Quantile regression with multiple proxy variables

Dongyoung Go | Ick Hoon Jin | Jongho Im

Data integration has become increasingly popular owing to the availability of multiple data sources. This study considered quantile regression estimation when a key covariate had multiple proxies across several datasets. In a unified estimation procedure, the proposed method incorporates multiple proxies that have both linear and nonlinear relationships with the unobserved covariates. The proposed approach allows the inference of both the quantile function and unobserved covariates and does not require the quantile function's linearity. Simulation studies have demonstrated that this methodology successfully integrates multiple proxies and reveals quantile relationships for a wide range of nonlinear data. The proposed method is applied to administrative data obtained from the Survey of Household Finances and Living Conditions provided by Statistics Korea, to specify the relationship between assets and salary income in the presence of multiple income records.

KEYWORDS
data integration, measurement error model, natural cubic spline, record linkage

1 | INTRODUCTION

With the emergence of multiple data sources, such as administrative, web-collected, and big data, data integration has become popular. For example, the integration of administrative and statistical data has been widely used in official statistics to improve data quality, data analyses, and data collection costs (Berg et al., 2021). As the demand for data integration increases, two kinds of statistical methodologies have been proposed—namely, record linkage and statistical matching. Record linkage assumes overlapped units, which allows the linkage of the same units between data sources, while statistical matching frequently links similar units between nonoverlapped datasets (Leulescu & Agafitei, 2013). In this study, we are interested in combining multiple values observed for the same attribute, a common issue in record linkage.

When multiple data sources are available, it is easy to have multiple observations on the same attribute due to nonresponse or measurement errors, different observation times, and mode effects with varying relation shape to the true attribute. To handle these multiple observations in statistical models, we treat them as multiple proxies of the unobserved true value. These proxies include various domains, from the case in which covariates are measured with errors (Carroll et al., 2006; Fuller, 1987) when the true covariate is simply unobserved (Filmer & Pritchett, 2001), to when the variable of interest is a conceptual variable (Mazumder, 2001; Solon, 1992; Zimmerman, 1992). Specifically, we consider the problem of estimating the quantile function in the presence of multiple proxies for true covariates. As in conventional linear regression, the quantile regression estimator's inconsistency in the absence of a true covariate is a commonly discussed issue in the literature (Brown, 1982; Carroll et al., 2006; He & Liang, 2000; Hausman et al., 2021; Montes-Rojas, 2011; Wei & Carroll, 2009).

Several studies have incorporated proxy variables in quantile regression estimation. He and Liang (2000)—considering the problem of estimating quantile regression coefficients in errors-in-variables models with a proxy variable—proposed an estimator in the context of linear and partially linear models. Wei and Carroll (2009) presented a nonparametric method for correcting bias caused by measurement error in the linear quantile regression model by constructing joint estimating equations that simultaneously hold for all quantile levels. Further, Firpo et al. (2017) proposed a semiparametric two-step estimator when repeated measures for the proxy are available. Schennach ((2007), (2008)) discussed identifying a
nonparametric quantile function under various settings when an instrumental variable is measured on all sampling units. Wang et al. (2012) modified the standard quantile regression objective function, tailoring it to the Gaussian measurement error model.

However, these studies are limited to combining multiple proxies in a nonunified framework and frequently require repeated observations of proxy variables. Additionally, most studies focus on proxy variables with limited formulas and have employed a classical measurement error model that assumes a zero-mean error is added to the unobserved covariate. To address these issues, we propose a novel Bayesian measurement error model by specifying how the proxies are generated from unobserved covariates. Motivated by the measurement error model approach (Clayton, 1992; Fuller, 1987; Lubotsky & Wittenberg, 2006; Richardson & Gilks, 1993), we demonstrate a Bayesian flexible combining method. Specifically, our approach uses a generalized additive model for the proxies, which allows them to be decomposed as a smoothing function of the true covariate and unobserved additive error. This provides a flexible and effective way to accommodate a wide range of proxy variables, including nonparametric spline functions that can be estimated within the model. Moreover, we employ a factor analysis-based measurement error model introduced in Fuller (1987) to combine multiple proxies in a unified framework. We treat the true covariate as a common factor in the factor analysis and construct a set of additive models. This approach allows the generation of synthetic values using a Bayesian approach (Clayton, 1992; Richardson & Gilks, 1993). However, following Berry et al. (2002), we use the Bayesian nonparametric quantile regression function as the outcome model with a natural cubic spline (Thompson et al., 2010) and penalized spline (Lang & Brezger, 2004), rather than linear parametric models that might be sensitive to model misspecification.

The proposed method has several advantages. First, we incorporate a flexible form of proxy variables in a wide range of different formulas, previously limited to additive measurement error. Second, the proposed method incorporates an arbitrary number of multiple proxies with an estimation of the true covariate, the unobserved relationship to the proxies, and quantile regression functions. Third, the proposed framework does not assume the linearity of the quantile regression function, which restricts the model’s flexibility.

The remainder of the paper is organized as follows: In Section 2, we describe the basic setup of the investigation. In Section 3, we introduce a method of combining multiple proxies and making inferences regarding nonparametric quantile regression likelihood. Further, the related priors and detailed Gibbs sampling steps are described. In Section 4, we present the simulation results for various simulation data. Extensive simulation studies on various datasets reveal the approach’s effectiveness in incorporating multiple proxies simultaneously, compared with the method of using one proxy variable directly and the previously proposed structural model. In Section 5, we propose the method to the public administrative dataset, which includes various economic features of 18,064 families, such as salary and property income. We apply this methodology to administrative data to study the quantile relationship between assets and actual salary income. In Section 6, we provide concluding remarks.

## 2 BASIC SETUP

Let \( \{y_i,x_i\}_{i=1}^{n} \) be a random sample of size \( n \), where \( y_i \) is the outcome variable and \( x_i \) is the explanatory covariate. Let \( g_p(x) \) be the \( p \)th quantile of the conditional distribution of \( y_i \), considering \( x_i \) such that

\[
P(y_i \leq g_p(x)|x_i) = p, \quad 0 < p < 1. \tag{1}
\]

Suppose that covariate \( x_i \) is not directly observed, but multiple proxies related to the covariate are observed from multiple data sources. These proxies are denoted as \( w_{ki}, k = 1,...,K \). We employ popular methods (e.g., Carroll et al., 2006; Delaigle et al., 2008; Li & Vuong, 1998) to estimate the quantile function \( g_p(x) \) when proxies are replicates of mismeasured variables, wherein only the mean zero measurement error is added to the covariate. However, these methods cannot be directly extended when multiple proxies exist in a different relationship to the unobserved \( x \).

To simultaneously account for different proxies, first, we write a set of regression models such that

\[
y_i = g_p(x_i) + \epsilon_i, \tag{2}
\]

\[
w_{1i} = x_i + u_{1i}, \tag{3}
\]

\[
w_{ki} = h_k(x_i) + u_{ki}, \quad k = 2,...,K, \tag{4}
\]

where the residual \( \epsilon_i \) in (2) follows an unspecified distribution that satisfies (1). The measurement errors \( u_{ki} \) in (3) and (4) are assumed to have a zero mean \( E(u_{ki}) = 0 \) and a constant variance \( \text{Var}(u_{ki}) = \sigma_k^2 \), and are further assumed to be independent of each other and distributed independently of \( x_i \) and \( y_i|x_i \). Equation (3) implicitly assumes a reference variable that is only exposed to sampling or additive measurement errors. This assumption is natural and popular when survey datasets exist (Berg et al., 2021; Fuller, 2009) and can be generalized to other deterministic functions of proxy and unobserved covariates, such as logarithms (Berry et al., 2002). We leave other relationships between proxies \( w_{ki} \) (\( k \geq 2 \)) and \( x \)
denoted by $h_k$ in (4) to take any form within generalized additive model. This includes nonparametric spline function as will be described in more detail later.

Lubotsky and Wittenberg (2006) used similar assumptions with (2) to (4) to estimate the regression coefficient for the linear regression of $y_i$ on $x_i$, with multiple proxies for $x_i$. They allowed nonzero covariance between the measurement errors; however, they assumed a linear relationship with all proxy variables $w_k$ for $k \geq 2$ in (4) and offered a lower bound on the regression coefficient of $y_i$ on $x_i$. Variable $w_1$ with additive error is the benchmark variable. The assumption of the existence of a benchmark variable is prevalent and necessary because it amounts to fixing the scale of unobserved $x$.

Intuitively, the proxy variables are not required to be independent of each other or $x_i$. However, because measurement errors $u_i$ are independent of each other and $y_i|x_i$, the proxies are conditionally independent considering the true covariate $x_i$. This model setup makes the conditional distribution between observed proxies independent of each other; this suggests that the likelihood of multiple proxy variables can be separately specified with a mixture representation based on conditional distribution given $x_i$ and their prior probability distribution. This makes Bayesian Gibbs sampling reasonable for the estimation method when $x_i$ is treated as an auxiliary variable and contributes to the generation of the observed data $(y_i, w_1, \ldots, w_n)$.

