Poisson Regression

Summary

The **Poisson Regression** procedure is designed to fit a regression model in which the dependent variable *Y* consists of counts. The fitted regression model relates *Y* to one or more predictor variables *X*, which may be either quantitative or categorical. The procedure fits a model using either maximum likelihood or weighted least squares. Stepwise selection of variables is an option. Likelihood ratio tests are performed to test the significance of the model coefficients. The fitted model may be plotted and predictions generated from it. Unusual residuals are identified and plotted.

Sample StatFolio: *Poisson reg.sgp*

Sample Data:

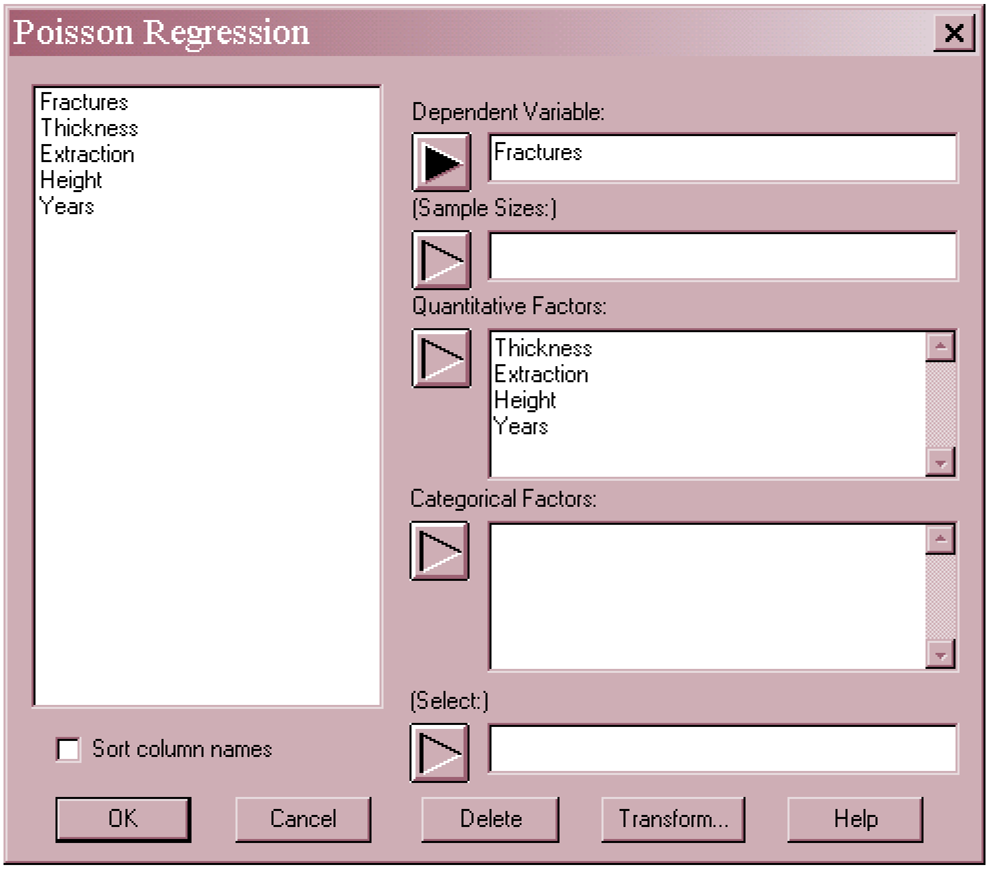
The file *mines.sgd* contains a set of data from Myers (1990) describing the number of injuries that occur in West Virginia coal fields. The data consist of *n* = 44 observations from different mines. A portion of the data is shown below:

Fractures	*Thickness*	*Extraction*	*Height*	*Years*
2	50	70	52	1
1	230	65	42	6
0	125	70	45	1
4	75	65	68	0.5
1	70	65	53	0.5
2	65	70	46	3
0	65	60	62	1
0	350	60	54	0.5
4	350	90	54	0.5
4	160	80	38	0
…	…	…	…	…

The dependent variable is *Fractures*, which tabulates the number of injuries in each mine. The other 4 columns are potential predictor variables that quantify various attributes of each mine.

Data Input

The data input dialog box requests information about the input variables:



Fractures
Thickness
Extraction
Height
Years

Dependent Variable: Fractures

(Sample Sizes:)

Quantitative Factors: Thickness, Extraction, Height, Years

Categorical Factors:

(Select:)

- **Dependent Variable**: a numeric variable containing the *n* values of dependent variable y_i. *Y* must consist of non-negative integer counts.

- **(Sample Sizes)**: optional sample sizes t_i corresponding to each count. If not specified, all t_i are set equal to 1.

- **Quantitative Factors**: numeric columns containing the values of any quantitative factors to be included in the model.

- **Categorical Factors**: numeric or non-numeric columns containing the levels of any categorical factors to be included in the model.

- **Select**: subset selection.

Statistical Model

The statistical model assumed for the data is that the values of the dependent variable *Y* follow a Poisson distribution of the form

