Multivariate, quantile regression model coefficient CIs using the empirical variance distribution, in the presence of collinearity

John P. D. Martin

Monday, May 2, 2016

Executive Summary

This paper investigates the full covariance matrix intercept CI estimator required for multivariate performance of the empirical variance distribution (evd) approach (Martin (1-4)), to estimate the confidence interval (CI) bounds of the quantile regression model coefficient estimates (Koencker & Bassett (5)), for unweighted data cases.

Analysis of the coverage accuracy of the multivariate quantile regression results is conducted by repeated sampling of several sample sizes to known regression models and error distributions. Importantly, the covariate correlation adjustment for model coefficient CIs for a given dataset will depend on the observed data range and the best coverage performance for the intercept model coefficient CI is found using the full covariance matrix expression of the covariate correlation.

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 coverage estimates are compared to default bootstrap quantile coverage calculated using 600 replicates (from the quantreg r package (6)). Also included in Table 1 is the variance inflation (adjustment) factor (VIF) $\frac{1}{(1-r_{12}^2)}$ affecting the slope model coefficient CI width, due to collinearity between the explanatory variables. In this paper, this VIF factor is shown to also be present in the full covariance matrix expression for the evd based intercept model coefficient CI but its effect is evenly balanced by a $cov(x_1, x_2)$ term in the numerator of the intercept CI estimator.

Table 1: Calculated VIF and Quantile regression confidence interval estimator coverage of sample size 1000 for (slope1, slope2, intercept) respectively, based on 1000 repeated samples

Regression $\mathbf{Y} = \beta\mathbf{X} + \varepsilon$	quantile	bootstrap coverage	evd_max coverage	bin_max coverage
x ~ (N(90,90))		slope CI VIF = 3		
$y = 1 * x + 1 * x^2 + rnorm(0, 10)$	0.1	(0.949,0.953,0.936)	(0.967,0.966,0.950)	(0.958,0.965,0.946)
“”	0.5	(0.956,0.953,0.952)	(0.962,0.961,0.967)	(0.958,0.956,0.963)
“”	0.9	(0.940,0.947,0.957)	(0.948,0.960,0.969)	(0.944,0.956,0.972)
x_1 ~ (N(90,90))				
x_2 ~ (Laplacian(-90,90))		slope CI VIF = 3		
$y = (x_1 + 0.5 * x_2)+$				
$(x_2 + 0.5 * x_1) + rnorm(0, 10)$	0.1	(0.943,0.954,0.944)	(0.959,0.961,0.947)	(0.953,0.963,0.938)
“”	0.5	(0.952,0.946,0.948)	(0.966,0.956,0.951)	(0.958,0.953,0.951)
“”	0.9	(0.947,0.952,0.957)	(0.943,0.954,0.952)	(0.941,0.951,0.951)
x_1 ~ exp(N(0,2))-1				
x_2 ~ Uniform(-30,60)		slope CI VIF = 1.36		
$y = (x_1) + (x_2 + x_1) + rnorm(0, 10)$	0.1	(0.966,0.951,0.949)	(0.952,0.963,0.949)	(0.950,0.961,0.958)
“”	0.5	(0.960,0.943,0.958)	(0.946,0.950,0.972)	(0.941,0.945,0.965)
“”	0.9	(0.956,0.948,0.931)	(0.946,0.960,0.944)	(0.948,0.954,0.945)

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.

It can be seen that for this modest sample size and significant VIFs ~ 1.3-3, the bootstrap estimator and the evd based confidence interval estimators exhibit nominal 95% coverage for slope and intercept model coefficient CI estimates in the presence of homoscedastic, independent errors and collinearity.

The first example illustrates polynomial quantile regression, with the data range constrained to mainly positive values between (-90,270), and so there is significant covariate correlation between x & x^2. The second example has linear bivariate quantile regression, with two variables each mostly positive(negative) with 33% mixing prior to quantile regression modelling, also resulting in significant VIF. The third example, uses asymmetric/symmetric variables with unequal mixing of the variables prior to quantile regression. In all cases, there is nominal 95% coverage performance.

