An empirical variance distribution approximation, using backtransformation from the percentile scale, for the quantile estimating function of continuous distributions

John P. D. Martin

Wednesday, October 7, 2015. fixed eqn 21 Nov 10 2018

Executive Summary

This paper demonstrates a quadratic polynomial proxy, to calculate quantile variance estimates of continuous distributions in the percentile scale, for the quantile estimating function (Koencker & Bassett (1)). The empirical variance distribution function created by the backtransformation of the percentile scale calculations to the original measurement scale for the proxy estimator, has a step function morphology, in common with the empirical cumulative distribution function.

Analysis of the accuracy of the results makes use of known results for consistent variance estimates of medians of unweighted samples, Martin (2,3) and quantile regression bootstrap results for "intercept only model". The derivation of closed form variance expressions in the percentile scale, is consistent with the expectation that suitably transforming the reference frame can improve estimation accuracy Miller (4).

Table 1 displays some quantile variance estimates for unweighted samples conducted in the percentile scale, and then backtransformed to the original scale for several common distributions; uniform, normal, log-normal and skewed bivariate normal. The empirical quantile variance estimates are compared to bootstrap quantile confidence intervals (from the quantreg r package (5)).

It can be seen that for modest samples, the empirical quantile variance estimates are generally asymmetric about the quantile estimate for the continuous distributions considered. However, the average standard error of the empirical variance distribtion is close to the calculated bootstrap values.

Table 1: Quantile regression estimates of intercept only model, bootstrap standard errors and proxy standard errors of unweighted samples

Distribution	quantile	est value	bootstrap std error	average proxy std error	asymm proxy std errors
random uniform(0,1)	0.25	0.2432	± 0.0126	± 0.0138	(-0.0171,0.0104)
""	0.5	0.4895	± 0.0123	± 0.0132	(-0.0116,0.0149)
""	0.75	0.7242	± 0.0196	± 0.0192	(-0.0103,0.0282)
random normal(0,1)	0.25	-0.6831	± 0.0459	± 0.0451	(-0.0530,0.0372)
""	0.5	-0.0300	± 0.0410	± 0.0433	(-0.0325,0.0541)
""	0.75	0.6237	± 0.0382	± 0.0367	(-0.0368,0.0366)
systematic uniform(0,1)	0.25	0.2510	± 0.0208	± 0.0217	(-0.0217,0.0217)
""	0.5	0.5013	± 0.0251	± 0.0241	(-0.0241,0.0241)
""	0.75	0.7516	± 0.0214	± 0.0217	(-0.0217,0.0217)
log-normal N(0.1,1)	0.25	0.5868	± 0.0249	± 0.0285	(-0.0315,0.0254)
""	0.5	1.0872	± 0.0404	± 0.0413	(-0.0381,0.0445)
""	0.75	2.1934	± 0.0779	± 0.0707	(-0.0734,0.0681)
$\frac{1}{3}$N(55,5.5)+$\frac{2}{3}$N(80,6)	0.25	59.78	± 1.5041	± 1.3028	(-1.6429,0.9628)
""	0.5	76.84	± 0.8308	± 0.7378	(-0.9610,0.5146)
""	0.75	81.67	± 0.4497	± 0.4447	(-0.5648,0.3245)

To highlight the benefit of the linearity in the unweighted sample distribution (2,3) to variance calculations, afforded by the transformation to the percentile scale. The third example in Table 1, is the highly artifical case of evenly spaced observations. In this rare linear example in the original scale, as expected, the quadratic polynomial proxy can closely match the bootstrapped linear programming variance estimate.

Importantly, the use of the quadratic polynomial proxy approach (i) employs the quantile estimating function to perform the backtransformation from the percentile scale, (ii) only the sample size and chosen quantile value are needed for calculations in the percentile scale (ie. no sorting required) and (iii) a density function has been derived which allows small samples estimates to be calculated self-consistently (without CLT assumptions).

The use of the empirical variance distribution approach to calculating the variance of the quantile regression residuals becomes a new method for calculating confidence intervals for fitted quantile regression coefficients.

Introduction

Quantiles (6) are an order statistic of a distribution defined by the equivalent probability amount contained under the cumulative distribution function up to the (ordered) value of the quantile point.

That is, x is a k-th q-quantile for a variable X if

$\Pr[X < x] \leq k/q$ or, equivalently, $\Pr[X \geq x] \geq (1 - k/q)$

So the 25th percentile point is the 25/100 (k/q) 100-quantile point where 25% of the probability under the cumulative density function has occurred.

An equivalent calculation of quantile points has been demonstrated (1) using least absolute deviation (LAD) regression of the following quantile estimation function

$$\min_{b \,\epsilon \mathbb{R}} \{\theta \,|x_t - b| + (1 - \theta) \,|x_t - b|\} \tag{1}$$

where $\theta \equiv k/q$ and x_t are the sample/population elements of X. As the absolute value functions in equation 1, create a piecewise linear function shape (convex polytope) to the estimating function, linear programming techniques are required to solve the minimsation problem. As such, closed form expressions for the standard error of the quantile estimates are not available from this approach.

In this paper, it is demonstrated by transformation to the percentile scale of a cumulative density function for unweighted samples (i) the above quantile estimation function can be closely approximated by a (smoothing) quadratic polynomial proxy with analytic coefficients (in this reference frame), (ii) by comparing the standard deviation of the derived density function of the quadratic polynomial proxy to existing consistent variance estimates of the median, the closed form CLT standard error estimates of the quadratic polynomial proxy appears to be fully calibrated and (iii) approximate standard error estimates for equation (1) may be obtained by (numerical) backtransformation of the quadratic polynomial proxy results using quantile regression (with an intercept only model) to the original measurement scale.

Quantile estimation function smoothing

One approach for estimated standard errors of the quantile estimation function solution is to concurrently calculate the standard errors of smoothed versions of the problem, Brown & Wang (7). Figure 1 demonstrates the potential for smoothed estimators to approximate the quantile estimation function (using a simple moving average smoothing example).

The black lines show the quantile estimating function results, the red lines show simple moving average smoothing and the blue markers indicate the sample values (from a N(0,1) distribution). Typically however

the smoothed estimator itself is a different solution for each quantile point and dataset, so only numerical solutions rather than analytic results are possible.