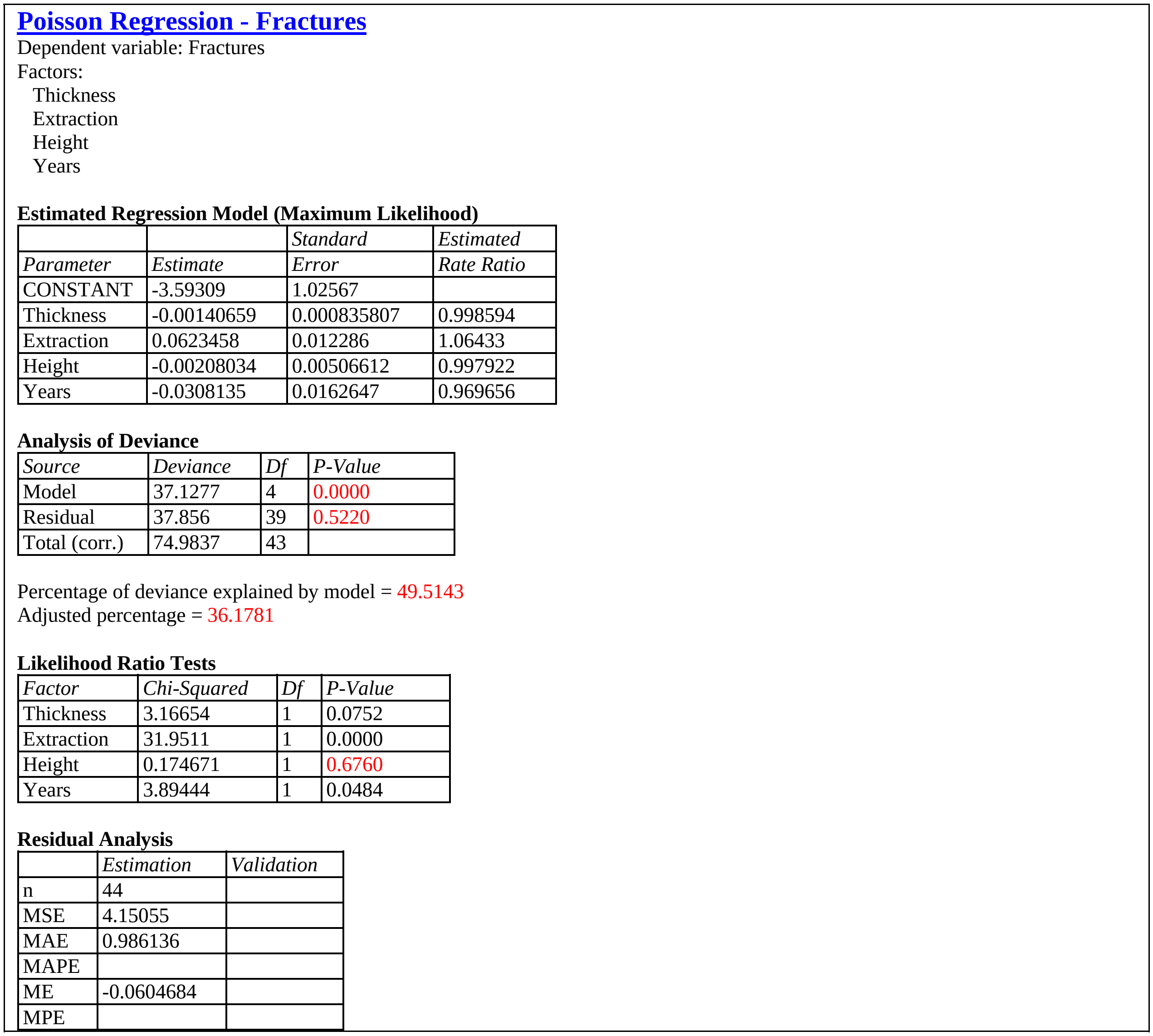
$$p(Y_i) = \frac{e^{-\lambda_i t_i}(\lambda_i t_i)}{Y_i!} \qquad (1)$$

where λ_i is the Poisson rate parameter at the settings of the predictor variables corresponding to the *i-th* observation. It is further assumed that the rate is related to the predictor variables through a log-linear link function of the form

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad (2)$$

Analysis Summary

The *Analysis Summary* displays a table showing the estimated model and tests of significance for the model coefficients. Typical output is shown below:



Poisson Regression - Fractures

Dependent variable: Fractures
Factors:
  Thickness
  Extraction
  Height
  Years

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Rate Ratio*
CONSTANT	-3.59309	1.02567	
Thickness	-0.00140659	0.000835807	0.998594
Extraction	0.0623458	0.012286	1.06433
Height	-0.00208034	0.00506612	0.997922
Years	-0.0308135	0.0162647	0.969656

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	37.1277	4	0.0000
Residual	37.856	39	0.5220
Total (corr.)	74.9837	43	

Percentage of deviance explained by model = 49.5143
Adjusted percentage = 36.1781

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Thickness	3.16654	1	0.0752
Extraction	31.9511	1	0.0000
Height	0.174671	1	0.6760
Years	3.89444	1	0.0484

Residual Analysis

	Estimation	*Validation*
n	44	
MSE	4.15055	
MAE	0.986136	
MAPE		
ME	-0.0604684	
MPE		

The output includes:

- **Data Summary:** a summary of the input data.

- **Estimated Regression Model:** estimates of the coefficients in the regression model, with standard errors and estimated rate ratios. The rate ratios are calculated from the model coefficients $\hat{\beta}_j$ by

$$\text{rate ratio} = \exp\left(\hat{\beta}_j\right) \tag{3}$$

  The rate ratio represents the percentage increase in the rate of events for each unit increase in X.

- **Analysis of Deviance:** decomposition of the deviance of the data into an explained (*Model*) component and an unexplained (*Residual*) component. *Deviance* compares the likelihood function for a model to the largest value that the likelihood function could achieve, in a manner such that a perfect model would have a deviance equal to 0. There are 3 lines in the table:

  1. **Total (corr.) –** the deviance of a model containing only a constant term, $\delta(\beta_0)$.

  2. **Residual –** the deviance remaining after the model has been fit.

  3. **Model –** the reduction in the deviance due to the predictor variables, $\delta(\beta_1,\beta_2,\ldots,\beta_k|\beta_0)$, equal to the difference between the other two components.

  The P-Value for the *Model* tests whether the addition of the predictor variables significantly reduces the deviance compared to a model containing only a constant term. A small P-Value (less than 0.05 if operating at the 5% significance level) indicates that the model has significantly reduced the deviance and is thus a useful for predictor for *Y*. The P-Value for the *Residual* term tests whether there is significant lack-of-fit, i.e., whether a better model may be possible. A small P-value indicates that significant deviance remains in the residuals, so that a better model might be possible.

- **Percentage of Deviance –** the percentage of deviance explained by the model, calculated by

$$R^2 = \frac{\delta(\beta_1, \beta_2, \ldots, \beta_k \mid \beta_0)}{\delta(\beta_0)} \tag{4}$$

  It is similar to an R-squared statistic in multiple regression, in that it can range from 0% to 100%. An adjusted deviance is also computed from

$$R^2_{adj} = \frac{\delta(\beta_1, \beta_2, \ldots, \beta_k \mid \beta_0) - 2p}{\delta(\beta_0)} \tag{5}$$

where p equals the number of coefficients in the fitted model, including the constant term. It is similar to the adjusted R-squared statistic in that it compensates for the number of variables in the model.

- **Likelihood Ratio Tests** – a test of significance for each effect in the fitted model. These tests compare the likelihood function of the full model to that of the model in which only the indicated effect has been dropped. Small P-values indicate that the model has been improved significantly by the corresponding effect.