3 | BAYESIAN ESTIMATION

This section briefly introduces the natural cubic spline and P-spline to be used for the quantile function $g_p(x)$ in (2). The proposed methodology for combining multiple proxies in Bayesian quantile regression is followed by explicit sampling steps.

3.1 | Splines for quantile regression

Although the proposed framework is applicable to other quantile regression functions, such as the polynomial function of covariate $x$ (Koenker & Bassett, 1978; Yu & Moyeed, 2001; Yu et al., 2003), polynomial quantile regression is frequently restricted because the degree of the polynomial must be chosen in advance, and data might have a limited local effect on the shape of the polynomial regression curve, especially for high quantiles (Thompson et al., 2010). To alleviate the parametric assumption and secure the model’s flexibility, we consider two popular spline methods for $g_p(x)$: Nonparametric cubic spline and the P-spline function. For details regarding the regression spline, refer to Brezger and Lang (2006), Carroll et al. (1999), Eilers and Marx (1996), and Green and Silverman (1993).

3.1.1 | Natural cubic spline

The natural cubic spline is a piecewise cubic polynomial function with continuous first and second derivatives at each knot and is linear beyond the boundary knots (Green & Silverman, 1993). Suppose that $t_1, \ldots, t_N$ are $N$ fixed knots covering a range of $x$ and $g = (g_1, \ldots, g_N)^T$ denotes the values of the natural cubic spline $g_p(x)$ at knots $t_1, \ldots, t_N; g_1(t_1), \ldots, g_N(t_N)$. As a desirable property of the natural cubic spline, there is a unique natural cubic spline function $g_p(x)$ with knots $t_1, \ldots, t_N$ satisfying $g(t_i) = g_i, i = 1, \ldots, N$ for any given value $g_1, \ldots, g_N$. Therefore, we can handle function $g(x)$ using its finite-length surrogate $g$. In terms of Bayesian inference, we can model $g_p(x)$ by giving the prior to $g$ and not $g_p(x)$. Following Green and Silverman (1993), the prior for $g$ is defined by a multivariate normal distribution as follows:

$$
\pi(g|\lambda) \propto \exp\left(-\frac{1}{2}g^T K g\right),
$$

(5)

where $K$ is the $N \times N$ matrix with rank($K$) = $N - 2$, a function of the difference between the knots defined by Eubank (1999). $\lambda$ contributes to the smoothness of curve $g$ and has a standard conjugate gamma prior:

$$
\pi(\lambda) = \lambda^{a_1 - 1} \exp(-\lambda) / \Gamma(a_1), \lambda > 0.
$$

The quadratic term $g^T K g$ in the exponential kernel is equivalent to the roughness penalty, $\int_0^1 g_p^2(t) dt = g^T K g$ (Green & Silverman, 1993) and is a natural choice because it corresponds to the penalty term in the penalized maximum likelihood. With this prior, the posterior log density of the function $g_p(x_i)$ is equal to the loss function in the regression context, with the roughness penalty added to the kernel of the log-likelihood function (Hoerl & Kennard, 1970; Green & Silverman, 1993; Tibshirani, 1996).
The final step in the Bayesian approach is defining the natural cubic spline function’s likelihood by changing the conventional polynomial part of the standard quantile regression likelihood to the natural cubic spline form (Thompson et al., 2010; Yu & Moyeed, 2001). The resulting likelihood takes the following form:

\[
L(y|g, x) = p^n(1-p)^n \exp\left\{ -\sum_{i=1}^{n} \rho_{\lambda}(y_i - g_p(x_i)) \right\}.
\]

(6)

Notably, we explicitly specify \( x \) in likelihood \( L(y|g, x) \) for generalization, where \( x \) is also an unknown variable.

### 3.1.2 Regression P-spline

Another general approach to spline fitting is a penalized spline or simply a P-spline. For a P-spline of degree \( l \) with \( N \) fixed knots, \( g_p(x) \) is defined by \( Z(x)^T \beta \) where \( Z(x) \) is \((N + l + 1)\) vector composed of B-spline basis functions evaluated at observation \( x \), and \( \beta \) is the coefficient of the basis functions (Eilers & Marx, 1996). A conventional basis is \( Z(x) = \left( 1, x, \ldots, (x - r_1), \ldots, (x - r_n) \right)^T \). Then, \( r_2, \ldots, r_{1 + N}, \sigma \) are the sizes of the jumps in the \( l \)th derivative of \( g(x) \) at the knots.

Eilers and Marx (1996) suggested a roughness penalty based on differences of adjacent spline coefficient to guarantee sufficient smoothness. In Bayesian analysis, the prior of \( \beta \) replaces the roughness penalty term of the penalized likelihood as their stochastic analogs (Lang & Brezger, 2004). Assuming a first-order random walk for \( \beta \), the joint conditional distribution of \( \beta \) is

\[
\pi(\beta|\lambda) \propto \exp\left( -\frac{1}{2} \beta^T K \beta \right),
\]

where \( K = \frac{1}{2} R^T R \) is \((N + l + 1) \times (N + l + 1)\) penalty matrix with rank \( K = N + l \) and \( R \) is a first-order difference matrix (Lang & Brezger, 2004; Waldmann et al., 2013). This prior on \( \beta \) induces a prior on \( g_p \), owing to the deterministic relationship between \( g_p \) and \( \beta, g_p(x) = Z(x)\beta \). The precision parameter \( \lambda \), again, contributes to the smoothness of curve \( g \) and has a standard gamma prior:

\[
\pi(\lambda) = \frac{\lambda^{a+1} \exp(-\frac{\lambda}{b})}{\Gamma(a)b^a}, \lambda > 0.
\]

The P-spline function’s likelihood can be defined by changing the conventional polynomial part of the standard quantile regression likelihood. The same result is obtained in (6). However, unlike the natural cubic spline, Waldmann et al. (2013) suggested exploiting the stochastic representation of the likelihood for more efficient sampling (Kozumi & Kobayashi, 2011).

### 3.2 Nonparametric quantile regression with multiple proxy variables

In this section, we describe the fully Bayesian approach to the problem setup in (2)-(4). Let the unknown parameters to be estimated \( \theta = (g, \lambda, x, \theta_2, \sigma^2) \), where \( \theta = (\theta_1, \ldots, \theta_K)^T \), \( \sigma^2 = (\sigma^2_1, \ldots, \sigma^2_K)^T \) and \( \theta_k \) is a parameter related to \( h_k(x) \). Without the loss of generality, the posterior density is

\[
p(\theta|y, w) \propto p(y|g, \lambda, x)p(w|x, \theta_2) \pi(g|\lambda) \pi(\lambda) \pi(\theta_2) \pi(\sigma^2) \pi(x),
\]

where \( w = (w_1, \ldots, w_K) \). For the variance parameter, we use the conjugate independent inverse gamma prior \( \pi(\sigma^2) = \prod_{k=1}^{K} \pi(\sigma^2_k) \). For the prior distribution of \( x \), the reasonable prior distribution and appropriate prior varies depending on the application (Berry et al., 2002). As an exemplary prior, we use a widely used hierarchical normal distribution, that is,

\[
\pi(x) \sim N(\mu_x, \sigma^2_x), \mu_x \sim N(0, \sigma^2_{\mu_x}), \sigma^2_x \sim IG(a_x, b_x).
\]

From these prior settings, we derive the complete conditional distributions for \( x \) up to the normalizing constant

\[
\pi(x|\theta_{-x}) \propto \exp\left( -\frac{1}{2} \sigma^2_x (x - \mu_x)^2 + \sum_{k=1}^{K} \frac{1}{\sigma^2_k} (w_k - h_k(x))^2 \right),
\]

where \( \mu_x \) and \( \sigma^2_x \) are the mean and variance of \( x \) respectively, and \( h_k(x) \) is the function of the \( k \)th proxy variable in the model. The resulting likelihood for the \( j \)th observation can be written as

\[
L(y|g, x) = p^n(1-p)^n \exp\left\{ -\sum_{i=1}^{n} \rho_{\lambda}(y_i - g_p(x_i)) \right\}.
\]
where \( \theta_k \) denotes all other parameters except \( x \). Other conditional distributions are derived from the conjugacy of their prior as follows:

\[
\begin{align*}
\pi(\sigma^2_k | \theta_{-k}) & \sim IG\left(\frac{n}{2} + a_k, b_k + \frac{1}{2} \sum_{i=1}^{n} (w_{ki} - h_k(x_i))^2\right), \\
\pi(\sigma^2_{\mu_k} | \theta_{-k}) & \sim IG\left(\frac{n}{2} + a_k, b_k + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu_k)^2\right), \\
\pi(\mu_k | \theta_{-k}) & \sim N\left(\frac{\sum_{i=1}^{n} x_i}{\sigma^2_{\mu_k}}, \frac{1}{\sigma^2_{\mu_k}}\right) \left(\frac{n}{\sigma^2_{\mu_k} + \sigma^2_{\mu_k}}\right)^{-1}. \end{align*}
\]

For the parameters \( \theta_k \) related to the relationship \( h_k(x) \) between \( w_k \) and \( x, h_k(x) \) can only contribute to its expectation of the conditional distribution, and the distribution's family remains the same because \( u_k \) is independent of \( x \). Therefore, the generalization of \( h_k(x) \) to an arbitrary functional form, such as a linear regression or natural cubic spline with a pre-existing Bayesian method for \( \theta_k \), is possible. We provide specific examples in the subsequent subsection. The priors for spline-related parameters \( g \) and \( \lambda \) differ depending on the regression spline function, as specified in Section 3.1.

