Quantile regression model coefficient CIs using the empirical variance distribution approximation

John P. D. Martin

Wednesday, December 16, 2015

Executive Summary

This paper investigates the application of the empirical variance distribution (evd) function (Martin (1,2)), to estimate the confidence interval bounds of the quantile regression model coefficient estimates (Koencker & Bassett (3)), for homoscedastic unweighted data.

Analysis of the coverage accuracy of the results is conducted by repeated sampling of several sample sizes to known regression models and error distributions. Table 1 displays the coverage performance of several quantile regression variance estimates of unweighted regression examples, for sample size 1000. The empirical variance distribution (evd) based estimates are compared to default bootstrap quantile confidence intervals calculated using 600 replicates (from the quantreg r package (4)).

It can be seen that for this modest sample size, the bootstrap estimator and the two evd based confidence interval estimators exhibit nominal 95% coverage for homoscedastic iid cases in slope estimates and >93% in the intercept estimates. For extreme quantiles in smaller samples or when homoscedastic iid error is not present, the coverage performance is lower.

Table 1: Quantile regression confidence interval estimator coverage of sample size 1000 for (slope,intercept) respectively, based on 1000 repeated samples

Regression $Y = \beta X + \varepsilon$	quantile	bootstrap coverage	evd_max coverage	bin_max coverage
$y = x + rnorm(0, 10)$	0.1	(0.947,0.944)	(0.956,0.965)	(0.954,0.960)
“”	0.5	(0.945,0.953)	(0.957,0.971)	(0.957,0.971)
“”	0.9	(0.950,0.952)	(0.958,0.960)	(0.960,0.957)
$y = x^2 + rnorm(0, 10)$	0.1	(0.953,0.947)	(0.969,0.937)	(0.966,0.933)
“”	0.5	(0.944,0.956)	(0.957,0.934)	(0.957,0.934)
“”	0.9	(0.957,0.947)	(0.962,0.928)	(0.960,0.924)
$y = x + urlaplace(0, 10)$	0.1	(0.950,0.948)	(0.958,0.961)	(0.951,0.958)
“”	0.5	(0.956,0.958)	(0.959,0.970)	(0.959,0.970)
“”	0.9	(0.944,0.951)	(0.957,0.957)	(0.953,0.958)
$y = x + \frac{x^2}{2} + rnorm(0, 10)$	0.1	(0.960,0.964,0.947)	(0.956,0.964,0.929)	(0.955,0.960,0.929)
“”	0.5	(0.955,0.958,0.959)	(0.962,0.961,0.943)	(0.957,0.952,0.934)
“”	0.9	(0.957,0.961,0.940)	(0.953,0.961,0.921)	(0.953,0.957,0.910)
$y = x + AR(1)$	0.1	(0.953,0.837)	(0.957,0.863)	(0.961,0.863)
“”	0.5	(0.950,0.798)	(0.962,0.835)	(0.962,0.835)
“”	0.9	(0.957,0.830)	(0.959,0.868)	(0.957,0.868)

In Table 1, the regression degrees of freedom model adjustment $\sqrt{n/(n-p-1)}$ for the evd based estimators was performed in the percentile scale.

The first example examines a location shift only model with gaussian noise. The second example involves a nonlinear location shift model analysed using polynomial quantile regression, providing a stronger test of the intercept confidence interval estimator. The third example, examines the impact on coverage performance of a

different iid noise distribution for which quantile regression is known to be a more efficient regression estimator. The fourth example, involving additive polynomial regression examines the coverage performance when multiple explanatory variables are present and illustrates the good coverage achievable if the homoscedastic iid model has correctly specified nonlinearities. Finally, to contrast the good performance for homoscedastic iid examples, the fifth example in Table 1, contains autocorrelation (AR(1)) between the errors terms of the sample observation. In this non-iid example, the coverage performance of the intercept confidence interval estimators is lower than desired for the bootstrap as well as the evd based estimators due to model error iid misspecification. In this final case, the slope confidence interval estimates remain at 95% as their calculation is robust against the non-iid behaviour.

Importantly, the use of the quadratic polynomial proxy approach applied to estimate sample quantile variances (1,2) has now been extended, to provide useful approximations of quantile regression model variances in homoscedastic unweighted cases.

Introduction

Quantiles (5) 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 (3) 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.

Another 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 (6). Consistent with that approach, Martin (1) identified an analytic quadratic polynomial smoothing function, in the percentile scale, for the quantile estimating function of unweighted samples. This analytic function, only requires the sample size and selected quantile value, to calculate the sample quantile confidence interval (CI) bounds in the percentile scale.

Backtransforming to the original measurement scale, using the CI bounds in the quantile estimation function calculations on the sample distribution results in the empirical variance distribution (evd). As shown in (1), the evd is an asymmetric stepped sample CI, in contrast to smooth symmetric bootstrap sample CIs, but similar in morphology to the discrete cumulative density function (cdf).

In Martin (2), several evd based sample quantile CI estimators were shown to have nominal 95% performance for samples sizes 50-100-1000, except for extreme quantiles in the smallest samples. Some improvement in the sample quantile CI coverage was also shown to be possible for these extreme cases, via use of quantile regression extrapolated 0th,100th quantile sample bounds.

In this current research, the evd based sample quantile CI estimators (2) have been trialled, assessed and adapted where required as quantile regression model coefficient CI estimators.

In this paper, the best performing evd based estimators for quantile regression model coefficient CIs of homoscedastic iid cases, will be presented along with the regression formulae required to transform the CI bounds to original scale values. The results are then compared to the known population slope and intercept regression values as well as default quantreg () bootstrap estimates.

In practice, this evd based model CI method to approximate quantile regression model coefficient CIs for homoscedastic iid cases is very easy to perform.

(i) Firstly, the quantile estimating function is used to perform the quantile regression modelling and derive the residuals distribution,
(ii) next, using only the sample size and the given quantile value $\theta \epsilon (0, 1)$, the evd variance estimator approach provides the confidence interval bound values (θ_{LB}, θ_{UB}) for the quantile regression residuals distribution,
(iii) the quantile estimating function is then used on the residuals distribution to backtransform the (quantile) confidence interval bound values (θ_{LB}, θ_{UB}) to the original measurement scale, and
(iv) simple regression formulae (analogous to the linear regression case) use the evd based approximations of the quantile regression residual CIs as input, to calculate evd based approximation estimates of the quantile regression model coefficient CIs.

Based on the present results, the regression degrees of freedom adjustment $\sqrt{\frac{n}{(n-p-1)}}$ required for model coefficient CIs can be applied in either step (ii) or (iv). There is weak evidence that performing the degrees of freedom adjustment in step (ii) has slightly better coverage for smaller sample sizes.

Sample quantile confidence interval evd based estimators

In (2), it was found that three sample quantile CI estimators calculated in the percentile scale before backtransformation to the original measurement scale,

(i) the binomial distribution,
(ii) the empirical variance distribution and
(iii) the total variance $E[Var(\theta|x)] + Var(E[\theta|x])$ confidence interval

all produced nominal 95% coverage for sample quantiles between 0.1-0.9 similar to bootstrap results.

binomial CI estimator

The binomial variance distribution formula is a total variance estimator, selfconsistently including the high probability of 0(1) occurring for repeated sampling of bernoulli experiments with probabilities close to 0(1). That is why it was described as the exact distribution for the estimate of proportions (7) from repeated sampling. It has an naturally skewed distribution reflecting the bounded distribution of the possible outcomes. For large n, CLT behaviour is observed.

In use for quantile variance estimation, in the percentile scale, the estimated quantile 2.5th & 97.5th points applicable to a 95% confidence interval are obtained by dividing the binomial cumulative distribution function by the sample size

$$F(k/n; n, \theta)_{binom} = \frac{1}{n} \sum_{i=0}^{\lfloor k \rfloor} \frac{n!}{k!(n-k)!} \theta^i (1-\theta)^{(n-i)} \tag{2}$$

where (i) θ is substituted for the proportion terms in the usual proportion formula as shown, and i are integers between 0 & n.

evd CI estimator

The empirical variance probability distribution, given in equation (3), is a rescaled normal distribution due to the quantile bounds (0,1). It is also only an estimator of the form $E[Var(\theta|x)]$ rather than a total variance estimator. This is because the quantile estimating function minimisation, that the quadratic polynomial proxy (1) mimics, is strictly a conditional variance of the observed sample, ie. if the observed quantile is 0 or 1, the calculated variance is zero.

$$f_{q-proxy}(b,\theta,n) = \left(\frac{1}{\int_0^1 \exp\{\frac{-(b-\theta)^2}{2\sigma^2_{CLT}}\}\ db} \right) \exp\{\frac{-(b-\theta)^2}{2\sigma^2_{CLT}}\} \tag{3}$$

where

$$\sigma_{CLT} = \sqrt{\frac{\theta(1-\theta)}{n}} \tag{4}$$

On backtransformation to the original measurement scale, the shape of the resulting empirical variance distribution (evd) using empirical cdf with interpolation, shares the step function character of the cumulative distribution function and in general exhibits some asymmetry compared to the convention of outputting symmetric bootstrap standard errors.

