Relate – Y and X

This procedure fits equations relating a *Response* variable Y and a *Predictor* variable X. Both linear and nonlinear models may be fit.

The data for this analysis consist of *n* values of two numeric variables. Let

y_i = *i*-th value of the *Response* variable

x_i = *i*-th value of the *Predictor* variable

Access

Highlight: one *Response* column and one *Predictor* column.

Select: *Relate* from the main menu.

Output Page 1: A plot of the fitted model.

Output Page 2: A plot of the fitted model with confidence limits for the mean response.

Output Page 3: Calculation of predicted values for Y given a value for X.

Output Page 4: Calculation of predicted values for X given a value for Y. This is often called the *Calibration* problem.

Output Page 5: Plot of the residuals from the fitted model.

Options

Unless transformations of the variables are requested, this procedure fits a linear model of the form

$$y = \beta_0 + \beta_1 x \tag{1}$$

In the above equation, β_1 represents the slope of the line, while β_0 represents the Y-intercept. You may request a transformation of one or both of the variables, in which case a linear model is fit after the transformations are applied.

To request a transformation of either variable:

1. Access the *Properties* dialog box for the variable to be transformed by double-clicking on the header of its column.

2. On the *Dist.* tab, select either a lognormal or a power normal distribution. If you select *Lognormal*, the logarithms of the data will be used to fit the model. If you select *Power normal*, the data will be raised to the indicated power when the model is fit.

Depending on the transformations selected, the fitted model will take one of the 9 forms shown in the table below:

Transformation on Y	*Transformation on X*	*Model*	*Equation*
none	none	Linear	$y = \beta_0 + \beta_1 x$
log	none	Exponential	$y = e^{(\beta_0 + \beta_1 x)}$
power - *p*	none	Power-Y	$y = (\beta_0 + \beta_1 x)^{1/p}$
none	log	Logarithmic-X	$y = \beta_0 + \beta_1 \ln(x)$
log	log	Multiplicative	$y = e^{\beta_0} x^{\beta_1}$
power - *p*	log	Power-Y Log-X	$y = (\beta_0 + \beta_1 \ln(x))^{1/p}$
none	power - *q*	Power-X	$y = \beta_0 + \beta_1 x^q$
log	power - *q*	Log-Y Power-X	$y = e^{(\beta_0 + \beta_1 x^q)}$
power - *p*	power - *q*	Double power	$y = (\beta_0 + \beta_1 x^q)^{1/p}$

