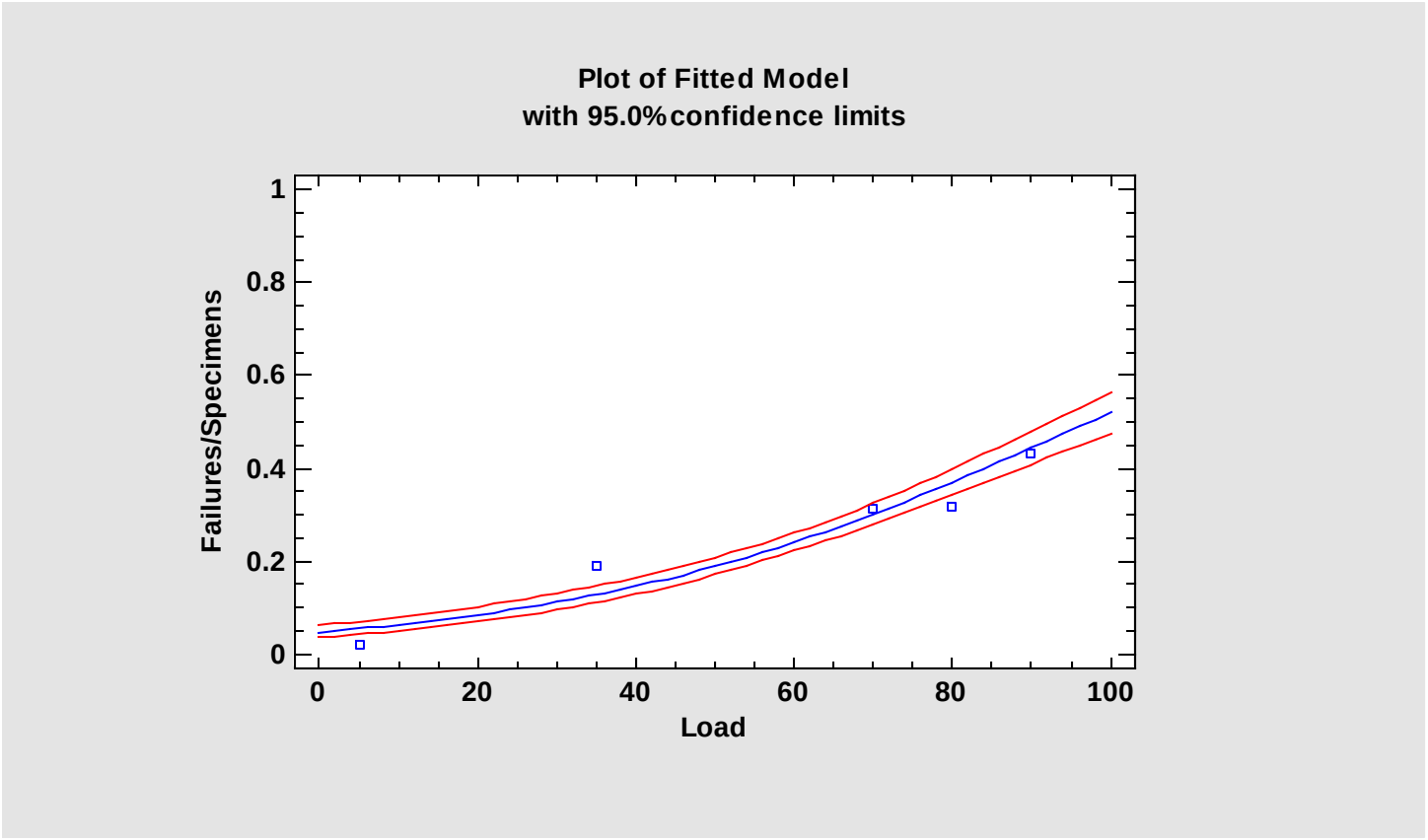
Data Input - Proportions



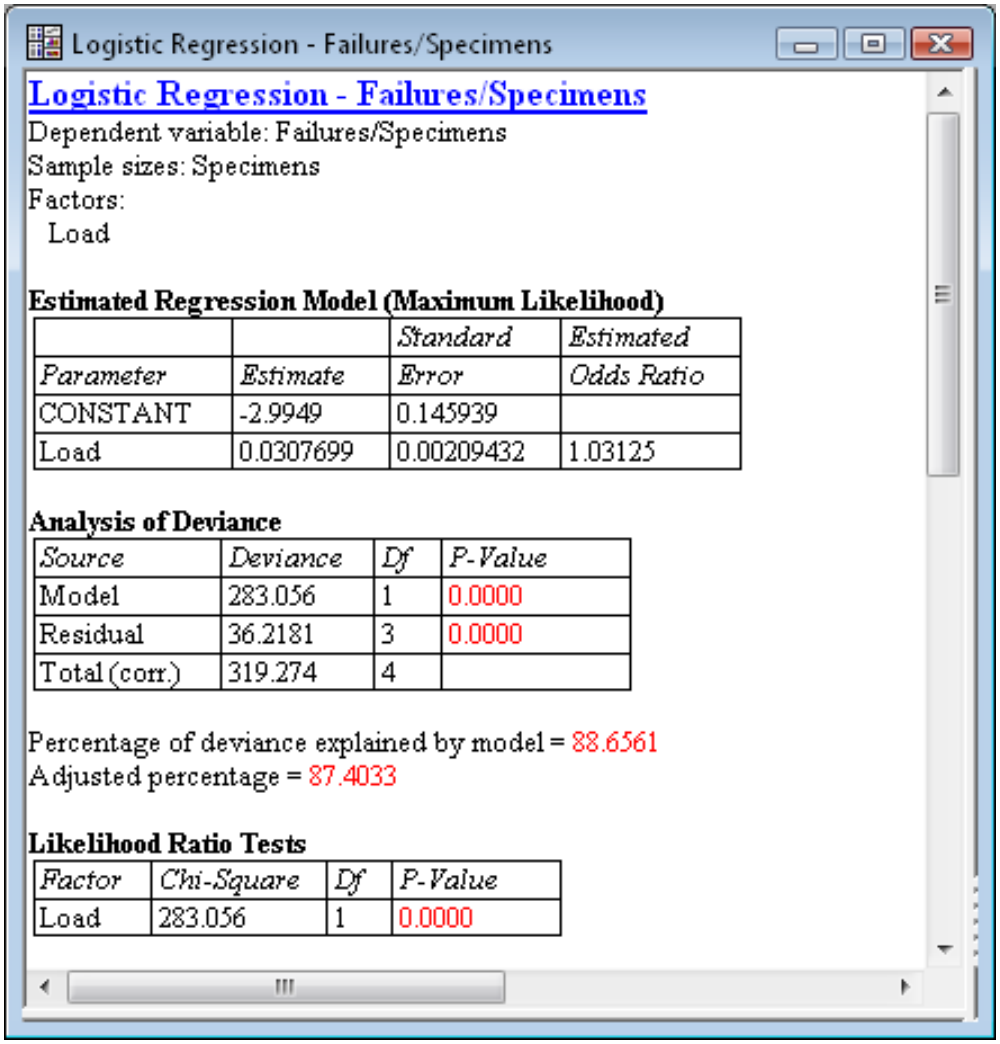
Analysis Options



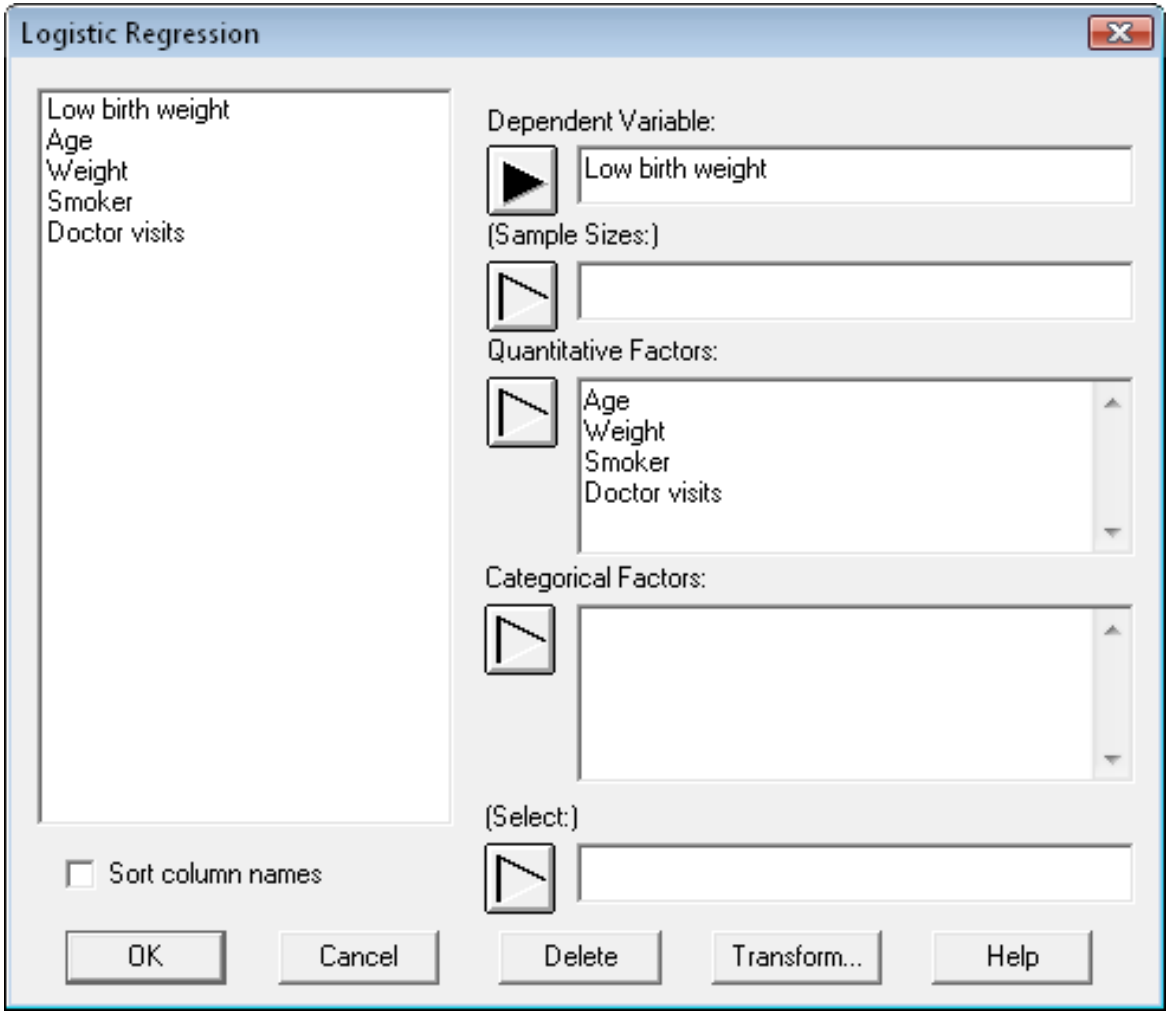
Plot of Fitted Model