As part of understanding the properties of the quantile estimation function for different continuous distributions and sample sizes, it is valuable to use a quasi-normalised version of the function where the minimum of the function always lies in the interval (0,1]. Equation (2) below specifies such a quasi-normalised quantile estimation function

$$\min_{b \in \mathbb{R}} \left(\frac{\theta \left| x_t - b \right| + (1-\theta) \left| x_t - b \right|}{\theta(1-\theta)(\frac{n(|\max(X)|+|\min(X)|)}{2})} \right) \tag{2}$$

$$\min_{b \in \mathbb{R}} \left(\frac{\frac{|x_t-b|}{(1-\theta)} + \frac{|x_t-b|}{\theta}}{\frac{n(|\max(X)|+|\min(X)|)}{2}} \right) \tag{3}$$

and the second pair of graphs in Figure 1 show (i) the conservation of the estimating function shape with quasi-normalisation and (ii) the consistent interval (0,1] in which the minimum occurs.

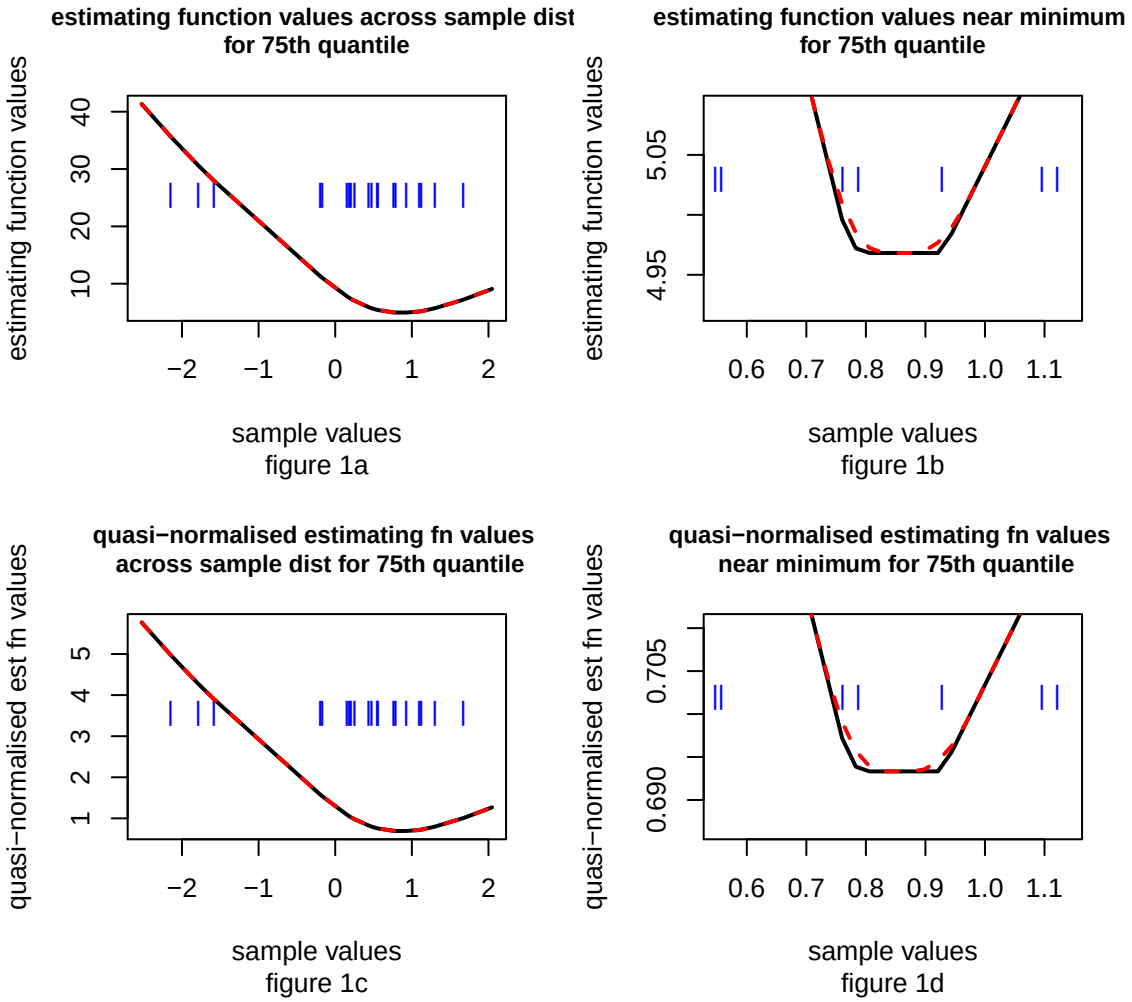


Figure 1; The quantile estimation function and a quasi-normalised version

Importantly, the denominator factor $\frac{n(|\max(X)|+|\min(X)|)}{2}$ is itself just a quasi simple arithmetic mean estimate of X. Its purpose (used in combination with the introduced $\frac{1}{\theta(1-\theta)}$ factor) is to ensure the quasi-normalised

quantile estimation function minimum (of unweighted samples) lies in the interval (0,1] regardless of the X distribution, other denominator choices such as Σx_t or $\Sigma |x_t|$ do not guarantee this behaviour.

Interestingly, when applying the quasi-normalised estimation function to quantile regression calculations, the function minimum may exceed 1 due to the lack of perfect correlation between the modelled values b & X.

Using the percentile scale for quantile estimation function

For the percentile scale for unweighted samples, the median (and q-quantile) are points on a linear scale. Each ordered point in this distribution has the following cdf values, using a basic median definition

