Semiparametrically efficient estimation in quantile regression of secondary analysis

Liang Liang,
Texas A&M University, College Station, USA

Yanyuan Ma,
Penn State University, University Park, USA

Ying Wei, and
Columbia University, New York, USA

Raymond J. Carroll
Texas A&M University, College Station, USA, and University of Technology, Sydney, Australia

Summary

Analysing secondary outcomes is a common practice for case–control studies. Traditional secondary analysis employs either completely parametric models or conditional mean regression models to link the secondary outcome to covariates. In many situations, quantile regression models complement mean-based analyses and provide alternative new insights on the associations of interest. For example, biomedical outcomes are often highly asymmetric, and median regression is more useful in describing the ‘central’ behaviour than mean regressions. There are also cases where the research interest is to study the high or low quantiles of a population, as they are more likely to be at risk. We approach the secondary quantile regression problem from a semiparametric perspective, allowing the covariate distribution to be completely unspecified. We derive a class of consistent semiparametric estimators and identify the efficient member. The asymptotic properties of the resulting estimators are established. Simulation results and a real data analysis are provided to demonstrate the superior performance of our approach with a comparison with the only existing approach so far in the literature.

Keywords

Biased samples; Case–control study; Heteroscedastic errors; Quantile regression; Secondary analysis; Semiparametric estimation
1. Introduction

In a cohort study of a disease, all individuals who are recruited are followed prospectively and their covariate information, i.e. risk factors related to the disease status, are collected even though only a small fraction of those individuals will develop the disease. As a consequence, cohort studies often demand very large sample sizes and can be both time consuming and costly, especially for rare diseases. The case–control design offers a cost-effective alternative to cohort designs. It separates the underlying source population into the case subpopulation that is formed by all diseased individuals, and the control subpopulation that is formed by all non-diseased individuals, and randomly samples a fixed number of cases and controls from them respectively. After the cases and controls have been identified, the covariates are ascertained. The resulting case–control sample is no longer a random sample of the underlying source population. The analysis for determining the relationship between disease status $D$ and the covariates is known as primary analysis of the case–control data. Prospective logistic regression that ignores the case–control sampling scheme is often used to perform the primary analysis, and Prentice and Pyke (1979) showed consistency and efficiency of the estimator for the slopes in the model when the rate of disease and the relationship between the covariates are both unspecified.

Besides the primary analyses, researchers often investigate the associations between the covariates and some secondary disease-related outcomes, which is known as the secondary analysis of case–control studies (Jiang et al., 2006; Lin and Zeng, 2009; Nagelkerke et al., 1995). Throughout the paper, we shall denote $Y$ as a secondary outcome, and $X$ as covariates. Direct use of routine analytic methods in secondary analyses, such as simply regressing $Y$ against $X$, generally produce biased estimates because a case–control sample is not a representative sample of the underlying source population. Well-developed approaches in secondary analysis that account for the case–control sampling nature commonly assume that the $Y$ given $X$ model is fully parametric or that the regression mean function $E(Y|X)$ is specified (Ghosh et al., 2013; Jiang et al., 2006; Lee et al., 1997; Li et al., 2010; Lin and Zeng, 2009; Ma and Carroll, 2016; Monsees et al., 2009; Richardson et al., 2007; Wei et al., 2013), e.g. $\logit\{\Pr(Y=1|X)\}=X^T\beta$ for a binary secondary outcome and $Y=X^T\beta+\varepsilon$ for a continuous secondary outcome.

However, in epidemiological studies, it is often of interest to study the quantiles of $Y$ as a function of $X$. In such cases, quantile regression (Koenker and Bassett, 1978) can provide a complete picture of the relationship between covariates and secondary outcome at any percentile. For example, Terry et al. (2007) found some prenatal and early life exposures influence only the upper quantiles of body mass index. Thus inferences that are focused on solely the mean of an outcome would miss or underestimate those important associations. In addition, quantile regression is more robust to skewed distributions and outliers.

Very limited work has been done on quantile regression in the context of secondary analysis. To our best knowledge, the only existing method is the weighted estimating equation approach that was proposed by Wei et al. (2016). The main idea of the weighted estimating equation approach is to construct weighted estimating functions by using both observed and pseudocounterfactual secondary outcomes, where the counterfactual secondary outcomes...
refer to the potential outcomes under alternative disease status, and the weights are respectively \( \Pr(D=1 \mid X) \) and \( \Pr(D=0 \mid X) \). Here \( D=1 \) denotes a case and \( D=0 \) denotes a control. The weighted estimating equation approach can be implemented via the simulated counterfactual outcome method SICO and kernel smoothing. SICO simulates the counterfactual outcomes directly, whereas kernel smoothing estimates the mean counterfactual outcome and thus avoids simulating pseudo-outcomes. The performances of kernel smoothing and SICO were found to be similar by Wei et al. (2016).

SICO relies on two assumptions:

a. the disease prevalence rate in the underlying source population is known, and

b. a linear additive quantile regression model is assumed for all quantiles, not just for a single quantile, although it allows quantile-specific coefficients.

The first assumption is needed to estimate the weights \( \Pr(D=d \mid X) \) in the source population, whereas the second assumption is required to simulate counterfactual outcomes. Of course the rate of disease in the source population is not always available, and specifying quantile models at all quantile levels simultaneously is much more restrictive than the original quantile regression model at one specific quantile level. Besides, as shown in the simulation study of Wei et al. (2016), those two assumptions do not help SICO to gain much efficiency compared with doing ordinary quantile regression only on the controls, which is a method that is reasonable when the disease is rare, but also inefficient.

In this paper, we relax both assumptions of SICO. Specifically, we allow the disease rate to be unknown in the source population and assume a quantile regression model only at one specific quantile level, using a semiparametric approach (Ma, 2010; Ma and Carroll, 2016). We show that the resulting estimator is generally consistent and asymptotically normal, and in addition that it is semiparametrically efficient in certain circumstances. We show that our new estimator provides much improved results over SICO with much weaker assumptions. Although the general approach here is similar to that in Ma and Carroll (2016), the treatments are very different in quantile regression from those in mean regression. In fact, this is the first attempt in the literature to use such semiparametric treatments in combination with the concept of a hypothetical population (Ma, 2010), which is discussed in detail in Section 2, and to handle quantile regression for case–control data.