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 (CLT) normal distribution form

$$f_{q-proxy}(b,\theta,n)_{CLT} \rightarrow \frac{1}{\sigma_{CLT}\sqrt{2\pi}} \exp\{\frac{-(b-\theta)^2}{2\sigma^2_{CLT}}\} \tag{5}$$

In use for quantile variance estimation, the effect of rescaling in equation (3), in the percentile scale, means the estimated quantile 2.5th & 97.5th points applicable to a 95% confidence interval are obtained by using a normal cumulative distribution function bounded to the interval [0,1]

$$F(q)_{evd} = \frac{1}{2}[1 + erf(\frac{q-\theta}{\sigma_{CLT}\sqrt{2}})]\ \epsilon\ [0,1] \tag{6}$$

evd plus $Var(E[\theta|x])$ approximation

Using the variance expression for indicator random variables, the $Var(E[\theta|x])$ term is approximated, in the percentile scale, by

$$Var(E[\theta|k/q]) = E(\theta)^2 Var(k/q)_\theta \tag{7}$$

$$\approx \begin{cases} \theta^2 Pr(\theta = 0) Pr(\theta > 0) & for\, \theta \leq 0.5 \\ (1-\theta)^2 Pr(\theta = 1) Pr(\theta < 1) & for\, \theta > 0.5 \end{cases} \tag{8}$$

where var(k/q) term is a population variance, in the percentile scale, and so is an unconditional variance contribution to the total variance.

In use for sample quantile variance estimation, the total variance estimator is obtained as the squared sum of the $E[Var(\theta|x)]$ and $Var(E[\theta|x])$ standard errors according to equations (6) & (7)

As mentioned in (2), the advantage of the empirical variance distribution based 95% confidence interval estimators is that the quantile values of the intervals can be determined very simply in the percentile scale using only the sample size and given quantile value.

Quantile regression model coefficients confidence interval estimators

A major difference between estimating regression model slope(s) and intercept coefficient confidence intervals and sample confidence intervals, is that the sample quantile confidence intervals are defined directly as points located on the measured sample distribution (using bootstrap approach) or cdf (using the evd approach).

However, for regression modelling the slope(s) and intercept coefficient confidence intervals, are typically a scaled value of the regression residuals variance. For example, for ordinary least squares regression with two explanatory variables,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{9}$$

the model slope(s) and coefficients are derived from the residuals sample variance (distribution) s_{res} via the linear regression relationships

$$s_{\beta_1} = \frac{s_{res}}{var(x_1)(1 - r_{12})} \tag{10}$$

$$s_{\beta_2} = \frac{s_{res}}{var(x_2)(1 - r_{12})} \tag{11}$$

$$s_{\beta_0} = s_{res}\sqrt{1 + s_{\beta_1}^2 \bar{x}_1^2 + s_{\beta_2}^2 \bar{x}_2^2} \tag{12}$$

$$r_{12} = \frac{cov(x_1, x_2)}{\sqrt{var(x_1)var(x_2)}} \tag{13}$$

Since the evd approximation to quantile estimating function, produces a smooth, differentiable approximation to the quantile estimating function (1), the obvious model slope variance estimators to trial with evd based model slope coefficient CI estimators for quantile regression are exactly equations (10) & (11). As the results shown in this paper indicate, this choice of functional relationship works well for the given evd based model slope CI estimators.

For the model intercept coefficient variance estimator, two obvious adaptions of (12) for evd based approximations of quantile regression model intercept variance (as well as others) were trialled.

$$s_{\beta_0}(\theta) = s_{res}\sqrt{1 + s_{\beta_1}^2 (x_1(\theta))^2 + s_{\beta_2}^2 (x_2(\theta))^2} \tag{14}$$

$$s_{\beta_0}(\theta) = s_{res}\sqrt{1 + s_{\beta_1}^2 median(x_1)^2 + s_{\beta_2}^2 median(x_2)^2} \tag{15}$$

where θ is the quantile value under quantile regression modelling. Examining the evd approximation CI results of a quadratic polynomial quantile regression example, very strongly indicates that equation (15) gives superior and symmetric coverage performance compared to equation (14) over the whole quantile range (0,1). That only the median sample estimate needs to involved, rather than individual θ sample estimates for the intercept CI, is probably a reflection of the equal slope observed for different quantile values in quantile regression results for homoscedastic data.

In this research, equations (10), (11) & (15) have been found to form a suitable basis to produce accurate evd based approximations to quantile regression model coefficient CIs for homoscedastic iid cases, from quantile regression residuals.

Importantly, to keep the evd based variance estimation maximising the use of quantile regression calculations. The residuals distribution variance is not calculated using sum of squares of errors of the residuals but is calculated by performing quantile regression calculations on the residual distribution using evd based estimates of (θ_{LB}, θ_{UB}), to then estimate the 2.5th & 97.5th bounds in the original measurement scale.

To properly acknowledge the loss of degrees of freedom, produced by the quantile regression model and hence higher uncertainty in estimated CIs, equations (10), (11) & (15) also will contain implicitly or explicitly a $\sqrt{n/(n-p-1)}$ factor. This correction can be done (i) in the percentile scale on equations (2) & (6) so the adjustment is implicitly in equations (10), (11) & (15), or (ii) equations (10), (11) & (15) can have the degree of freedom adjustment explicitly included. Both versions of this adjustment have been trialled for evd based CI estimators.

assessing/adapting evd based estimators for use as quantile regression model coefficient CI estimators

As a result of investigations of the coverage performance of the above three sample CI estimators equations (2), (6) & (7), as quantile regression model coefficient CI estimators, some shortcomings were identified with the direct use of evd based sample quantile estimators as quantile regression model CI estimators

1. the evd plus $Var(E[\theta|x])$ total variance approximation produced overcoverage due to the $Var(E[\theta|x])$ term
2. the use of extrapolated 0th, 100th quantile bounds in (2) produced overcoverage
3. the asymmetry of the residuals distribution CI when using raw evd based estimators in the original scale led to undercoverage
4. the use of mean evd standard errors calculated from the average of the asymmetric residuals distribution CIs also led to undercoverage. Noting that in (1) there appeared to evidence of reasonable agreement between bootstrap sample quantile CIs and mean evd based sample quantile CIs.

These observations directly led to the most likely candidates for evd based quantile regression model CIs, being to take the maximum half CI of the asymmetric evd CIs (using equations (2) or (6)) of the residuals distribution and use that as a symmetric confidence interval estimate. This approach was expected to cause some overcoverage but that is much more palatable than undercoverage.

Some evidence of the need for symmetrisation for quantile regression model coefficient CIs is that the repeated sampling produced symmetrical distributions for the point estimates of the slope(s) and intercept obtained from the quantile regression results. Whereas in (1,2), under repeated sampling, there was a clear asymmetry in the histogram about a given sample quantile, for non-linear sample distributions.

evd approximations to the regression residuals variance/CI

Given the above findings, the field of useful evd based estimators for quantile regression model coefficient CI estimators was thus reduced to the following short list of estimators of the residuals CI

1. bin_max_perc; binomial distribution based evd CI estimator with the regression degrees of freedom adjustment in the percentile frame

determine 2.5th & 97.5th percentile bounds of residuals using

qbinom(0.025,(n-p-1),q)/(n-p-1)

qbinom(0.975,(n-p-1),q)/(n-p-1)

then calculate the maximum half CI in the original scale of the residuals distribution using quantile regression

max_evd_perc = max(abs(bin_UB-point estimate),abs(bin_LB-point estimate))

2. bin_max_org; binomial distribution based evd CI estimator with the regression degrees of freedom adjustment in the original scale

determine 2.5th & 97.5th percentile bounds of residuals using

qbinom(0.025,n,q)/n

qbinom(0.975,n,q)/n

then calculate the maximum half CI in the original scale of the residuals distribution using quantile regression

max_bin_org = max(abs(bin_UB-point estimate),abs(bin_LB-point estimate))

then use $\sqrt{(n/n-p-1)}$ factor to inflate max_bin_org

3. evd_max_perc; evd distribution with regression degrees of freedom adjustment conducted in the percentile frame

determine variance in percentile scale of residuals distribution using

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

determine 2.5th & 97.5th percentile bounds of residuals using

max(qnorm(0.025,q,σ_{CLT}(n-p-1)),0)

min(qnorm(0.975,q,σ_{CLT}(n-p-1)),1)

then calculate the maximum half CI in the original scale of the residuals distribution using quantile regression

max_evd_perc = max(abs(evd_UB-point estimate),abs(evd_LB-point estimate))

4. max_evd_org; evd distribution with regression degrees of freedom adjustment conducted later in the original scale

determine variance in percentile scale of residuals distribution using

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

determine 2.5th & 97.5th percentile bounds of residuals using

max(qnorm(0.025,q,σ_{CLT}(n)),0)

min(qnorm(0.975,q,σ_{CLT}(n)),1)

then calculate the maximum half CI in the original scale of the residuals distribution using quantile regression

max_evd_org = max(abs(evd_UB-point estimate),abs(evd_LB-point estimate))

then use $\sqrt{(n/n-p-1)}$ factor to inflate max_evd_org

estimating model slope(s) and intercept CIs

The output of the evd based estimators of the residuals variance/CI were then used to estimate the model slope(s) and intercepts using equations (10), (11) & (15). The coverage performance was then assessed and compared to the known population regression model and default quantreg bootstrap estimates using 600 replicates as shown in the next section.

Assessing coverage performance for different homoscedastic datasets

In the following figures, the coverage performance of evd based model coefficients CI estimators (based on 1000 resamples) for five quantile regression cases. Three samples size n=50,100,1000 were trialled.

(i) y = 1 * runif(n,-10,10) + rnorm(n,0,10);

A simple linear relationship between the dependent variable (Y) and explanatory variable (runif(n,-10,10)) where the quantile regression for different quantiles results in a set of parallel regression lines.

(ii) y = 1 * runif(n,-10,10)^2 + rnorm(n,0,10);

Where the quantile regression is performed as quadratic polynomial fit but results in a set of diverging curved regression lines. This model fit provided a sensitive test of the evd based model intercept CI estimator.