1/n , 2/n , 3/n , ... , n/n

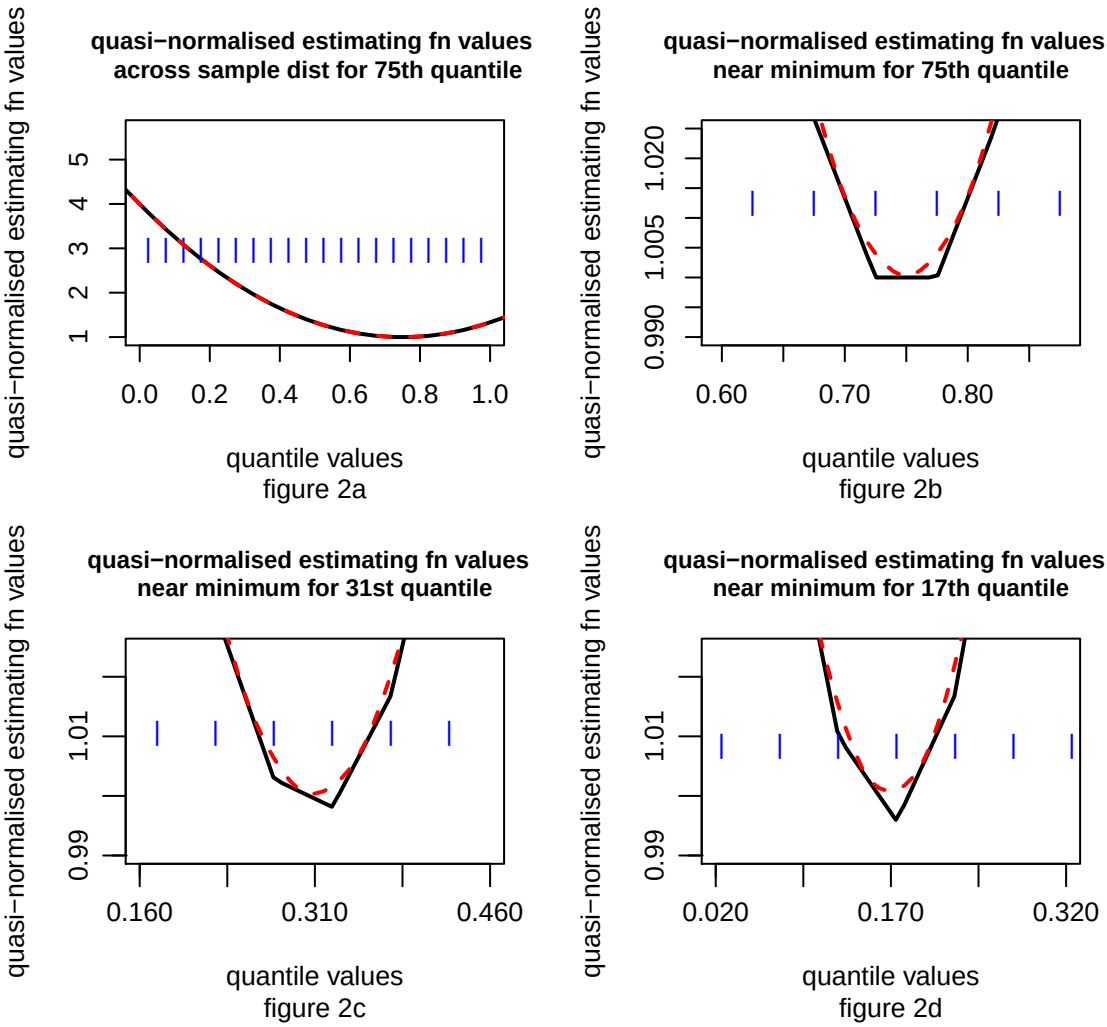
It is also the case that other median definitions are used, a particularly useful definition for use with the quantile estimation function, in the percentile scale is

1/2n, 3/2n, 5/2n, ... , (2n-1)/2n

which describes the empirical cumulative distribution function with interpolation $\frac{(h_i+h_i-1)}{2}$ and the sample endpoints are assigned half percentile weights as a continuity correction.

Figure 2, shows the quasi-normalised quantile estimation function which is contained in the interval [0,1], the empirical cumulative distribution function with interpolation, for an unweighted sample of 20 data points with three target percentiles 75th, 31st & 17th.

The black lines show the quasi-normalised estimating function values for a given choice of θ and the red lines indicate the behaviour of smoothing using a simple moving average. The blue markers are the (equally spaced) quantile values of the data points using the empirical cumulative distribution function with interpolation definition, in the percentile scale. Importantly, the equal spacing of the data points in this reference frame leads to equivalent smoothing function behaviour near the quantile estimating function regardless of the choice of θ.



As can be seen, by using the quantile definition where the empirical cumulative distribution function with interpolation (ecdf_int) is assigned as the percentiles of the data points

$$\min_{b\,\epsilon(0,1)} \left(\frac{\frac{|k/q-b|}{(1-\theta)} + \frac{|k/q-b|}{\theta}}{\frac{n(|\max(k/q)|+|\min(k/q)|)}{2}} \right) \in (1-1/2n, 1] \tag{4}$$

$$b_{min} \in [\theta - 1/2n, \theta + 1/2n] \tag{5}$$

as respectively, (i) each segment of the piecewise linear quantile estimation function shape has (percentile) width 1/n in the percentile scale for unweighted samples and (ii) the equally spaced sequence of data points in the percentile scale makes the denominator factor a true arithmetic sum which causes the estimation function minimum to be ~ 1.

Observing from figure 2 and all other attempted θ choices in the percentile scale interval (0,1), for the smoothed quantile estimation function with ecdf_int, using a simple moving average it is the case that

$$\min_{b\,\epsilon(0,1)} (smoothed\, quasinormalised\, function) \equiv 1 \tag{6}$$

$$b_{min}(smoothed\, quasinormalised\, function) = \theta \tag{7}$$

Curve fitting the coefficients of the smoothed quasi-normalised estimation function as a quadratic polynomial for several quantile points, using ecdf_int definition for the data points in the (percentile scale) interval [0,1], gives the following results (with R-squared ~ 1) and easy assignment of analytic functions for the quadratic polynomial coefficients

```
##    quantile_pt   sq_coeff   lin_coeff  intercept
## 1    0.0100000 101.009554   -2.019655   1.018776
## 2    0.1000000  11.111051   -2.222162   1.112065
## 3    0.2500000   5.333304   -2.666638   1.333791
## 4    0.3333333   4.499976   -2.999976   1.500386
## 5    0.4000000   4.166644   -3.333311   1.667024
## 6    0.5000000   3.999978   -3.999978   2.000344
## 7    0.6000000   4.166644   -4.999977   2.500358
## 8    0.6666667   4.499976   -5.999976   3.000386
## 9    0.7500000   5.333304   -7.999971   4.000458
## 10   0.9000000  11.111051  -19.999940  10.000954
## 11   0.9900000 101.009554 -199.999453 100.008675
```

Quantile value	$\frac{1}{\theta(1-\theta)}b^2$	$\frac{-2}{(1-\theta)}b$	$\frac{1}{(1-\theta)}$
0.01	10000/99	-200/99	100/99
0.1	100/9	-20/9	10/9
0.25	15/3	-8/3	4/3
1/3	9/2	-6/2	3/2
0.4	25/6	-10/3	5/3
0.5	4/1	-4/1	2/1
0.6	25/6	-5/1	5/2
2/3	9/2	-6/1	3/1
0.75	15/3	-8/1	4/1
0.9	100/9	-20/1	10/1
0.99	10000/99	-200/1	100/1

Hence it is the case that the simple quadratic polynomial

$$\min_{b\,\epsilon(0,1)} \left(\frac{1}{\theta(1-\theta)}b^2 + \frac{-2}{(1-\theta)}b + \frac{1}{(1-\theta)} \right) \in [1, \text{inf}) \tag{8}$$

is an excellent approximation to the quasi-normalised quantile estimation function in the percentile scale, using the ecdf_int assignment of percentile values to the known number of data points.

