Semi and Non-Parametric Econometrics

Sébastien COUBE, Jean-Michel ROUFOSSE et Benjamin PHAN

Introduction

We studied an article from Victor Chernozhukov and Iván Fernández-Val on the inference for extremal conditional quantile models[1]. This article uses extreme values (EV) theory in order to derive inference for quantile regression in extremal quantiles, that is to say when $\tau \sim O(1/T)$, T being the number of individuals in the sample.

This article shows that normal confidence intervals are too tight when it comes to extreme quantiles. Then, their coverage deteriorates. Bootstrap estimators do not work neither, because central limit theorem does not apply in extremal quantiles.

EV theory gives good confidence intervals, but they rely on renormalizations that cannot be inferred from the data. The point of this article is to find self-normalized confidence intervals in order to compute those confidence intervals.

We applied and discussed the two methods that we found in the article. The first one relies on Pickands estimators in order to compute the estimators of the parameters of the quantile regression and their confidence intervals. The second one relies on sampling and self-normalized extreme values to provide confidence intervals. We tried to apply both of them. However it seems that the Pickands estimators are quite erratic. In addition, we suspect that there is a mistake in the Pickands estimator $\hat{\xi}$ defined in the equation 3.15 in the article $\left(\hat{\xi} = \frac{-1}{ln(2)} ln \frac{\bar{X_T'}\hat{\beta}(4\tau_T) - \hat{\beta}(\tau_T)}{\bar{X_T'}\hat{\beta}(2\tau_T) - \hat{\beta}(\tau_T)}\right)$ as it does not correspond to the expression of this estimator appearing in other works (where we can find $\hat{\beta}(2\tau_T)$ instead of $\hat{\beta}(\tau_T)$ in the numerator). Our attempts to reproduce the method using the Pickands estimators were not very convincing so we preferred to focus on the results of the second method.

We chose to study the impact of various variables on the heart rate.

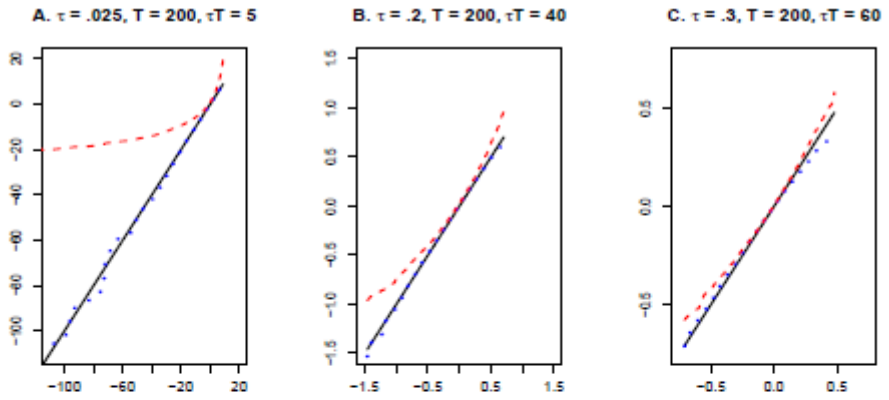
I Choice of the method

In order to understand the method better, we tried to reproduce the figure 1 from Fernández-Val and Chernozukhov's article. This led us to think carefully about the methods that are implemented in the R package quantreg.

I.1 Presentation of the figure 1

The authors want to show that the EV approach for confidence intervals is accurate, unlike the normal approximation, using a Monte-Carlo experiment and quantile-quantile plots (qqplot). All what we can assume about a qqplot is that it is an increasing curve. If two vectors are sampled according to the same distribution, the qqplot will look like the bisector. Here, for each $\tau = 0.025, 0.2, 0.3,$ 10000 samples Y of size 200 were drawn from a $Cauchy(1,1)$ distribution. Then, for each of those Y, $\widehat{q_\tau(Y)}$ was estimated with quantile regression.
Then, using EV and normal approximation, the authors deduced two distributions for $\widehat{q_\tau(Y)}$.
Then, they compared these two distributions and the empirical distribution of $\widehat{q_\tau(Y)}$. We can see that the EV samples are close to the bisector, and that the normal samples are far from the bisector. So the EV seems much better than the normal approximation.

[1] *Inference for Extremal Conditional Quantile Models, with an Application to Market and Birthweight Risks*, 26 Dec 2009 https://arxiv.org/abs/0912.5013



. **Quantiles of the true law of QR vs. quantiles of EV and normal laws**. The figure is based on a simple design with $Y = X + U$, where U follows a Cauchy distribution and $X = 1$. The solid line "——" shows the actual quantiles of the true distribution of QR with quantile index $\tau \in \{.025, .2, .3\}$. The dashed line "- - -" shows the quantiles of the conventional normal law for QR, and the dotted line "....." shows the quantiles of EV law for QR. The figure is based on 10,000 Monte Carlo replications and plots quantiles over the 99% range.

Figure 1: Toy example form the beginning of the article

I.2 Reproduction of the results

We tried to reproduce the left-hand plot in figure 1. Questions quickly arose about the method we had to use. Indeed, the Frisch-Newton method fits the plots from the article, while the Barrodale-Roberts method does not.
We simulated 10000 samples of $T = 200$ iid $Cauchy(1, 1)$ variables and estimated the quantile of order $\tau = 0.025$ with the Frisch-Newton ('fn') and Barrodale-Roberts ('br') methods from the 'quantreg' package. We computed the normal quantiles with the formula:

$$\hat{q_\tau} - q_\tau \sim \mathcal{N}\left(0, \frac{1}{\sqrt{n}} \frac{\sqrt{\tau}(1-\tau)}{f(q_\tau)^2}\right) \tag{1}$$

with f and q_τ respectively the density function and the theoretical quantile of order τ of distribution $(Cauchy(1, 1))$. We used a simplified formula : indeed, this model is a location model.
We also computed the theoretical EV quantile with the formula (2.5) of the article:

$$\hat{q_\tau} - q_\tau \sim q_\tau \left(\Gamma_k^{-\xi} - k^{-\xi}\right) \tag{2}$$

with $k = \tau T$, Γ_k a Gamma variable of parameter k and $\xi = 1$ in the case of a Cauchy distribution.

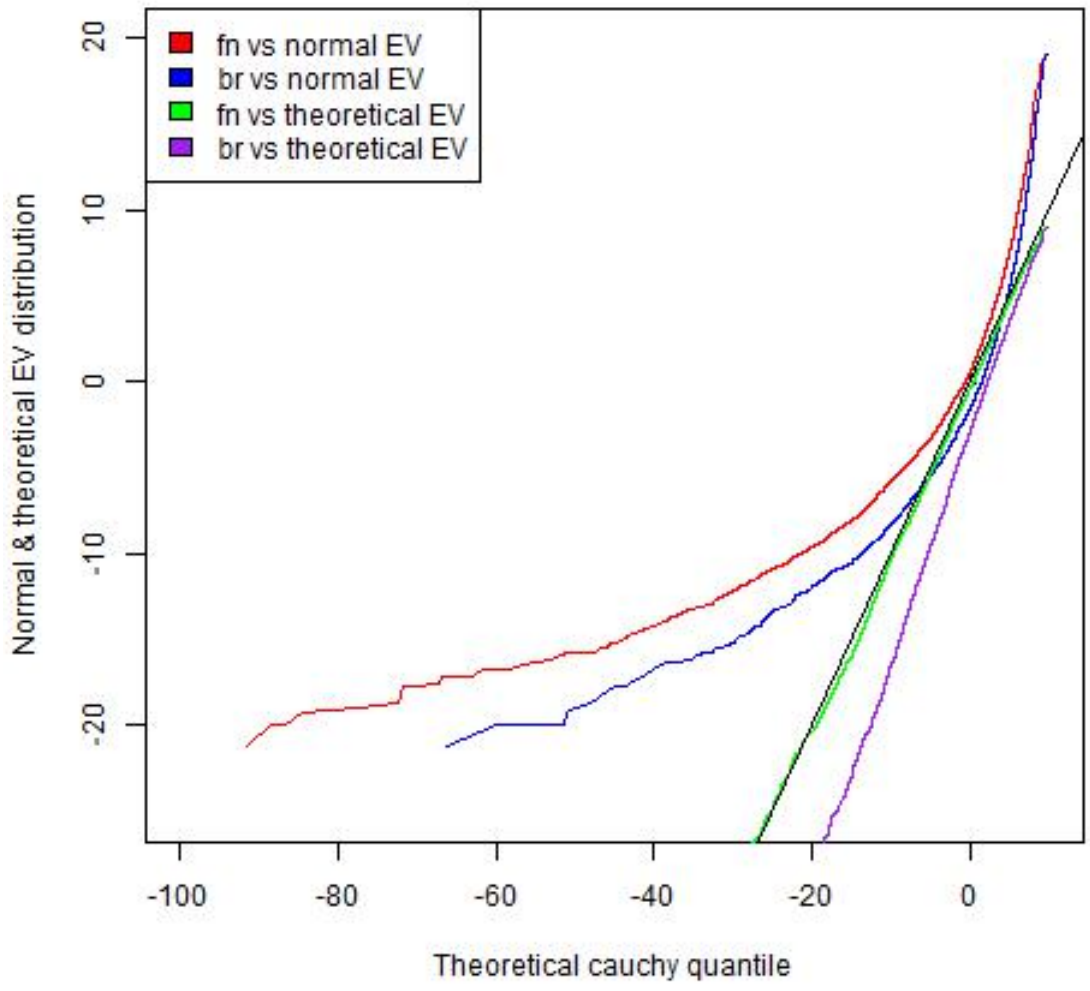


Figure 2: Trying to reproduce the example with method 'fn' and 'br' of the package quantreg ($\tau = 0.025$)