A key benefit of this Bayesian approach is that the smoothing spline's observations are generated from the posterior; thus, we can estimate the entire posterior distribution of \( g \), which was difficult in Lubotsky and Wittenberg (2006). Furthermore, an additional assumption is required to combine proxies and identify the unobserved covariate \( x \) (Aigner et al., 1984; Lubotsky & Wittenberg, 2006). However, in the Bayesian framework, treating unobserved \( x \) as a latent variable and placing its prior probability distribution, which corresponds to the structural approach in the literature (Fuller, 1987), is natural. Although the regression function \( g_p \) is the primary focus of interest, the joint posterior distribution is a powerful tool that enables the inference of the unobserved covariate \( x_i \) and its unobserved relationship with proxies \( h_k(x) \).

### 3.3 | Gibbs sampling step

The proposed framework for combining multiple proxies can be completed by defining the formula of \( h_k \) and the distribution of the measurement error \( u_k \), which can be defined by the researcher. Provided that a benchmark variable is available to fix the scale of unobserved \( x \), the proposed method can accommodate a wide range of proxies. This includes a polynomial parametric relationship between \( w_k \) and \( x \) for interpretability or based on prior knowledge of \( w_k \) and \( x \), while it is also possible to use a nonparametric spline model to accommodate a nonlinear relationship with prior assumptions but a higher data requirement. For each case, we formulate the entire problem in the Bayesian framework and present the Metropolis-Hastings steps (Gamerman & Lopes, 2006).

#### 3.3.1 | Natural cubic spline with quadratic proxy

Suppose we have two proxies: one with an additive error and another with a quadratic relationship.

\[
\begin{align*}
y_j & = g_1(x_j) + e_j, \\
w_{1j} & = x_j + u_{1j}, \\
w_{2j} & = x_j^2 + u_{2j},
\end{align*}
\]

where \( u_{kj} \sim N(0, \sigma^2_k), k = 1,2 \). Here, parameter \( \theta_2 \) for \( h_2(x) = \alpha = (a_0, a_1, a_2) \). We use normal prior for \( \pi(\alpha) \sim N(\mu_\alpha, V_\alpha) \). Following Thompson et al. (2010), we model a quantile function of covariate \( x \) using the natural cubic spline, with \( N \) evenly spaced fixed knots covering a range of \( x \).

A Gibbs sampling algorithm for the quantile regression model is constructed by sampling each component of \( \theta \) from the full conditional distributions. Following Thompson et al.’s (2010) initialization, \( g^{(0)} \) is set as the posterior mean value of the quantile regression curve (Yu & Moyeed, 2001) at \( t_1, \ldots, t_N \), and \( \lambda^{(0)} \) is obtained by applying generalized cross-validation of the usual smoothing spline (Green & Silverman, 1993). Additionally, \( x^{(0)} \) is set as a multiple proxies \( w_1 \) because it is a more reliable proxy in the initialization step with no information regarding \( \alpha \).

One iteration of the Gibbs sampling algorithm at iteration \( t \) is as follows:

1. Generate candidate \( g^* \) from the multivariate normal distributions,

\[
g^* | g^{(t-1)} \sim \text{MVN}(g^{(t-1)}, \Sigma),
\]

2. Sample \( \lambda^{(t)} \) from the posterior distribution of \( \lambda \).

3. Sample \( x^{(t)} \) from the posterior distribution of \( x \).

4. Sample \( g^{(t)} \) from the posterior distribution of \( g \).

5. Sample \( \alpha^{(t)} \) from the posterior distribution of \( \alpha \).

6. Sample \( \theta^{(t)} \) from the posterior distribution of \( \theta \).

7. Update \( w^{(t)} \) using the sampled values of \( x^{(t)} \) and \( g^{(t)} \).

8. Update \( e^{(t)} \) using the sampled values of \( g^{(t)} \) and \( w^{(t)} \).

9. Update \( u^{(t)} \) using the sampled values of \( e^{(t)} \) and \( w^{(t)} \).

10. Update \( h^{(t)} \) using the sampled values of \( x^{(t)} \) and \( g^{(t)} \).

11. Update \( \theta^{(t)} \) using the sampled values of \( h^{(t)} \) and \( g^{(t)} \).

12. Update the model parameters using the sampled values of the latent variables.
and accept $g^*$ with probability,

$$r = \min \left\{ 1, \frac{L(y|g^{(t-1)}, x^{(t-1)}), \pi(g^*|\lambda)q(g^{(t-1)}|g^*)}{L(y|g^{(t-1)}, x^{(t-1)}), \pi(g^{(t-1)}|\lambda)q(g^*|g^{(t-1)})} \right\},$$

where $q$ represents the proposal density function.

2. Generate candidate $\lambda^*$ from the log-normal distribution,

$$\eta^* \sim N(\log(\lambda^{(t-1)}), \sigma^2_\eta),$$

where $\lambda^* = \exp(\eta^*)$, and accept $\lambda^*$ with probability,

$$r = \min \left\{ 1, \frac{\pi(g^{(t)}|\lambda^*)\pi(\lambda^*)q(\lambda^{(t-1)}|\lambda^*)}{\pi(g^{(t)}|\lambda^{(t-1)})\pi(\lambda^{(t-1)}|\lambda^*)} \right\}.$$

3. Generate $x^*$ from the multivariate normal distribution,

$$x^*|x^{(t-1)} \sim \text{MVN}(x^{(t-1)}, \Sigma_w),$$

and accept $x^*$ with probability,

$$r = \min \left\{ 1, \frac{L(y|g^{(t)}, x^*)\pi(w_1|w_2, x^*)\pi(w_2|x^*)\pi(x^*|\mu^{(t-1)}, (\sigma^2_\alpha)^{(t-1)})q(x^{(t-1)}|x^*)}{L(y|g^{(t)}, x^{(t-1)})\pi(w_1|w_2, x^{(t-1)})\pi(w_2|x^{(t-1)})\pi(x^{(t-1)}|\mu^{(t-1)}, (\sigma^2_\alpha)^{(t-1)})q(x^{(t-1)}|x^*)} \right\}.$$

From the independent assumption, this step can be separated as generating $x_i^*$ from the normal $x_i^*|x_i^{(t-1)} \sim N(x_i^{(t-1)}, \sigma^2_{x_i})$ with acceptance probability reduced to the term of the ith data.

4. Sample $(\sigma^2_\alpha)^{(t)} \sim \text{Inv-Gamma}\left(\frac{a}{2} + \alpha_1, b_2 + \frac{1}{2} \sum_{i=1}^n \left( w_{2i} - \mu_i^{(t)} \right)^2 \right)$, where Inv-Gamma($a, b$) indicates the inverse gamma with shape parameter $a$ and scale parameter $b$, and $a_1$ and $b_2$ are the corresponding parameters for the prior of $\sigma^2_\alpha$.

5. Sample $(\sigma^2_\beta)^{(t)} \sim \text{Inv-Gamma}\left(\frac{a_2}{2} + \alpha_2, b_2 + \frac{1}{2} \sum_{i=1}^n \left( \mu_{2i}^{(t-1)} - \alpha_2 x_i^{(t-1)} \right)^2 \right)$ with $a_2$ and $b_2$ be the parameters for the prior of $\sigma^2_\beta$.

6. Sample $(\sigma^2_{\varepsilon})^{(t)} \sim \text{Inv-Gamma}\left(\frac{a_3}{2} + \alpha_3, b_3 + \frac{1}{2} \sum_{i=1}^n \left( \mu_i^{(t)} - \alpha_3 x_i^{(t-1)} \right)^2 \right)$ with $a_3$ and $b_3$ be parameters for the prior of $\sigma^2_{\varepsilon}$.

7. Sample $\mu^{(t)} \sim \text{N}(M_x, V_x)$, where

$$V_x = \left( \frac{n}{\sigma^2_{\varepsilon}} \right)^{-1} \text{ and } M_x = \left( \frac{\sum_{i=1}^n x_i^{(t-1)} + nM_x}{\sigma^2_{\varepsilon}} \right) \left( \frac{n}{\sigma^2_{\varepsilon}} \right)^{-1} \text{ with } M_x \text{ and } \sigma^2_{\varepsilon} \text{ the prior mean and variance for } \mu_x.$$

8. Sample $a^{(t)} \sim \text{MVN}(\mu_x, V_x)$, where

$$V_x = \left( \frac{X^T X}{(\sigma^2_a)^{(t)}} + V^{-1}_a \right)^{-1} \text{ and } \mu_x = \left( \frac{X^T X}{(\sigma^2_a)^{(t)}} + V^{-1}_a \right)^{-1} \left( \frac{X^T W_1}{(\sigma^2_a)^{(t)}} + V^{-1}_a \mu_a \right) \text{ with } M_a \text{ and } V_a \text{ as the prior mean vector and covariance matrix for } a \text{ and } X \text{ as the vector of } x_i^{(t)} \text{, } i = 1, \ldots, n.$$

Steps 1–3 require the metropolis–hastings algorithm, and the other steps can be sampled from the conjugate distribution. The inference regarding the unobserved regressor, quantile spline function, or regression coefficient is based on these posterior samples.