Since the quantile estimation function already provides the minimum value b in the original measurement scale for a given θ (without sorting the dataset). The main use of the quadratric polynomial proxy in the percentile scale is then to derive an (approximate) estimate for the quantile variance. As will be shown in the next section, this quantile variance estimate also does not need sorting of the dataset as only the number of data points and the ecdf_int defintion is required for the calculation.

Variance estimates for quantiles by derivation of a density function

Given the quadratic polynomial proxy for the quasi-normalised estimation function is differentiable, it is possible to create the following density function by transformation of the expression

Firstly, since equation (8) in the interval (0,1) has a minimum value of 1, the following expression (with -1 added) is also a valid function for finding the minimum percentile b for a given θ

$$\min_{b\,\epsilon(0,1)} \left(\frac{1}{\theta(1-\theta)} b^2 + \frac{-2}{(1-\theta)} b + \frac{1}{(1-\theta)} - 1 \right) \; \in [0, \inf) \tag{9}$$

Secondly, the expression can be inverted from a minimisation problem to a maximisation problem by using a multiplicative factor of -1

$$\max_{b\,\epsilon(0,1)} \{ - \left(\frac{1}{\theta(1-\theta)} b^2 + \frac{-2}{(1-\theta)} b + \frac{1}{(1-\theta)} - 1 \right) \} \; \in (-\inf, 0] \tag{10}$$

The function is then transformed to form likelihood scale values, ie. lying in the interval (0,1) via exponentiation.

$$\max_{b\,\epsilon(0,1)} \exp\{ - \left(\frac{1}{\theta(1-\theta)} b^2 + \frac{-2}{(1-\theta)} b + \frac{1}{(1-\theta)} - 1 \right) \} \; \in [0, 1] \tag{11}$$

Finally, a density functional form can be defined from equation 11 noting (i) the quadratic nature of the exponent, a property that is known to be shared by the normal distribution and (ii) including the sample size as a multiplicative factor consistent with the asymptotic $\sqrt{n}$ behaviour expected for population quantiles (1). In particular,

$$f_{q-proxy}(b, \theta, n) \propto \exp\{ - \left(\frac{1}{\theta(1-\theta)} b^2 + \frac{-2}{(1-\theta)} b + \frac{1}{(1-\theta)} - 1 \right) \frac{n}{2} \} \; \in [0, 1] \tag{12}$$

Using the standard normal transformation, it is then simple to derive the mean and central limit theorem (CLT) variance of equation (12)

$$- \left(\frac{(b-\mu)^2}{2\sigma^2_{CLT}} \right) = - \left(\frac{1}{\theta(1-\theta)} b^2 + \frac{-2}{(1-\theta)} b + \frac{1}{(1-\theta)} - 1 \right) \frac{n}{2} \tag{13}$$

first by expanding the LHS of equation (13)

$$- \left(\frac{b^2}{2\sigma^2_{CLT}} + \frac{-2\mu b}{2\sigma^2_{CLT}} + \frac{\mu^2}{2\sigma^2_{CLT}} \right) = - \left(\frac{1}{\theta(1-\theta)} b^2 + \frac{-2}{(1-\theta)} b + \frac{1}{(1-\theta)} - 1 \right) \frac{n}{2} \tag{14}$$

and then equating coefficients

$$\frac{1}{2\sigma^2_{CLT}} \equiv \frac{n}{2\theta(1-\theta)} \tag{15}$$

$$\frac{-2\mu}{2\sigma^2_{CLT}} \equiv \frac{-2n}{2(1-\theta)} \tag{16}$$

$$\frac{\mu^2}{2\sigma^2_{CLT}} \equiv \frac{n}{2(1-\theta)} - \frac{n}{2} = \frac{\theta n}{2(1-\theta)} \tag{17}$$

This derived density function gives the following compact solution for the population quantile and its CLT variance of the quadratic polynomial proxy for the quantile estimation function in the percentile scale

$$\mu = \theta \tag{18}$$

$$\sigma_{CLT} = \sqrt{\frac{\theta(1-\theta)}{n}} \tag{19}$$

Since the density function is found to be proportional to a normal distribution depending on θ and n, the proportionality factor needed in equation (12) is also identified by integration over the interval [0,1], giving the density function the compact form

$$f_{q-proxy}(b,\theta,n) = \left(\frac{1}{\int_0^1 \exp\{\frac{-(b-\theta)^2}{2\sigma^2_{CLT}}\}\ db} \right) \exp\{\frac{-(b-\theta)^2}{2\sigma^2_{CLT}}\} \tag{20}$$

and in the large n limit, the density function of the quadratic polynomial proxy for the quantile estimating function (in the percentile scale) converges to a normal distribution form

$$f_{q-proxy}(b,\theta,n)_{CLT} \rightarrow \frac{1}{\sqrt{2\pi\theta(1-\theta)n}} \exp\{\frac{-(b-\theta)^2}{2\sigma^2_{CLT}}\} \tag{21}$$

Comparison to analytic median jackknife variance results in percentile scale

While equations(20,21) are a very compact result, and may add alternative insights to the coverage research done on $var(proportion\, p) = \sqrt{\frac{p(1-p)}{n}}$ (which has the same CLT variance form to the quadratic polynomial proxy for quantiles) by Agresti et al (8) and earlier authors, the above variance estimator will be assessed for accuracy in this paper in the following two ways.

1. Within the percentile scale, known results for the CLT jacknife variance estimate of median can be used to assess if equations (20,21) are accurate for medians or require further calibration.
2. Numerical comparison of the variance results using equations (20,21) after backtransformation of the percentile scale results to the original measurement scale to bootstrap variance results for quantile estimation function for several continuous distributions.

Test 1 checks will be conducted in the rest of this section and test 2 checks in the following section.

Martin (2015) presented a derivation of jackknife variance estimates of median in the percentile scale which can then be backtransformed to the original measurement scale providing consistent estimates of median variance for unweighted samples.

The proof involved series expansions of powers of $1/n$. However, since that paper used the standard (1/n, 2/n, ... n/n) quantile definition and this quantile paper has used the ecdf_int definition (1/2n, 3/2n, 5/2n, ... (2n-1)/2n) the proof has been repeated in appendix A to show (i) that the ecdf_int quantile definition doesn't change the outcome and (ii) that a series expansion approach for the jackknife variance estimate is not necessary.

Comparing the jackknife variance of the median to the quadratic polynomial proxy CLT variance estimate for the median point (from equation (19))

$$\sigma^2_{CLT}(median\, based\, on\, quadratic\, polynomial) = \frac{(\frac{1}{2})(\frac{1}{2})}{n} \tag{22}$$

$$= \frac{1}{4n} \tag{23}$$

$$\sigma^2_{jk\,median\,in\,percentile\,scale} = \frac{1}{4(n-1)} \tag{24}$$

$$\rightarrow \frac{1}{4n}\, for\, large\, n \tag{25}$$

there is good agreement, so no further calibration adjustment to the estimator in the CLT limit for median value is required.