For extreme quantiles in smaller samples the evd based model coefficient CIs coverage performance is lower.

Introduction

Quantiles (7) 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] ≤ k/q or, equivalently, Pr[X ≥ x] ≥ (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 (5) 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 (8). 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 Martin (3), the evd based sample quantile CI estimators (1,2) were trialled, assessed and adapted where required as quantile regression model coefficient CI estimators. Since the evd approximation to quantile estimating function, produces a smooth, differentiable approximation to the quantile estimating function (1), the evd based CI estimators trialled for homoscedastic iid cases were analogous to the ordinary least squares regression estimators where the median has replaced the mean estimate in the variance intercept formula.

In Martin (4), the evd based CI estimators trialled for homoscedastic iid cases were adapted successfully to the case of (univariate) linear expanding horn heteroscedasticity. This extension was achieved firstly by using auxiliary regression of the quantile regression residuals and the evd approach to estimate seperate homoscedastic variance contributions to the slope and intercept coefficient CI. Then weighted medians, using auxiliary regression model, were used to replace the ordinary unweighted medians in the intercept coefficient CI formula, since using bootstrap estimates the minimum intercept variance occurred approximately when the intercept was close to the weighted median. In that paper, the evd based results of (3,4) were pointed out to have good performance for univariate quantile regression but that more research was needed for the multivariate case particularly on the quantile regression covariate correlation analogue to r_{12}. However, if r_{12} was known to be small then some bivariate quantile regression cases also displayed nominal 95% coverage performance.

In this paper, the coverage properties of the full covariance matrix evd approximation, where the quantile model coefficients β_{qr}

$$\mathbf{Y} = \beta_{\mathbf{qr}}\mathbf{X} + \varepsilon \tag{2}$$

are obtained from the quantile estimation function (5), and the variance estimate is approximated from (i) the empirical variance distribution (evd) estimate of the quantile regression residuals variance and (ii) the (linear) regression covariance matrix of the covariates

$$COV(\beta_{qr}) \approx (\mathbf{X}'\mathbf{X})^{-1} s_{res}^2 \tag{3}$$

is examined for bivariate quantile regression cases under repeated sampling for known population regression distributions and compared to boostrap estimates.

Covariance matrix terms for evd based Quantile regression model coefficient CIs

In Martin (3,4), the slope model coefficient CI estimators for univariate cases are derived from the asymmetric empirical variance distribution (evd) approximations of the quantile regression residuals sample variance (distribution) s_{res} and var(x),

Univariate case

For quantile regression with one explanatory variable,

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{4}$$

the model slope CI estimates may be derived from empirical variance distribution (evd) approximations of the quantile regression residuals model variance (distribution) $s_{maxhalfCI}$ via the relationship

$$s_\beta = \frac{s_{maxhalfCI}}{\sqrt{var(x)}} \tag{5}$$

while the evd based intercept CI estimator is of the form

$$s_{\beta_0}(\theta) = \sqrt{s_{maxhalfCI}^2 + \frac{s_{meanhalfCI}^2}{var(x)}(0 - median(x))^2} \tag{6}$$

where the $s_{meanhalfCI}$ evd based residuals variance in the slope related term of (6), provides better coverage performance for the evd based intercept CI estimator based on recent (linear expanding heteroscedasticity) results (4).

$s_{meanhalfCI}$ is a symmetrised variance estimate using the mean of the asymmetric evd s_{res} CI.

$s_{maxhalfCI}$ is a symmetrised variance estimate using the maximum half CI of the asymmetric evd s_{res} CI.

Multivariate case

For quantile regression with two explanatory variables,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{7}$$

the model slope(s) CI estimates may be derived from empirical variance distribution (evd) approximations of the quantile regression residuals model variance (distribution) $s_{maxhalfCI}$ via the relationships

$$s_{\beta_1} = \frac{s_{maxhalfCI}}{\sqrt{var(x_1)(1 - r_{12}^2)}} \tag{8}$$

$$s_{\beta_2} = \frac{s_{maxhalfCI}}{\sqrt{var(x_2)(1 - r_{12}^2)}} \tag{9}$$

$$r_{12} = \frac{cov(x_1, x_2)}{\sqrt{var(x_1)var(x_2)}} \tag{10}$$

where $s_{maxhalfCI}$ is derived from the evd based maximum half CI estimate (3,4), in the presence of homoscedastic iid errors.