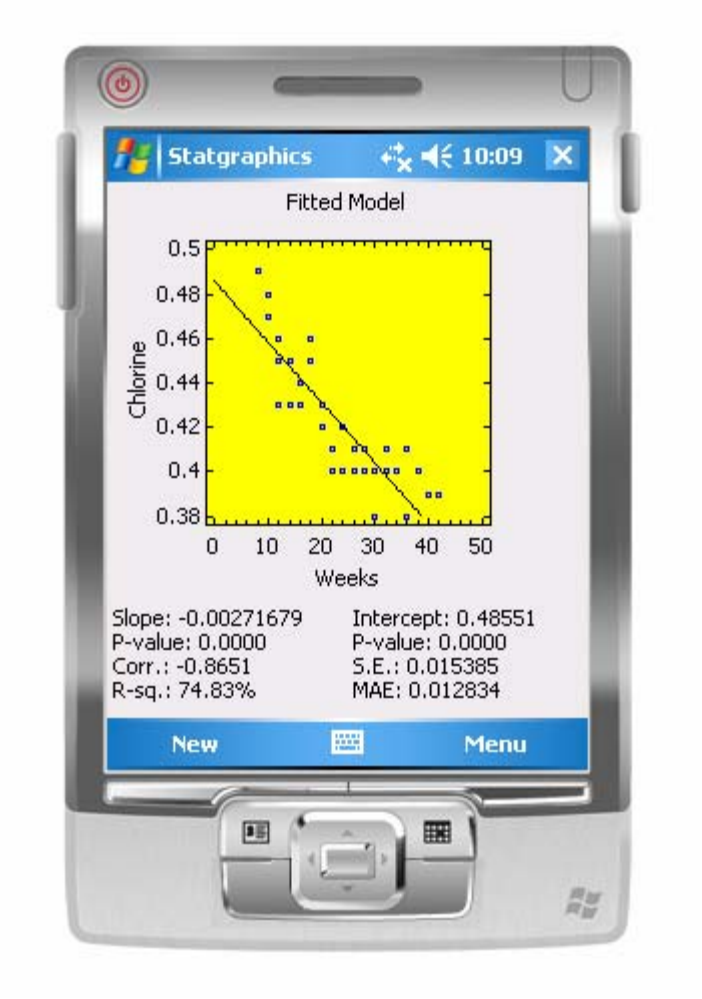
Sample Data

The text entitled Applied Regression Analysis, third edition by Draper and Smith (Wiley, 1998) contains a sample of $n = 44$ measurements of the age and amount of chlorine in samples of a product. The data is contained in the file *chlorine.sgm*. The first several rows of that file are shown below:

Row	Chlorine	Weeks
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

Fitted Model

The initial output displayed shows the fitted model, together with the data.



Also displayed are several statistics:

(1) *Slope*: the estimated slope of the line $\hat{\beta}_1$. A P-value is also given for a two-sided test of the null hypothesis that the true slope equals 0. Small P-values indicate that the slope is significantly different from 0.

(2) *Intercept*: the estimated intercept $\hat{\beta}_0$. A P-value is also given for a two-sided test of the null hypothesis that the true intercept equals 0. Small P-values indicate that the intercept is significantly different from 0.

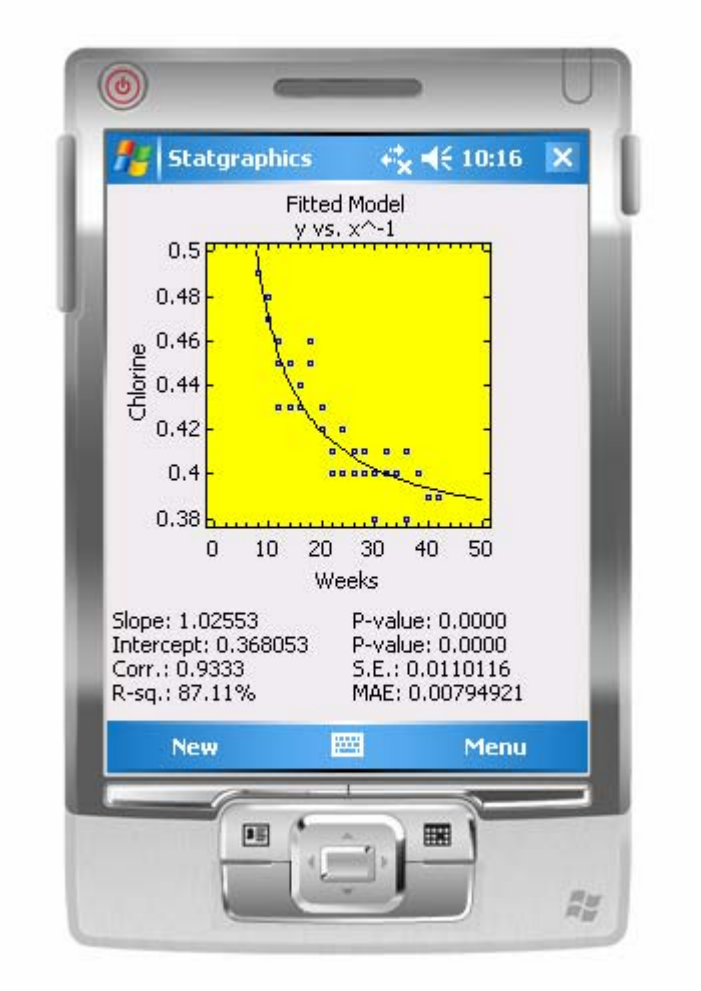
(3) *Corr.*: the estimated correlation coefficient between Y and X. The correlation coefficient *r* ranges from -1 to +1 and measures the strength of the linear relationship between the variables.

(4) *R-sq.*: the coefficient of determination R^2. R^2 ranges between 0% and 100% and measures the percentage of the variability in Y that has been explained by the regression on X.

(5) *S.E.*: the standard error of the regression $\hat{\sigma}$. This value estimates the standard deviation of repeated values of Y at the same X and is used to calculate prediction limits.

(6) *MAE*: mean absolute error. This is the average absolute value of the residuals from the fit.