### 3.3.2 P-spline with arbitrary nonlinear proxy

Suppose we have two proxies—one with an additive error and another with a nonlinear relationship.
\[ y_i = g_p(x_i) + e_i, \]

\[ w_{i1} = x_i + u_{i1}, \]

\[ w_{i2} = h_2(x_i) + u_{i2}, \]

where \( u_{i1} \sim N(0, \sigma^2), k = 1.2. \)

Following Waldmann et al. (2013), we used a stochastic representation of the likelihood \( L(y|x), \) that is, \( y \sim N(g_p(x) + As, B \frac{\pi}{t}) \). where \( A = \frac{1-2p}{p+1}, B = \frac{p^2}{p+1} \) and \( s \sim \text{Exp}(\tilde{\nu}) \), with the conjugate gamma prior on \( \tilde{\nu} \) \( \sim \text{Ga}(a_0, b_0). \) This hierarchical representation of the likelihood in (6) enables efficient Gibbs sampling. For additional details, refer to Kozumi and Kobayashi (2011) and Waldmann et al. (2013).

As we have two nonlinear functions \( g_p \) and \( h_2 \) to be fitted, we use two P-splines. To discriminate the coefficients of the basis function for \( g_p \) and \( h_2 \), we use subscripts \( g_p \) and \( h_2. \) For a P-spline with \( N \) fixed knots, \( g_p \) is defined by \( \beta_p, \) and for any realization of \( \beta_p \), there exists a corresponding realization \( g_p(x) = Z(\mu)\beta_p. \) As both splines \( g_p \) and \( h_2 \) share covariate \( x, \) they share the same knots and penalty matrix \( K, \) which reduces the computation. Thereafter, the prior is as specified in Section 3.1; \( \pi(\beta_p) \propto \text{exp}(-\frac{1}{2}g_p_{\beta_p}^T K_{\beta_p} g_p_{\beta_p}), \pi(\beta_h) \propto \text{exp}(-\frac{1}{2}h_2_{\beta_h}^T K_{\beta_h} h_2_{\beta_h}), \pi(\lambda_\beta) \sim \text{Ga}(a_0, b_0), \) \( \pi(\beta_\mu) \sim \text{Ga}(a_0, b_0). \)

One iteration of the Gibbs sampling algorithm at iteration \( t \) as follows:

1. Generate \( x_i^* \) from the normal distribution, \( x_i^* \mid x_i^{(t-1)} \sim N(x_i^{(t-1)}, \sigma_{x_i}^{2}), \) and accept \( x_i^* \) with probability,

\[
r = \min \left\{ \frac{N(\gamma_i)Z(x_i^*)^T g_p(x_i^{(t-1)}) + A_x^{(t-1)} B g_p(x_i^{(t-1)})}{N(\gamma_i)Z(x_i^{(t-1)})^T g_p(x_i^{(t-1)}) + A_x^{(t-1)} B g_p(x_i^{(t-1)})}, \frac{N(\gamma_i)Z(x_i^{(t-1)})^T g_p(x_i^*) + A_x^{(t-1)} B g_p(x_i^*)}{N(\gamma_i)Z(x_i^{(t-1)})^T g_p(x_i^*) + A_x^{(t-1)} B g_p(x_i^*)} \right\}
\]

2. Sample \( \beta^{(t)}_p \sim N(M_p, V_p), \) where

\[
V_p = \left( \frac{s_i^{(t-1)}K + g_p^T Z D^{-1} Z}{B} \right)^{-1}, \quad M_p = V_p^{-1} \left( \frac{g_p^T Z D^{-1} (y - A s^{(t-1)})}{B} \right)
\]

with \( D = \text{diag}(s_i^{(t-1)}, ..., s_i^{(t-1)}) \) and \( Z \) represents the design matrix with \( Z(x_i^{(t)}), i = 1, ..., n. \)

3. Sample \( s_i^{(t)} \sim \text{Inv-Gauss}(\sqrt{\frac{\rho^2_{(t)}}{2B}} \frac{2B_{s_i^{(t)}}}{B}), \) where \( \text{Inv-Gauss}(a,b) \) is an inverse gaussian distribution with mean parameter \( a \) and shape parameter \( b. \)

4. Sample \( \sigma^{(t)}_x \sim \text{Inv-Gauss}(\sqrt{\frac{\sigma^2_{x_i}^{(t)}}{2B}} \frac{2B_{\sigma^2_{x_i}^{(t)}}}{B}), \) where \( \text{Inv-Gauss}(a,b) \) is an inverse gaussian distribution with mean parameter \( a \) and shape parameter \( b. \)

5. Sample \( \sigma^{(t)}_\mu \sim \text{Inv-Gauss}(\sqrt{\frac{\sigma^2_{x_i}^{(t)}}{2B}} \frac{2B_{\sigma^2_{x_i}^{(t)}}}{B}), \) where \( \text{Inv-Gauss}(a,b) \) is an inverse gaussian distribution with mean parameter \( a \) and shape parameter \( b. \)

6. Sample \( \mu_{(t)}^{(t)} \sim N(M_p, V_p), \) where

\[
V_p = \left( \frac{n}{\sigma^2_{x_i}^{(t-1)} + 1} \right)^{-1}, \quad M_p = V_p^{-1} \left( \frac{n}{\sigma^2_{x_i}^{(t-1)} + 1} \right)
\]

with \( M_p \) and \( \sigma^2_{x_i} \) the prior mean and variance for \( \mu. \)

7. Sample \( \sigma^2_{x_i}^{(t)} \sim \text{Inv-Gamma}\left(\frac{1}{2} + a_0, b_0 + \frac{1}{2} \sum_{i=1}^n (x_i^{(t)} - \mu_{\text{p}}^{(t)})^2\right) \) with \( a_0 \) and \( b_0 \) as the parameters for the prior of \( \sigma^2_{x_i} \)

8. Sample \( \sigma^2_{x_i}^{(t)} \sim \text{Inv-Gamma}\left(\frac{1}{2} + a_0, b_0 + \frac{1}{2} \sum_{i=1}^n (w_{i1} - x_i^{(t)})^2\right) \) with \( a_0 \) and \( b_0 \) as the parameters for the prior of \( \sigma^2_{x_i} \)

9. Sample \( \beta_{h}^{(t)} \sim N(M_p, V_p), \) where

\[
V_p = \left( \frac{1}{2} + a_0, b_0 + \frac{1}{2} \sum_{i=1}^n (w_{i1} - x_i^{(t)})^2\right) \) with \( a_0 \) and \( b_0 \) as the parameters for the prior of \( \lambda_{\beta}. \)

10. Sample \( \lambda_{\beta}^{(t)} \sim \text{Ga}(a_0, b_0 + 0.5 \text{rank}(\lambda_{\beta}^{(t-1)} K_{\beta}), b_0 + 0.5 \rho_{\beta}^{(t)} \lambda_{\beta}^{(t-1)} K_{\beta}^{(t)}), \) where \( a_0 \) and \( b_0 \) are the corresponding parameters for the prior of \( \lambda_{\beta}. \)
11. Sample \( \sigma_2^2(i) \sim \text{Inv-Gamma} \left( \frac{3}{2} + a_2, b_2 + \frac{1}{2} \sum_{j=1}^{n} \left( w_{2j} - Z(x_j) \right)^T \rho_0(i)^{-1} \right)^2 \) with \( a_2 \) and \( b_2 \) as the parameters for the prior of \( \sigma_2^2 \).

The inference regarding the unobserved regressor, quantile spline function, or regression coefficient is based on these posterior samples.

4 | SIMULATION

We conduct a simulation study to empirically evaluate the proposed method using various datasets. This simulation has the following three purposes: to evaluate the flexibility of the proposed method in various datasets by considering several different error types, to compare the proposed method with an alternative approach, and to evaluate the effect of different types of proxies and effect of the number of proxies on the proposed method.

We use the dataset studied by Waldmann et al. (2013) and Yue and Rue (2011). To match the scale between each dataset, we scale the range of \( x \) to \([-5,5]\) for each simulation using the appropriate affine transformation. We simulate the datasets using the following formulae:

- Dataset1: \( y_i = 0.4x_i + 0.5 \sin(2.7x_i) + 1.1/(1 + x_i^2) + \epsilon_i \);
- Dataset2: \( y_i = \sin(2(4x_i - 2)) + 2 \exp(-(-16^2)(x - 0.5)^2) + ((3x_i)/2)\epsilon_i \),

and the quantile functions for each dataset are given by

- Dataset1: \( g_p(x_i) = 0.4x_i + 0.5 \sin(2.7x_i) + 1.1/(1 + x_i^2) + F^{-1}(p) \);
- Dataset2: \( g_p(x_i) = \sin(2(4x_i - 2)) + 2 \exp(-(-16^2)(x - 0.5)^2) + ((3x_i)/2)F^{-1}(p) \),

where \( F^{-1} \) is the cumulative distribution function of the distribution from where the error \( \epsilon_i \) is sampled. For the error distribution \( F \), we consider three different error distributions as follows: standard normal distribution, Student's \( t \) distribution with two degrees of freedom, and gamma distribution with shape 4 and rate 1. Figure 1 displays a data structure of both datasets with each error term, on \( p \in \{0.1, 0.25, 0.5, 0.75, 0.9\} \).