(iii) y = 1 * runif(n,-10,10) + urlaplace(n,0,10);

A second simple linear relationship between the dependent variable (Y) and explanatory variable (runif(n,-10,10)) but with different error distribution. To illustrate that quantile regression and hence any useful evd based model coefficient CI estimators are robust to different types of homoscedsatic iid error distributions.

(iv) y = 1 * runif(n,-10,10) + 0.5 * runif(n,-10,10)^2 + urlaplace(n,0,10);

A multiple explanatory variables quantile regression with nonlinear relationship to more fully test the evd based model coefficient CI estimators.

(v) y = 1 * runif(n,-10,10) + AR(1); with a non-iid scale factor of 0.5

A homoscedastic non-iid regression example, to see if only the evd based model intercept coefficient CIs but not the slope coefficients fails to produce nominal 95% coverage. With a non-iid scale factor of 0.5, the effective sample size of variance of the error term is $\sqrt{n/2}$ rather than $\sqrt{n}$.

In the figures below, the coverage is calculated for 15 quantile points (0.025, 0.05 ,0.1 ,0.2 ,0.25 ,0.3 ,0.4 ,0.5 ,0.6 ,0.7 ,0.75 ,0.8 ,0.9 ,0.95 ,0.975). The black points and lines indicate the default quantreg bootstrap estimates (R=600). The blue lines indicate the evd_max CI estimators, estimator 3 are included on each subfigure x(a), x(b), x(c), estimator 4 are included on each subfigure x(d), x(e), x(f). The red lines indicate the bin_max CI estimators, estimator 1 are included on each subfigure x(a), x(b), x(c), estimator 2 are included on each subfigure x(d), x(e), x(f). This overlap of estimators and subfigures allows a visual comparison of the effects of sample size, degrees of freedom adjustment placement and estimator type.

Figures 1 & 2 display the coverage performance of bootstrap, bin_max & evd_max for the slope and intercept respectively of model (i).

Figures 3 & 4 display the coverage performance of bootstrap, bin_max & evd_max for the slope and intercept respectively of model (ii).

Figures 5 & 6 display the coverage performance of bootstrap, bin_max & evd_max for the slope and intercept respectively of model (iii).

Figures 7, 8 & 9 display the coverage performance of bootstrap, bin_max & evd_max for the two slopes and intercept respectively of model (iv).

Figures 10 & 11 display the coverage performance of bootstrap, bin_max & evd_max for the slope and intercept respectively of model (v).

Generally, the figures for the homoscedastic iid cases (i)-(iv) are similar in their characteristics.

1. for n=1000, nominal 95% coverage for bootstrap, evd_max, bin_max estimators for all quantiles (0.025-0.975)
2. for n=100, lower slope CI coverage for quantiles (0.025,0.05,0.95,0.975)
3. for n=50, lower slope CI coverage for quantiles (<.1,>.9) for all three CI estimators
4. bin_max has moderate slope CI undercoverage for quantile ranges (.1-.3), (.7-.9)
5. evd_max has slope CI overcoverage for quantile ranges (.1-.3), (.7-.9)
6. for nonlinear model cases (ii) & (iv), the intercept CI has 93% coverage for evd_max & bin_max. The default bootstrap estimates are nominal 95% coverage.

For the non-iid case (v), figure 11 shows undercoverage of the intercept CI estimators for default bootstrap, evd_max, bin_max estimators

