

Article

Pertinent Prediction Intervals in Linear Regression

Dimitris N. Politis

Department of Mathematics and Halicioglu Data Science Institute, University of California San Diego, La Jolla, San Diego, CA 92093, USA; dpolitis@ucsd.edu

Abstract

In linear regression, a point predictor $\hat{Y}_f$ of a future response Y_f associated with a regressor value of interest $\underline{x}_f$ can easily be constructed. Since $\hat{Y}_f$ will always incur a prediction error, it is desirable to accompany the point predictor by a prediction interval, say $C(\underline{x}_f)$, that will contain the target Y_f with a pre-specified high probability, e.g., 90%. An estimated prediction interval, say $\hat{C}(\underline{x}_f)$, is called *pertinent* if its construction incorporates the variability of all estimators that are employed in the prediction problem. So far, pertinent prediction intervals have only been constructed via some form of bootstrap. However, resampling can be quite computationally expensive since the estimation/prediction problem has to be re-calculated on a large number of pseudo-scatterplots, each having the same sample size as the original one. The paper at hand proposes a short-cut that directly employs the asymptotic normal distribution of relevant estimators—as opposed to a bootstrap histogram—in order to capture their variability. The resulting prediction interval achieves pertinence without full-scale resampling, thus offering computational savings of orders of magnitude.

Keywords: bootstrap; coverage; prediction; regression; uncertainty

1. Introduction

Consider regression data $\{(Y_1, \underline{X}'_1), \dots, (Y_n, \underline{X}'_n)\}$. The response Y_i is assumed univariate while the regressor $\underline{X}_i$ is a $p \times 1$ vector. Note that the entries of $\underline{X}_i$ are in general real-valued but they can also be discrete, i.e., taking values on a countable subset of the real line; however, we will not consider categorical regressors as in [1].

Suppose that the $\{\underline{X}_1, \dots, \underline{X}_n\}$ were chosen (or randomly generated—see Assumption 1 in what follows) as taking the values $\{\underline{x}_1, \dots, \underline{x}_n\}$ respectively, and consider the usual linear regression model:

$$Y_i = \underline{x}'_i \beta + Z_i, \text{ for } i = 1, \dots, n, \quad (1)$$

where the Z_i are independent, identically distributed (i.i.d.) from some distribution F with $EZ_i = 0$ and $EZ_i^2 = \sigma^2$.

The goal is to predict a future response Y_f associated with a particular regressor value $\underline{x}_f$ of interest. Furthermore, we wish to construct a data-based prediction interval (PI), say $\hat{C}(\underline{x}_f)$, so that for some chosen $\alpha \in (0, 1)$ we have

$$P(Y_f \in \hat{C}(\underline{x}_f) | \underline{X}_f = \underline{x}_f) \approx 1 - \alpha \quad (2)$$

where the approximation is typically asymptotic as more data accrue. A review of the state-of-the-art of PI construction in different settings is given by [2].



Academic Editor: Wei Zhu

Received: 7 May 2026

Revised: 16 June 2026

Accepted: 17 June 2026

Published: 25 June 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

At first glance, this exercise seems to be a straightforward application of employing the quantiles of the conditional distribution $F_f(y) = P(Y_f \leq y | \underline{X}_f)$ which in many settings admits a consistent estimator, say $\hat{F}_f(y)$. Letting $G^{-1}(\gamma) = \inf\{x : G(x) \geq \gamma\}$ denote the quantile inverse of a distribution G , we can construct the ‘oracle’ equal-tailed PI $C(\underline{x}_f) = [F_f^{-1}(\alpha/2), F_f^{-1}(1 - \alpha/2)]$, and its estimated version $\hat{C}(\underline{x}_f) = [\hat{F}_f^{-1}(\alpha/2), \hat{F}_f^{-1}(1 - \alpha/2)]$. If $F_f(y)$ is continuous and strictly increasing for y in the neighborhood of the two quantiles of interest, i.e., for γ equal to $\alpha/2$ and $1 - \alpha/2$, then

$$P(Y_f \in C(\underline{x}_f) | \underline{X}_f = \underline{x}_f) = 1 - \alpha. \quad (3)$$

and

$$P(Y_f \in \hat{C}(\underline{x}_f) | \underline{X}_f = \underline{x}_f) \rightarrow 1 - \alpha \text{ as } n \rightarrow \infty \quad (4)$$

since $\hat{F}_f^{-1}(\gamma)$ will be a consistent estimator of $F_f^{-1}(\gamma)$; see e.g., Lemma 1.2.1 of [3].

Nevertheless, despite the asymptotic validity of Equation (4), the quantile-based PI $\hat{C}(\underline{x}_f)$ is well-known to suffer from finite-sample undercoverage, i.e.,

$$P(Y_f \in \hat{C}(\underline{x}_f) | \underline{X}_f = \underline{x}_f) < 1 - \alpha \text{ for finite } n. \quad (5)$$

This undercoverage has been illustrated empirically time and again—see e.g., [4] and the references therein—but recently some theoretical explanations have also been given; see [5,6]. The implicit reason for the finite-sample undercoverage is that by plugging-in the estimated quantiles as if they were true, the PI $\hat{C}(\underline{x}_f)$ fails to capture the variability of these estimated quantiles, i.e., it is neglecting part of the variance of the problem at hand. In other words, the quantile-based PI $\hat{C}(\underline{x}_f)$ is not *pertinent*; this term was coined by [7] to describe PIs that encapsulate the variability of all estimated quantities figuring in the PI construction.

Starting with the work of [8], there has been increased interest in PIs obtained via *conformal prediction*; see also [9]. Conformal prediction was originally developed for i.i.d. data. Modelling the regression pairs $(Y_1, \underline{X}'_1), \dots, (Y_n, \underline{X}'_n)$ as i.i.d., a conformal prediction PI $\tilde{C}(\underline{X}_f)$ can be constructed that satisfies

$$P(Y_f \in \tilde{C}(\underline{X}_f)) \geq 1 - \alpha. \quad (6)$$

However, the above finite-sample coverage guarantee is on the average, over all possible regressor values $\underline{X}_f$. If the practitioner desires $1 - \alpha$ coverage conditionally on a particular regressor value $\underline{X}_f = \underline{x}_f$, then the exchangeability of the pairs $(Y_i, \underline{X}'_i)$ breaks down and the finite-sample guarantee of (6) is lost; see [10,11]. There is an active body of literature developing variations of the conformal prediction theme to produce PIs that are asymptotically valid conditionally on $\underline{X}_f = \underline{x}_f$ as in Equation (2); see e.g., [12–14], as well as [15] and the references therein. However, none of these variations yields pertinent PIs; see [6] for a discussion.

Note that if the regression errors Z_i can be assumed to be Gaussian, then a pertinent PI with exact coverage is available; see Equation (A1) in the Appendix A. However, in the Big Data Era, the assumption of Gaussianity is typically seen to be unjustifiable, and often just plain wrong. Without having to resort to unnecessarily restrictive assumptions such as Gaussianity, the only general way so far to obtain pertinent PIs has been via some form of bootstrap. Furthermore, to achieve pertinence, the resampling procedure must be carefully crafted so that it captures the variability of all estimated quantities including the variability of estimating the point predictor; see e.g., Ch. 3.7 of [4].

Although the bootstrap approach gives generally good results in finite-samples (as well as asymptotically), it can be quite computationally expensive as it involves re-fitting

the regression over M bootstrap pseudo-samples; note that each pseudo-sample entails a new scatterplot of n data points. In the paper at hand, we describe a shortcut that gives pertinent PIs based on a limited Monte Carlo simulation as opposed to full-scale resampling. The new method is introduced in Section 3 in the linear regression framework that is outlined in Section 2. Finally, Section 4 reports the results of a small numerical experiment, and provides some practical recommendations. A brief Conclusions section is also provided as well as an Appendix A to intuitively describe the notion of pertinence.