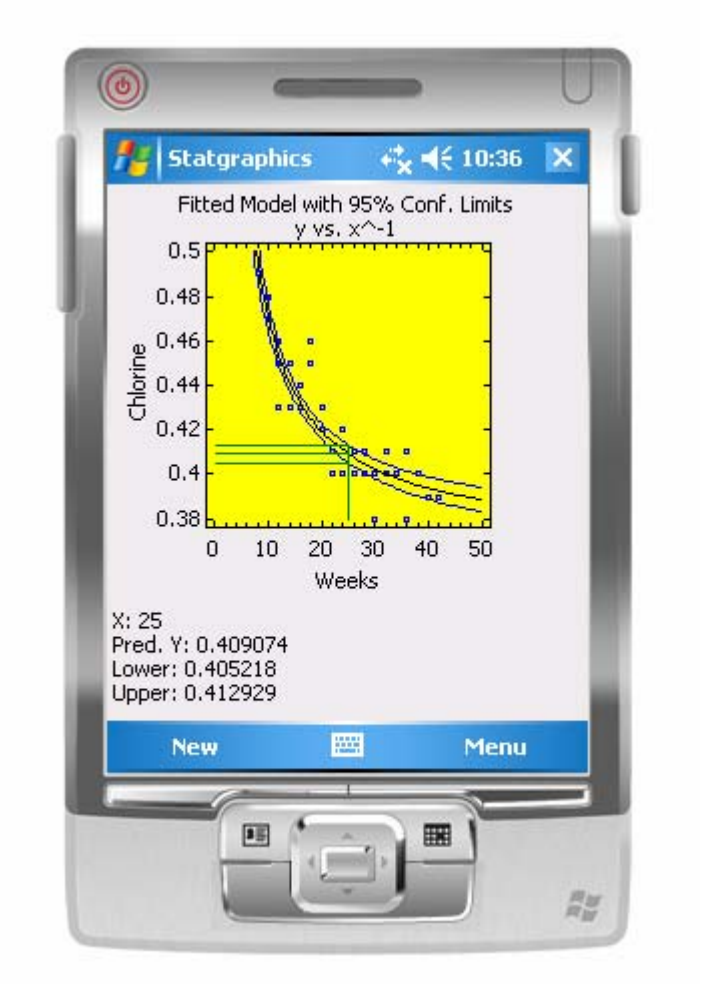
For the current data, a better fit is obtained if the weeks is transformed by taking the reciprocal of X (power = -1). The resulting fit is shown below:



All of the statistics displayed at the bottom correspond to the fitted linear model after the variables have been transformed.

Limits

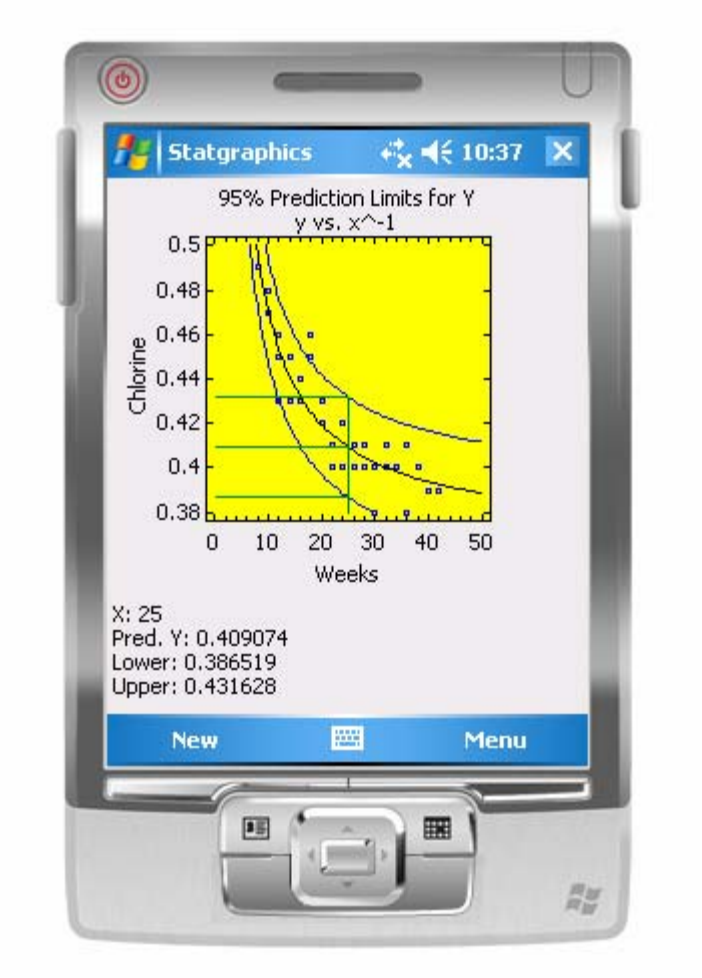
This page displays a plot of the fitted model with confidence limits for the mean value of Y at any given X.