model (i): slope CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

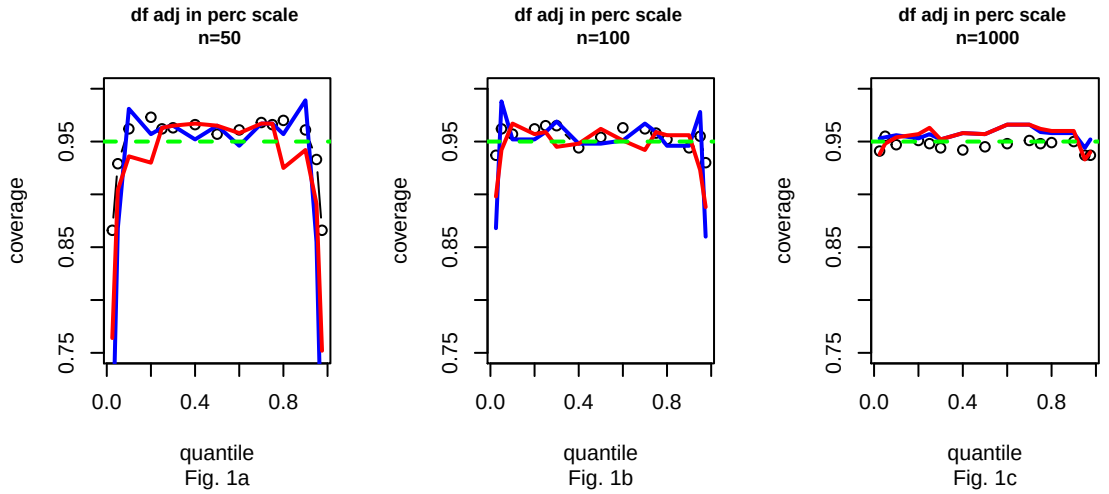

Fig. 1a Fig. 1b Fig. 1c

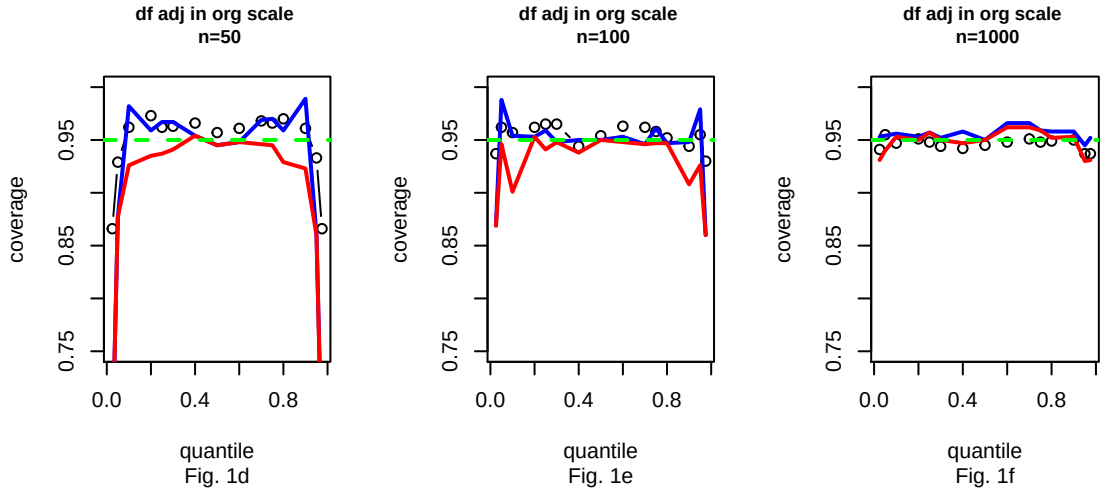

Fig. 1d Fig. 1e Fig. 1f

model (i): intercept CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

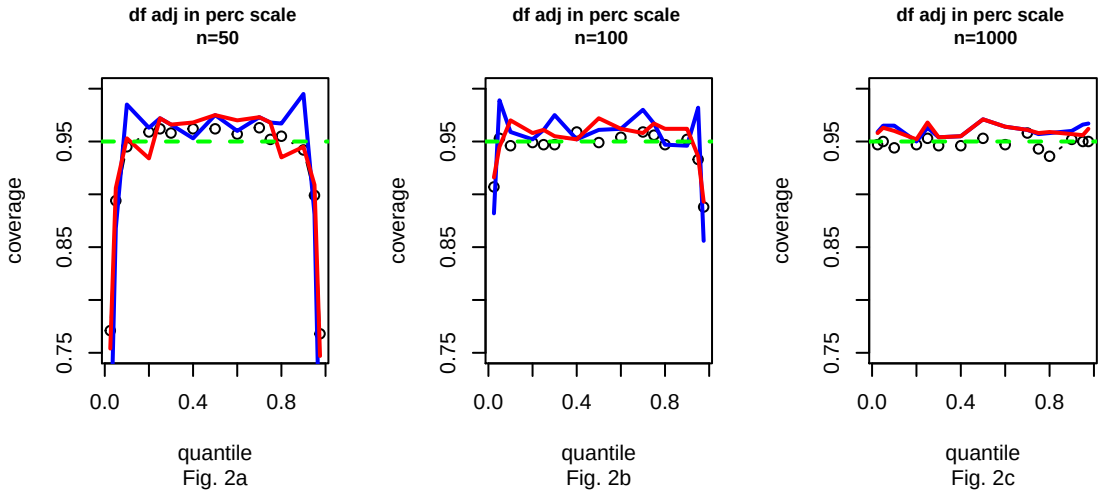

Fig. 2a

Fig. 2b

Fig. 2c

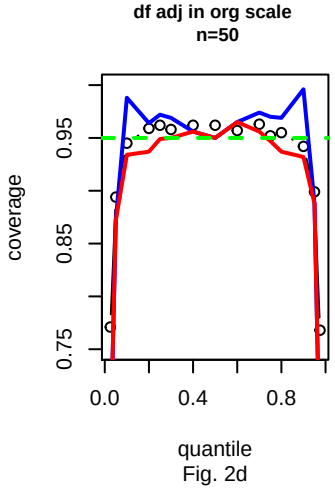

Fig. 2d

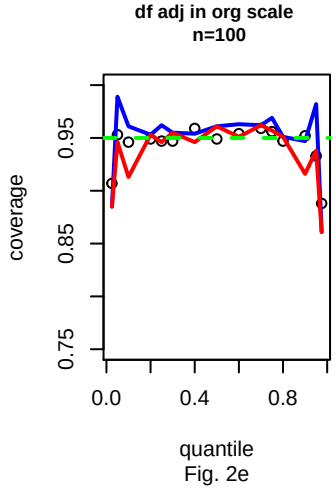

Fig. 2e

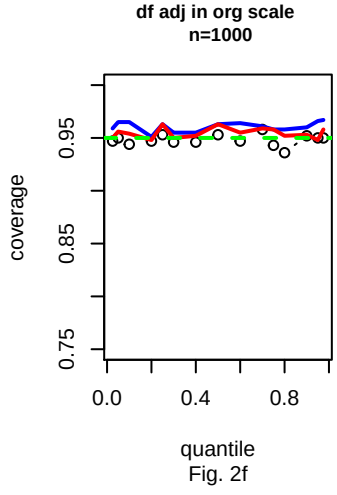

Fig. 2f

model (ii): slope CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

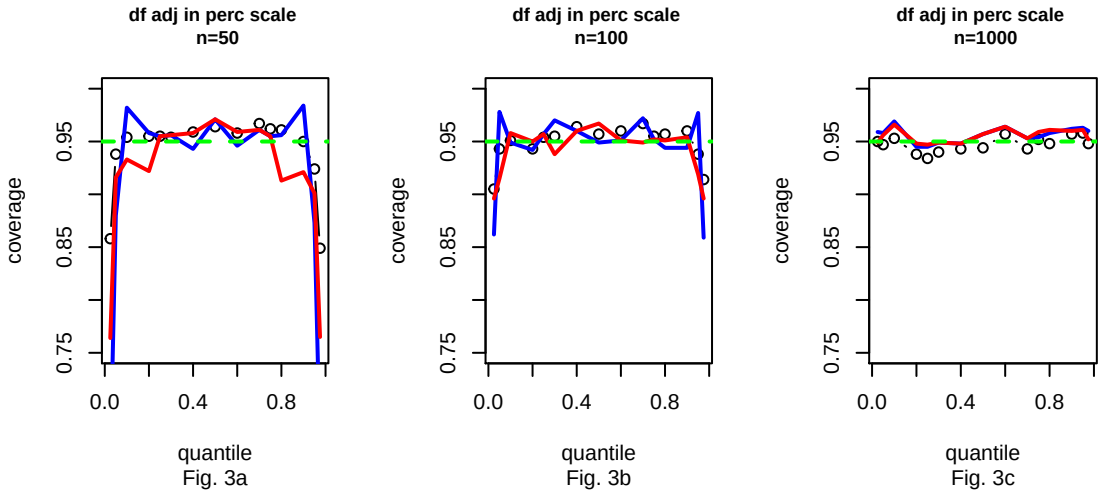

Fig. 3a

Fig. 3b

Fig. 3c

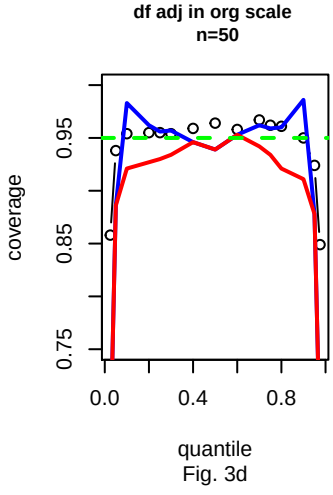

Fig. 3d

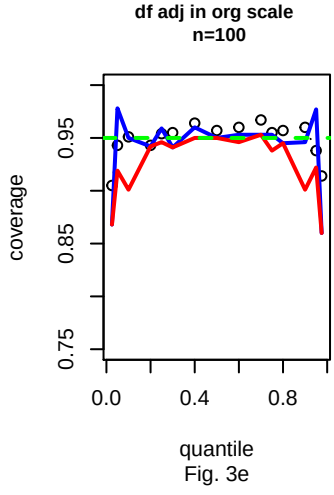

Fig. 3e

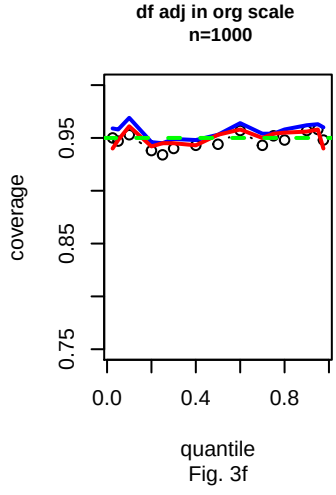

Fig. 3f

model (ii): intercept CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

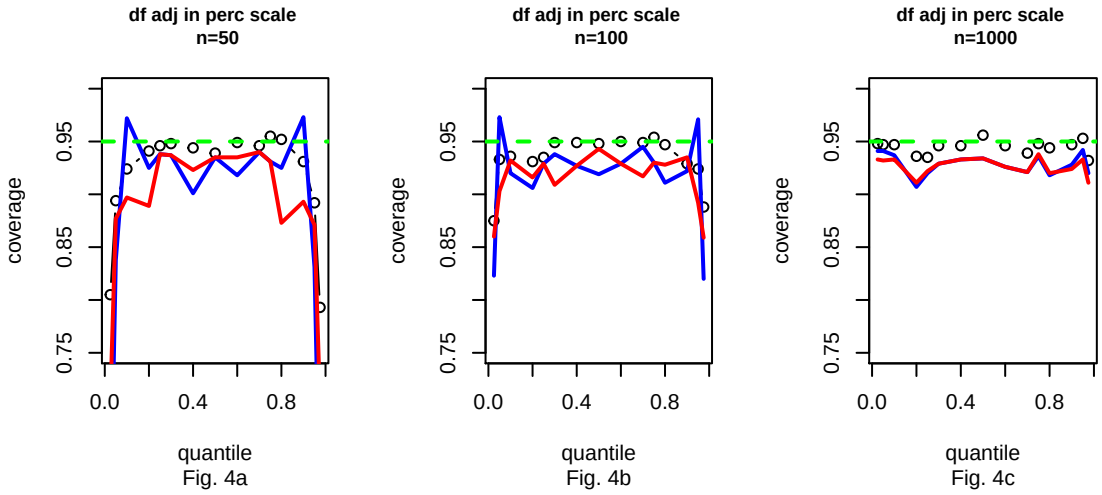

Fig. 4a

Fig. 4b

Fig. 4c

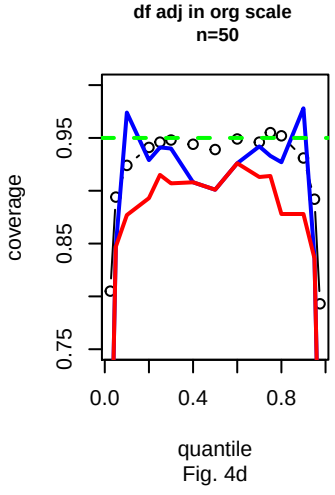

Fig. 4d

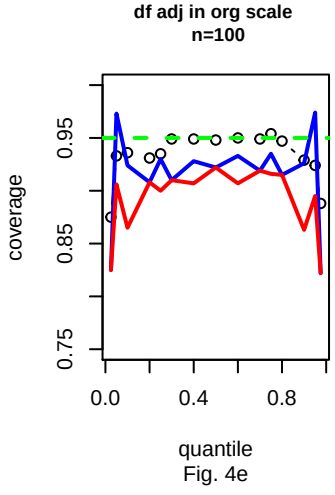

Fig. 4e