- **Residual Analysis** – if a subset of the rows in the datasheet have been excluded from the analysis using the *Select* field on the data input dialog box, the fitted model is used to make predictions of the *Y* values for those rows. This table shows statistics on the prediction errors, defined by

  $$e_i = y_i - \hat{\lambda}_i t_i \tag{6}$$

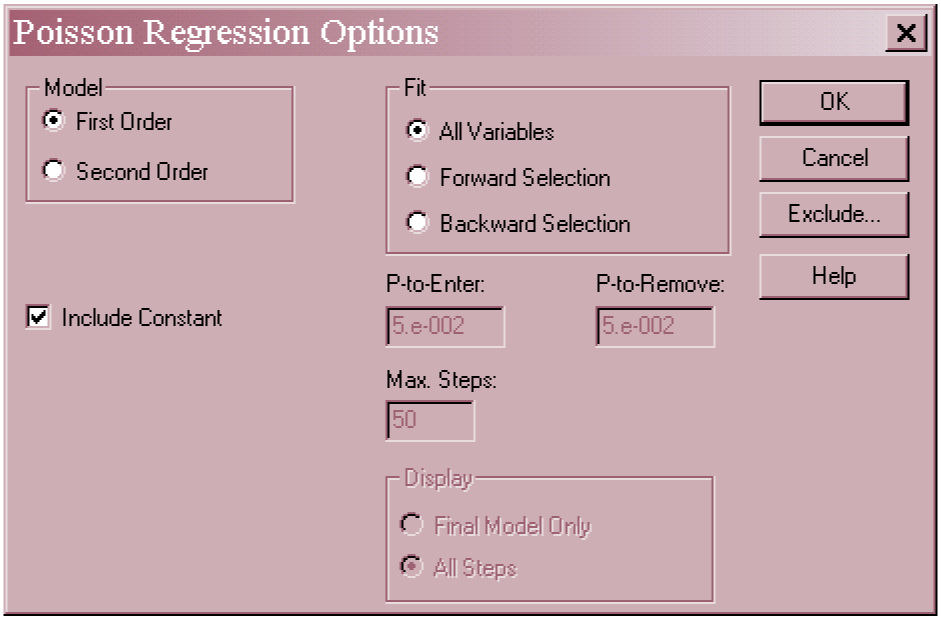
  Included are the mean squared error (MSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the mean error (ME), and the mean percentage error (MPE). These validation statistics can be compared to the statistics for the fitted model to determine how well that model predicts observations outside of the data used to fit it.

The fitted model for the sample data is

$$\hat{\lambda} = \exp\left(-3.593 - 0.001407 Thickness + 0.06235 Extraction - 0.002080 Height - 0.03081 Years\right)$$

The regression explains about 49.5% of the deviance of a model with only a constant. The P-value for a couple of the variables is above 0.05, indicating that they might reasonably be removed from the model.

Analysis Options



Poisson Regression Options

Model: ● First Order ○ Second Order

Fit: ● All Variables ○ Forward Selection ○ Backward Selection

☑ Include Constant

P-to-Enter: 5.e-002

P-to-Remove: 5.e-002

Max. Steps: 50

Display: ○ Final Model Only ● All Steps

- **Model**: order of the model to be fit. First order models include only main effects. Second order models include quadratic effects for quantitative factors and two-factor interactions amongst all variables.

- **Include Constant**: If this option is not checked, the constant term β_0 will be omitted from the model.

- **Fit**: specifies whether all independent variables specified on the data input dialog box should be included in the final model, or whether a stepwise selection of variables should be applied. Stepwise selection attempts to find a parsimonious model that contains only statistically significant variables. A *Forward Stepwise* fit begins with no variables in the model. A *Backward Stepwise* fit begins with all variables in the model.

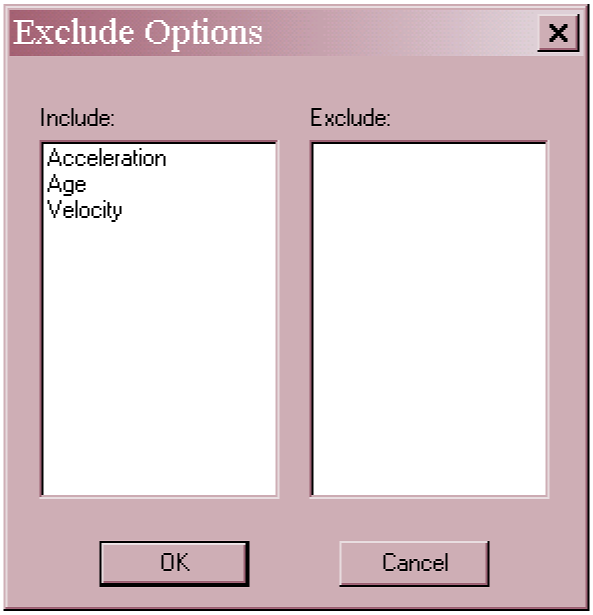
- **P-to-Enter** - In a stepwise fit, variables will be entered into the model at a given step if their P-values are less than or equal to the *P-to-Enter* value specified.

- **P-to-remove** - In a stepwise fit, variables will be removed from the model at a given step if their P-values are greater than the *P-to-Remove* value specified.

- **Max Steps:** maximum number of steps permitted when doing a stepwise fit.

- **Display**: whether to display the results at each step when doing a stepwise fit.

- **Exclude**: Press this button to exclude effects from the model. A dialog box will be displayed:



Acceleration
Age
Velocity

Double click on an effect to move it from the *Include* field to the *Exclude* field or back again.

