Statistics and Computing (2026) 36:66
https://doi.org/10.1007/s11222-025-10774-9

ORIGINAL PAPER

Global test for covariate significance in quantile regression

Tomáš Mrkvička[1] · Konstantinos Konstantinou[2] · Mikko Kuronen[3] · Mari Myllymäki[3]

Received: 28 March 2025 / Accepted: 8 November 2025

Abstract

Quantile regression is used to study effects of covariates on a particular quantile of the data distribution. Here we are interested in the question whether a covariate has any effect on the entire data distribution, i.e., on any of the quantiles. To this end, we treat all the quantiles simultaneously and consider global tests for the existence of the covariate effect in the presence of nuisance covariates. This global test for covariate significance in quantile regression can be used as the extension of linear regression or as the extension of distribution comparison in the sense of Kolmogorov-Smirnov test or as the extension of partial correlation. The proposed method is based on pointwise coefficients, permutations and global envelope tests. The global envelope test serves as the multiple test adjustment procedure controlling the family-wise error rate and provides the graphical interpretation which automatically shows the quantiles or the levels of categorical covariate responsible for the rejection. The Freedman-Lane permutation strategy showed liberality of the test for extreme quantiles, therefore we propose four alternatives that work well even for extreme quantiles and are suitable in different conditions. One of the strategies is suitable in a general situation, while others under more specific conditions. We show asymptotic exactness of the proposed permutation procedures and present a simulation study to inspect the performance of these strategies, and we apply the chosen strategies to two data examples.

Keywords Distribution comparison · Global envelope test · Multiple comparison problem · Permutation test · Significance testing · Simultaneous testing

1 Introduction

Quantile regression is used in many research fields to model the quantiles or full conditional distribution of the response variable rather than the mean and variance when assumptions of the ordinary linear model do not hold. The proposed statistical tool, the global test for covariate significance in quantile regression, can serve in three scenarios. First, when mean regression does not explain the dependence between the response and covariates satisfactory, the quantile regression may be used to explain this dependence for any quantile of the response variable. However, when the inference is made for several quantiles, it is usually done quantile-wise without correction for multiple testing. Such an inference leads to a multiple testing problem, which – if overlooked – can lead to erroneous conclusions. The global test for covariance significance solves this multiple testing problem and, therefore, it can be used to depict if there is the dependence of any quantile of the response variable on the covariates.

The second scenario, where the global quantile regression is of interest, is, when two or more distributions should be compared, as in the Kolmogorov-Smirnov test, but the distributions depend on extra nuisance covariates. In such a scenario, the data from different distributions are accompanied by a categorical covariate, and global test in quantile regression, where the interesting categorical covariate is

✉ Mari Myllymäki
mari.myllymaki@luke.fi

Tomáš Mrkvička
mrkvicka.toma@gmail.com

Konstantinos Konstantinou
konkons@chalmers.se

Mikko Kuronen
mikko.kuronen@luke.fi

[1] Department of Data Science and Computing Systems, University of South Bohemia, Studentská 1668, České Budějovice 37005, Czech Republic

[2] Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Chalmersgatan 4, Gothenborg 41296, Sweden

[3] Natural Resources Institute Finland (Luke), Latokartanonkaari 9, Helsinki 00790, Finland

Published online: 19 January 2026

supplemented with the nuisance covariates, can be used to investigate whether there is any difference between the distributions.

The third scenario of interest is where the partial correlation of two variables with the presence of nuisance variables is of interest. When partial correlation is computed, both parametrically and nonparametrically, it summarizes the dependence in one number. On the other hand, the proposed inference provides the information on dependence for any quantile through the multiple quantile regressions that are bound together in one global testing procedure with correct multiple testing control.

In the following, we first provide a motivational example, which shows the benefits of global test in quantile regression in the first scenario, and then review the current and proposed techniques as well as provide an outline for the rest of this paper.

1.1 Motivational example

We investigated the effect of the price of gold, oil, and uranium on the log returns of EUR/USD exchange rates in order to show different information that can be gained by using global test in quantile regression. To remove the effect of inflation from the prices, the prices of gold, oil, and uranium were computed as residuals of a simple exponential model that was fitted to the original prices. The data contains 3201 observations.

The top row of Figure 1 shows the common output of quantile regression for quantiles ranging from 0.01 to 0.99 (Koenker 2022). The red lines show the mean regression coefficients and the dashed red lines their confidence intervals. Since the mean confidence intervals cover 0 in all cases, the mean regression does not reveal any effect of any covariate on response. The corresponding p-values are 0.636, 0.285, and 0.539, respectively. The black lines show the quantile regression coefficients, and the gray area around them represents their pointwise confidence bands. Performing the inference by pointwise confidence bands obtained from multiple quantile regressions suggest that there is an effect of oil on lower and upper quantiles and that there can be effect of gold on upper quantiles and uranium on lower quantiles. Due to the multiple testing, we cannot trust these effects. The global test in quantile regression helps us to answer the question of dependence.

To account for the multiple testing problem, we applied the global tests for the three variables with RLS, RQ, and RQ permutation strategies, which will be introduced in Section 4. The result of the global quantile regression test is shown in the second row of Figure 1. The black lines show again the quantile regression coefficients and the gray area around 0 is the global envelope of the global test in quantile regression. The global envelope represents the acceptance region of the test. The coefficients outside the envelope are significant coefficients under the global significance level of 0.05. Thus the effects of oil (global p-value=0.0004) and uranium (global p-value=0.006) are confirmed, whereas the effect of gold is not (global p-value=0.48).

In quantile regression, the user is often interested in the estimation of the effect of a certain covariate together with its confidence interval plotted simultaneously for all quantiles (the top row of Figure 1). The confidence bands help to understand the analysis results. The proposed global test in quantile regression (the bottom row of Figure 1) offers the same level of information but at the global significance level. In our example, we can judge that the increase in oil price significantly reduces the volatility of log returns. For low quantiles, the coefficient of oil is significantly positive, and for high quantiles it is negative. The increase in uranium prices decreases the regression coefficients significantly only for low quantiles, meaning the increased possibility for a big fall in the exchange rate.

1.2 Current and proposed techniques

In this paper, we are interested in making inferences for all quantiles simultaneously, along with a graphical interpretation that could be used to determine for which quantiles the effect of a covariate is present. This global inference can be viewed as an extension of quantile regression: while quantile regression tests the effect of the covariate locally at a specific quantile, the proposed inference tests the effect globally for all quantiles. We propose to solve the studied problem of simultaneous inference for quantile regression by a method that consists of the estimation of pointwise regression coefficients first, then using a permutation strategy that provides the resampled pointwise regression coefficients under the null hypothesis, and finally applying the global envelope test as a solution for multiple testing and graphical interpretation. The global envelope test provides the family-wise error rate control of multiple tests (Myllymäki et al. 2017). Therefore, the proposed global test in quantile regression provides the same control. The methods for pointwise estimation of confidence intervals are summarized, e.g., by Koenker (2005) and implemented in the R package quantreg (Koenker 2022), with visualization.

The studied problem can also be solved by testing the effect of a covariate for all quantiles pointwisely by methods reviewed, e.g., in Koenker (1994), and applying a multiple correction method, e.g., the Holm-Bonferroni correction (Holm 1979) in order to solve the multiple testing problem. Also, recently, new methods for simultaneous confidence bands were developed analytically. For example Belloni et al. (2014) and (Koenker et al. 2018, Chapter15) discuss the simultaneous confidence bands for a quantile process $\boldsymbol{\beta}(\tau)$ on $[0, 1]$ based on asymptotic theory. These bands are however,

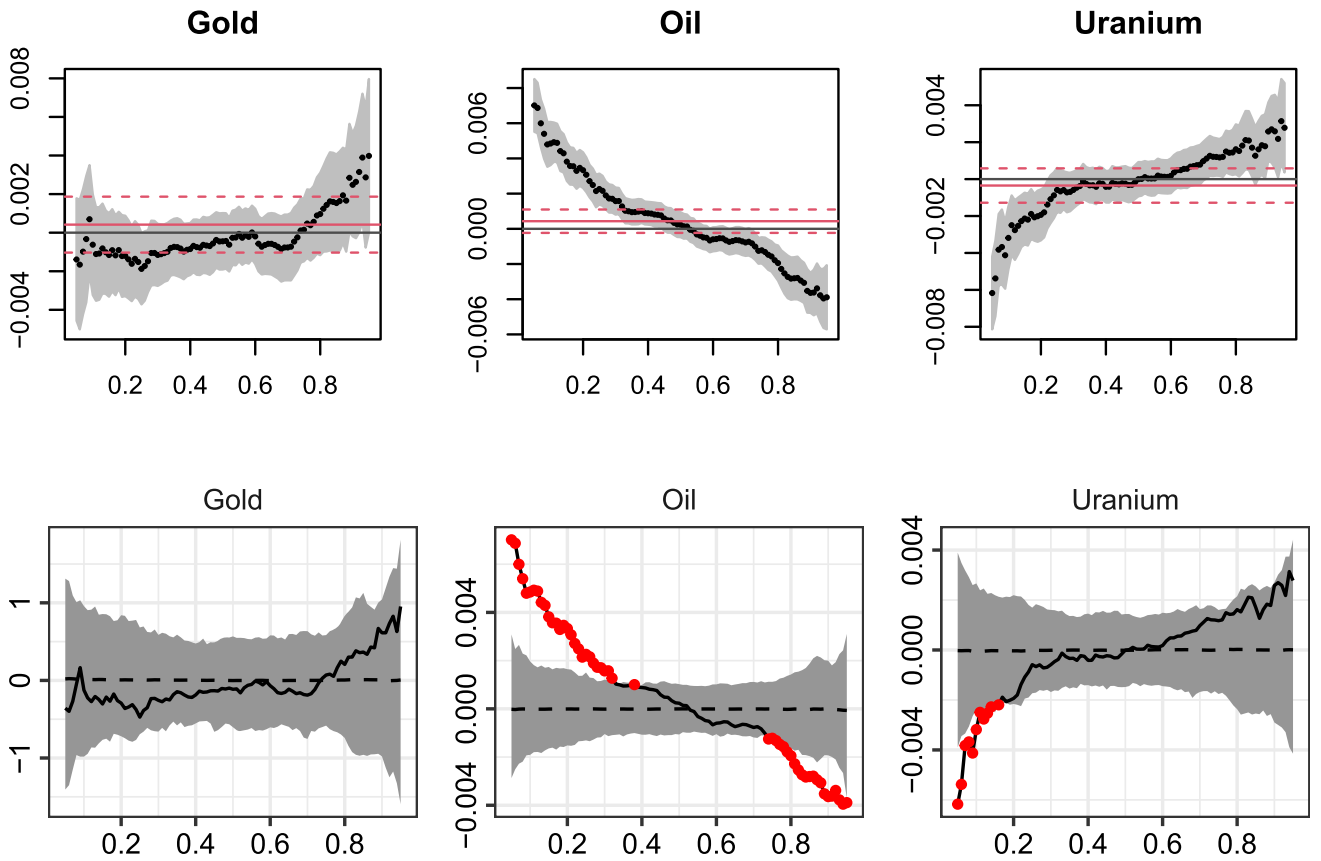


Fig. 1 95% pointwise confidence bands (gray bands, top row) and 95% global envelopes (gray bands, bottom row) for price of gold, oil or uranium as the interesting covariate and having all others as nuisances. The three global envelopes are based on 2499 permutations and the RLS, RQ or RQ permutation strategy, respectively

valid only under complex regularity conditions. On the other hand, Peng and Fine (2009) proposed a cumulative approach in order to summarize the covariates effect of all quantiles in one number. This can be used to deduce if the effect is globally significant, but it can not be used to infer which quantiles are significant. Another global problem was considered in Khmaladze (1982) and Koenker and Xiao (2002), namely the constantness of the effect of all covariates. This test is usually used to check the assumption of location shift (i.e., the effect of covariates for all quantiles is constant) or location-scale shift form (i.e., the covariates affect only mean and variance of the response distribution). This test cannot be used to globally test the significance of a covariate, due to the difference in null hypotheses. Also the global test in quantile regression models were recently developed, e.g., Zheng et al. (2015). These models concentrate on simultaneous model parameter estimation. On the contrary, we use local quantile regression models and bind them through a global testing procedure.

In order to achieve global inference for quantile regression, we rely on permutation methods in this paper. Cade and Richards (2006) used the Freedman-Lane (FL) permutation strategy (Freedman and Lane 1983) for the quantile regression. This strategy is regarded as the most precise method in testing a covariate effect of a univariate or functional linear models in the presence of nuisance covariates (Anderson and Robinson 2001; Anderson and Ter Braak 2003; Winkler et al. 2014). Cade and Richards (2006) also proposed an improvement of the FL procedure for quantile regression, which we also investigate in this paper. They used it for testing with a univariate test statistic which reflects the location or scale of the distribution only. Ditzhaus et al. (2021) proposed to use permutations for quantile regression, too, but they proposed only simple permutation of the data, i.e., the strategy of the one-way ANOVA problem (even thought this was applied for a factorial design of two-way ANOVA). Similarly as Cade and Richards (2006), they concentrated on univariate test statistics such as the median or interquartile distance.

Instead of the permutations, it is also possible to use nonparametric Bootstrap tests to achieve the desired global inference. (Davison and Hinkley 1997, p. 161) explains that the difference between these two approaches is that Bootstrap

allows for resampling with replacement, whereas permutations do not. That is due to the fact that a nonparametric Bootstrap test also has to resample from the null distribution here. Since the difference is small and Davison and Hinkley (1997) claims that there is also not much difference in the results of these two approaches, and since most of the literature relies on permutations, we will investigate here only the permutation approaches.

Here, we are interested in testing the effect for all quantiles at the global significance level α. We investigate the suitability of various permutation strategies for the given aim. It turns out that the FL permutation strategy does not perform well for global test in quantile regression, due to its liberality for extreme quantiles. Therefore, we propose several alternative permutation strategies, which perform better for extreme quantiles.

Our method for solving the problem of multiple testing is based on global envelope tests (Myllymäki et al. 2017; Mrkvička et al. 2022; Myllymäki and Mrkvička 2024) recently developed for spatial statistics and functional data analysis. This method allows to use a functional (or multivariate) test statistic and have the global significance level α. Besides, it allows us to draw the $100(1-\alpha)\%$ global envelope that represents the acceptance region under the null model of no effect of a certain covariate under the presence of other covariates. If the observed effect of the covariate is not fully contained in the global envelope, the test is significant at the global significance level α. Further, the test shows the quantiles which are the reason for a potential rejection of the null hypothesis, suggesting how the covariate affects the distribution of the response variable.

Since the global envelope test is based on ranks, it has no assumptions on the distribution of the functional test statistics, neither the homogeneity of the distribution of the test statistic along its domain. The only assumption is the exchangeability of the test statistic under the permutation strategy. That is, the global envelope test is exact, i.e., the type I error is precisely α, according to Lemma 1 in Myllymäki et al. (2017), if the permutations are exchangeable. Some of the studied permutation strategies fulfill the exchangeability, but some do not. For instance, the famous FL permutation strategy does not satisfy the exchangeability in the presence of nuisance covariates. Therefore, we study via a simulation study which of the alternative permutation strategies match the best the preset significance level and have the highest power. We also define a new permutation strategy that complies with the prescribed significance level in all cases.