In Section 2, we introduce the secondary quantile regression model and the framework that we use, and we provide necessary identifiability conditions. In Section 3, we construct the semiparametrically efficient estimator through a conventional semiparametric approach. The implementation of the resulting estimator is described in Section 4, whereas its asymptotic properties are discussed in Section 5. In Section 6, we demonstrate the superiority of our semiparametrically efficient estimator over existing approaches via various simulation studies. Section 7 illustrates the practical application of our approach through the analysis of a colorectal cancer data set. Section 8 contains a short discussion. Technical details and proofs are given in Appendix A and in the on-line supplement.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from
2. Model and framework

2.1. Model

Recall that we denote disease status \( D \), with \( D=0 \) representing a control and \( D=1 \) representing a case. Let \( \mathbf{X} \) be the covariates of interest, which can be either continuous or discrete. Other than \( \mathbf{X} \), a secondary outcome \( Y \) is available, e.g. an important biomarker or characterization of the disease. The disease risk is related to the covariates \( \mathbf{X} \) and the secondary outcome \( Y \) through a logistic regression model:

\[
\Pr^{\text{true}}(D = d \mid \mathbf{X} = \mathbf{x}, Y = y) = \frac{\exp \{d(\alpha_c + \alpha_1^T\mathbf{x} + \alpha_2y)\}}{1 + \exp (\alpha_c + \alpha_1^T\mathbf{x} + \alpha_2y)}, \tag{1}
\]

where \( c \) is a constant. Here and throughout the text, the superscript ‘true’ is used to denote a model in the underlying source population from which we obtain the case–control sample.

Let \( q_{\tau, Y \mid \mathbf{X}} \) denote the \( \tau \)th conditional quantile of \( Y \) given \( \mathbf{X} \) in the underlying source population. We assume that it can be written as

\[
q_{\tau, Y \mid \mathbf{X}}^{\text{true}}(y, \mathbf{x}, \tau, \beta_{\tau}) = \beta_{\tau, c} + \mathbf{x}^T\beta_{\tau}, \tag{2}
\]

where \( \tau \in (0, 1) \). Here, we assume that model (2) holds for only one specific quantile level \( \tau \) of interest determined by the application, whereas the SICO method of Wei et al. (2016) assumes model (2) for all \( \tau \in (0, 1) \).

Model (2) can be written alternatively as

\[
Y = \beta_{\tau, c} + \mathbf{x}^T\beta_{\tau} + \epsilon_{\tau},
\]

where \( \epsilon_{\tau} \) has its \( \tau \)th conditional quantile given \( \mathbf{X} \) equal to 0, whereas its distribution is otherwise not specified, i.e. \( \int u_{\tau} \eta_{2}(\epsilon_{\tau}, \mathbf{x})\,d\epsilon_{\tau}=0 \), where \( u_{\tau} \equiv \mathbb{1}(\epsilon_{\tau} < 0) - \tau \), and \( \eta_{2}(\epsilon_{\tau}, \mathbf{x}) \) denotes the density function of \( \epsilon_{\tau} \) conditional on \( \mathbf{X} = \mathbf{x} \) in the true population, which is unknown. In addition, the distribution of \( \mathbf{X} \) is also unspecified.

Suppose that we draw a case–control sample from the underlying source population model (1)–(2) with the cases being oversampled. Subsequently, the association between \( \mathbf{X} \) and \( Y \) in the case–control sample may differ dramatically from the true association in the underlying source population. The main difficulty in secondary analysis is that the case–control sample is not a representative sample of the underlying source population. To overcome this problem, we adopt the general concept of the hypothetical population as in Ma (2010); Ma...
and Carroll (2016). Such a hypothetical population has the same case-to-control ratio as the case–control sample and is connected with the true population through the fact that they share the same distribution given disease status. Ma (2010) proved the first-order asymptotic equivalence between case–control sampling and random sampling in the hypothetical population. Her result permits us to view the case–control sample as a prospective random sample taken from this hypothetical population. The exact form of the hypothetical population for the secondary quantile regression problem is given in Section 2.2.

2.2. Hypothetical population

We consider the case that the rate of disease in the true population, \( \pi_1 \equiv \Pr[D=1] > 0 \), is unknown, and that it can be rare or common. Let \( \pi_0 = 1 - \pi_1 \). The goal is to estimate \( \alpha = (\alpha_x, \alpha_y^1, \alpha_y^2)^T \) and \( \beta = (\beta_{\tau,c}, \beta_{\tau,e}^1)^T \). According to Ma (2010), the case–control sample can be treated as a random sample from a hypothetical population. The hypothetical population has the disease to non-disease ratio \( n_1/n_0 \), and it retains the same joint distribution of \((X, Y)\) given disease status \( D \) as in the true source population. The density function of a random observation \((X, Y, D)\) from the hypothetical population is

\[
f_{X,Y,D}(x,y,d,\beta,\alpha,\eta_1,\eta_2) = f_D(d)f_{X,Y\mid D}(x,y,d)
\]

\[
= \frac{n_d}{n}f_{X,Y\mid D}(x,y,d)
\]

\[
= \frac{n_d}{n} \frac{\eta_1(x)\eta_2(e_x,x)}{\int \eta_1(v)\eta_2(e_x,v) f_{X,Y\mid D}(x,y,d,x,\alpha) \, \mu(v) \, \mu(y) \, dv \, dy}
\]

\[
= \frac{n_d}{n} \eta_1(y)\eta_2(y - \beta_{\tau,c} - x^T \beta_{\tau,e}) H(d,y,x,\alpha)
\]

\[
\int \eta_1(v)\eta_2(v - \beta_{\tau,c} - \tau^T \beta_{\tau,e}) H(d,y,v,x,\alpha) \, \mu(v) \, \mu(y) \, dv \, dy
\]

where \( n = n_0 + n_1 \), \( H(\cdot) \) is defined in equation (1) and \( \mu \) denotes a Lebesgue measure for a continuous random variable and a counting measure for a discrete random variable. The probability density and mass functions of \( X \) and \( Y \) given \( X \) are \( \eta_1 \) and \( \eta_2 \) respectively. Define \( e_x = Y - \beta_{\tau,c} - X^T \beta_{\tau,e} \). We have \( \eta_1, \eta_2 \geq 0, \int \eta_1(x) \, \mu(x) = 1, \int \eta_2(e_x, x) \, \mu(e_x) \, \mu(x) = 1 \) and \( \int \eta_2(e_x, x) \, \mu(e_x) \, \mu(x) = 1 \). However, both \( \eta_1 \) and \( \eta_2 \) have unknown forms. We use the notation \( \eta_1 \) and \( \eta_2 \) instead of \( f_{X,Y}^{\text{true}} \) and \( f_{Y}^{\text{true}} \) to emphasize that they are treated as infinite dimensional nuisance parameters in our approach.

Since the observations can be viewed as a random sample from the hypothetical population with model (3), we use classical semiparametric analysis (Bickel et al., 1993; Tsiatis, 2007) to derive what we call a locally efficient semiparametric estimator for \( \alpha \) and \( \beta \). In Section 3, we describe the general approach of the estimator proposed, but leave technique details to Appendices A.1 and A.2.
2.3. Identifiability

Before deriving our locally efficient semiparametric estimator, it is useful first to establish the identifiability of \( a \) and \( \beta \). We assume the following conditions.

**Assumption 1**—For any \( \delta > 0 \), there exists \( K > 0 \) such that \( \lim_{x^{(i)} \to \pm \infty} \Pr(e_\tau \leftarrow K|x) < \delta \), where \( x^{(i)} \) is the \( i \)th element of \( x \).