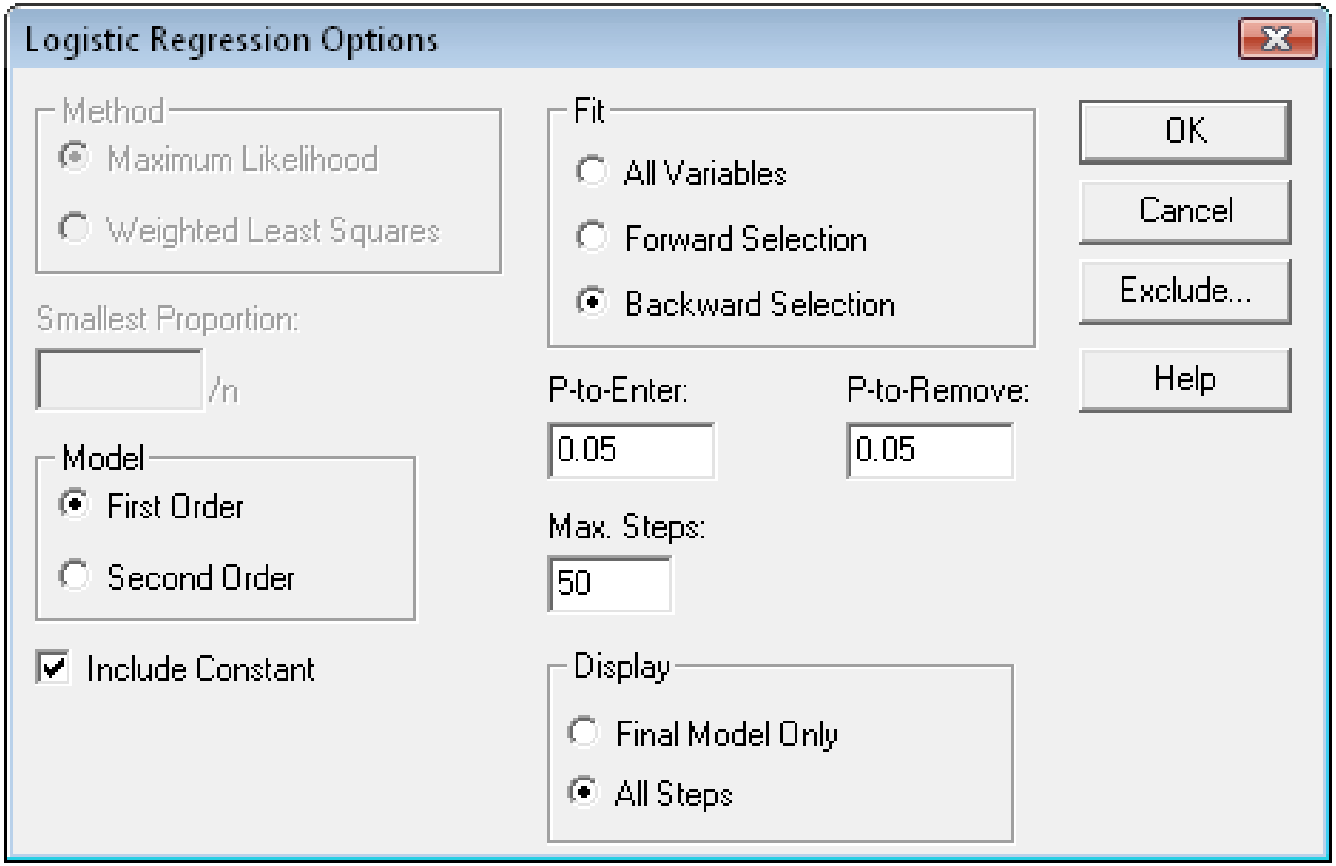
Statistical Results



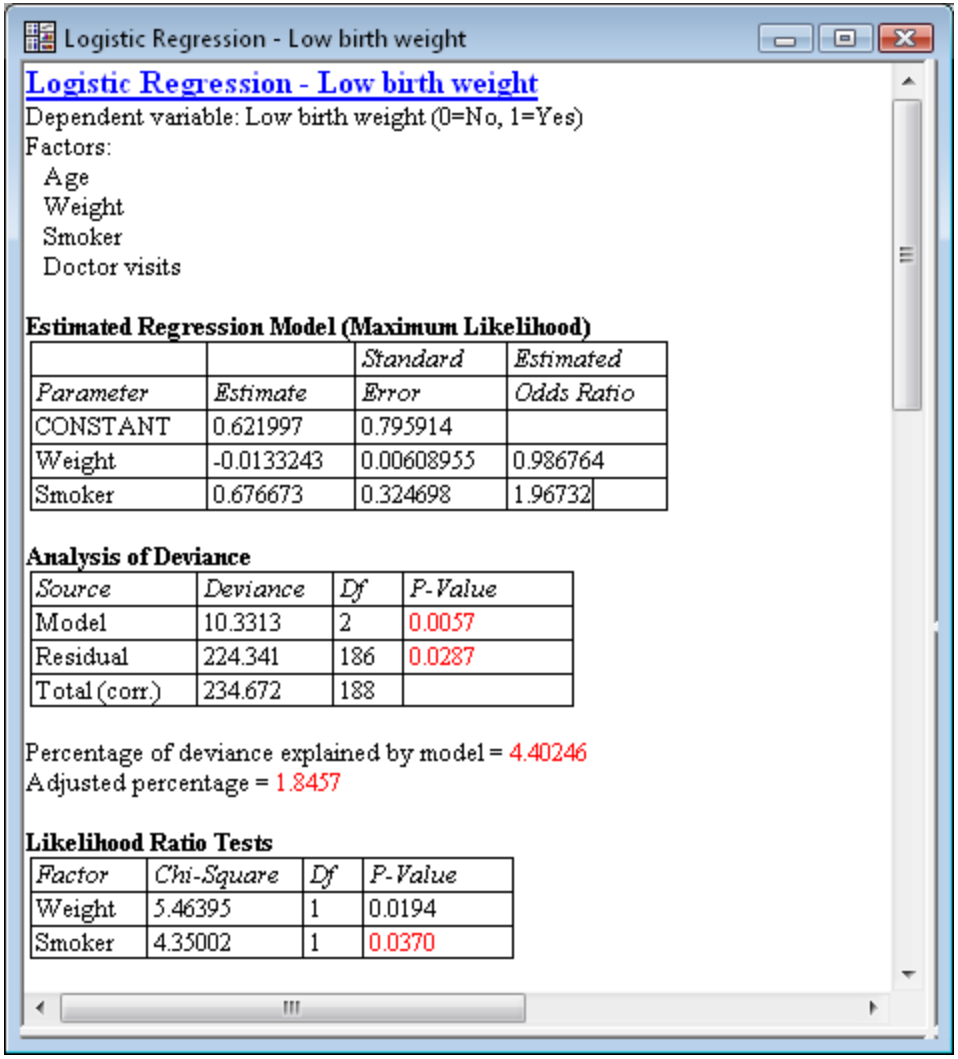
Data Input - Binary



Analysis Options



Analysis Summary



Example 7: Poisson Regression