Dataset2 follows the heteroscedastic structure, as the error generated from the error distribution is multiplied by \( x_i \), and the resulting quantile curves are no longer parallel to each other as in Dataset1. For the error distribution, \( t \) is a heavy-tailed distribution, which may cause the dataset to include extreme outliers. The gamma distribution with shape 4 and rate 1 has a nonzero expectation skewed to the right. This causes the resulting quantile function to shift to positive values for higher quantiles. Similar examples were analyzed in Fenske et al. (2011), Kottas and Krnjačić (2009), Taddy and Kottas (2010), Waldmann et al. (2013), and Yue and Rue (2011).

The proposed method can adapt multiple proxies with arbitrary relationships to the covariates. To assess a proxy's effect, further, we suppose that the actual covariates \( x \) are not directly observed, and that proxies \( w_{k}, k = 1, 2, 3 \) are observed. For the relationship \( h_k \) between proxy \( w_k \) and unobserved covariate \( x \), we consider three different types as follows: Identity, polynomial, and smooth nonlinear function. That is, we have

\[
\begin{align*}
    w_{31} &= x_i + u_{31}, \\
    w_{32} &= x_0 + a_1 x_i + a_2 x_i^2 + u_{32}, \\
    w_{33} &= \sin(12(x_i + 0.1))/(x_i + 0.1) + u_{33},
\end{align*}
\]

with \( \alpha = (3, 0.25, 0.75) \) and \( u_{3j} \sim N(0,1), k = 1, 2, 3 \). The nonlinear example \( h_3 \) is based on Hastie et al. (2009) and is generated from smoothing splines. The parameter for the quadratic coefficient \( \alpha \) is determined such that the ratio of variance in the error components to the total variance in the proxy variables in \( w_2 \) is roughly matched to that in \( w_3 \).

Most studies that have considered the presence of a proxy variable have focused on mean regression and assumed the existence of replicates of the benchmark variable, which is a stricter assumption than the case considered here (Carroll et al., 2006; Wei & Carroll, 2009). Instead, we adjust Carroll et al.’s (1999) approach to be suitable for quantile regression problems. The method uses a partially Bayesian approach, which estimates the moment function of unknown \( x \) using \( w \) and estimates the spline function by minimizing the conventional penalized likelihood with the given estimated moments. This method uses a two-step estimation procedure that uses the information of \( w \) to estimate \( x \), and the relationship between \( x \) and \( y \) is estimated thereafter.

Consequently, five estimators were considered for comparison in each simulation.

1. Model without measurement error (woME): The benchmark estimator with true covariate \( x \) directly observed without error.
2. Structural estimator (Structural): The estimator calculated using Carroll et al.’s (1999) model, with all related prior settings equal to Bayesian estimator with multiple proxies (BEMP).
3. Bayesian estimator with polynomial proxies (BEMP-poly): The proposed estimator that incorporates two proxies \( w_1, w_2 \), but not \( w_3 \).
4. Bayesian estimator with nonlinear proxies (BEMP-nonlinear): The proposed estimator that incorporates two proxies \(w_1, w_3\), but not \(w_2\).

5. Bayesian estimator with all proxies (BEMP-all): The proposed estimator that incorporates all proxies \(w_1, w_2,\) and \(w_3\).

For the woME model, we used the natural cubic spline method applied in quantile regression using Thompson et al.’s (2010) model. For conciseness and unity, we present only the performance of the BEMP with a natural cubic spline. The performance of a BEMP with a P-spline is similar to that of a BEMP with a natural cubic spline and is presented in Appendix S1. We used the same 30 knots for the computation of \(g\) as those in Thompson et al. (2010).

For each of the generated datasets, the number of observations is fixed at \(n = 1000\) and the quantile functions on a fixed grid \(p \in \{0.1, 0.25, 0.5, 0.75, 0.9\}\) are estimated. We assume identical MCMC sampling settings for all the five models. We set the number of iterations to 300,000 and took every 50th sample after discarding the first 50,000 steps as the burn-in period. The convergence is satisfactory, and an average of 5,000 posterior samples are used for the point estimation. After running 100 Monte Carlo (MC) simulations, we report their average and standard deviation.

We use two popular metrics to compare the estimators (Fan, 1992; Gelfand & Ghosh, 1998; Härdle, 1986):

- Mean squared error (MSE):

\[
MSE(\hat{g}_p) = \frac{1}{n} \sum_{i=1}^{n} (g_p(x_i) - \hat{g}_p(x_i))^2.
\]

- Posterior predictive loss (PPL):

\[
PPL_m(\hat{g}_p) = \sum_{i=1}^{n} \sigma_{g_p}(x_i) + \frac{m}{m+1} \sum_{i=1}^{n} (g_p(x_i) - \hat{g}_p(x_i))^2,
\]
where $\hat{g}_p(x_i)$ is the estimate of $g_p(x_i)$ and $\sigma^2_\hat{g}_p(x_i)$ is the posterior predictive distribution’s variance for $g_p(x_i)$.

The first component in the PPL is a penalty term for model complexity, and the second is a term for goodness-of-fit. $m$ is the weight term that controls this trade-off; the smaller value of $m$ gives more penalty to complex model. We use $m = 1, \infty$ suggested in Gelfand and Ghosh (1998).

### 4.1 Simulation result

Table 1 summarizes the simulation results for Dataset1 with normal errors: All the models that we test exhibit superior performance to the naive model, which directly treats the proxy with error $w_2$ as a true covariate (see Appendix S1). The proposed BEMP outperforms the comparative method across all metrics. For the models using two proxies, BEMP-poly outperforms BEMP-nonlinear, though we matched the ratio of variance in the error components to the total variance in the proxy variables. This is a predictable result because BEMP-nonlinear uses two sets of spline parameters, and the number of parameters to be estimated may adversely affect the resulting performance (Hastie et al., 2009). However, notably, BEMP-all exhibits the most optimal performance, even better than BEMP-poly. The superior performance of BEMP-all is interesting because the number of parameters in BEMP-all is larger than that in BEMP-nonlinear. We believe that this is because when the amount of proxy used increases, the increased amount of information aids the estimation of each other and provides a more precise estimation. In other words, a more precise estimation of $x$ makes it easier to estimate the spline parameter and vice versa. A similar discussion can be found in Berry et al. (2002) and Lubotsky and Wittenberg (2006), which argued for the effect of using multiple types of information. To investigate the model’s performance in various aspects, we also provide a comparison of the computation time of the proposed method in Table 2. Despite the fact that the proposed algorithm requires the estimation of a large number of parameters due to its application in the absence of true covariates, the use of a conjugate prior makes the proposed method efficient enough in terms of the number of parameters to be estimated.

### Table 1 Monte Carlo means (standard errors) of MSE, PPL1, and PPL∞ for homogeneous Dataset1 with standard normal distributed error.

| Quantile | woME                  | Structural            | BEMP-poly             | BEMP-nonlinear      | BEMP-all             |
|----------|-----------------------|-----------------------|-----------------------|---------------------|----------------------|
| $p = 0.1$| MSE 0.029 (0.013)     | 0.163 (0.038)        | 0.09 (0.031)          | 0.146 (0.06)        | 0.063 (0.027)        |
|          | PPL1 0.08 (0.009)     | 0.204 (0.041)        | 0.115 (0.019)         | 0.137 (0.034)       | 0.1 (0.017)          |
|          | PPL∞ 0.094 (0.015)    | 0.287 (0.053)        | 0.161 (0.033)         | 0.214 (0.062)       | 0.132 (0.028)        |
| $p = 0.25$| MSE 0.019 (0.008)    | 0.129 (0.03)         | 0.053 (0.018)         | 0.097 (0.035)       | 0.035 (0.015)        |
|          | PPL1 0.05 (0.005)     | 0.15 (0.024)         | 0.069 (0.009)         | 0.089 (0.019)       | 0.061 (0.009)        |
|          | PPL∞ 0.059 (0.009)    | 0.216 (0.034)        | 0.097 (0.017)         | 0.134 (0.035)       | 0.079 (0.016)        |
| $p = 0.5$ | MSE 0.016 (0.007)    | 0.113 (0.029)        | 0.04 (0.017)          | 0.087 (0.04)        | 0.03 (0.014)         |
|          | PPL1 0.038 (0.005)    | 0.124 (0.021)        | 0.053 (0.008)         | 0.071 (0.015)       | 0.045 (0.007)        |
|          | PPL∞ 0.046 (0.008)    | 0.186 (0.032)        | 0.074 (0.016)         | 0.113 (0.034)       | 0.06 (0.013)         |
| $p = 0.75$| MSE 0.016 (0.01)     | 0.126 (0.032)        | 0.071 (0.027)         | 0.155 (0.035)       | 0.058 (0.038)        |
|          | PPL1 0.04 (0.005)     | 0.143 (0.019)        | 0.063 (0.011)         | 0.088 (0.015)       | 0.055 (0.015)        |
|          | PPL∞ 0.049 (0.009)    | 0.209 (0.03)         | 0.096 (0.025)         | 0.165 (0.032)       | 0.083 (0.033)        |
| $p = 0.9$ | MSE 0.03 (0.026)     | 0.167 (0.044)        | 0.129 (0.025)         | 0.159 (0.034)       | 0.128 (0.036)        |
|          | PPL1 0.059 (0.012)    | 0.203 (0.028)        | 0.084 (0.011)         | 0.094 (0.016)       | 0.08 (0.013)         |
|          | PPL∞ 0.075 (0.024)    | 0.293 (0.043)        | 0.146 (0.021)         | 0.175 (0.032)       | 0.143 (0.03)         |

Note: Bold entries denote the best performance among models relying on proxy variables, excluding “woME”.

