Worm plot to diagnose fit in quantile regression

Stef van Buuren[1,2]
[1] Department of Statistics, TNO Quality of Life, The Netherlands
[2] Department of Methodology & Statistics, Faculty of Social Sciences, University of Utrecht, The Netherlands

Abstract: The worm plot is a series of detrended Q-Q plots, split by covariate levels. The worm plot is a diagnostic tool for visualizing how well a statistical model fits the data, for finding locations at which the fit can be improved, and for comparing the fit of different models. This paper shows how the worm plot can be used in conjunction with quantile regression. No parametric distributional assumptions are needed to create the worm plot. We fitted both an LMS and a quantile regression model on Dutch height data. The worm plot shows that the quantile regression model is superior to the LMS model in terms of fit. At the same time, it also contains a warning that the particular quantile model used may actually overfit the data. The resulting quantile curves are wiggly at the extremes, and appear less well suited for drawing growth diagrams. The paper concludes that the worm plot is a natural diagnostic tool for quantile regression.

1 Introduction

Quantile regression (QR) (Koenker, 2005; Koenker and Bassett, 1978) provides a flexible way to describe how a given quantile $\tau(0 < \tau < 1)$ of an outcome y changes as a result of one or more covariates x. Various goodness-of-fit tests are available for the linear quantile regression model, for example, likelihood ratio and Langrange multiplier tests (Koenker and Machado, 1999). While such tests statistics are very useful overall aggregates of fit, they typically do not provide much insight at what points the largest deviations between the data and model occur. Without such information, it may be difficult to identify the parts of the model that need improvement.

The worm plot is a diagnostic tool that was developed to steer the model selection process in fitting age-conditional growth reference curves (van Buuren and Fredriks, 2001). The tool consists of a number of detrended Q-Q plots (Wilk and Gnanadesikan, 1968), split according to age. A model that fits the data well is characterized by 'flat worms'. The worm plot is sensitive to local deviations of the

---

Address for correspondence: Stef van Buuren, Department of Statistics, TNO Quality of Life, PO Box 2215, 2301 CE Leiden, The Netherlands. E-mail: Stef.vanBuuren@tno.nl

model. Figure 1 is an example of a worm plot for a badly fitting LMS model on the height of Dutch boys. Locations at which the worms are not flat suggest specific changes to the model (see Table 1). The worm plot was developed as an adjunct to the LMS model (Cole, 1988; Cole and Green, 1992), where it was used to guide the smoothness parameters in the L-, M- and S-curves. Scripts and examples can be found at the website of this journal and at www.stefvanbuuren.nl.
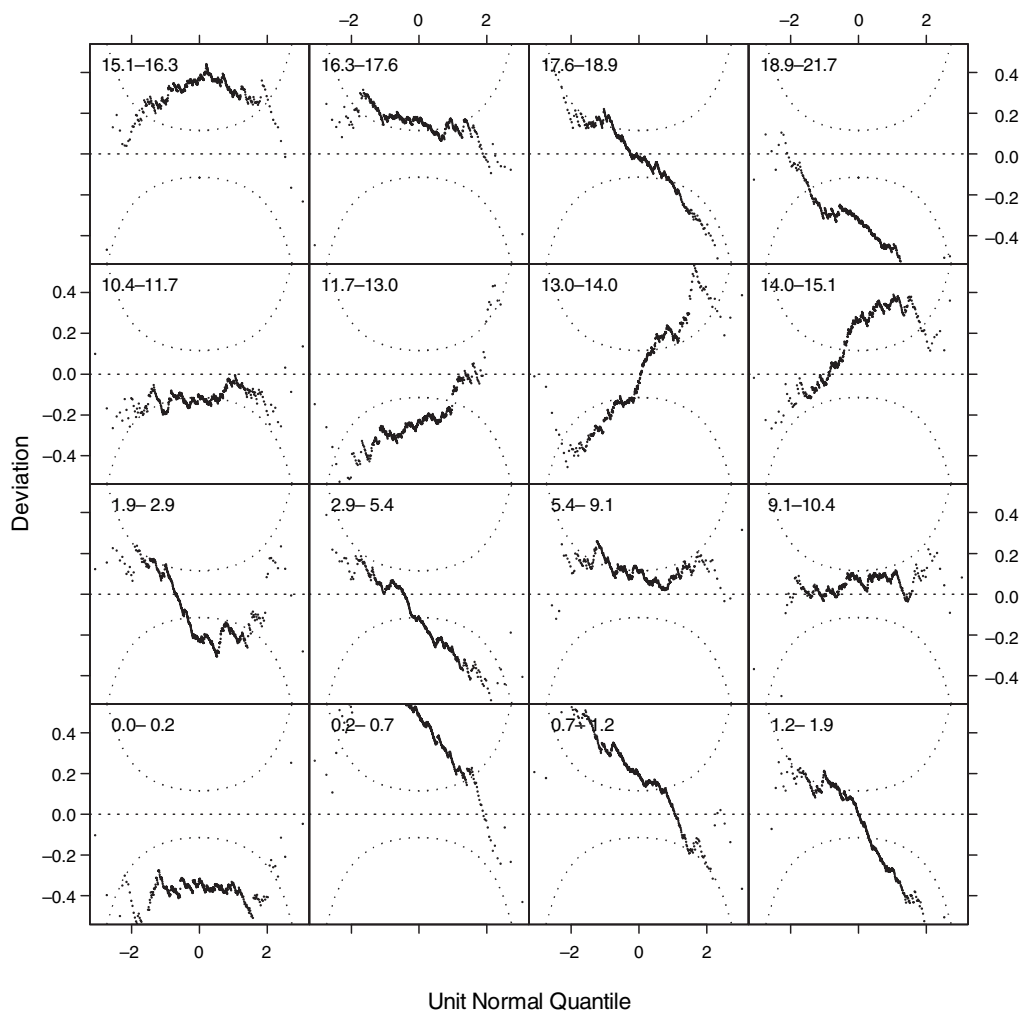