In Martin (3,4) a simplistic version of the intercept model coefficient CI estimator for bivariate cases was trialled which contained r_{12} in two slope related terms. In hindsight, the weaker performance of the Martin (3,4) bivariate intercept estimator can be attributed to an incomplete use of the covariance matrix eqn (3).

Complete use of the covariance matrix in the intercept model coefficient CI estimator, for the evd approximation for quantile regression (adapted from ordinary least squares linear regression), involves firstly the assumption

$$median(y) \approx \hat{\beta}_0 + \hat{\beta}_1 median(x_1) + \hat{\beta}_2 median(x_2) \tag{11}$$

where the equivalent result for ordinary least squares linear regression is exact. The impact of the approximation in eqn (11) is not significant since the actual estimate used and required is the variance estimator and $var(\hat{\beta}_0(\theta)) \approx var(\hat{\beta}_0(\theta + \delta\theta))$. Given eqn (11), a good approximation of the variance of $\hat{\beta}_0$ is

$$\begin{aligned} var(\beta_0) \approx&\ var(median(y) - \hat{\beta}_1 median(x_1) - \hat{\beta}_2 median(x_2)) \\ =&\ var(median(y)) - 2 * cov(median(y), median(x_1) * \hat{\beta}_1) - 2 * cov(median(y), median(x_2) * \hat{\beta}_2) \\ &+\ var(\hat{\beta}_1) median(x_1)^2 + var(\hat{\beta}_2) median(x_2)^2 + cov(\hat{\beta}_1, \hat{\beta}_2) median(x_1) median(x_2) \\ \approx&\ var(median(y)) + var(\hat{\beta}_1) median(x_1)^2 + var(\hat{\beta}_2) median(x_2)^2 +\ cov(\hat{\beta}_1, \hat{\beta}_2) median(x_1) median(x_2) \end{aligned} \tag{12}$$

Usings eqns (8), (9), (10), (12) and Martin (3,4), covariance matrix consistent model intercept CI estimates for homoscedastic iid error cases, may be usefully obtained from evd approximations of the quantile regression residuals model variance via the relationship

$$s_{\beta_0}(\theta) = \sqrt{(s^2_{maxhalfCI} + \frac{s^2_{meanhalfCI}(var(x_2)median(x_1)^2 + var(x_1)median(x_2)^2 - 2cov(x_1, x_2)median(x_1)median(x_2))}{var(x_1)var(x_2)(1 - r^2_{12})})} \tag{13}$$

It should be noted that, variance inflation factors (VIF) are rarely considered for intercept model coefficient CI estimates except when comparing different models. The dearth of interest in VIFs for intercept estimates can be seen to arise from the strict relationship between r_{12}, $var(x_1)$, $var(x_2)$ and $cov(x_1, x_2)$ in (10). This strict relationship plays an important role in balancing the numerator and denominator magnitudes of eqn (13) as $r_{12} \to 1$, to leave the intercept model coefficient CI width relatively insensitive to the VIF. This theoretical behaviour coupled with good coverage performance under repeated quantile regression sampling simulation, is an strong test of the validity of evd based quantile regression modelling variance estimation approach, eqn (3).

The inconsistent evd based results found by Martin (3,4) for bivariate quantile regression intercept model coefficient CI estimates can be related to the approximation used for the intercept CI width in those papers

$$s_{\beta_0}(\theta) \sim \sqrt{s^2_{maxhalfCI} + \frac{s^2_{meanhalfCI}median(x_1)^2}{var(x_1)(1 - r^2_{12})} + \frac{s^2_{meanhalfCI}median(x_2)^2}{var(x_2)(1 - r^2_{12})}} \tag{14}$$

where it is now understood that the $cov(x_1, x_2)$ term was not consistently included. In the limit, $cov(x_1, x_2)$ & $r_{12} \to 0$, eqn (14) would have been adequate for particular examples.