Due to the nonparametric nature of the global envelope test, we can test continuous, categorical effects, interactions, and within the categorical effect also the differences between the groups via the joined functional test statistics (Mrkvička et al. 2017).

The rest of the paper is organized as follows. Section 2 gives the necessary background on quantile regression and global envelope tests. Section 3 explains the proposed global test. Section 4 describes the different permutation strategies to generate simulations under the null model of no effect of the interesting covariate and shows the theoretical properties of the proposed tests. The performance of the permutation strategies together with the global test is then investigated in Section 5. The permutational approaches are compared with the pointwise confidence band corrected by the Holm-Bonferroni correction only because the other solutions introduced in the literature require regularity conditions or provide only cumulative or summarized information about the problem. Section 6 applies the chosen tests to analyze another data set, in the second scenario mentioned above. Section 7 is for discussion of the results and extensions. The implementation of the proposed method is available in the `R` package `GET` (Myllymäki and Mrkvička 2024) (function `global_rq`). It can be downloaded from `CRAN` (https://cran.r-project.org/package=GET), together with a vignette for global test in quantile regression.

2 Notation and background

2.1 Linear quantile regression

Classical linear regression models focus on modeling the conditional expectation of a response variable $\mathbf{Y}$ given a set of covariates $\mathbf{X}$. In linear regression, the mean response is modeled as a linear combination of the regression parameters $\boldsymbol{\beta}$ and the covariates $\mathbf{X}$, i.e., $\mathbb{E}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, and estimation of the regression coefficients is performed by minimizing the sum of squared residuals. However, linear regression models are often insufficient either due to violations of the linear model's assumptions or due to the interest being in the tails of the distribution rather than its mean. Hence, analysis of covariate effects across the conditional distribution of the response variable requires more flexible statistical modeling than traditional linear regression only.

Quantile regression, introduced by Koenker and Bassett (1978), focuses on the modeling of the conditional quantiles of the response variable. That is, for any $\tau \in [0,1]$, the $\tau-$quantile of the conditional distribution of the response Y_i given a set of covariates $\mathbf{X}_i$,

$$\begin{aligned} Q_{Y_i|\mathbf{X}_i}(\tau) &= \inf\{y : F_{Y_i|\mathbf{X}_i}(y) \geq \tau\} \\ &= \mathbf{X}_i^T \boldsymbol{\beta}(\tau),\ i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $F_{Y_i|\mathbf{X}_i}$ is the conditional cumulative distribution function of Y_i given $\mathbf{X}_i$, and $\boldsymbol{\beta}(\tau)$ is the regression coefficient of the model for the τ-quantile. For instance, the quantile

regression for $\tau = 0.5$ defines the linear model for the conditional median, a robust alternative to the standard linear model.

Unlike classical linear regression, which has a closed formula for the estimator of the regression coefficients, estimating the parameters of quantile regression requires solving an optimization problem. The regression coefficients $\boldsymbol{\beta}(\tau)$ are estimated by minimizing the expected loss

$$\hat{\boldsymbol{\beta}}(\tau) = \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}) \tag{2}$$

where $\rho_\tau(u) = u\,(\tau - \mathbf{1}(u < 0))$, i.e., $\rho_\tau(u) = u\tau$ if $u \geq 0$ and $-u(1-\tau)$ if $u < 0$. The optimization problem in Equation (2) can efficiently be solved by linear programming methods (Dantzig 2016; Portnoy and Koenker 1997). We used the R library quantreg (Koenker 2022) for the estimation of $\hat{\boldsymbol{\beta}}(\tau)$.

2.2 Inference for quantile regression

Studying the effect of the covariates of interest on quantiles of the conditional distribution of the response requires inference of the quantile regression process $\boldsymbol{\beta}(\tau)$ on $[0, 1]$. In the literature, there exist three main approaches to construct confidence intervals for $\hat{\boldsymbol{\beta}}(\tau)$.

The first approach assumes that under some mild conditions, the estimated regression quantiles are asymptotically normal (Koenker 2005). Calculating the standard error requires the estimation of the so called sparsity function $s(\tau) = [f(F^{-1}(\tau))]^{-1} = \frac{d}{d\tau} F^{-1}(\tau)$, where f is a probability density function such that $f = F'$. The sparsity function $s(\tau)$ can be estimated by $\hat{s}_n(\tau) = [\hat{F}_n^{-1}(\tau + h_n) - \hat{F}_n^{-1}(\tau - h_n)]/2h_n$ where h_n is a bandwidth which tends to zero as $n \to \infty$ and needs to be selected, and $\hat{F}_n$ is the empirical cumulative distribution function, or by kernel smoothing. The estimator $\hat{s}_n(\tau)$ is unstable when the assumption that the errors are iid is violated. In the case with non-iid errors, a Huber estimate of the limiting covariance matrix needs to be computed (Koenker 1994). This case can be treated by assuming that $Q_{Y_i|\mathbf{X}_i}(\tau)$ is locally linear in $\mathbf{X}_i$ (Koenker and Machado 1999). For the remainder of the paper, we refer to this method as the "NID" method. The quantreg package recommends using the NID method for data with more than 1000 datapoints as this method is very fast (Chen and Wei 2005). On the contrary, this method is not ideal for small samples, as the methods for automatic bandwidth selection, for instance, the method in Hall and Sheather (1988), tend to give large bandwidths, which often result in violations of the local linearity assumption.

The second class of methods is the rank-score methods, which construct the confidence intervals by the inversion of the rank-score test (Gutenbrunner et al. 1993; Koenker 1994; Koenker and Machado 1999). The rank-score methods avoid the estimation of the sparsity function and are more robust to model assumptions. However, those methods require solving a parametric linear programming problem. Therefore, this approach is slow for large samples as its computational complexity is exponential in p and n (Chen and Wei 2005; Kocherginsky et al. 2005). In the quantreg package, the rank method is used by default for small samples ($n < 1000$).

The third method for constructing confidence intervals is based on resampling strategies (Efron 1979). Most common methods are based on bootstrapping the pairs of the response and explanatory variables (Hahn 1995) or bootstrapping the residuals (Bickel and Freedman 1981). In the residual bootstrap exchangeability of the residuals needs to be assumed. Recently, there have been a lot of research using bootstrap techniques for estimating standard errors in the quantile regression setting (Parzen et al. 1994; He and Hu 2002; Kleiner et al. 2014).

However, all the methods above concern local inference, but we are interested in simultaneous inference for $\boldsymbol{\beta}(\tau)$, $\tau \in [0, 1]$. In this paper, we propose an inference method using permutation-based global envelopes test. The proposed test is compared with the Holm-Bonferroni adjusted local NID test (see above).

Another important question in quantile regression is if the effect of all covariates can be considered constant for all quantiles. It was studied in Koenker and Xiao (2002). They proposed tests for the hypothesis that a linear model specification is of the location shift or location-scale shift form. The tests are based on the approach proposed by Khmaladze (1982).

2.3 Quantile regression for modeling distributions

There are tests to test differences between two distributions. The two-sample Kolmogorov-Smirnov test is maybe the most well-known. Here we only remark that the global test in quantile regression with a categorical predictor can also be used to solve the problem of finding the differences between the distributions (two or more), not only when the categorical predictor is the only covariate of the model but also in the presence of further nuisance covariates. The proposed global test in quantile regression can determine not only if there is a difference at the global significance level, but it can also determine which τs are responsible for the rejection.

2.4 Global envelope tests

Global envelope tests are non-parametric Monte-Carlo tests for multivariate or functional summary statistics (Myllymäki et al. 2017). Let $\mathcal{T} = (\tau_1, \ldots, \tau_d)$ be the vector of d discrete

values of quantile levels where the statistic is evaluated. Further, let $\mathbf{T}_0 = (T_{01}, \ldots, T_{0d}) = (T_0(\tau_1), \ldots, T_0(\tau_d))$ stand for the d-dimensional discretization of the empirical statistic and $\mathbf{T}_1, ..., \mathbf{T}_s$ be the corresponding statistics for s data sets simulated under the "null model". The tests are global in the sense that the test is performed simultaneously for all $\tau \in \mathcal{T}$, i.e., the family-wise error rate is controlled by the prespecified significance level α. The advantage of the global envelope test is that it allows for graphical interpretation of the test result by a global envelope that represents the acceptance region of the test: A $100(1-\alpha)\%$ global envelope is a band $(\mathbf{T}^\alpha_{\text{low}}, \mathbf{T}^\alpha_{\text{upp}})$ with $\mathbf{T}^\alpha_{\text{low}} = (T^\alpha_{\text{low},1}, \ldots, T^\alpha_{\text{low},d})$ and $\mathbf{T}^\alpha_{\text{upp}} = (T^\alpha_{\text{upp},1}, \ldots, T^\alpha_{\text{upp},d})$, constructed under the null model, such that the probability that $\mathbf{T}_0$ is completely within the envelope is equal to $1-\alpha$. Therefore, the empirical test statistic $\mathbf{T}_0$ goes outside the given $100(1-\alpha)\%$ global envelope for some τ if and only if the global test rejects the null hypothesis ($p < 0.05$). The τ's where $\mathbf{T}_0$ goes outside the envelope are responsible for the rejection of the test.

Global envelopes are constructed by ranking the statistics $\mathbf{T}_0, \ldots, \mathbf{T}_s$ based on a ranking measure E. The ranking is then used to identify the $\alpha(s+1)$ most extreme vectors. Examples of the ranking measures, which allows for one-to-one correspondence between formal and graphical results, are the extreme rank length measure (Narisetty and Nair 2016; Myllymäki et al. 2017), the continuous rank measure (Hahn 2015), and the area measure (Mrkvička et al. 2022). For a more rigorous description of the available ranking measures, you are referred to Myllymäki and Mrkvička (2024) and references therein. Now, let $E_i < E_j$ be interpreted as $\mathbf{T}_i(\tau)$ is more extreme than $\mathbf{T}_j(\tau)$ and let $E_{(\alpha)} \in \mathbb{R}$ be the largest E_i such that

$$\sum_{i=0}^{s} \mathbf{1}(E_i < E_{(\alpha)}) \leq \alpha(s+1)$$

and let $I_{(\alpha)}$ denote the set of vectors less than or as extreme as $E_{(\alpha)}$. Then, a $100(1-\alpha)\%$ global envelope based on the measure E is given by

$$\left(T^{(\alpha)}_{\text{low}\,k}, T^{(\alpha)}_{\text{upp}\,k}\right) = \left(\min_{i \in I_{(\alpha)}} T_{ik}, \max_{i \in I_{(\alpha)}} T_{ik}\right) \quad \text{for } k = 1, ..., d.$$

The validity of global envelope tests is independent of the distribution or potential inhomogeneity of the distribution of the test statistic along its domain. However, in order for the global envelopes to achieve desired type I errors, the test statistics $\mathbf{T}_0, ..., \mathbf{T}_s$ must be exchangeable. The exchangeability depends on the permutation strategy used to obtain the replications of the test statistic under the null model.

Any functional measure E can be used to rank the statistics $\mathbf{T}_0, ..., \mathbf{T}_s$, but only those which satisfies the one to one correspondence between formal results and their graphical interpretation represented by the global envelope are considered in this work.

3 Global test in quantile regression

Assume the quantile regression model

$$Q_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(\tau) = \mathbf{X}\boldsymbol{\beta}(\tau) + \mathbf{Z}\boldsymbol{\gamma}(\tau) \text{ for all } \tau \in \mathcal{T}, \tag{3}$$

where $Q_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(\tau) = (Q_{Y_1|\mathbf{X}_1,\mathbf{Z}_1}(\tau), \ldots, Q_{Y_n|\mathbf{X}_n,\mathbf{Z}_n}(\tau))$ is a $n \times 1$ vector of conditional τ-quantiles of $Y_1, \ldots, Y_n$, $\mathbf{X}$ is a $n \times p$ matrix of the interesting covariates, $\mathbf{Z}$ is a $n \times q$ matrix of nuisance covariates, $\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \ldots, \beta_p(\tau))$ and $\boldsymbol{\gamma}(\tau) = (\gamma_1(\tau), \ldots, \gamma_q(\tau))$ are the corresponding parameter vectors of dimensions $p \times 1$ and $q \times 1$, respectively, and $\mathcal{T} = \{\tau_1, \ldots, \tau_d\}$ is a discrete set of quantiles we are interested in. The null hypothesis of interest is

$$H_0 : \beta_j(\tau) = 0 \text{ for all } j = 1, \ldots, p \text{ and } \tau \in \mathcal{T}. \tag{4}$$

This null hypothesis is studied throughout the whole paper. Our aim is to construct a test with the family-wise error rate control for all β_j, $j = 1, \ldots, p$ and $\tau \in \mathcal{T}$, i.e., global quantile regression test of significance of covariates contained in $\mathbf{X}$. We propose the following strategy for this purpose:

Algorithm 1 Global inference for quantile regression (3) using permutation schemes

1. For observed data, compute the test vector

$$\mathbf{T}_0 = (\beta_1(\tau_1), \ldots, \beta_1(\tau_d), \ldots, \beta_p(\tau_1), \ldots, \beta_p(\tau_d)) \tag{5}$$

containing all the coefficients of the vectors

$$\boldsymbol{\beta}(\tau_1), \ldots, \boldsymbol{\beta}(\tau_d)$$

rearranged for better visualization.
2. Simulate s replicates of data under the null hypothesis (4).
3. Compute the test vectors for the s simulated data, and obtain $\mathbf{T}_1, \ldots, \mathbf{T}_s$.
4. Apply a global envelope test to $\mathbf{T}_0, \mathbf{T}_1, \ldots, \mathbf{T}_s$.

Global envelope testing provides a global p-value, the graphical interpretation that determines the τ's and the elements of the vector $\boldsymbol{\beta}$ that are responsible for the rejection in the global test (see the data study examples for a detailed description of the graphical interpretation). Since we observe all parameters in $\mathbf{T}_0$, we perform - simultaneously with the global test - a post-hoc test in cases when the covariate is categorical. This means that all levels of the categorical covariate are tested to have a different effect than the reference level. The generation of the data under the null hypothesis (4) is

a critical part of the test; in the following section, we will describe different alternatives for this purpose.

We remark here that the global envelope test produces the acceptance and rejection regions for the global null hypothesis, whereas usually the pointwise confidence intervals for the parameters of the model are obtained in quantile regression procedures.

4 Permutation strategies for quantile regression

In the following, we introduce six permutation strategies as candidates for producing simulations under the null hypothesis (4). We note that exchangeability of the test statistics $\mathbf{T}_0, \mathbf{T}_1, \ldots, \mathbf{T}_s$ is satisfied only for the permutation strategy for categorical covariates described in Section 4.3.