Example: Backward Stepwise Fit of Second-Order Model

In order to find a parsimonious but useful predictor for the number of *Fractures*, a second order model was considered. This model adds additional terms involving quadratic effects such as *Thickness*^{2} and interactions between each pair or variables, represented by crossproducts such as *Thickness*Extraction*. To keep the model from containing a lot of insignificant terms, a stepwise approach was used. Two such approaches are available:

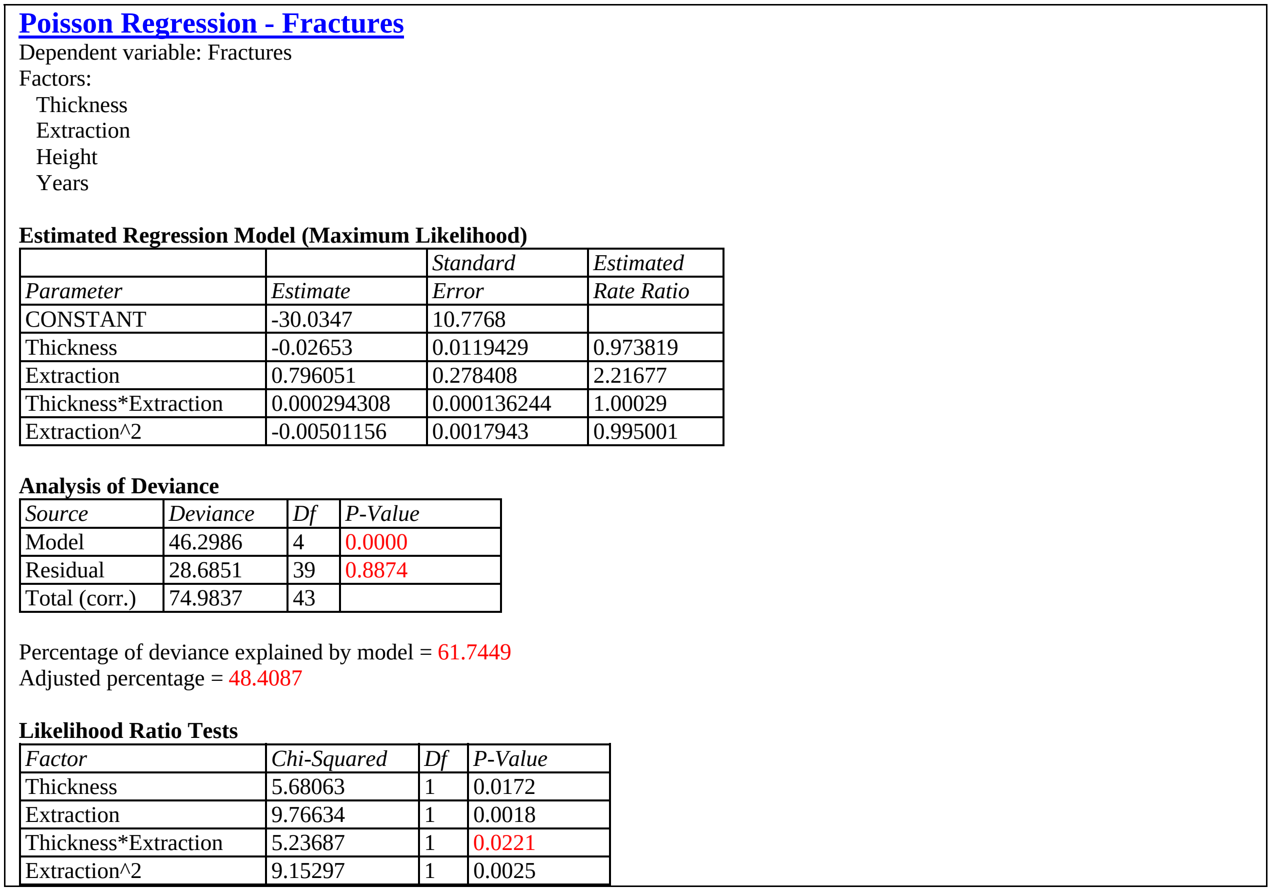
- **Forward selection** – Begins with a model involving only a constant term and enters one variable at a time based on its statistical significance if added to the current model. At each step, the algorithm brings into the model the variable that will be the most statistically significant if entered. As long as the most significant variable has a P-value less than or equal to that specified on the *Analysis Summary* dialog box, it will be brought into the model. When no variable has a small enough -value, variable selection stops. In addition, variables brought into the model early in the procedure may be removed later if their P-value falls below the *P-to-remove* criterion.

- **Backward selection** – Begins with a model involving all the variables specified on the data input dialog box and removes one variable at a time based on its statistical significance in the current model. At each step, the algorithm removes from the model the variable that is the least statistically significant. If the least significant variable has a P-value greater than that specified on the *Analysis Summary* dialog box, it will be removed from the model. When all remaining variables have small P-values, the procedure stops. In addition, variables removed from the model early in the procedure may be re-entered later if their P-values reach the *P-to-enter* criterion.

The following output shows the result of a backwards stepwise fit:

```
Stepwise factor selection
Method: backward selection
P-to-enter: 0.05
P-to-remove: 0.05

Step 0:
    14 factors in the model.  29 d.f. for error.
    Percentage of deviance explained =  68.75%     Adjusted percentage =  28.74%

Step 1:
    Removing factor Years with P-to-remove = 0.931068
    13 factors in the model.  30 d.f. for error.
    Percentage of deviance explained =  68.74%     Adjusted percentage =  31.40%

Step 2:
    Removing factor Height*Height with P-to-remove = 0.667761
    12 factors in the model.  31 d.f. for error.
    Percentage of deviance explained =  68.49%     Adjusted percentage =  33.82%

Step 3:
    Removing factor Thickness*Thickness with P-to-remove = 0.785169
    11 factors in the model.  32 d.f. for error.
    Percentage of deviance explained =  68.39%     Adjusted percentage =  36.39%

Step 4:
    Removing factor Thickness*Years with P-to-remove = 0.847819
    10 factors in the model.  33 d.f. for error.
    Percentage of deviance explained =  68.35%     Adjusted percentage =  39.01%

Step 5:
    Removing factor Extraction*Years with P-to-remove = 0.688459
    9 factors in the model.  34 d.f. for error.
    Percentage of deviance explained =  68.13%     Adjusted percentage =  41.46%

Step 6:
    Removing factor Height with P-to-remove = 0.529659
    8 factors in the model.  35 d.f. for error.
    Percentage of deviance explained =  67.60%     Adjusted percentage =  43.60%

Step 7:
    Removing factor Extraction*Height with P-to-remove = 0.957829
    7 factors in the model.  36 d.f. for error.
    Percentage of deviance explained =  67.60%     Adjusted percentage =  46.26%

Step 8:
    Removing factor Years*Years with P-to-remove = 0.402248
    6 factors in the model.  37 d.f. for error.
    Percentage of deviance explained =  66.66%     Adjusted percentage =  47.99%

Step 9:
    Removing factor Thickness*Height with P-to-remove = 0.39377
    5 factors in the model.  38 d.f. for error.
    Percentage of deviance explained =  65.69%     Adjusted percentage =  49.69%

Step 10:
    Removing factor Height*Years with P-to-remove = 0.0852434
    4 factors in the model.  39 d.f. for error.
    Percentage of deviance explained =  61.74%     Adjusted percentage =  48.41%

Final model selected.
```

