Reference guide for
# Reference Value Advisor v2.1

**NATIONAL
VETERINARY
SCHOOL
TOULOUSE**

February 2012

ReferenceValue Advisor is a set of Excel macros that compute reference intervals from data contained in spreadsheet. It closely follows the CLSI guideline [1].
It also allows to study how covariates are linked to reference intervals.

Data
The data should be contained in columns whose first line indicates the name of the analyte for which a reference interval is to be computed.

A click on the button gives the following menu
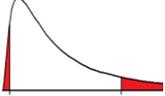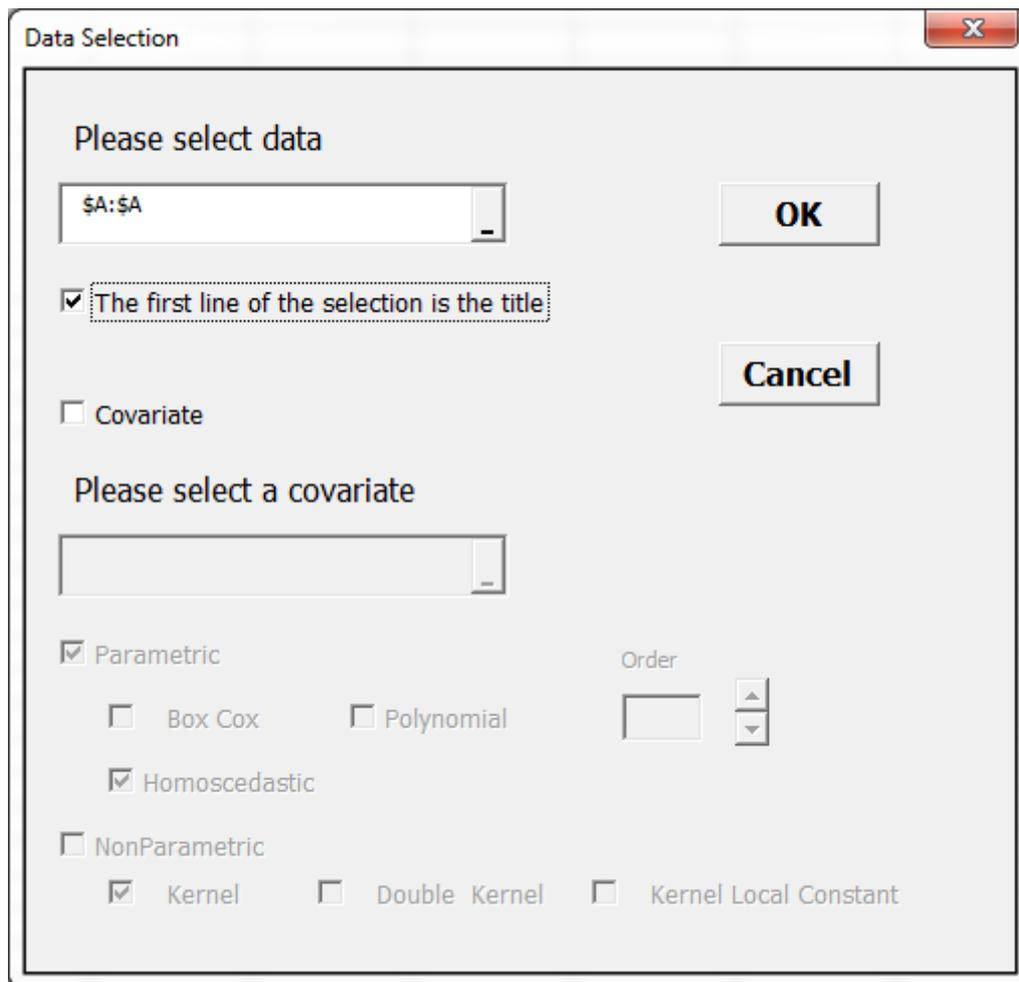
Select the data with the mouse and click on the OK button.

Results given by Reference Value Advisor
Reference Value Advisor creates
- for each analyte a spreadsheet containing the report of the analysis,
- for all analytes a spreadsheet containing information about outliers,

As an example open the file Example
This file contains three columns: Creat, Urea and age.

Selecting the columns Creat and Urea produces three spreadsheets: Report for Creat, Report for Urea, Outliers Analysis.
The spreadsheet Outliers Analysis contains outliers' information for both Creat and Urea.

Let us see now in detail the information given in the report spreadsheet.
This spreadsheet contains 2 pages reproduced hereafter.

Reference Value Advisor V2.1	**Results for Creat**				
	Date		13/02/2012		
	Performed by		Didier		

	Untransformed data		Box-Cox transformed data		
Method	Standard	Robust	Standard	Robust	Nonparametric
N	1460	1460	1460	1460	1460
Mean	97.7		32.4		
Median	97.0	97.4	32.2	32.4	
SD	15.4	15.5	7.1	7.1	
Mininum	50	50	8.1	8.1	
Maximum	149	149	54.5	54.5	
λ_1 coefficient Box-Cox			-37.842	-37.842	
λ_2 coefficient Box-Cox			0.808	0.808	
P-Value **Anderson-Darling**/	0.013		0.228		
Symmetry test for Robust		0.000		0.000	
Outliers **Dixon**					
Outliers **Tukey**	0	0	0	0	
Suspect data **Tukey**	8	8	7	7	
Lower limit of reference interval	67.4	67.1	68.6	68.6	69.0
Upper limit of reference interval	128.0	127.7	129.1	129.0	128.5
90% CI for lower limit	66.4	65.9	67.6	67.6	68.0
	68.6	68.3	69.6	69.6	70.0
90% CI for upper limit	126.8	126.6	127.8	127.8	127.0
	129.2	128.9	130.3	130.2	131.0

Comments

Suspect data detected according to Tukey. IFCC-CLSI C28-A3 recommends that unless these data
are known to be aberrant observations, the emphasis should be on retaining rather than deleting them.
The sample size is large enough to compute the nonparametric reference interval: [69 ; 128.5].
The 90% CI of one (or more) limit is larger than recommended in IFCC-CLSI C28-A3.