Figure 1   Worm plot of the *z*-scores for height of Dutch boys (LMS model 0051R). The plot consists of detrended Q-Q plots in 16 age groups of equal size, ordered from the lower-left panel to the upper-right panel. Model 0051R fits badly in almost all ages

Table 1 Interpretation of various patterns in the worm plot

Shape	Moment	If the	then the
intercept	mean	worm passes above the origin,	fitted mean is too small.
		worm passes below the origin,	fitted mean is too large.
slope	variance	worm has a positive slope,	fitted variance is too small.
		worm has a negative slope,	fitted variance is too large.
parabola	skewness	worm has a U-shape,	fitted distribution is too skew to the left.
		worm has an inverted U-shape,	fitted distribution is too skew to the right.
S-curve	kurtosis	worm has an S-shape on the left bent down,	tails of the fitted distribution are too light.
		worm has an S-shape on the left bent up,	tails of the fitted distribution are too heavy.

QR and LMS have similar objectives, but differ in methodology. Both approaches aim to fit smooth quantiles of *y* by functions that vary smoothly in *x*. The difference is that QR is nonparametric and free of distributional assumptions, whereas the LMS model assumes that *y* is distributed normally after an age-dependent Box-Cox transformation. QR is more general and has the potential of identifying features in the data that might go undetected by the LMS method.

The conventional Q-Q plot portrays how the empirical and theoretical quantiles differ. Application of the worm plot to QR is non-standard because QR lacks an underlying distribution from which the theoretical quantiles can be taken. This paper proposes a distribution-free version of the Q-Q plot that does not require a parametric distribution. The central idea is as follows: if the quantile model fits the data, then the distribution of the empirical quantiles must be uniform. The Q-Q plot provides a way to assess uniformity, and if it holds, then the quantile model provides an adequate description of the data.

The present paper will explore the use of the worm plot within the context of QR. The purpose of the worm plot is three-fold:

(a) to investigate how well the model fits the data;
(b) to identify locations at which the fit could be improved;
(c) to compare the fit of quantile models with that of parametric models.

The paper first describes the substantive problem of interest. Subsequently, both the LMS and the QR model are applied to the data. The worm plot provides a means to compare the fit of both models. The discussion concludes the paper with some thoughts about the relative merits of the worm plot, and its role within the context of QR.

2 Fourth Dutch Growth Study

Cross-sectional data on height, weight and head circumference of Dutch children between ages 0 and 21 years were collected within the Fourth Dutch Growth Study (Fredriks, van Buuren, Burgmeijer and Meulmeester *et al*., 2000; Fredriks,
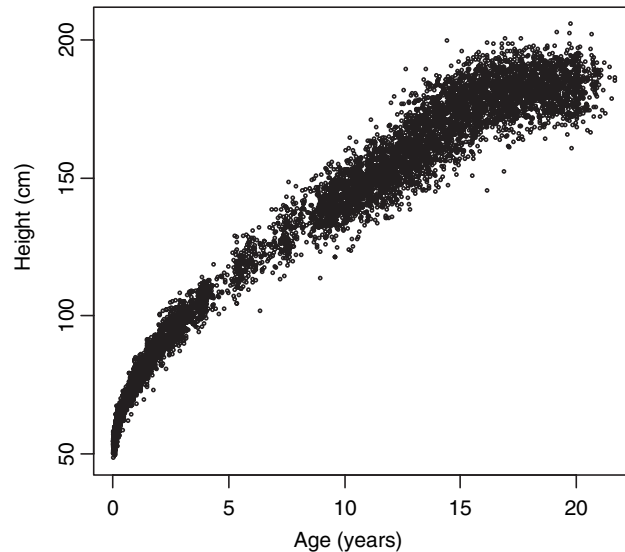
Figure 2 Length/height (cm) against age (years) for Dutch boys (Fredriks, van Buuren, Burgmeijer and Meulmeester, 2000)

van Buuren, Wit and Verloove-Vanhorick, 2000). This study is a follow-up to earlier nation-wide studies performed in 1955, 1965 and 1980. The design was stratified according to province, municipal size, sex and age in order to arrive at a sample that is representative of children of Dutch origin. The realized sample size of healthy Dutch boys without known growing problems was equal to 7482 boys. Proportionally more children at infancy were sampled so as to have more details in periods of rapid growth. The primary goal of the study was to update the Dutch 1980 growth references.