I.3 Comparison of the methods of the 'quantreg' package

We can observe (figure 7 in annexes) that with $\tau = 0.025$ and $T = 200$ (so $k = 5$) the 'br' method corresponds *exactly* to z_6, the 6^{th} ($(k+1)^{th}$) smallest value (i.e. the empiricial 6/200 quantile), while 'fn' *roughly* corresponds to z_5, the 5^{th} (k^{th}), with a mean absolute error of 0.47. Actually the 'fn' method gives an estimation almost systematically between z_5 and z_6

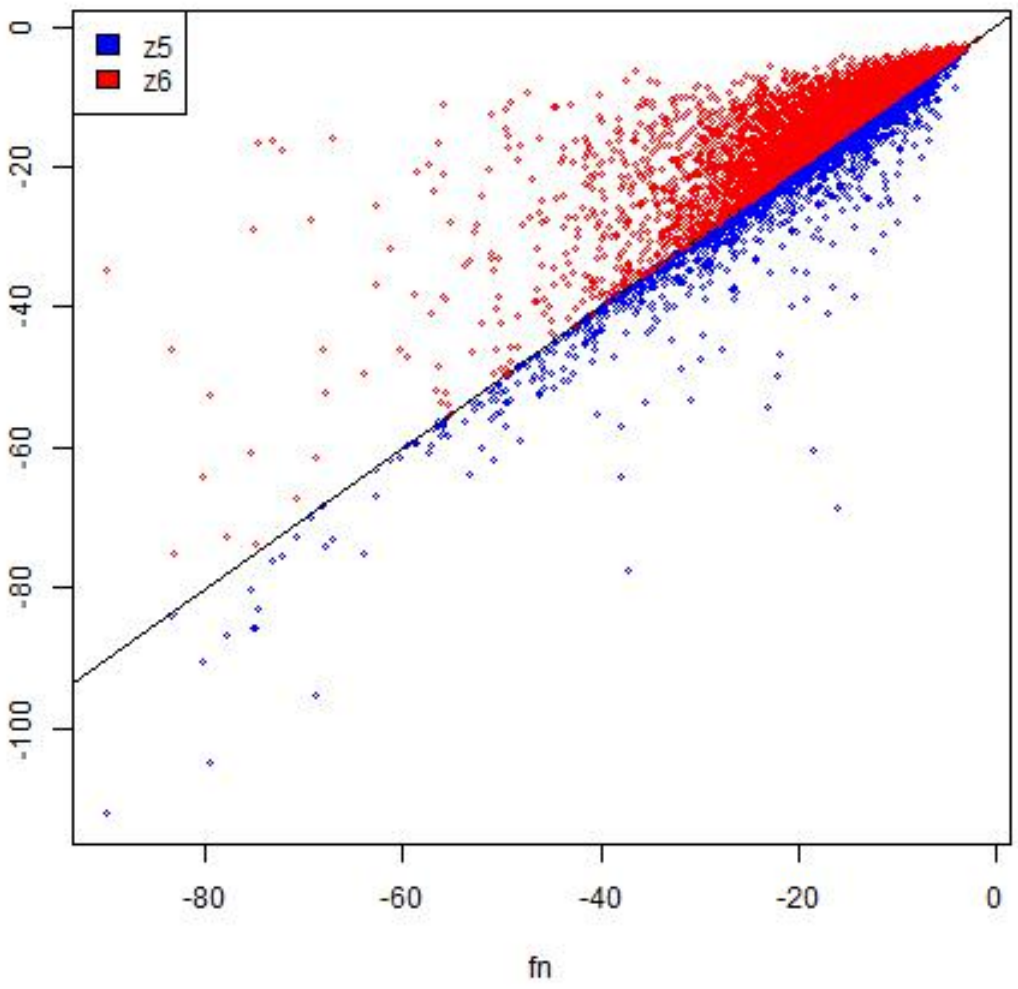


Figure 3

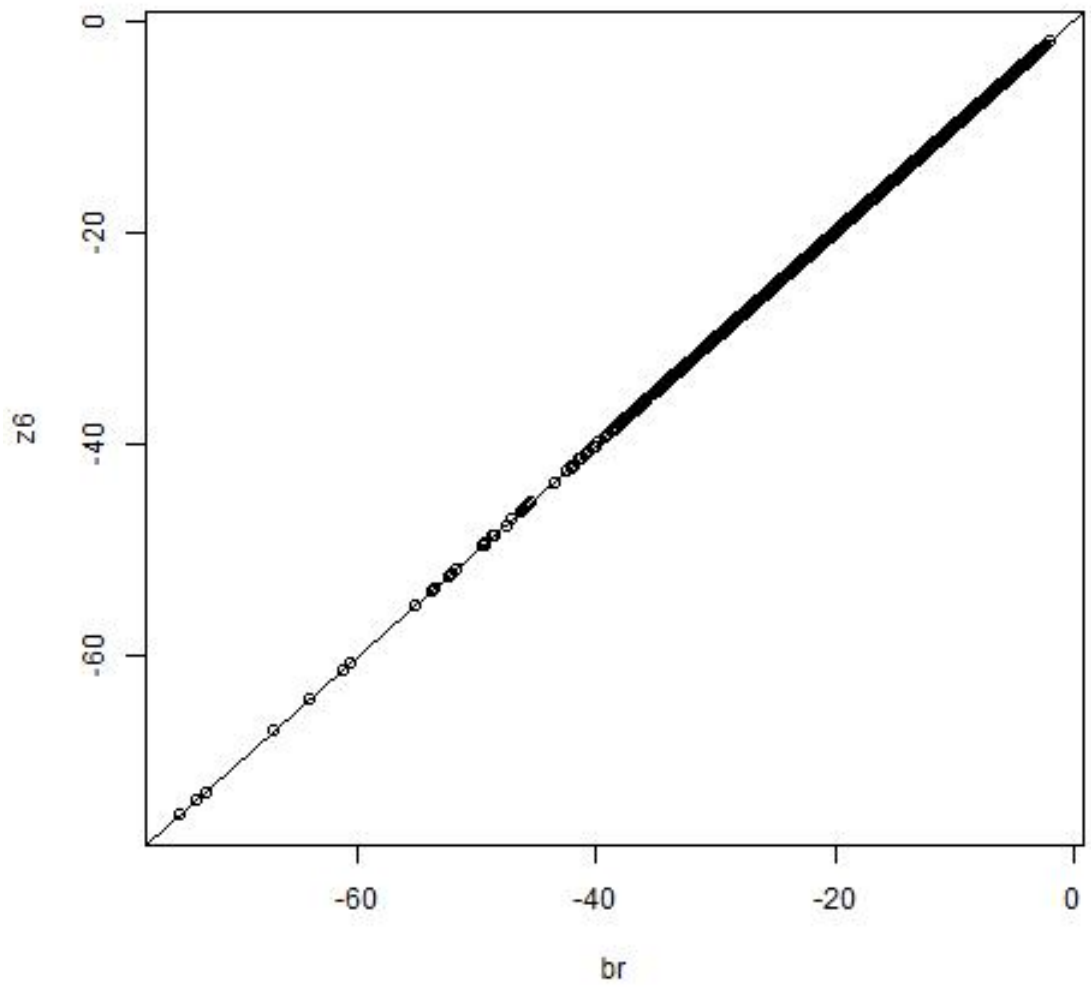


Figure 4

However, when we subtract a *small* $\epsilon > 0$ to τ (i.e. sufficiently small such that $\epsilon T < 1$) we observe that both 'fn' and 'br' method correspond *exactly* to the 5^{th} value (the mean absolute error is 0 with the 'br' method and about 10^{-5} with the 'fn' method).

Therefore, we can make the hypothesis that the 'br' method uses a definition of the cdf with a strict inequality while the 'fn' method uses a loose inequality. However, when τT is an integer, the 'fn' method seems to be slightly biased. This bias is positive for $\tau < 0.5$ and negative for $\tau > 0.5$ (cf figure 8 in annexes)

Moreover, we notice that the mean of z_6 is really close to the true quantile of order $5/T$ of the Cauchy law (of parameters $(1, 1)$)while z_5 is not as close. As we know, the empirical quantile is not an unbiased estimator of the true quantile in general[2]. And it seems that the empirical quantile of order $6/T$ is a better estimator. However, it is a bit disputable to talk about *unbiased* estimator as the expectations of empirical quantiles are not defined for a Cauchy law.

This an important question in our case. Even though with fixed τ the empirical quantile converges to the true quantile as T tends to infinity, we *can not* make this assumption as we have a τ varying inversely proportionally to T. To understand better this phenomenon and determine how to use the

[2]https://stats.stackexchange.com/questions/76259/demonstration-of-sample-quantile-bias

regression module, we modified some parameters. Moreover, the bias of empirical quantiles is in general greater when the quantile is extreme and the distribution has a strong variance (the variance is not even defined for a Cauchy distribution).
Instead of taking $T = 200$ like in the article, we took $T = 5000$. Our experience confirms our precedent results. z_6 is a better estimator of the true quantile of order $5/T$ (the true quantile is -318 and z_6 has a mean of -316 while z_5 has a mean of -394). Still, z_6 does not seem to be a perfect estimator.
We also tried to replace the Cauchy distribution by a Normal distribution of variance=1. In this case, z_6 seems to be slightly positively biased (the true quantile is -2.09 and z_6 has a mean of -2.06) and z_5 seems to be slightly negatively biased (mean of -2.11). We can assume that taking z_6 as estimator of the quantile of order $5/T$ is a trick that does not work with any law. We also notice that the bias is relatively small compared to the case with a Cauchy distribution.