The standard method

The standard method assumes the data distribution is Gaussian and the 95% reference interval is computed as

$$\left[\hat{m} - t_{N-1}^{0.975}\hat{\sigma}\sqrt{\frac{N+1}{N}}; \hat{m} + t_{N-1}^{0.975}\hat{\sigma}\sqrt{\frac{N+1}{N}}\right] \qquad \text{(eq. 1)}$$

where

N is the sample size named as N in the output table,

$\hat{m}$ is the sample mean. Its value is given in the Mean row,

$\hat{\sigma}$ is the sample standard deviation whose value is given in the SD row,

$t_{N-1}^{0.975}$ is the 0.975 quantile of a Student distribution with N-1 degrees of freedom.

The lower (resp. upper) limit of the reference interval is given in the row entitled Lower limit of reference interval (resp. Upper limit of reference interval) for the columns named Untransformed data.

The robust method

The robust method assumes that the data distribution is symmetric without being necessarily Gaussian. It is built using robust statistics (median and median absolute deviation) that are less sensitive to outliers. This reference interval is obtained by a process that iteratively gives a smaller weight to data far from the central location. The method is fully explained in [2] and the tuning parameters that need to be used for this computation are given in [1].

The 95% reference interval is computed with the robust method is given by

$$\left[T - t_{N-1}^{0.975}SD; T + t_{N-1}^{0.975}SD\right] \qquad \text{(eq. 2)}$$

where

N is the sample size named as N in the output table,

T is an estimate of the central location of the data distribution. Its value is given in the Median row,

SD is an estimate of the data dispersion whose value is given in the SD row

$t_{N-1}^{0.975}$ is the 0.975 quantile of a Student distribution with N-1 degrees of freedom.

The lower (resp. upper) limit of the reference interval is given in the row entitled Lower limit of reference interval (resp. Upper limit of reference interval).

The robust method is sensitive to asymmetry of the distribution. For this reason, the data distribution should be first carefully inspected.

Confidence intervals of the limits

Confidence intervals (CI) of the reference limits are computed for all methods.

For the standard method the 90% CI is obtained using parametric bootstrap when N≤20.

In all other cases a nonparametric bootstrap is used.

Reference Interval Advisor spends most of the computation time to compute these confidence intervals.

The 90% confidence interval of the lower (resp. upper) reference limit is given in the rows 90% CI for lower limit (resp. 90% CI for upper limit).

The width of the confidence intervals depends on the sample size N and on the data dispersion. When N is low, the confidence intervals are wide while large N give narrow confidence intervals. The confidence interval of the reference limits should be given whatever the sample size used for the computation. They give useful information on the precision of the computed reference interval.

The generalized Box-Cox transformation

The generalized Box-Cox transformation columns contains information about the data after the following Box-Cox transformation

$$h_\lambda(x) = \begin{cases} \dfrac{(x+\lambda_1)^{\lambda_2} - 1}{\lambda_2} & \text{if } \lambda_2 \neq 0 \\ \log(x+\lambda_1) & \text{otherwise} \end{cases}$$

The parameters λ_1 and λ_2 that enter in the Box-Cox transformation are chosen so that the transformed data distribution is as Gaussian as possible. The λ values given in the λ coefficient **Box-Cox** rows optimize the likelihood. An interval is first build using equation (eq. 1) on the Box-Cox transformed data. This interval is then back-transformed using $h_\lambda^{-1}(.)$ to give the reference interval given in the rows Lower/Upper limit of reference interval for the columns Box-Cox transformed data. The function $h_\lambda^{-1}(.)$ is given by

$$h_\lambda^{-1}(y) = \begin{cases} (1+\lambda_2 y)^{1/\lambda_2} - \lambda_1 & \text{if } \lambda_2 \neq 0 \\ \exp(y) - \lambda_1 & \text{otherwise} \end{cases}.$$

Normality and symmetry tests

Normality is required by the standard method for an accurate computation of the reference interval. The row p-value **Anderson-Darling/symmetry test for Robust** gives the result of the normality test for both the untransformed and the Box-Cox transformed data. When it is little than 0.05, normality should be questioned. Non normality can be due to a non-normal distribution or presence of outliers. There exist other tests for normality such as Kolmogorov-Smirnov test, Chi-square test… We deliberately chose to retain only the Anderson-Darling test as it mainly looks for deviation from normality in the tails of the distribution which are of primary importance for the reference interval determination.

The robust method required the data distribution to be symmetrical about its median. For the robust method, the result of the symmetry test is contained in the row p-value **Anderson-Darling/symmetry test for Robust**. When this P-value is less than 0.05, symmetry should be questioned and the robust method should not be used. This test is completely described in [4].

Outliers

An outlier is a value that has a small probability to have been observed in the reference population. Outliers can distort a reference interval determination. There exist three kinds of outliers:

1. data that are the results of a mistake. These data should be removed from the analysis.
2. data that have been collected on a patient that does not belong to the reference population. If the reference population contains values for healthy patients, other results such as analytes concentrations/ clinical examination can help to detect these patients. When detected, these data should be removed from the analysis.
3. data that are far from the other. This situation is touchier than the two others as it can be explained by 1°), 2°) or by a wrong choice of distribution to describe the data. It can be identified by a large number of values detected as suspect or as outliers. When the chosen distribution does not fit the data, it is reasonable to change the distribution not the data.
   For this reason, we suggest
   a)   to use the robust method [2] described hereafter that is less sensitive to the shape of the chosen distribution.

5

b)   to compute the reference interval after removing these data and to compare the reference intervals obtained with and without the potential outliers. If the reference interval does not change too much then keeps this reference interval.

c)   to get more data (at least 120) so that the nonparametric method can be used. As we will see, the nonparametric method does not assume any specific shape for the data distribution.

d)   to change the distribution. As the current version of this code does not propose other distributions, a solution is to consult your preferred statistician.

The three lines Outliers Dixon, Outliers Tukey and Suspect data Tukey should help you to detect such values.

The row Outliers **Dixon** gives the results of the tests for the minimum and the maximum values of the untransformed/Box-Cox transformed data. The corresponding cells table can contain Min, Max or Min/Max respectively meaning that the minimum, the maximum, both the minimum and the maximum values are outliers. When these cells are empty, the Dixon test did not detect that these values as outliers. This test is known to be rather insensitive.

The rows Outliers **Tukey** and Suspect data **Tukey** respectively contain the number of outliers and suspect data identified using the Tukey method [3].
Tukey differentiated "mild" and "extreme" outliers that we here respectively named suspect and outliers data.
If Q1, Q2, Q3 respectively represent the first, second and third quartiles of the distribution and IQ=Q3-Q1 is the inter-quartile range, outliers are data smaller than Q1-3 IQ or greater than Q3+3*IQ and suspect data are data contained in the interval [Q1-3 IQ; Q1-1.5 IQ] or in [Q3+1.5*IQ ; Q3+3*IQ].
The suspect data and the outliers are respectively written in orange and red cells in the columns written on the rhs of the spreadsheet.