Figure 2 is a scatter plot of height versus age. The modeling task consists of describing this cloud of points by age-conditional quantiles/centiles or, alternatively, by age-conditional *z*-scores. The use of *z*-scores is popular in growth because that scale is capable of displaying differences in the extremes of the distribution, and because *z* is linear in *y* under normality.

For clinical applications, the extremes of the growth chart are the most relevant area. Screening procedures may set criteria for referral as low as -3 SD (van Buuren *et al.*, 2004).

3   LMS model

Suppose a measurement y_i and one or more covariates x_i are available for a sample of n persons. A classic assumption in human growth is that y is normally distributed given x. Let $\mu(x)$ and $\sigma(x)$ respectively denote the mean and standard deviation of y at different levels of x. The quantile function of y for $0 < \tau < 1$ can then be written

as $Q(\tau|x) = \mu(x) + \sigma(x)\Phi^{-1}(\tau)$, where Φ^{-1} is the inverse of the standard normal distribution function. If normality holds, then a proportion τ of the sample will lie below the curve $Q(\tau|x)$, whereas a proportion of $1 - \tau$ is located above the curve.

The simple normal quantile model may fail to fit the data for two reasons. First, if $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ are biased estimates of $\mu(x)$ and $\sigma(x)$ at some x, then $Q(\tau|x)$ may not split the sample at proportions τ and $1-\tau$ at this x. Bias in $\mu(x)$ and $\sigma(x)$ could easily arise because one would generally want these functions to be smooth in x. Second, the normal distribution may be a poor approximation of the empirical data at some x, which could also result in a failure to split the sample into proportions τ and $1 - \tau$. This can occur if y is skewed, for example, for body weight or body mass index.

The development of the LMS model was motivated by the need to model skewed distributions. The LMS model writes the quantile function as

$$Q(\tau|x) = \mu(x)(1 + \lambda(x)\sigma(x)\Phi^{-1}(\tau))^{1/\lambda(x)}, \tag{3.1}$$

where $\mu(x), \sigma(x)$ and $\lambda(x)$ can be interpreted as curves of the median, coefficient of variation and Box-Cox power parameter, respectively, each varying smoothly in x. The measurements can be expressed as

$$z_i = \frac{(y_i/\mu(x_i))^{\lambda(x_i)} - 1}{\lambda(x_i)\sigma(x_i)}, i = 1, \ldots, n. \tag{3.2}$$

For a well-fitting model, we expect $z_i \sim N(0, 1)$ for all x in the reference sample, that is, the sample from which $\mu(x), \sigma(x)$ and $\lambda(x)$ are estimated. In practice, the fit can be reduced due to excessive mass (clumping) of values in some region of the distribution, for example, by rounding to whole centimeters. As the transformation approach assumes smoothness and continuity of density, it may fail to pick up such violations in empirical data.

The worm plot is a diagnostic check whether the normality holds for a given LMS solution. More in particular, the primary modeling task in the LMS model is to specify the degrees of freedom E_μ, E_σ and E_λ for the fitted curves $\hat{\mu}(x), \hat{\sigma}(x)$ and $\hat{\lambda}(x)$. Choosing only a few degrees of freedom provides very smooth reference curves, but these curves may fail to faithfully represent the true pattern across x. On the other hand, specifying a high value puts great fidelity on the data at the expense of smoothness. The worm plot is a sensitive tool that enables the user to decide on the number of degrees of freedom. Basically, the user increases the degrees of freedom E_μ, E_σ or E_λ, and elects the solution above which the worm plot does not further improve.