The algorithm began with 14 effects. After 10 steps, the number of effects in the model had been reduced to 4. A summary of the final model is shown below:



Poisson Regression - Fractures

Dependent variable: Fractures

Factors:
  Thickness
  Extraction
  Height
  Years

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Rate Ratio*
CONSTANT	-30.0347	10.7768	
Thickness	-0.02653	0.0119429	0.973819
Extraction	0.796051	0.278408	2.21677
Thickness*Extraction	0.000294308	0.000136244	1.00029
Extraction^2	-0.00501156	0.0017943	0.995001

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	46.2986	4	0.0000
Residual	28.6851	39	0.8874
Total (corr.)	74.9837	43	

Percentage of deviance explained by model = 61.7449
Adjusted percentage = 48.4087

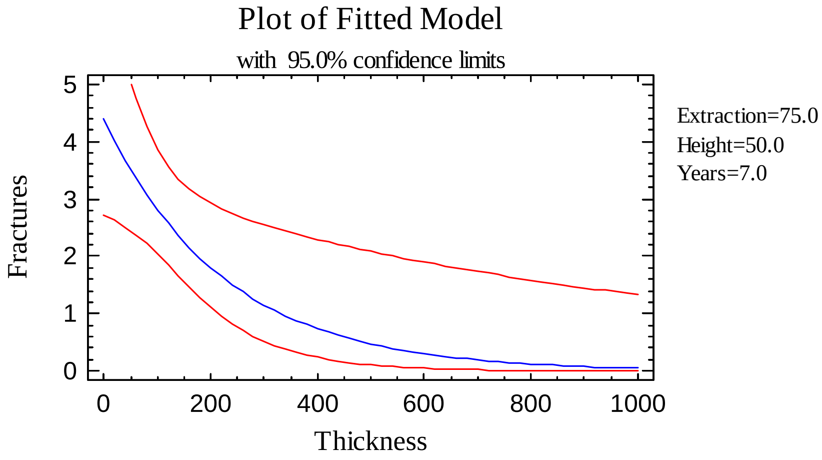
Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Thickness	5.68063	1	0.0172
Extraction	9.76634	1	0.0018
Thickness*Extraction	5.23687	1	0.0221
Extraction^2	9.15297	1	0.0025

The final model involves only 2 variables: *Thickness* and *Extraction*. It contains main effects for both variables, an interaction between the 2 variables, and a quadratic effect for *Extraction*. The percentage of deviance explained by the model has increased to approximately 61.7%.

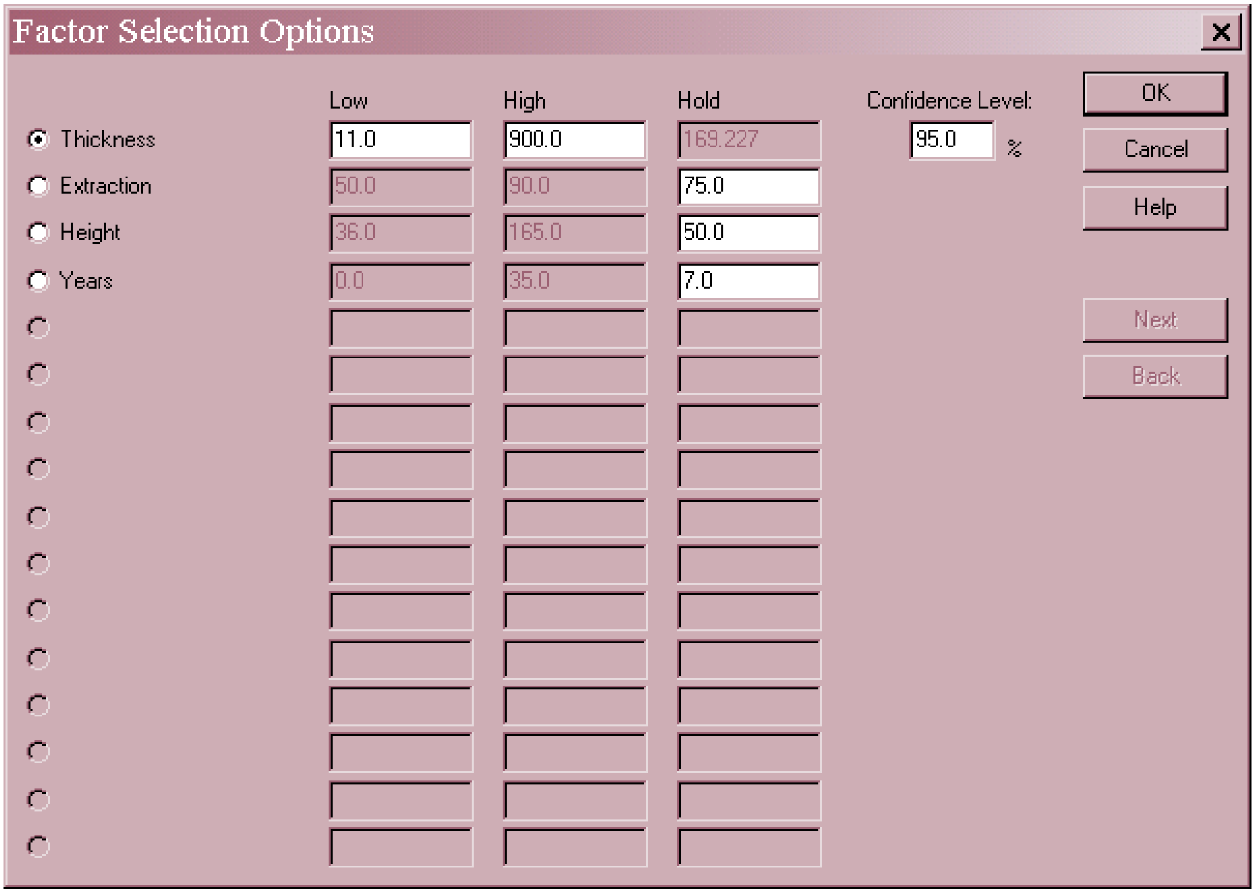
Plot of Fitted Model

The *Plot of Fitted Model* displays the estimated mean rate $\hat{\lambda}(X)$ versus any single predictor variable, with the other variables held constant.