However, when we add explicative variables to the model, it seems that there is no problem in the estimation of the parameters with either method and we do not need to subtract an ϵ to τ because both methods yield the same results. Their behavior was correct and we did not worry ourselves more about this.

We reproduced an example that was presented in the beginning of the article in order to understand better the advantages of EV approach. However, we used theoretical EV distribution. When it comes to data obtained with a survey, we do not have this information. Then, we need to implement the so called self-normalized *EV.*
We also wanted to test the estimators on toys examples that would be a little bit more difficult than estimating a quantile.

II Implementation of self-normalized extreme-value confidence intervals

II.1 Explanation of the method

The canonically-normalized EV confidence intervals cannot be implemented because they rely on a renormalization that cannot be calculated without precise information about the distribution of the nuisance parameters.
The objective is to establish a confidence interval using the following formula:

$$\lim_{T\to\infty} P\left(\psi'\widehat{\beta_\tau} - \widehat{c}_{1-\alpha/2}/\widehat{A} \leq \psi'\beta_\tau \leq \psi'\widehat{\beta_\tau} - \widehat{c}_{\alpha/2}/\widehat{A}\right) = 1-\alpha$$

Let's define a little bit more the terms.

- β and $\widehat{\beta}$ are respectively the true and the estimated coefficients.
- ψ is any non-random vector. In particular, we can set $\psi = (0, ..., 0, 1, 0, ..., 0)$ in order to select a coefficient.
- $\widehat{A}$ is the renormalisation. It is precisely the issue with canonical EV statistics: here, $\widehat{A}$ is a function of the data.
- $\widehat{c}_q$ is the empirical q-quantile of a collection of statistics obtained through resampling.

The two last terms are calculated using a factor m. This factor has a role in the computation of renormalizing constants. Using a *rule of the thumb* (an expression used by the authors of the article), it is set to: $1 + (d+p)/\tau * T$, d being the dimension of the regressors, T the number of regressors, and p is a small integer -according to the authors of the article, taking it between 2 and 20 changes little to the result. Important point, m is supposed to be bigger than 1.
Now that we introduced m, the renormalization is defined as follows:

$$A = \frac{\sqrt{\tau T}}{\bar{X}'(\widehat{\beta}_{m\tau} - \widehat{\beta}_\tau)}$$

with $\bar{X}$ being the mean of the regressors X Here, we can see that it is important to have $m > 1$ so that $A > 0$.
$\widehat{c}_q$ is computed using subsamples of X. For B_T subsamples[3] of size b , one computes the analogues of A and β_τ. The difference is that τ is replaced by $\tau_b = \tau * T/b$. Since b is small before T, τ_b has to be bigger than τ. Then, *in each subsample* one has to compute

$$\widehat{A_b} = \frac{\sqrt{\tau_b b}}{\bar{X}'(\widehat{\beta}_{m\tau_b} - \widehat{\beta}_{\tau_b})}$$

$$\widehat{V} = \widehat{A_b}\psi(\beta_{tau_b\ subsample} - \beta_{tau_b\ complete\ sample})$$

c_q is then selected as the q^th empirical quantile of of the collection $\{\widehat{V}_1 ... \widehat{V}_{B_T}\}$ and used in the confidence interval.

One can note that b has to be smaller than T and $m > 1$. Then $m\tau_b$ will be big when compared to τ : it becomes a problem when it comes to estimate this confidence interval for a relatively high τ.

II.2 Implementation of the method on a simple example

In order to verify that the method we implemented gives sensible results, we constructed a simple example with known parameters. We simulated $(Y_i, X_i)_{i\in[\![1,1000]\!]}$ iid with $X \sim \mathcal{U}[0, 10]$ and $Y = X + \sqrt{X} + \epsilon$ with $\epsilon|X \sim \mathcal{N}(0, \sqrt{X})$. This example, yet simple, induces heteroscedasticity as the variance of $Y|X$ depends on X. We have the following regression model:

$$Y = \beta_0 + \beta_1\sqrt{X} + \beta_2 X + \nu \tag{3}$$

We have

$$q_{Y|X}(\tau) = \sqrt{X} + X + q_{\mathcal{N}(0,\sqrt{X})}(\tau)$$
$$q_{Y|X}(\tau) = \sqrt{X} + (1 + q_{\mathcal{N}(0,1)}(\tau))X$$

($|X| = X$ *a.s.*)

[3] subsets of "small" size drawn without replacement

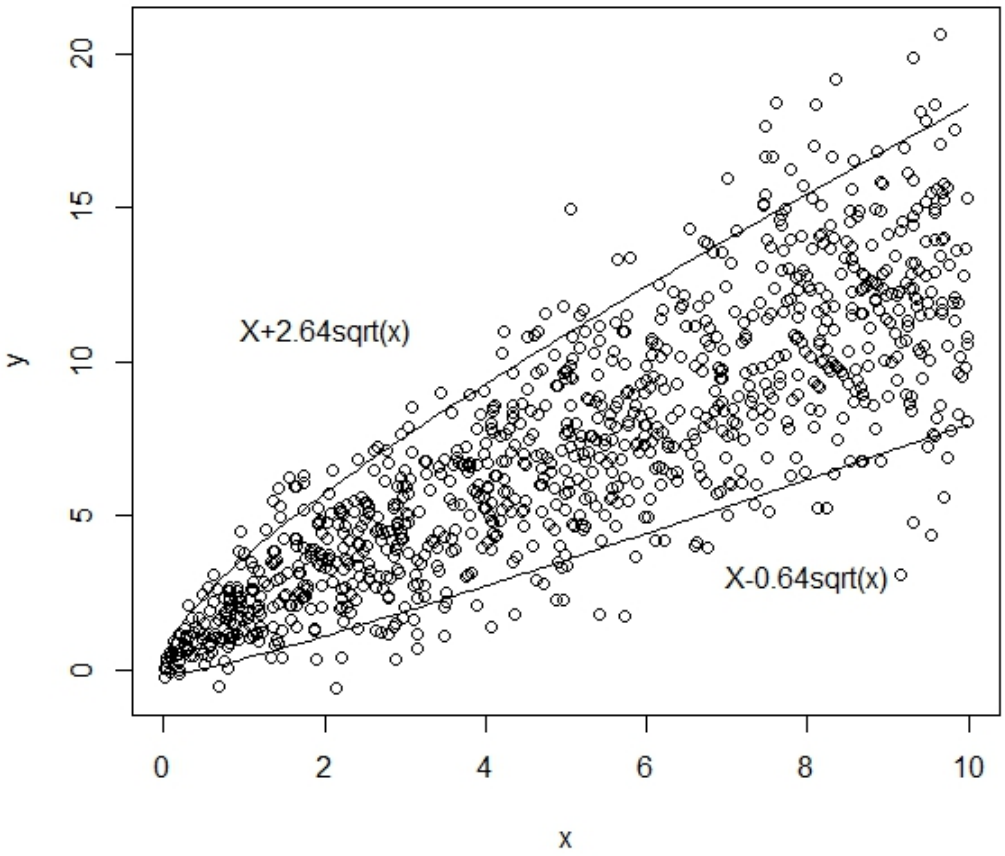


Figure 5: Simulation for our example

Therefore a quantile regression of our model should give: $\beta_0(\tau) = 0$, $\beta_1(\tau) = 1$ and $\beta_2(\tau) = 1+q_{\mathcal{N}(0,1)}(\tau)$. For instance, we should get $\beta_2(0.05) \approx -0.64$ and $\beta_2(0.95) \approx 2.64$ as $q_{\mathcal{N}(0,1)}(0.05) \approx -1.64$ and $q_{\mathcal{N}(0,1)}(0.05) \approx 1.64$.

Knowing that, we repeated several times the previous simulation estimating each time the parameters β_j and their 90% confidence intervals. Fortunately, the confidence intervals contain the expected values about 90% of times which is what was supposed to occur.

In our bootstrap R code, we have two parts:

- Sampling of the SN-QR statistic for the full sample size : in which we chose $b = 100$ and $B_T = 200$
- Calculating the confidence interval :

In our function, since we have $b = 100$ and $B_T = 200$ (as a reminder, $T = 1000$), the maximum quantile we can use in our regression is $\frac{b}{T} = \frac{100}{1000} = 0.1$. This explains why we are interested in values of τ in the following set : $\tau \in \{0.001, 0.002, 0.003, 0.004, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.08, 0.09\}$. We decided to only take these 12 values because we chose to perform a total of 100 [4] simulations for each value of τ, and the bootstrapping part of our code takes time. Time results would have been worse if we hadn't parallelized a great deal of the bootstrapping.

[4] according to the CLT and the fact that each experience is a $Bern(0, \alpha)$, this should be enough to deal with the variance

Moreover, modeling the normal inference method wasn't straightforward. In our case, we could not use the simplified formula for normal inference and we did not manage to implement the complete formula. We used the standard deviation estimates given by the package in order to compute the confidence intervals as interval length between bounds provided by the package would sometimes be close to 10^2. As a result, we decided to compute the bounds ourselves, using :

$$upper\_bound := \beta_{i\in[|0:2|]} - q_{90\%_{\mathcal{N}(0,1)}} * se(\beta_{i\in[|0:2|]})$$
$$lower\_bound := \beta_{i\in[|0:2|]} + q_{90\%_{\mathcal{N}(0,1)}} * se(\beta_{i\in[|0:2|]})$$

Here we have $q_{90\%_{\mathcal{N}(0,1)}} \approx 1.64$ and $se(\beta_{i\in[|0:2|]})$ being the standard error of the estimator $\beta_{i\in[|0:2|]}$, quantity calculated thanks to the quantile regression. We quickly noticed that results from our modelization of the normal inference were extremely similar to the article's results, thus showing our approach was successful.