Importantly, to keep the evd based variance estimation maximising the use of quantile regression calculations. The residuals distribution variances $s_{maxhalfCI}$ and $s_{meanhalfCI}$ are not calculated using sum of squares of errors of the residuals but are 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. As noted in (3,4) $s_{maxhalfCI}$ term applies to the numerator of the slope CI estimates and the constant term of the intercept CI estimator giving respectively, good coverage performance and comparability to bootstrap results for the minimum intercept CI width. The $s_{meanhalfCI}$ modification was introduced in (4) for the covariate dependent terms of the intercept CI estimator giving improved 95% coverage performance under linear expanding horn heteroscedasticity.

Algorithm for evd based estimators for use as quantile regression model coefficient CI estimators in homoscedastic iid error cases

In practice, this evd based model CI method to approximate quantile regression model coefficient CIs for homoscedastic iid errors contains the following steps.

(a) Firstly, the quantile estimating function is used to perform the quantile regression modelling of $\mathbf{Y} = \beta\mathbf{X} + \varepsilon$ and derive the residuals distribution,

(b) the validity of homoscedasticity iid errors in the residuals is assessed,

(c) 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 homoscedastically transformed quantile regression residuals distribution,

(d) the quantile estimating function is then used on the residuals distribution to backtransform the evd based model slope coefficent(quantile) confidence interval bound values (θ_{LB}, θ_{UB}) to the original measurement scale.

(e) for the slope CI and constant term of the intercept CI estimator, the **maximum half CI** of the backtransformed results is obtained,

(f) for the slope terms of the intercept CI estimator, the **mean half CI** of the backtransformed results is obtained,

(g) depending on the number of covariates and following eqn (3), $var(x_1)$, $var(x_2)$, $cov(x_1, x_2)$, r_{12} and VIF etc are obtained,

(h) for the numerator of the slope terms of the intercept CI estimators the medians of the covariate data values are obtained using quantile estimating function,

(i) the regression formula eqns (8),(9) and (13), are used to calculate evd based approximation estimates of the quantile regression model intercept coefficient CIs. For other point estimates besides the intercept, replace median values in the RHS of (13) with the covariate value(s) differences $(x_i - median(x_i))$ of interest.

In the implemented code, the regression degrees of freedom adjustment $\sqrt{\frac{n}{(n-p-1)}}$ required for model coefficient CIs has been applied in the percentile frame (θ_{LB}, θ_{UB}). This is in contrast, to (3) where two versions of regression degrees of freedom adjustment were assessed.

In the presence of linear expanding horn heteroscedastic id errors (4), the above algorithm needs to be extended to include auxiliary regression of the quantile regression residuals, weighted median(y) evd variance and the calculation of weighted medians of the covariates.

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 three homoscedastic quantile regression cases. The coverage performance is calculated by reference to the known population regression model and comparison to default quantreg bootstrap estimates using 600 replicates.

Consistent with previous papers (3,4), both evd_max and bin_max evd based estimator results are presented. These evd estimator options represent fine tuning investigations of the confidence interval bound values (θ_{LB}, θ_{UB}) to deal with the application of a population distribution function (evd) to single sample estimation. For quantiles ~ 0.05, in small samples, the two options evd_max(bin_max) give over(under) estimates of the CIs respectively, and so provide a nice constrast in evaluating performance. In principle, a finite sampling version, eg. jackknife etc, of the evd_max estimator should be derived as the definitive evd based estimator and would provide insight on the {maximum half CI}, {mean half CI} contributions to the CI terms as finite sample adjustments.