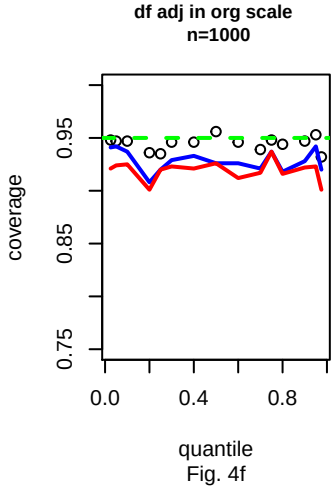

Fig. 4f

model (iii): slope CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

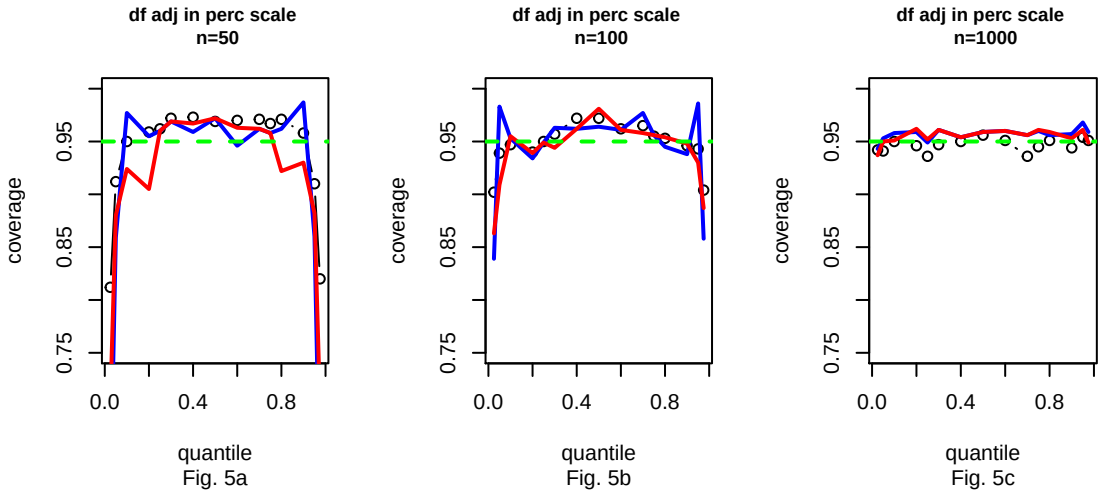

Fig. 5a

Fig. 5b

Fig. 5c

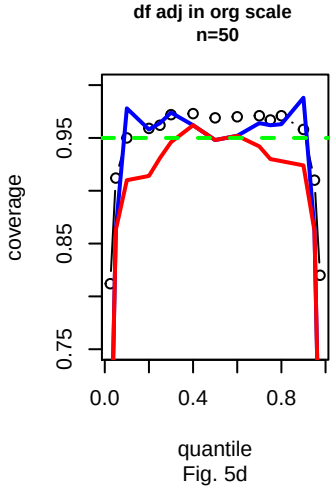

Fig. 5d

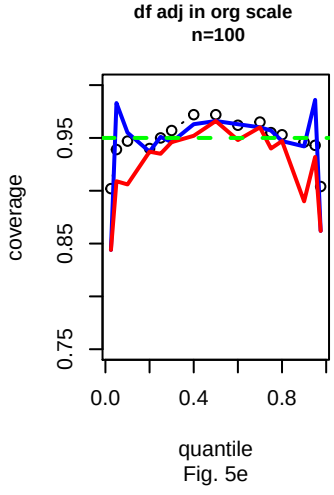

Fig. 5e

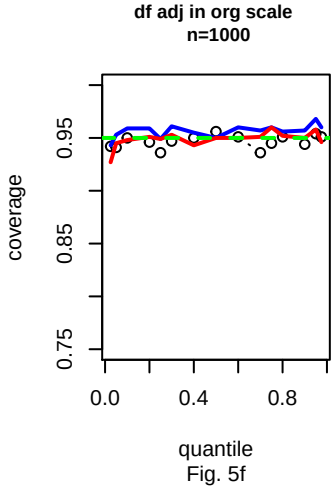

Fig. 5f

model (iii): intercept CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

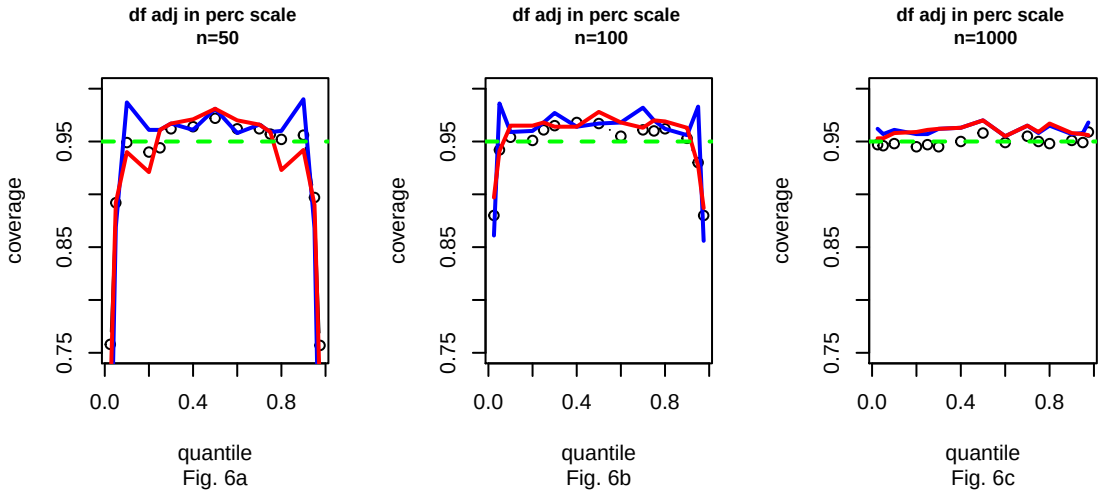

Fig. 6a

Fig. 6b

Fig. 6c

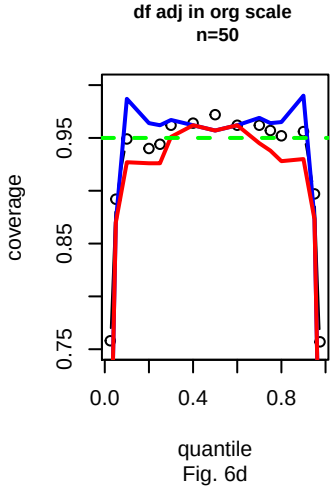

Fig. 6d

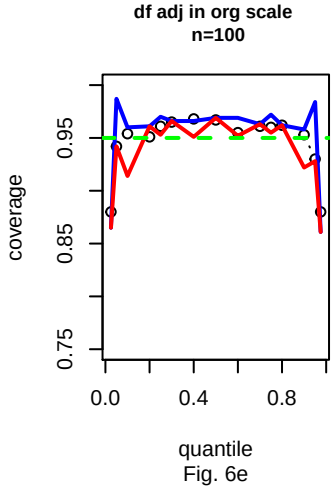

Fig. 6e

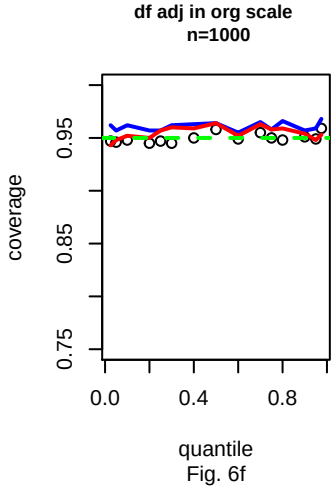

Fig. 6f

model (iv): linear term slope CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

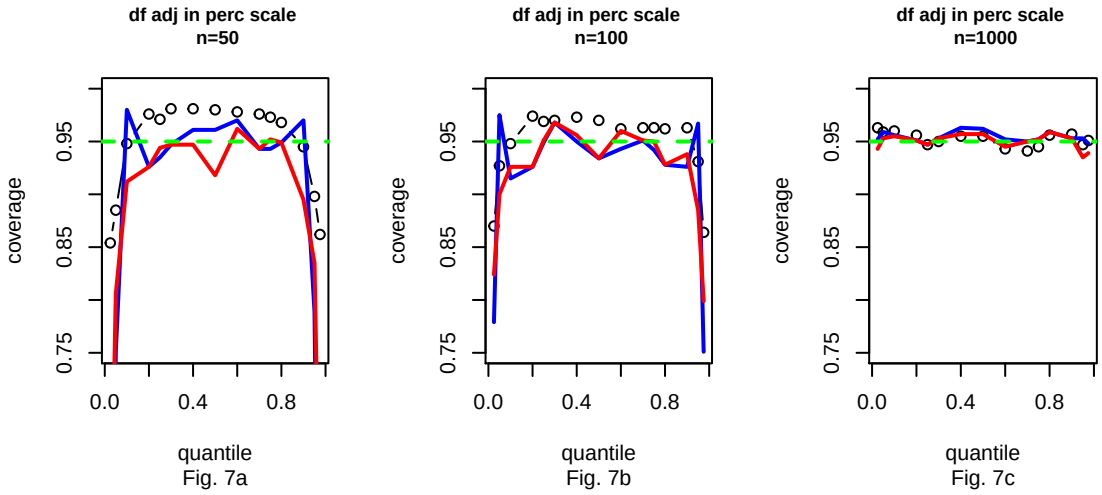

Fig. 7a

Fig. 7b

Fig. 7c

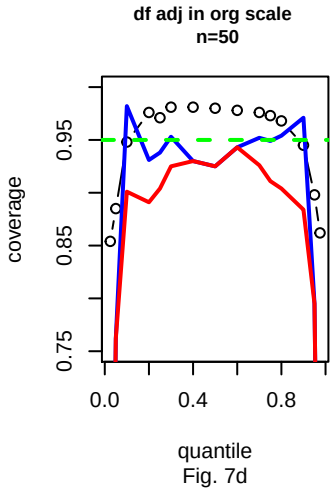

Fig. 7d

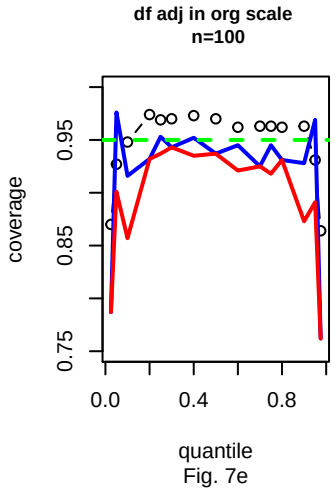

Fig. 7e

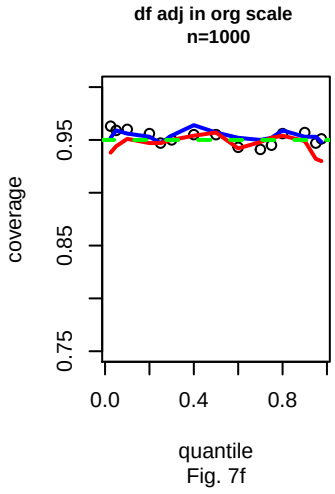

Fig. 7f

model (iv): quadratic term slope CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