From the result figure, we can see that the coverage obtained from extreme-value confidence intervals is very close to 90%, for low values of τ, like $\tau \in \{0.001, 0.002, 0.003, 0.004, 0.005\}$, as well as higher values of τ like $\tau \in \{0.05, 0.08, 0.09\}$. This cannot be said for the normal inference method. Indeed, despite the fact that for $\tau \in \{0.05, 0.08, 0.09\}$ the normal inference method gives us coverages around 90%, when we are interested in extreme values like for $\tau \in \{0.001, 0.002, 0.003, 0.004, 0.005\}$, the coverage thus obtains ranges from 0.3% to 0.75%, therefore resembling the results in the article.

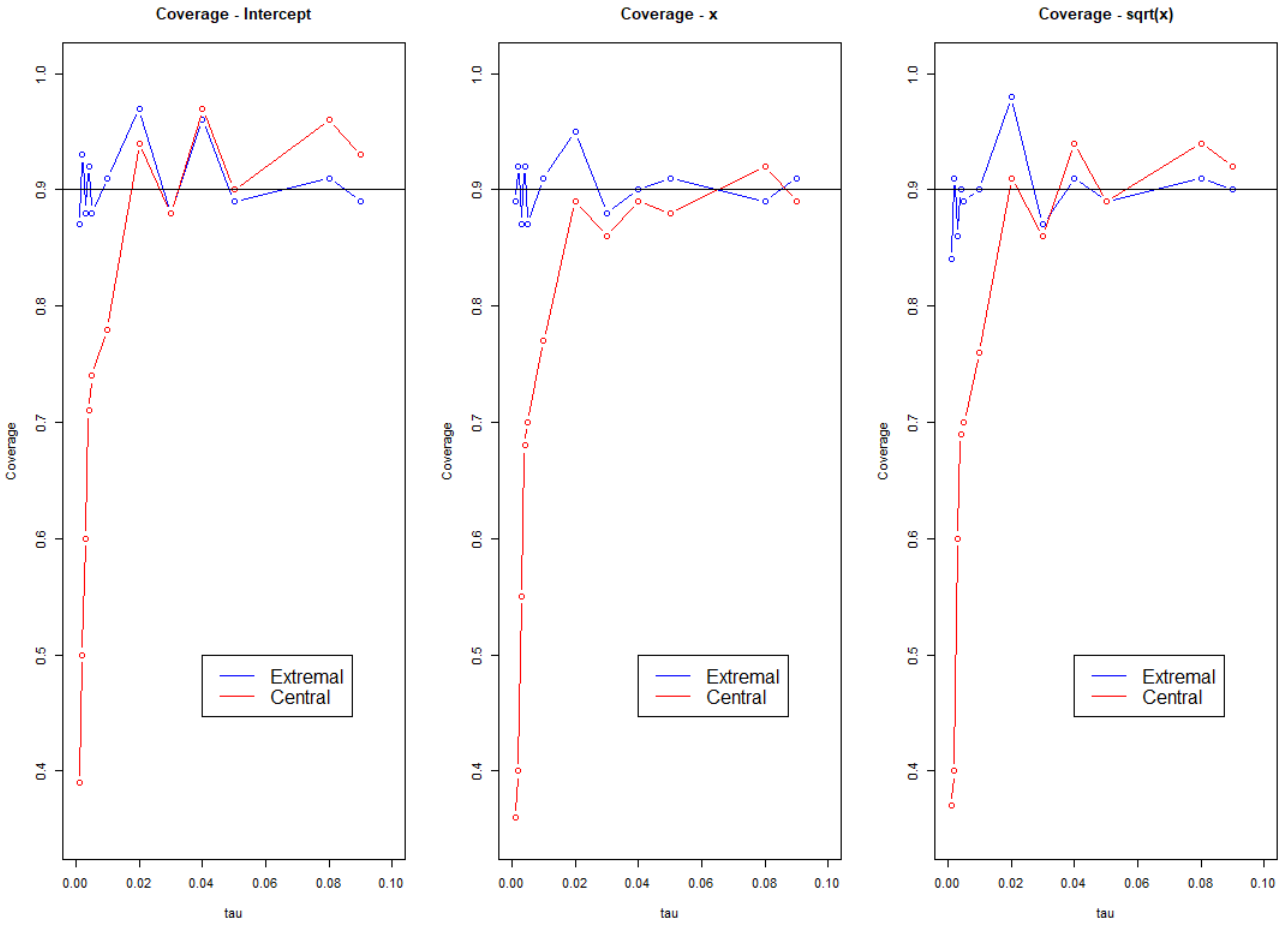


Figure 6: Coverage of extremal confidence intervals and normal confidence intervals of the model $Y = X + \sqrt{X} + \epsilon$ with $\epsilon|X \sim \mathcal{N}(0, \sqrt{X})$. Based on 1,000 repetitions.

III Applying the method to extremal quantile regression of pulse rate

III.1 Description of the database

We decided to apply these regressions techniques to study quantiles of heart rates from individuals part of the *National Longitudinal Study of Adolescent to Adult Health*. These respondents have been surveyed through their adolescence and young adulthood, and at the time of the questioning in 2008, corresponding to the the wave IV, subjects were then between 24 and 34 years old. A survey[5] of roughly 1000 questions was used to create the database we will be using in this study. These questions cover a wide range of subjects, establishing social and economic backgrounds and also covering the person's diet, health habits, illness history etc ...

We used three general regressors : sex, BMI[6], and age.

- In the last 30 days, how often have you felt that difficulties were piling up so high that you could not overcome them? The question had 5 answers, that we put into 3 groups : a group of respondents that declared it never happened, one that declared it almost never happened, and a mixing of 3 low-frequencies groups : "sometimes", "often", and "very often". We had to mix those groups in order to avoid singular design matrices during resampling.

- How many close friends do you have? This question was a binned quantitative variable. (Close friends include people whom you feel at ease with, can talk to about private matters, and can call on for help.)

For the obvious reason that a regular physical training may have an effect on cardio-vascular capacity, we used sport-related covariates. Unfortunately, this part of the survey was poorly designed, and some of the questions are catch-all questions that commingle sports that have little in common.[7] All of these questions had integer answers.

- In the past seven days, how many times did you bicycle, skateboard, dance, hike, hunt, or do yard work?

- In the past seven days, how many times did you roller blade, roller skate, downhill ski, snow board, play racket sports, or do aerobics?

- In the past seven days, how many times did you participate in strenuous team sports such as football, soccer, basketball, lacrosse, rugby, field hockey, or ice hockey?

- In the past seven days, how many times did you participate in individual sports such as running, wrestling, swimming, cross-country skiing, cycle racing, or martial arts?

- In the past seven days, how many times did you participate in gymnastics, weight lifting, or strength training?

- In the past seven days, how many times did you play golf, go fishing or bowling, or play softball or baseball?

- In the past seven days, how many times did you walk for exercise?

- On the average, how many times per week do you use a physical fitness or recreation center in your neighborhood?

[5] `http://www.thearda.com/Archive/Files/Codebooks/ADDHW4H_CB.asp`

[6] body mass index

[7] We did not use one of these variables : *In the past seven days, how many times did you play golf, go fishing or bowling, or play softball or baseball?* In addition of being an infamous catch-all, those who gave a positive answer are few ; this led us to singluar design matrices during resamplings.

We also added food-related regressors targeting sugar, fat and salt-heavy consumptions. Those questions also had integer answers.

- How many times in the past seven days did you eat food from a fast food restaurant, such as McDonald's, Burger King, Wendy's, Arby's, Pizza Hut, Taco Bell, or Kentucky Fried Chicken or a local fast food restaurant?

The next questions ask about what you generally eat and drink at home and away from home.

- In the past seven days, how many regular (non-diet) sweetened drinks did you have? Include regular soda, juice drinks, sweetened tea or coffee, energy drinks, flavored water, or other sweetened drinks.
- In the past seven days, how many diet or low-calorie drinks did you have? Include diet sodas, unsweetened tea or coffee, or other drinks sweetened with artificial sweeteners.

We ended this group of lifestyle-related regressors by a section of drug variables. Unfortunately, there was no question about long-term coffee and tea consumption. All of the variables were binary or integers.

- During the past 30 days, on how many days did you drink?
- Did you have an alcoholic drink (beer, wine, or liquor) within the past 24 hours?
- During the past 30 days, on the days you smoked, how many cigarettes did you smoke each day?
- Did you drink a caffeinated beverage (e.g., coffee, tea or soda) in the past 24 hours?

We eventually added variables that indicate if the individuals have proper healthcare.

- The next questions are about health insurance and health services. Which of the following best describes your current health insurance situation?
  The question has many answers we merged into 3 categories: insured, not insured, and Medicaid
- How long ago did you last have a routine check-up? This question was a binned qualitative variable, that we used as a quantitative variable.