Graphs
Box-plot

Creat
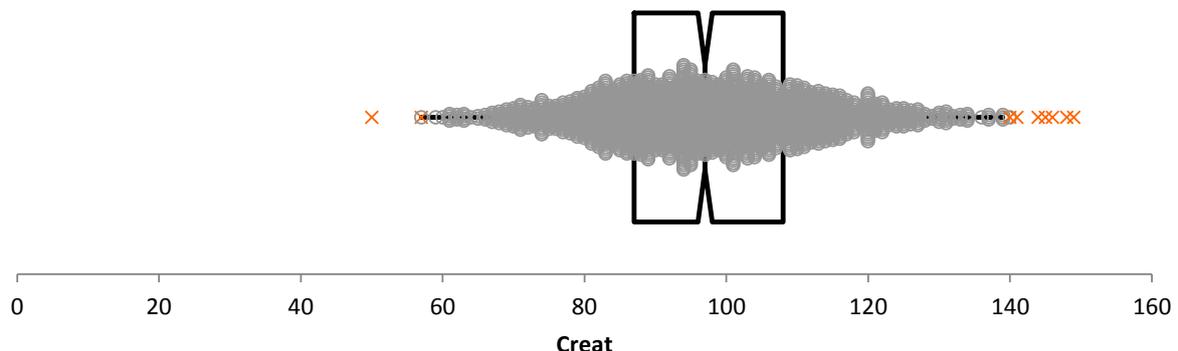
Covariates

When you expect the reference interval of your data to change with some covariates, you can try to build a reference interval for each value of the covariates. RefValAdv allows you to do such a thing but it only admits continuous covariates such as age, weight etc…

A lot of methods can be used to build reference intervals with covariates. The main differences between these methods rests on the knowledge you have on the distribution of your data: 1) you know, or you expect, that for each value of your covariates, your data are normally distributed up to some transformation 2) you do not know the distribution of your data or you expect your data to be non-normally distributed.
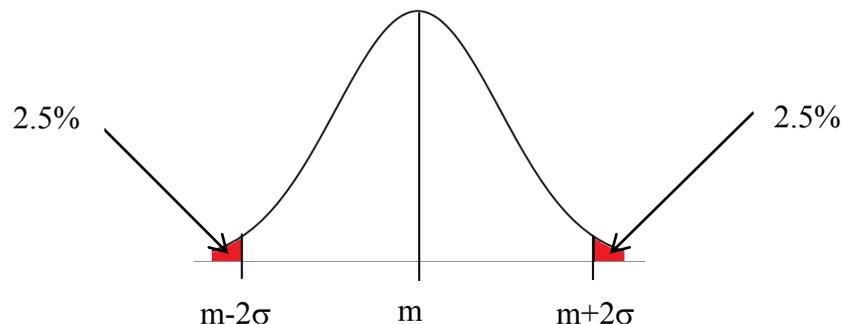
1) Parametric models

The distribution of data is described using the Gaussian distributions. Such descriptions are called models.

Recall that when the data distribution is normal with a population mean m and a population standard deviation σ, then the reference interval

$$[m-2\sigma \; ; m+2\sigma] \tag{eq. 3}$$

contains the values that can be observed in 95% of healthy individuals.

For parametric Gaussian models, we expect the population mean m to change with the covariates.

Let us take an example to fix ideas. Assume that we are interested in the building of reference intervals of creatinine for dogs of different ages. Assume that we know that the creatinine population mean is a linear function of age: $m = \theta_1 + \theta_2 \times Age.$

If the dispersion of creatinine (standard deviation σ) does not change with age, the reference interval for a given age is built as previously by:

$$[\theta_1 + \theta_2 \times Age \; -2\sigma \; ; \theta_1 + \theta_2 \times Age \; +2\sigma].$$
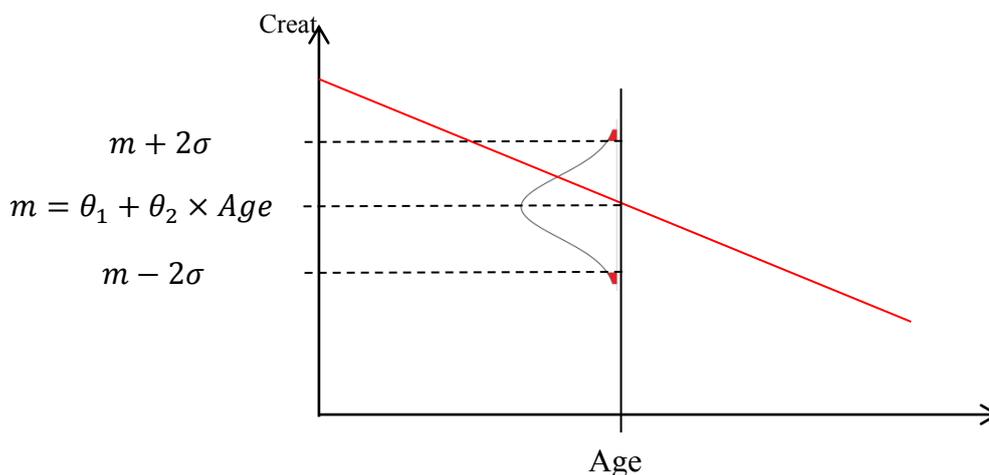
Assume now that instead of Age, we have p covariates denoted by $X_1, ..., X_P$. If the population mean m depends linearly on these covariates we have:

$$m = \theta_1 + \theta_2 \times X_1 + \cdots + \theta_{p+1} \times X_p. \qquad \text{(eq. 4)}$$

Under the assumption that the standard deviation σ does not change with these covariates, the reference interval becomes

$$[\theta_1 + \theta_2 \times X_1 + \cdots + \theta_{p+1} \times X_p - 2\sigma ; \theta_1 + \theta_2 \times X_1 + \cdots + \theta_{p+1} \times X_p + 2\sigma].$$

When the average relationship between the data and a covariate seems to be nonlinear, a polynomial can be used instead. As an example, if the creatinine population mean is a polynomial function of degree p of age:

$$m = \theta_1 + \theta_2 \times Age + \theta_3 \times Age^2 + \cdots + \theta_{p+1} \times Age^p$$

Such models with standard deviation that does not change with covariates are called *homoscedastic*.