Three samples size n=50,100,1000 were trialled and three distributions taking note to include the intercept inside the covariate distributions.

model (i) $y = 1 * x + 1 * x^2 + rnorm(n, 0, 10)$

where x = rnorm(n,90,90). A polynomial quantile regression relationship between the dependent variable (Y) and explanatory variable (rnorm(n,90,90)) with homoscedastic iid noise. The choice of a mostly positive data range ~(-90,270) for the explanatory variable results in significant covariate correlation between x & x^2.

model (ii) $y = 1 * (rnorm(n, 90, 90) + 0.5 * urlaplace(n, -90, 90)) + 1 * (0.5 * rnorm(n, 90, 90) + urlaplace(n, -90, 90)) + rnorm(n, 0, 10)$

A linear relationship between the dependent variable (Y) and two collinear random variables. The similarity and mixing of the random variables results in significant collinearity providing a strong test of the covariate correlation estimate.

model (iii)

$y = 1 * (exp(rnorm(n, 0, 2)) - 1) + 1 * (exp(rnorm(n, 0, 2)) - 1 + runif(n, -30, 60)) + rnorm(n, 0, 10)$

A bivariate example where the covariate correlation estimate needs to deal with uneven dependence between the variables and that one explanatory variable has a skewed distribution.

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).

In the graphs, the blue lines indicate the evd_max CI estimators, and the red lines indicate the bin_max CI estimators described in (3), which includes degrees of freedom correction in percentile scale. This overlap of estimators and subfigures allows a visual comparison of the effects of sample size, and estimator type. The black circles are quantreg (default) bootstrap estimates with 600 replicates. The green dashed line indicates 95% coverage level.

In subfigures (a-c) contain the first slope coefficient CI coverage, (d-f) contain the intercept coefficient CI coverage & (g-i) contain the second slope coefficient CI coverage for the indicated sample sizes.

Figures 1,2 & 3, display respectively, the coverage performance of bootstrap, bin_max & evd_max for the slope and intercept respectively of model (i), (ii) & (iii) as a function of the sample size.

model (i): beta1, beta0, beta2 CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