### Table 2 Comparison of computation efficiency between natural cubic spline without measurement error and BEMP-poly with natural cubic spline method in Dataset1 with normal errors.

|             | CPU    | # Params |
|-------------|--------|---------|
| woME        | 6.64   | 31      |
| BEMP-poly   | 27.56  | 1038    |
| BEMP-nonlinear | 32.80 | 1066    |
| BEMP-all    | 41.77  | 1070    |

Note: CPU: CPU time in seconds cost by a 30,000 steps on a 3.22-GHz personal computer. # Params: number of parameters needed to be estimated.
The results are presented in Table 3, which summarizes the simulation results for Dataset1 with the Student t distribution error. With t distribution, the performance of BEMP-nonlinear exhibits a sensitive result to the outliers induced from heavy-tailed error compared to BEMP-poly. Consequently, in some cases, BEMP-all exhibits inferior performance to BEMP-poly, which uses true polynomial structures for $h_2$. However, it still demonstrates the effect of using multiple proxies, with BEMP-all outperforming BEMP-nonlinear significantly.

The results with the gamma-distributed error in Table 4 also present a similar trend. BEMP-all exhibits the most optimal performance, revealing that the proposed method has an efficient estimator with more information gained from combining proxy variables. We present the visualization of the gamma distribution error in Figure 2, where the proposed method exhibits the largest performance gap between the woMEs. Figure 2a reveals that the BEMP-nonlinear approach fails to capture the true relationship’s overall shape compared with BEMP-quad and worsens BEMP-all's performance. However, BEMP-nonlinear still works better than the structural method, and the performance drop in BEMP-all is

---

**Table 3** Monte Carlo means (standard errors) of MSE, PPL1, and PPL$_\infty$ for homogeneous Dataset1 with Student t distributed error.

| Quantile | woME   | Structural | BEMP-poly | BEMP-nonlinear | BEMP-all |
|----------|--------|------------|-----------|----------------|----------|
| $p = 0.1$ | MSE    | 0.123 (0.076) | 0.319 (0.147) | **0.161 (0.271)** | 0.311 (0.202) | 0.181 (0.11) |
|          | PPL$_1$ | 0.192 (0.051) | 0.719 (0.19) | **0.232 (0.174)** | 0.259 (0.108) | 0.197 (0.062) |
|          | PPL$_\infty$ | 0.253 (0.087) | 0.884 (0.218) | **0.304 (0.307)** | 0.414 (0.208) | 0.296 (0.114) |
| $p = 0.25$ | MSE | 0.031 (0.017) | 0.164 (0.042) | **0.064 (0.027)** | 0.118 (0.057) | 0.053 (0.023) |
|          | PPL$_1$ | 0.072 (0.01) | 0.23 (0.047) | **0.089 (0.023)** | 0.117 (0.031) | 0.083 (0.017) |
|          | PPL$_\infty$ | 0.086 (0.018) | 0.313 (0.059) | **0.121 (0.035)** | 0.175 (0.058) | 0.112 (0.027) |
| $p = 0.5$ | MSE | 0.019 (0.009) | 0.128 (0.029) | 0.039 (0.014) | 0.083 (0.036) | 0.033 (0.016) |
|          | PPL$_1$ | 0.042 (0.005) | 0.152 (0.026) | 0.058 (0.008) | 0.077 (0.016) | 0.052 (0.008) |
|          | PPL$_\infty$ | 0.051 (0.009) | 0.222 (0.033) | 0.077 (0.014) | 0.12 (0.033) | 0.07 (0.015) |
| $p = 0.75$ | MSE | 0.035 (0.023) | 0.172 (0.048) | **0.089 (0.037)** | 0.159 (0.035) | 0.098 (0.051) |
|          | PPL$_1$ | 0.056 (0.012) | 0.215 (0.036) | 0.077 (0.017) | 0.093 (0.014) | **0.076 (0.019)** |
|          | PPL$_\infty$ | 0.074 (0.023) | 0.299 (0.052) | 0.124 (0.033) | 0.174 (0.03) | **0.123 (0.044)** |
| $p = 0.9$ | MSE | 0.118 (0.08) | 0.321 (0.229) | 0.17 (0.117) | 0.213 (0.176) | 0.164 (0.092) |
|          | PPL$_1$ | 0.12 (0.045) | 0.752 (0.237) | 0.12 (0.116) | 0.139 (0.103) | **0.108 (0.056)** |
|          | PPL$_\infty$ | 0.178 (0.082) | 0.917 (0.311) | 0.207 (0.172) | 0.246 (0.19) | **0.189 (0.101)** |

Note: Bold entries denote the best performance among models relying on proxy variables, excluding “woME”.

---

**Table 4** Monte Carlo means (standard errors) of MSE, PPL1, and PPL$_\infty$ for homogeneous Dataset1 with gamma distributed error.

| Quantile | woME   | Structural | BEMP-poly | BEMP-nonlinear | BEMP-all |
|----------|--------|------------|-----------|----------------|----------|
| $p = 0.1$ | MSE    | 0.038 (0.024) | 0.167 (0.047) | 0.126 (0.032) | 0.144 (0.04) | **0.125 (0.038)** |
|          | PPL$_1$ | 0.064 (0.011) | 0.247 (0.047) | 0.085 (0.014) | 0.088 (0.017) | **0.078 (0.013)** |
|          | PPL$_\infty$ | 0.083 (0.022) | 0.323 (0.059) | 0.148 (0.028) | 0.162 (0.036) | **0.141 (0.03)** |
| $p = 0.25$ | MSE | 0.032 (0.015) | 0.159 (0.045) | 0.065 (0.028) | 0.103 (0.036) | **0.061 (0.026)** |
|          | PPL$_1$ | 0.053 (0.008) | 0.224 (0.037) | 0.068 (0.011) | 0.082 (0.023) | **0.062 (0.012)** |
|          | PPL$_\infty$ | 0.068 (0.015) | 0.3 (0.048) | 0.099 (0.025) | 0.136 (0.04) | **0.093 (0.024)** |
| $p = 0.5$ | MSE | 0.045 (0.02) | 0.176 (0.054) | 0.058 (0.027) | 0.152 (0.082) | **0.056 (0.035)** |
|          | PPL$_1$ | 0.069 (0.013) | 0.273 (0.044) | 0.084 (0.017) | 0.136 (0.048) | **0.079 (0.029)** |
|          | PPL$_\infty$ | 0.092 (0.022) | 0.368 (0.054) | 0.114 (0.029) | 0.206 (0.087) | **0.108 (0.045)** |
| $p = 0.75$ | MSE | 0.093 (0.04) | 0.252 (0.088) | **0.119 (0.079)** | 0.327 (0.142) | 0.148 (0.082) |
|          | PPL$_1$ | 0.126 (0.029) | 0.463 (0.072) | 0.177 (0.068) | 0.274 (0.076) | **0.176 (0.06)** |
|          | PPL$_\infty$ | 0.174 (0.047) | 0.597 (0.089) | **0.238 (0.105)** | 0.438 (0.145) | 0.251 (0.098) |
| $p = 0.9$ | MSE | 0.205 (0.101) | 0.428 (0.177) | **0.256 (0.182)** | 0.472 (0.178) | 0.344 (0.176) |
|          | PPL$_1$ | 0.254 (0.068) | 0.951 (0.171) | **0.318 (0.13)** | 0.401 (0.102) | **0.327 (0.102)** |
|          | PPL$_\infty$ | 0.361 (0.116) | 1.176 (0.2) | **0.454 (0.217)** | 0.633 (0.189) | **0.501 (0.188)** |

Note: Bold entries denote the best performance among models relying on proxy variables, excluding “woME”.
moderate, revealing a fitted line like that of BEMP-quad. For the extreme quantiles, BEMP-all and BEMP-quad fail to capture the curvature on the right side of the domain in $p = 0.1$, whereas the methods in $p = 0.9$ fail to capture the curvature on the left side of the domain.

For Dataset2, the simulation results are summarized in Tables 5–7. While for Dataset2, the heteroscedasticity and fluctuating function with large higher derivatives make the estimation difficult (Figure 2b), and the overall result remains the same. In this dataset, the performance of BEMP-all outperforms across all error distributions. Generalizing this limited simulation and suggesting that BEMP-all outperforms in all cases is difficult, but in practice, we rarely have information regarding the true polynomial structure of the proxy or the quality of each proxy; remarkably, the result is sufficiently encouraging to generally use the BEMP-all method.