2. Linear Regression Framework

We can re-write the linear regression model of Equation (1) as

$$\underline{Y}_n = X\underline{\beta} + \underline{Z}_n$$

where $\underline{Y}_n = (Y_1, \dots, Y_n)'$ and $\underline{Z}_n = (Z_1, \dots, Z_n)'$ are $n \times 1$ random vectors, $\underline{\beta}$ is a $p \times 1$ deterministic parameter vector with $p < n$, and X is an $n \times p$ design matrix with i th row given by vector $\underline{X}_i'$. Throughout the paper we will also assume:

Assumption 1. Assume that the design matrix X is either deterministic or that inference is conducted conditionally on the $\underline{X}_i'$ vectors that are assumed independent to the error vector $\underline{Z}_n$. Hence, without this being explicitly denoted, all probabilities and expectations will be tacitly understood to be conditional on the event $\{\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n\}$ where $\{\underline{x}_1, \dots, \underline{x}_n\}$ are some fixed values.

Consequently, the coverage of a PI will be evaluated conditionally on $\{\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n\}$ as well as $\underline{X}_f = \underline{x}_f$ as compared to the probability measure P_2 of [6].

Let $\hat{\underline{\beta}}$ be an estimator of $\underline{\beta}$ that is linear in the data $\underline{Y}_n$, i.e., $\hat{\underline{\beta}} = B\underline{Y}_n$ for some $p \times n$ matrix B ; the matrix B will typically depend on X but this will not be explicitly denoted. For example, if X has full rank, we can let $B = (X'X)^{-1}X'$ so that $\hat{\underline{\beta}}$ is the Least Squares (LS) estimator. If X is not of full rank, we can let $B = (X'X + \lambda I_p)^{-1}X'$ so that $\hat{\underline{\beta}}$ is a ridge regression estimator; here I_p is the $p \times p$ identity matrix and λ is a (small) positive constant. Other examples are readily available in the literature.

Let $\hat{\underline{\beta}}^{(i)}$ be the same estimator based on the dataset that has the pair $(Y_i, \underline{X}_i')$ omitted. In other words, $\hat{\underline{\beta}}^{(i)} = B^{(i)}\underline{Y}_n^{(i)}$ where $\underline{Y}_n^{(i)}$ is $\underline{Y}_n$ with Y_i omitted, and $B^{(i)}$ is B with its i th column omitted. The predictive and fitted residuals ($\tilde{z}_i$ and z_i respectively) corresponding to data point Y_i are defined in the usual manner, i.e., $\tilde{z}_i = Y_i - \underline{x}_i'\hat{\underline{\beta}}^{(i)}$ and $z_i = Y_i - \underline{x}_i'\hat{\underline{\beta}}$.

Remark 1 (LS residuals). In the LS case where $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}_n$ the predictive residuals can be easily obtained without re-fitting the regression since $\tilde{z}_i = z_i / (1 - h_i)$ where $h_i = \underline{x}_i'(X'X)^{-1}\underline{x}_i$ is the i th diagonal element of the ‘hat’ matrix $X(X'X)^{-1}X'$; see Theorem 10.1 of [16]. In this setting, we can also use the so-called studentized residuals $\hat{z}_i = z_i / \sqrt{1 - h_i}$ as defined by [17]. Assuming that the regression has an intercept term, Equation (10.12) of [16] further implies $1/n \leq h_i \leq 1$, yielding the ordering $|z_i| \leq |\hat{z}_i| \leq |\tilde{z}_i|$ for all i , i.e., the predictive residuals have the largest scale while the fitted residuals have the smallest.

Recall that the L_2 -optimal point predictor of Y_f given $\underline{X}_f = \underline{x}_f$ is the conditional expectation $E(Y_f | \underline{X}_f = \underline{x}_f) = \underline{x}_f'\underline{\beta}$. Since $\underline{\beta}$ is unknown, our practical point predictor of Y_f will be $\hat{Y}_f = \underline{x}_f'\hat{\underline{\beta}}$. In order to construct a pertinent PI, the goal is to approximate the distribution of the so-called ‘root’ $Y_f - \hat{Y}_f$. If we were to do so via bootstrap, one would proceed by resampling the residuals and creating new pseudo-scatterplots via the (fitted) model (1). In this setting, the transformation-based approach of [4] suggested resampling the predictive residuals $\tilde{z}_i$ —as opposed to the fitted residuals z_i —in order to alleviate the

undercoverage of bootstrap PIs first pointed out by [18]. The predictive residuals will be found useful in what follows as well.

3. Pertinent Prediction Intervals

Consider again the root $Y_f - \hat{Y}_f = Y_f - \underline{x}_f' \hat{\beta}$. Model (1) implies that $Y_f = \underline{x}_f' \beta + Z_f$ where $Z_f \sim F$. Hence,

$$Y_f - \hat{Y}_f = Z_f + \underline{x}_f' \beta - \underline{x}_f' \hat{\beta} = Z_f + E_f$$

where $E_f = \underline{x}_f' \beta - \underline{x}_f' \hat{\beta}$ is the estimation error. Standard regression conditions typically ensure that

$$E \underline{x}_f' \hat{\beta} = \underline{x}_f' \beta + o(1/\sqrt{n}) \quad \text{and} \quad \text{Var}(\underline{x}_f' \hat{\beta}) = O(1/n) \quad (7)$$

as $n \rightarrow \infty$. For instance, if $\hat{\beta}$ is the LS estimator in a well-specified model, then $E \hat{\beta} = \beta$. Further conditions on the design matrix are available to ensure the 2nd part of Equation (7); see [16].

Note that Equation (7) implies that $E_f \rightarrow 0$ in probability as $n \rightarrow \infty$. It is tempting to treat E_f as negligible, and hence construct a simple quantile-based PI for Y_f as

$$[\hat{Y}_f + \hat{F}^{-1}(\alpha/2), \hat{Y}_f + \hat{F}^{-1}(1 - \alpha/2)] \quad (8)$$

where $\hat{F}$ is a (uniformly) consistent estimator of F . For example, $\hat{F}$ can be the empirical distribution function (e.d.f.) of the residuals (fitted, predictive or studentized) after centering them to ensure that $\hat{F}$ is a distribution with mean zero. If F is continuous and strictly increasing in the neighborhood of the two quantiles of interest, namely $F^{-1}(\alpha/2)$ and $F^{-1}(1 - \alpha/2)$, then the latter are estimated consistently by $\hat{F}^{-1}(\alpha/2)$ and $\hat{F}^{-1}(1 - \alpha/2)$ respectively, implying that the quantile-based PI (8) has asymptotic coverage $1 - \alpha$. However, PI (8) will be referred to as *naive* in what follows because it ignores the term E_f and its variability, and typically exhibits finite-sample *undercoverage*.

Recall that $\hat{\beta}$ is a function of $\underline{Z}_n$ which is independent of Z_f . Hence, the distribution of $Z_f + E_f$ is given by $G = F * F_E$ where $*$ denotes convolution and F_E is the distribution of E_f . Consequently,

$$P(G^{-1}(\alpha/2) \leq Y_f - \hat{Y}_f \leq G^{-1}(1 - \alpha/2) | \underline{X}_f = \underline{x}_f) = 1 - \alpha \quad (9)$$

assuming G is continuous and strictly increasing at its tail quantiles $G^{-1}(\alpha/2)$ and $G^{-1}(1 - \alpha/2)$. Equation (9) then suggest the 'oracle' PI

$$[\hat{Y}_f + G^{-1}(\alpha/2), \hat{Y}_f + G^{-1}(1 - \alpha/2)]. \quad (10)$$

The above oracle PI has exact coverage $1 - \alpha$ but it involves the two unknown distributions, F and F_E . To construct a practical pertinent PI we propose to replace F and F_E by respective estimators $\hat{F}$ and $\hat{F}_E$. $\hat{F}$ was already mentioned; as regards F_E , recall that regularity conditions typically ensure that $\hat{\beta}$ is asymptotically normal, and therefore

$$\frac{E_f}{\sqrt{\text{Var}(\underline{x}_f' \hat{\beta})}} \Rightarrow N(0, 1) \quad \text{as } n \rightarrow \infty. \quad (11)$$

Coupled with the fact that $\text{Var}(\underline{x}_f' \hat{\beta}) = \sigma^2 \underline{x}_f' B B' \underline{x}_f$ we are led to defining $\hat{F}_E$ as a normal distribution with mean zero and variance $\hat{\sigma}^2 \underline{x}_f' B B' \underline{x}_f$; here, $\hat{\sigma}^2$ is a consistent estimator of σ^2 , e.g., the sample variance of the residuals (fitted, predictive or studentized). Although $\hat{F}$ may be a discontinuous distribution, the continuity of $\hat{F}_E$ implies that our estimator $\hat{G} = \hat{F} * \hat{F}_E$ is a continuous, strictly increasing distribution that is (uniformly) consistent for G .

