QTEGRA DATA PROCESSING ALGORITHMS

Basic Mathematical Methods

Qtegra Evaluation

Author: Lars Fabian Paape

Version: 0.2

Status: Draft

Date: 23 Octobers 2013

Content

1 Document History

Version	Status	Date	Modifier	Comments
0.1	Draft	22-August-2013	Lars Fabian Paape	First version documenting the Qtegra-wide used basic mathematical methods.
0.2	Draft	23-October-2013	Lars Fabian Paape	Plotting Error Estimates section added.

2 References

[1]: Not yet documented, Rev. A.2, E. Wapelhorst, 2012
[1]: Not yet documented, Rev. A.2, E. Wapelhorst, 2012
[3]: Regression Coefficient, Wolfram MathWorld, www.mathworld.com
[4]: Estimating Errors in Least-Squares Fitting, TDA Progress Report 42-122, Nasa , 1995
[5]: Least-Squares Fitting, Wolfram MathWorld, www.mathworld.com
[6]: Statistical Definitions, Wikipedia, the free encyclopedia, http://www.wikipedia.org
[7]: Theorie zur linearen Regression, Department of Chemistry University Basel, 2000

3 Purpose

This document describes the system-wide used basic mathematical methods. It includes information about simple statistical calculations, linear and polynomial regression as well as Gaussian error estimation for the before mentioned regression analysis.

4 Basic Mathematical Methods

4.1 Statistical Calculations

4.1.1 Average

For k given values $x_0, x_1, ..., x_k$ the average or mean value $\bar{x}$ is defined by

$$\bar{x} = \frac{\sum_{i=0}^{k} x_i}{k}.$$

Excel-Formula: AVERAGE

4.1.2 Standard Deviation (SD)

For k given values $x_0, x_1, ..., x_k$ the corrected sample standard deviation s is defined by

$$s = \sqrt{\frac{1}{k-1}\sum_{i=0}^{k}(x_i - \bar{x})^2}.$$

Excel-Formula: STDEV

4.1.3 Relative Standard Deviation (RSD)

With $\bar{x} \neq 0$ the relative standard deviation (RSD) of the k values $x_0, x_1, ..., x_k$ is given by

$$RSD = \frac{s}{\bar{x}} \cdot 100\%.$$

Excel-Formula: STDEV/AVERAGE

Application Note: If the area of the software where the before mentioned formulas apply allows exclusion of values the value k and the set of used values is defined by the included values.

4.2 Linear and Quadratic LQF

For given data pairs $(x_0,y_0),(x_1,y_1),...,(x_k,y_k)$ and the polynomial function $f_l(x) = a_l x^l + \cdots + a_1 x + a_0$ the solution of a least squares fit is given as the set of parameters $a_0,a_1,...,a_l$ which minimizes the sum:

$$\sum_{i=0}^{k} [y_i - f_l(x)]^2 = \text{Min!}$$

or in case of weighting:

$$\sum_{i=0}^{k} w_i [y_i - f_l(x)]^2 = \text{Min!}$$

4.2.1 Linear Least Squares Fit (LLQF)

The curve used for solving the least squares fit problem is given as $f(x) = a_1 x + a_0$. This kind of least squares fit is also commonly known as the determination of the linear regression curve. Its parameters a_1, a_0 can be easily determined by calculating:

$$a_0 = \frac{\sum_{i=0}^{k} y_i \sum_{i=0}^{k} x_i^2 - \sum_{i=0}^{k} x_i \sum_{i=0}^{k} x_i y_i}{k\sum_{i=0}^{k} x_i^2 - (\sum_{i=0}^{k} x_i)^2}$$

$$a_1 = \frac{k\sum_{i=0}^{k} x_i y_i - \sum_{i=0}^{k} x_i \sum_{i=0}^{k} y_i}{k\sum_{i=0}^{k} x_i^2 - (\sum_{i=0}^{k} x_i)^2}$$

Weighting can be easily applied to the formula.

4.2.1.1 Regression Coefficient

Using linear regression the slope a_1 of the regression function $f(x) = a_1 x + a_0$ is called regression coefficient.