### 4.2 Estimation of nonlinear proxy

To empirically examine the proposed model's performance in estimating the proxy relationship and further validate the assumption regarding the effect of multiple proxies, we assess the model estimation of $h_3(x)$, the nonlinear relationship between the unobserved covariate and proxy. We evaluate the same metrics used in Section 4.1 for the posterior result of $h_3(x)$.

---

**FIGURE 2** Fitted line for estimators in Dataset1 (a) and Dataset2 (b), with gamma distributed error in $p_0 = 0.1, 0.5, 0.9$. The dotted black line represents the true quantile function.

**TABLE 5** Monte Carlo means (standard errors) of MSE, PPL1, and PPL∞ for heterogeneous Dataset2 with standard normal distributed error.

| Quantile | Standard normal error | woME       | Structural | BEMP-poly | BEMP-nonlinear | BEMP-all |
|----------|-----------------------|------------|------------|-----------|----------------|----------|
| $p = 0.1$ | MSE               | 0.048 (0.018) | 0.481 (0.06) | 0.333 (0.041) | 0.429 (0.066) | 0.225 (0.049) |
|          | PPL1               | 0.067 (0.008) | 0.371 (0.054) | 0.192 (0.02) | 0.233 (0.032) | 0.138 (0.023) |
|          | PPL∞               | 0.089 (0.016) | 0.616 (0.076) | 0.359 (0.04) | 0.449 (0.064) | 0.252 (0.047) |
| $p = 0.25$ | MSE              | 0.022 (0.01) | 0.32 (0.057) | 0.204 (0.037) | 0.296 (0.053) | 0.112 (0.029) |
|          | PPL1              | 0.04 (0.006) | 0.248 (0.046) | 0.125 (0.018) | 0.16 (0.026) | 0.078 (0.014) |
|          | PPL∞              | 0.05 (0.011) | 0.412 (0.07) | 0.227 (0.036) | 0.308 (0.053) | 0.135 (0.028) |
| $p = 0.5$ | MSE             | 0.018 (0.009) | 0.158 (0.032) | 0.094 (0.02) | 0.166 (0.036) | 0.063 (0.016) |
|          | PPL1            | 0.035 (0.005) | 0.156 (0.033) | 0.072 (0.009) | 0.096 (0.018) | 0.054 (0.008) |
|          | PPL∞            | 0.043 (0.009) | 0.233 (0.046) | 0.119 (0.019) | 0.18 (0.035) | 0.085 (0.015) |
| $p = 0.75$ | MSE           | 0.021 (0.008) | 0.221 (0.042) | 0.112 (0.023) | 0.2 (0.048) | 0.07 (0.015) |
|           | PPL1          | 0.041 (0.005) | 0.198 (0.038) | 0.088 (0.012) | 0.119 (0.023) | 0.063 (0.008) |
|           | PPL∞          | 0.051 (0.008) | 0.303 (0.056) | 0.144 (0.023) | 0.22 (0.047) | 0.098 (0.015) |
| $p = 0.9$ | MSE        | 0.033 (0.012) | 0.488 (0.1) | 0.282 (0.05) | 0.437 (0.145) | 0.155 (0.06) |
|           | PPL1        | 0.064 (0.009) | 0.356 (0.071) | 0.191 (0.027) | 0.251 (0.07) | 0.125 (0.032) |
|           | PPL∞        | 0.082 (0.014) | 0.608 (0.116) | 0.331 (0.052) | 0.466 (0.143) | 0.205 (0.062) |

Note: Bold entries denote the best performance among models relying on proxy variables, excluding “woME”.

---
Table 6 summarizes the result for Dataset1 and Dataset2 for all types of error distribution. We verify that the proposed method successfully estimates the nonlinear relationship between the proxy and true covariate with moderate performance across all metrics. Additionally, BEMP-all exhibits superior performance to BEMP-nonlinear in estimating the proxy’s nonlinear relationship. This result supports our assumption that information from other proxies improves the estimation of the proxy’s current relationship. The improved estimation of the relationship between the proxy and covariate directly affects the unobserved covariate’s estimation, which might be important for the estimated quantile function’s performance.

| Quantile | Student t error | Structural | BEMP-poly | BEMP-nonlinear | BEMP-all |
|----------|-----------------|------------|-----------|----------------|---------|
| p = 0.1  | MSE             | 0.116 (0.06) | 0.553 (0.146) | 0.323 (0.117) | 0.39 (0.149) | 0.285 (0.145) |
|          | PPL_1           | 0.131 (0.035) | 0.62 (0.188) | 0.198 (0.057) | 0.222 (0.072) | 0.182 (0.071) |
|          | PPL_\infty     | 0.194 (0.063) | 0.887 (0.241) | 0.358 (0.116) | 0.418 (0.146) | 0.326 (0.143) |
| p = 0.25 | MSE             | 0.035 (0.016) | 0.339 (0.064) | 0.222 (0.048) | 0.296 (0.075) | 0.137 (0.043) |
|          | PPL_1           | 0.055 (0.008) | 0.308 (0.052) | 0.14 (0.024)  | 0.167 (0.036) | 0.095 (0.023) |
|          | PPL_\infty     | 0.071 (0.015) | 0.492 (0.072) | 0.252 (0.047) | 0.312 (0.073) | 0.162 (0.044) |
| p = 0.5  | MSE             | 0.022 (0.01)  | 0.161 (0.042) | 0.096 (0.021) | 0.167 (0.037) | 0.066 (0.019) |
|          | PPL_1           | 0.039 (0.006) | 0.18 (0.039)  | 0.076 (0.01)  | 0.102 (0.018) | 0.058 (0.009) |
|          | PPL_\infty     | 0.05 (0.01)   | 0.263 (0.056) | 0.125 (0.02)  | 0.187 (0.036) | 0.09 (0.019)  |
| p = 0.75 | MSE             | 0.036 (0.026) | 0.251 (0.06)  | 0.132 (0.024) | 0.183 (0.081) | 0.086 (0.024) |
|          | PPL_1           | 0.056 (0.013) | 0.26 (0.05)   | 0.107 (0.013) | 0.123 (0.038) | 0.077 (0.014) |
|          | PPL_\infty     | 0.075 (0.026) | 0.377 (0.074) | 0.174 (0.024) | 0.214 (0.078) | 0.12 (0.025)  |
| p = 0.9  | MSE             | 0.097 (0.091) | 0.544 (0.166) | 0.308 (0.101) | 0.341 (0.146) | 0.22 (0.106)  |
|          | PPL_1           | 0.123 (0.051) | 0.64 (0.163)  | 0.236 (0.06)  | 0.227 (0.073) | 0.173 (0.058) |
|          | PPL_\infty     | 0.167 (0.095) | 0.938 (0.217) | 0.386 (0.109) | 0.394 (0.145) | 0.28 (0.11)   |

Note: Bold entries denote the best performance among models relying on proxy variables, excluding “woME”.

Table 7 summarizes the result for Dataset1 and Dataset2 for all types of error distribution. We verify that the proposed method successfully estimates the nonlinear relationship between the proxy and true covariate with moderate performance across all metrics. Additionally, BEMP-all exhibits superior performance to BEMP-nonlinear in estimating the proxy’s nonlinear relationship. This result supports our assumption that information from other proxies improves the estimation of the proxy’s current relationship. The improved estimation of the relationship between the proxy and covariate directly affects the unobserved covariate’s estimation, which might be important for the estimated quantile function’s performance.

| Quantile | Gamma error | woME       | Structural | BEMP-poly | BEMP-nonlinear | BEMP-all |
|----------|-------------|------------|------------|-----------|----------------|---------|
| p = 0.1  | MSE         | 0.053 (0.022) | 0.497 (0.11) | 0.362 (0.059) | 0.414 (0.106) | 0.217 (0.051) |
|          | PPL_1       | 0.071 (0.012) | 0.446 (0.1)  | 0.215 (0.03) | 0.226 (0.053) | 0.138 (0.024) |
|          | PPL_\infty  | 0.096 (0.022) | 0.69 (0.14)  | 0.396 (0.059) | 0.432 (0.106) | 0.245 (0.049) |
| p = 0.25 | MSE         | 0.036 (0.017) | 0.265 (0.075) | 0.163 (0.036) | 0.202 (0.072) | 0.101 (0.03)  |
|          | PPL_1       | 0.056 (0.009) | 0.29 (0.076) | 0.114 (0.018) | 0.122 (0.034) | 0.08 (0.015)  |
|          | PPL_\infty  | 0.072 (0.016) | 0.429 (0.103) | 0.197 (0.036) | 0.221 (0.07)  | 0.131 (0.029) |
| p = 0.5  | MSE         | 0.046 (0.023) | 0.163 (0.056) | 0.109 (0.03) | 0.117 (0.056) | 0.088 (0.033) |
|          | PPL_1       | 0.062 (0.013) | 0.268 (0.058) | 0.093 (0.017) | 0.093 (0.031) | 0.077 (0.018) |
|          | PPL_\infty  | 0.084 (0.024) | 0.356 (0.068) | 0.148 (0.031) | 0.151 (0.059) | 0.119 (0.034) |
| p = 0.75 | MSE         | 0.076 (0.042) | 0.317 (0.091) | 0.161 (0.06) | 0.177 (0.085) | 0.131 (0.059) |
|          | PPL_1       | 0.092 (0.022) | 0.465 (0.093) | 0.14 (0.035) | 0.141 (0.046) | 0.114 (0.031) |
|          | PPL_\infty  | 0.129 (0.042) | 0.643 (0.113) | 0.218 (0.065) | 0.229 (0.087) | 0.178 (0.06)  |
| p = 0.9  | MSE         | 0.153 (0.062) | 0.824 (0.209) | 0.31 (0.108) | 0.4 (0.171) | 0.26 (0.096) |
|          | PPL_1       | 0.157 (0.038) | 1.12 (0.261) | 0.247 (0.063) | 0.26 (0.088) | 0.188 (0.053) |
|          | PPL_\infty  | 0.227 (0.067) | 1.522 (0.318) | 0.399 (0.116) | 0.459 (0.173) | 0.318 (0.1) |

Note: Bold entries denote the best performance among models relying on proxy variables, excluding “woME”.