The following lemma puts it all together.

Lemma 1. Assume conditions ensuring (7) and (11), and the (uniform) consistency of $\hat{F}$ and $\hat{\sigma}^2$ for F and σ^2 respectively. If G is continuous and strictly increasing at its tail quantiles $G^{-1}(\alpha/2)$ and $G^{-1}(1 - \alpha/2)$, then

$$P(\hat{G}^{-1}(\alpha/2) \leq Y_f - \hat{Y}_f \leq \hat{G}^{-1}(1 - \alpha/2) | \underline{X}_f = \underline{x}_f) \rightarrow 1 - \alpha \quad (12)$$

as $n \rightarrow \infty$.

Equation (12) leads us to construct our novel pertinent PI as

$$[\hat{Y}_f + \hat{G}^{-1}(\alpha/2), \hat{Y}_f + \hat{G}^{-1}(1 - \alpha/2)]. \quad (13)$$

Note that the Central Limit Theorem (CLT) of Equation (11), as well as the approximations of F and σ^2 by $\hat{F}$ and $\hat{\sigma}^2$, do require a reasonably large sample. But even with n as small as 100 these approximations can be good enough for practical application; see Section 4 for some numerical evidence. Moreover, recall that assumption (7) implies that $E_f = O_P(1/\sqrt{n})$ which is *not* negligible with n about 100—therefore the importance of actually including it in the PI construction; the following remark elaborates.

Remark 2 (Asymptotic validity). Lemma 1 implies that the pertinent PI (13) has asymptotic coverage $1 - \alpha$ conditionally on $\underline{X}_f = \underline{x}_f$. Although consistency is a *sine qua non*, it does not tell the whole story; recall that even the naive PI (8) is consistent. The PI (13) is called pertinent because it captures the variance of the estimation error E_f which the naive PI (8) treats as negligible. As shown in the simulation results of Section 4, the pertinent PI (13) alleviates the finite-sample undercoverage of the naive PI (8).

In principle, the convolution $\hat{G} = \hat{F} * \hat{F}_E$ could be carried out numerically; however, the following Monte Carlo algorithm makes its implementation straightforward.

Monte Carlo Algorithm for Pertinent Prediction Intervals

1. Compute the estimator $\underline{\hat{\beta}} = BY_n$, the point predictor $\hat{Y}_f = \underline{x}_f' \underline{\hat{\beta}}$, and the residuals of choice (fitted, predictive or studentized).
2. Compute an estimator $\hat{\sigma}^2$ of σ^2 ; e.g., $\hat{\sigma}^2$ could be the sample variance of the chosen residuals. [In the case of Least Squares, the unbiased estimator $(n - p)^{-1} \sum_{i=1}^n z_i^2$ could also be entertained.]
3. Compute an estimator $\hat{F}$ of F ; e.g., $\hat{F}$ can be the e.d.f. of the residuals of choice after centering to ensure that $\hat{F}$ is a distribution with mean zero.
4. Choose a large integer M , e.g., $M \geq 1000$, and generate $Z_1^*, \dots, Z_M^*$ i.i.d. from $\hat{F}$; the Z_i^* s can be thought of as pseudo-residuals. [If $\hat{F}$ is an e.d.f. that puts mass $1/n$ on values $a_1, \dots, a_n$ (say), then generating $Z_1^*, \dots, Z_M^*$ can be done by drawing randomly (with replacement) from the set $\{a_1, \dots, a_n\}$.]
5. Generate $E_1^*, \dots, E_M^*$ i.i.d. from a normal distribution with mean zero and variance $\hat{\sigma}^2 \underline{x}_f' B B' \underline{x}_f$.
6. Define $W_i = Z_i^* + E_i^*$ for $i = 1, \dots, M$, and compute $\hat{G}^{-1}(\gamma)$ as the γ sample quantile of $W_1, \dots, W_M$, i.e., $\hat{G}^{-1}(\gamma) = W_{[\gamma M]}$ where $W_{[1]} \leq W_{[2]} \leq \dots \leq W_{[M]}$ are the order statistics.
7. Use the above $\hat{G}^{-1}(\gamma)$ with γ equal to $\alpha/2$ and $1 - \alpha/2$ to construct the pertinent PI (13).

Remark 3 (Choice of residuals). In the LS setting of Remark 1, note that our Equation (7) implies that $h_i \rightarrow 0$ as $n \rightarrow \infty$, and hence all residuals (fitted, predictive or studentized) are asymptotically equivalent. Therefore, asymptotic considerations can not help us in choosing which residuals to

employ for the construction of the estimators $\hat{\sigma}^2$ and $\hat{F}$ in the above algorithm. Note, however, that the validity of Equation (7) hinges on either p being finite, or at least that $p \ll n$. For example, if p is of the same order of magnitude as n , then the h_i s will not be close to 0 even for large n , and the fitted, predictive and studentized residuals will remain distinct. The numerical experiment of Section 4 will provide some finite-sample guidance on our choice of residuals and the effect of the relative magnitude of p to n .

Remark 4 (Computational cost as compared to bootstrap). Note that sampling from $\hat{F}$ is effectively a bootstrap procedure. However, in the above Monte Carlo algorithm, sampling from $\hat{F}$ is done only once, i.e., we are drawing just one bootstrap sample of size M . Suppose that the estimators $\hat{\beta}$, $\hat{\sigma}^2$ and $\hat{F}$ can be computed in $O(n^\zeta)$ operations for some $\zeta \geq 1$. Then, the above algorithm has computational cost of $O(n^\zeta) + O(M)$ operations. By contrast, the pertinent version of the residual-based bootstrap requires drawing M resamples each of size n from $\hat{F}$; using each of the M resamples as a set of pseudo-residuals, a new pseudo-scatterplot is created upon which $\hat{\beta}$, $\hat{\sigma}^2$, etc. are recomputed—see the MF/MB procedure in Part II of [4] for more details. Hence, the pertinent version of the residual-based bootstrap requires $O(Mn^\zeta)$ operations which is more expensive by orders of magnitude. To elaborate, consider the realistic case where $n^\zeta \geq 1000$; since M is also taken not less than 1000, the result is that the above Monte Carlo algorithm can be a thousand times faster than the pertinent version of the residual bootstrap.

To elaborate, consider the popular example of LASSO regression that can be computed in $O(p^3 + p^2n)$ operations as long as $p < n$; see [19]. Suppose that $p \sim \sqrt{n}$ in which case LASSO regression is computable in $O(n^\zeta)$ operations with $\zeta = 2$. Let $M = 1000$, and that the sample size satisfies $n > 33$. Then, we have $M < n^2$ and our Monte Carlo algorithm is computable in $O(n^2)$ operations, i.e., same order as computing a single LASSO regression. By contrast, the residual-based bootstrap requires $O(Mn^2)$ operations, i.e., it is like fitting 1000 LASSO regressions.

4. Numerical Results

We now conduct a small numerical experiment with three goals: (a) to show that the pertinent PI (denoted PPI) of Equation (13) has better finite-sample coverage than the naive PI of Equation (8); (b) to provide some finite-sample guidance on which type of residuals to employ in the PPI algorithm; and (c) to see the effect of the magnitude of p (relative to n) in our choice of residuals, and the resulting PPI performance.