In *heteroscedastic* models, the standard deviation is free to change with covariates. Only one kind of heteroscedasticity has been implemented in this version of Reference Value Advisor:

$$\sigma^2 = S^2(1 + a \times m)^2$$

where s and a are two parameters.

Notice that when a = 0, this model becomes homoscedastic.

For heteroscedastic models, when the population mean *m* depends on covariates as in eq. 4, the variance σ^2 is given by

$$\sigma^2 = S^2 \left(1 + a \times \left(\theta_1 + \theta_2 \times X_1 + \cdots + \theta_{p+1} \times X_p\right)\right)^2.$$

The reference interval obtained with this heteroscedasticity is

$$\left[\theta_1 + \theta_2 \times X_1 + \cdots + \theta_{p+1} \times X_p - 2\sqrt{S^2 \left(1 + a \times \left(\theta_1 + \theta_2 \times X_1 + \cdots + \theta_{p+1} \times X_p\right)\right)^2} ;\right.$$

$$\left.\theta_1 + \theta_2 \times X_1 + \cdots + \theta_{p+1} \times X_p + 2\sqrt{S^2 \left(1 + a \times \left(\theta_1 + \theta_2 \times X_1 + \cdots + \theta_{p+1} \times X_p\right)\right)^2}\right] \qquad \text{(eq. 5)}$$

which has exactly the same structure as eq. 1.

These reference intervals are built assuming that both the population mean m and the standard deviation σ are known which is never the case in practice. These values are estimated on a sample of individuals. We thus only have sample estimates that give information about the actual value of the parameters but with an error measured by the standard error (se).

When, in (eq. 5), estimates are used instead of the parameters that enters in m and σ, the general form of the reference change a little bit to account for such an error of estimation. A precise description of these changes is beyond the scope of this short explanation.

Box-Cox transformation

When for each value of the covariate the data distribution is not Gaussian, a data transformation can be used to "normalize" the distribution. Several families of transformations are used by statisticians among which, the Box-Cox transformation is likely the most famous.

The Box-Cox transformation with parameter λ (lambda) is given by

$$\frac{y^\lambda - 1}{\lambda}.$$

When λ=1 this transformation is linear implying that there is no need to transform the data to normalize their distribution.

When λ goes to 0, this transformation becomes the natural logarithm: *ln*(y). This suggests that taking the natural logarithm of the data should normalize their distribution.

Model validation
In a parametric Gaussian model, it is assumed that
- the mean population m depend on covariates as described by the model
- the dispersion of the data about the mean population m changes (or does not change) as described by the model
- for each value of the covariates, the data (y) is normally distributed.

Such assumptions need to be checked carefully. A simple graphical way to check these assumptions is to represent the weighted residuals

$$\epsilon = \frac{y - m}{\sigma}$$

as a function of covariates and/or as a function of the fitted values.

When these three assumptions hold, the weighted residuals should be in a band with a constant width about 0.

Reference for the technical part
The considered model can be written as

$$h_\lambda(Y) = X\theta + \epsilon \qquad \text{(eq. 6)}$$

where

$h_\lambda(.)$ is a vectorial function equal to identity when no Box-Cox transformation is chosen and to $\frac{y^\lambda - 1}{\lambda}$. for Box-Cox transformed data.

Y is the n-vector containing the data

X is a nxp whose columns contain the covariates

θ is a p-vector of parameters to be estimated

ϵ is an unobserved random variable assumed to be $N(0; \sigma^2 C)$

C is a diagonal matrix equal to the identity matrix I for homoscedastic models, and to diag$[(I + aX\theta)(I + aX\theta)]$ for heteroscedastic models.

The estimates of $\lambda, \theta, \sigma^2, a$ are obtained by maximization of the likelihood.

Estimating the limits of reference interval remain to find two limits L and U so that

$$P(L \le h_\lambda(Y^*) \le U) = 0.95$$

where Y^* is a new observation obtained from model given in eq.6.

For linear models (ie homoscedastic without transformation) the result is exact while the Delta-method is used when the model is non-linear. The 90% confidence interval of the reference band is obtained using the Delta-method. A complete description of the method can be found in [5].

How to do in practice

A click on the button gives the following menu
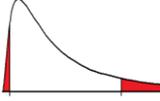
Select the data (analyte concentration) in the spreadsheet
Clicking in the "Covariate" box allows selecting one or several covariates.
The box "parametric" declares that a parametric model is to be selected.
The box "Homoscedastic" allows toggling between homo and heteroscedastic models
The box "Polynomial" allows to select the degree of the polynom. When nothing is selected, a polynom of degree 1 is selected by default.

Let us first try a heteroscedastic model with a Box-Cox transformation.
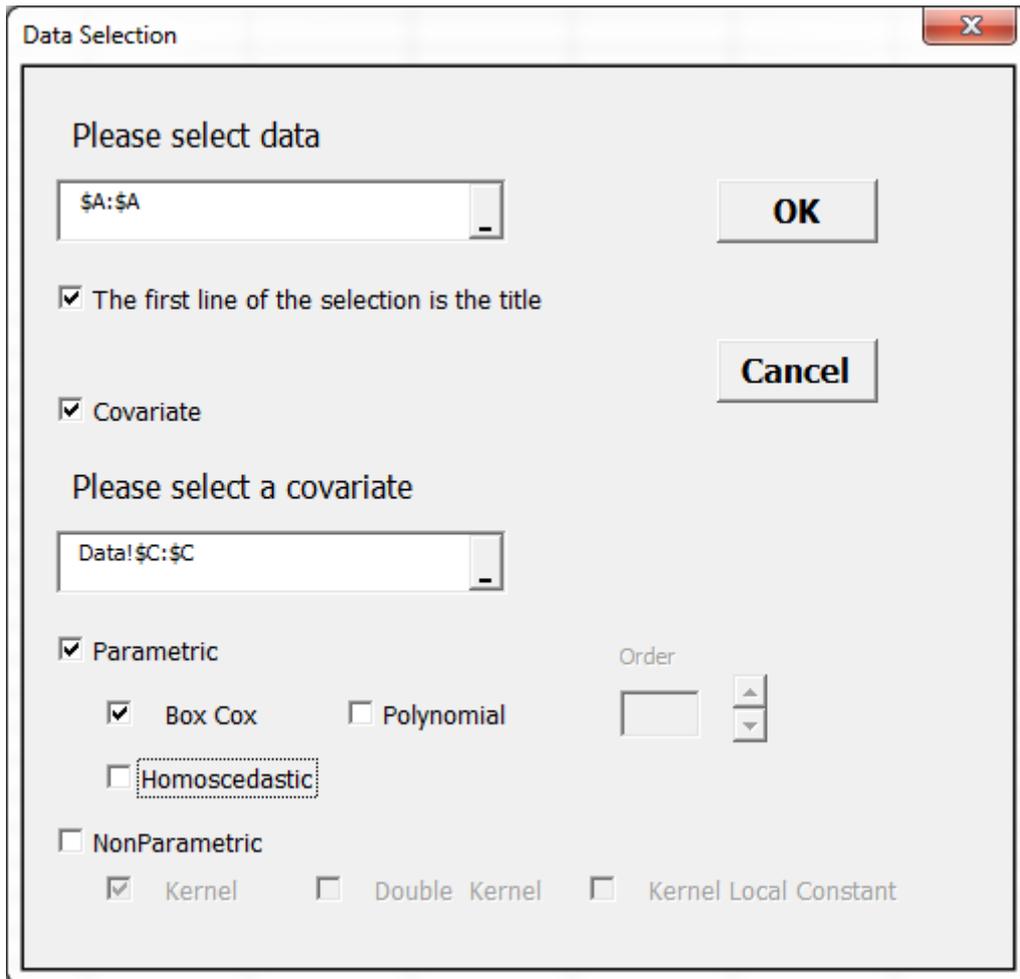
Using this model we declare that