**Assumption 2**—\( a_1 + \beta_1 a_2 \neq 0 \) and \( a_2 \neq 0 \).

Assumption 1 ensures that the left-hand tail of \( e_\tau \) given \( x \) is not too heavy when an arbitrary element of \( x \) diverges to \( \pm \infty \). This is a natural condition to guarantee that the \( \tau \) th quantile signal \( \beta_{t,c} + X^\top \beta_\tau \) can be separated from the noise \( e_\tau \). Assumption 2 ensures that the logistic regression model (1) indeed depends on the value of \( \beta_\tau \). When it is violated, we find that \( \beta_{t,c} \) and \( a_2 \) are no longer identifiable. See sections S.2 and S.3 in the on-line supplement for detailed proofs. The identifiability result is stated in proposition 1 below, whereas its proof is provided in section S.1 in the supplement.

**Proposition 1**—Under assumptions 1 and 2, the parameters \( a \) and \( \beta \) are identifiable.
case–control sampling scheme, which enables consistent estimation of all parameters in both model (1) and model (2). Among the class of all semiparametric estimators, the optimal estimator, which is usually referred to as the semiparametrically efficient estimator, is the estimator with smallest variance.

We adopt a geometric approach (Bickel et al., 1993; Tsiatis, 2007) to derive the semiparametrically efficient estimator. Specifically, we consider a Hilbert space $\mathcal{H}$ that consists of all $p$-dimensional measurable functions with mean 0 and finite variance and define the inner product of two arbitrary functions in $\mathcal{H}$ to be their covariance. We then decompose the Hilbert space $\mathcal{H}$ as $\mathcal{H} = \Lambda \oplus \Lambda^\perp$, where $\Lambda$ is the nuisance tangent space and $\Lambda^\perp$ is the orthogonal complement of $\Lambda$. The semiparametrically efficient estimator can be solved from $\sum_{i=1}^{n} S_{\text{eff}}(X_i, Y_i, D_i; \theta, \eta) = 0$, where $S_{\text{eff}}$ is the projection of the score function $S_{\theta}$ onto $\Lambda^\perp$. Consequently, $S_{\text{eff}}$ is usually called the efficient score function.

Under the hypothetical population model (3), the nuisance tangent space has the form $\Lambda = \{g(e, x) - E[g(e,x) | d]: E_{\text{true}}(g)=0, E_{\text{true}}\{ u_g(e,x) | x=0, \text{almost surely}\} \}$. Its orthogonal complement is

$$\Lambda^\perp = \{h(d, e, x): E(h) = 0, E(h - E(h | D) | e, x) = 0, \text{almost surely} \}.$$ 

The detailed derivation of $\Lambda$ and $\Lambda^\perp$ is provided in Appendix A.1.

The projection of the score function $S_{\theta}$ onto $\Lambda^\perp$ is very mathematically involved. Here we list the final form of the efficient score function $S_{\text{eff}}$, while deferring all the technical details to Appendix A.2. In particular, we show that $S_{\text{eff}}(X, Y, D; \theta, \eta) = S(X, Y, D; \theta, \eta) - g(Y - \beta_{r,c} - X^T \beta_r, X) - (1 - D)v_0 - Dv_1$, where
\[ \pi_0 \equiv \text{pr}^{\text{true}}(D = 0) = \int \eta_1(x) \eta_2(e, x) H(0, x, y) d\mu(x) d\mu(y), \] (5)

\[ \pi_1 \equiv \text{pr}^{\text{true}}(D = 1) = \int \eta_1(x) \eta_2(e, x) H(1, x, y) d\mu(x) d\mu(y). \]

\[ b_0 \equiv E\{f_D | X, Y(1, X, Y) | D = 0\}, \]

\[ b_1 \equiv E\{f_D | X, Y(0, X, Y) | D = 1\}, \]

\[ c_0 \equiv E(S | D = 0) - E\{E(S | e, X) | D = 0\}, \]

\[ c_1 \equiv E(S | D = 1) - E\{E(S | e, X) | D = 1\}, \]

\[ \kappa(x, y) \equiv \left( \sum_{d = 0}^1 \eta_d H(d, x, y)/(n\pi_d) \right)^{-1}, \]

\[ t_1(x) \equiv E_{\text{true}}\{u^2 \kappa(X, Y) | x\}^{-1}, \]

\[ t_2(x) \equiv E_{\text{true}}\{u^2 E(S | e, X) | x\} - (c_0/b_0)E_{\text{true}}\{u^2 f_D | X, Y(0, X, Y) | x\}, \]

\[ t_3(x) \equiv -b_0^{-1}E_{\text{true}}\{u^2 f_D | X, Y(0, X, Y) | x\}, \]

\[ a(x) \equiv t_1(x)[t_2(x) + t_3(x)u_0], \]

\[ u_0 \equiv [1 - E\{u^2 t_1(X) \kappa(X, Y) | D = 0\}]^{-1}E\{u^2 t_1(X) t_2(X) \kappa(X, Y) | D = 0\}, \]

\[ u_1 \equiv -(n_0/n_1)u_0, \]

\[ v_0 \equiv (\pi_1/b_0)(u_0 + c_0), \]

\[ v_1 \equiv -(\pi_0/b_0)(u_0 + c_0), \]

\[ g(e, x) \equiv E(S | e, X) - u^a \kappa(x, y) - v_0 f_D | X, Y(0, X, Y) - v_1 f_D | X, Y(1, X, Y). \]

Although we provide the explicit expression of the efficient score \( S_{\text{eff}} \), some of the quantities in expressions (5), such as the expectations conditional on \( x \) in the underlying source population, need to be estimated non-parametrically. The details are provided in Section 4.

The semiparametrically efficient estimator is then obtained by solving
\[
\sum_{i=1}^{n} \{S(X_i, Y_i, D_i) - g(Y_i - X_i^T \beta, X_i)\} - n_0v_0 - n_1v_1 = 0. \quad (6)
\]

The estimating equation (6) involves the disease and non-disease rate, \(\pi_1\) and \(\pi_0 = 1 - \pi_1\), which are assumed to be unknown. However,

\[
\pi_0 = \int H(0, X, Y, \alpha)f_{Y|X}(y, x, \beta)f_{X}(x)d\mu(x)d\mu(y)
\]

\[
= \int \sum_d N_d/(N\pi_d)H(d, X, Y, \alpha)f_{Y|X}(y, x, \beta)f_{X}(x)d\mu(x)d\mu(y)
\]

\[
= \int \sum_d N_d/(N\pi_d)H(d, X, Y, \alpha)f_{X}(x, y, \beta)d\mu(x)d\mu(y)
\]

\[
= E\left\{n_0H(0, X, Y, \alpha)/\pi_0 + n_1H(1, X, Y, \alpha)/(1 - \pi_0)\right\}.
\]

\[
(7)
\]

Moreover, \(\pi_0 = \text{pr}^{\text{true}}(D=0)\) is the unique solution to equation (7) if \(\text{pr}\{H(0, X, Y, \alpha) > 0\} > 0\).