Throughout this section, we will focus on the LS estimator $\hat{\beta} = (X'X)^{-1}X'Y_n$ where X is an $n \times p$ matrix having a column of 1s as its first column.

The structure of our numerical experiment is as follows:

1. Choose p and n ; for our main results we will employ the following combinations for (p, n) : (10,50), (20,50), (20,100), (40,100).
2. Generate all the elements of X (other than its first column) as i.i.d. standard normal.
3. Choose the error distribution F ; for our purposes we will take F as either standard normal or Laplace (i.e., two-sided exponential) with mean zero and variance one.
4. Choose $\underline{\beta}$, the α level, and the regressor of future interest $\underline{x}_f$; for our purposes, $\underline{\beta}$ is generated as $3 \cdot (U_1, \dots, U_p)'$ where $U_1, \dots, U_p$ are i.i.d. Uniform $[0, 1]$, the α level is chosen as 0.10, and $\underline{x}_f$ is a row of 1s.
5. Generate $Z_1, \dots, Z_n$ i.i.d. from F , and use them to generate $Y_1, \dots, Y_n$ from model (1).
6. Fit the data via LS and compute $\hat{\beta}$ and $\hat{Y}_f$; also compute the three sets of residuals: fitted, studentized and predictive.
7. Calculate four PIs for Y_f : PI.naive of Equation (8), and three versions of the PPI of Equation (13): using the fitted residuals (PPI.fitted), using the studentized residuals (PPI.stud), and using the predictive residuals (PPI.pred). The three PPIs employ the Monte Carlo algorithm using $M = 1000$. Compute the lengths (LEN) of the four PIs.

8. Further generate $Z_{n+1}, \dots, Z_{n+500}$ i.i.d. from F , and use them to generate 500 realizations of Y_f to be used to test the PI coverage. The k th realization $Y_f^{(k)}$ is generated as $Y_f^{(k)} = \underline{x}_f' \beta + Z_{n+k}$.
9. Calculate the coverage (CVR) of each PI using the 500 realizations of Y_f , i.e., CVR of a PI is $500^{-1} \sum_{k=1}^{500} \mathbf{1}\{Y_f \in \text{PI}\}$ where $\mathbf{1}$ denotes the indicator function.
10. Repeat steps 5 to 9 many times—in our case 200. Each of these $j = 1, \dots, 200$ replications has its own $\text{CVR}[j]$ and $\text{LEN}[j]$ associated with each PI.

Finally:

- (a) Calculate the average CVR of each method over the 200 replications.
- (b) Plot a histogram of $\text{CVR}[j]$ for $j = 1, \dots, 200$ to portray the CVR variability, and compute what percentage of these CVRs fall below the target level of $1 - \alpha$.
- (c) Calculate the average LEN of each method over the 200 replications; also calculate the variability of LEN via its sample standard deviation.

To elaborate on the PI constructions:

- PI.naive and PPI.fitted employ as $\hat{F}$ the e.d.f. of the fitted residuals. The latter further uses $\hat{\sigma}^2$ to be the unbiased estimator $(n - p)^{-1} \sum_{k=1}^n z_k^2$. Recall that having a column of 1s in X ensures $\sum_{k=1}^n z_k = 0$.
- PPI.stud employs $\hat{F}$ the e.d.f. of the centered studentized residuals, i.e.,

$$\hat{F}(a) = n^{-1} \sum_{i=1}^n \mathbf{1}\{\hat{z}_i - n^{-1} \sum_{k=1}^n \hat{z}_k \leq a\}.$$

It also employs $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n [\hat{z}_i - n^{-1} \sum_{k=1}^n \hat{z}_k]^2$.

- Finally, PPI.pred employs $\hat{F}$ the e.d.f. of the centered predictive residuals, i.e.,

$$\hat{F}(a) = n^{-1} \sum_{i=1}^n \mathbf{1}\{\tilde{z}_i - n^{-1} \sum_{k=1}^n \tilde{z}_k \leq a\}.$$

It also employs $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n [\tilde{z}_i - n^{-1} \sum_{k=1}^n \tilde{z}_k]^2$.

The main results of our experiment are presented in Tables 1–9. There were also nine pages of CVR histograms associated with the results of each of the nine tables. To save space, we only present five representative ones in Figures 1–5; the omitted histograms were very similar to the ones included here. Note that each of these histograms is an approximation of the probability distribution of the random variable

$$P(Y_f \in \text{PI} | \underline{X}_f = \underline{x}_f, \underline{Y}_n). \quad (14)$$

Recall that per Assumption 1 the above probability is also conditional on the event $\{\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n\}$ without it being explicitly denoted.

Table 1. Performance of 90% PI constructions: mean coverage levels (CVR), % of CVRs that are <0.9, mean PI length (with standard deviation in parentheses); case of Normal errors.

$p = 10, n = 50$	CVR	% < 0.9	Length (St.Dev.)
PI.naive	0.765	97%	2.79 (0.37)
PPI.fitted	0.837	80.5%	3.31 (0.42)
PPI.stud	0.874	60%	3.62 (0.46)
PPI.pred	0.913	29%	4.06 (0.51)

Table 2. Entries as in Table 1; case of Laplace errors.

$p = 10, n = 50$	CVR	% < 0.9	Length (St.Dev.)
PI.naive	0.812	90%	2.74 (0.48)
PPI.fitted	0.865	66%	3.24 (0.55)
PPI.stud	0.889	48%	3.52 (0.60)
PPI.pred	0.917	27%	4.00 (0.68)

Table 3. Entries as in Table 1; case of Normal errors.

$p = 10, n = 50$	CVR	% < 0.9	Length (St.Dev.)
PI.naive	0.680	100%	2.45 (0.32)
PPI.fitted	0.845	72%	3.41 (0.43)
PPI.stud	0.899	39%	4.01 (0.53)
PPI.pred	0.960	8%	5.27 (0.70)