Confidence limits for $\lambda(X)$ are included on the plot.

Pane Options



	Low	High	Hold
● Thickness	11.0	900.0	169.227
○ Extraction	50.0	90.0	75.0
○ Height	36.0	165.0	50.0
○ Years	0.0	35.0	7.0
○			
○			
○			
○			
○			
○			
○			
○			
○			
○			
○			
○			

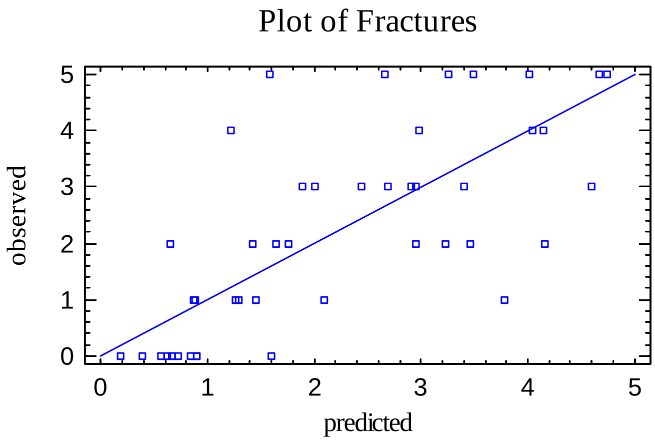
Confidence Level: 95.0 %

- **Factor:** select the factor to plot on the horizontal axis.
- **Low and High:** specify the range of values for the selected factor.
- **Hold:** values to hold the unselected factors at.
- **Confidence Level:** percentage used for the confidence limits. Set to 0 to suppress the limits.
- **Next** and **Back:** used to display other factors when more than 16 are present.

The estimated rate of fractures decreases from a high of almost 4.5 to a low of nearly 0 as the *Thickness* of the mine increases, at *Extraction* = 75, *Height* = 50, and *Years* = 7.

Observed Versus Predicted

The *Observed versus Predicted* plot shows the observed values of *Y* on the vertical axis and the predicted mean values $\hat{\lambda}_i t_i$ on the horizontal axis.



If the model fits well, the points should be randomly scattered around the diagonal line.

Predictions

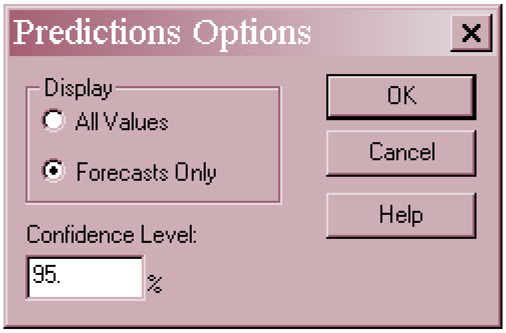
The fitted regression model may be used to predict the outcome of new samples whose predictor variables are given. For example, suppose a prediction is desired for a mine with *Thickness* = 100, *Extraction* = 70, *Height* = 50, and *Years* = 10. A new row could be added to the datasheet with these values for the predictor variables, but the entry for *Fractures* would be left blank. The *Predictions* pane would then display:

Predictions for Fractures

	Observed	*Fitted*	*Lower 95.0% CL*	*Upper 95.0% CL*
Row	*Value*	*Value*	*for Prediction*	*for Prediction*
45		1.24396	0.846319	1.82844

The table shows the fitted value $\hat{\lambda}_i t_i$, together with approximate 95% confidence intervals.

Pane Options



Display
○ All Values
● Forecasts Only

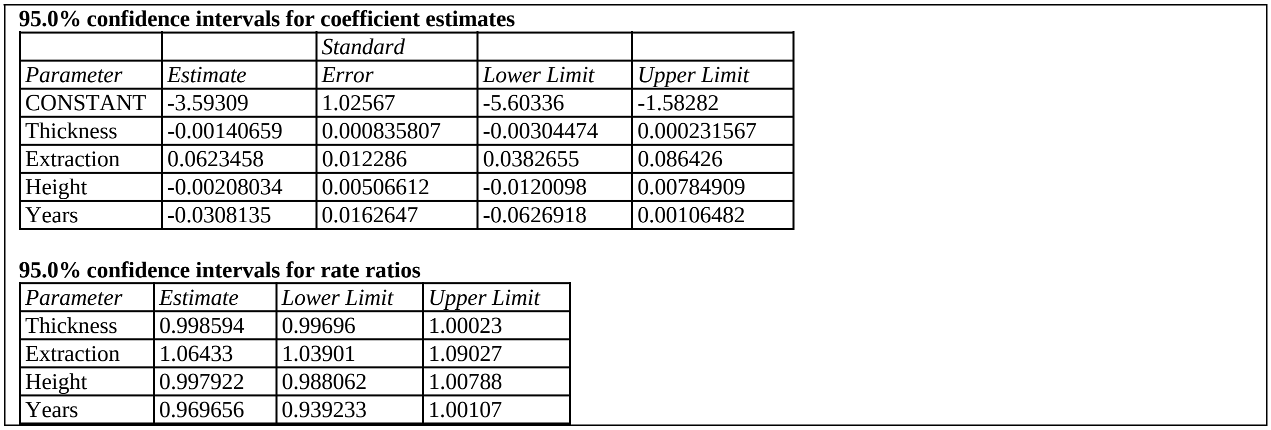
Confidence Level: 95. %

- **Display**: display *All Values* (predictions for all rows in the datasheet), or *Forecasts Only* (predictions for rows with missing values for *Y*).

- **Confidence Level:** percentage used by the confidence intervals.

Confidence Intervals

The *Confidence Intervals* pane shows the potential estimation error associated with each coefficient in the model, as well as for the rate ratios.