4.2.2 Quadratic Least Squares Fit (QLQF)

Uses the same basic algorithm with the difference that $\mathrm{f}(x)$ is defined as $\mathrm{f}(x) = a_2 x^2 + a_1 x + a_0$. Solving the following system of linear equations solves the problem:

$$A \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{k} 1 & \sum_{i=0}^{k} x_i & \sum_{i=0}^{k} x_i^2 \\ \sum_{i=0}^{k} x_i & \sum_{i=0}^{k} x_i^2 & \sum_{i=0}^{k} x_i^3 \\ \sum_{i=0}^{k} x_i^2 & \sum_{i=0}^{k} x_i^3 & \sum_{i=0}^{k} x_i^4 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{k} y_i \\ \sum_{i=0}^{k} x_i y_i \\ \sum_{i=0}^{k} x_i^2 y_i \end{bmatrix}.$$

Weighting can be easily applied to the formula.

4.2.3 Polynomial Least Squares Fit s(PLQF)

This system of equations shown under 4.2.2 can easily be extended to polynomials of grade>2 (see Vandermonde matrix).

4.2.4 Exponential Fit

For given data pairs $(x_0,y_0),(x_1,y_1),...,(x_k,y_k)$ together with the exponential function $\mathrm{f}\ (x) = a_1 e^{a_0 x}$ the solution of a least squares fit with parameters a_0, a_1 can be calculated by using the logarithmic transformation

$$\ln(\mathrm{f}\ (x)) = \ln(a_1) + a_0 x = a_0 x + a_1^{'} \text{ with } a_1^{'} = \ln(a_1)$$

and solving the linear regression for the transformed problem.

4.2.5 Coefficient of Determination

For n given values $y_0, y_1,..., y_n$ the general definition of the coefficient of determination R^2 is given by

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}}$$

with

$$SS_{res} = \sum_{i=0}^{n} \left[y_i - \mathrm{f}_l\right]^2$$

and

$$SS_{tot} = \sum_{i=0}^{n} \left[y_l - \bar{y}\right]^2 .$$

R^2 can be seen to be related to the unexplained variance, since the second term compares the unexplained variance (variance of the model's errors) with the total variance (Wikipedia).

Note: In linear least squares regression with an estimated intercept term, R^2 equals the square of the Pearson correlation coefficient between the observed and modeled data values of the dependent variable.

Excel-Document: Qtegra Data Processing Regression Sample.xlsx

4.3 Error Estimation and Confidence Intervals

4.3.1 Error estimation for linear regression

Let

$$\hat{y} = bx + a$$

be the formula resulting for the linear regression where $(x_i, y_i)_{1..n}$ are the defining data points. The residual e_i is defined as $e_i = y_i - \hat{y}_i$ and the sum of squared errors (SSE) is given by

$$SSE = \sum_{i=1}^{n} e_i^2.$$

Since errors are obtained after calculating two regression parameters from the data, errors have $n - 2$ degrees of freedom. Therefore the mean squared errors (MSE) result in

$$s_e{}^2 = MSE = \frac{SSE}{n-2}.$$

Using the MSE the standard error for the parameters a and b can be written as

$$s_a = s_e \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{(\sum_{i=1}^{n} x_i - n\bar{x}^2)}\right)},$$

$$s_b = \frac{s_e}{\sqrt{(\sum_{i=1}^{n} x_i - n\bar{x}^2)}}.$$

The error of the estimated linear function value at x can be calculated by

$$error(x) = \sqrt{(s_a)^2 + (x \cdot s_b)^2} \text{ (Gaussian Error Estimation).}$$

4.3.2 Error Estimation for Exponential Fit

The exponential function $f(x; a, b)$ is given by

$$f(x; a, b) = ae^{bx}.$$

Using the substitution $a' = \ln(a)$ the formula can be written as $f(x; a', b) = e^{bx+a'}$ and

$$y' = \ln\left(f(x; a', b)\right) = bx + a'.$$

Using logarithmic transformation the estimation of the parameters a, b can be solved as a linear regression problem. By using the error estimations for the linear regression problem one gets the following standard errors:

$$s_{a'} = s_e \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{(\sum_{i=1}^{n} x_i - n\bar{x}^2)}\right)},$$

$$s_b = \frac{s_e}{\sqrt{(\sum_{i=1}^{n} x_i - n\bar{x}^2)}}.$$

Using Gaussian error estimation the standard error for the parameter a results in

$$s_a = s_{a'} e^{a'} = e^{a'} s_e \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{(\sum_{i=1}^{n} x_i - n\bar{x}^2)}\right)}.$$

The error of the estimated function value at x can be calculated by

$$error(x) = \sqrt{\left(e^{bx+a'} \cdot s_{a'}\right)^2 + \left(e^{bx+a'} \cdot x \cdot s_b\right)^2}$$ (Gaussian Error Estimation).