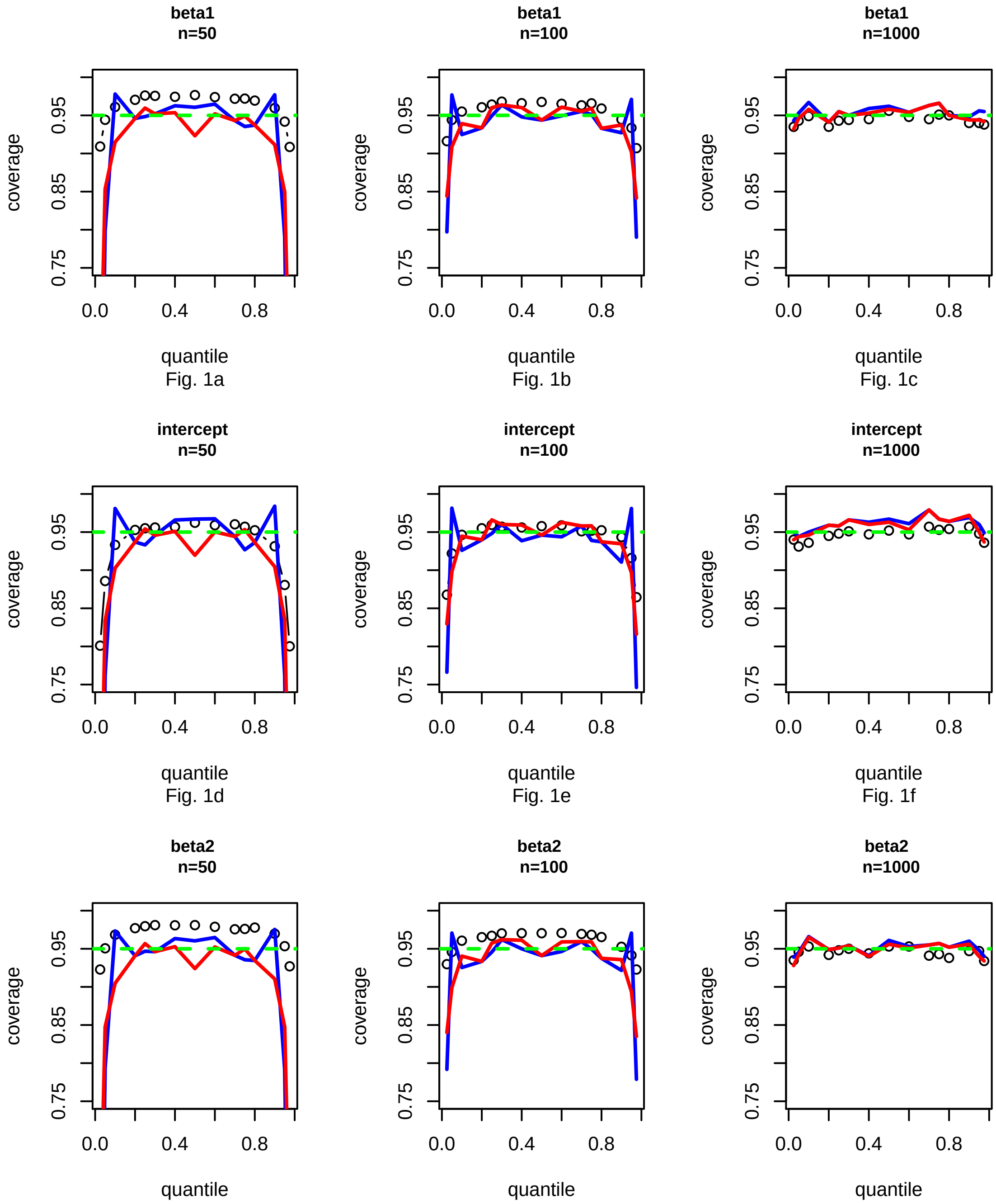


Fig. 1a

Fig. 1b

Fig. 1c

Fig. 1d

Fig. 1e

Fig. 1f

Fig. 1g

Fig. 1h

Fig. 1i

In figure 1, for n=1000, nominal 95% coverage is observed for bootstrap, evd_max, bin_max estimators for all quantiles (0.025-0.975) in slope and intercept CIs. For smaller samples, bin_max and evd_max based CIs estimates exhibit poorer coverage for extreme quantiles, while default bootstrap exhibited slight overcoverage.