Numerical calculations for unweighted sample quantiles of several continuous distributions

The algorithm for using the quadratic polynomial proxy as the quantile variance estimator is

1. Use the quantile estimation function (or quasi-normalised version) in the original measurement scale to determine X distribution values for each given θ. In practice, this can be done using the "quantreg" r package to perform quantile regression of the X distribution against an "intercept only" model.
2. Use the given sample size and quantile point θ to determine the quadratic polynomial proxy variance estimate or confidence interval in the percentile scale (using equations (18,19,21) and/or equations (18,20))
3. Use the quantile estimation function (or quasi-normalised version) in the original measurement scale to determine X distribution values for each given θ value of the variance estimate or confidence interval obtained in step 2

The bootstrap quantile variance estimates will be obtained by using the "quantreg" r package results in step 1 and calling the summary.boot option.

Using the quadratic polynomial proxy for the variance estimate of the quantile estimating function, means that in the percentile scale, the variance distribution of possible theta values is smooth, as shown in figures 3a & c for $\theta = 0.025$ & 0.5, for a sample size of 1000 from a N(0,1) dist. However, on backtransformation to the original measurement scale, as shown in figures 3b & d, the variance distribution becomes a stepped curve (due to the linear interpolation of the ecdf_int used in the calculation).

This "empirical variance distribution" has the step function behaviour of sample cdfs, corresponding to the density of the sample distribution (blue markers) and so may be considered a natural companion distribution of the empirical cdf whereas the bootstrap variance estimates from quantile regression are smoothed distributions.

Figure 3 is followed by tables of (i) the results of the quantile regression fit (intercept only model to N(0,1) distribution) and (ii) backtransformed quadratic polynomial proxy variance estimates for more quantile values 0.025,.1,.25,.5,.75,.9 & .975. The asymmetry in the ± one standard error confidence interval for n=1000 is apparent in the results for the quadratic polynomial proxy variance estimates. However, the average standard error of the interval estimate is reasonably similar to the bootstrap values. For smaller sample sizes and quantiles in the tail of the sample distribution, the asymmetry in the confidence interval of the empirical variance distribution will only get stronger.

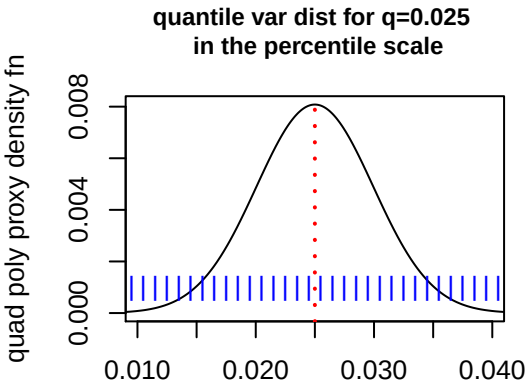


figure 3a

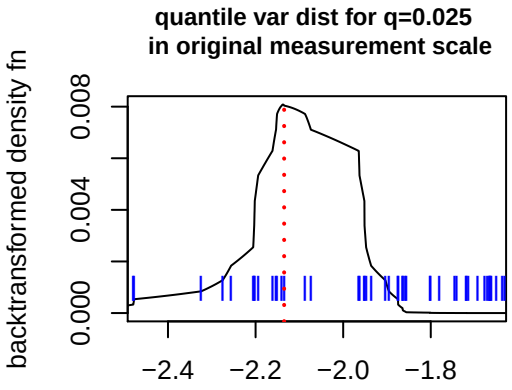


figure 3b

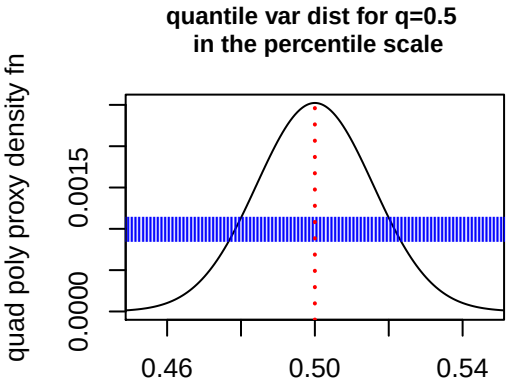


figure 3c

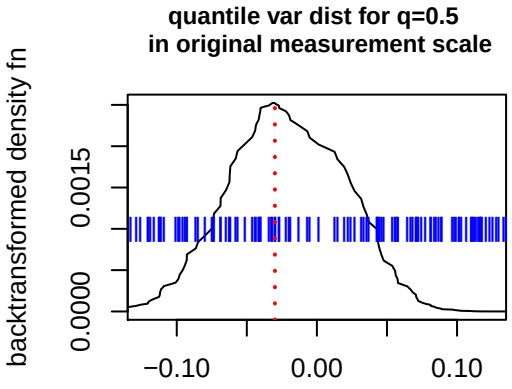


figure 3d

```
## [1] "N(0,1) dist - n=1000"
```

```
## [1] "quantile regression fit - rq"
```

```
##             beta0
## 0.025 -2.13479155
## 0.1   -1.29784640
## 0.25  -0.68306316
## 0.5   -0.02995135
## 0.75   0.62365748
## 0.9    1.22807862
## 0.975  1.94485995
```

```
## [1] "quantile regression fit - summary.rq(...,se='boot')"
```

```
##             beta0  std_error     t_value        Pr
## 0.025 -2.13505859 0.10774669 -19.8155379 0.0000000
## 0.1   -1.30133885 0.05566373 -23.3785796 0.0000000
## 0.25  -0.68333872 0.04588480 -14.8924856 0.0000000
## 0.5   -0.03143517 0.04009836  -0.7839513 0.4332545
## 0.75   0.62378242 0.03814910  16.3511695 0.0000000
## 0.9    1.22926668 0.04200756  29.2629838 0.0000000
## 0.975  1.94986875 0.08433364  23.1208896 0.0000000
```

```
## [1] "quadratic polynomial proxy - backtransformation fit"
```

```
##       lower_std_dist upper_std_dist average_std_value
## 0.025    -0.05925445     0.17209288        0.11567366
## 0.1      -0.04704810     0.04981353        0.04843081
## 0.25     -0.05296750     0.03716536        0.04506643
## 0.5      -0.03244573     0.05408806        0.04326690
## 0.75     -0.03678176     0.03661655        0.03669915
## 0.9      -0.02647153     0.04438942        0.03543047
## 0.975    -0.07721293     0.06372868        0.07047080
```

In Table 1 and the appendix, backtransformed quadratic polynomial proxy (empirical) variance estimates of quantiles for unweighted samples are compared to bootstrap quantile confidence intervals for five examples of continuous distributions.