4.1 Freedman-Lane (FL)

Several approximate permutation methods have been proposed to test the significance of one or more regression coefficients in univariate and functional linear regression models for conditional means. Freedman-Lane procedure (Freedman and Lane 1983) has been found to be the method that is closest to being exact, i.e., reaching the nominal significance level (Anderson and Robinson 2001; Anderson and Ter Braak 2003). In the following, we explain how the replicates of data under the null hypothesis (4) are obtained in the Freedman-Lane permutation scheme. The general idea of the method is to permute the residuals of the reduced model, which does not contain the interesting covariates.

New data $\mathbf{Y}^*$ are generated by the following steps:

1. Fit the reduced model

$$Q_{\mathbf{Y}|\mathbf{Z}}(\tau) = \mathbf{Z}\boldsymbol{\gamma}(\tau) \text{ for all } \tau \in \mathcal{T} \tag{6}$$

to obtain the estimated coefficients $\widehat{\boldsymbol{\gamma}}(\tau)$.
2. Compute the residuals

$$\epsilon_i(\tau) = Y_i(\tau) - \mathbf{Z}_i^T \widehat{\boldsymbol{\gamma}}(\tau) \tag{7}$$

of the model (6) for $i = 1, \ldots, n$ and $\tau \in \mathcal{T}$.
3. Permute the rows of the $n \times d$ residual matrix $\boldsymbol{\epsilon}$ to produce the permuted residual matrix $\boldsymbol{\epsilon}^*$.
4. Construct the permuted data

$$\mathbf{Y}^*(\tau) = \mathbf{Z}\widehat{\boldsymbol{\gamma}}(\tau) + \boldsymbol{\epsilon}^*(\tau) \text{ for every } \tau \in \mathcal{T}, \tag{8}$$

where $\boldsymbol{\epsilon}^*(\tau)$ correspond to columns of $\boldsymbol{\epsilon}^*$.

4.2 Freedman-Lane with removal of zero residuals (FL+)

Cade and Richards (2006) suggested an enhancement to the permutation strategy of Freedman and Lane (1983) in the case of quantile regression. Their adjustment excludes from the permutations the zero residuals that are inherent in the quantile regression. That is, in step 4. of the Freedman-Lane simulation (see Section 4.1), for every τ, new permuted data $\mathbf{Y}^{**}(\tau)$ are constructed from the $\mathbf{Y}^*(\tau)$ of Equation (8) in the Freedman-Lane permutation by removing $q - 1$ elements corresponding to zero residuals. The new data $\mathbf{Y}^{**}(\tau)$ will have only $n - q + 1$ observations.

4.3 Within categorical nuisance (WN)

In the case that the quantile regression model (3) includes only categorical nuisance covariates, it is possible to employ simple permutations of the response variable within each level of the categorical nuisance covariates: Assume that there is a categorical nuisance covariate which has K levels. If there are more than one categorical nuisance covariate, every group of the first nuisance covariate can be decomposed into smaller groups according to the second nuisance covariate, etc. The decomposition then forms a new categorical covariate, say, with K levels. Because of the decomposition, the interactions of the nuisance factors are always present in the permutations. New data are in this case generated as follows:

1. Split the data into subsets based on the K levels of $\mathbf{Z}$. Let $(\mathbf{Y}^{(k)}, \mathbf{X}^{(k)}, \mathbf{Z}^{(k)})$, with $k = 1, \ldots, K$, be the K subsets.
2. Within each subset $k = 1, \ldots, K$, permute the elements of each $\mathbf{Y}^{(k)}$ to produce $\mathbf{Y}^*(k)$ and consequently $\mathbf{Y}^*$.

Note that under the null hypothesis (4), the distributions are equal inside each subset and, thus, the permutations of step 2 are exchangeable under (4).

4.4 Simple permutation with removal of the location effect of the nuisance covariates (RL)

In this permutation scheme, the mean effect of nuisance covariates is removed using a linear model and residuals of the fitted model are then permuted to simulate under the null hypothesis. We adjust Algorithm 1 for this procedure as specified in Algorithm 2.

Note here that due to the specificity of the quantile regression, it is necessary to always include the intercept in the interesting covariates in X in steps 2. and 4. of this algorithm. This also holds for the next two algorithms (Sections 4.5 and 4.6).

Algorithm 2 Global inference for quantile regression (3) with removal of the location effect of the nuisance covariates (RL).

1. Fit the mean linear model
$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}_Z.$$
2. Fit the quantile regression model for the residuals of the linear model from 1.,
$$Q_{\boldsymbol{\epsilon}_Z|\mathbf{X}}(\tau) = \mathbf{X}\boldsymbol{\beta}(\tau) \text{ for all } \tau \in \mathcal{T}. \quad (9)$$
The test vector $\mathbf{T}_0$ is specified according to Formula (5) from the estimated coefficients of the model (9).
3. Permute the residuals $\boldsymbol{\epsilon}_Z$ to obtain simulated data $\boldsymbol{\epsilon}^*_Z$. Repeat this s times.
4. Compute the test vectors for the s simulated data, and obtain $\mathbf{T}_1, \ldots, \mathbf{T}_s$.
5. Apply a global envelope test to $\mathbf{T}_0, \mathbf{T}_1, \ldots, \mathbf{T}_s$.

4.5 Simple permutation with removal of the location and scale effect of the nuisance covariates (RLS)

In this permutation scheme, the scaling of the residuals is added to Algorithm 2 in order to remove the scale of the nuisance effect. That is, the permutation scheme is as in the Algorithm 2 with step 1. replaced with

1' Fit the mean linear model $\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}'_Z$, then fit the mean linear model $\text{abs}(\boldsymbol{\epsilon}'_Z) = \mathbf{Z}\boldsymbol{\omega} + \boldsymbol{\epsilon}''_Z$. Set $\boldsymbol{\epsilon}_Z = \boldsymbol{\epsilon}'_Z/(\mathbf{Z}\boldsymbol{\omega})$.

4.6 Simple permutation with removal of the quantile effect of the nuisance covariates (RQ)

In this permutation scheme, effects of nuisance covariates are removed using a quantile regression model, and residuals of the fitted model are then permuted to simulate under the null hypothesis (4). The permutation scheme is as in Algorithm 2 with changing of steps 1. and 2. with

1'' Fit the quantile regression model
$$Q_{\mathbf{Y}|\mathbf{Z}}(\tau) = \mathbf{Z}\boldsymbol{\gamma}(\tau) \text{ for all } \tau \in \mathcal{T},$$
from where the residuals $\boldsymbol{\epsilon}_Z = (\boldsymbol{\epsilon}_Z(\tau_1), \ldots, \boldsymbol{\epsilon}_Z(\tau_d))$ are obtained.

2'' Consider d quantile regression models for the residuals $\boldsymbol{\epsilon}_Z(\tau_1), \ldots, \boldsymbol{\epsilon}_Z(\tau_d)$,
$$Q_{\boldsymbol{\epsilon}_Z(\tau_1)|\mathbf{X}}(\tau_1) = \mathbf{X}\boldsymbol{\beta}(\tau_1), \ldots, Q_{\boldsymbol{\epsilon}_Z(\tau_d)|\mathbf{X}}(\tau_d) = \mathbf{X}\boldsymbol{\beta}(\tau_d). \quad (10)$$
Compute the test vector $\mathbf{T}_0$ according to Formula (5) from the estimated coefficients of the models (10).

In this permutation scheme, similarly to the FL+ scheme, the different data are used for different τ's, but the permutations are kept the same.

All the permutation strategies fit one quantile regression for each quantile and each permutation.

4.7 Theoretical results

In this section, we first show when the proposed methods are exact, i.e., achieve the prescribed level α and then when they are asymptotically exact.

Definition 1 The test statistics $(\mathbf{T}_0, \ldots, \mathbf{T}_s) \in S^{s+1}$ are called exchangeable, if their joint probability distribution is permutation invariant, i.e.,
$$Pr\{(\mathbf{T}_0, \ldots, \mathbf{T}_s) \in A\} = Pr\{(\mathbf{T}_{\sigma(0)}, \ldots, \mathbf{T}_{\sigma(s)}) \in A\}$$
for any measurable set $A \subset S^{s+1}$ and any permutation σ.

The following Theorem is a reformulation of Lemma 1 proven in Myllymäki et al. (2017).

Theorem 1 *Let $(\mathbf{T}_0, \ldots, \mathbf{T}_s) \in S^{s+1}$ be exchangeable multivariate test statistic, i.e. $S = \mathbb{R}^d$, and let E be an unequivocal functional ranking measure, i.e.,*
$$Pr\{E(\mathbf{T}_i) < E(\mathbf{T}_j) \text{ or } E(\mathbf{T}_j) < E(\mathbf{T}_i)\} = 1 \quad \forall i \neq j$$
and it assigns smaller values to the more extreme vector from the given set of vectors. Then the Monte Carlo test with $p = \frac{1}{s+1}(1 + \sum_{i=1}^{s} \mathbf{1}(E(\mathbf{T}_i) < E(\mathbf{T}_0))$ rejects the null hypothesis at the prescribed significance level α, provided that $\alpha(s+1)$ is an integer.

Corollary 2 *Let the multivariate measure E be the cont or area measure defined in (Myllymäki and Mrkvička 2024, AppendixA). Let the global test in quantile regression be performed 1) without nuisance covariates or 2) with categorical nuisance covariates and the permutation strategy WN. Then the global test in quantile regression is exact.*

Proof The cont and area measures are unequivocal multivariate ranking measures. See their definition in Myllymäki and Mrkvička (2024). The global test in quantile regression is a Monte Carlo test with multivariate test statistic $\mathbf{T}_0$ defined in (5). Since the setting of the global test in quantile regression is with no nuisance or categorical nuisance with WN permutation strategy, the vector $(\mathbf{T}_0, \ldots, \mathbf{T}_s)$ is exchangeable under the null hypothesis (4). Thus, due to Theorem 1 the test is exact.

Definition 2 The sequence of test statistics $(\mathbf{T}^n_0, \ldots, \mathbf{T}^n_s) \in S^{s+1}$ is called asymptotically exchangeable for $n \to \infty$

if their limiting joint probability distribution is permutation invariant, i.e.,

$$\lim_{n\to\infty} Pr\{(\mathbf{T}_0^n,\ldots,\mathbf{T}_s^n)\in A\} = \lim_{n\to\infty} Pr\{(\mathbf{T}_{\sigma(0)}^n,\ldots,\mathbf{T}_{\sigma(s)}^n)\in A\} \quad (11)$$

for any measurable set $A\subset S^{s+1}$ and any permutation σ.

Lemma 3 *If a sequence $(\mathbf{T}_0^n,\ldots,\mathbf{T}_s^n)$ converges in distribution to limit $(\mathbf{T}_0,\ldots,\mathbf{T}_s)$ with a continuous and exchangeable distribution, then the sequence is asymptotically exchangeable.*

Proof Let A be a measurable set. Since distribution of the limit is continuous, A is its continuity set. Thus

$$\lim_{n\to\infty} Pr\{(\mathbf{T}_0^n,\ldots,\mathbf{T}_s^n)\in A\} = Pr\{(\mathbf{T}_0,\ldots,\mathbf{T}_s)\in A\} = Pr\{(\mathbf{T}_{\sigma(0)},\ldots,\mathbf{T}_{\sigma(s)})\in A\} = \lim_{n\to\infty} Pr\{(\mathbf{T}_{\sigma(0)}^n,\ldots,\mathbf{T}_{\sigma(s)}^n)\in A\}$$

for all measurable sets A.

Theorem 4 *Let $(\mathbf{T}_0^n,\ldots,\mathbf{T}_s^n)\in S^{s+1}$ be asymptotically exchangeable multivariate test statistics and let E be an unequivocal multivariate ranking measure. Then the sequence of Monte Carlo tests with $p^n = \frac{1}{s+1}(1+\sum_{i=1}^s \mathbf{1}(E(\mathbf{T}_i^n) < E(\mathbf{T}_0^n))$ is asymptotically exact, i.e.,*

$$\lim_{n\to\infty} \mathbb{E}\{\mathbf{1}(p^n\le\alpha)\} = \alpha,$$

provided that $\alpha(s+1)$ is an integer.

Proof Let $A_j^n = 1+\sum_{i=0,i\neq j}^s \mathbf{1}(E(\mathbf{T}_i^n) < E(\mathbf{T}_j^n)$ be the rank of the multivariate statistic $\mathbf{T}_j^n$ among the $\mathbf{T}_i^n$s. The unequivocal condition implies that $(A_0^n,\ldots,A_s^n)$ is a permutation of $(1,\ldots,s+1)$ almost surely. Then, for $j=0,\ldots,s$,

$$\lim_{n\to\infty} Pr(A_j^n = k) = \frac{1}{s+1} \text{ for } k=1,\ldots,s+1,$$

because the ranks $A_0^n,\ldots,A_s^n$ are asymptotically exchangeable. Immediately, this implies that

$$\lim_{n\to\infty} \mathbb{E}\{\mathbf{1}(p^n\le\alpha)\} = \lim_{n\to\infty} Pr\{A_0^n\le\alpha(s+1))\} = \frac{\alpha(s+1)}{s+1} = \alpha.$$

Corollary 5 *Let the multivariate measure E be cont or area measure. Assume that the response and interesting variables are continuous. Let the global test in quantile regression be performed with*

1. *RL permutation strategy and the effect of nuisance covariates is only on mean or*
2. *RQ permutation strategy.*

Then the global test in quantile regression is asymptotically exact.

Proof Under true nuisance parameters, steps 2-5 of Algorithm 2 reduce to the test without nuisance covariates, which is exact according to Corollary 2 and its test statistics $\mathbf{T}_0^{k,\infty},\ldots,\mathbf{T}_s^{k,\infty}$ are exchangeable. Here k denotes the number of data used in steps 2-5 and ∞ is used to express that step 1 of Algorithm 2 is estimated with an infinite number of data. The global test in quantile regression described in Algorithm 2 first removes the effect of nuisance covariates using linear regression estimator for every τ. This estimator is strongly consistent, i.e., $\gamma^n(\tau)$ converges almost surely to the true parameter $\gamma(\tau)$. We next show that the quantile regression estimates $\beta^k(\gamma^n)$ computed from the fixed number of data k used in steps 2-5 and n number of data used in step 1 converge in distribution to $\beta^k(\gamma)$ as $n\to\infty$ (i.e., the test statistics $\mathbf{T}_0^{k,n},\ldots,\mathbf{T}_s^{k,n}$ converge to $\mathbf{T}_0^{k,\infty},\ldots,\mathbf{T}_s^{k,\infty}$).