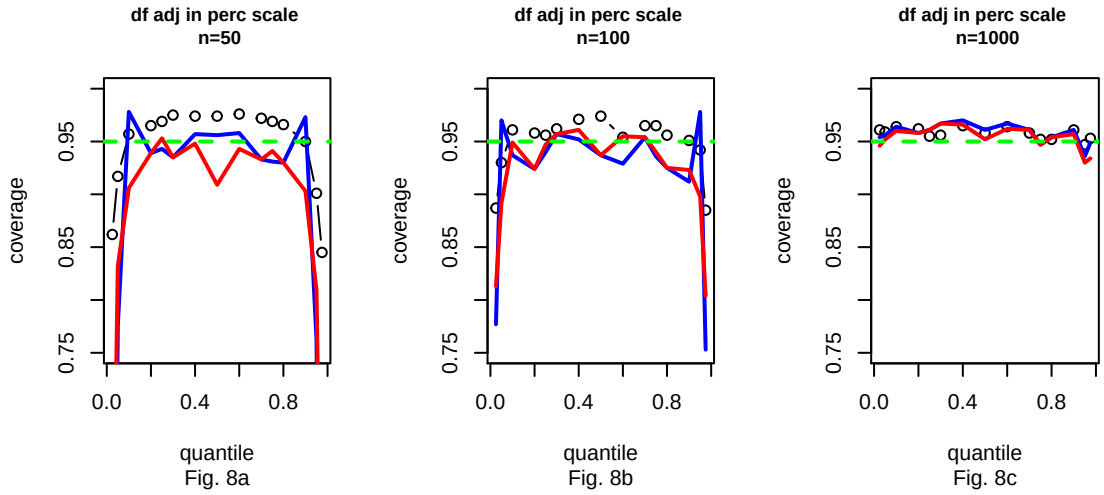

Fig. 8a

Fig. 8b

Fig. 8c

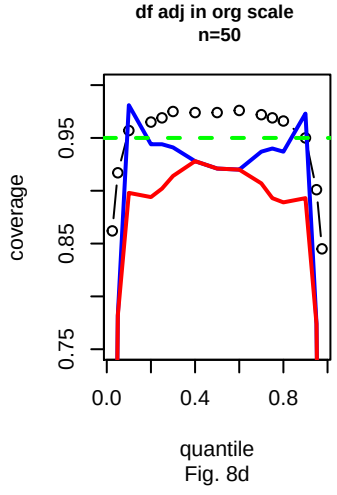

Fig. 8d

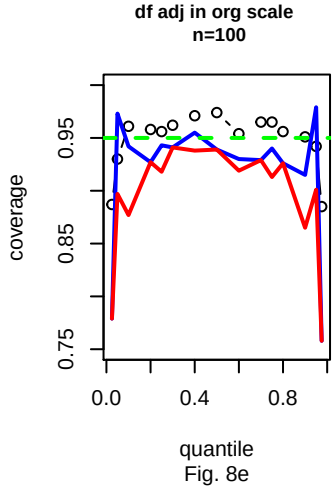

Fig. 8e

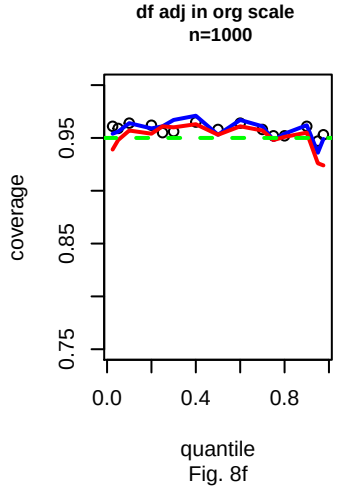

Fig. 8f

model (iv): intercept CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

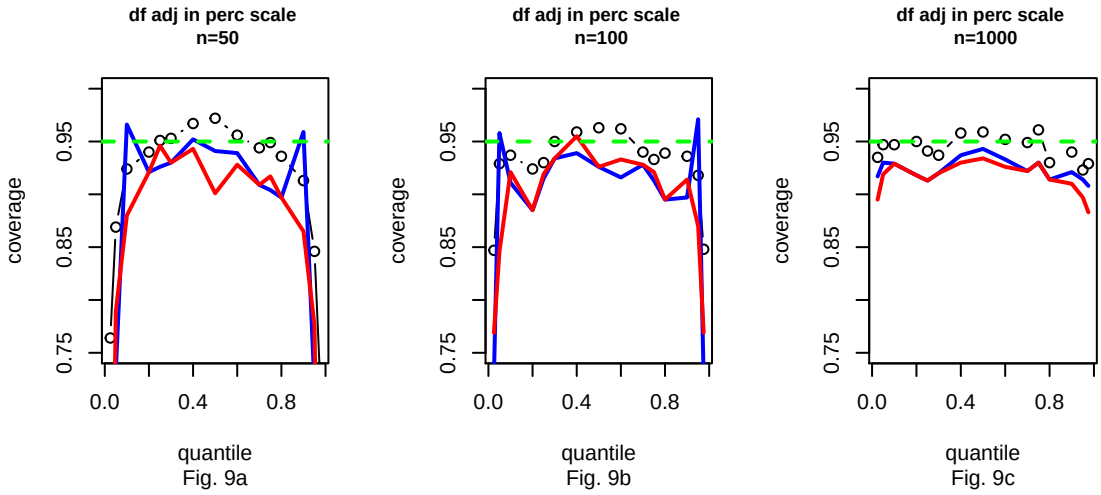

Fig. 9a

Fig. 9b

Fig. 9c

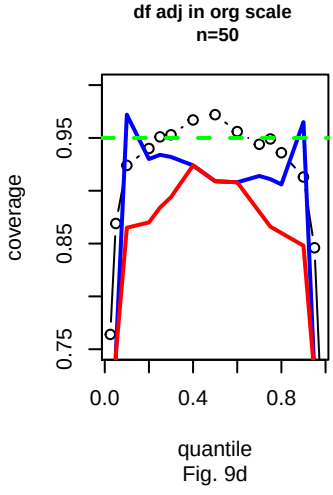

Fig. 9d

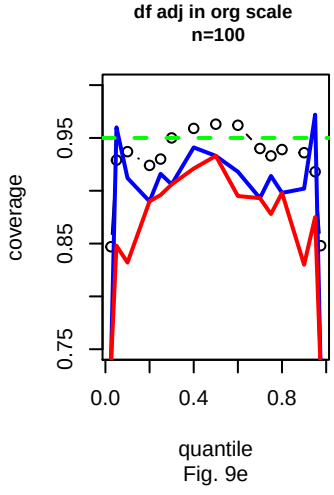

Fig. 9e

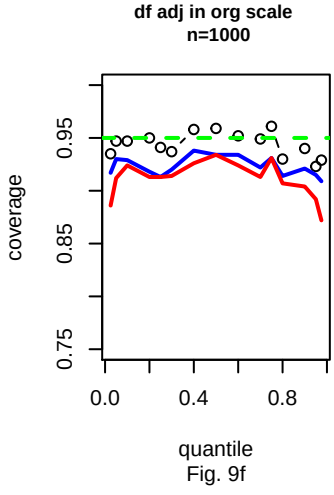

Fig. 9f

model (v): slope CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

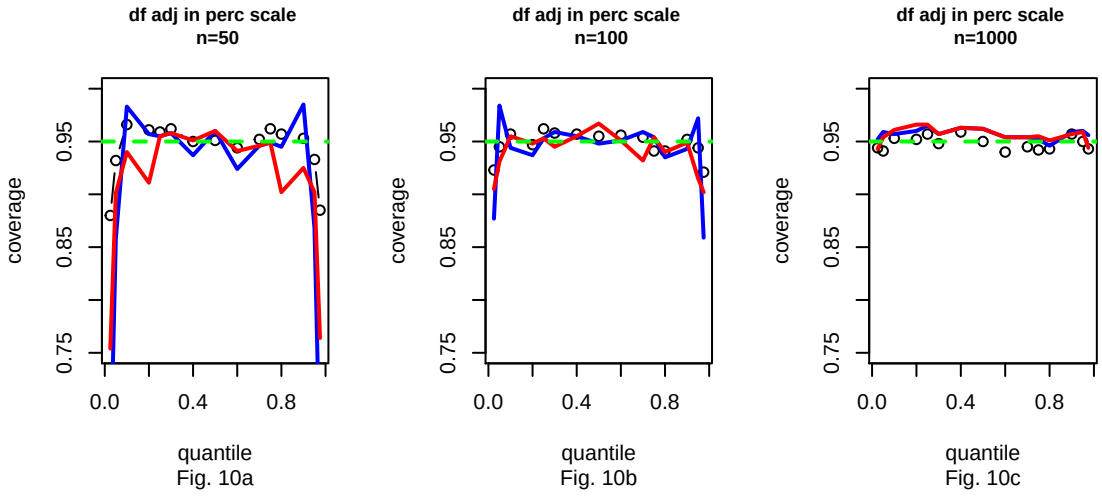

Fig. 10a

Fig. 10b

Fig. 10c

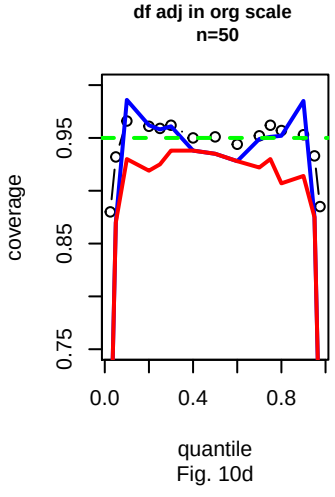

Fig. 10d

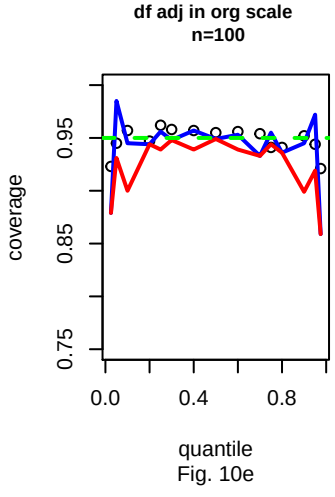

Fig. 10e

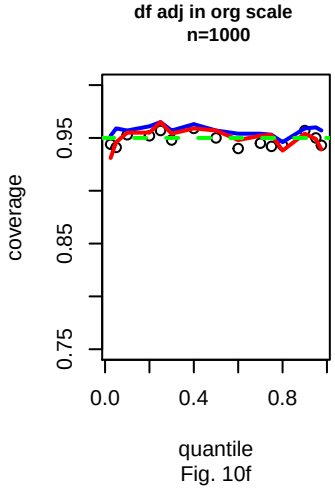

Fig. 10f

model (v): intercept CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

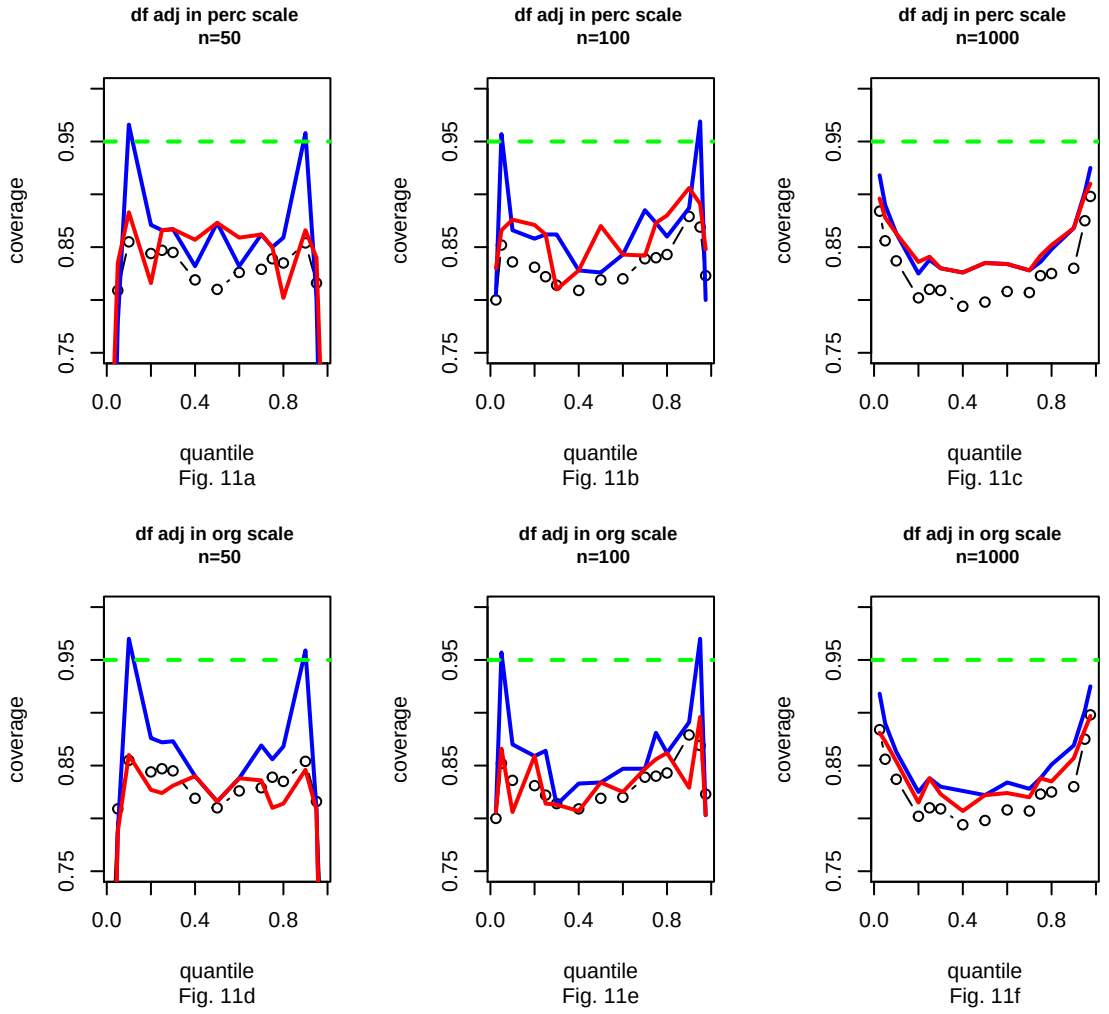

Fig. 11a Fig. 11b Fig. 11c

Fig. 11d Fig. 11e Fig. 11f

Comparing coverage performance for a small sample "real world" example

To round off the CI performance comparison, appendix A, contains a comparison of the evd based model coefficient CI estimator (evd_symm_max using degrees of freedom adjustment in percentile scale) to the quantreg default bootstrap and rank inversion CI estimates for the Engel food expenditure dataset [8]. Providing this example also allows, the r markdown version of this paper [9], to conveniently contain a concise version of the evd based model coefficient CI estimator algorithm rather than the lengthy repeated sampling version.

This Engel dataset when analysed in log-log form provides an interesting example of near homoscedastic iid performance to underline the difference in performance expected for CI estimators between "ideal" error distributions and real world examples.

Conclusions

For homoscedastic iid unweighted cases, both the evd_max and bin_max evd based approximations to quantile regression model coefficient CIs show good performance for model slope CIs coverage for moderate samples n=1000. For smaller samples, the coverage performance is weaker for extreme quantiles.