Figure 3 shows the worm plot resulting from the final LMS model applied to the height data of the Dutch boys, and where x was defined as age of the child. The solution sets $E_\mu = 10, E_\sigma = 6$ and $E_\lambda = 0$. Increasing E_μ, E_σ or E_λ did not appreciably improve the worm plot. All worms are relatively flat, and the worm plot indicates that the LMS model fits the data quite well.
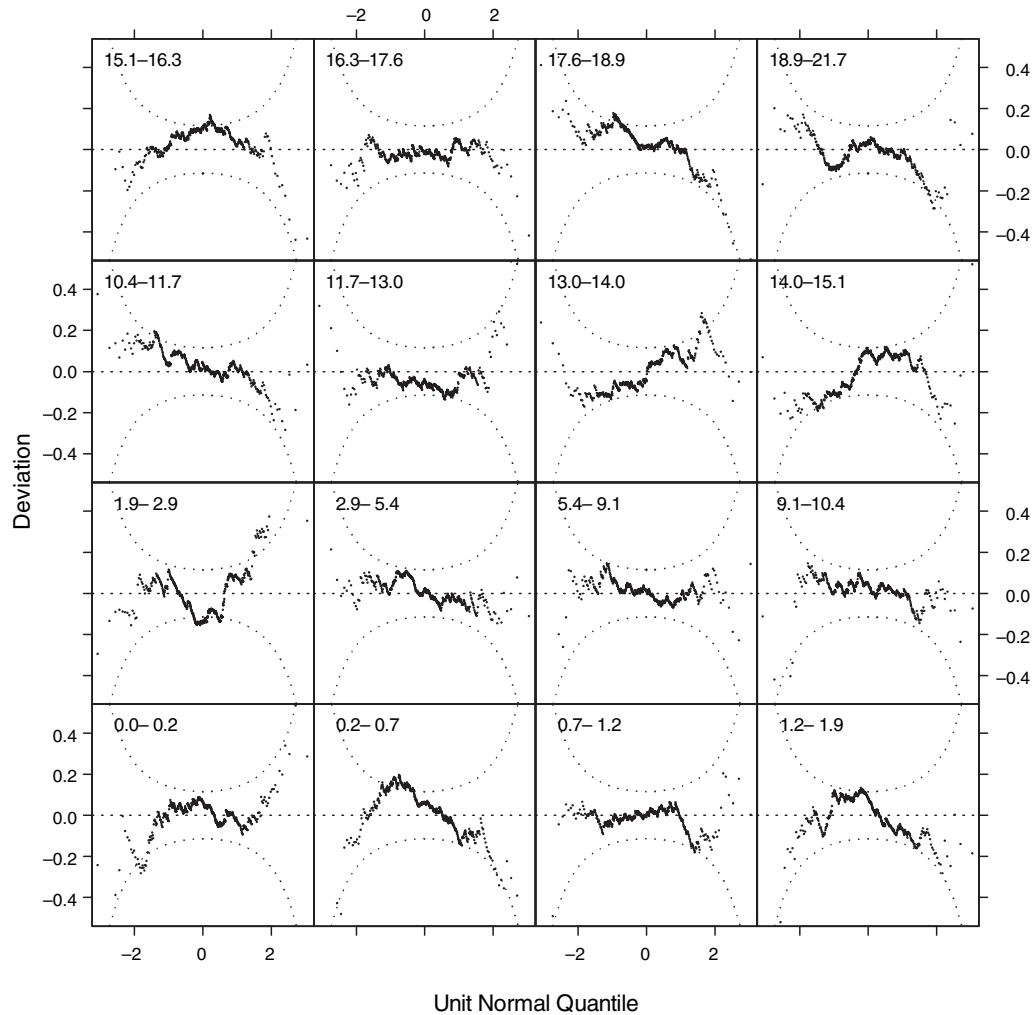
Figure 3   Worm plot for length/height of Dutch boys for the final LMS model 0106R. The model fits well and was used to draw the Dutch reference diagram

The LMS model is restricted to modeling variation in location, shape and skewness. This means that the model may potentially miss features in the data, for example, kurtosis, that cannot be modeled by the Box-Cox power transformation. One approach is to extend the parametric family to include other shapes. This route was followed by Rigby and Stasinopolous, whose implementation of the Generalized Additive Models for Location, Scale and Shape (GAMLSS) method is a flexible and general generalization of LMS (Rigby and Stasinopolous, 2005).

4 Quantile regression

Quantile regression is an alternative approach that abandons any parametric assumptions. The QR model can be formulated as follows: suppose that we minimize the function

$$\sigma(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - \beta) \tag{4.1}$$

over the scalar β, where $\rho_\tau(u) = (\tau - I(u < 0))$ is the 'check function'. The optimal solution $\hat{\beta}$ is equal to the τth sample quantile of $y = \{y_1, \ldots, y_n\}$, so $\hat{\beta} = F_y^{-1}(\tau)$.

Koenker and Bassett (1978) extended this optimization interpretation of ordinary sample quantiles to linear parametric models for conditional quantile functions. QR replaces β by a linear function of the covariates x as

$$\sigma(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta), \tag{4.2}$$

where β is a vector of p elements. The estimate $x_i'\hat{\beta}$ can be interpreted as the τth sample quantile conditional function of y given x.

5 Worm plot for quantile regression

In section 3, we distinguished between two possible reasons for lack-of-fit in the LMS model: bias in the estimates of $\mu(x), \sigma(x)$ and $\lambda(x)$ due to smoothness requirements, and misspecification of the family of conditional distribution. QR is more general than LMS in the sense that it does not require specification of the (age-) conditional distribution. Thus, misspecification of the distribution does not present a problem in QR. Bias can still occur, though, if the component $x_i'\beta$ misspecified. The worm plot provides a way to diagnose this problem.