The score function $s(\epsilon,\boldsymbol{\beta}) = \sum_{i=1}^k \rho_\tau(\epsilon_i - \mathbf{X}_i^T\boldsymbol{\beta})$ is k-Lipschitz continuous with respect to the response ϵ. Thus, the sequence of functions $\boldsymbol{\beta}\mapsto s(\epsilon^n,\boldsymbol{\beta})$ converges uniformly to $\boldsymbol{\beta}\mapsto s(\epsilon^\infty,\boldsymbol{\beta})$, where ϵ^n is the residual computed from n data (i.e., with γ^n) and ϵ^∞ is the residual computed using $\gamma^\infty=\gamma$. Since we assume that response and interesting variables are continuous, the probability that the quantile regression computed in steps 2-5 of Algorithm 2 has a unique solution is one (Koenker 2005, p. 31). Since $s(\epsilon,\boldsymbol{\beta})$ is convex, it follows that the minimum of the score function is well separated (cf. van der Vaart 2007, Theorem 5.56). Since all other assumptions of Theorem 5.56 of van der Vaart (2007) are also satisfied in our case, $\beta^k(\gamma^n)$ converges in distribution to $\beta^k(\gamma)$. Because the distribution of $\beta^k(\gamma)$ is continuous, then asymptotic exchangeability follows from Lemma 3 in the first situation (RL) of this Corollary. In the second situation (RQ), the strong consistency of quantile regression (Bassett and Koenker 1986) is used instead of the strong consistency of linear regression. Because the cont and area measures are unequivocal multivariate ranking measures, asymptotic exactness follows from Theorem 4.

Remark 1 The above theorem was proven for continuous response and interesting variables. This continuity assumption is needed to obtain the uniqueness of the quantile regression $\beta^k(\gamma)$. (Koenker 2005, p. 31) discuss the existence and uniqueness of solutions in more detail, noting that the quantile regression may achieve multiple solutions for a discrete interesting variable. However, the simulation study below shows that the proposed methods work approximately also in such cases.

Remark 2 Often, the extreme rank length measure (ERL) is used in the global envelope test. This measure can theoretically achieve ties, thus the above results do not hold for this measure. Nevertheless, practically there are no ties and therefore the above results hold for this measure approximately, as it is shown in the simulation study.

The asymptotic exactness was not shown for RLS procedure, since in step 1' the effect of nuisance covariates on variability is estimated from absolute residuals. Therefore, this procedure should be taken as an approximation.

The above results showed asymptotic behavior of the proposed tests. In the next Section, we will explore the finite sample behavior of all proposed procedures.

5 Simulation study

We assumed the quantile regression model (3) and studied the performance of the global test for the hypothesis (4) under different permutation schemes (see Table 1). The performance was investigated in terms of power and type I errors. Additionally, the permutation-based methods were also compared with Holm-Bonferroni corrected p-values obtained using the NID method as implemented in the quantreg package as well as the minimum pointwise p-value without any correction.

In this section, we first show how well the procedures can distinguish the differences between the two distributions with a nuisance covariate when they differ in the tails. It is an instance of the second possible usage of a global test in quantile regression. Since the FL method gives almost equivalent results to the FL+ method, we show results only for FL+. These methods showed surprisingly high liberality, and therefore we further studied what is the reason for this liberality. It turned out that the reason is the high liberality of FL methods applied to single extreme quantiles. Therefore, we added into the study also the FL methods without extreme quantiles. Further, we studied how well the procedures can distinguish the differences between the two distributions with a nuisance covariate when they differ in scale or in the shape of the distribution. Finally, we study how the correlation between nuisance and interesting covariate affects the type I error rate of different permutation procedures.

In each experiment, the interesting covariate X influences the distribution of the response variable Y. In addition, the nuisance covariates Z and Z_1 affect the response distribution. We considered three different nuisance effects, namely location shift, location-scale shift, and shape shift effects. To investigate the validity of the permutation strategies in case of model misspecification, we designed scenarios where the underlying assumptions of the permutation strategy are not met. For instance, we considered the performance of the permutation strategy based on the nuisance location shift assumption, when the nuisance affects the shape of the response distribution.

Our observations consist of realizations of X, Z, Z_1 and Y from their corresponding distributions. In all tests below, unless otherwise specified, we used the following choices:

- All the global envelope tests (first six tests of Table 1) were based on 1000 permutations.
- We considered 10 equally spaced quantiles τ varying from 0.01 up to 0.99.
- In addition, we considered the variation of FL+ method also with 10 quantiles τ but on the interval from 0.1 to 0.9. This variation did not consider the extreme quantiles τ and it was denoted by an asterisk(*) in the figures.

We performed the first set of experiments as in Section 5.1 also with 100 equally spaced quantiles τ varying from 0.01 to 0.99. The results were similar to those with 10 τ values with respect to their significance level, except for the PH procedure. (The NC method was not included to the experiment.) The PH procedure had lower empirical significance levels with 100 τ values than with 10 τ values: it was conservative in the cases where it was exact for 10 τ values, but it persisted to be liberal in cases where it was liberal for 10 τ values. Therefore and for the reason of faster computing time, we present below the results only for the case of 10 τ values as specified above.

5.1 Sensitivity to differences in the tails of the distributions

In the first two simulation experiments, X was categorical with two levels and the two distributions corresponding to the levels of X differed in the tails. For the nuisance covariate, we considered different alternatives. It was either categorical or continuous. In Experiment (I), it affected either the location or location and scale of the response distribution, while in Experiment (II) we considered a "noise" nuisance covariate affecting the shape of the response distribution. More precisely, in Experiment (I),

$$\begin{cases} X \sim \text{Bernoulli}(0.5) \\ Y' \mid X \sim \begin{cases} N(0,1) & \text{if } X = 0 \\ t_4 & \text{if } X = 1 \end{cases} \\ Z \sim F_Z \\ Y = (1 + aZ)Y' + bZ \end{cases} \tag{I}$$

where $a, b \in \mathbb{R}$ and F_Z is the distribution of the nuisance variable for which we considered the following four alternatives:

Table 1 Description and abbreviations of the tests investigated in the simulation study. The first six methods are based on global test in quantile regression (GQR) with different permutation strategies

Test description	Abbreviation
GQR using the Freedman-Lane permutation	FL
GQR using the extension of the Freedman-Lane permutation	FL+
GQR using the permutation that removes the location nuisance effect	RL
GQR using the permutation that removes the location-scale nuisance effect	RLS
GQR using the permutations for categorical nuisance	WN
GQR using the permutation that removes the quantile nuisance effect	RQ
Pointwise p-values adjusted using Holm-Bonferroni method	PH
Minimum pointwise p-value	NC

(Ia) Continuous Z with effect on the location, $F_Z = \text{Unif}(0, 1.5)$, $a = 0, b = 1$
(Ib) Continuous Z with effect on the location and the scale, $F_Z = \text{Unif}(0, 1.5)$, $a = 1, b = 1$
(Ic) Categorical Z with effect on the location, $F_Z = \text{Bernoulli}(0.5)$, $a = 0, b = 0.1$
(Id) Categorical Z with effect on the location and the scale, $F_Z = \text{Bernoulli}(0.5)$, $a = 0.1, b = 0.1$.

In Experiment (II),

$$\begin{cases} X \sim \text{Bernoulli}(0.5) \\ Y' \mid X \sim \begin{cases} N(0,1) & \text{if } X = 0 \\ t_4 & \text{if } X = 1 \end{cases} \\ Z \sim F_Z \\ Z_1 \sim \text{Unif}(0, 1.5) \\ \epsilon \sim N(1, 0.04) \\ Y = \begin{cases} \epsilon & \text{if } Z < Z_1 \\ Y' & \text{otherwise} \end{cases} \end{cases} \tag{II}$$

where both Z and Z_1 are nuisance covariates and F_Z is the distribution of the nuisance covariate Z with the following two alternatives:

(IIa) Continuous Z with $F_Z = \text{Unif}(0, 1)$,
(IIb) Categorical Z with $F_Z = \text{Bernoulli}(0.5)$.

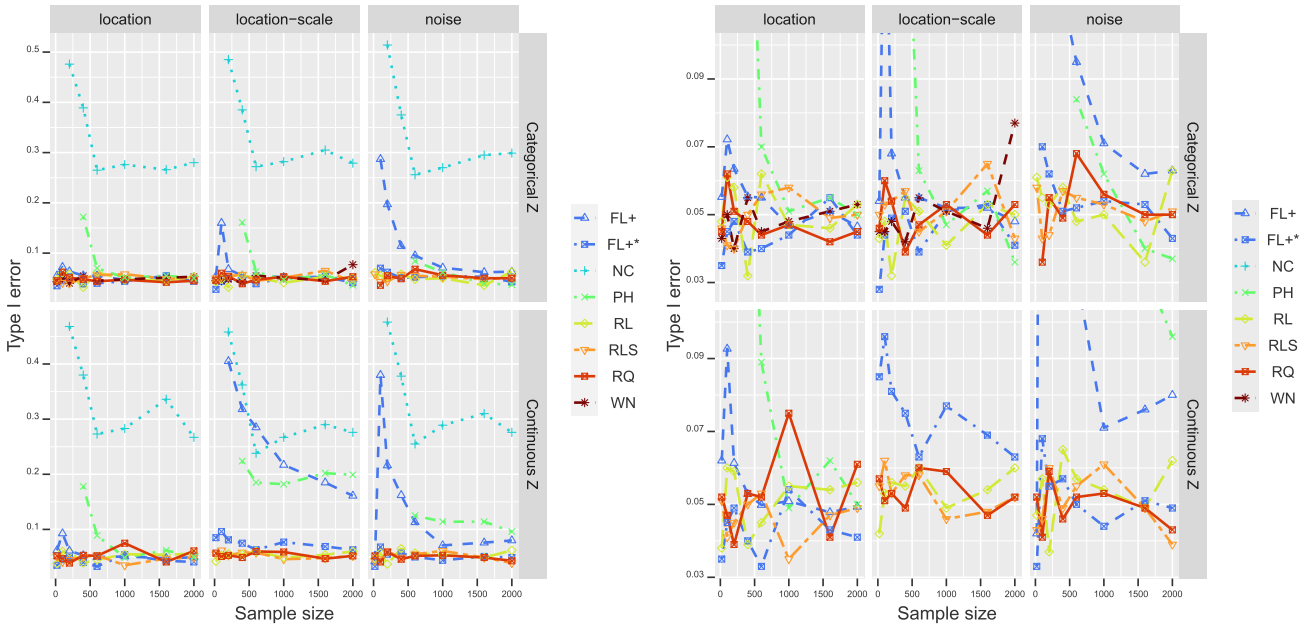
For all cases of Experiments (I) and (II), we simulated two data sets with $M = 100000$ datapoints, one for testing the empirical significance level (D_{sign}) and one for testing for power of the tests (D_{power}). For both datasets, we first simulated M realisations of the interesting covariate X from the Bernoulli(0.5) distribution and M realisations of the nuisance covariate Z from F_Z. For D_{power}, we then simulated the response variable Y as specified above. For D_{sign}, the only difference in the construction was that the values of Y' of Experiments (I) and (II) were simulated from $N(0, 1)$, both for $X = 0$ and $X = 1$, making the two distributions to coincide. We then used simple random sampling without replacement to obtain samples of size $N = 10, 50, 100, 200, 300, 500, 800, 1000$. For each sample size N, we drew 1000 independent samples. For each sample of data, we then performed the tests of Table 1.

5.1.1 Empirical significance levels

Figure 2 shows the empirical significance levels. It is evident that the test based on the FL+ permutation is extremely liberal in the presence of continuous nuisance covariates with location-scale shift or noise nuisance effects. The results are similar for the FL permutation and hence are omitted to increase the readability of Figure 2. Moreover, a similar behavior is observed for the PH test for small sample sizes (less than 500). For large sample sizes (more than 500), the overall behaviour of the method is unpredictable. Furthermore, as expected, the NC test is liberal. In contrast, the empirical significance levels of the RL, RLS, RQ and WN tests were close to the nominal level, independently of the type of the nuisance effect or the sample sizes.

5.1.2 Power

Next the power of those methods that achieved nominal significance levels was studied (see Figure 3). We investigated the power only for the cases and samples sizes where their empirical significance levels were approximately 5%. The results suggest that the global envelope tests (the first six test of Table 1) are generally more powerful than the PH test and the FL+* method. However, the RQ method was an exception; it had lower power than PH test for sample sizes less than 1000. This is likely because the quantile effect is poorly estimated for extreme quantiles. The FL+* method is naturally less powerful as it does not consider the extreme quantiles ($\tau \in [0.1, 0.9]$), and the distributional differences



(a) All results visible (b) Zoomed view

Fig. 2 Empirical significance levels for Experiments (I) and (II) among 1000 simulated samples of different sizes (x-axis) for the different tests of Table 1 (different line types). The nuisance covariate Z is either categorical or continuous with location, location-scale (Experiment (I)) or noise (Experiment (II)) effects on the response. Results are based on 10 values for τ varying from 0.01 to 0.99, except the FL+* test that considers 10 different values of τ's in the range [0.1,0.9]

between the two groups in Experiments (I) and (II) were in the tails. In the case of continuous location-scale effect the RLS permutation outperformed the RL permutation. On the contrary, under model misspecification, i.e., noise effect, the RL permutation outperformed the RLS permutation. Finally, for location effects (first column of Figure 3) it is unclear which method is the best as the FL+, RL and RLS methods had equally high power, and also WN was equally powerful in the case of categorical Z.

5.1.3 Liberality of Freedman-Lane and pointwise *p*-values

To investigate the source of liberality in the FL, FL+ and PH tests, we performed local tests, i.e., tests for single τ in the setup of Experiment (I). In each such test, only one quantile τ is considered, and the behavior of the methods is studied. The individual quantiles considered here were $\tau = 0.01, 0.05, 0.1, 0.2, 0.5$. For a categorical nuisance covariate, the resulting significance levels are shown in Figure 4 and for a continuous nuisance covariate the corresponding results are displayed in Figure 5. The tests based on the FL and FL+ permutations were extremely liberal for extreme quantiles and the pointwise test was liberal for extreme quantiles and small samples sizes. In the case of continuous nuisance with location-scale effects, the liberality was more apparent. On the other hand, the test based on the FL and FL+ permutations achieved correct significance levels for non-extreme quantiles and hence they are suitable for global testing when quantile range excludes the most extreme quantiles. For instance, in the case of median regression the use of the FL and FL+ permutations can be justified. Also for sample sizes larger than 500, it seems acceptable to exclude only quantiles < 0.1 (and > 0.9).