Response variable is a count.

mining.sgd

	Injuries	Thickness	Extraction	Height	Years
	per subregion	inner burden (feet)	percent of previous seam	lower seam (feet)	since mine opened
1	2	50	70	52	1
2	1	230	65	42	6
3	0	125	70	45	1
4	4	75	65	68	0.5
5	1	70	65	53	0.5
6	2	65	70	46	3
7	0	65	60	62	1
8	0	350	60	54	0.5
9	4	350	90	54	0.5
10	4	160	80	38	0
11	1	145	65	38	10
12	4	145	85	38	0
13	1	180	70	42	2
14	5	43	80	40	0
15	2	42	85	51	12
16	5	42	85	51	0
17	5	45	85	42	0
18	5	83	85	48	10

mining B C

Poisson Model

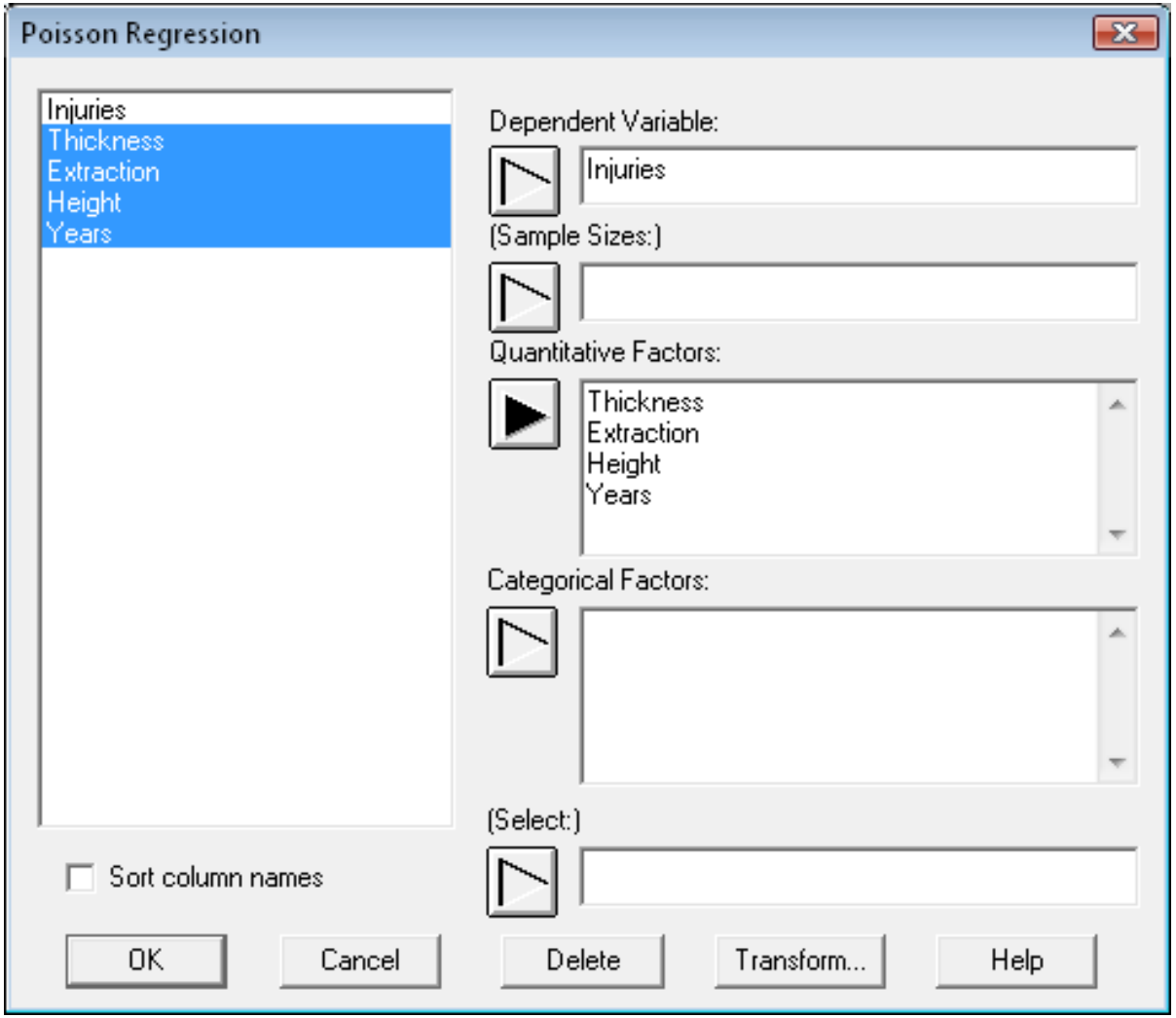
Values of the response variable are assumed to follow a Poisson distribution:

$$p(Y) = \frac{\lambda^{Y} e^{-\lambda}}{Y!}$$

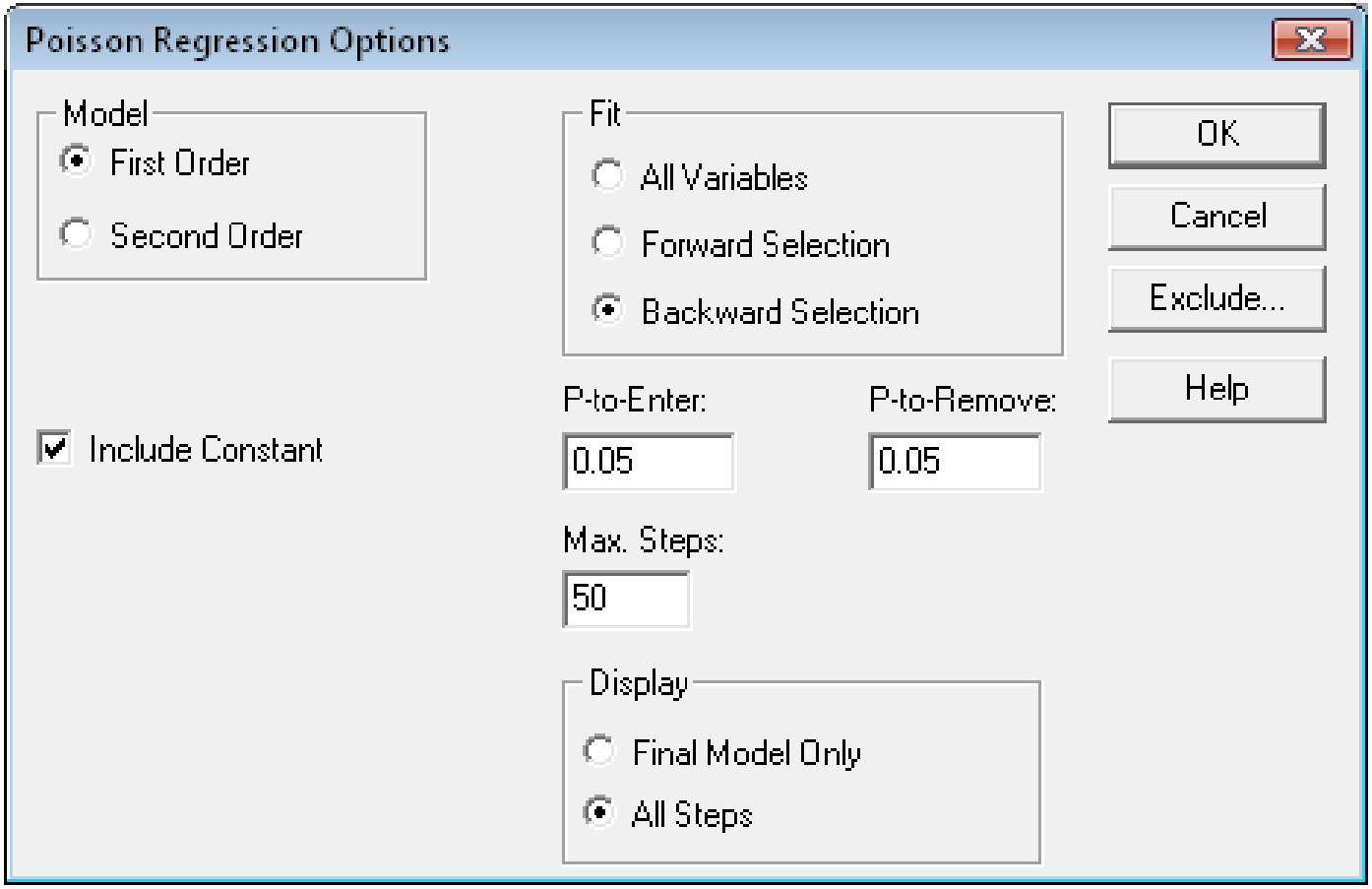
The rate parameter λ is related to the predictor variables through a log-linear link function:

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

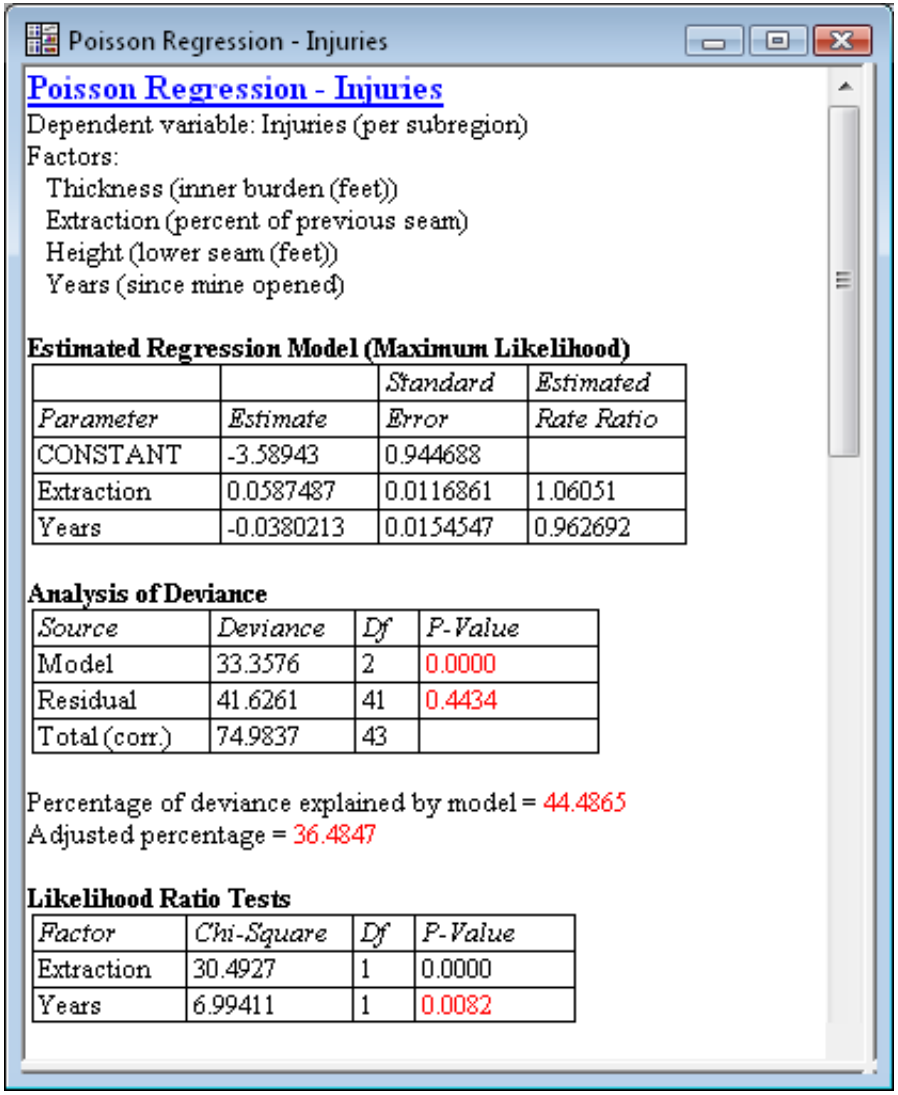
Data Input


Poisson Regression
Injuries
Thickness
Extraction
Height
Years
Dependent Variable:
Injuries
(Sample Sizes:)
Quantitative Factors:
Thickness
Extraction
Height
Years
Categorical Factors:
(Select:)
Sort column names
OK
Cancel
Delete
Transform...
Help

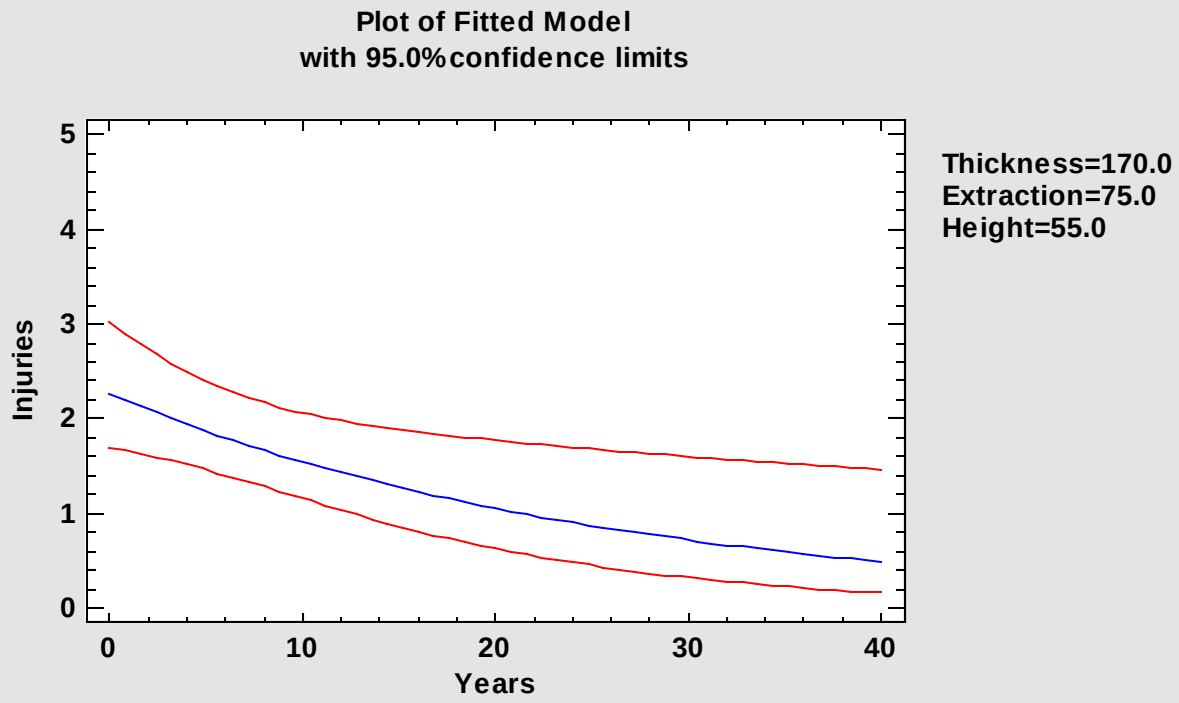
Analysis Options



Statistical Results



Plot of Fitted Model



References

- Applied Logistic Regression (second edition) – Hosmer and Lemeshow, Wiley, 2000.
- Applied Regression Analysis (third edition) – Draper and Smith, Wiley, 1998.
- Applied Linear Statistical Models (fifth edition) – Kutner et al., McGraw-Hill, 2004.
- Classical and Modern Regression with Applications (second edition) – Myers, Brookes-Cole, 1990.

More Information

Go to www.statgraphics.com

Or send e-mail to info@statgraphics.com

STATPOINT TECHNOLOGIES, INC.