The rationale of the worm plot applied to QR is as follows. In the simple univariate model (4.1), the best estimate $\hat{\beta}$ is equal to the τth sample quantile of y. Plotting $F_y(\hat{\beta})$ against τ (that is, the *probability-probability plot*) or plotting $\hat{\beta}$ against $F_y^{-1}(\tau)$ (that is, the *quantile-quantile plot*) reveals that the points will be on the line of identity. For the more general model (4.2) these plots are not informative, but instead we may plot $F_y(x_i'\hat{\beta})$ against $\tau|x_i$, or alternatively, $x_i'\hat{\beta}$ against $F_{y|x}^{-1}(\tau|x_i)$, the theoretical quantile at covariate level x_i. As above, one would expect that all points $(x_i'\hat{\beta}, F_{y|x}^{-1}(\tau|x_i))$ are located on the line of identity. Or, put in other words, if the model fits the data we expect that $F_y(x_i'\hat{\beta})$ has a uniform distribution at each level of x_i. In practice this may

not be true for some x_i as $\hat{\beta}$ is a global estimate that need not be optimal at each local x_i.

Figure 4 shows an example where this occurs. Figure 4(a) portrays a highly nonlinear relation between x and y, inappropriately modeled by a linear QR for $\tau = 0.50$. The regression line divides the sample globally into two halves, but there is gross lack-of-fit in the middle and the extreme regions of x. Figure 4(b) is a detrended
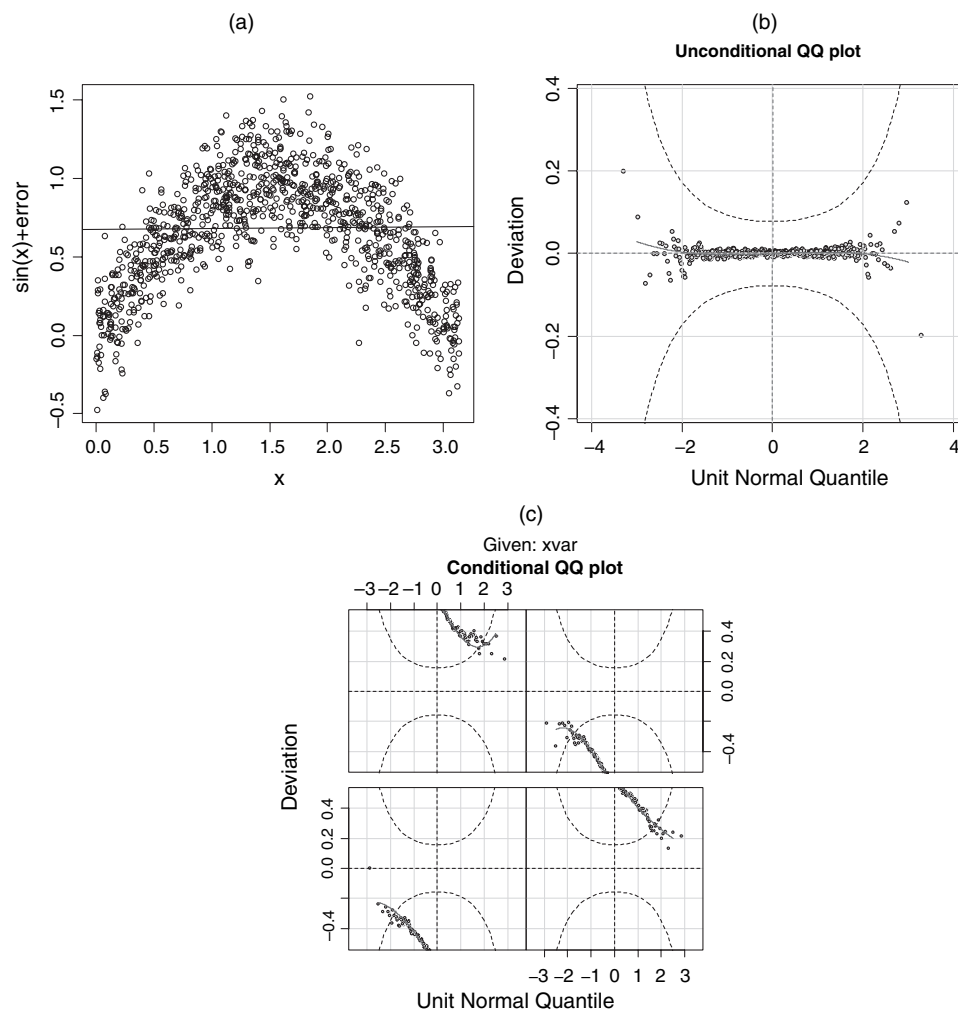

Figure 4   Demonstration that Q-Q plot is not informative for conditional models. Figure(a) is a nonlinear relation modeled by an inappropriate linear quantile model ($\tau = 0.5$). The regression line divides the sample into two halves, but there is clear misfit in the middle and the extreme regions of x. Figure(b), the detrended unconditional Q-Q plot, does not indicate any misfit. Figure(c), the conditional Q-Q plot (or worm plot) is sensitive to this type of misfit

Q-Q plot, which wrongly conveys the message that everything is OK. The worm plot (Figure 4(c)) conditions on x, and tells a very different story. The worm plot easily identifies lack-of-fit, and is sensitive to model deviations at the level of the covariates x. Where the conventional Q-Q plot diagnoses fit to the constant model (4.1), the worm plot is the natural companion of the QR model (4.2). Note that the strong nonlinearity in Figure 4 will also show up in a conventional diagnostic graph of the residuals against age. Plotting such diagnostic graphs is useful to detect deviations in the mean structure of the model and the data. However, such plots are generally less informative than the worm plot about any lack-of-fit in the variance, skewness and kurtosis of the residuals.