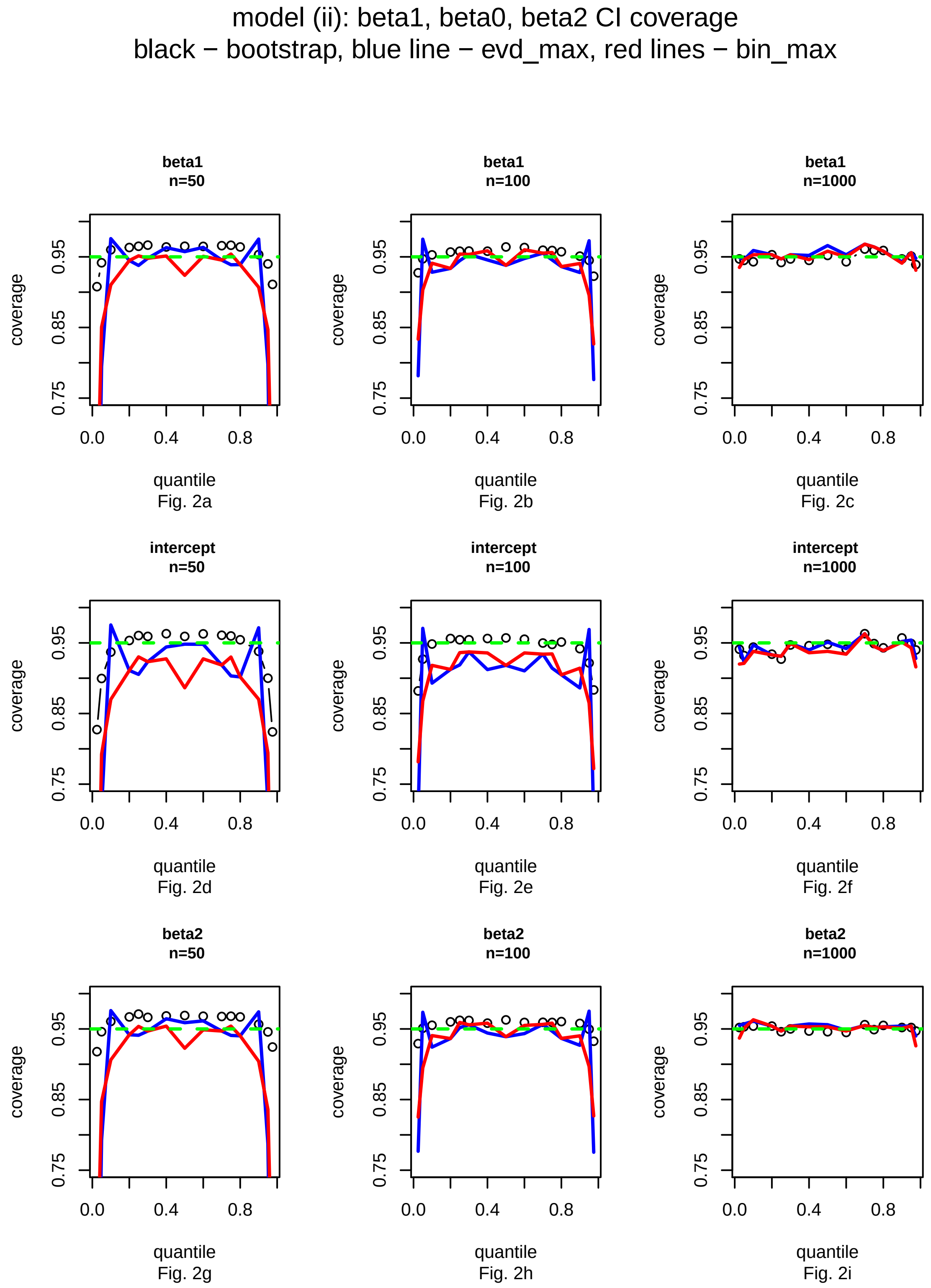
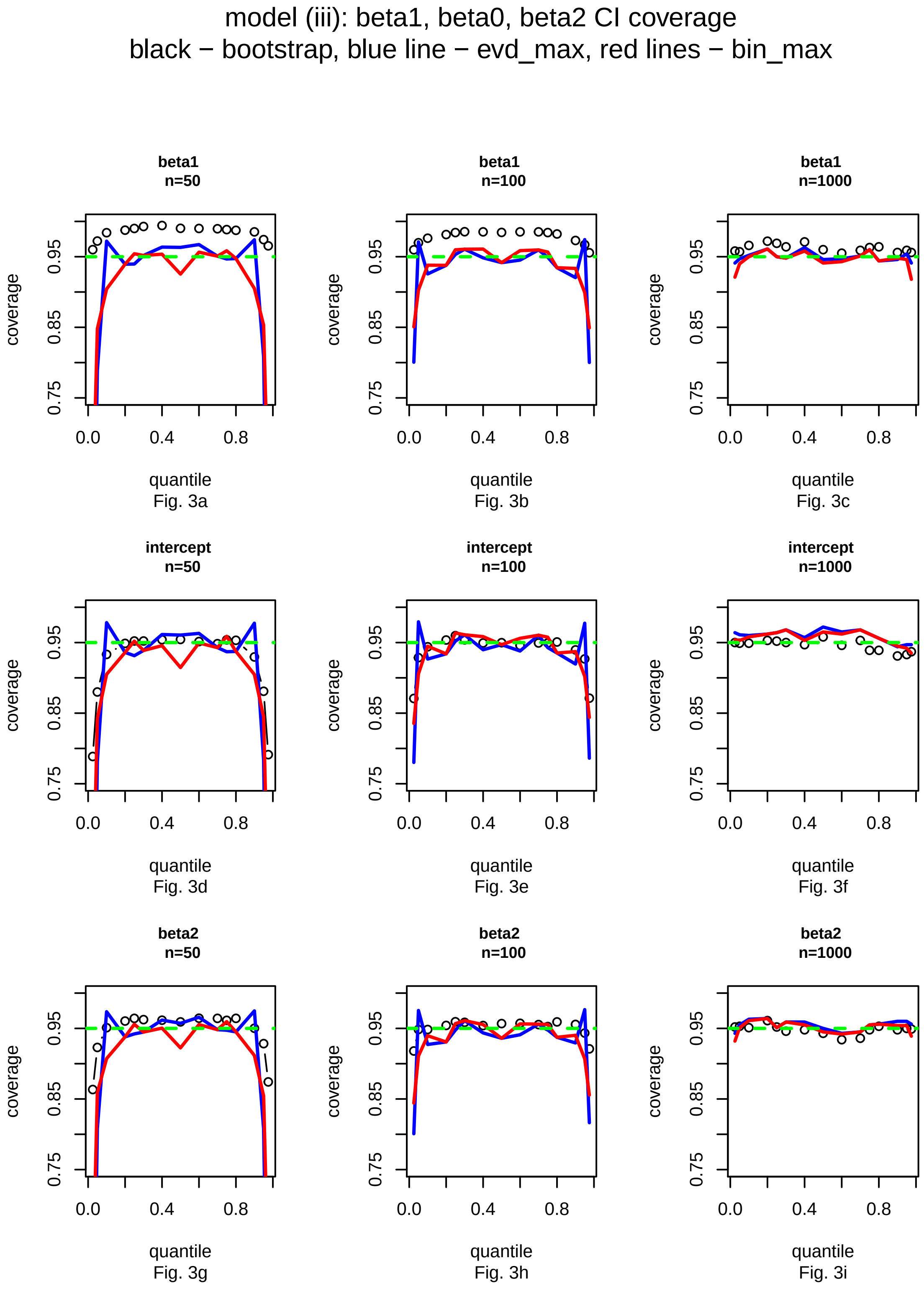