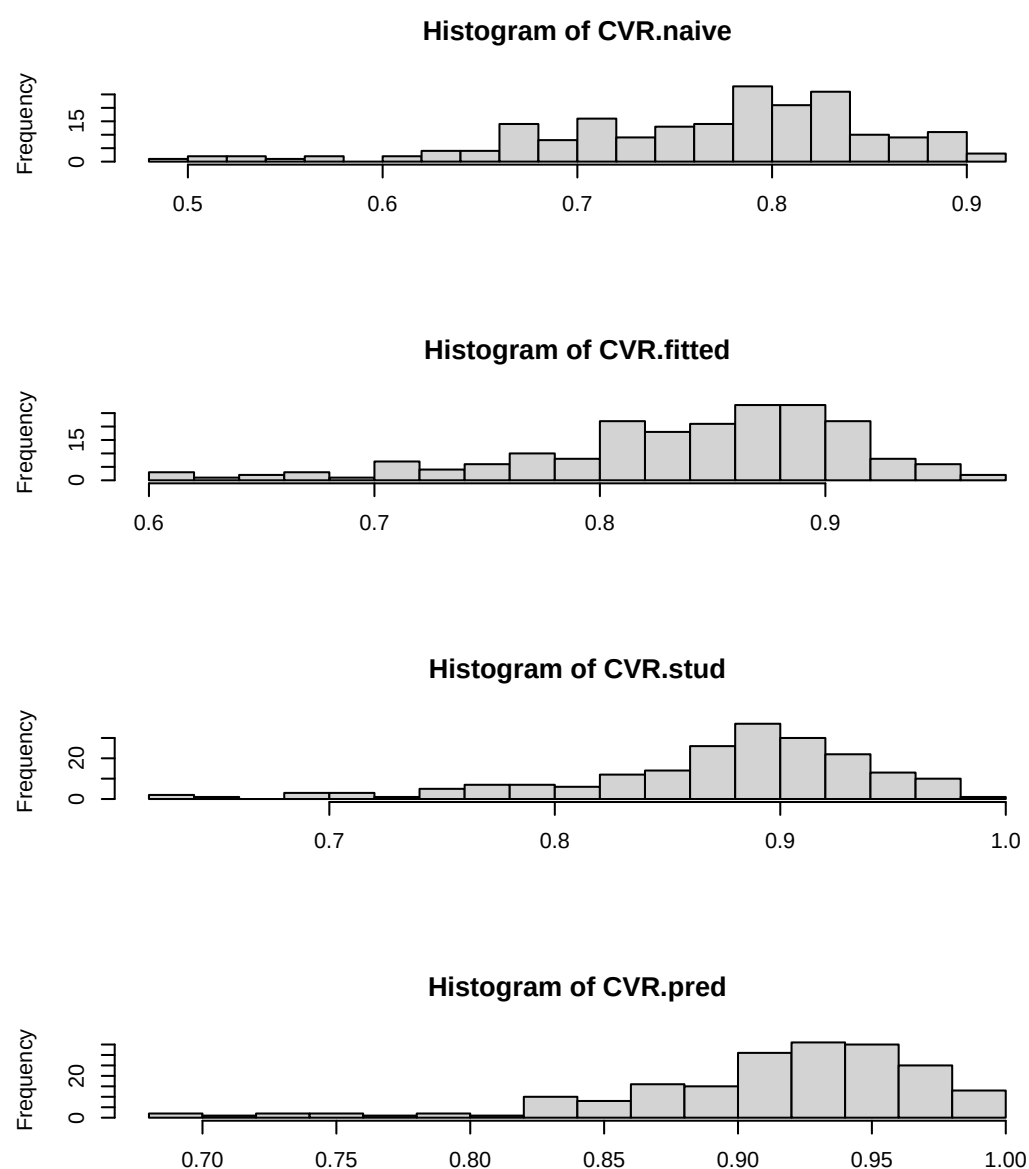
**Figure 1.** CVR histograms; case $p = 10, n = 50$; Normal errors.

Table 4. Entries as in Table 1; case of Laplace errors.

$p = 10, n = 50$	CVR	% < 0.9	Length (St.Dev.)
PI.naive	0.639	99%	2.40 (0.41)
PPI.fitted	0.842	60%	3.71 (0.59)
PPI.stud	0.888	37%	4.35 (0.69)
PPI.pred	0.950	10%	5.71 (0.99)

Table 5. Entries as in Table 1; case of Normal errors.

$p = 10, n = 50$	CVR	% < 0.9	Length (St.Dev.)
PI.naive	0.791	97.5%	2.87 (0.27)
PPI.fitted	0.860	73%	3.38 (0.28)
PPI.stud	0.893	44%	3.71 (0.32)
PPI.pred	0.925	16%	4.13 (0.35)

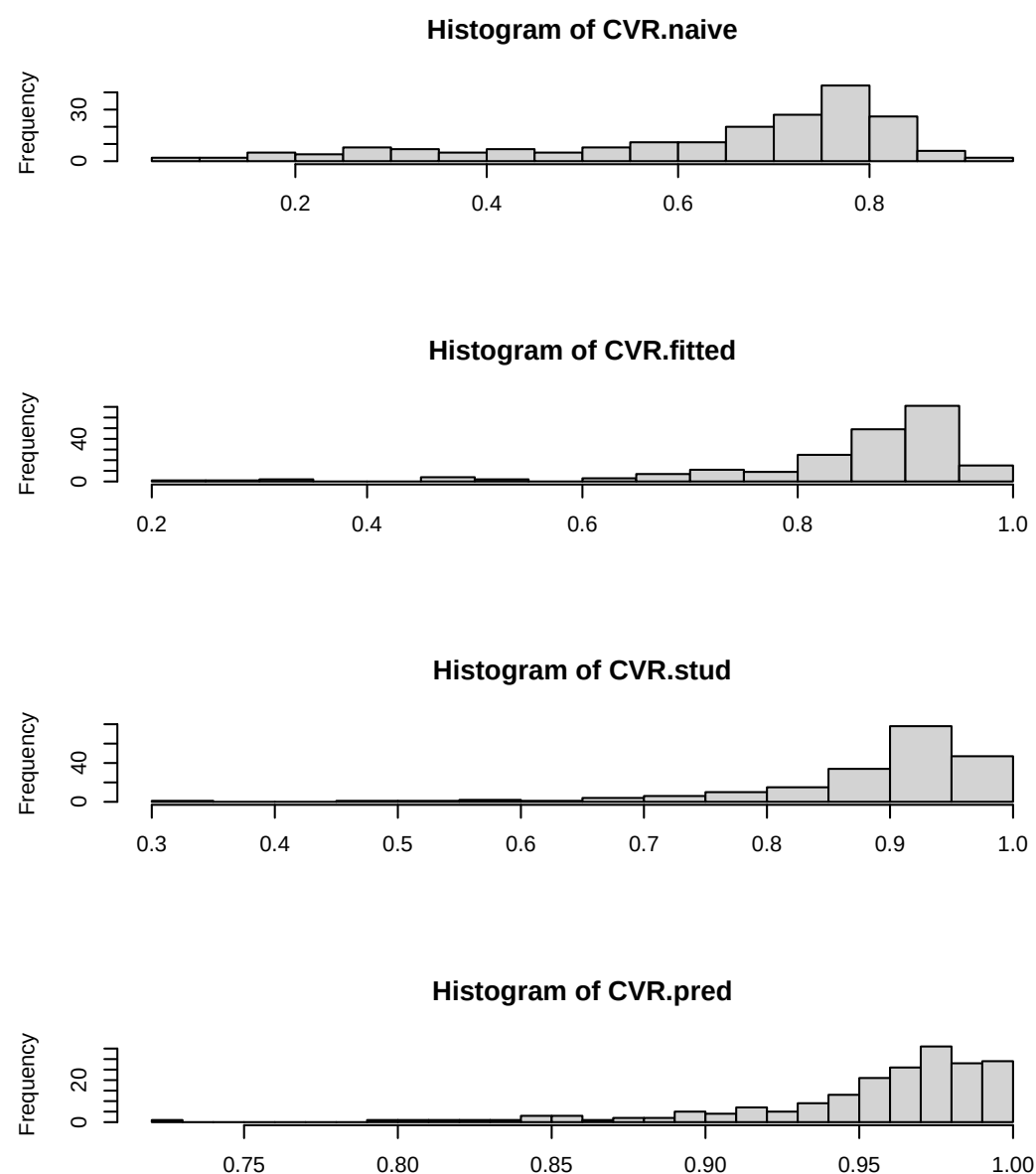
**Figure 2.** CVR histograms; case $p = 20, n = 50$; Laplace errors.

Table 6. Entries as in Table 1; case of Laplace errors.

$p = 10, n = 50$	CVR	% < 0.9	Length (St.Dev.)
PI.naive	0.832	87%	2.77 (0.35)
PPI.fitted	0.868	55%	3.37 (0.40)
PPI.stud	0.895	31.5%	3.66 (0.41)
PPI.pred	0.927	16.5%	4.10 (0.51)

Table 7. Entries as in Table 1; case of Normal errors.