The confidence level for the bounds is determined by the *Confidence* setting on the *Pred*. tab for the response variable's *Column Properties* dialog box. Lines are drawn and predictions displayed at a selected value of X. To change that value, click on *Menu* - *Options* – *Statistics* and specify a different value on the dialog box.

Predict

This page displays a plot of the fitted model with predictions limit for observed values of Y at any given X.

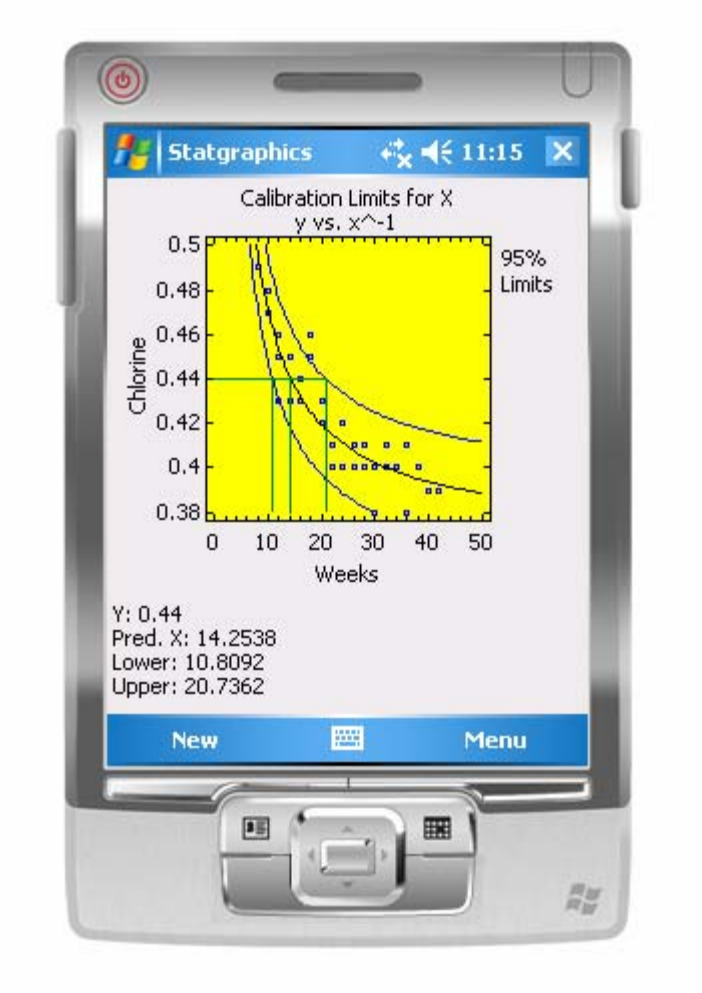


The percentage level for the bounds is determined by the *Percentiles* setting on the *Pred.* tab for the *Response* variable's *Column Properties* dialog box. That tab can also be used to specify one-sided bounds rather than two-sided limits.

Lines are drawn and predictions displayed at a selected value of X. To change that value, click on *Menu* - *Options* – *Statistics* and specify a different value on the dialog box.

Calibrate

This page displays a plot of the fitted model with prediction limits for X given an observed value of Y.