It is straightforward to calculate the points needed to construct the worm plot. First, k regression quantile models are fitted to x_i using various τ. We used the grid $\tau \in \{0.001, 0.002, \ldots, 0.01, 0.02, \ldots, 0.98, 0.99, 0.991, \ldots, 0.999\}$ with $k = 117$ elements. Inclusion of extreme τ centiles allows calculation of $F_y(x_i'\hat{\beta})$ for values in the tails. Note that this requires a relatively large sample size. For smaller samples, one could truncate the extremes to, say, 0.01–0.99 at the expense of loss of discrimination beyond the truncation point. As a second step, the corresponding reference table is calculated, which stores the distribution of y at selected x-values in terms of the k quantiles. The quantiles at these x-values can be easily calculated by prediction. The choice of x-values in the reference table will normally depend on the problem at hand, and the spacing should be fine enough to allow for essentially unbiased linear interpolation between the x-values. Third, for each data point (y_i, x_i) the empirical probability $F_y(x_i'\hat{\beta})$ is calculated from the reference table by double linear interpolation (first over x, then over k). Fourth, both scales in the Q-Q plot are transformed, without loss of generality, towards the z-score scale. The worm plot can then be drawn using the points $(\Phi^{-1}(\tau), \Phi^{-1}(F_y(x_i'\hat{\beta})) - \Phi^{-1}(\tau))$ conditional on x. The minimum and maximum values on the horizontal axis are equal to $\Phi^{-1}(0.001) = -3.09$ and $\Phi^{-1}(0.999) = 3.09$. Note that this step may be bypassed to produce a P-P plot, which retains the quantile scale. The P-P plot is generally more sensitive to differences in the center of the distribution, and less suited to detect differences in the extremes (Wilk and Gnanadesikan, 1968).

6   Application

Wei *et al.* suggest applying (4.2) to calculate references curves of height (Wei *et al.*, 2006). These authors propose modeling the nonlinear relation between age (x) and height (y) by replacing x by a cubic B-splines basis, where the interior knots are located at $\{0.2, 0.5, 1.0, 1.5, 2.0, 5.0, 8.0, 10.0, 11.5, 13.0, 14.5, 16.0\}$. Spacing of the knots was chosen for more flexibility during infancy and in the pubertal growth spurt period. Application of B-splines reduces the nonlinear relation to a linear problem. Calculation of the solution is straightforward.

Figure 5 is the worm plot of the QR model. The worm plot indicates an extremely good fit between the model and the data. All worms are very flat, perhaps with the exception of the age range 1.9 to 2.9 years, where the upper empirical quantiles are above the theoretical quantiles. It is possible to improve upon this aspect of the solution by including an extra knot at the age of 3 years. Compared to Figure 3, the fit to the data of the B-spline model is clearly superior to the LMS method. The QR model thus provides very accurate description of the quantiles in the reference sample.
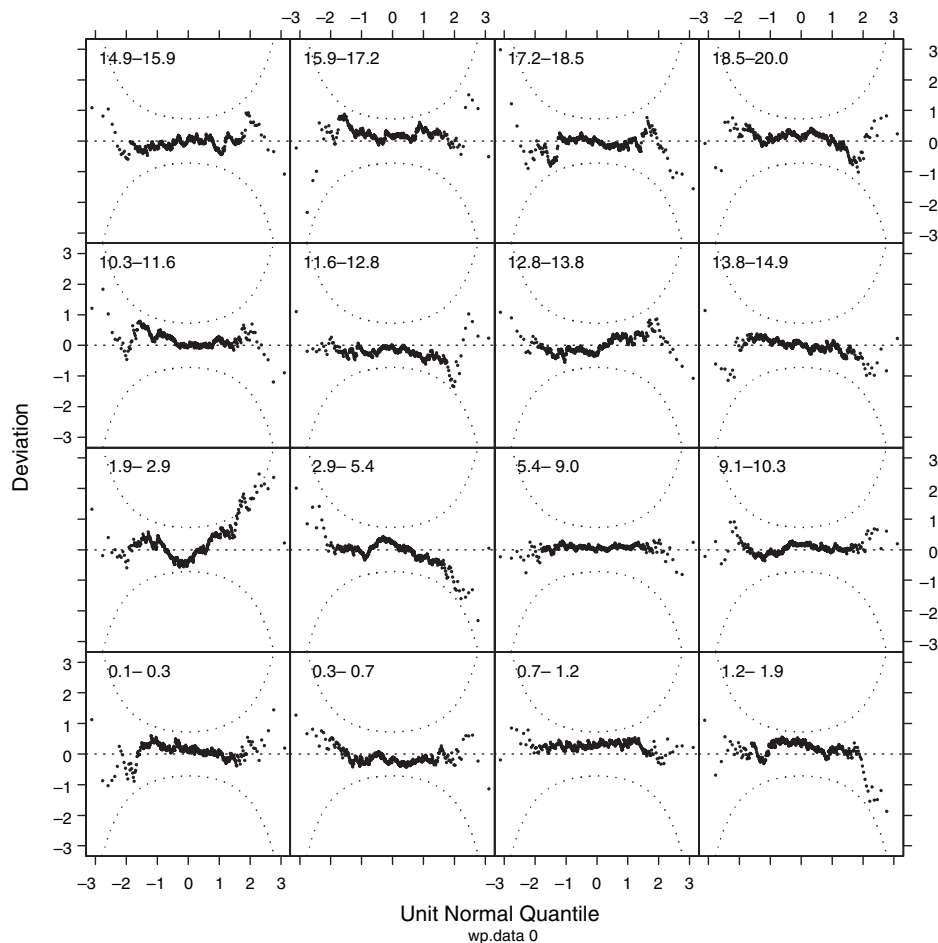