Hence, we can obtain a consistent estimator of \(\pi_0\) through solving

\[
\pi_0 = \sum_{i=1}^{n} \frac{H(0, X_i, Y_i, \alpha)}{n_0H(0, X_i, Y_i, \alpha)/\pi_0 + n_1H(1, X_i, Y_i, \alpha)/(1 - \pi_0)}.
\]

Denote the resulting estimator by \(\hat{\pi}_0\). We then estimate \(\pi_1\) by \(\hat{\pi}_1 = 1 - \hat{\pi}_0\).

4. Algorithm

Some other quantities that are involved in the estimating equation (6) depend on the unknown distributions of \(X\) and \(Y|X\), i.e. \(\eta_1\) and \(\eta_2\). All those quantities can be estimated non-parametrically if we know the exact form of the score function in the underlying source population, i.e. \(S\), which is defined in equation (4). Unfortunately, \(S\) itself also relies on the unknown probability density function \(\eta_2\). Here, we propose an algorithm based on a posited score function \(S^*\). It is obtained by replacing \(\eta_2\) in equation (4) with an arbitrary density function \(\eta_2^*\), whose \(\tau\)th quantile is 0. The resulting estimator of \(\theta\) is

a. consistent and asymptotically normal, and
b. it is further efficient if the posited density function is the truth.
We call an estimator with properties (a) and (b) locally efficient throughout the text. Here is the detailed algorithm.

**Step 1:** posit a model \( n_2^*(e, x) \) for \( n_2(e, x) \) which has \( r \) th quantile 0, and calculate equation (4), calling the result \( S^* \).

**Step 2:** solve \( \hat{\pi}_0 = \sum_1^n \{ H(0, X_i, Y_i) / \hat{\pi}_0 + n H(1, X_i, Y_i) / (1 - \hat{\pi}_0) \}^{-1} \) to obtain \( \hat{\pi}_0 \).

**Step 3:** set \( \hat{\pi}_1 = 1 - \hat{\pi}_0 \) and

\[
\hat{\pi}_1 = \hat{\pi}(X, Y) = \{ \sum_d n_d H(d, X_i, Y_i) / (n \hat{\pi}_d) \}^{-1},
\]

\[
\tilde{f}_0 = \tilde{f}_{D | X, Y}(0, X_i, Y_i) = n_0 H(0, X_i, Y_i) \hat{\pi}_0 / (n \hat{\pi}_0).
\]

\[
\tilde{f}_{11} = \tilde{f}_{D | X, Y}(1, X_i, Y_i) = n_1 H(1, X_i, Y_i) \hat{\pi}_0 / (n \hat{\pi}_0).
\]

\[
\hat{\mu}_{sl} = \hat{E}(S^* | e_{x, r} X) = \sum_d n_d H(d, X_i, Y_i) S^*(d, X_i, Y_i) \hat{\pi}_d / (n \hat{\pi}_d).
\]

\[
\hat{b}_0 = \sum_{i=1}^n \tilde{f}_{i} 0_i / \sum_{i=1}^n \tilde{f}_{i} 0_i.
\]

\[
\hat{b}_1 = \sum_{i=1}^n \tilde{f}_{i} 1_i / \sum_{i=1}^n \tilde{f}_{i} 1_i.
\]

\[
\hat{\epsilon}_0 = \sum_{i=1}^n \{ S^*(0, X_i, Y_i) - \hat{\mu}_{sl} \} \hat{f}_i 0_i / \sum_{i=1}^n \hat{f}_i 0_i.
\]

\[
\hat{\epsilon}_1 = \sum_{i=1}^n \{ S^*(1, X_i, Y_i) - \hat{\mu}_{sl} \} \hat{f}_i 1_i / \sum_{i=1}^n \hat{f}_i 1_i.
\]

**Step 4:** perform a non-conventional weighted version of non-parametric kernel regression to form

\[
\hat{E}_{true}(u_{x}^{2} | x) = \frac{\sum_d \hat{\pi}_d | d \sum_{i=1}^n I(D_i = d) u_{x}^{2} \hat{f}_{i} K_H(X_i - x)}{\sum_d \hat{\pi}_d | d \sum_{i=1}^n I(D_i = d) K_H(X_i - x)}
\]

and \( \hat{f}_1(x) = \hat{E}_{true}(u_{x}^{2} | x)^{-1} \).

**Step 5:**

a. perform non-parametric regression using the data \( (X, u_{x}, \hat{\mu}_{sl}) \) with \( D = 0 \) to obtain \( \hat{E}(u_{x} \hat{\mu}_{sl} | x, D = 0) \); similarly, perform non-parametric regression using the data \( (X, u_{x}, \hat{\mu}_{sl}) \) with \( D = 1 \) to obtain \( \hat{E}(u_{x} \hat{\mu}_{sl} | x, D = 1) \);

b. form \( \hat{E}_{true}(u_{x} \hat{\mu}_{sl} | x) = \sum_d \hat{\pi}_d \hat{E}(u_{x} \hat{\mu}_{sl} | x, D = d) \hat{I}_{X|x}(x, d) / \hat{I}_{X|D}(x, d) \).

**Step 6:**
Step 7:

a. perform non-parametric regression using the data \((X_i, u_{t,i}, \hat{f}_i)\) with \(D_i=0\) to obtain \(\hat{E}(u_{t,i}, \hat{f}_i | x; 0)\); similarly, perform non-parametric regression using the data \((X_i, u_{t,i}, \hat{f}_i)\) with \(D_i=1\) to obtain \(\hat{E}(u_{t,i}, \hat{f}_i | x; 1)\);

b. form \(\hat{E}_{true}(u_{t,i}, \hat{f}_i | x) = \Sigma_d \hat{p}_d \hat{E}(u_{t,i}, \hat{f}_i | x; d) \hat{I}_M(x; d)/\Sigma_d \hat{p}_d \hat{I}_M(x; d)\).