Two common symmetric distributions

(i) uniform distribution unif(0,1) of sample 1000, and

(ii) the standard normal distribution N(0,1) of sample 1000

an artificial linear smooth symmetric distribution in the measurement scale

(iii) systematic unif(0,1) of random sample size 415 using start/skip selection

and two common skewed distributions

(iv) a log-normal distribution exp(N(0.1,1)) of sample 1000, and

(v) a continuous bivariate normal distribution 1/3(N(55,5.5))+2/3(N(80,6)) of sample 270. This example is a continuous distribution analogue of the old faithful geyser waiting duration data located as data(faithful) on the r library(datasets).

This range of datasets has different smoothness and nonlinearity of the cdf across the quantile range [0,1]. It is seen that the empirical variance distribution has asymmetric standard errors for samples sizes 270-1000 in comparison to the bootstrap calculations except for the artifical linear distribution example. For this artifical example (iii), the original scale cdf is linear and the empirical variance distribution is expected to be symmetric and very closely match the bootstrap results, which is the case.

In general, comparing the "empirical variance distribution standard error estimates using the percentile scale"" to the bootstrap calculations, results in good agreement for the "average of the asymmetric standard errors" from the empirical variance distribution of the unweighted quantiles across all five examples. The transformation of the quantile variance calculation to the linear ecdf_int scale allows the "smoothed quantile estimating function approach" to consider the full characteristic of the sample distribution and the back-transformation from the percentile scale to the original measurement scale handles the nonlinearity in the observed data distribution.

Table 1: Quantile regression estimates of intercept only model, bootstrap standard errors and proxy standard errors of unweighted samples

Distribution	quantile	est value	bootstrap std error	average proxy std error	asymm proxy std errors
random uniform(0,1)	0.025	0.0026	± 0.0032	± 0.0015	(-0.0019,0.0012)
""	0.10	0.0778	± 0.0102	± 0.0070	(-0.0072,0.0068)
""	0.25	0.2432	± 0.0126	± 0.0138	(-0.0171,0.0104)
""	0.5	0.4895	± 0.0123	± 0.0132	(-0.0116,0.0149)
""	0.75	0.7242	± 0.0196	± 0.0192	(-0.0103,0.0282)
""	0.9	0.8818	± 0.0101	± 0.0099	(-0.0095,0.0103)
""	0.975	0.9711	± 0.0043	± 0.0033	(-0.0033,0.0034)

Distribution	quantile	est value	bootstrap std error	average proxy std error	asymm proxy std errors
random normal(0,1)	0.025	-2.1348	± 0.1078	± 0.1157	(-0.0593,0.1721)
“”	0.10	-1.2979	± 0.0557	± 0.0484	(-0.0471,0.0498)
“”	0.25	-0.6831	± 0.0459	± 0.0451	(-0.0530,0.0372)
“”	0.5	-0.0300	± 0.0410	± 0.0433	(-0.0325,0.0541)
“”	0.75	0.6237	± 0.0382	± 0.0367	(-0.0368,0.0366)
“”	0.9	1.2281	± 0.0420	± 0.0354	(-0.0265,0.0444)
“”	0.975	1.9449	± 0.0843	± 0.0705	(-0.0772,0.0637)
systematic uniform(0,1)	0.025	0.0271	± 0.0074	± 0.0072	(-0.0072,0.0072)
“”	0.10	0.1018	± 0.0136	± 0.0144	(-0.0144,0.0144)
“”	0.25	0.2510	± 0.0208	± 0.0217	(-0.0217,0.0217)
“”	0.5	0.5013	± 0.0251	± 0.0241	(-0.0241,0.0241)
“”	0.75	0.7516	± 0.0214	± 0.0217	(-0.0217,0.0217)
“”	0.9	0.9009	± 0.0144	± 0.0144	(-0.0144,0.0144)
“”	0.975	0.97547	± 0.0079	± 0.0072	(-0.0072,0.0072)
log-normal N(0.1,1)	0.025	0.0158	± 0.0074	± 0.0120	(-0.0065,0.0175)
“”	0.10	0.3513	± 0.0163	± 0.0160	(-0.0141,0.0178)
“”	0.25	0.5868	± 0.0249	± 0.0285	(-0.0315,0.0254)
“”	0.5	1.0872	± 0.0404	± 0.0413	(-0.0381,0.0445)
“”	0.75	2.1934	± 0.0779	± 0.0707	(-0.0734,0.0681)
“”	0.9	3.7931	± 0.1308	± 0.1051	(-0.0889,0.1214)
“”	0.975	6.8360	± 0.8589	± 0.9175	(-0.1238,1.7112)
$\frac{1}{3}$N(55,5.5)+$\frac{2}{3}$N(80,6)	0.025	48.16	± 1.6878	± 1.5209	(-2.2188,0.8230)
“”	0.10	52.97	± 0.6888	± 0.6439	(-0.8096,0.4781)
“”	0.25	59.78	± 1.5041	± 1.3028	(-1.6429,0.9628)
“”	0.5	76.84	± 0.8308	± 0.7378	(-0.9610,0.5146)
“”	0.75	81.67	± 0.4497	± 0.4447	(-0.5648,0.3245)
“”	0.9	85.93	± 0.7630	± 0.5499	(-0.3975,0.7022)
“”	0.975	91.42	± 0.9172	± 1.1671	(-1.4353,0.8989)

Conclusions

The transformation of the quantile variance calculation to the percentile scale for the quantile estimating function of unweighted samples, allows (i) analytical variance estimates to be derived consistent with median jackknife variance estimates, (ii) good confidence interval agreement with quantile regression bootstrap calculations on backtransformation to the original measurement scale, (iii) no requirement for sorting of the dataset when used in conjunction with the quantile estimating function.

While the approach has shown promise, the validation checks have mostly involved regression modelling with an “intercept only model”. Thus the use of the exponentiation step in equation (12) only has only been really tested for quantile analysis of a single distribution, rather than multivariate quantile regression.

Preliminary coverage work, using the percentile scale to calculate the variance of quantile regression residuals, confirms good results for the confidence intervals of the intercept coefficients but potentially lower coverage for the slope coefficients estimates. It may be use of the beta distribution (8) rather than exponentiation is also a good approach.