III.2 Results

In practice, the τ we used for regression were between 0.01 and 0.1, and 0.9 and 0.99. We used $b = 2000$ and $BT = 500$. There are 5000 individuals in the base.
Most of the time, we did not find major discrepancies between EV intervals (blue) and normal intervals (red). In most cases, the EV interval even seems to be tighter than the normal interval. However, for$\tau = .99$ EV confidence intervals are much wider. We can assume that it is hard to estimate effects for the highest quantiles perhaps because of unobserved variables such as chronic diseases. The normal inference may be optimistic for the high end of the distribution.
There is a lot to say about the coefficients.

III.2.a BMI, age and sex

As expected, BMI has a positive effect. For the lowest quantiles, being a woman seems to have an important effect, and to lower pulse rate by 4 bpm. Age seems to increase pulse rate, which can be explained in terms of life cycle : for individuals between 24 and 34 years, the effect of age itself can be reinforced by a more sedentary life.

III.2.b Psychology and wellbeing

As for the feeling of being overwhelmed by difficulties, there is a strong effect for lower quantiles. The reference is the category of people who think they "almost never" feel they cannot cope with difficulties. Surprisingly, the other two categories ("never", "sometimes and often") have lower pulse rate. We would rather expect chronic stress to increase the pulse rate. In order to deal with this strange finding, we could maybe look towards declarative bias.
Declaring to have many close friends seems to have greater effects on extreme quantiles on both sides ; we cannot explain.

III.2.c Drinks, food and smoking

Having eaten fast food or drunk sodas (light or not) within the past 7 days induces a really small increase of the heart rate for either the low quantiles but it is not significant as 0 is in the confidence intervals almost every time. These variables does not seem significant either for the high quantiles. However, caffeinated drinks increase a little the heart rate for the most extremely low quantiles ($\tau < 0.4$) while decreasing it for the high quantiles.
Concerning alcohol, the consumption within the last 24h increases the heart rate. This may be because people who have consumed alcohol recently are drunk or have hangover which have a direct impact on their health state, increasing their heart rate. Meanwhile, we observe that the consumption of alcohol within the past 30 days rather decrease the heart rate, especially for the high quantiles.
Smoking cigarettes does not have any significant effect on the pulse rate for either end of the distribution.

III.2.d Sports and physical activity

For low quantiles, sport-related variables are found to be negative. Yet, depending on the type of sport, that effect is found to be more or less strong. For example, walking does tend to decrease the heart rate, but however with a small effect. Martial arts and individual sports requiring endurance, such as running or swimming have an outstanding effect. We think that this coefficient means that there are marathoners, triathletes, heavy swimmers, whose pulse rate is really low, in the lower quantiles.

Interestingly enough, walking has a stronger effect of decreasing one's heart rate if one is in the highest end of the heart rate distribution. This shows that for those who have an elevated heart rate even such a small physical activity like walking can decrease it. However, looking at the high endurance sport effect, we see an effect almost equal to zero for hight quantiles. We explain this by the following : there are so few athletic people with high heart rates in our data base that the only reasonable regression coefficient is a close to 0 value. On the other hand, for people with poor physical condition, walking for exercise is a soft sport that can be practiced more easily than other sports.

We can indeed confirm the intuitive idea that sports has a stabilizing effect on one's heart rate. Not only does doing more intense sports reduces it for the lowest end of our distribution, but even on the other tail of the distribution do we have a reducing effect of the heart rate even with such an activity as walking. Sports can finally be linked with a lesser amount of stress in individuals, in addition to having better cardio-vascular capacities.

III.2.e Medical attention

Surprisingly, being without an insurance or relying on medicaid has little influence. We can explain this by the fact that respondents are relatively young.
For both ends of the distribution, there is a surprising result : the farther the last routine check-up, the lower the pulse rate. This might be because of endogeneity : if an individual knows that his pulse rate is high (which is really likely because of smartphones), he might consider to go and see the doctor for a checkup.

IV Annexes

IV.1 Figures

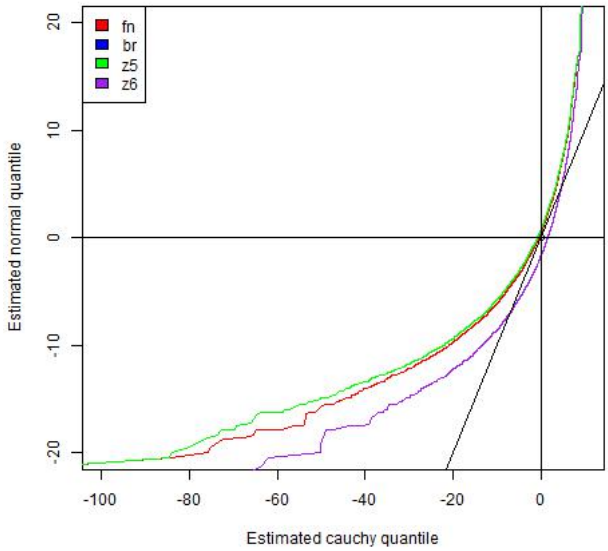


Figure 7: Comparison between fn, br, z5 and z6 ($\tau = 0.025$)

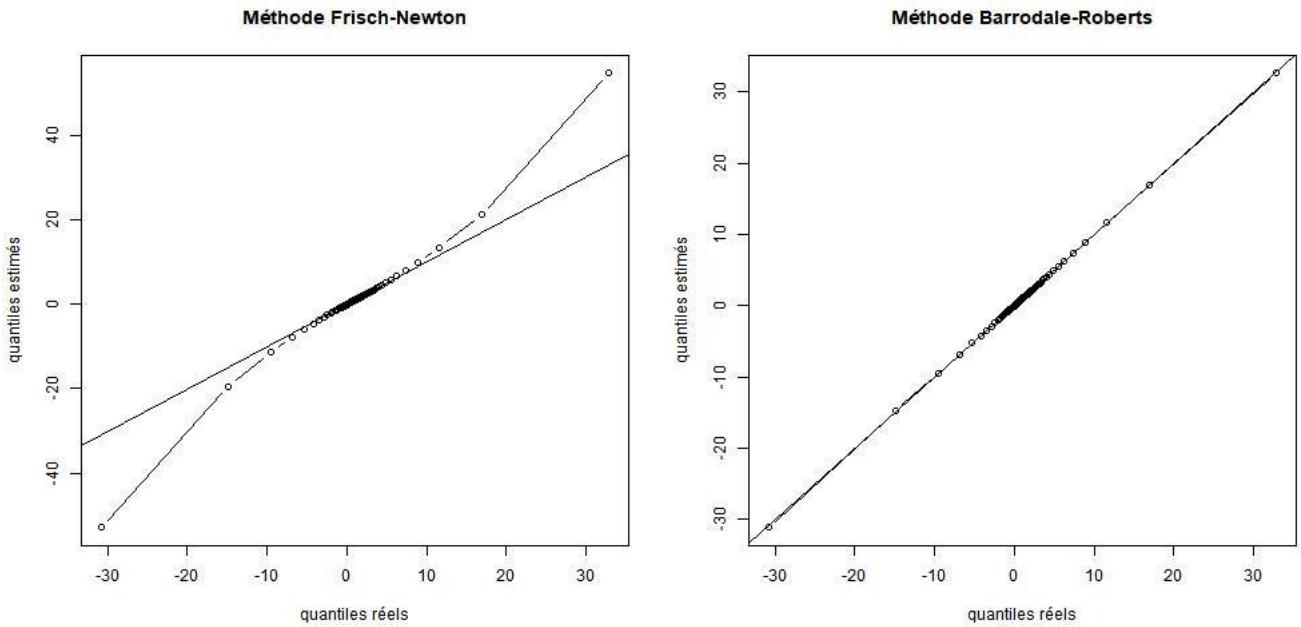


Figure 8: Frisch-Newton and Barrodale-Roberts estimators function of τ

IV.2 Results

These figures show point estimates, extremal 90% confidence intervals in blue, and normal 90% confidence intervals in red.