Figure 5   Worm plot of the quantile regression model of Wei *et al.* (2006) that codes age into a B-splines bases with 15 internal knots. The fit of this model to the data is extremely good
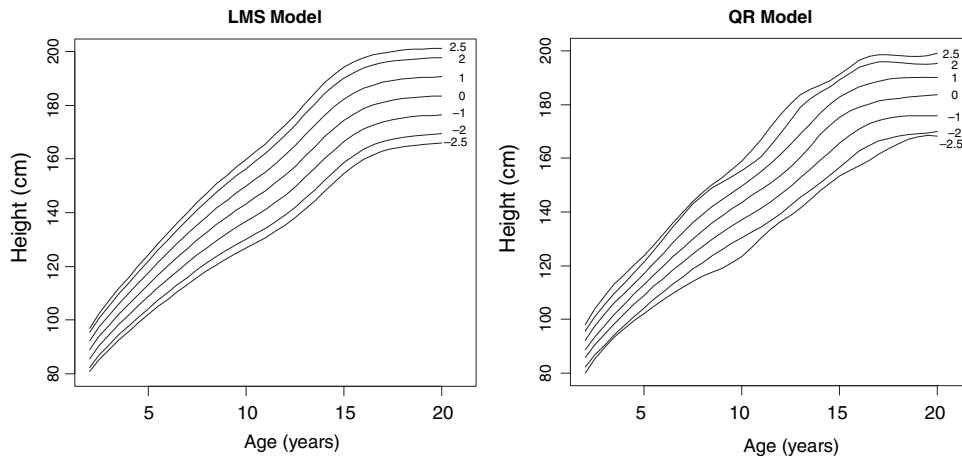
''buuren'' --- 2007/12/3 --- page 372 --- #10

Figure 6 References diagrams under the LMS model of Cole and Green (1992) and the quantile regression model of Wei *et al.* (2006)

The other side of the coin is that the resulting reference curves of the QR model look quite wiggly. Figure 6 is a side-by-side plot of the reference curves under the LMS and QR models for the standard quantiles published on the Dutch reference curves. The discrepancies between both models grow for the more extreme reference curves, even with a large sample size ($n = 7482$). In the QR model, the smoothness of the reference curves is related to τ, with the smoothest curve occurring at $\tau = 0.5$.

The quantile model can be molded arbitrarily close to the data by taking a more liberal B-spline basis. Though this property is attractive for data description, the resulting model may actually be too good to be true. More specifically, if a new sample of the same size was collected, it is quite likely that the reference curves will look different. A sign that this occurs here can be found in the worm plot. Less than 1% of the points fall outside the 95% confidence intervals, where we expect 5% under the true model. It thus appears that this particular QR model is partly fitting random noise due to sampling in the data. Observe that this conclusion pertains to the data at hand and to the model with interior knots as proposed by Wei *et al.* (2006). It is well possible that fitting more restrictive models to these data could provide a solution that is less dependent on sampling fluctuations.

7 Conclusion

The worm plot is a natural companion to QR. It allows the user to investigate how well the model fits the data, to identify locations at which the fit could be improved, and to compare the fit of quantile models with that of parametric models. The worm plot is easy to calculate, and requires no more than routines for QR, linear

interpolation and plotting. No parametric distributional assumptions are needed to create the worm plot.

We demonstrated the worm plot to compare two fitting models for reference curves, the LMS and the QR method. Both are similar approaches to construct reference curves. The main difference is that the LMS assumes, for each level of x, that y follows a distribution. The worm plot showed that the QR model gives a much better fit to the Dutch height data than the LMS model. The worm plot also suggests, however, that the particular QR model as fitted here is perhaps 'too good to be true'. Note that overfitting can also occur with the LMS model (for example, by substantially increasing the degrees of freedom), and would manifest itself in a similar way through the worm plot. In either case, a more restrictive model is needed to avoid sampling fluctuations.

Several authors have suggested the QR model as a method to fit reference curves (Gannoun *et al.*, 2002; He, 1997; Heagerty and Pepe, 1999; Kapitula and Bedrick, 2005; Wei *et al.*, 2006). The QR model does not assume a distribution, and is more general than the LMS model. QR is rapidly entering mainstream statistics, and the theoretical properties are well studied. QR is easily applied to multivariate predictors, and constraints can be applied to the coefficients. The number of implementations in statistical software is growing.