References

1. Koencker, R. W. & Bassett G., Econometrica, 1978, vol. 46, issue 1, pages 33-50
2. Martin J.P.D., 2015, http://figshare.com/articles/Improved_jackknife_estimates_for_median_variance_in_equally_weighted_samples_using_percentile_scale_based_calculations/1332463
3. Martin J.P.D., 2015, http://figshare.com/articles/effect_of_sample_size_and_repeated_values_for_jackknife_CI_estimates_of_unweighted_median/1332464
4. Miller, R.G. (1974) The Jackknife–a review, Biometrika 61, pp1-15
5. Koencker, R. W., Portnoy S. et al, https://cran.r-project.org/web/packages/quantreg/quantreg.pdf
6. https://en.wikipedia.org/wiki/Quantile
7. Brown, B. M. and Wang, Y.-G. (2005). Standard errors and covariance matrices for smoothed rank estimators. Biometrika 92 149-158. MR2158616
8. Agresti A. and Caffo B., The American Statistician, Vol. 54, No. 4, (Nov., 2000), pp. 280-288

Appendix A - Jackknife variance estimate for median of unweighted samples in the percentile scale

In the drop one unit jackknife variance approach for such a distribution (ignoring continuity corrections), the (n-1) subsampled cdf values assigned to each ordered point are

1/2(n-1), 3/2(n-1), 5/2(n-1), ... , (2n-3)/2(n-1)

Performing the drop one unit jackknife variance calculation about the median point in this reference frame.

If n is even,

For the lower half of the ordered units, the drop one unit jackknife estimate of the median point which is an interpolated point, has the value using the ecdf_int defintion

$$jk\_lower = ((2(\frac{n}{2} + \frac{1}{2}) - 1) - 2) \cdot \frac{1}{2(n-1)} \tag{26}$$

$$= \frac{n-2}{2(n-1)} \tag{27}$$

For the upper other half of the ordered units, the drop one jackknife estimate of median point has the numerator value unchanged

$$jk\_lower = (2(\frac{n}{2} + \frac{1}{2}) - 1) \cdot \frac{1}{2(n-1)} \tag{28}$$

$$= \frac{n}{2(n-1)} \tag{29}$$

The mean of the drop jackknife estimates, for n even, is

$$\frac{1}{n} \cdot (\frac{n}{2} \cdot (jk\_lower + jk\_upper)) = \frac{1}{n} \cdot (\frac{n}{2} \cdot (\frac{n-2}{2(n-1)} + \frac{n}{2(n-1)})) \tag{30}$$

$$= \frac{1}{4(n-1)} \cdot (n - 2 + n) \tag{31}$$

$$= \frac{1}{4(n-1)} \cdot (2(n-1)) \tag{32}$$

$$= \frac{1}{2} \tag{33}$$

which is the full sample median estimate

$$\therefore bias\_jk\_n\_even = 0 \tag{34}$$

If n is odd,

The expressions for jk_lower and jk_upper in edcf_int definition are the same but the mean of the drop one jackknife estimates, for n odd, has one extra data point for jk_lower

$$\frac{1}{n} \cdot ((floor(\frac{n}{2}) + 1)jk\_lower + floor(\frac{n}{2})jk\_upper)$$
$$= \frac{1}{n} \cdot (floor(\frac{n}{2} + 1) \cdot (\frac{n-2}{2(n-1)}) + floor(\frac{n}{2}) \cdot (\frac{n}{2(n-1)})) \quad (35)$$

$$= \frac{1}{2n(n-1)} \cdot (floor(\frac{n}{2}) \cdot ((n-2) + n) + (n-2)) \quad (36)$$

$$= \frac{1}{2n(n-1)}(floor(\frac{n}{2}) \cdot 2(n-1) + (n-2)) \quad (37)$$

$$= \frac{1}{2} + \frac{n-2}{2n(n-1)} \quad (38)$$

$$\approx \frac{1}{2} + \frac{1}{2n} + ... \quad (39)$$

which differs from the full sample median estimate (as expected for n odd)

$$\therefore bias\_jk\_n\_odd \approx +\frac{1}{2n} + ... \quad (40)$$

The next step in the jackknife estimation is to calculate the variance of the distribution of the jackknife estimates.

The variance of the jackknife estimates, for n even is

$$(n-1) \cdot \frac{1}{n} \sum (jk\_est - jk\_mean)^2 = \frac{n-1}{n} \cdot \frac{n}{2} \cdot ((jk\_lower - \frac{1}{2})^2 + (jk\_upper - \frac{1}{2})^2) \quad (41)$$

$$= \frac{n-1}{n} \cdot \frac{n}{2} \cdot ((\frac{n-2}{2(n-1)} - \frac{1}{2})^2 + (\frac{n}{2(n-1)} - \frac{1}{2})^2) \quad (42)$$

$$= \frac{1}{8(n-1)} \cdot (((n-2) - (n-1))^2 + (n - (n-1))^2) \quad (43)$$

$$= \frac{1}{8(n-1)} \cdot ((-1)^2 + (1)^2) \quad (44)$$

$$= \frac{1}{8(n-1)} \cdot ((-1)^2 + (1)^2) \quad (45)$$

$$= \frac{1}{4(n-1)} \quad (46)$$

The variance of the jackknife estimates, for n odd is

$$(n-1) \cdot \frac{1}{n} \sum (jk\_est - jk\_mean)^2 = \frac{n-1}{n} \cdot ((floor(\frac{n}{2}) + 1) \cdot ((jk\_lower - \frac{1}{2})^2$$
$$+ (floor(\frac{n}{2}) \cdot (jk\_upper - \frac{1}{2})^2))) \quad (47)$$

$$= \frac{n-1}{n} \cdot ((floor(\frac{n}{2}) + 1) \cdot ((\frac{n-2}{2(n-1)} - \frac{1}{2})^2 + (floor(\frac{n}{2}) \cdot (\frac{n}{2(n-1)} - \frac{1}{2})^2))) \tag{48}$$

$$= \frac{n-1}{n} \cdot (floor(\frac{n}{2}) \cdot ((\frac{n-2}{2(n-1)} - \frac{1}{2})^2 + (\frac{n}{2(n-1)} - \frac{1}{2})^2) + (\frac{n-2}{2(n-1)} - \frac{1}{2})^2) \tag{49}$$

$$= \frac{1}{4n(n-1)} \cdot (floor(\frac{n}{2}) \cdot (((n-2) - (n-1))^2 + (n - (n-1))^2) + ((n-2) - (n-1))^2) \tag{50}$$

$$= \frac{1}{4n(n-1)} \cdot (floor(\frac{n}{2}) \cdot ((-1)^2 + (1)^2) + (-1)^2) \tag{51}$$

$$= \frac{1}{4n(n-1)} \cdot (floor(\frac{n}{2}) \cdot 2 + 1) \tag{52}$$

$$= \frac{1}{4n(n-1)} \cdot n \tag{53}$$

$$= \frac{1}{4(n-1)} \tag{54}$$

Therefore, the jackknife variance estimate of the unweighted median, in the percentile scale, is of the common form (for n odd or even)