Fig. 2a Fig. 2b Fig. 2c

Fig. 2d Fig. 2e Fig. 2f

Fig. 2g Fig. 2h Fig. 2i

In figure 2, for n=1000, nominal 95% coverage is observed for bootstrap, evd_max, bin_max estimators for all quantiles (0.025-0.975) in slope and intercept CIs. For smaller samples, bin_max and evd_max based slope CIs estimates exhibit poorer coverage for extreme quantiles, bin_max and evd_max based intercept CIs estimates exhibit small undercoverage and default bootstrap exhibits slight overcoverage for slope CIs.



model (iii): beta1, beta0, beta2 CI coverage
black − bootstrap, blue line − evd_max, red lines − bin_max

beta1 n=50 — Fig. 3a

beta1 n=100 — Fig. 3b

beta1 n=1000 — Fig. 3c

intercept n=50 — Fig. 3d

intercept n=100 — Fig. 3e

intercept n=1000 — Fig. 3f

beta2 n=50 — Fig. 3g

beta2 n=100 — Fig. 3h

beta2 n=1000 — Fig. 3i

In figure 3, for n=1000, nominal 95% coverage is observed for bootstrap, evd_max, bin_max estimators for all quantiles (0.025-0.975) in slope and intercept CIs. For smaller samples, bin_max and evd_max based CIs estimates exhibit poorer coverage for extreme quantiles, while default bootstrap exhibited overcoverage for the slope CI of the skewed explanatory variable.

Conclusions

For homoscedastic, independent error, unweighted cases, full use of the covariance matrix provides improved evd based approximations to quantile regression model intercept coefficient CIs. Good coverage performance for model slope and intercept coefficient CIs is observed in the presence of significant collinearity, for moderate (n=1000) to large samples.

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. Martin J.P.D., 2015, http://dx.doi.org/10.6084/m9.figshare.2055882
4. Martin J.P.D., 2016, http://dx.doi.org/10.6084/m9.figshare.2055873
5. Koencker, R. W. & Bassett G., Econometrica, 1978, vol. 46, issue 1, pages 33-50
6. Koencker, R. W., Portnoy S. et al, https://cran.r-project.org/web/packages/quantreg/quantreg.pdf
7. https://en.wikipedia.org/wiki/Quantile
8. Brown, B. M. and Wang, Y.-G. (2005). Standard errors and covariance matrices for smoothed rank estimators. Biometrika 92 149-158. MR2158616