$p = 10, n = 50$	CVR	% < 0.9	Length (St.Dev.)
PI.naive	0.676	100%	2.55 (0.26)
PPI.fitted	0.854	61%	3.82 (0.36)
PPI.stud	0.890	44.5%	4.34 (0.41)
PPI.pred	0.963	9.5%	5.60 (0.53)

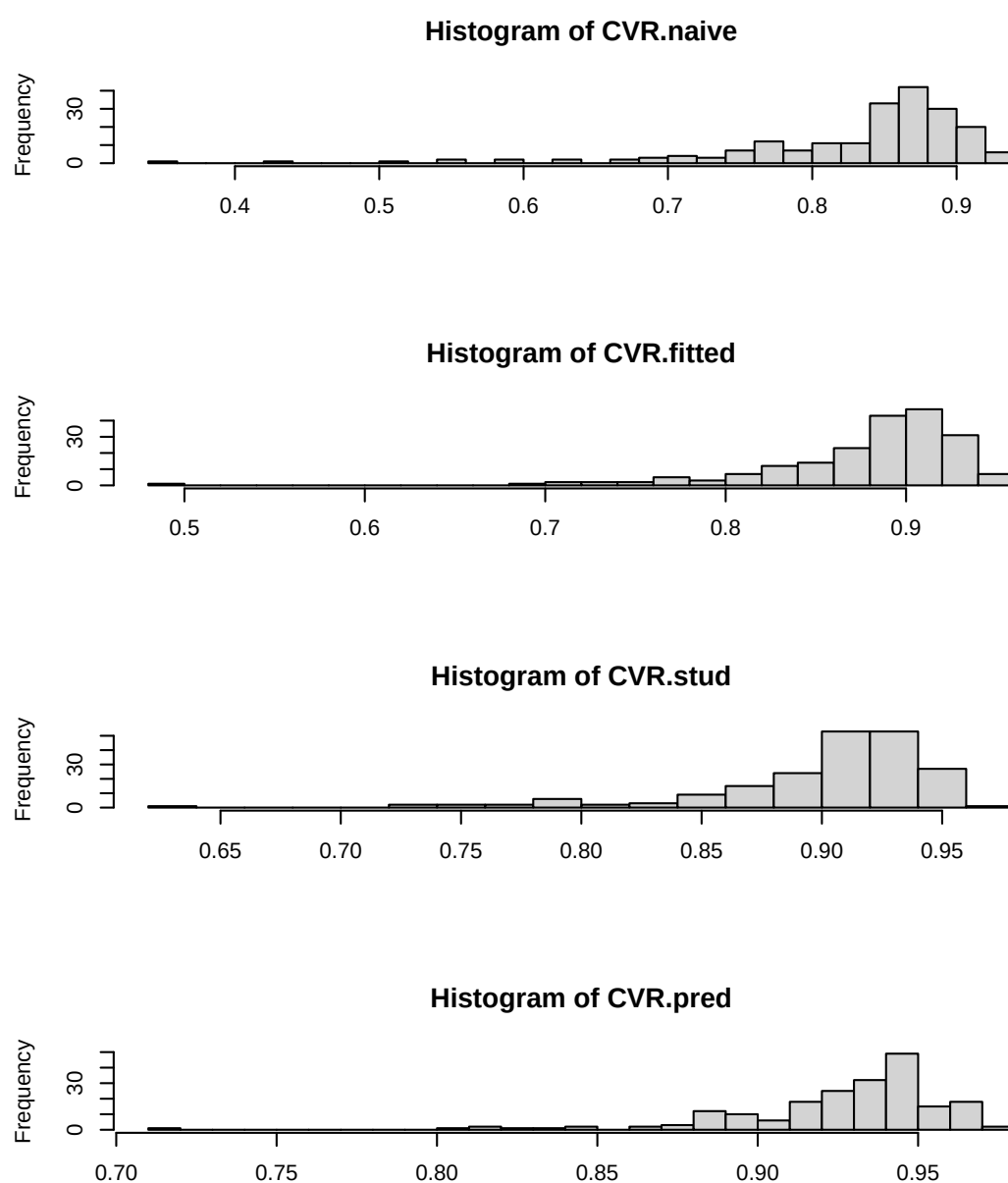
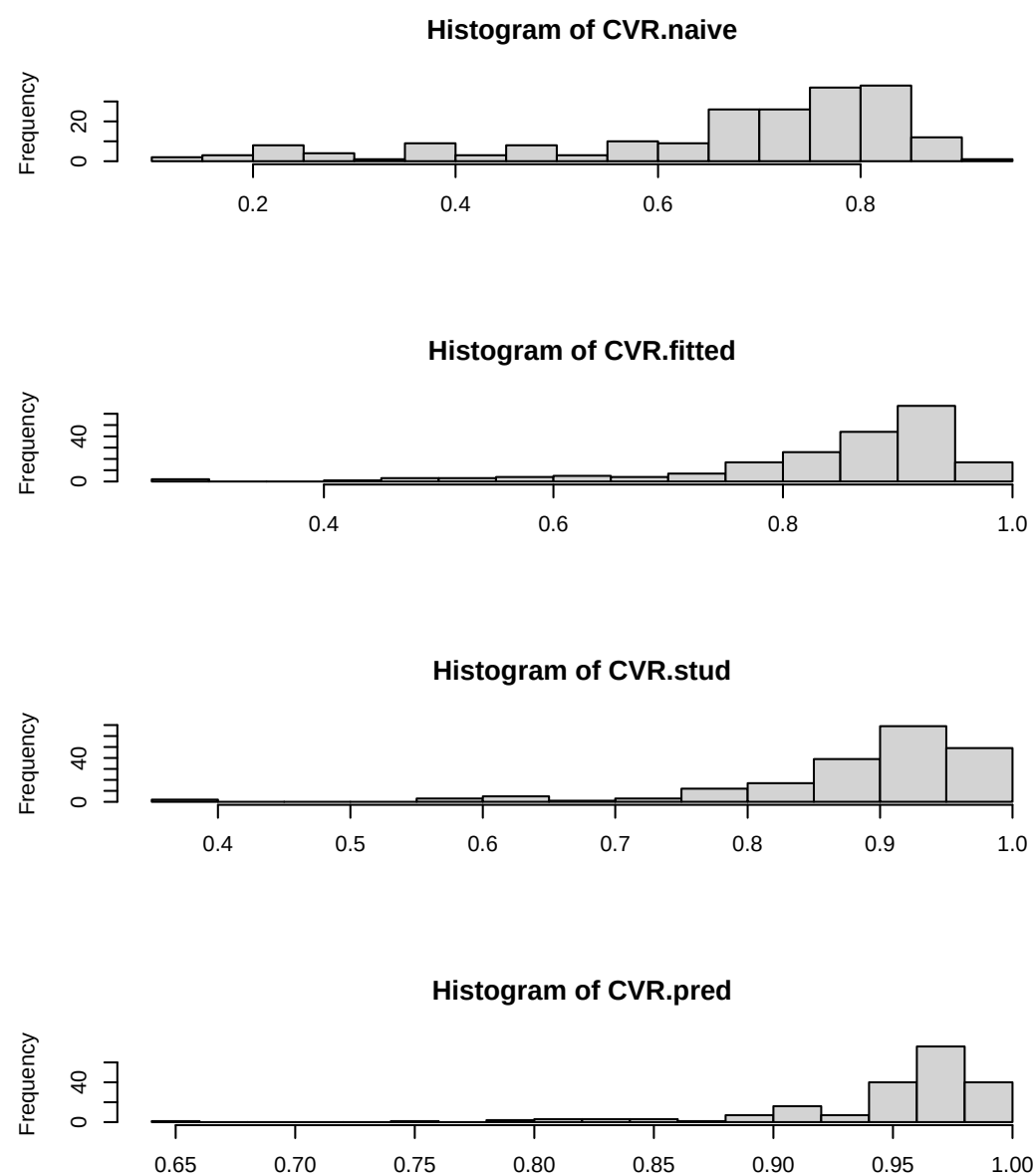
**Figure 3.** CVR histograms; case $p = 20, n = 100$; Normal errors.

Table 8. Entries as in Table 1; case of Laplace errors.

$p = 10, n = 50$	CVR	% < 0.9	Length (St.Dev.)
PI.naive	0.667	99%	2.47 (0.35)
PPI.fitted	0.845	57%	3.75 (0.53)
PPI.stud	0.888	39%	4.26 (0.59)
PPI.pred	0.950	10%	5.54 (0.79)

Table 9. Entries as in Table 1; case of Normal errors.