$$\frac{\text{Creat}^\lambda - 1}{\lambda} \sim N\left(\theta_1 + \theta_2 \times \text{Age};\ S^2\left(1 + a \times (\theta_1 + \theta_2 \times \text{Age})\right)^2\right) \qquad \text{(eq. 7)}$$

Depending of your computer, for this amount of data the computation can be a bit long. Clicking on OK produces a single spreadsheet containing a table with the parameters estimate and four plots.

Table 1: This table gives the parameters estimate for model described in eq. 7.

Reference Value Advisor V2.1				Date	17/02/2012
				Performed by	Didier

Regression

Dep. var: Creat   N = 1460   Squared Multiple R   0.099

	Parameter	se	T	P(2-tail)
Intercept	26.205	0.138	189.312	0.000
Age	-0.018	0.001	-12.620	0.000
a	0.000	0.001	0.003	0.997
S2	5.750			

lambda	test for lambda=1	0.604	0.001	-513.051	0.000
	test for lambda=o			783.515	0.000

Before beginning the interpretation of these results, we have to check that our model approximately describe the data by having a look on figure 1 that represents the weighted residuals versus Age.
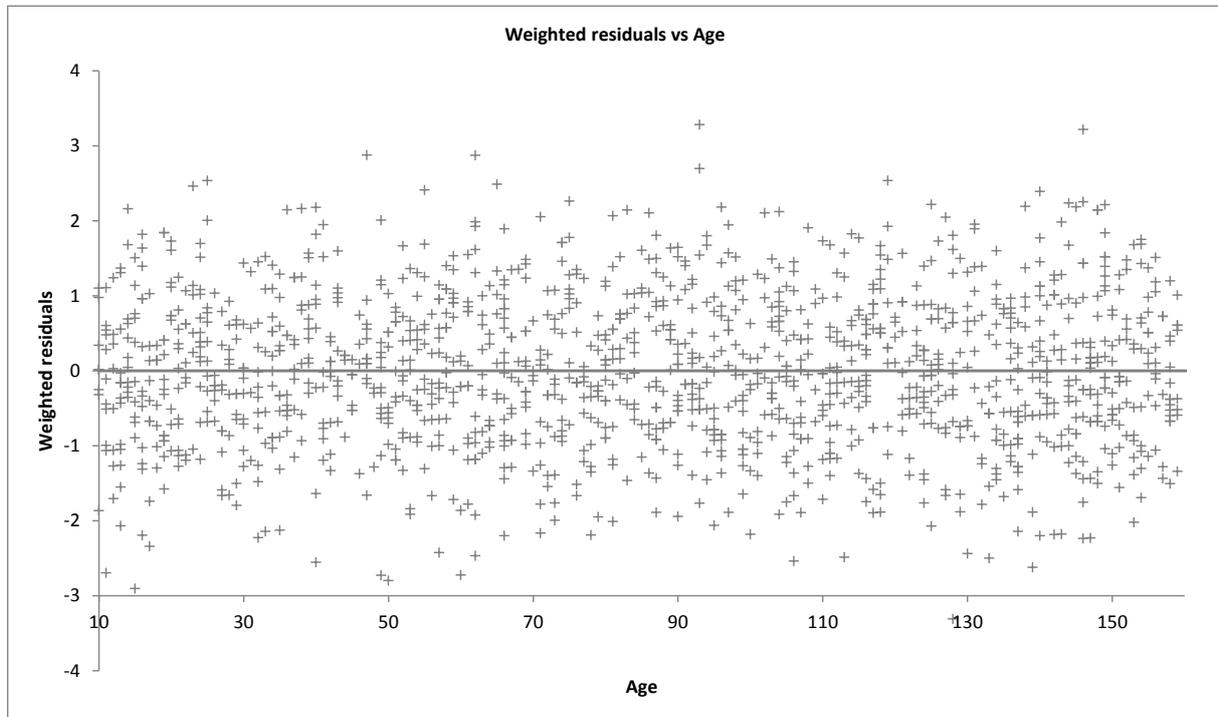

Figure 2: Weighted residuals of the regression model described by eq. 7. If the model is reasonable, the residuals should be centered about 0. No specific shape should appear. A specific shape can be a U shape or a conic shape. Here the residuals seem to be well distributed around 0.

As we can see, the weighted residuals seem to be well distributed around zero. We can thus read the results contained in table 1.

From this table we obtain an estimate of lambda (λ): 0.604. The actual value of lambda is different of 1 (P<0.001) implying that a Box-Cox is needed, and it is different of zero (P<0.001). There is no need to log-transform the creatinine concentrations.
Next, we can see that the estimate of a = 2.544×10^{-6} is so small that we can't ensure that, accounting to its uncertainty (se = 0.00075), its actual value is different of 0 (P=0.997).