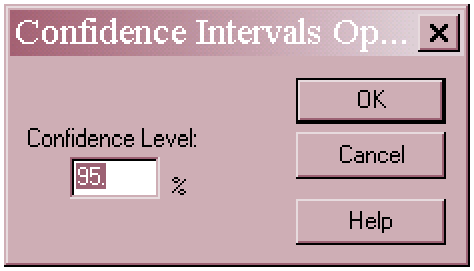
95.0% confidence intervals for coefficient estimates

		Standard		
Parameter	*Estimate*	*Error*	*Lower Limit*	*Upper Limit*
CONSTANT	-3.59309	1.02567	-5.60336	-1.58282
Thickness	-0.00140659	0.000835807	-0.00304474	0.000231567
Extraction	0.0623458	0.012286	0.0382655	0.086426
Height	-0.00208034	0.00506612	-0.0120098	0.00784909
Years	-0.0308135	0.0162647	-0.0626918	0.00106482

95.0% confidence intervals for rate ratios

Parameter	*Estimate*	*Lower Limit*	*Upper Limit*
Thickness	0.998594	0.99696	1.00023
Extraction	1.06433	1.03901	1.09027
Height	0.997922	0.988062	1.00788
Years	0.969656	0.939233	1.00107

Pane Options

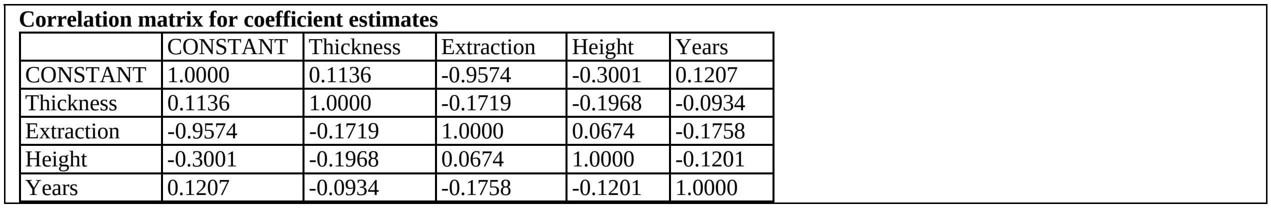


Confidence Level: 95. %

- **Confidence Level:** percentage level for the confidence intervals.

Correlation Matrix

The *Correlation Matrix* displays estimates of the correlation between the estimated coefficients.



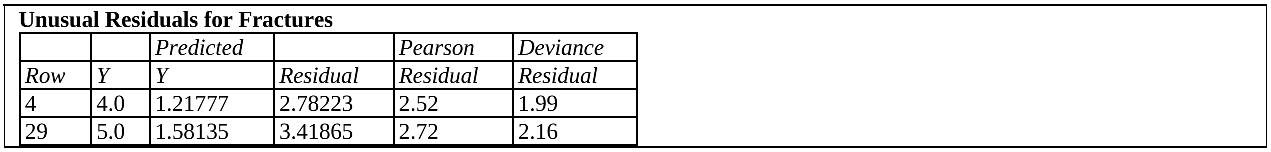
Correlation matrix for coefficient estimates

	CONSTANT	Thickness	Extraction	Height	Years
CONSTANT	1.0000	0.1136	-0.9574	-0.3001	0.1207
Thickness	0.1136	1.0000	-0.1719	-0.1968	-0.0934
Extraction	-0.9574	-0.1719	1.0000	0.0674	-0.1758
Height	-0.3001	-0.1968	0.0674	1.0000	-0.1201
Years	0.1207	-0.0934	-0.1758	-0.1201	1.0000

This table can be helpful in determining how well the effects of different independent variables have been separated from each other.

Unusual Residuals

Once the model has been fit, it is useful to study the residuals to determine whether any outliers exist that should be removed from the data. The *Unusual Residuals* pane lists all observations that have unusually large residuals.



Unusual Residuals for Fractures

		Predicted		*Pearson*	*Deviance*
Row	*Y*	*Y*	*Residual*	*Residual*	*Residual*
4	4.0	1.21777	2.78223	2.52	1.99
29	5.0	1.58135	3.41865	2.72	2.16

The table displays:

- **Row –** the row number in the datasheet.

- **Y –** the observed value of Y.

- **Predicted Y –** the fitted value $\hat{\lambda}_i t_i$.

- **Residual** – the difference between the observed and predicted values defined by

$$e_i = y_i - \hat{\lambda}_i t_i \tag{7}$$

- **Pearson Residual –** a standardized residual in which each residual is divided by an estimate of its standard error:

$$r_i = \frac{e_i}{\sqrt{\hat{\lambda}_i t_i}} \tag{8}$$

- **Deviance Residual –** a residual that measures each observation’s contribution to the residual deviance:

$$d_i = \text{sgn}(r_i)\sqrt{2\left\{ y_i \ln\left(\frac{y_i}{\hat{\lambda}_i t_i}\right) - y_i + \hat{\lambda}_i t_i \right\}} \tag{9}$$

  The sum of squared deviance residuals equals the deviance on the *Residuals* line of the analysis of deviance table.

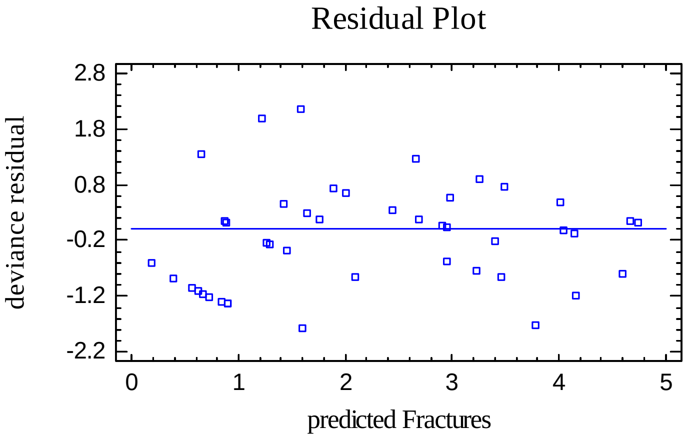
The table includes all rows for which the absolute value of the Pearson residual is greater than 2.0. The current example shows 2 residuals that exceed 2.5, but none that exceed 3.0.

Residual Plots

As with all statistical models, it is good practice to examine the residuals. The *Poisson Regression* procedure includes various type of residual plots, depending on *Pane Options*.

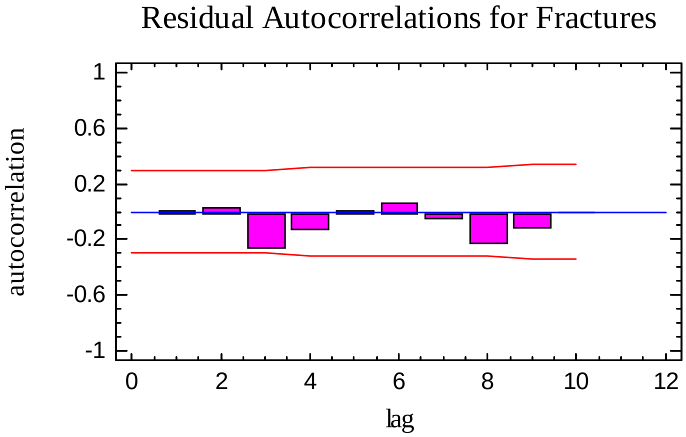
Scatterplot versus Predicted Value

This plot is helpful in visualizing whether the variability of the residuals is constant or depends on the predicted value.