5.2 Sensitivity to effects on the scale of the distribution

The performance of the methods was studied in two further cases where X was still categorical, but it affected the scale of the response distribution. The conditional response distribution was defined through a t_{df} distribution where the degrees of freedom df were controlled by the realisations of X. As the normal distribution coincides with the t_{df} distribution as $df \to \infty$, the contrast between the standard normal and the t_4 distribution (as studied in Section 5.1) is larger than the contrast between t_{df} distributions with df simulated from a Poisson distribution with mean 3. Similarly to previous experiments, location, location-scale and noise effects were added to the response distribution. In Experiment (III),

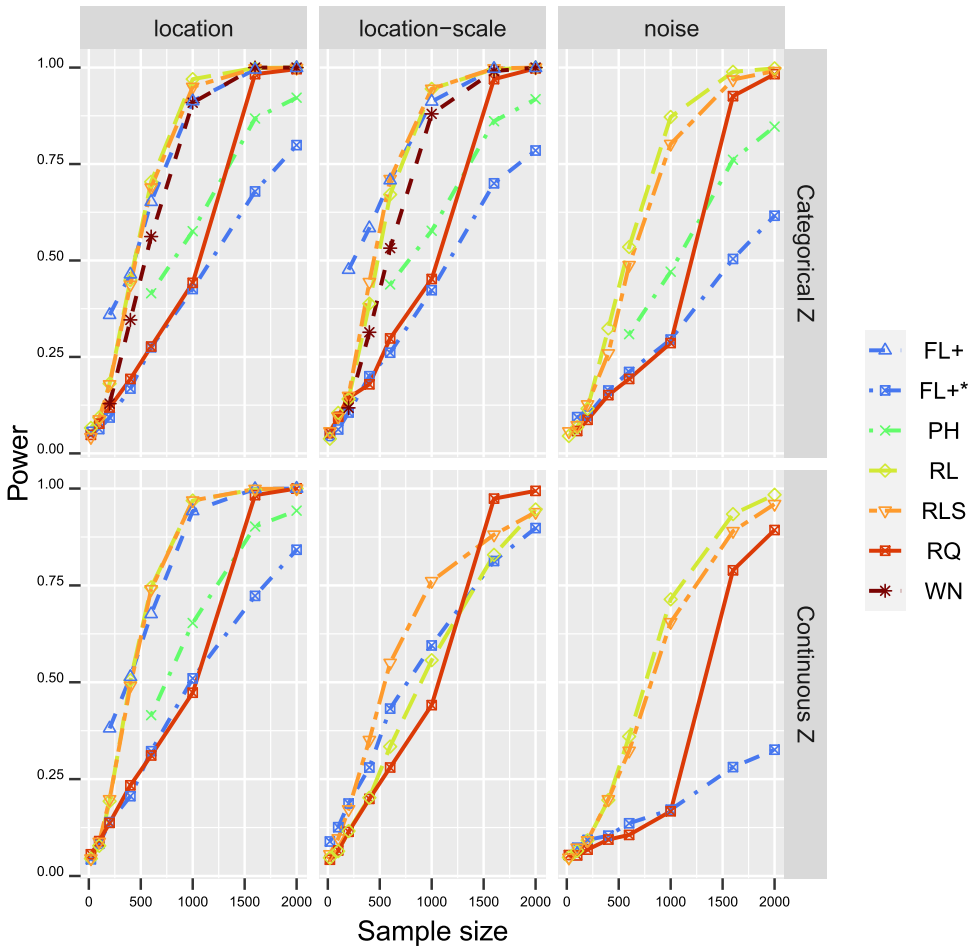


Fig. 3 Power for Experiments (I) and (II) among 1000 simulated samples of different sizes (x-axis) for the different tests of Table 1 (different line types). The nuisance covariate Z is either categorical or continuous with location, location-scale (Experiment (I)) or noise (Experiment (II)) effects on the response. Results are based on 10 values for τ varying from 0.01 to 0.99, except the FL+* test that considers 10 different values of τ's in the range [0.1,0.9

$$\begin{cases} X \sim \max(\text{Poisson}(3), 1) \\ Y' \mid X \sim t_X \\ Z \sim F_Z \\ Y = (1 + aZ)Y' + bZ \end{cases} \quad \text{(III)}$$

As before, $a, b \in \mathbb{R}$ are parameters controlling the size of the nuisance effect and F_Z is the distribution of the nuisance variable, for which we considered the the same cases (Ia)-(Id) as in Experiment (I). In Experiment (IV),

$$\begin{cases} X \sim \max(\text{Poisson}(3), 1) \\ Y' \mid X \sim t_X \\ Z \sim F_Z \\ Z_1 \sim \text{Unif}(0, 1.5) \\ \epsilon \sim N(1, 0.04) \\ Y = \begin{cases} \epsilon & \text{if } Z < Z_1 \\ Y' & \text{otherwise} \end{cases} \end{cases} \quad \text{(IV)}$$

where Z and Z_1 are nuisance covariates and F_Z is the distribution of the nuisance covariate Z with the two cases (IIa)-(IIb) as in Experiment (II).

As shown in Figure 6 (left), the FL+ and PH tests are again liberal when extreme quantiles are considered. Regarding the significance levels, the other tests also behaved similarly as in the previous experiments: the NC method was highly liberal and RL, RLS, RQ and WN were fine (Figure 6).

The power of the methods was also investigated (see Figure 7). As in Section 5.1.2, we only considered the samples sizes and methods with a significance level of approximately 5%. In the presence of location and location-scale effects the RQ permutation had the highest power, with the RL and RLS being the least powerful methods.

5.3 Sensitivity to effects on the shape of the distribution

Finally, we considered the case where X is either discrete or continuous and influences the shape of the response distribution while the nuisance covariate Z influences the scale of the response distribution. This Experiment (V) is in detail as follows:

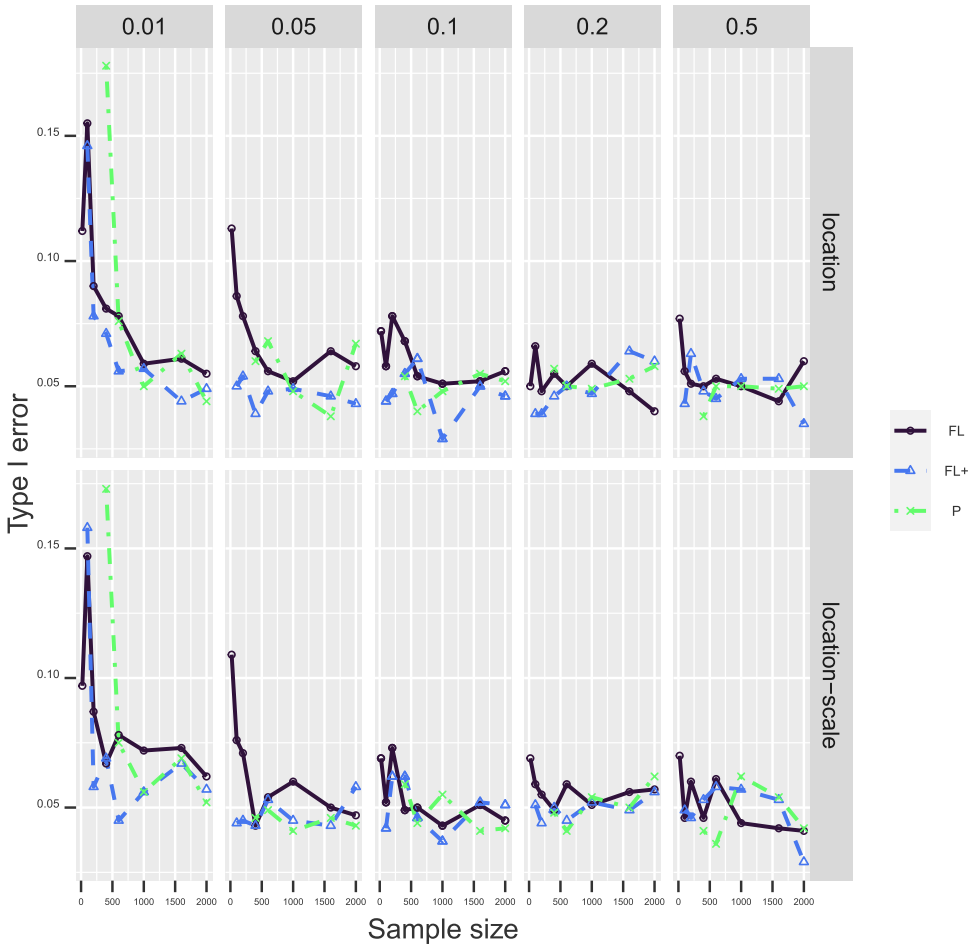


Fig. 4 Empirical significance levels for the simulation Experiment (I) among 1000 simulated samples of different sizes (x-axis) for the Freedman-Lane based permutation strategies and pointwise p-value (different line types). The nuisance covariate Z is categorical with location or location-scale effects on the response. Quantiles considered are $\tau = \{0.01, 0.05, 0.1, 0.2, 0.5\}$ (columns)

$$\begin{cases} X \sim F_X \\ Z \sim \text{Unif}(0.5, 2) \\ Y \sim \text{Gamma}(X, Z) \end{cases} \tag{V}$$

where $F_X = \text{Unif}(4, 5)$ if X is continuous and F_X takes values 4.7 and 5 with equal probabilities if X is categorical.

We studied at the empirical significance levels in this experiment by simulating the interesting covariate X having no effect on the response distribution, i.e., the data (Y) were simulated from the Gamma distribution with shape parameter $shape = 4.5$. Again the FL+ and the PH tests were liberal when extreme quantiles τ were considered, while the tests with the RL, RLS and RQ permutations achieved correct significance level for all sample sizes (Figure 8).

For testing the power of the tests, the response variable Y was simulated from a Gamma distribution where the shape parameter was defined through the interesting covariate X as specified in (V). As earlier, we considered only the sample sizes and methods whose empirical significance levels were approximately 5%. Figure 9 shows the results. The RQ test had low power for small samples, while the other methods were equivalent in terms of power.

5.4 Sensitivity to correlation of the interesting and nuisance covariates

Finally, we studied the performance of the permutation methods in the case where the interesting covariate X and the nuisance covariate Z are correlated. In this Experiment (VI), we had

$$\begin{cases} A, B, C \sim \text{Unif}(0, 1) \\ X = (1 - c)A + cC, \text{ with } 0 \le c \le 1 \\ Z = (1 - c)B + cC, \text{ with } 0 \le c \le 1 \\ Y \sim \text{Gamma}(4 + X, 1 + Z) \end{cases} \tag{VI}$$

We considered the cases with $c = 0, 0.3, 0.5, 0.7$. In this setup X and Z are positively correlated with correlation given by $\text{cor}(X, Z) = \frac{c^2}{1+2c^2-2c}$. Therefore, increasing c towards 1, increases the correlation between X and Z, while X and Z are independent when $c = 0$. To increase the readability of Figure 10 showing the results, only Type I errors lower than 0.3 are shown. For instance, under this model misspecification, the RL permutation strategy led to the more liberal test the larger the correlation between X and Z was. This is because the RL permutation fails to remove the com-

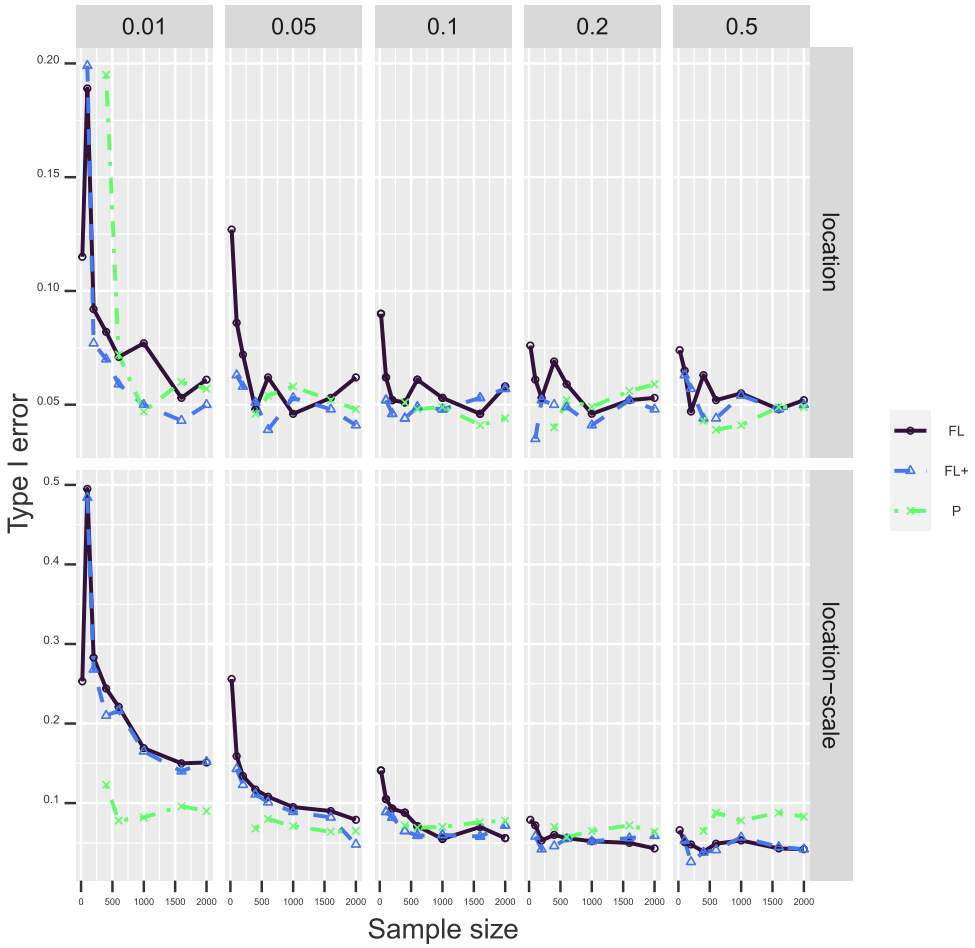


Fig. 5 Empirical significance level for the simulation Experiment (I) among 1000 simulated samples of different sizes (x-axis) for the Freedman-Lane based permutation strategies and pointwise p-value (different line types). The nuisance covariate Z is continuous with location or location-scale effects on the response. Quantiles considered are $\tau = \{0.01, 0.05, 0.1, 0.2, 0.5\}$ (columns)

plete nuisance effect, here a location-scale effect, from the response Y. Hence, there is still a significant effect of the nuisance Z present on the residuals $\boldsymbol{\epsilon}_Z$. Now, as the correlation between X and Z increases, the effect of X on $\boldsymbol{\epsilon}_Z$ becomes significant, causing the test to be more liberal. On the contrary, the permutation tests that correctly remove the nuisance effects (RLS and RQ) were conservative with increasing correlation, which can result in low power. Finally, only the extension of the Freedman-Lane test without considering extreme quantiles (FL+*) achieved the significance level close to the nominal level for all levels of correlation and larger sample size.