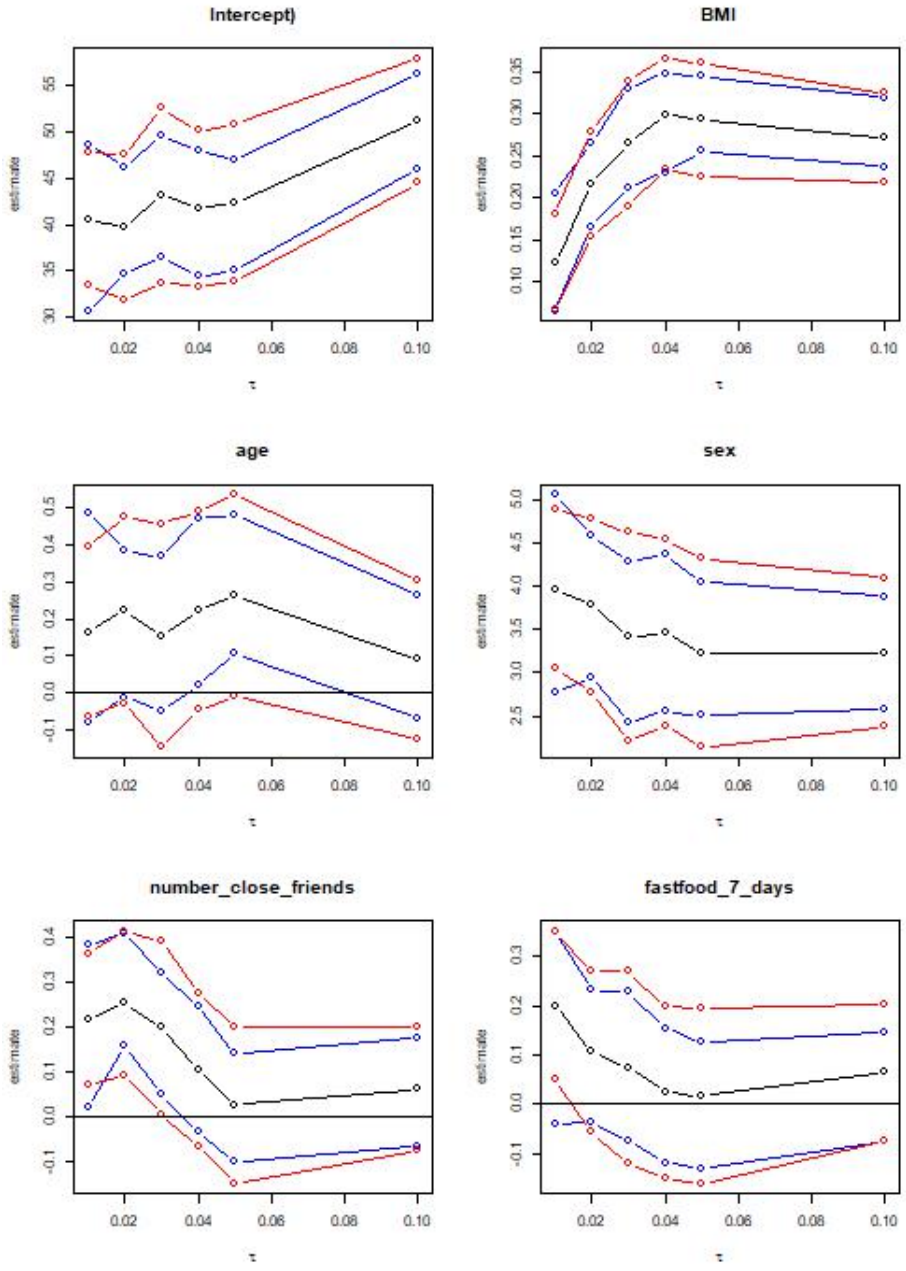


Figure 9: Regression for small quantiles

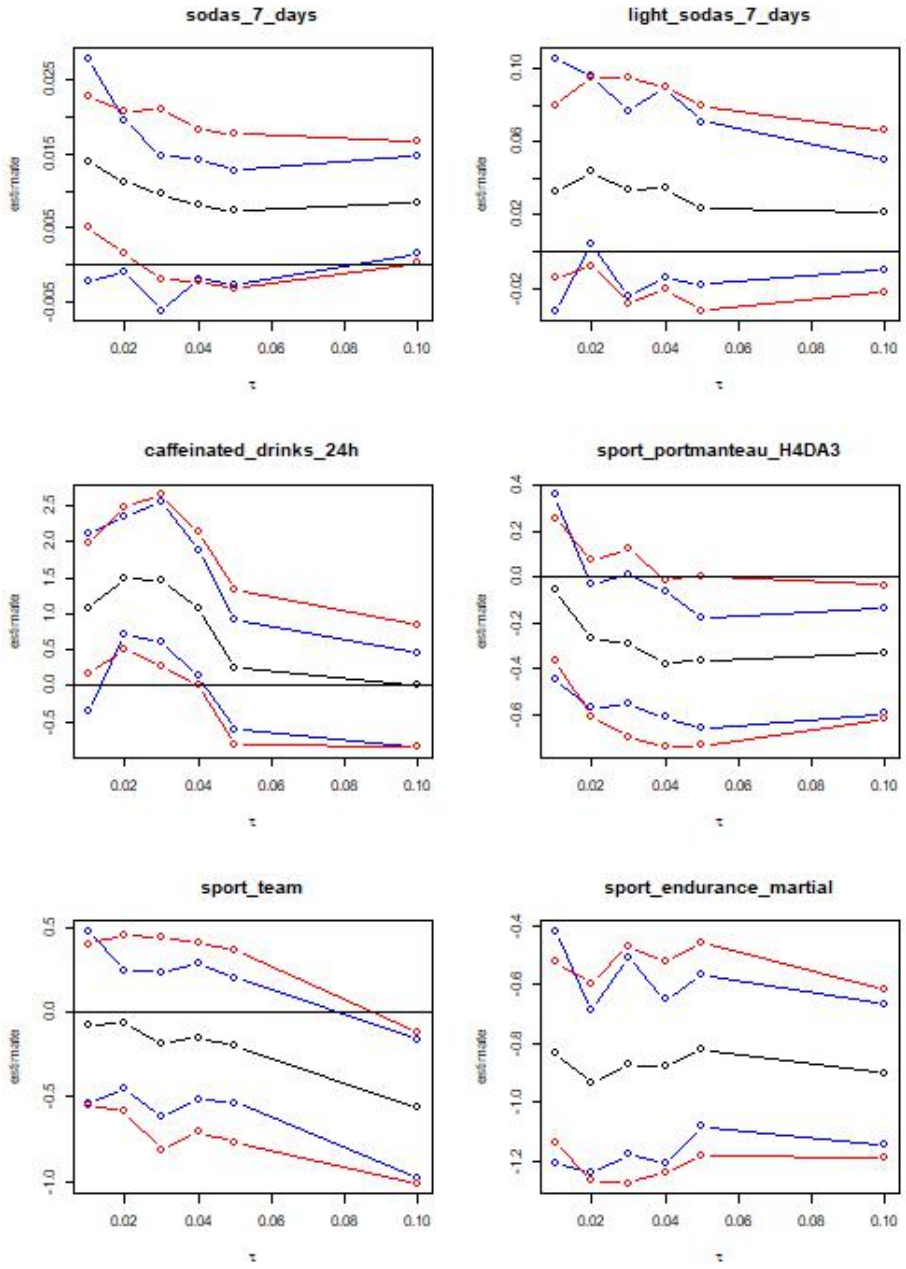


Figure 10: Regression for small quantiles

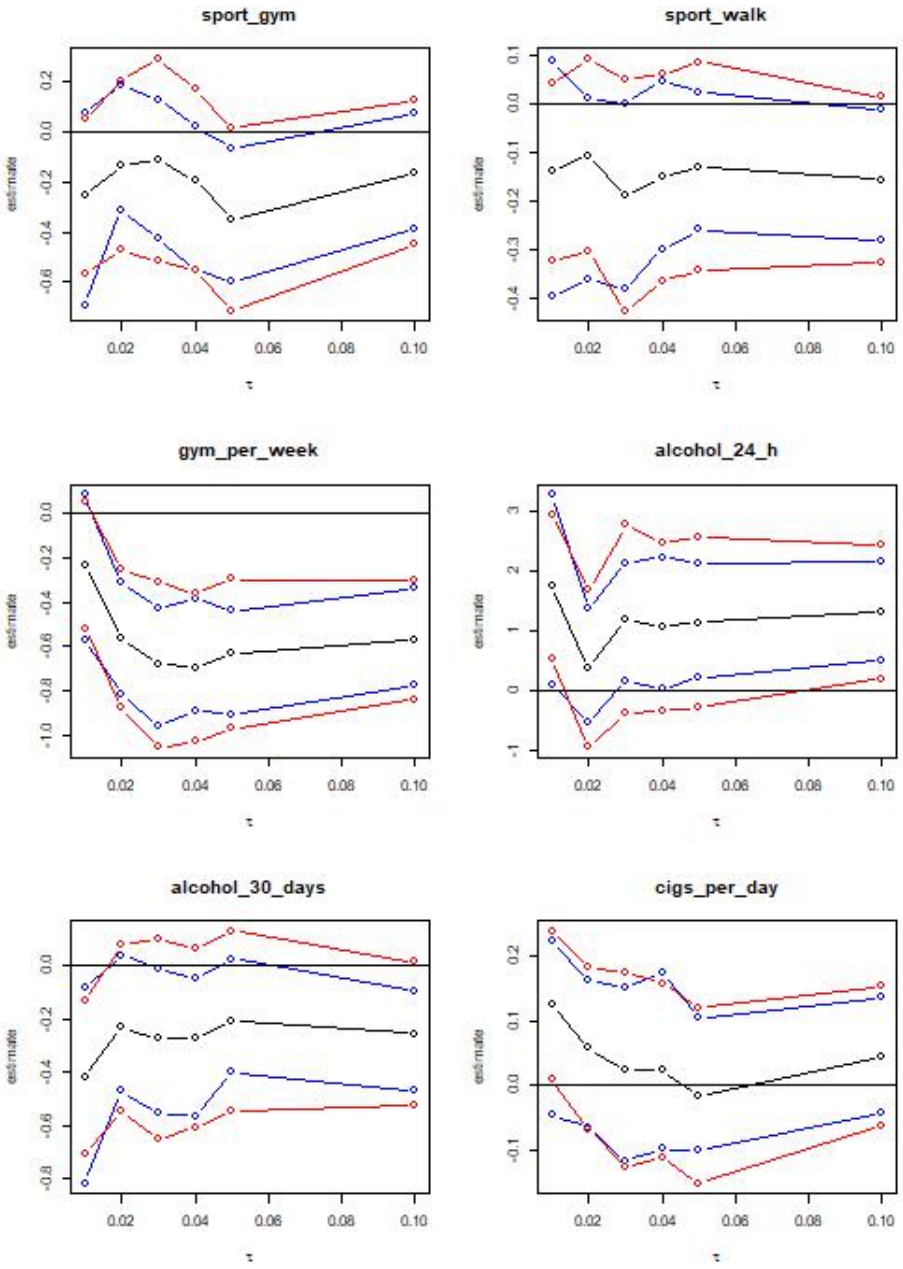