$p = 10, n = 50$	CVR	% < 0.9	Length (St.Dev.)
PI.naive	0.866	73%	3.12 (0.29)
PPI.fitted	0.887	60.5%	3.28 (0.29)
PPI.stud	0.897	51%	3.36 (0.30)
PPI.pred	0.908	38%	3.45 (0.31)

**Figure 4.** CVR histograms; case $p = 40, n = 100$; Laplace errors.

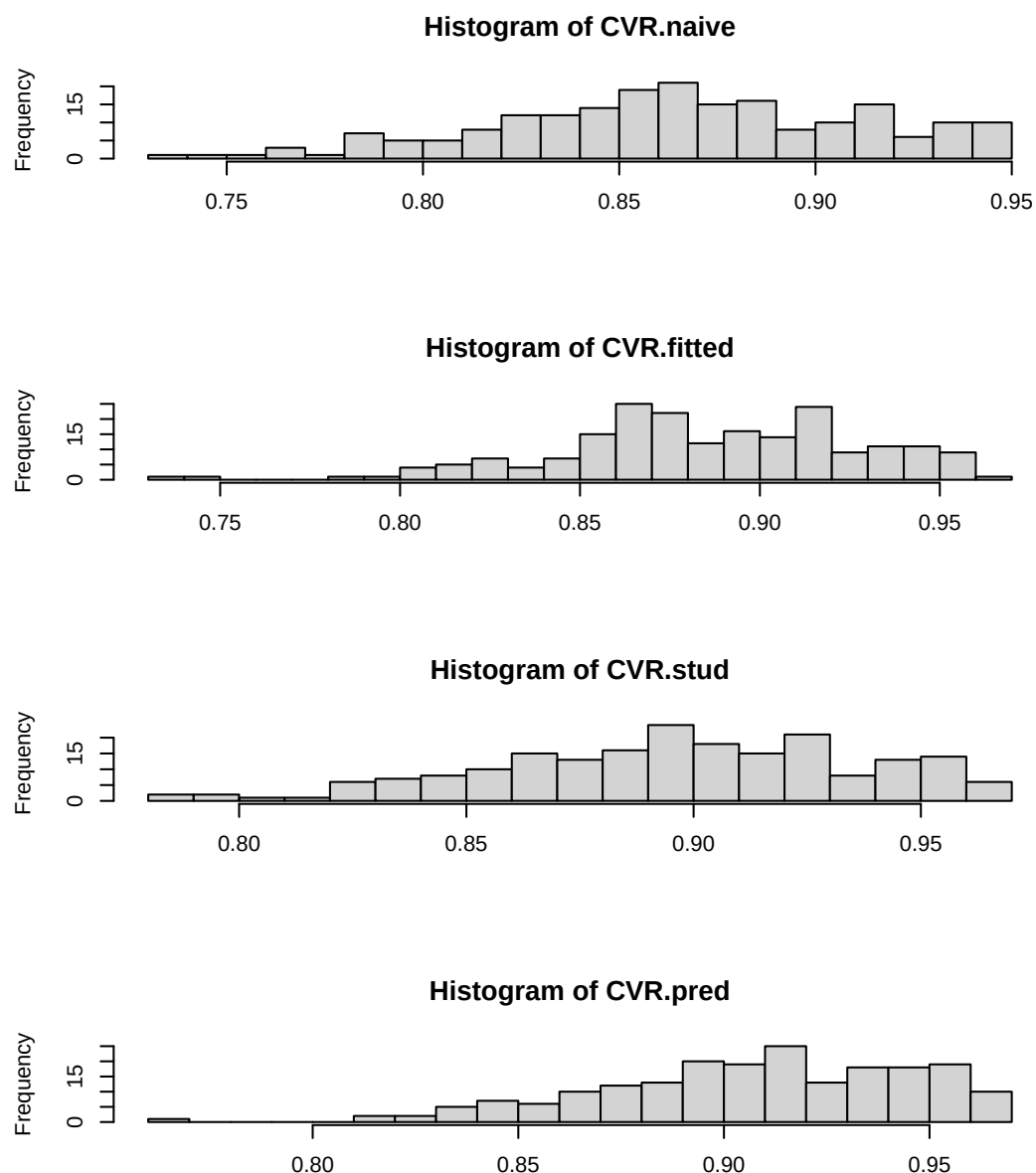


Figure 5. CVR histograms; case $p = 5$, $n = 100$; normal errors.

Being a function of the sample $\underline{Y}_n$, the quantity in Equation (14) is a random variable that varies across samples. Our numerical experiment tries to encapsulate this variability in an analogous way as in Figure 2 of [6]. Note that the mean CVR reported in the first column of Tables 1–9 gives the center of each of these histograms, namely

$$P(Y_f \in \text{PI} | \underline{X}_f = \underline{x}_f) \quad (15)$$

which is the expected value of the random variable (14) averaged over all possible $\underline{Y}_n$ samples.

From the numerical findings of Tables 1–9 we can draw several interesting conclusions:

- The undercoverage of PI.naive is strongly illustrated in all Tables and Figures.
- All three of our proposed PPIs have larger coverage than PI.naive; furthermore, we observe the ordering:

$$\text{CVR}[\text{PI.fitted}] < \text{CVR}[\text{PI.stud}] < \text{CVR}[\text{PI.pred}]$$

which is not unexpected in view of Remark 1 that compares the scales of the three types of residuals.

- It is clear that PI.fitted falls short of achieving acceptable coverage for finite samples; this can be attributed to the known tendency of the fitted residuals to underestimate the scale of the errors—see [20].
- The preferable methods seems to be either PPI.stud or PPI.pred; the latter appears to overcover in several instances so it can be thought of as conservative.
- The variability of the CVRs that is illustrated by the histograms in Figures 1–4 is striking. Also interesting is the high proportion of CVRs (conditional on the sample) that are below the nominal level over the 200 replications. PPI.pred is more conservative in that respect as well, i.e., having the smallest proportion of PIs that undercover.
- In our Tables 1–8, the ratio p/n was either 0.2 or 0.4 which can not be considered small; neither a sample size of 50 or 100 can be considered particularly large. Despite the fact that these setups are quite challenging, both PPI.stud and PPI.pred provided reliable performances (with the latter being more conservative).
- Interestingly, there is no appreciable difference of PI performance going from normal errors to Laplace even though in the latter case the LS estimator does not coincide with the minimum maximum likelihood estimator. The only side-effect of the heavier tails of the Laplace distribution appears to be a slightly heavier left tail in some of our histograms.

As discussed in Remark 3, the standard asymptotics kick in when $p/n \rightarrow 0$ in which case the fitted, studentized and predictive residuals become approximately equal to each other. In that case the performance of all three PPIs is expected to be good, although PPI.stud and PPI.pred will still be preferable over the other two PIs. To corroborate this claim, we offer the results of an additional simulation in the ‘easy’ case of $p = 5$ and $n = 100$, i.e., when p/n is small; as Table 9 shows, PPI.stud and PPI.pred have mean CVRs of 0.897 and 0.908 respectively that are practically equivalent to the nominal 90%. So are two methods PPI.stud and PPI.pred equivalent in this ‘easy’ case?

The above question allows us to think harder on our design goals for a successful PI construction. Until now, our target was to have CVR (15), i.e., $P(Y_f \in \text{PI} | \underline{X}_f = \underline{x}_f)$ close to the nominal 90% (and ideally not below the nominal). However, Figure 5 shows the histogram of the CVR (14) that has Equation (15) as its center. In order for the center to be about 90%, the CVR (14) should take values both below and above 90%. Indeed, as shown in Corollary 4.1 of [21], PIs based on the residual bootstrap will exhibit large-sample histograms with their median at the nominal level, i.e., asymptotically the histogram will have 50% of occurrences below the nominal and 50% above.