Step 8:

a. form \(\hat{I}_2(x) = \hat{E}_{true}(u_{t,i}, \hat{f}_i, \hat{f}_0 | x) - (\hat{E}_0/\hat{h}_0)\hat{E}_{true}(u_{t,i}, \hat{f}_0 | x)\) and \(\hat{\gamma}_2(x) = -\hat{b}_0^{-1} \hat{E}_{true}(u_{t,i}, \hat{f}_0 | x)\);

b. form \(\hat{E}(u_{t,i}, \hat{f}_1(x), \hat{f}_0(x), \hat{y}, \hat{f}_0, y; D=0) = \Sigma_{i=1}^n u_{t,i} \hat{f}_1(X_i) \hat{f}_0(Y_i) \hat{f}_0/\Sigma_{i=1}^n \hat{f}_0; \hat{E}(u_{t,i}, \hat{f}_1(x), \hat{f}_0(x), \hat{y}, \hat{f}_0, y; D=0) = \Sigma_{i=1}^n u_{t,i} \hat{f}_1(X_i) \hat{f}_0(Y_i) \hat{f}_0/\Sigma_{i=1}^n \hat{f}_0; \hat{E}(u_{t,i}, \hat{f}_1(x), \hat{f}_0(x), \hat{f}, y; D=0)\)

and \(\hat{u}_0 = [1 - \hat{E}(u_{t,i} t_1(x) t_2(x) y; D=0)]^{-1} \hat{E}(u_{t,i} t_1(x) t_2(x) y; D=0)\);

c. form \(\hat{u} = (n_0/n_1) \hat{u}_0, \hat{v}_0 = (\hat{p}_0/\hat{h}_0)(\hat{u}_0 + \hat{u}_0)\) and \(\hat{v}_1 = - (\hat{p}_0/\hat{h}_0)(\hat{u}_0 + \hat{u}_0)\);

d. form \(\hat{a}(x) = \hat{t}_1(x) (\hat{I}_2(x) + \hat{\gamma}_2(x) \hat{u}_0)\);

e. form \(\hat{g}(x) = \hat{\mu}_{\hat{a}, \hat{u}_0, \hat{v}_0, \hat{v}_1}(X_i) X_i - \hat{v}_0 \hat{f}_0 - \hat{v}_1 \hat{f}_0 \hat{f}_0\);

f. form \(\hat{S}_{eff}(D_i, X_i, Y_i, S) = S - \hat{g}_i - \hat{v}_{D_i}\) and solve the corresponding estimating equation.

This algorithm is designed for continuous \(X\). When \(X\) is discrete, simply replace the various non-parametric regressions with the corresponding averages. In step 4, \(K(y) = K(-y/h)\), where \(h\) is a bandwidth and \(K(\cdot)\) is a kernel function. Non-parametric kernel regression is also needed in steps 5(a), 6(a) and 7(a). These steps are the only steps that involve a bandwidth \(h\), which is a tuning parameter that needs to be selected. Theoretical requirements and empirical methods to select \(h\) are provided in regularity condition 4 in Section 5 and in the practical guidelines in Section 6.

5. Asymptotics

The asymptotic distribution of the estimator proposed is given in theorem 1 below, with proof provided in Appendix A.3. We assume the following regularity conditions.

**Condition 1**—There is a constant \(0 < C < \infty\) such that \(\lim_{h \to 0} n_0/n_1 = C\).

**Condition 2**—The univariate kernel function is a probability density function with support \((-1, 1)\) and order \(r\), i.e. \(\int K(x) x^r dx = 0\) if \(1 \leq r < \infty\). The \(d\)-dimensional
kernel function, still represented with $K$, is a product of $d$ univariate kernel functions, i.e.

$$K(x) = \prod_{i=1}^{d} K(x_i)$$

for a $d$-dimensional $x$.

**Condition 3**—For $d = 1, 0$, $f_{X|D}(x|D=d)$, $E(u_{\tau}^{\gamma} | X, D = d)$, $E(u_{\tau} \beta_{1} | X, D=d)$ and $E(u_{\tau} \beta_{2} | X, D=d)$ have compact support and have continuous $r$th derivatives.

**Condition 4**—The bandwidth $h=n^{-w}$ where $1/(2d)>w>1/(4r)$, where $d$ is the dimension of $x$. This allows the optimal bandwidth $h=O(n^{-1/(2r+d)})$ as long as we choose a kernel of order $2r>d$.

Conditions 1–4 are typical assumptions to ensure the consistency of the non-parametric kernel estimators that were built in Section 4 and the subsequent semiparametric estimator of $\theta$. Specifically, condition 1 is a general assumption in all case–control studies of the type that we are considering. It ensures that the number of cases and controls is proportional in the case–control sample. Consequently, the total sample size $n$ approaching $\infty$ implies that both $n_0$ and $n_1$ approach $\infty$ at the same rate. Conditions 2 and 4 are standard requirements on kernel function $K$ and bandwidth $h$. Condition 3 is the smoothness assumption on the functions that are needed to be estimated non-parametrically.

**Theorem 1**—Under the regularity conditions 1–4 that are listed above, the estimator $\hat{\theta}$ that is obtained from solving the estimating equation

$$\sum_{i=1}^{n} S_{\text{eff}}^{*}(D, X, Y, \theta) = 0$$

satisfies

$$n^{1/2}(\hat{\theta} - \theta) \rightarrow N(0, A^{-1}B(A^{-1})^T)$$

when $n \rightarrow \infty$, where $A = E(\partial S_{\text{eff}}^{*}(D, X, Y, \theta)/\partial \theta^T)$ and $B = \text{cov}(S_{\text{eff}}^{*}(D, X, Y, \theta))$. If $\eta_{2}^{*} = \eta_{2}$, then $A=-B$ and the estimator is semiparametrically efficient.

The results in theorem 1 ensure the efficiency of $\hat{\theta}$ only when $\eta_{2}^{*} = \eta_{2}$. If $\eta_{2}^{*} \neq \eta_{2}$, $\hat{\theta}$ is only guaranteed to be a root-$n$-consistent estimator. Thus, it is of great value if $\eta_{2}^{*} = \eta_{2}$ indeed. Unfortunately, because $\eta_{2}^{*}$ is a purely posited model, there is no way to ensure that it is $\eta_{2}$. In practice, a feasible approach to obtain an $\eta_{2}^{*}$ that is close to $\eta_{2}$ is to perform a standard quantile regression first based on the control only data to assess $\eta_{2}$, and then to use this assessment to aid in positing a suitable model $\eta_{2}^{*}$.

### 6. Simulation study

**6.1. Overview**

In this section, we study the finite sample performance of our estimator via various simulations. We considered two different types of secondary models:

a. a location–scale model and
b. a more general model.

We describe these studies in Sections 6.2 and 6.3 respectively.

We compared five estimators:

a. our semiparametric approach with a correctly specified posited model for $\eta_2$;
b. our semiparametric approach with the misspecified posited model for $\eta_2$;
c. ordinary quantile regression using only controls, which has negligible bias when the rate of disease is rare but is not efficient because it does not use all the data;
d. SICO with a correctly specified rate of disease in the source population;
e. SICO with an incorrectly specified rate of disease in the source population.