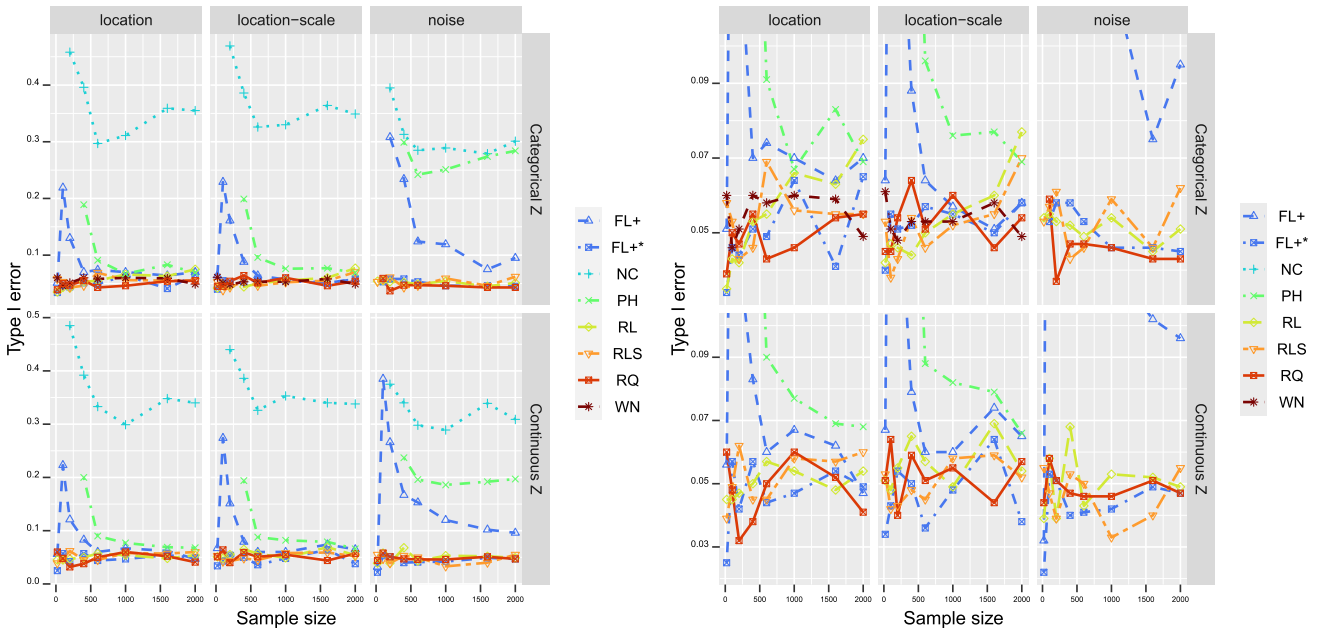
Further, we studied the behaviour of the methods in the simulation setup of the first two experiments, modified to the case where X and Z are correlated. These two cases are detailed as follows:

$$\begin{cases} A, B, C \sim \text{Unif}(0, 1) \\ X = \text{round}((1-c)A + cC), \text{ with } 0 \le c \le 1 \\ Y' \mid X \sim \begin{cases} N(0,1) & \text{if } X = 0 \\ t_4 & \text{if } X = 1 \end{cases} \\ Z = 1.5 \cdot ((1-c)B + cC), \text{ with } 0 \le c \le 1 \\ Y = (1 + aZ)Y' + bZ \end{cases} \quad \text{(VII)}$$

$$\begin{cases} A, B, C \sim \text{Unif}(0, 1) \\ X = \text{round}((1-c)A + cC), \text{ with } 0 \le c \le 1 \\ Y' \mid X \sim \begin{cases} N(0,1) & \text{if } X = 0 \\ t_4 & \text{if } X = 1 \end{cases} \\ Z = 1.5 \cdot ((1-c)B + cC), \text{ with } 0 \le c \le 1 \\ Z_1 \sim \text{Unif}(0, 1.5) \\ \epsilon \sim N(1, \sigma_\epsilon^2) \\ Y = \begin{cases} \epsilon & \text{if } Z < Z_1 \\ Y' & \text{otherwise} \end{cases} \end{cases} \quad \text{(VIII)}$$

We considered values $c = 0.3, 0.4, 0.5, 0.9$. As before, increasing c towards 1 increases the dependency between X and Z.

The empirical significance levels for Experiments (VII) and (VIII) are shown in Figure 11. According to the results, the RL permutation strategy is liberal when the assumption of the test, i.e., the effect of nuisance is only in location, is not satisfied (cases of location-scale and noise). That is caused by the fact that the RL method filters away only the location effect of the nuisance, i.e., the residuals $\boldsymbol{\epsilon}_Z$ still contain other effects of the nuisance. As a result, if the interesting and the nuisance covariates are correlated, the interesting covariate also affects $\boldsymbol{\epsilon}_Z$. This remaining effect causes a significant



(a) All results visible (b) Zoomed view

Fig. 6 Empirical significance levels for Experiments (III) and (IV) among 1000 simulated samples of different sizes (x-axis) for the different tests of Table 1 (different line types). The nuisance covariate Z is either categorical or continuous with location, location-scale (Experiment (III)) or noise (Experiment (IV)) effects on the response. Results are based on 10 values for τ varying from 0.01 to 0.99, except the FL+* test that considers 10 different values of τ's in the range [0.1,0.9]

result when the interesting covariate is tested by quantile regression. The same can be seen for the RLS permutation strategy when the assumption of the test, i.e., the effect of nuisance is only in location and scale, is not satisfied. This effect is not presented in the RQ permutation strategy, nevertheless, all three methods appear to be conservative with increasing correlation between interesting and nuisance covariates.

The results of Experiments (VI) and (VII) suggest that the effect of nuisance covariates must be rigorously tested. Suppose the interest is not on the tails of the distribution. In that case, the FL+ permutation test without extreme quantiles appears to be a good choice: it achieved the correct significance level independently of the amount of correlation between X and Z in our experiments.

5.5 Summary

From the above experiments, we give the following recommendations:

- The pointwise minimum p-value is extremely liberal for the global test, as the multiple testing problem is not considered. The PH test seems to be liberal even though the Holm-Bonferroni correction for multiple testing is conservative. Thus, these tests are not recommended.
- The Freedman-Lane-based global quantile tests should be avoided when extreme quantiles are considered. If the interval for quantiles is (0.1,0.9), then the Freedman-Lane global quantile tests should be avoided with less than 500 data.
- In the presence of only one categorical nuisance, the WN method is recommended, because it satisfies the exchangeability.
- The RL, RLS methods are liberal when X and Z are correlated and the assumptions of the effect of nuisance covariates on data are not satisfied.
- The RQ permutation can have lower power for small sample sizes as the quantile effect is badly estimated for extreme quantiles.
- If the nuisance influences only the location, then the RL permutation is recommended and if it further influences the scale then the RLS permutation is recommended. If the effect is unknown, then the RQ permutation is recommended.
- The effect of nuisance covariates must be rigorously tested in order to choose the appropriate permutation strategy. This is possible via the Khmaladze test implemented in the quantreg package. Since the Khmaladze test is recommended for non-extreme quantiles only and we are usually interested in all quantiles, the visual

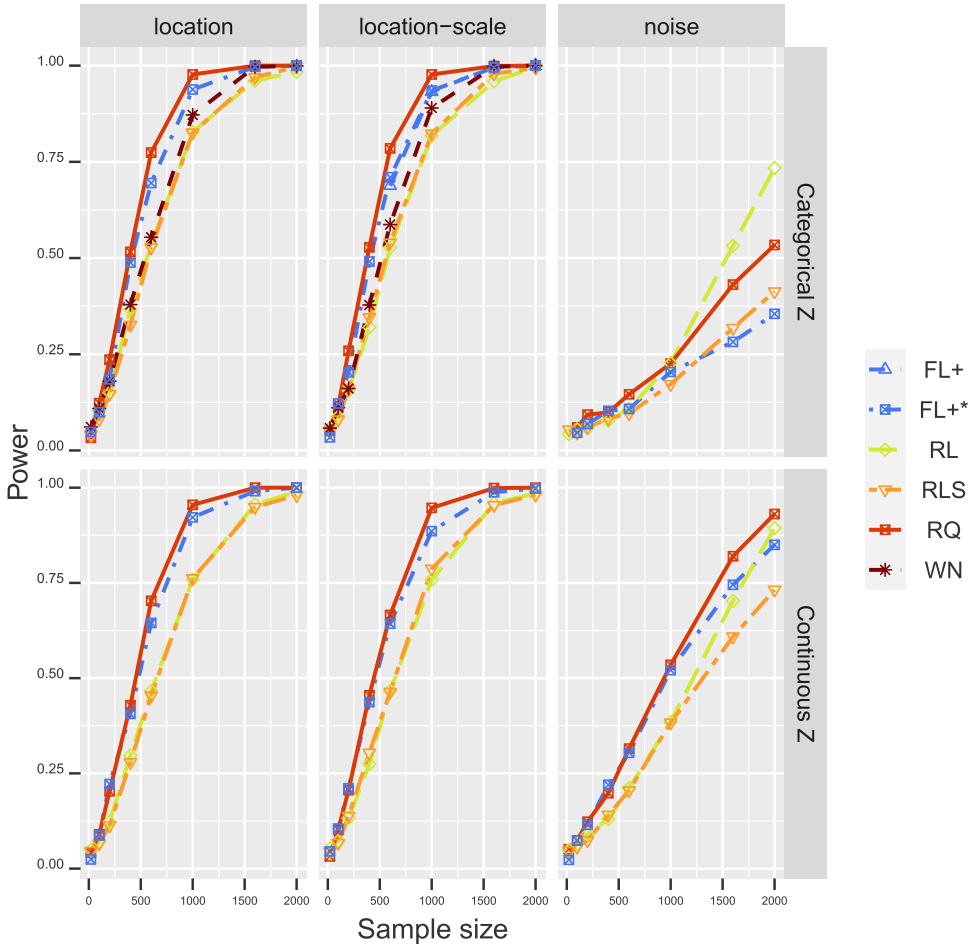


Fig. 7 Power for Experiments (III) and (IV) among 1000 simulated samples of different sizes (x-axis) for the different tests of Table 1 (different line types). The nuisance covariate Z is either categorical or continuous with location, location-scale (Experiment (III)) or noise effects (Experiment (IV)) on the response. Results are based on 10 values for τ varying from 0.01 to 0.99, except the FL+* test that considers 10 different values of τ's in the range [0.1,0.9]

inspection of the pointwise confidence bands provided by the quantreg package can serve as an exploratory tool for checking what kind of effect the nuisance covariate has on the response. The constant coefficients with respect to quantiles correspond to the location shift, and the linear behavior of the coefficients corresponds to the location-scale shift.

- In cases when a user is not sure which permutation method is the best to be applied, he/she can choose the RQ method, which always complies with the correct size. The only thing the user risks with this choice is having a lower power than the test with a more appropriate permutation strategy for his/her dataset.

6 Data examples

6.1 Forest stand age with respect to forest naturalness

In the Finnish national forest inventory (NFI), naturalness of the forest is evaluated in the field from three criteria, namely structure, deadwood and human action. Myllymäki et al. (2023) studied properties of the forest structure within the structural naturalness classes evaluated in the field, and we also inspect only this structural naturalness. Namely, we investigated the distributions of stand age in the three naturalness groups 'natural', 'near-natural' and 'non-natural' in the Finnish Lapland, excluding the northernmost part. The study region corresponds to 'North' of (Myllymäki et al. 2023, Figure1). Here, for simplicity, we restricted our attention to plots on rich mineral soils. Only plots in forest land that were completely located within a single stand and had at least five measured trees were considered (see details in Myllymäki et al. 2023). Because the stand age depends potentially on the dominant species, we included as the nuisance covariate the dominant species as a variable with three categories 'Broadleaf', 'Conifer' and 'Mixed' as defined in Myllymäki et al. (2022). Numbers of plots in each category are shown in Table 2.

Our quantile regression model is

$$Age \sim constant + naturalness + species$$

where naturalness is our interesting factor and species is the nuisance. According to the quantile regression fit (Figure

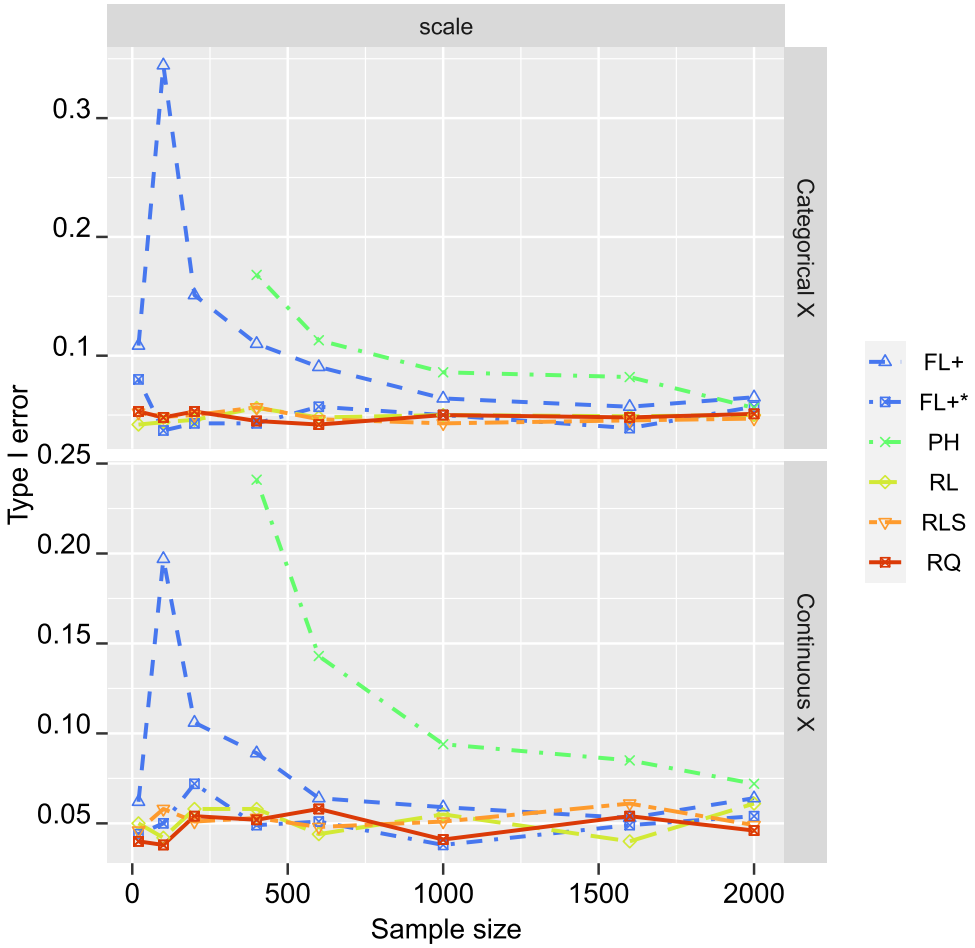


Fig. 8 Empirical significance level for Experiment (V) among 1000 simulated samples of different sizes (x-axis) for the different tests of Table 1 (different line types). The interesting covariate X is either categorical or continuous and the nuisance covariate Z is continuous with scale effect. Results are based on 10 values for τ varying from 0.01 to 0.99, except the FL+* test that considers 10 different values of τ's in the range [0.1,0.9]

12, rows 1-2), the effect of the nuisance effect, i.e., dominant species, appears to be location-scale shift, since the estimated coefficients (row 2) appear to be linear in τ. Therefore, to test for the differences between the distributions of stand age in the natural, near-natural and non-natural forests, we applied the permutation algorithm RLS of Section 4.5. Figure 12 (row 3) shows the results of this test based on 2499 permutations and one hundred equidistantly distributed quantiles. The gray zone shows the global envelope, while the estimated coefficients are shown by a black solid line, overlaid with red dots when outside the envelope. Note here that the global test of naturalness contains both functional coefficients shown in row 3 of Figure 12. Thus, the test corresponds to the ANOVA test of the effect of the categorical covariate with the reference group being the non-natural stands. This two-way ANOVA for whole distributions is tested using 2×100 pointwise tests, one hundred for the contrast between near-natural and non-natural stands and one hundred for the contrast between natural and non-natural stands. The p-value of the global test is 0.0004. When the estimated coefficients are outside the global envelope for some quantiles and coefficients, the global test is significant (p-value ≤ 0.05). Further, the test identifies both the significant quantiles and the corresponding coefficient under the global test. Here, the coefficients of near-natural and natural stands show the difference to non-natural reference group simply for all quantiles. It can be seen that both the near-natural and natural stands are uniformly (for all quantiles) older than non-natural stands.