Thus, this model seems to be too complex for these data. We are going to simplify it by setting a=0 *i.e.* we choose the following homoscedastic model:

$$\frac{\text{Creat}^\lambda - 1}{\lambda} \sim N(\theta_1 + \theta_2 \times \text{Age}; S^2) \qquad \text{(eq. 8)}$$

Table 2: This table gives the parameters estimate for model described in eq. 8.

Regression

Dep. var: Creat   N = 1460   Squared Multiple R   0.098

	Parameter	*se*	*T*	*P(2-tail)*
Intercept	26.205	0.138	189.312	0.000
Age	-0.018	0.001	-12.620	0.000
S2	5.751			
lambda test for lambda=1	0.604	0.001	-513.049	0.000
test for lambda=0			783.516	0.000

Analysis of variance

	SS	*Df*	*MS*	*F*	*P*
Regression	915.904	1	915.904	159.253	0.000
Residual	8385.316	1458	5.751		
Total	9301.220	1459			

Let us now comment this output.

The output contains 2 tables: first a table named "**Regression**" giving the parameters estimate and because the model is homoscedastic, an overall table named "**Analysis of variance**" giving the decomposition of variability.

We need first to examine the table "**Analysis of variance**". It says to us that the chosen model explains significantly variations of Creatinine (P<0.000).

The percentage of variability explained by this model is given by "Squared Multiple R". Actually only 9.8% of the variability is explained by the model.

The same information is contained in figure 2 that represents the data (creatinine) versus the fitted values. As the model significantly explains some variability, we can look at the table "Regression".

The lambda parameter is estimated to be: 0.604 with a standard error of 0.001. As already seen, lambda is significantly different of 1 *i.e.* there is a need for a Box-Cox transformation, and of 0 *i.e.* there is no need to log-transform the creatinine data.

The first two lines of this table respectively entitled "Intercept" and "Age" give information on the corresponding parameter. From these lines we get estimates for θ_1 (26.205) and θ_2 (-0.018). The corresponding standard errors are contained in the next column. The column *P(2-tail)* gives the result of the t-test on nullity of the parameters: both θ_1 and θ_2 are significantly different of 0.

The linear model applies to the Box-Cox transformed data:
$$\frac{\text{Creat}^{0.604} - 1}{0.604} = 26.205 - 0.018 \times \text{Age}$$
From this model, we can use the following back-transformation to obtain the fitted values:
$$\text{Creat} = [1 + 0.604 \times (26.205 - 0.018 \times \text{Age})]^{1/0.604}$$
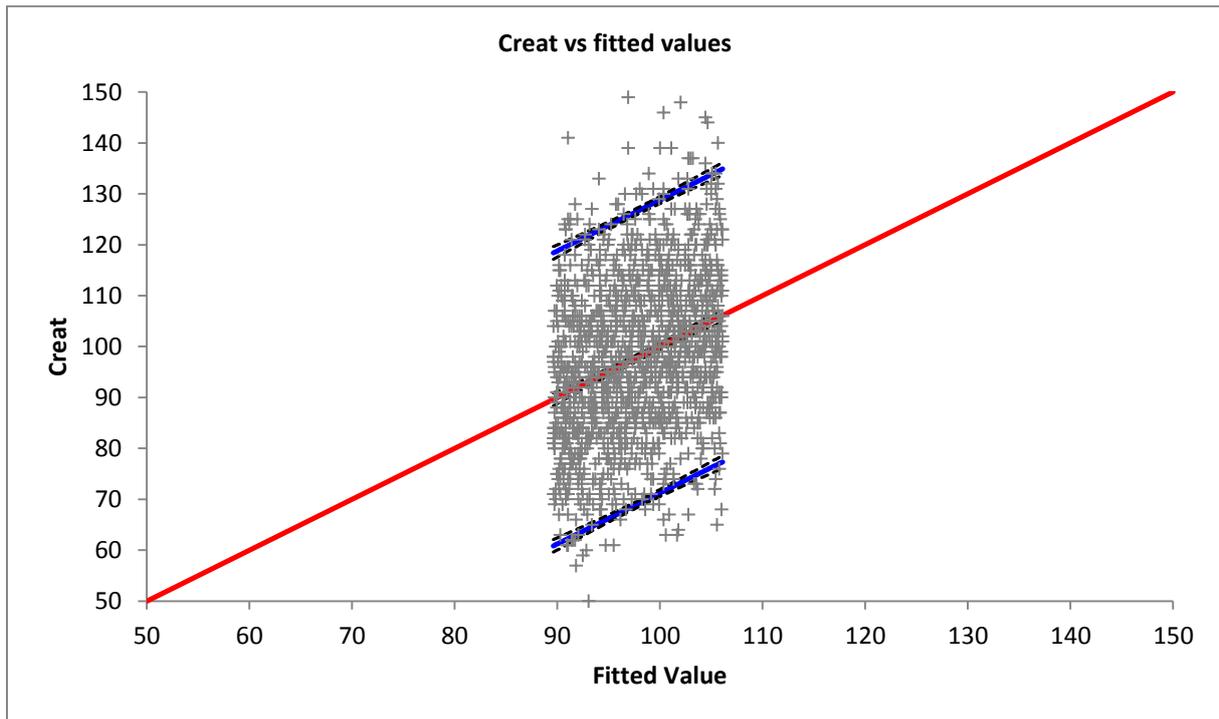These values are contained in the column named Ref 0.5.

Figure 2: The data are represented versus the fitted value obtained with the model. The red line represents the first bisector line. The magnitude of dispersion on the y-axis gives information on the actual dispersion of the data. The magnitude of dispersion on the x-axis gives information on what the model explains. We can see here that even if the model explains some dispersion, it does not explain a large part of the dispersion. Squared multiple R quantifies how much dispersion the used model actually explains. It is here equal to 9.8 %