6.2. Location–scale secondary model

We consider the location–scale quantile regression model

$$Y = 0.5 + X + (1 + 0.2X)e^\gamma$$

$$= 0.5 + F^{-1}(\tau) + [1 + 0.2F^{-1}(\tau)]X + (1 + 0.2X)(e^\gamma - F^{-1}(\tau)).$$

where $X$~uniform(0, 1) and $e^\gamma$~$F$. Here we set $F$ to be the standard normal distribution and the standardized gamma distribution with shape parameter $k=2.4$. In the location–scale model, the conditional quantile of $Y$ is linear in $X$ at any quantile level. For a given $\tau$, the quantile regression coefficients are

$$\beta_{\tau, c} = 0.5 + F^{-1}(\tau),$$

$$\beta_{\tau} = 1 + 0.2F^{-1}(\tau),$$

and the quantile regression error is $e_\tau = (1+0.2X)(e^\gamma - F^{-1}(\tau))$. A similar model was considered in Wei et al. (2016). We set the logistic regression model to be $pr(D=1 | Y, X) = H(a_c + X + 0.5 Y)$, where $a_c$ is chosen to achieve a 4.5% rate of disease.

For each setting described above, we simulated 1000 data sets, each with 1000 cases and 1000 controls, and applied the algorithm that was discussed in Section 4 for $\tau=0.1, 0.25, 0.5, 0.75, 0.9$ with two different posited models for $\eta_2$. The first posited model is the true conditional density of $Y$ given $X$, which is the oracle case and serves as a benchmark. Specifically, we set $\eta_2(e_{\tau}, x)$ to be $N(- (1 + 0.2X)\Phi^{-1}(\tau), (1 + 0.2X)^2)$, where $\Phi$ is the distribution function of the standard normal distribution, when $F$ is the standard normal distribution. We set

$$\eta_2(e_{\tau}, x) = \frac{k^{1/2}}{(1 + 0.2x)^{\frac{k}{2}}(k)} \left\{ \frac{t}{1 + 0.2x} + F^{-1}(\tau) \right\}^{k-1} \exp \left\{ -k^{1/2} \left\{ \frac{t}{1 + 0.2x} + F^{-1}(\tau) \right\} \right\}.$$
when $F$ is the standardized gamma distribution. The second posited model is simply $N[\Phi^{-1}(\tau), 1]$. Clearly, such a posited model for $\eta_2$ has the $\tau$th quantile 0, and the second element in $S^*$ has a simple and clear form: $(y-\beta_c x e^{-\beta_c x +\Phi^{-1}(\tau)})(1, x)^T$. It is a misspecified model because both the normal and the gamma cases for $F$ are actually heteroscedastic. Theoretically, any arbitrary posited model will produce a consistent semiparametric estimator, whereas the resulting estimator will be more efficient if the posited model is closer to the truth.

To solve the estimating equation that was described in Section 4, we minimize its sum of squares by using the Nelder–Mead method (Nelder and Mead, 1965). The same method is used in estimating $\pi_0$. Here, we did not use methods such as the Newton–Raphson method to solve the estimating equations because of the non-smoothness that is caused by the nature of quantile regression. When performing the non-parametric regressions, we use a bandwidth $h = cn_0^{-1/3}$, where $c$ is a constant and $n_0$ is the number of controls. We have tested the performance of our estimator by using various values of $c$ between 0.5 and 1.5 and found that the results were similar. Here we report only the results when $c=1$. When studying variants of SICO, we used 100 replications of the method of Wei et al. (2016), and, since SICO requires a rate of disease and the rate of disease in the source population is often unknown, we experimented in both cases where this rate is correctly (4.5%) and incorrectly (1%) specified.

The results are summarized in Table 1, with the mean estimates over the simulation, standard deviation of the estimates and mean-squared efficiency compared with the controls only analysis. We display only the results for the slope $\beta_c$. The results for the intercept are similar, but the mean-squared error comparisons are lower. First, the two SICO estimates are not much better than the controls only analysis, neither of which achieved more than a 35% gain in mean-squared error efficiency. Second, our semiparametric methods dominate all three competitors, with up to a 100% gain in efficiency compared with the controls only analysis. This is striking because both our semiparametric estimates also estimate the rate of disease and do not treat it as known. Third, there is only a slight gain in efficiency when the posited model for $\eta_2$ is correct, indicating robustness of our method, hence alleviating the need for extensive effort to construct a correct working model.

Further to perform inference using our estimators, we can use the bootstrap to assess the variability of our estimator, which is a typical approach in quantile regression problems. In Table 2, we summarize the average of the estimated standard deviations of our estimators based on 300 bootstrap samples, together with the 95% confidence intervals. We can see that, except for a few cases, the results are reasonably accurate in that the bootstrap estimates are reasonably close to the sample standard deviations, whereas the coverages are close to the nominal levels. Large bootstrap sample sizes can certainly be used. We further performed hypothesis testing based on our results and the bootstrap standard deviation and the type I error rate and power as a function of the distance between $\beta_c$ under $H_1$ and $H_0$ are given in Fig. 1.
6.3. Non-location–scale secondary model

We next examine the performance of the five methods that were described in Section 6.1 in non-location–scale secondary models. Systematic linear quantile models that are assumed in the SICO estimation no longer hold here. Specifically, we considered the following simulation settings. First, we set the secondary model to be

\[ Y = 0.5 + X + 0.5(1 + X^2)^{3/4} \{e^\tau - F^{-1}(\tau)\}, \]

where \( X \sim \text{uniform}(0, 1) \) and \( e^\tau \sim F \). Again, \( F \) is either the standard normal distribution \( \mathcal{N}(0, 1) \), or the standardized gamma distribution with shape parameter 2.4. By design, the conditional quantile function of \( Y \) is linear in \( X \) only at the target quantile level \( \tau \) with \( (\beta_{\tau,c}, c, \beta_{\tau}) = (0.5, 1) \). In our simulations, we consider three target quantiles: \( \tau = 0.25, 0.5, 0.75 \).

Second, we consider a relatively rare disease rate of 4.5\%, a rare disease rate of 1.0\% and a common disease rate of 10.0\%. Third, the logistic regression model was \( \text{pr}(D = 1 | Y, X) = H(a_c + a_1 X + a_2 Y) \), where \( a_1 = 1 \) and \( a_2 = 0.50 \). The intercept \( a_c \) is chosen to achieve the specific rates of disease given above. All the other settings are identical to those in Section 6.2.

The results are presented in Tables 3–5. The broad conclusions from these simulations are much the same as those given in Section 6.2. First, our semiparametric methods uniformly dominate the controls only analysis and both SICO analyses, and this is quite noticeable for estimating the slope. Second, when the rate of disease is 10\%, the controls only analysis and SICO when setting its required disease rate to 1.0\% have some modest biases. Third, there is only a slight gain in efficiency when the model posited for \( \eta_2 \) is correct, once again indicating robustness of our method, hence alleviating the need for extensive effort to construct a correct working model.

7. Data analysis

The consumption of red meat is known to be positively associated with colorectal cancer. Although red meat can provide necessary nutrition such as protein, vitamins and minerals, it can also produce MeIQx, which is a carcinogenic heterocyclic amine, if cooked at high temperatures for a long duration. We thus analyse the median relationship between MeIQx and red meat consumption.