For another example, we switched the roles of naturalness and dominant species. Naturalness is now a nuisance factor, and the associated coefficients shown in row 1 of Figure 12 do not appear to be linear. Therefore, the location-scale shift can not be assumed and we used the RQ permutation strategy. We used again 2499 permutations and the same τs as earlier. Figure 12 (row 4) shows the results of this global test, i.e., the two-way ANOVA test for whole distributions with conifer dominated stands being the reference group. The p-value of the global test is 0.0004. It can be seen that there is a significant effect for quantiles between 0.65 and 0.8 for the difference between mixed and conifer plots. This can be interpreted as mixed stands being significantly younger than the conifer stands but only for older stands (not very old). Similarly, it can be seen that there is a significant effect for quantiles between 0.3 and 0.85 for the

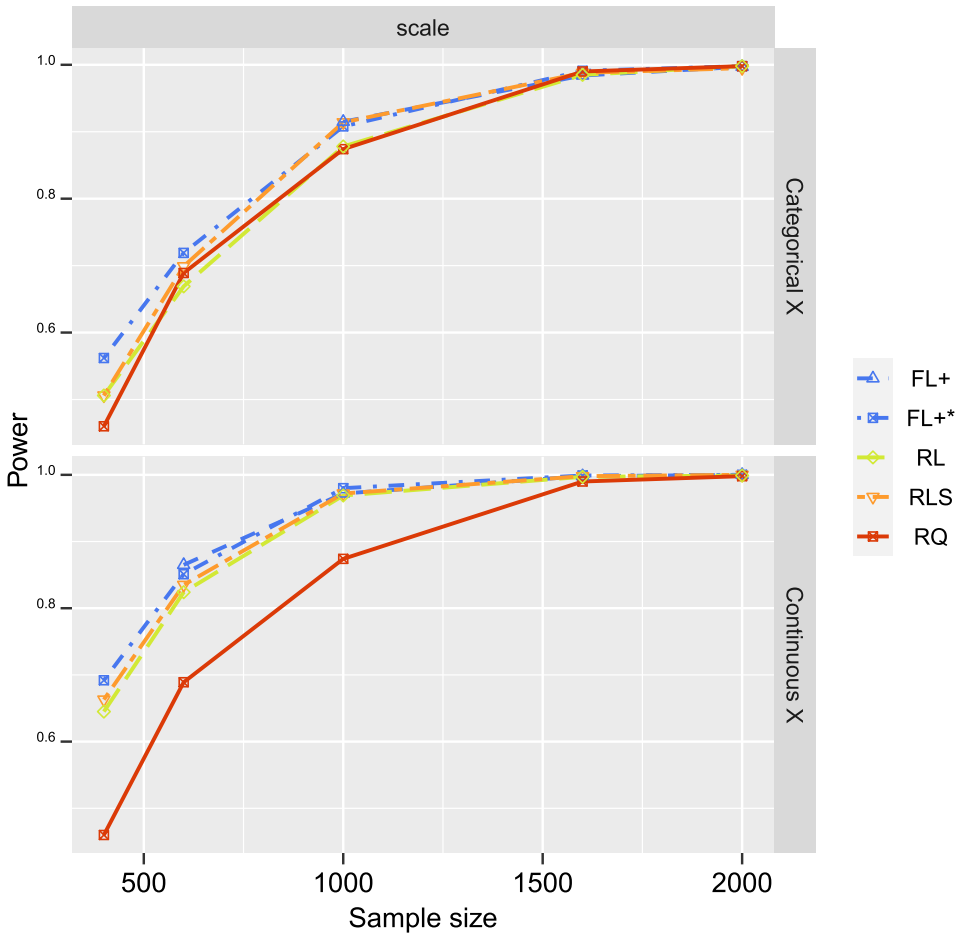


Fig. 9 Power for Experiment (V) among 1000 simulated samplesof different sizes (x-axis) for the different tests of Table 1 (different line types). The interesting covariate X is either categorical or continuous and the nuisance covariate Z is continuous with scale effect. Results are based on 10 values for τ varying from 0.01 to 0.99, except the FL+* test that considers 10 different values of τ's in the range [0.1,0.9]

Table 2 Numbers of NFI plots in total and in the different naturalness groups (0 = natural; 1 = near-natural; 2 = non-natural)

Dominant species	0	1	2
Broadleaf	29	9	81
Conifer	59	36	341
Mixed	55	23	140

difference between broadleaf and conifer stands. This can be interpreted as broadleaf stands being significantly younger than conifer stands except for very young and very old stands. In other words, the stand age distribution of broadleaf dominated forests is more skewed to the right than the distribution of conifer dominated forests, but the ranges are equal. As one can see from the above example, the global test in quantile regression allows for a rich interpretation of the differences between the distributions after accounting for the nuisance effects.

6.2 Motivational example

In the motivational example (see Section 1.1), the RLS permutation strategy was chosen for gold as an interesting covariate because the prices of uranium and oil appear to be location scale shifts; their coefficients behave quite linearly with respect to τ. On the other hand, the effect of gold seems to be non-linear. Thus, the RQ permutation strategy was used for oil and uranium.

6.3 Running time

To compare the computational cost of all studied methods, we show the computational times for all methods for both examples described above in Table 3. The computational time is rather similar for all methods relying on permutations, since the permutations are the most demanding job. The PH and NC methods do not rely on permutations and therefore their computational cost is in different orders.

7 Conclusions and discussion

In this paper, we studied the possibilities to test the significance of a covariate in global test in quantile regression, i.e., simultaneously for all the quantiles. We realized first that the pointwise p-values traditionally used in quantile regression

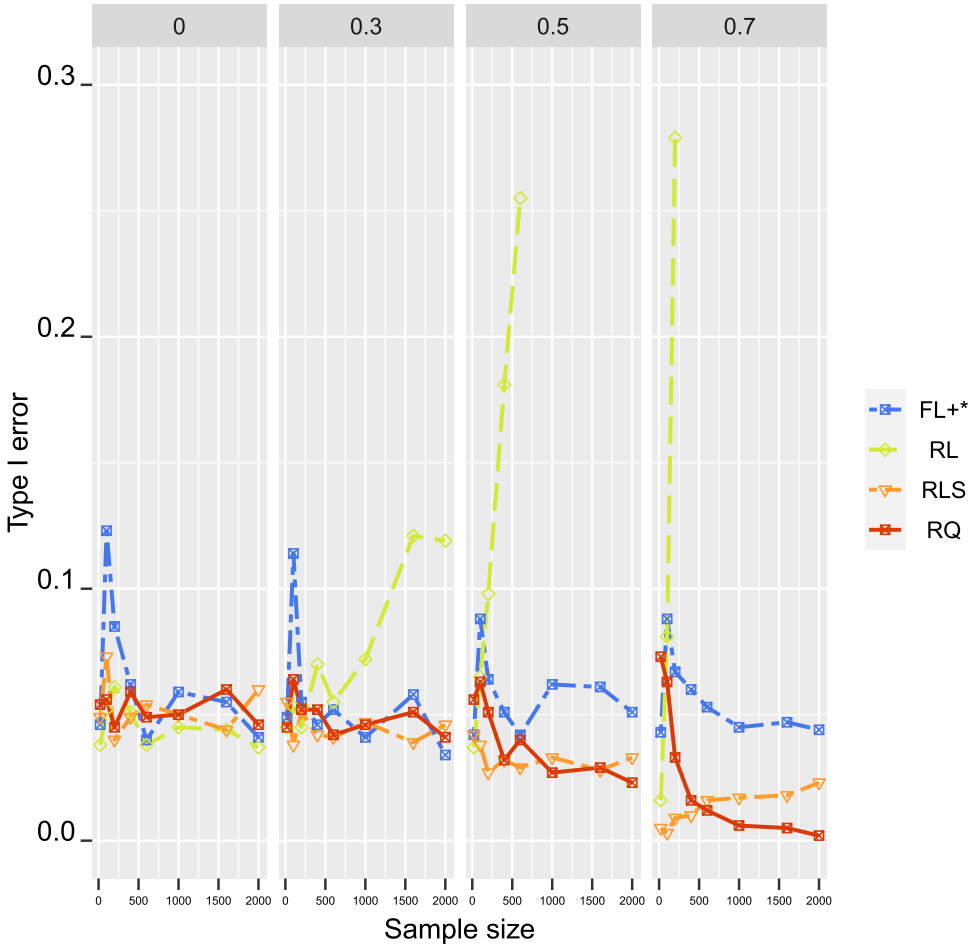


Fig. 10 Empirical significance levels for Experiment (VI) among 1000 simulated samples of different sizes (x-axis) for the different tests of Table 1 (different line types). The nuisance covariate Z is continuous with scale effect. The different columns correspond to the results for different values of c. Results are based on 10 values for τ varying from 0.01 to 0.99, except the FL+* test that considers 10 different values of τ's in the range [0.1,0.9]

Table 3 Runtime in seconds of the global tests in quantile regression for 90 and 100 quantiles for the gold and forest examples, respectively, and 2499 permutations

Permutation strategy	Forest	Gold
FL	320	681
FL+	371	868
WN	328	N.A.
RL	367	557
RLS	346	521
RQ	247	446
PH/NC	0.4	0.5

are seriously liberal for extreme quantiles, so much that even the conservative multiple testing adjustment (cf. PH of Table 1) does not correct the liberality. Therefore, we decided to rely on pointwise permutation tests with the global envelope test as the multiple testing adjustment procedure.

The choice of the permutation strategy is the crucial point in permutation tests. Surprisingly, it turns out that the traditionally used Freedman-Lane permutation strategies are also liberal for extreme quantiles. Therefore, we proposed other permutation strategies that seem to work well even for extreme quantiles. These strategies are based on evaluating the type of influence of data by nuisance covariates. If this influence is only in location, the permutation with the removal of the location effect is recommended. If this influence is in location and scale, the permutation with the removal of the location and scale effect of nuisance covariates is recommended. If this influence is more general, then the permutation with the removal of the quantile effect of the nuisance covariates is recommended. Thus, the permutation with removal of the quantile effect is a relatively safe choice.

The recommended methods were conservative when a correlation between nuisance and interesting covariates was present, and the assumptions of these methods about the effect of nuisance covariates on the data were satisfied. We believe that this is always the case, as if the model is correctly specified ϵ_Z will not contain any nuisance effect, and hence if X and Z are highly correlated, X will have no effect on ϵ_Z, which will lead to a conservative test. On the other hand, the RL and RLS methods seem to be extremely liberal when the correlation of interesting and nuisance covariates is present, and the assumptions of these methods about the effect of nui-

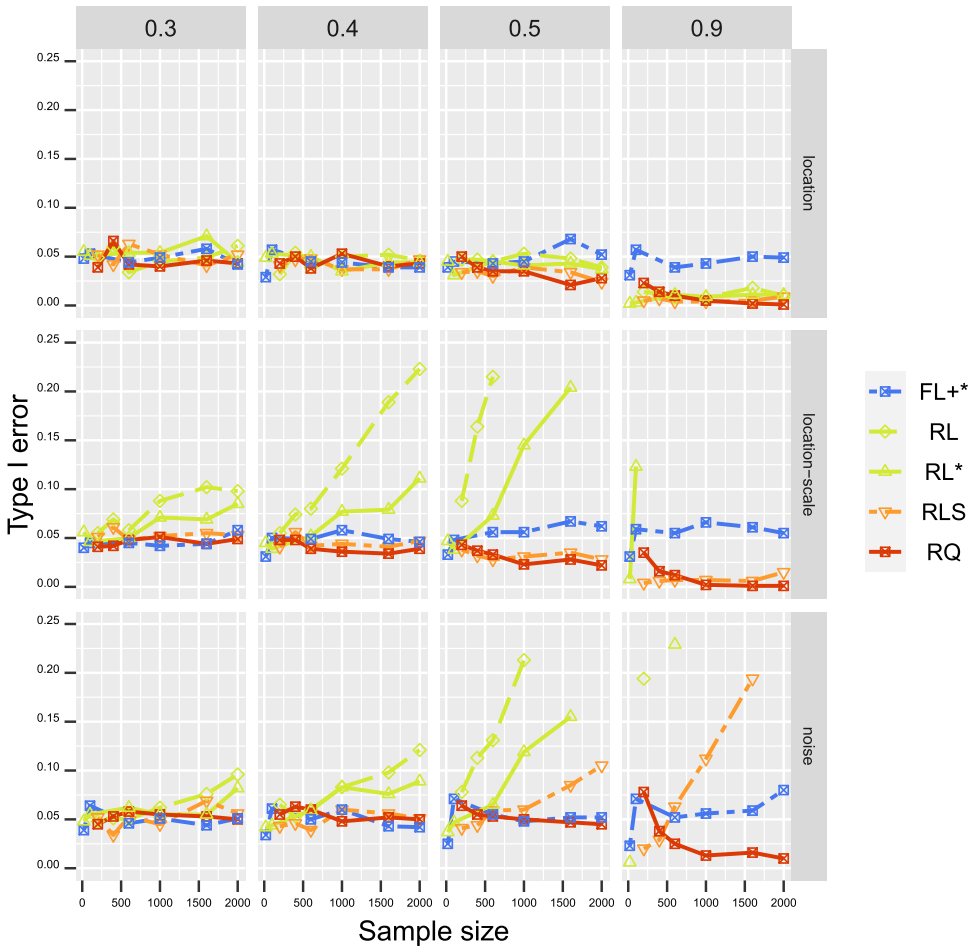


Fig. 11 Empirical significance levels for Experiments (VII) and (VIII) among 1000 simulated samples of different sizes (x-axis) for the different tests of Table 1 (different line types). The nuisance covariate Z is continuous with location, location-scale (Experiment (VII)) or noise (Experiment (VIII)) effects on the response. The different columns correspond to the results for different values of c. Results are based on 10 values for τ varying from 0.01 to 0.99, except the FL+* test that considers 10 different values of τ's in the range [0.1,0.9]

sance covariates on the data are not satisfied. This behavior makes the assumption of the effect of nuisance covariates on the data critical for choosing the permutation strategy. The reason for carefully checking the kind of nuisance effect is that the safe method, permutation with removal of the quantile effect of the nuisance covariates, can have lower power than the other proposed methods for a smaller amount of data.