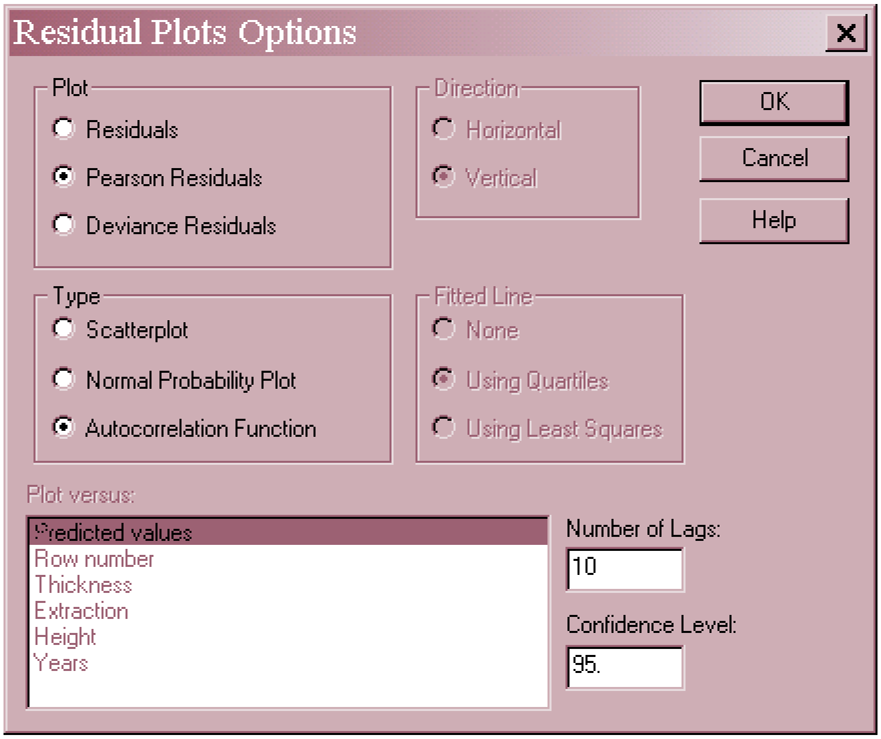
Residual Autocorrelations

This plot calculates the autocorrelation between residuals as a function of the number of rows between them in the datasheet.



It is only relevant if the data have been collected sequentially. Any bars extending beyond the probability limits would indicate significant dependence between residuals separated by the indicated “lag”.

Pane Options



Residual Plots Options

Plot
- ○ Residuals
- ● Pearson Residuals
- ○ Deviance Residuals

Direction
- ○ Horizontal
- ● Vertical

Type
- ○ Scatterplot
- ○ Normal Probability Plot
- ● Autocorrelation Function

Fitted Line
- ○ None
- ● Using Quartiles
- ○ Using Least Squares

Plot versus:
- Predicted values
- Row number
- Thickness
- Extraction
- Height
- Years

Number of Lags: 10

Confidence Level: 95.

- **Plot:** the type of residuals to plot:

  1. *Residuals* – the observed values minus the fitted values.
  2. *Studentized residuals* – the residuals divided by their estimated standard errors.
  3. *Deviance Residuals* – residuals scaled so that their sum of squares equals the residual deviance.

- **Type:** the type of plot to be created. A *Scatterplot* is used to test for curvature. A *Normal Probability Plot* is used to determine whether the model residuals come from a normal distribution (normality is not expected in this procedure). An *Autocorrelation Function* is used to test for dependence between consecutive residuals.

- **Plot Versus**: for a *Scatterplot*, the quantity to plot on the horizontal axis.

- **Number of Lags**: for an *Autocorrelation Function*, the maximum number of lags. For small data sets, the number of lags plotted may be less than this value.

- **Confidence Level:** for an *Autocorrelation Function*, the level used to create the probability limits.

Influential Points

In fitting a regression model, all observations do not have an equal influence on the parameter estimates in the fitted model. Those with unusual values of the independent variables tend to have more influence than the others. The *Influential Points* pane displays any observations that have high influence on the fitted model:

Influential Points for Fractures

Row	*Leverage*
25	0.437161
30	0.367098

Average leverage of single data point = 0.113636

The table displays all points with high *leverage*. Leverage is a statistic that measures how distant an observation is from the mean of all *n* observations in the space of the *independent* variables. The higher the leverage, the greater the impact of the point on the fitted values $\hat{y}$. Points are placed on the list if their leverage is more than 3 times that of an average data point.

The observation with the highest leverage in the sample data is row #25, although it is only about 4 times the average leverage.

Save Results

The following results may be saved to the datasheet:

1. *Predicted Values* – the fitted values $\hat{\lambda}_i t_i$ corresponding to each row of the datasheet.
2. *Lower Limits* – the lower confidence limits for $\hat{\lambda}_i t_i$.
3. *Upper Limits* – the upper confidence limits for $\hat{\lambda}_i t_i$.
4. *Residuals* – the ordinary residuals.
5. *Pearson Residuals* – the standardized Pearson residuals.
6. *Deviance Residuals* – the deviance residuals.
7. *Leverages* – the leverages for each row.

Calculations

Let λ_i = the estimated model evaluated at the settings of the predictor variables in row *i*.

Likelihood Function

$$L = \prod_{i=1}^{n} \frac{[\lambda_i t_i]^{y_i} \exp(-\lambda_i t_i)}{y_i!} \qquad (10)$$

Deviance

$$\delta(\hat{\beta}) = \frac{L(\hat{\beta})}{\prod_{i=1}^{n} \frac{y_i^{\,y_i} \exp(-y_i)}{y_i!}} \qquad (11)$$

Leverage

$$h_i = diag\left\{X_i'\left(X'WX\right)^{-1} X_i\right\} w_i \qquad (12)$$

$$\bar{h} = \frac{p}{n} \qquad (13)$$