We apply our semiparametric approach to a case–control sample of 640 cases with colorectal adenoma and 665 healthy controls. The data were sampled from a larger population-based cohort study: the prostate, lung, colorectal and ovarian cancer screening trial. This cohort study recruited a total of 33971 participants, among whom 10\% developed at least one histologically verified colorectal adenoma (Peters et al., 2003).

In our analysis, we set the covariate \( X \) to be red meat consumption in grams and set the secondary outcome \( Y \) to be the MeIQx that is produced during the cooking of red meat. Because both \( X \) and \( Y \) are heavily skewed in their original measurement scales, we transformed \( X \) and \( Y \) by first adding 1.0 and taking logarithms, and then dividing by their...
respective standard deviations. We fitted models (1) and (2) to this case–control data set with \( \tau = 0.5 \) by using three approaches, i.e. the controls only approach, SICO and our semiparametric approach that was described in Section 4. Although the true rate of disease \( \pi_1 \) is known in this example, we performed a sensitivity analysis of SICO assuming rates of 1.0%, 4.5% and 10.0%, where 10% is the true rate. Our locally efficient semiparametric approach is applied under the posited model \( N(-\Phi^{-1}(\tau), 1) \) and estimates the rate of disease.

The result is summarized in Table 6. All three approaches showed a positive association between red meat consumption and MeIQx or, equivalently, higher consumption of red meat leads to higher exposure to MeIQx. We define efficiency as the ratio of variances relative to the controls only estimator. As in Sections 6.2 and 6.3, once again our semiparametric estimator dominates in terms of variance. For the semiparametric method, we further performed a bootstrap analysis to evaluate the estimation variability and the results for \( \beta_0 \) and \( \beta_X \) are respectively 0.175 and 0.034, which are reasonably close to their sample standard deviations.

8. Discussion

Quantile regression is in wide use. We considered quantile regression in the secondary analysis of case–control studies, making minimal model assumptions. We specify only a linear relationship between covariates at a given quantile in the secondary model, whereas the covariate distribution is modelled completely non-parametrically. We showed that, despite these weak assumptions, the problem is identifiable excluding a few cases. We developed a class of consistent semiparametric estimators and identified the most efficient member. The superiority of our semiparametric estimator over the only other existing estimator in the literature, SICO, was demonstrated both theoretically and numerically, although SICO makes much stronger assumptions.

In the case that all quantile levels are of interest, our method is also applicable. For example, consider model (2) to be true for all \( \tau \), where we assume that the parameter, which is written \( \beta(\tau) \), is a smooth function of \( \tau \). We can first select a set of distinct quantile levels, say \( 0 < \tau_1 < \ldots < \tau_K < 1 \), and estimate each \( \beta_{\tau_j} \) by using the approach that was discussed in Section 4, and then smooth over those estimates to estimate the function \( \beta(\tau) \).

The implementation of the algorithm that was discussed in Section 4 involves several non-parametric regressions, which may be subject to the curse of dimensionality when the dimension of the covariates increases. One possible future work would be to employ dimension reduction techniques such as single-index models to limit this problem.

Although we have described our model and method in the linear quantile regression framework, the extension to the non-linear case is obvious and straightforward. A more interesting extension is to improve the efficiency of our locally efficient semiparametric estimator further by imposing certain parametric structures on the regression error \( \epsilon_\tau \), say \( \epsilon_\tau = \exp(X^T \zeta)\epsilon^{*}_\tau \), where \( \epsilon^{*}_\tau \) has \( \tau \)th quantile 0. The general methodology of our approach can easily be extended here, but asymptotic properties would need to be re-established.
An interesting extension of the work here is to consider several quantiles $\tau_k, k = 1, \ldots, K$, simultaneously. For this, we can form a set of estimating equations for each $\tau_k$, and then combine them and solve for all the parameters, invoking the generalized method of moments, if necessary. In case we need to test whether a subvector of the parameters is 0, we recommend the use of a similar procedure to that in the simulation, using the bootstrap to assess the estimation variance.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

Ma’s research was partially supported by National Science Foundation grant DMS-1608540. Liang and Carroll’s research was supported by a grant from the National Cancer Institute (U01-CA057030). Wei’s research was supported by a National Institutes of Health grant (R01 HG008980-01).

**References**


Appendix A: Sketch of technical arguments

A.1. Nuisance tangent space $\Lambda$ and its orthogonal complement $\Lambda^\perp$

The nuisance tangent space $\Lambda$ with nuisance parameter $\eta = (\eta_1, \eta_2)$ is a subspace of the Hilbert space $\mathcal{H}$. It is defined as the mean-squared closure of parametric submodel nuisance tangent spaces, where a parametric submodel nuisance tangent space with a finite dimensional parameter $\gamma = (\gamma_1^T, \gamma_2^T)^T$ is the set of all elements of the form $\mathbf{W} \mathbf{S}_\gamma$. Here $\mathbf{W}$ is an arbitrary $p \times q$ matrix with $p = \text{dim}(\Theta)$ and $q = \text{dim}(\gamma)$, and $\mathbf{S}_\gamma$ is the nuisance score function, i.e. $\mathbf{S}_\gamma = (\mathbf{S}_{\gamma 1}, \mathbf{S}_{\gamma 2})^T$, with:

$$\begin{align*}
\mathbf{S}_{\gamma 1} &= \eta_1(x; \gamma_1)^{-1} \partial \eta_1(x; \gamma_1) / \partial \gamma_1 - \mathbf{\pi}_d^{-1} E[\eta_1(x; \gamma_1)^{-1} \partial \eta_1(x; \gamma_1) / \partial \gamma_1 | d], \\
\mathbf{S}_{\gamma 2} &= \eta_2(x, x; \gamma_2)^{-1} \partial \eta_2(x, x; \gamma_2) / \partial \gamma_2 - \mathbf{\pi}_d^{-1} E[\eta_2(x, x; \gamma_2)^{-1} \partial \eta_2(x, x; \gamma_2) / \partial \gamma_2 | d].
\end{align*}$$

$\eta_1(\cdot)$ is an arbitrary density function, i.e. $\eta_1 \geq 0$ and $\int \eta_1(x) \, d\mu(x) = 1$, whereas $\eta_2(\cdot, x)$ is an arbitrary density function with $\tau$th quantile 0, i.e. $\eta_2 \geq 0$, $\int \eta_2(x, x) \, d\mu(x) = 1$ and $\int u(x) \eta_2(x, x) \, d\mu(x) = 0$. It is easy to show that the nuisance tangent space $\eta = (\eta_1, \eta_2)$ can be written as $\Lambda = \Lambda_1 \oplus \Lambda_2$, where

$$\begin{align*}
\Lambda_1 &= \{g(x) - E[g(X) | d] : g(x) \in \mathcal{R}^p, E_{\text{true}}[g(X)] = 0\}, \\
\Lambda_2 &= \{g(x, x) - E[g(x, x) | d] : g(x, x) \in \mathcal{R}^p, E_{\text{true}}[g(x, x) | x] = 0, E_{\text{true}}[u(x) g(x, x) | x] = 0, \text{almost surely}\}. 
\end{align*}$$