Figure 11: Regression for small quantiles

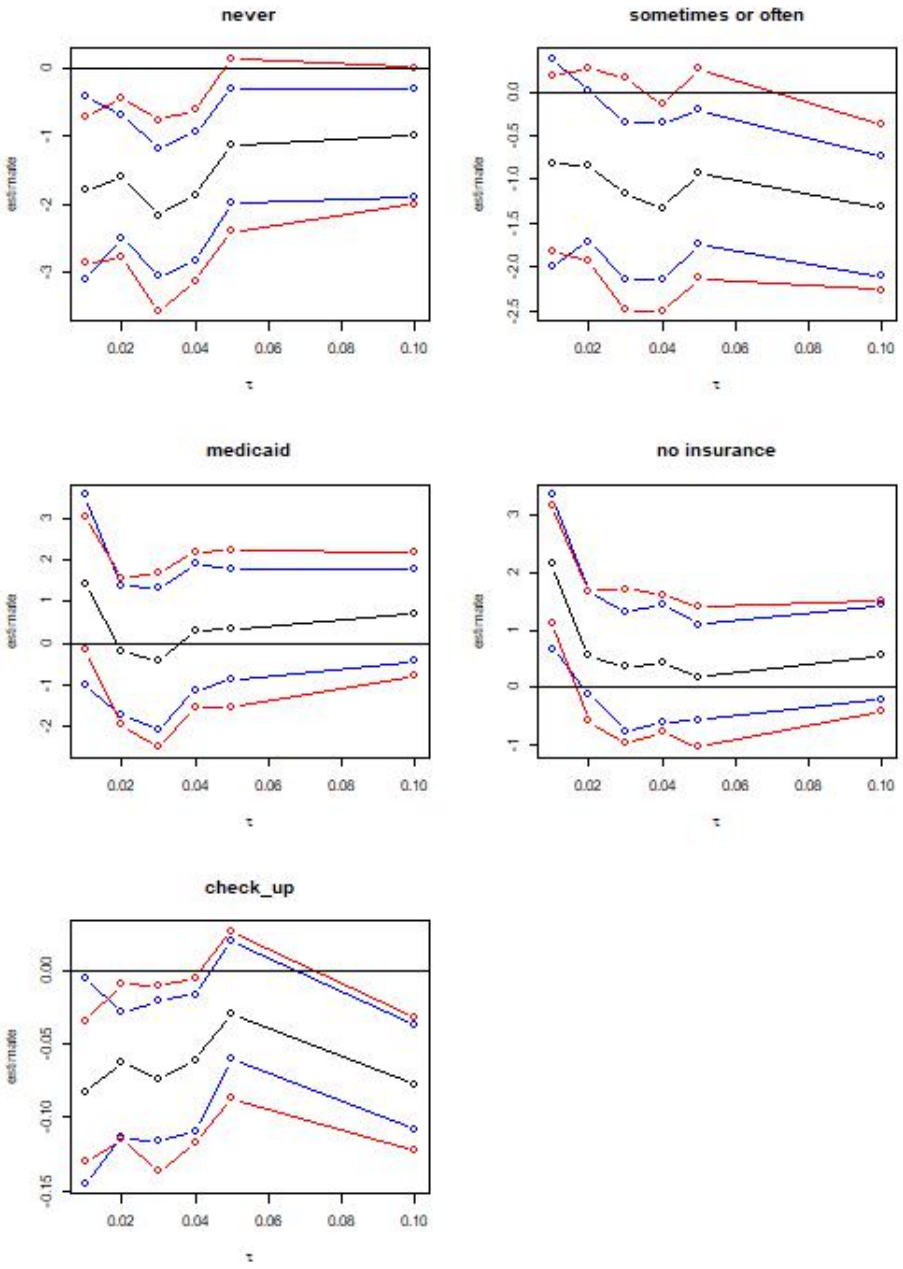


Figure 12: Regression for small quantiles

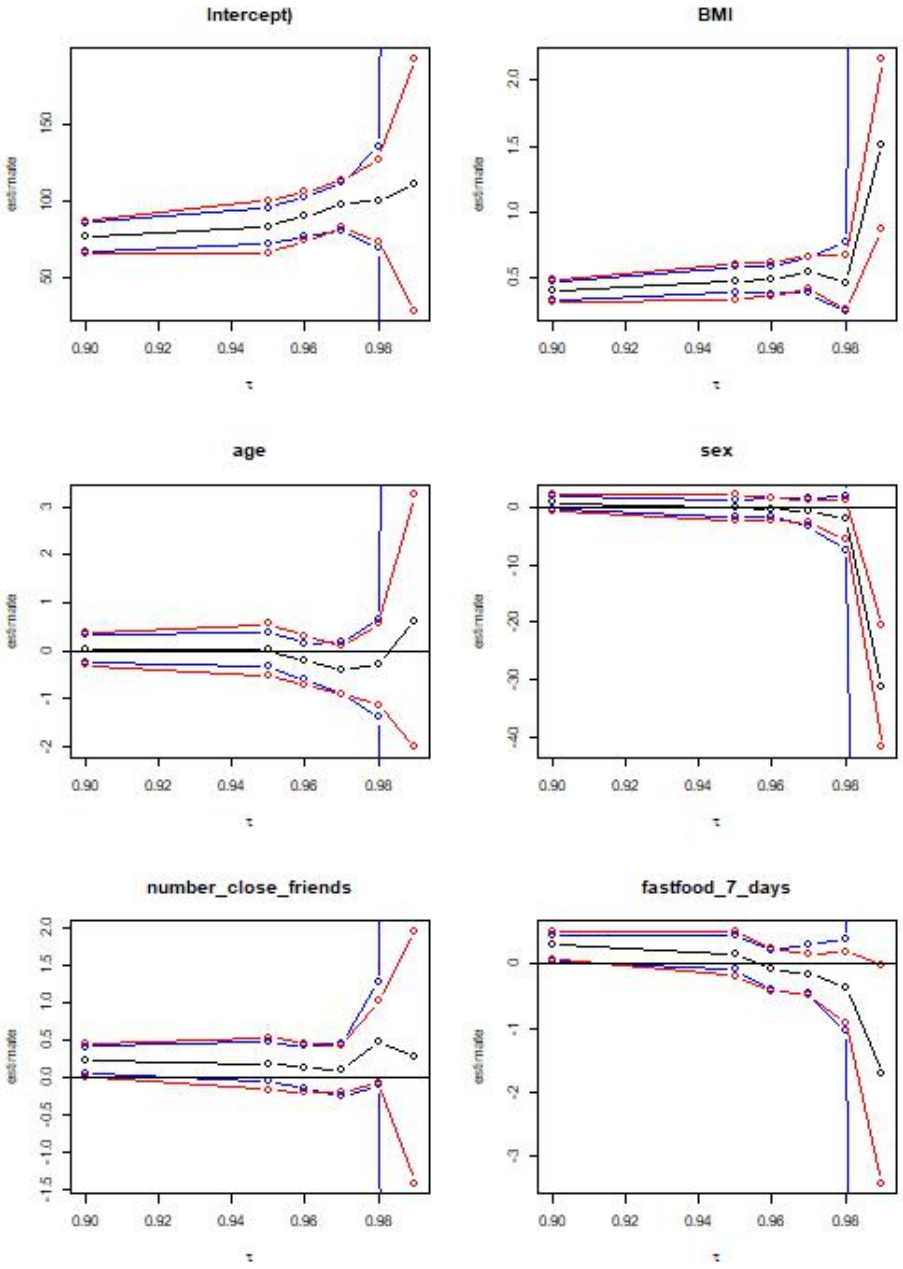


Figure 13: Regression for high quantiles

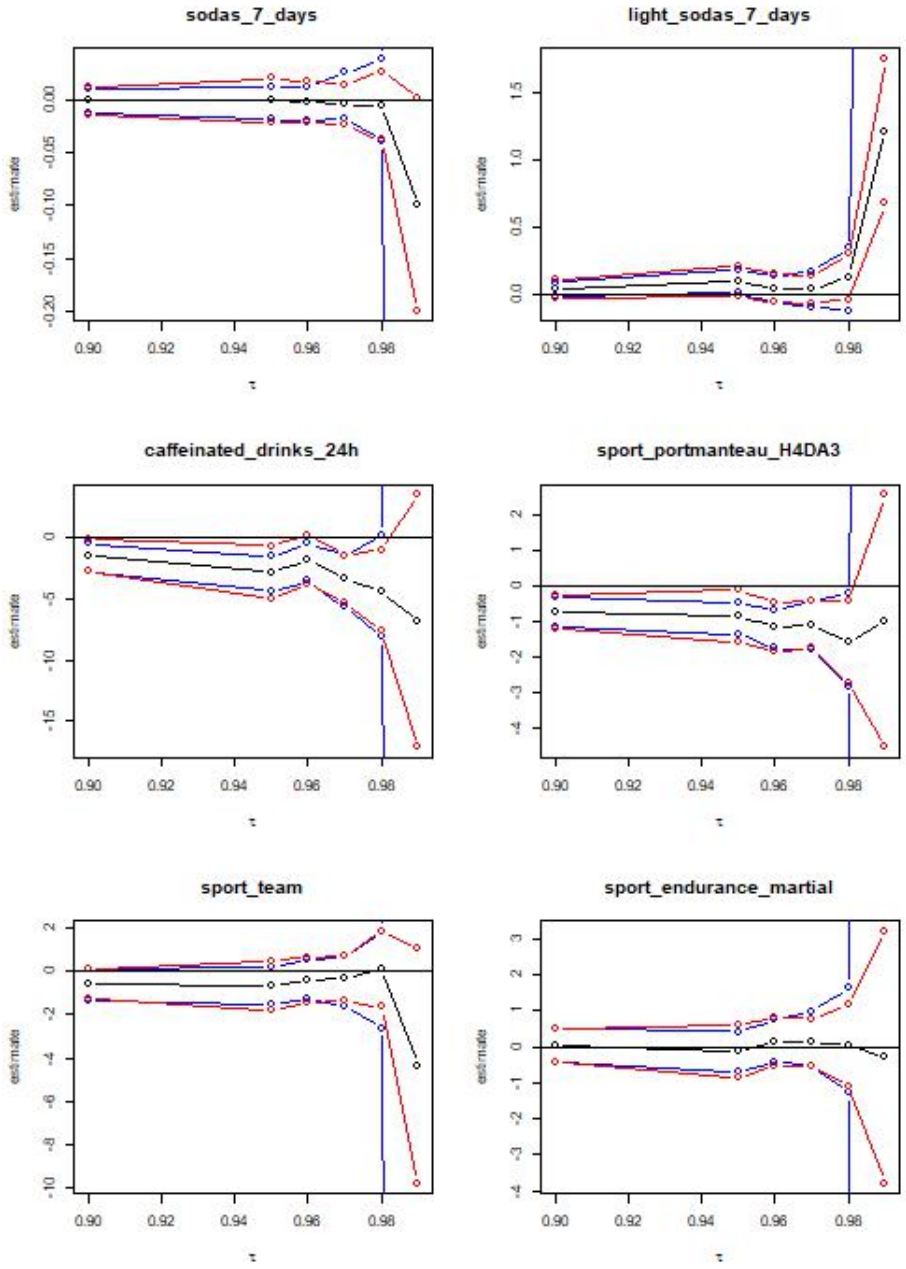