For model intercept CI estimates, the evd_max and bin_max evd based approximations have good coverage for linear cases but slight undercoverage (93%) for nonlinear examples, with the same sample size dependence as for model slope CI estimates.

It would be worthwhile investigating, (i) minor improvements to equation (15) for model intercept CI estimates and (ii) heteroscedastic consistent standard error extensions of the current approach.

References

1. Martin J.P.D., 2015, http://dx.doi.org/10.6084/m9.figshare.1566828
2. Martin J.P.D., 2015, http://dx.doi.org/10.6084/m9.figshare.1591019
3. Koencker, R. W. & Bassett G., Econometrica, 1978, vol. 46, issue 1, pages 33-50
4. Koencker, R. W., Portnoy S. et al, https://cran.r-project.org/web/packages/quantreg/quantreg.pdf
5. https://en.wikipedia.org/wiki/Quantile
6. Brown, B. M. and Wang, Y.-G. (2005). Standard errors and covariance matrices for smoothed rank estimators. Biometrika 92 149-158. MR2158616
7. Agresti, A. and Coull, B. A. 1998 The American Statistician, vol. 52, p119-126. doi:10.2307/2685469. JSTOR 2685469
8. https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf
9. https://github.com/johnpdmartin/sampling-investigations/blob/master/quantile_regression_model_coefficient_CIs_using_the_empirical_variance_distribution_approximation.Rmd

Appendix A: Comparison of model coefficient CIs for Engel dataset

The Engel dataset is a small sample example of the heteroscedastic relationship between food expenditure and household income and is included in the quantreg package as an quantile regression example.

In log linear form, by taking the logarithms of both food expenditure and household income prior to quantile regression analysis, the error distribution is almost homoscedastic in nature, as shown in Figure A1.

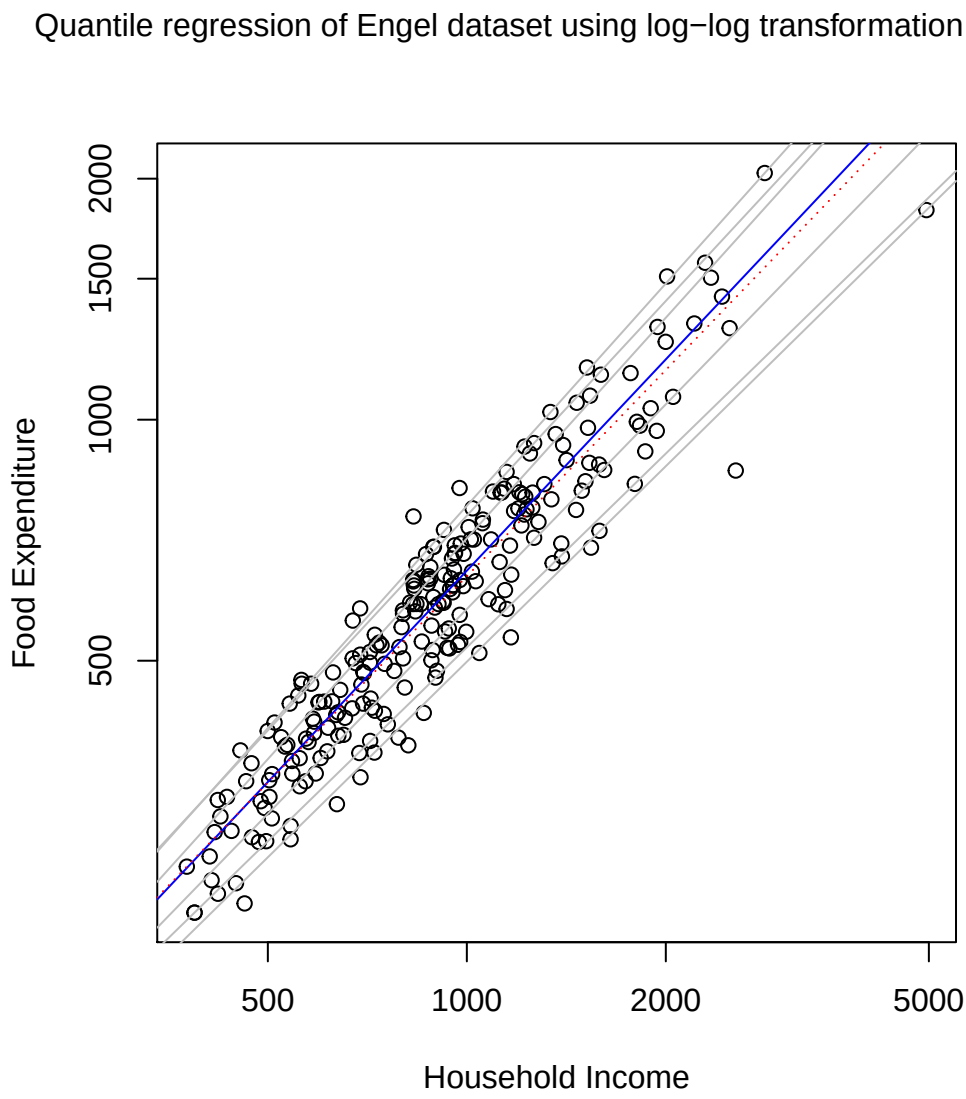

Figure A1: Engel dataset in log-linear form

Figure A2, shows the model coefficents CI estimates of the intercept and slope, using

1. rank inversion
2. quantreg package default bootstrap (delete-d-group jackknife)
3. evd_symm_max with the degrees of freedom calculated in the percentile scale

where the outline of the bootstrap CIs is given in all the sub-graphs.