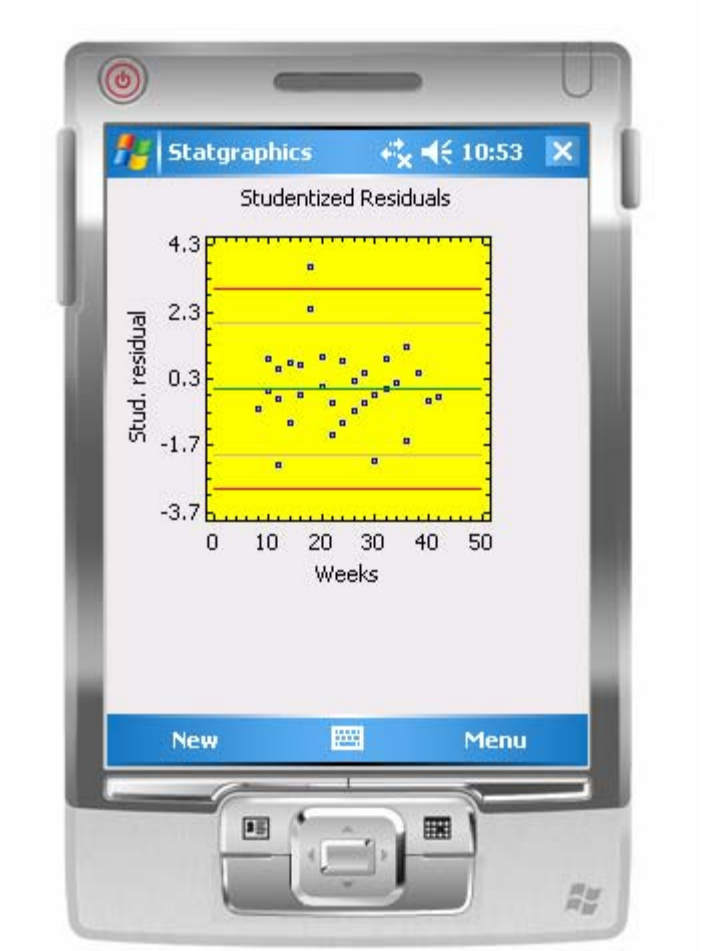


This page is similar to the *Predictions* page, except that predictions are made for X given Y instead of Y given X. This is often called the *calibration problem*, since a typical application is one in which standard samples with known values of X are used to fit a model, and then that model is used to estimate the true value of X given measurements of Y taken from new samples.

Lines are drawn and predictions displayed at a selected value of Y. To change that value, click on *Menu* - *Options* – *Statistics* and specify a different value on the dialog box.

Residuals

This page displays a plot of the Studentized residuals from the fitted model versus the values of X.



The *i*-th residual e_i is defined as the difference between the observed value of Y and the value $\hat{y}_i$ predicted by the fitted model:

$$e_i = y_i - \hat{y}_i \tag{2}$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3}$$

The *i*-th Studentized deleted residual, which is displayed in the plot, reexpresses e_i in terms of how many standard deviations it is away from the fitted model, when that model is fit using all observations *except* the *i*-th. This prevents a large residual from masking its presence by having too great an influence on the fitted line.

Horizontal reference lines are drawn at 0, 2-sigma, and 3-sigma.

Calculations

Least Squares Estimates

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{4}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{5}$$

where

$$S_{xx} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \tag{6}$$

$$S_{xy} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) \tag{7}$$

Correlation Coefficient

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} \tag{8}$$

where

$$S_{YY} = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \tag{9}$$

R-squared

$$R^2 = r^2 \tag{10}$$

Limits

$$\text{Confidence limits: } \hat{y} \pm t_{\alpha/2,n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\left(x - \bar{x}\right)^2}{S_{xx}}} \tag{11}$$

$$\text{Prediction limits: } \hat{y} \pm t_{\alpha/2,n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{\left(x - \bar{x}\right)^2}{S_{xx}}} \tag{12}$$

Standard Error

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}\right)^2}{n-2}} \tag{13}$$

Mean Absolute Error

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}|}{n} \tag{14}$$

Inverse Predictions

$$\hat{x}_{new} = \frac{y_{new} - \hat{\beta}_o}{\hat{\beta}_1} \tag{15}$$

Lower and upper limits for x_{new} are found using Fieller's approach, which solves for the values of $\hat{x}_{new}$ at which the prediction limits

$$\hat{y} \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(\hat{x}_{new} - \bar{x}\right)^2}{S_{XX}}} \tag{16}$$

are equal to y_{new}.

Studentized Residuals

$$d_i = \frac{e_i}{\hat{\sigma}_i\sqrt{\left(1 - h_i\right)}} \tag{17}$$

where

$$h_i = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{S_{xx}} \tag{18}$$

and

$$\hat{\sigma}_i = \sqrt{\frac{\left(n-2\right)\hat{\sigma}^2 - \frac{\left(x_i - \bar{x}\right)^2}{1 - h_i}}{n-3}} \tag{19}$$