Figure 14: Regression for high quantiles

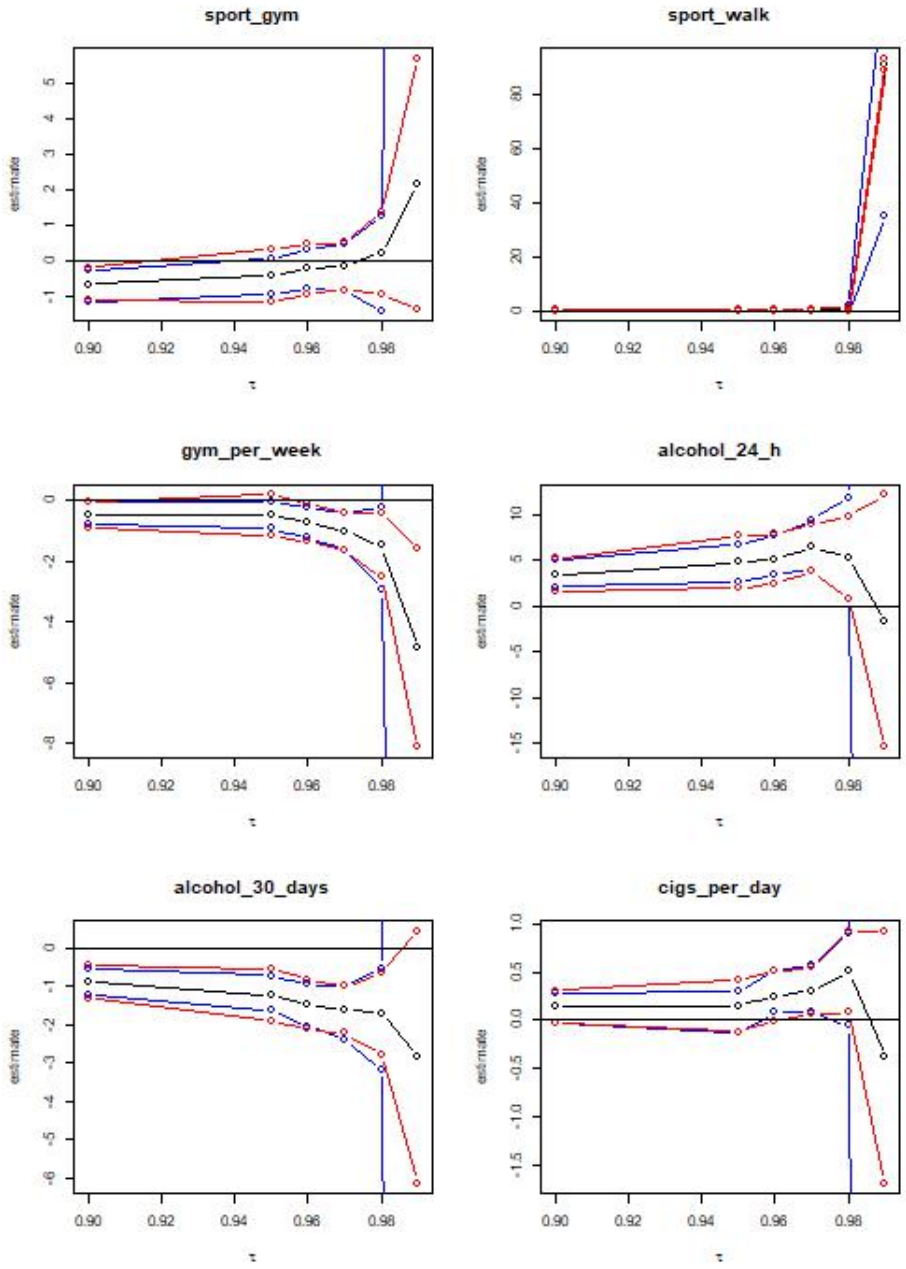


Figure 15: Regression for high quantiles

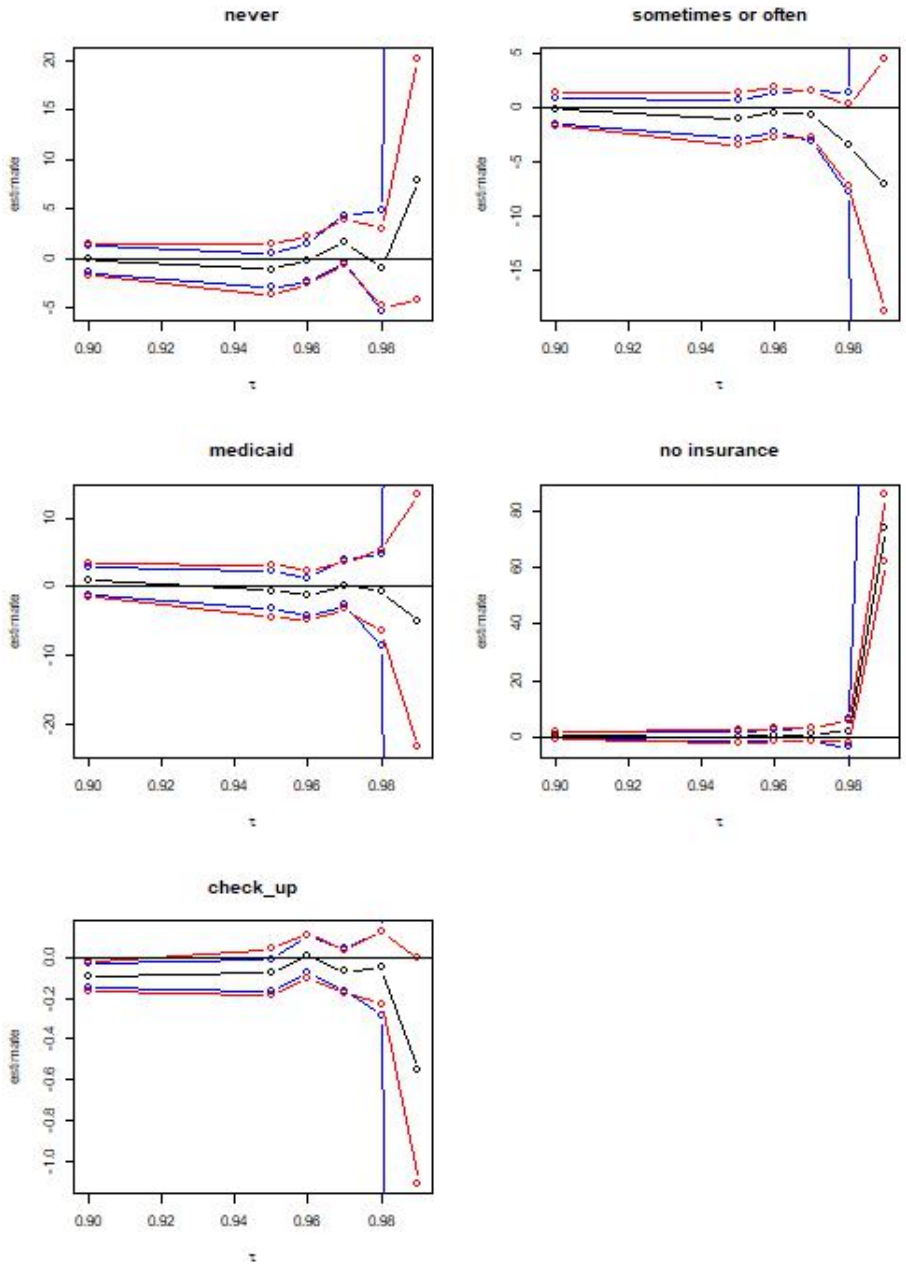


Figure 16: Regression for high quantiles