The rank inversion method was replaced as the quantreg default quantile regression CI estimator by the bootstrap method. As seen in figure A2, for this quasi-iid homoscedastic case, the bootstrap and evd_symm_max estimates are similar for many quantiles except for minor differences for regions near 0.4-0.5 and 0.6-0.8. The difference occur in both the slope and intercept CI estimates suggesting a breakdown of the iid conditions for the evd_symm_max method. The older rank inversion method exhibits smaller CIs compared to the other two methods for quasi-iid homoscedastic conditions.

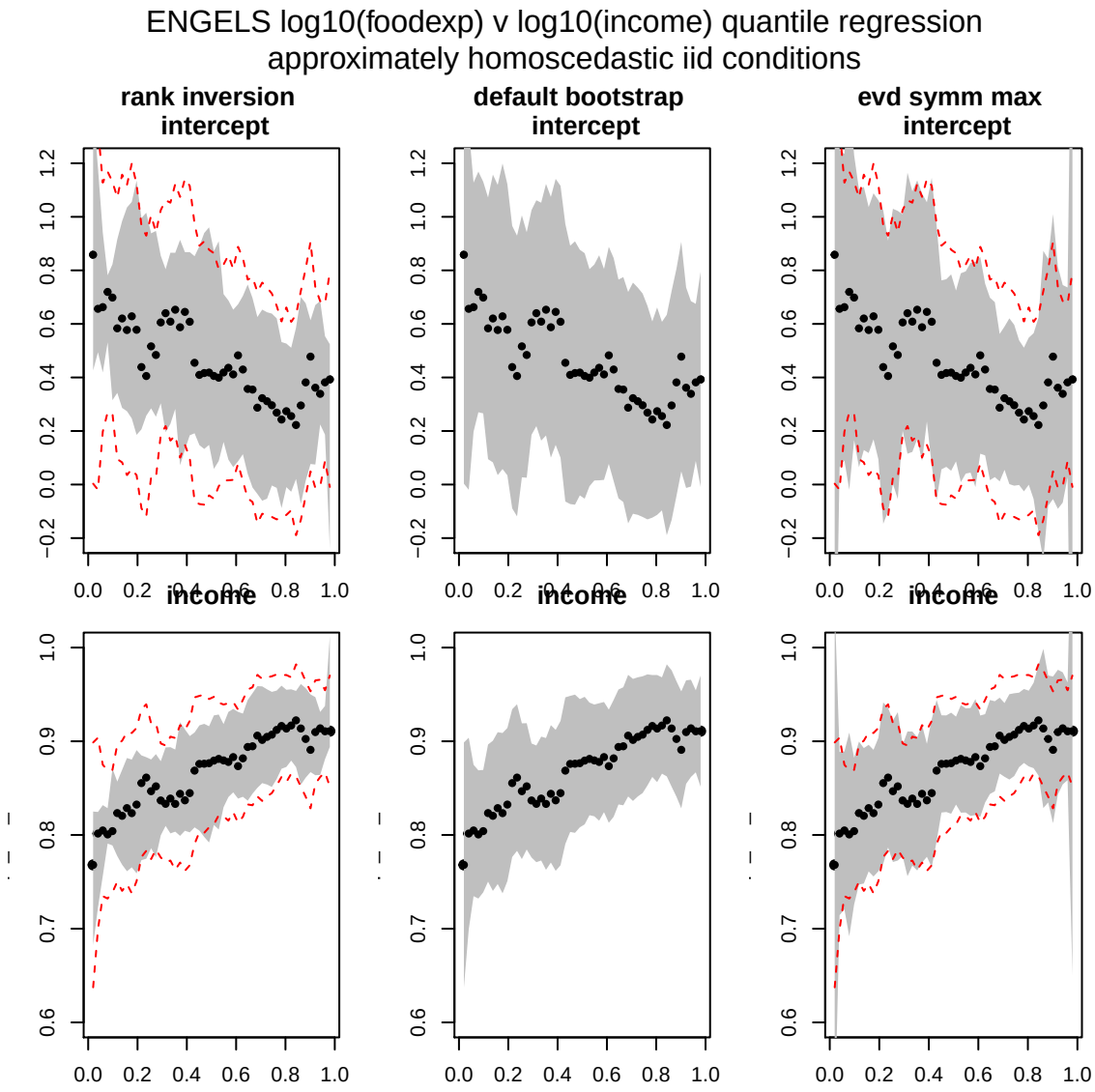

Figure A2: Estimated intercept and slope model coefficient CIs for log-log transformed Engel dataset

For completeness, figure A3 gives the comparative performance of the three CI estimators for the original heteroscedastic scattered dataset. It can be seen, the delete-d-group jackknife CI estimates are larger than both the evd_symm_max and rank inversion CI estimates. So the evd based approach will need further development for non-iid cases.

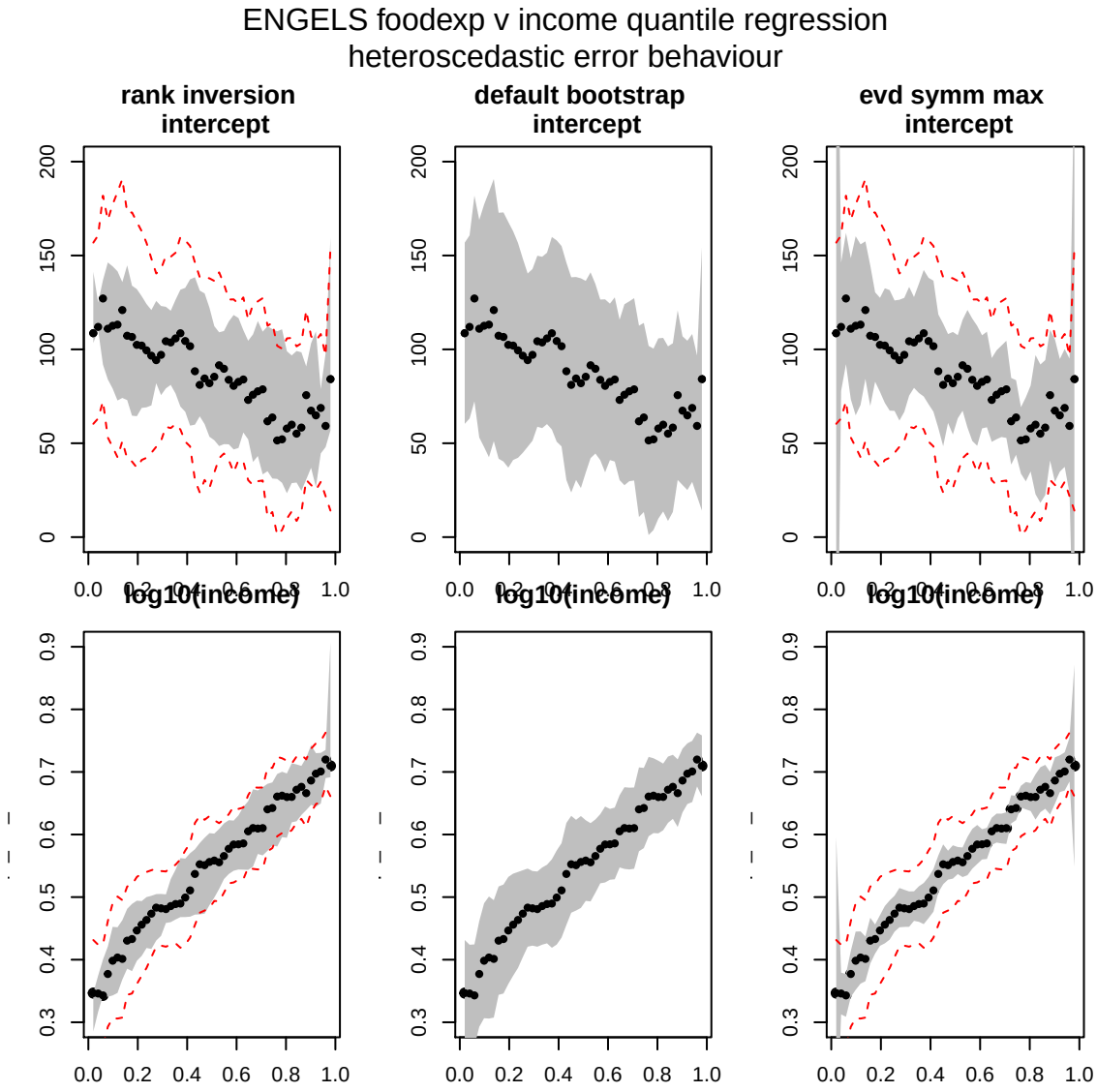

Figure A3: Estimated model coefficient CIs for non-transformed Engel dataset