4.3.3 Confidence Intervals for Error Estimates

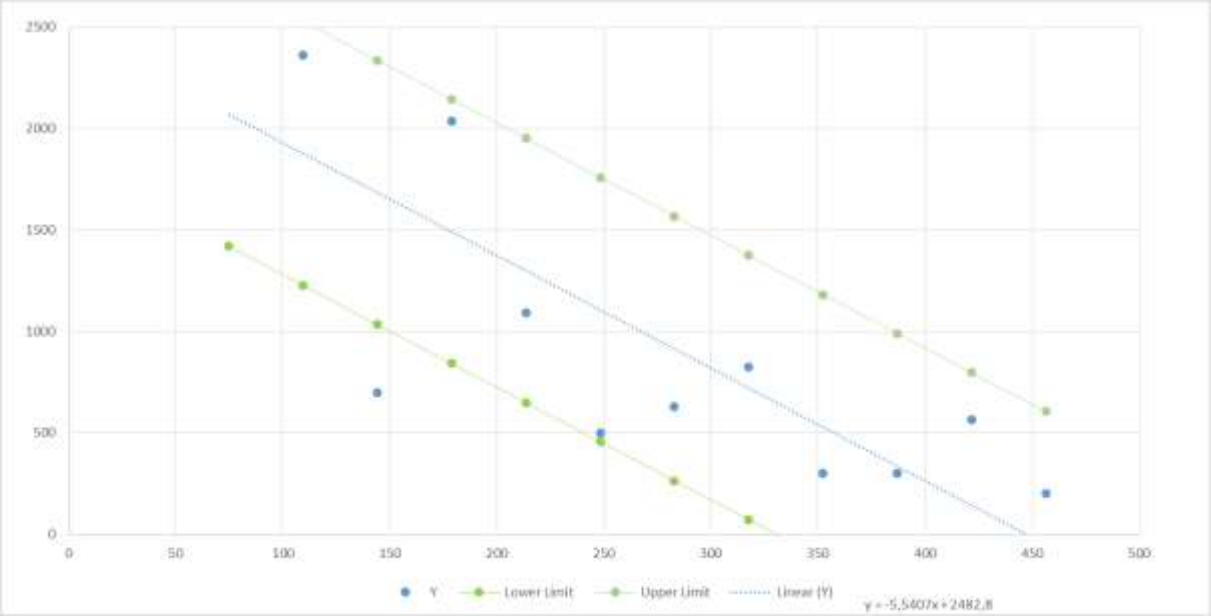
The $100(1-\alpha)\%$ confidence intervals for a and b can be computed using $t[1-\alpha/2, n-2]$ - the 1-α/2 quantile of the t distribution with n-2 degrees of freedom. The confidence intervals are

$$[a - ts_a, a + ts_a] \text{ and } [b - ts_b, b + ts_b].$$

Note: If a confidence interval includes zero then the regression parameter cannot be considered different from zero at the $100(1-\alpha)\%$ confidence level. In case of a preset parameter (i.e. forcing through a value) there are n-1 degrees of freedom.

4.3.4 Plotting Error Estimates

Having a Gaussian Error Estimation $error(x)$ for the given fit function $f(x)$ and using $t[1-\alpha/2, n-2]$ - the 1-α/2 quantile of the t distribution with n-2 degrees of freedom the following plot shows the function and its error band at a $100(1-\alpha)\%$ confidence delta by plotting the interval $[f(x) - t \cdot error(x), f(x) + t \cdot error(x)]$ over x.



Plot from *Excel-Document:* Qtegra Data Processing Regression Sample.xlsx

Note: eQuant displays the error band in a light green color, see example.