QQplot of Weighted residuals

Creat vs Age with BC transformation

15
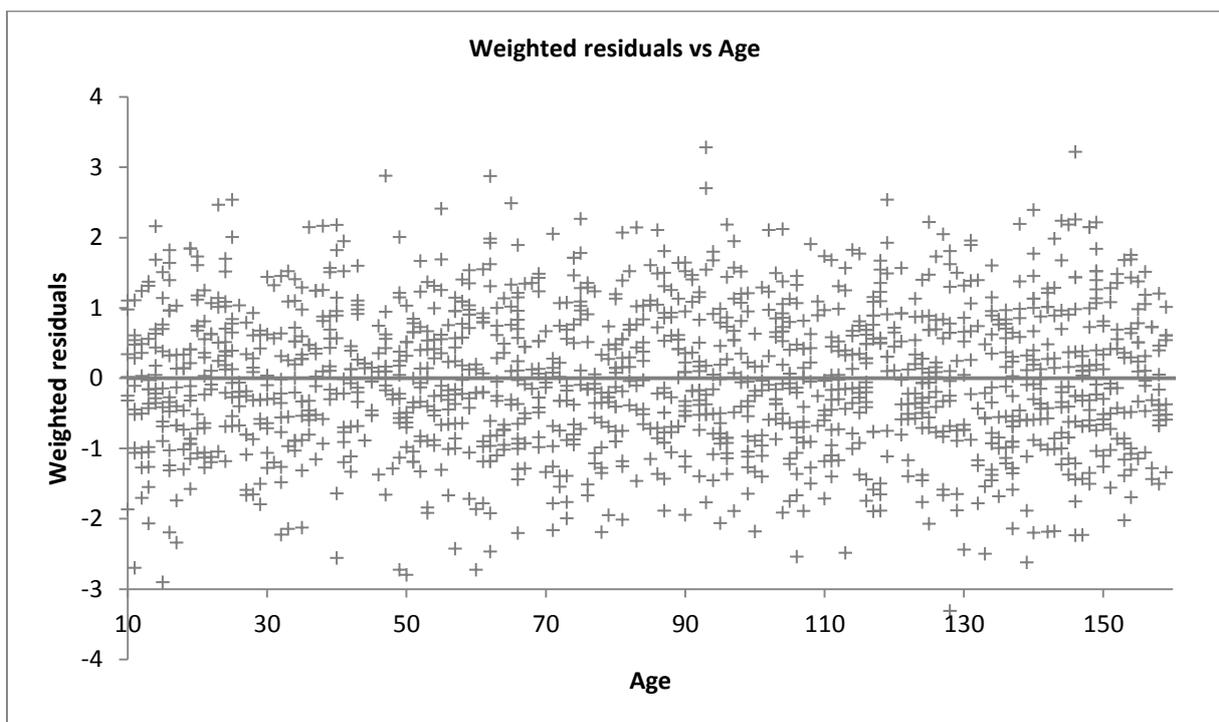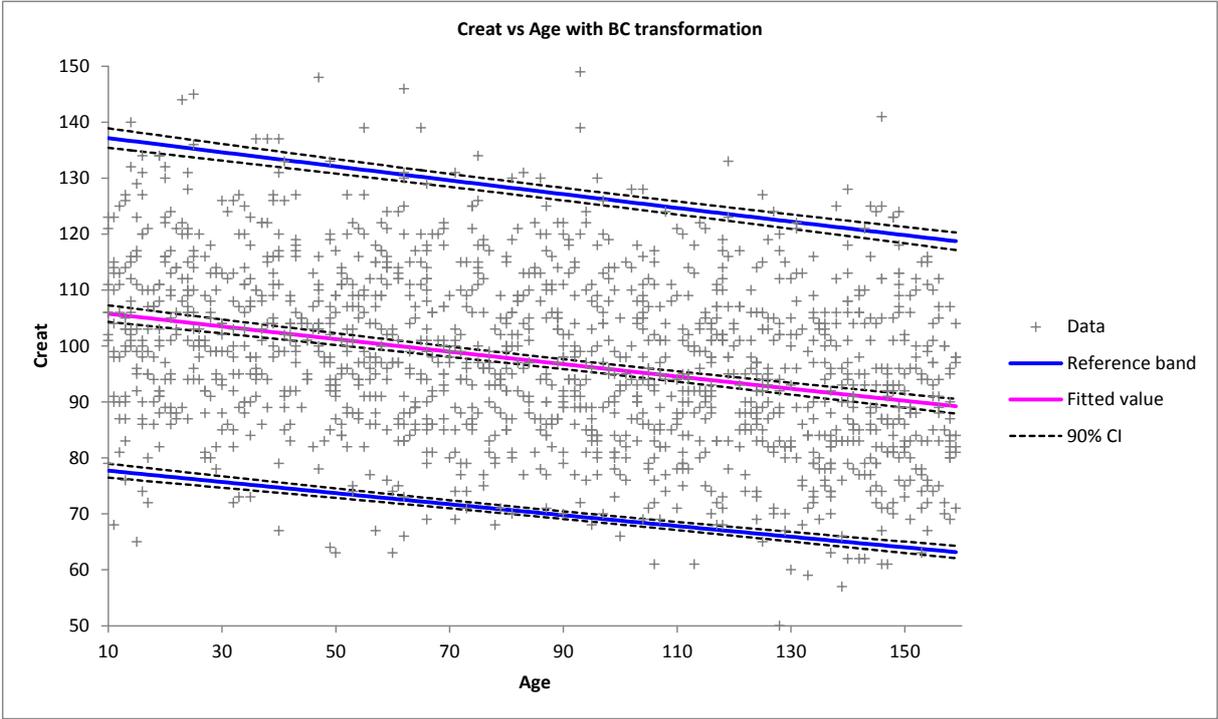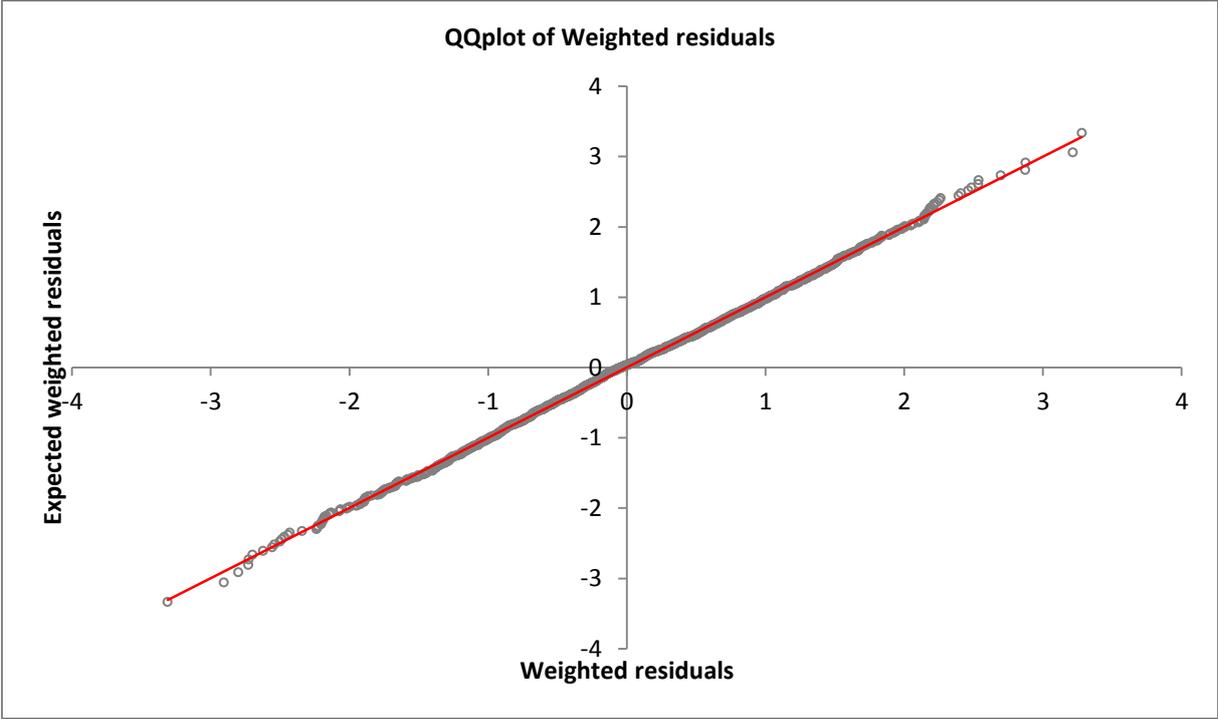
2) Non Parametric models