However, using QR for reference curves is not without drawbacks. The QR model lacks an explicit formula that allows one to convert a measurement into quantile or z-score. The availability of such a formula was one of the requirements set forth by a World Health Organization expert committee (Borghi *et al.*, 2006). The QR model also does not allow for estimating a quantile or z-score beyond the most extreme tabled quantiles. Such extrapolations beyond the data are desirable for tracking children with extreme growth. The reference curves produced by the quantile model are irregular near the extremes, and are generally less aesthetically pleasing. We may prevent wiggly curves by including an extra smoothing step over the joint quantiles, as in the model of Healy, Rasbash and Yang (1988). An alternative is to vary the amount of smoothing depending on τ. This takes out the irregularity of the extreme quantiles, but the method is still capable of detecting gross changes in the conditional quantile function. Some have pointed that QR has the potential of generating quantile curves that cross, which creates a problem in the interpretation. Solutions for this have been developed (He, 1997), but no such solutions were needed here. In the present application, no problems showed up with quantiles that crossed.

Although the QR model gives a much better fit to the data, the LMS model seems preferable for the purpose of constructing growth diagrams. Use of the LMS model implies an assumption that the natural phenomena and the observation process should provide smooth distributions that extrapolate beyond the data. Under this assumption, the LMS model is a parsimonious description of the real world, provides stable estimates of extreme conditional quantiles, and comes with a simple formula for calculating z-scores. The case may be different for outcomes other than height (for which normality is well established), though the LMS model has been proven

versatile for many outcomes. The worm plot will help in revealing the outcomes for which the LMS model does not fit, and where a more general model like GAMLSS or QR is called for.

The QR model could have a great future as a diagnostic tool for checking parametric models. Quantile functions may guide model choices in parametric models for location, shape, skewness and kurtosis. Plotting quantile functions on top of parametrically-fitted functions will suggest locations where the parametric model cuts too many corners, and where a more general parametric model is needed. The real strength of QR is that it reveals points at which distributional assumptions go astray. The worm plot and regression quantiles are different, but intimately related, tools that will point in the same direction. Both will find stable features in that data which may otherwise go undetected. Regression quantiles are useful to visualize deviations between (smoothed) empirical and modeled probabilities, whereas the worm plot suggests where adjustments in the parametric model can be made. The combined use of QR and the worm plot will open up a way for a new set of tools for model exploration.

Acknowledgement

References

Borghi E, de Onis M, Garza C, Van den Broeck J, *et al*. (2006) Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. *Statistics in Medicine*, **25**, 247–65.

Cole TJ (1988) Fitting smoothed centile curves to reference data (with discussion). *Journal of the Royal Statistical Society, Series A*, **151**, 385–418.

Cole TJ and Green PJ (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305–319.

Fredriks AM, van Buuren S, Burgmeijer RJ, Meulmeester JF, *et al*. (2000) Continuing positive secular growth change in The Netherlands 1955–1997. *Pediatric Research*, **47**, 316–23.

Fredriks AM, van Buuren S, Wit JM and Verloove-Vanhorick SP (2000) Body index measurements in 1996–7 compared with 1980. *Archives of Diseases in Childhood*, **82**, 107–12.

Gannoun A, Girard S, Cuinot C and Saracco J (2002) Reference curves based on non-parametric quantile regression. *Statistics in Medicine*, **21**, 3119–135.

He X (1997) Quantile curves without crossing. *The American Statistician*, **51**, 186–92.

Heagerty PJ and Pepe MS (1999) Semiparametric estimation of regression quantiles with applications. *Journal of the Royal Statistical Society, Series C*, **48**, 533–51.

Healy MJR, Rasbash J and Yang M (1988) Distribution-free estimation of age-related centiles. *Annals of Human Biology*, **15**, 17–22.

Kapitula LR and Bedrick EJ (2005) Diagnostics for the exponential normal growth curve model. *Statistics in Medicine*, **24**, 95–108.

Koenker R (2005) *Quantile regression*. Cambridge: Cambridge University Press.

Koenker R and Bassett G (1978) Regression quantiles. *Econometrica*, **46**, 33–50.

Koenker R and Machado JAF (1999) Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**, 1296–310.

Rigby RA and Stasinopolous DM (2005) Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society, Series C*, **54**, 507–54.

van Buuren S and Fredriks AM (2001) Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259–277.

van Buuren S, van Dommelen P, Zandwijken GR, Grote FK, *et al*. (2004) Towards evidence based referral criteria for growth monitoring. *Archives of Diseases in Childhood*, **89**, 336–41.

Wei Y, Pere A, Koenker R and He X (2006) Quantile regression methods for reference growth charts. *Statistics in Medicine*, **25**, 1369–382.

Wilk MB and Gnanadesikan R (1968) Probability plotting methods for the analysis of data. *Biometrika*, **55**, 1–17.

"buuren" --- 2007/12/3 --- page 376 --- #14