The data study examples show how one can choose the appropriate permutation strategy. They also show that if the pointwise tests are significant, the global test can be significant as well, but it does not have to. Considering the computational time of the proposed procedure, it is dependent on the chosen number of permutations. For every permutation, the quantile regression model must be evaluated, and that is the task where the algorithm spent the most time. Therefore, the computational time does not depend on the permutation strategy. E.g., the computational time for the motivational example was 10 minutes on a usual computer and 4 minutes for the forest example.

The proposed tests are useful if we are interested in the existence of the effect of a covariate on the data distribution in at least one quantile. They are also useful if several data distributions are compared, but the data are attached with nuisance covariates. An example is that the distribution of a statistic is compared for different health statuses, but every person for which the statistic is computed is attached with various covariates like age or sex.

One of the advantages of the global envelope test used on the pointwise permutation tests here is that it provides the graphical output, which automatically detects the quantiles responsible for the potential rejection. Also, it automatically detects which levels of the categorical covariates differ from the overall mean across all levels. Another advantage of the global envelope test here is its nonparametric nature, which causes the adjustment procedure to be valid for any test statistic withoutthe necessity of computing its asymptotic variances.

The only problem in this kind of permutation procedure is the assumption of exchangeability of the test vector under the permutation strategy. It is known that when nuisance covariates are present, the exchangeability can not be reached even for linear models where the mean value is modeled. For these models, the Freedman-Lane procedure is well accepted and

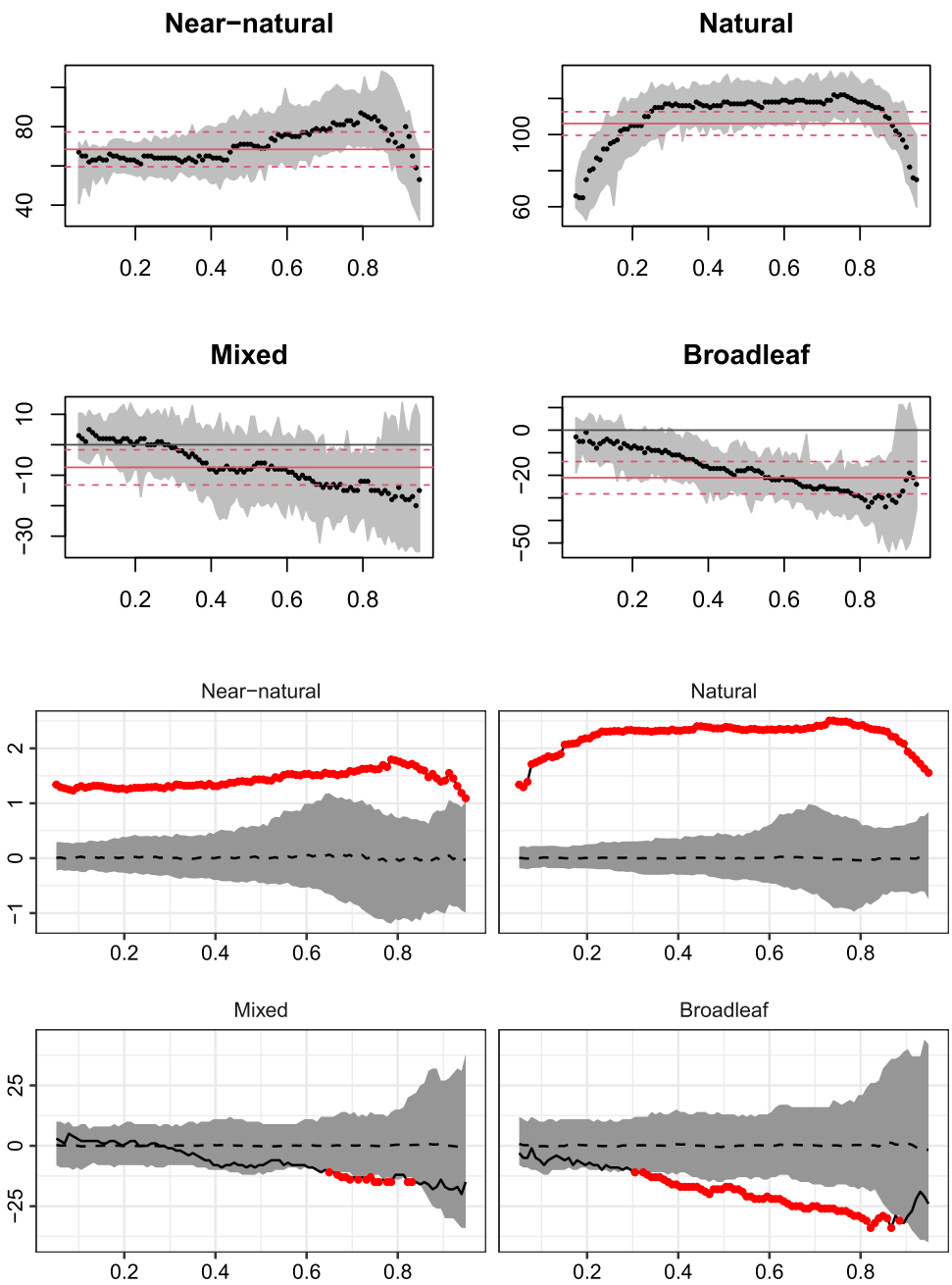


Fig. 12 95% pointwise confidence bands (rows 1-2) and 95% global envelopes (rows 3 and 4) for the effect of naturalness or dominant species on stand age. The global envelope on row 3 is based on the RLS permutation strategy testing the effect of naturalness accounting for the dominant species as a nuisance. Row 4 is based on the RQ permutation strategy testing the effect of dominant species accounting for the naturalness as nuisance

the exactness of such tests is studied via simulations. We followed here the same strategy for quantile regression. By our simulation study, we showed that even though our proposed permutation strategies do not reach exchangeability, their empirical significant levels were very close to the nominal level or below it (conservativeness). The conservativeness of our procedures appeared only when the nuisance and interesting covariates were correlated.

Some of the proposed permutation methods construct different new permuted data for each quantile. This may appear strange from the quantile regression point of view. On the other hand, the essence of the proposed methodology is using local quantile regression for every quantile independently and applying the quantile-wise permutation method. That gives many quantile-wise tests that are seriously dependent. On top of that, the universal correction for multiple testing is applied through the global envelope test. Such a multiple testing adjustment can be used for any kind of tests in cases when they are based on the same permutations. The global envelope test further uses the information about the correlation between quantile-wise tests that arises from the common permutations in order to bring more power to the adjustment.

The proposed procedures were studied only in the cases of main effect models. It is possible to apply our methods

also in the case of studying interactions but the proposed permutation strategies would have to be slightly changed: the main effects considered as the nuisance effects would have to appear also in step 2. of the proposed procedures even thought their effect was already removed in step 1. This adjusted procedure was not rigorously analyzed yet and therefore it remains for our future work.

We remark here that quantile regression allows for modeling the heteroscedasticity. On the other hand, the null hypothesis of the global test is that there is no effect of the interesting covariate in any quantile $\tau \in \mathcal{T}$, so consequently, under the null hypothesis, there is no heteroscedasticity with respect to the interesting covariate. Thus, the permutation strategy does not need to take into account heteroscedasticity in the interesting covariate. However, the effect of the nuisance covariates must be considered, and the corresponding permutation strategy must be chosen, as it was discussed above. The results shown in Figure 2 suggest that the presence of heteroscedasticity in the nuisance covariates increases the liberality of the FL procedures, and it would be worth investigating remedies for this heteroscedasticity issue in the FL procedures, e.g., through weighted permutation strategies and weighted quantile regression. This would require estimating the variability for weighting. Since it is a wide topic, we left these possibilities for further research.

Acknowledgements The project was partly supported by the ERC CZ grant LL2407 of the Ministry of Education, Youth and Sport of the Czech Republic. The authorswish to acknowledge CSC – IT Center for Science, Finland, for computational resources. The authors also thank Aila Särkkä for valuable feedback and suggestions.

Funding Open access funding provided by Natural Resources Institute Finland. The authors gratefully acknowledge financial support from *ERC CZ of the Ministry of Education, Youth and Sport of the Czech Republic (grant LL2407)*, *Swedish Research Council* and *the European Union - NextGenerationEU in the Research Council of Finland project (Grant number 348154) under the Research Council of Finland's flagship ecosystem for Forest-Human-Machine Interplay—Building Resilience, Redefining Value Networks and Enabling Meaningful Experiences (UNITE) (Grant number 357909)*.

Code Availability The implementation of the proposed method with an example code is available in the R package GET.

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

References

Anderson, M., Ter Braak, C.: Permutation tests for multi-factorial analysis of variance. J. Stat. Comput. Simul. **73**(2), 85–113 (2003). https://doi.org/10.1080/00949650215733

Anderson, M.J., Robinson, J.: Permutation tests for linear models. Australian & New Zealand Journal of Statistics **43**(1), 75–88 (2001). https://doi.org/10.1111/1467-842X.00156

Bassett, G.W., Koenker, R.W.: Strong consistency of regression quantiles and related empirical processes. Economet. Theor. **2**(2), 191–201 (1986). https://doi.org/10.1017/S0266466600011488

Belloni, A., Chernozhukov, V., Hansen, C.: Inference on treatment effects after selection among high-dimensional controls. Rev. Econ. Stud. **81**(2(287)), 608–650 (2014)

Bickel, P.J., Freedman, D.: Asymptotic theory for the bootstrap. Ann. Stat. **9**(6), 1196–1217 (1981)

Cade, B.S., Richards, J.D.: A permutation test for quantile regression. J. Agric. Biol. Environ. Stat. **11**(1), 106–126 (2006)

Chen, C. and Y. Wei. 2005. Computational issues for quantile regression. *Sankhyā: The Indian Journal of Statistics*: 399–417

Dantzig, G.: Linear programming and extensions. Princeton University Press, Linear programming and extensions (2016)

Davison, A.C., Hinkley, D.V.: Bootstrap Methods and their Application. Cambridge University Press, Cambridge (1997)

Ditzhaus, M., R. Fried, and M. Pauly. 2021. Qanova: quantile-based permutation methods for general factorial designs. *TEST*

Efron, B.: Bootstrap methods: Another look at the jackknife. Ann. Statist. **7**(1), 1–26 (1979)

Freedman, D., Lane, D.: A nonstochastic interpretation of reported significance levels. Journal of Business and Economic Statistics **1**(4), 292–298 (1983)

Gutenbrunner, C., Jurečková, J., Koenker, R., Portnoy, S.: Tests of linear hypotheses based on regression rank scores. Journal of Nonparametric Statistics **2**(4), 307–331 (1993)

Hahn, J.: Bootstrapping quantile regression estimators. Economet. Theor. **11**(1), 105–121 (1995)

Hahn, U. 2015. A note on simultaneous Monte Carlo tests. Technical report, Centre for Stochastic Geometry and advanced Bioimaging, Aarhus University

Hall, P., Sheather, S.J.: On the distribution of a studentized quantile. J. Roy. Stat. Soc.: Ser. B (Methodol.) **50**(3), 381–391 (1988)

He, X., Hu, F.: Markov chain marginal bootstrap. J. Am. Stat. Assoc. **97**(459), 783–795 (2002)

Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**(2), 65–70 (1979)

Khmaladze, E.V.: Martingale approach in the theory of goodness-of-fit tests. Theory of Probability & Its Applications **26**(2), 240–257 (1982). https://doi.org/10.1137/1126027

Kleiner, A., A. Talwalkar, P. Sarkar, and M.I. Jordan. 2014. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*: 795–816

Kocherginsky, M., He, X., Mu, Y.: Practical confidence intervals for regression quantiles. J. Comput. Graph. Stat. **14**(1), 41–55 (2005)

Koenker, R. 1994. Confidence intervals for regression quantiles. In *Asymptotic statistics: proceedings of the fifth prague symposium, held from September 4–9, 1993*, pp. 349–359. Springer

Koenker, R.: Quantile Regression. Press, Cambridge U (2005)

Koenker, R.: quantreg: Quantile Regression. R package version **5**, 94 (2022)

Koenker, R. and G. Bassett. 1978. Regression quantiles. *Econometrica: journal of the Econometric Society*: 33–50. https://doi.org/10.2307/1913643

Koenker, R., Chernozhukov, V., He, X., Peng, L.: Handbook of Quantile Regression. Chapman & Hall (2018)

Koenker, R., Machado, J.A.: Goodness of fit and related inference processes for quantile regression. J. Am. Stat. Assoc. **94**(448), 1296–1310 (1999)

Koenker, R., Xiao, Z.: Inference on the quantile regression process. Econometrica **70**(4), 1583–1612 (2002). https://doi.org/10.1111/1468-0262.00342

Mrkvička, T., Myllymäki, M., Hahn, U.: Multiple Monte Carlo testing, with applications in spatial point processes. Stat. Comput. **27**(5), 1239–1255 (2017). https://doi.org/10.1007/s11222-016-9683-9

Mrkvička, T., Myllymäki, M., Kuronen, M., Narisetty, N.N.: New methods for multiple testing in permutation inference for the general linear model. Stat. Med. **41**(2), 276–297 (2022), https://doi.org/10.1002/sim.9236. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.9236

Myllymäki, M. and T. Mrkvička. 2024. GET: Global envelopes in R. *Journal of Statistical Software 111*(3): 1–40. https://doi.org/10.18637/jss.v111.i03

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., Hahn, U.: Global envelope tests for spatial processes. J. R. Statist. Soc. B **79**, 381–404 (2017), https://doi.org/10.1111/rssb.12172. arXiv:1307.0239 [stat.ME]

Myllymäki, M., Tuominen, S., Kuronen, M., Packalen, P., Kangas, A.: The relationship between forest structure and naturalness in the Finnish national forest inventory. (2023). https://doi.org/10.1093/forestry/cpad053

Narisetty, N.N., Nair, V.J.: Extremal depth for functional data and applications. J. Am. Stat. Assoc. **111**(516), 1705–1714 (2016)

Parzen, M.I., Wei, L.J., Ying, Z.: A resampling method based on pivotal estimating functions. Biometrika **81**(2), 341–350 (1994)

Peng, L., Fine, J.P.: Competing risks quantile regression. J. Am. Stat. Assoc. **104**(488), 1440–1453 (2009)

Portnoy, S., Koenker, R.: The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. Stat. Sci. **12**(4), 279–300 (1997)

van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press (2007)

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E.: Permutation inference for the general linear model. Neuroimage **92**, 381–397 (2014). https://doi.org/10.1016/j.neuroimage.2014.01.060

Zheng, Q., Peng, L., He, X.: Globally adaptive quantile regression with ultra-high dimensional data. Ann. Stat. **43**(5), 2225–2258 (2015)