In such models, the distribution of the data is estimated as a weighted mixture of simple distributions. In this version of RefValAdv a single covariate can be used. Non parametric models can be used only when the number of data is sufficient. These models should thus be used with large datasets.

These models only assume that data are independent *i.e.* follow-up data should not be analyzed using these methods. They do not assume any specific relationship between the covariate and the analyte for which the reference band is to be built. In this respect, they are very robust to a departure to usual distributional assumptions required to build reference bands.

Four methods can be used to build nonparametric reference bands.

1) The Kernel method use a Gaussian kernel for the x-axis
2) The double Kernel method use a Gaussian kernel for x-axis and an uniform kernel for the y-axis (cf [])
3) A local constant kernel smoother's method
4) A local linear kernel smoother's method

A full explanation of these methods is beyond the scope of this short note (cf [6] and [7]). However, if the sample size is large enough, the reference bands obtained with these methods can be used safely.

Creat vs Age using method Kernel
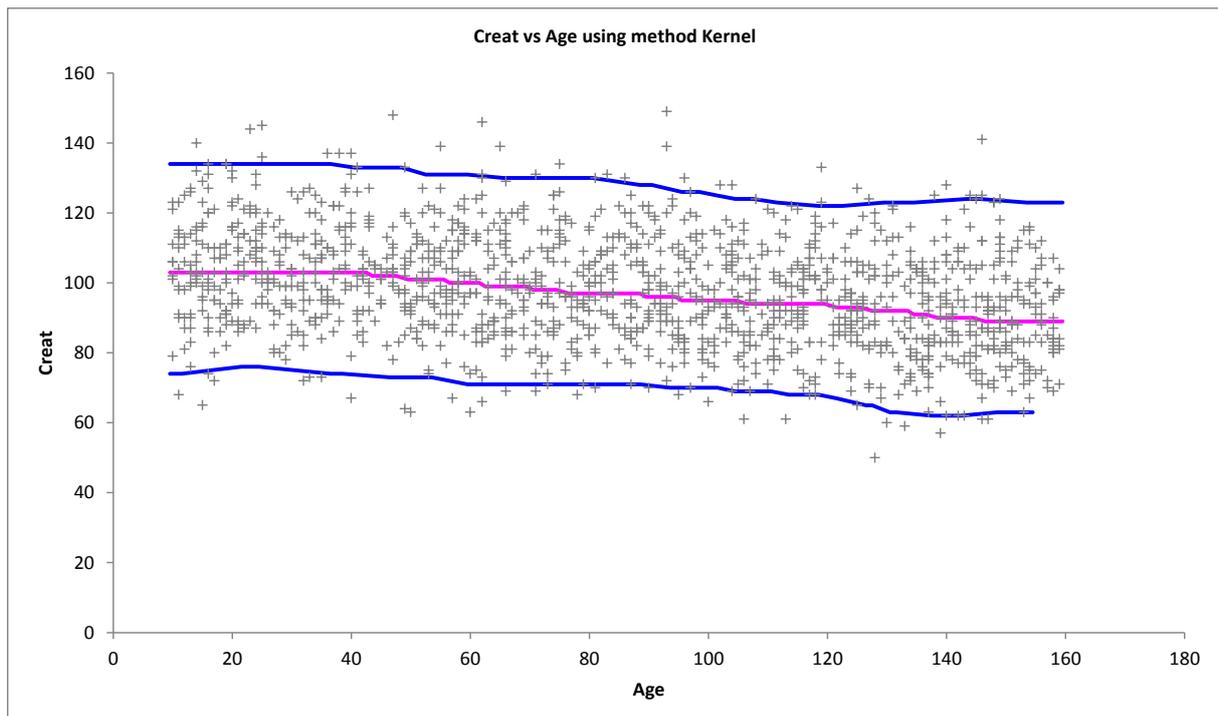
References

[1] CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline. Third ed. Wayne, PA: CLSI, 2008.

[2] Horn PS. A biweight prediction interval for random samples. J Am Stat Assoc 1988;83:249-256.

[3] John W. Tukey. "*Exploratory Data Analysis*". Addison-Wesley, Reading, MA. 1977.

[4] McWilliams TP. A Distribution-Free Test for Symmetry Based on a Runs Statistic. J Am Stat Assoc 1990;85:1130-1133.

[5] Davidian, M., and Giltinan, D.M. (1995), *Nonlinear Models for Repeated Measurement Data*, New York: Chapman and Hall.

[6] Gannoun, A., Girard, S., Guinot, C., Saracco, J. Reference curves based on nonparametric quantile regression. Stat. in Med., 2002; 21, 3119-3155.

[7] Gannoun, A., Girard, S., Guinot, C., Saracco, J. Sliced inverse regression in reference curves estimation. CSDA, 2004; 46(1), 103-122.