Table 8 summarizes the result for Dataset1 and Dataset2 for all types of error distribution. We verify that the proposed method successfully estimates the nonlinear relationship between the proxy and true covariate with moderate performance across all metrics. Additionally, BEMP-all exhibits superior performance to BEMP-nonlinear in estimating the proxy’s nonlinear relationship. This result supports our assumption that information from other proxies improves the estimation of the proxy’s current relationship. The improved estimation of the relationship between the proxy and covariate directly affects the unobserved covariate’s estimation, which might be important for the estimated quantile function’s performance.
presents the MSE between the posterior samples of the unobserved covariate
Monte Carlo means (standard errors) of MSE, PPL₁, and PPL₂, evaluated for h₃ in Dataset1 (top) and Dataset2 (bottom).

4.3 Estimation of unobserved covariate

Although the estimation of the regression function \( g_\* \) is the primary focus in most cases, the inference of the mismeasured covariate \( x \) can be another important interest. To further investigate the proposed method, we examine the posterior samples to estimate the unobserved covariate \( x \). Table 9 presents the MSE between the posterior samples of the unobserved covariate \( x \) and their true value in Dataset1 with a normal error. Notably, with the naive approach using \( w_\* \) instead of \( x \), the MSE is \( \sigma^2 \) in expectation, which is 1 in our simulation. The MSE of the estimation of \( x \) exhibits patterns consistent with the other parameters. For all the models that we test, the MSE is smaller than that of the naive approach. BEMP-nonlinear and structural methods exhibit similar performances, whereas adding a nonlinear proxy still boosts the estimation of \( x \).

This is evident in that BEMP-all generally outperforms BEMP-quad in all the quantiles, which, again, demonstrates multiple proxies’ effectiveness. The results from the other cases exhibit similar patterns (Appendix S1).
We apply the proposed method to a real dataset that includes asset and income variables. Statistics Korea released microdata from the Survey of Household Finances and Living Conditions (SFLC), incorporating administrative data obtained from other government institutions. The released dataset includes basic demographic variables for 18,064 families and various economic features, such as salary income, property income, assets, asset management plans, debt, and debt repayment capacity for each family unit collected in 2020.

This application aims to determine the quantile relationship between assets and true salary income. Income data provide critical information for a wide range of policies. However, administrative salary income is prone to measurement error, so the direct use of this information can precipitate misleading inferences (Davern et al., 2005; Moore et al., 2000). We consider using administrative salary income and property income as two types of proxies: one is exposed to additive error, and the other is a correlated proxy. These values are suitable for use as proxies because it is reasonable to assume that property income and salary income have a high correlation (Lerman & Yitzhaki, 1985). Consequently, the model is described as follows:

\[
\begin{align*}
\text{asset}_i & = g_p(\text{true salary income}_i) + e_i, \\
\text{administrative salary income}_i & = \text{true salary income}_i + u_{1i}, \\
\text{property income}_i & = \alpha_0 + \alpha_1 \text{true salary income}_i + u_{2i}.
\end{align*}
\]

Notably, we assume that we do not observe true salary income, but instead observe multiple proxies, administrative income and property income. To model the correlated proxy property income, we utilize a linear regression for the relationship between the proxy and
covariate with parameter $\alpha$. As a preprocessing step, we eliminate the extreme quantiles (i.e., 0.999 and 0.001 percentiles in terms of each variable). After preprocessing, the data comprises 11,317 family units. Further, we attempt a log transform for asset variables to alleviate data skewness and improve model convergence. Following Thompson et al.’s (2010) suggestion, we take $N = 30$ knots equally spaced over the range of variables, which is log-transformed administrative salary income.

We investigate the effect of true salary income on asset in different quantile $p \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Figure 3 presents the resulting quantile lines. In Figure 3, the fitted relation’s spread is not parallel, implying that a heterogeneous effect exists. The fitted quantile function of $p = 0.9$ presents a larger gap between the other fitted quantile functions, which indicates that the conditional distribution of asset is not symmetrical and right-skewed. More interestingly, for higher quantiles, such as $p = 0.9$ and $p = 0.75$, the fitted line in the lower level of administrative salary income ($\leq 5k$) generally curved upward, with its highest point in administrative salary income $= 3k$.

This outcome might be attributable to various reasons. For example, people with their assets not structurally proportional to salary income or those with the upper end of asset value might structurally misreport their salary income, which is in line with the literature (Moore et al., 2000; Stocké, 2006; Valet et al., 2019). In either case, nonparametric quantile regression can derive a novel insight that was impossible through parametric regression and is more informative in combining proxies than the naïve approach.

6 | CONCLUSION

This study proposes a Bayesian quantile regression estimation method that integrates multiple proxies obtained from multiple datasets. A simulation study on various datasets demonstrates that the proposed method can accommodate multiple proxies with linear and nonlinear relationships with the true covariate. We demonstrate that the proposed method is promising for capturing the underlying relationship, effective in incorporating multiple proxies simultaneously, and making reliable estimations of unobserved covariates and their relationships with proxies. Further, we presented an application of this methodology using a public SFLC dataset and provided the underlying relationship between assets and salary income in the presence of multiple income records.

This study has some limitations. We adopted a spline function for quantile regression; thus, the fitted regression line’s behavior tends to be erratic near the boundaries, which is a well-known characteristic of the spline approach (Hastie et al., 2009). Further, noteworthy, both natural cubic spline quantile regression (Thompson et al., 2010) and P-spline quantile regression (Lang & Brezger, 2004) use the asymmetric Laplace distribution (ALD) for errors (Yu & Moyeed, 2001). While the choice of ALD as a working likelihood is based on its efficient Bayesian inference in the proposed model and the result was reasonably flexible for estimation purposes, the efficiency of ALD as a working likelihood comes at the expense of direct posterior inference validity, and reasoning based on credible intervals obtained from the posterior should be regarded with caution (Yang et al., 2016). With a similar framework, alternative approaches such as that of Dunson and Taylor (2005), Kottas and Krnjajić (2009), and Taddy and Kottas (2010) can also be applied to settings with multiple proxies. We leave the extension of this parametric assumption for future research. Finally, the flexible estimation of the quantile and proxy functions can lead to an increased number of parameters, which may result in instability. The proposed method relies on the presence of benchmark variable to stabilize the estimation by fixing the scale and mitigate these issues. The relaxation of this assumption and its impact on the robustness of the method are left as a topic for future research.

ACKNOWLEDGEMENTS

We thank the editor, associate editor, and one reviewer for their constructive comments. Ick Hoon Jin is partially supported by the Yonsei University Research Fund of 2019-22-0210 and by the National Research Foundation (NRF) Korea (NRF-2020R1A2C1A01009881). Jongho Im’s research is also supported by the NRF Korea (NRF-2021R1C1C1014407).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jongho Im https://orcid.org/0000-0001-8362-4756

REFERENCES

Aigner, D. J., Hsiao, C., Kapteyn, A., & Wansbeek, T. (1984). Latent variable models in econometrics. Handbook of Econometrics, 2, 1321–1393.

Berg, E., Im, J., Zhu, Z., Colin, L.-B., & Li, J. (2021). Integration of statistical and administrative agricultural data from Namibia. Statistical Journal of the IAOS, 37, 557–578.

Berry, S. M., Carroll, R. J., & Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. Journal of the American Statistical Association, 97(457), 160–169.

Brezger, A., & Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. Computational Statistics & Data Analysis, 50(4), 967–991.
Yu, K., & Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters, 54*(4), 437–447.

Yue, Y. R., & Rue, H. (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis, 55*(1), 84–96.

Zimmerman, D. J. (1992). Regression toward mediocrity in economic stature. *The American Economic Review, 82*(3), 409–429.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

---

**How to cite this article:** Go, D., Hoon Jin, I., & Im, J. (2023). Quantile regression with multiple proxy variables. *Stat, 12*(1), e547. [https://doi.org/10.1002/sta4.547](https://doi.org/10.1002/sta4.547)