By the Law of Large Numbers, such a histogram will have progressively smaller and smaller range since asymptotically

$$P(Y_f \in \text{PI} | \underline{X}_f = \underline{x}_f, \underline{Y}_n) \rightarrow E(P(Y_f \in \text{PI} | \underline{X}_f = \underline{x}_f, \underline{Y}_n)) = P(Y_f \in \text{PI} | \underline{X}_f = \underline{x}_f)$$

where the above expectation is taken over the distribution of $\underline{Y}_n$. In spite of the above, it seems preferable that a desirable finite-sample histogram of $P(Y_f \in \text{PI} | \underline{X}_f = \underline{x}_f, \underline{Y}_n)$ would have less mass below the nominal 90% (as compared to above the nominal), i.e., only a minority of the CVRs (14) would be below the nominal 90%. In this sense, PPI.pred appears preferable to PPI.stud even here since, as shown in Table 9, PPI.stud and PPI.pred have respectively 51% and 38% of occurrences below the nominal.

5. Conclusions

An interesting problem in statistics is the construction of prediction intervals (PI) in linear regression without assuming Gaussianity which is often not justifiable. We desire PIs that have (a) asymptotic conditional validity, and (b) are ‘pertinent’, i.e., are able to

capture/incorporate the variability associated with estimated quantities employed in the PI construction. Such PIs in the past have been traditionally constructed via resampling.

The paper at hand proposes a short-cut that directly employs the asymptotic normal distribution of relevant estimators—as opposed to a bootstrap histogram—in order to capture their variability. The resulting prediction interval achieves the property of pertinence without full-scale resampling, thus offering computational savings of orders of magnitude over the bootstrap.

Asymptotic conditional validity of the new approach is shown under general conditions using either fitted, studentized or predictive residuals. Numerical work confirms that the new method has good finite-sample performance as well, and suggests the use of predictive residuals in order to construct conservative PIs, i.e., resulting into the smallest proportion of PIs that undercover.

Funding: This research was partially supported by NSF grant DMS 24-13718.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: Many thanks are due to three anonymous reviewers for their constructive suggestions that helped improve the paper.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. An Intuitive Explanation of Pertinence

The idea of pertinent intervals, i.e., intervals that possess the property of pertinence, is meant to describe prediction intervals that capture/incorporate the variability of estimated quantities in their construction. To explain it intuitively, consider the textbook setup of linear regression with Gaussian errors, i.e., model (1) with $Z_i \sim \text{i.i.d. } N(0, \sigma^2)$, and $\hat{\beta}$ being the Least Squares estimator.

Let $S^2 = \text{RSS}/(n - p)$, where RSS is the Residual Sum of Squares, i.e., $\sum_{i=1}^n z_i^2$. Then, conditioning on $X_f = x_f$ and on X (as in Assumption 1), we can define an exact $(1 - \alpha)100\%$ PI for Y_f as

$$\hat{Y}_f \pm t_{n-p}(\alpha/2)S\sqrt{1 + x_f'(X'X)^{-1}x_f}; \quad (\text{A1})$$

see Equation (5.27) in [16].

A different practitioner might use the $\alpha/2$ standard normal quantile $z(\alpha/2)$ instead of the t -distribution quantile $t_{n-p}(\alpha/2)$, and construct the PI

$$\hat{Y}_f \pm z(\alpha/2)S\sqrt{1 + x_f'(X'X)^{-1}x_f}. \quad (\text{A2})$$

Under standard assumptions on the structure of the design matrix X ensuring that the minimum eigenvalue of $X'X$ diverges to infinity, we have $x_f'(X'X)^{-1}x_f \rightarrow 0$ as $n \rightarrow \infty$ (with p held constant or at least $p \ll n$). One can then consider the simplified prediction interval

$$\hat{Y}_f \pm z(\alpha/2)S. \quad (\text{A3})$$

Under the normality of F , all three PIs, namely (A1), (A2) and (A3), are asymptotically valid. However, interval (A1) is the optimal one with exact finite-sample coverage $1 - \alpha$. Moreover, interval (A2) has better finite-sample coverage than interval (A3) as it captures the estimation variability of the linear model, i.e., the term $\text{Var}(x_f'\hat{\beta}) = \sigma^2 x_f'(X'X)^{-1}x_f$ which interval (A3) treats as negligible. That is why interval (A3) can be called ‘naive’, while interval (A2) can be called ‘pertinent’.

In general, an exact interval such as (A1) will not be available because of potential non-normality of the errors. Furthermore, if F is not Gaussian, then neither of the two intervals (A2) and (A3) will be valid, not even asymptotically. As mentioned already, in the absence of Gaussianity, the quantile-based PI (8) will be asymptotically valid under standard assumptions but it will be ‘naive’ as it neglects the variability inherent in estimation. To improve upon it and come up with a pertinent PI, past literature typically required a bootstrap procedure as described by [17]; see also Section 3.6 of [4]. The paper at hand proposes a new alternative method to construct a pertinent PI in linear regression without full-scale resampling.

References

1. Tian, H.; Huang, L.; Cheng, C.-Y.; Zhang, L. Regression models with ordered multiple categorical predictors. *J. Stat. Comput. Simul.* **2018**, *88*, 3164–3178. [\[CrossRef\]](#)
2. Tian, Q.; Nordman, D.J.; Meeker, W.Q. Methods to Compute Prediction Intervals: A Review and New Results. *Stat. Sci.* **2022**, *37*, 580–597. [\[CrossRef\]](#)
3. Politis, D.N.; Romano, J.P.; Wolf, M. *Subsampling*; Springer: New York, NY, USA, 1999.
4. Politis, D.N. *Model-Free Prediction and Regression: A Transformation-Based Approach to Inference*; Springer: New York, NY, USA, 2015.
5. Bai, Y.; Mei, S.; Wang, H.; Xiong, C. Understanding the under-coverage bias in uncertainty Estimation. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Virtually, 6–14 December 2021.
6. Wang, Y.; Politis, D.N. Model-free bootstrap and conformal prediction in regression: Conditionality, conjecture testing, and pertinent prediction intervals. *J. Nonparametr. Stat.* **2026**, *38*, 311–345. [\[CrossRef\]](#)
7. Politis, D.N. Model-free Model-fitting and Predictive Distributions. *Test* **2013**, *22*, 183–221. [\[CrossRef\]](#)
8. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: Berlin/Heidelberg, Germany, 2005.
9. Shafer, G.; Vovk, V. A Tutorial on Conformal Prediction. *J. Mach. Learn. Res.* **2008**, *9*, 371–421.
10. Lei, J.; Wasserman, L. Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Ser. B* **2014**, *76*, 71–96. [\[CrossRef\]](#)
11. Foygel Barber, R.; Candés, E.J.; Ramdas, A.; Tibshirani, R.J. The limits of distribution-free conditional predictive inference. *Inf. Inference* **2021**, *10*, 455–482.
12. Lei, J.; G’Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-free Predictive Inference for Regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. [\[CrossRef\]](#)
13. Romano, Y.; Patterson, E.; Candés, E.J. Conformalized quantile regression. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
14. Chernozhukov, V.; Wüthrich, K.; Zhu, Y. Distributional Conformal Prediction. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2107794118. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Duchi, J.C. A Few Observations on Sample-Conditional Coverage in Conformal Prediction. *arXiv* **2025**, arXiv:2503.00220.
16. Seber, G.A.F.; Lee, A.J. *Linear Regression Analysis*, 2nd ed.; Wiley: New York, NY, USA, 2003.
17. Stine, R.A. Bootstrap prediction intervals for regression. *J. Am. Stat. Assoc.* **1985**, *80*, 1026–1031. [\[CrossRef\]](#)
18. Efron, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Stat. Assoc.* **1983**, *78*, 316–331. [\[CrossRef\]](#)
19. Efron, B.; Hastie, T.; Johnstone, J.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499. [\[CrossRef\]](#)
20. Foygel Barber, R.; Candés, E.J.; Ramdas, A.; Tibshirani, R.J. Predictive inference with the jackknife+. *Ann. Stat.* **2021**, *49*, 486–507. [\[CrossRef\]](#)
21. Zhang, Y.; Politis, D.N. Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees. *Inf. Inference* **2023**, *12*, 157–209. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.