$$\sigma_{jk\,median\,in\,percentile\,scale} = \frac{1}{4(n-1)} \tag{55}$$

Appendix B - Results of quantile regression, and bootstrap & back-transformed quadratic polynomial proxy quantile variance calculations

```
## [1] "unif(0,1) dist - n=1000"
```

```
## [1] "quantile regression fit - rq"
```

```
##            beta0
## 0.025 0.02601124
## 0.1   0.07783348
## 0.25  0.24319823
## 0.5   0.48953543
## 0.75  0.72419012
## 0.9   0.88184184
## 0.975 0.97106848
```

```
## [1] "quantile regression fit - summary.rq(...,se='boot')"
```

```
##            beta0   std_error    t_value           Pr
## 0.025 0.02605549 0.003234597   8.055249 2.220446e-15
## 0.1   0.07918112 0.010220640   7.747178 2.309264e-14
## 0.25  0.24338662 0.012615345  19.292902 0.000000e+00
## 0.5   0.48964047 0.012277121  39.882354 0.000000e+00
## 0.75  0.72501748 0.019613240  36.965716 0.000000e+00
## 0.9   0.88184733 0.010103679  87.279824 0.000000e+00
## 0.975 0.97115708 0.004338639 223.839082 0.000000e+00
```

```
## [1] "quadratic polynomial proxy - backtransformation fit"
```

```
##       lower_std_dist upper_std_dist average_std_value
## 0.025   -0.001880320    0.001196404       0.001538362
## 0.1     -0.007163878    0.006780625       0.006972251
## 0.25    -0.017080727    0.010421660       0.013751194
## 0.5     -0.011584627    0.014896005       0.013240316
## 0.75    -0.010249780    0.028228207       0.019238994
## 0.9     -0.009485524    0.010331737       0.009908631
## 0.975   -0.003263098    0.003367797       0.003315447
```

```
## [1] "systematic uniform - n=415"
```

```
## [1] "quantile regression fit - rq"
```

```
##            beta0
## 0.025 0.02713593
## 0.1   0.10175128
## 0.25  0.25098199
## 0.5   0.50130447
## 0.75  0.75162695
## 0.9   0.90085766
## 0.975 0.97547301
```

```
## [1] "quantile regression fit - summary.rq(...,se='boot')"
```

```
##            beta0   std_error    t_value           Pr
## 0.025 0.02713593 0.007425563   3.654393 2.909308e-04
## 0.1   0.10175128 0.013577844   7.493920 4.090062e-13
## 0.25  0.25098199 0.020785887  12.074635 0.000000e+00
```

```
## 0.5   0.50130447 0.025139977  19.940530 0.000000e+00
## 0.75  0.75162695 0.021413135  35.101210 0.000000e+00
## 0.9   0.90085766 0.014408664  62.521942 0.000000e+00
## 0.975 0.97547301 0.007865662 124.016645 0.000000e+00
```

```
## [1] "quadratic polynomial proxy - backtransformation fit"
```

```
##       lower_std_dist upper_std_dist average_std_value
## 0.025   -0.007220841    0.007220841       0.007220841
## 0.1     -0.014441681    0.014441681       0.014441681
## 0.25    -0.021662522    0.021662522       0.021662522
## 0.5     -0.024069469    0.024069469       0.024069469
## 0.75    -0.021662522    0.021662522       0.021662522
## 0.9     -0.014441681    0.014441681       0.014441681
## 0.975   -0.007220841    0.007220841       0.007220841
```

```
## [1] "log-normal N(0.1,1) dist - n=1000"
```

```
## [1] "quantile regression fit - rq"
```

```
##           beta0
## 0.025 0.1826479
## 0.1   0.3512814
## 0.25  0.5868445
## 0.5   1.0872189
## 0.75  2.1934302
## 0.9   3.7930709
## 0.975 6.8360109
```

```
## [1] "quantile regression fit - summary.rq(...,se='boot')"
```

```
##            beta0  std_error  t_value Pr
## 0.025  0.1872052 0.01575034 11.88579  0
## 0.1    0.3515415 0.01630564 21.55950  0
## 0.25   0.5882277 0.02490510 23.61877  0
## 0.5    1.0904539 0.04040664 26.98700  0
## 0.75   2.1938350 0.07789814 28.16287  0
## 0.9    3.8169323 0.13080612 29.18007  0
## 0.975 10.3908863 0.85894144 12.09732  0
```

```
## [1] "quadratic polynomial proxy - backtransformation fit"
```

```
##       lower_std_dist upper_std_dist average_std_value
## 0.025   -0.006541497     0.01748509        0.01201329
## 0.1     -0.014096904     0.01780984        0.01595337
## 0.25    -0.031511303     0.02538231        0.02844680
## 0.5     -0.038129846     0.04446660        0.04129822
## 0.75    -0.073416702     0.06807324        0.07074497
## 0.9     -0.088860601     0.12139430        0.10512745
## 0.975   -0.123785866     1.71117908        0.91748247
```

```
## [1] "bivariate normal $\frac{1}{3}$N(55,5.5)+$\frac{2}{3}$N(80,6) dist - n=270"
```

```
## [1] "quantile regression fit - rq"
```

```
##          beta0
## 0.025 48.15886
## 0.1   52.97092
## 0.25  59.77709
## 0.5   76.83587
```

```
## 0.75  81.66641
## 0.9   85.92700
## 0.975 91.42332
```

```
## [1] "quantile regression fit - summary.rq(...,se='boot')"
```

```
##          beta0 std_error   t_value Pr
## 0.025 48.15886 1.6878166  28.53323  0
## 0.1   53.00392 0.6888204  76.94882  0
## 0.25  59.77709 1.5041189  39.74226  0
## 0.5   76.83992 0.8308053  92.48849  0
## 0.75  81.66641 0.4496774 181.61111  0
## 0.9   86.13580 0.7630104 112.88941  0
## 0.975 91.42332 0.9172243  99.67390  0
```

```
## [1] "quadratic polynomial proxy - backtransformation fit"
```

```
##       lower_std_dist upper_std_dist average_std_value
## 0.025     -2.2188193      0.8229995         1.5209094
## 0.1       -0.8096290      0.4781064         0.6438677
## 0.25      -1.6428598      0.9627517         1.3028058
## 0.5       -0.9609622      0.5146009         0.7377816
## 0.75      -0.5647841      0.3245227         0.4446534
## 0.9       -0.3975341      0.7022067         0.5498704
## 0.975     -1.4353267      0.8989092         1.1671180
```