We can further write

$$\Lambda = \{g(x, x) - E[g(x, x) | d] : g \in \mathcal{R}^p, E_{\text{true}}[g] = 0, E_{\text{true}}[u(x) g(x, x) | x] = 0, \text{almost surely}\}.$$

For any $h \in \mathcal{R}^p$, if $h \in \Lambda^\perp_1$, then

$$0 = E(h^T [g(X) - E(g(X) | D)]) = E((h - E(h | D))^T [g(X) - E(g(X) | D)]) = E((h - E(h | D))^T g(X)) = E(E(h - E(h | D) | X)^T g(X)).$$

Therefore, $E(h - E(h | D | X) \Sigma_{d \delta} \int f_{X, Y, D} \, d\mu(y)/\eta_1(x) = c$ almost surely, for some constant $c$. Note that $E[h - E(h | D | X)] = 0$; we further obtain...
\[ 0 = \int E(h - E(h \mid D) \mid x) \sum_d \int f_{X,Y,D}(x,y,d) d\mu(y) d\mu(x) = \int c_1(x) d\mu(x) = c. \]

Hence, \( c = 0 \) and \( E(h - E(h \mid D) \mid \epsilon, X) = 0 \) almost surely.

For any \( h \in \Lambda_1^\perp \), if \( h \in \Lambda_2^\perp \), then

\[ 0 = E(h^T \{ g(\epsilon, X) - E(g(\epsilon, X) \mid D) \}) = E((h - E(h \mid D))^T \{ g(\epsilon, X) - E(g(\epsilon, X) \mid D) \}) = E(E(h - E(h \mid D) \mid \epsilon, X)^T g(\epsilon, X)). \]

Consequently, \( E(h - E(h \mid D) \mid \epsilon, X) \sum_d f_{X,Y,D}(X, Y, d) \eta_1(X) \eta_2(\epsilon, X) = u_\epsilon a(X) + c(X) \) almost surely. Since \( h \in \Lambda_1^\perp \), we have \( 0 = E(h - E(h \mid D) \mid \epsilon, X) = E(E(h - E(h \mid D) \mid \epsilon, X) \mid X) \), almost surely. Thus,

\[ 0 = \int E(h - E(h \mid D) \mid \epsilon, X) \sum_d f_{X,Y,D}(X, Y, d) \]

\[ = \int \{ u_\epsilon a(X) + c(X) \} \eta_1(X) \eta_2(\epsilon, X) d\mu(y) \]

\[ = \frac{\int c(X) \eta_1(X) d\mu(y)}{\int \sum_d f_{X,Y,D}(X, Y, d) d\mu(y)} \]

which implies that \( c(X) = 0 \) almost surely.

As a result, \( E(h - E(h \mid D) \mid \epsilon, X) \sum_d f_{X,Y,D}(X, Y, d) \eta_1(X) \eta_2(\epsilon, X) = u_\epsilon a(X) \) almost surely or, equivalently, \( E(h - E(h \mid D) \mid \epsilon, X) \sum_d n_d H(d, X) \eta_1(X) \eta_2(\epsilon, X) = e\epsilon a(X) \) almost surely.

Hence,

\[ \Lambda^\perp = \Lambda_1^\perp \cap \Lambda_2^\perp = \left\{ h(d, \epsilon, x) : E(h) = 0, E(h - E(h \mid D) \mid \epsilon, x) \sum_d \frac{n_d H(d, x, \alpha)}{n\alpha_d} = a(x) u_\epsilon, \text{ almost surely } \forall a \right\}. \]
A.2. Efficient score function $S_{\text{eff}}$

We now derive the efficient score $S_{\text{eff}}$ through orthogonally decomposing $S_{\theta}$ into a function in $\Lambda$ and a function in $\Lambda^\perp$.

We write $S_{\theta} = S - E(S \mid D) = g(e, x) - E(g \mid D) + S_{\text{eff}}$, where $E_{\text{true}}(u \mid g(x)) = 0$. We alternatively write $S_{\text{eff}} = S - g(e, x) - E(S - g \mid D)$, and $S_{\text{eff}}$ satisfies

$$E\{S_{\text{eff}} - E(S_{\text{eff}} \mid D) \mid \epsilon, x\} = \sum_d n_d H(d, x, y) - 1 = a(x) u \tau.$$

and $E(S_{\text{eff}} \mid d) = 0$ automatically; hence we can ignore the second requirement $E(S_{\text{eff}}) = 0$. The property $E(S_{\text{eff}} \mid d) = 0$ also simplifies the first requirement to

$$E(S_{\text{eff}} \mid \epsilon, x) = \sum_d n_d H(d, x, y) - 1 = a(x) u \tau.$$

This gives

$$a(x) u \tau \left( \sum_d n_d H(d, x, y) - 1 \right)^{-1} = E(S - g \mid \epsilon, x) - E\{E(S - g \mid D) \mid \epsilon, x\}.$$

From our model (3), we have

$$f_D \mid X, Y(d, x, y) = \frac{n_d H(d, x, y)}{n_d \pi_d} \left( \sum_d n_d H(d, x, y) - 1 \right)^{-1}.$$

Here, explicitly,

$$\pi_0 = \text{pr}_{\text{true}}(D = 0) = \int \eta_1(x) \eta_2(e, x) H(0, x, y) d\mu(x) d\mu(y),$$

$$\pi_1 = \text{pr}_{\text{true}}(D = 1) = \int \eta_1(x) \eta_2(e, x) H(1, x, y) d\mu(x) d\mu(y).$$

To simplify the notation, in the following calculation we denote
\[ b_0 = E\{f_D \mid X, Y(1, X, Y) \mid D = 0\}, \]
\[ b_1 = E\{f_D \mid X, Y(0, X, Y) \mid D = 1\}, \]
\[ c_0 = E(S \mid D = 0) - E\{E(S \mid \epsilon, X) \mid D = 0\}, \]
\[ c_1 = E(S \mid D = 1) - E\{E(S \mid \epsilon, X) \mid D = 1\}, \]
\[ \kappa(x, y) = \sum_{d = 0}^{1} \frac{1}{n_d} \frac{H(d, x, y) H(\mu_x)}{\pi_d}, \]
\[ u_0 = E\{u\mid a(X)\kappa(X, Y) \mid D = 0\}, \]
\[ u_1 = E\{u\mid a(X)\kappa(X, Y) \mid D = 1\}, \]
\[ v_0 = E(S - g \mid D = 0), \]
\[ v_1 = E(S - g \mid D = 1). \]

Simple calculation shows that \( \pi_0 + \pi_1 = 1, b_0 n_0 = b_1 n_1, c_0 n_0 + c_1 n_1 = 0 \) and \( v_0 n_0 + v_1 n_1 = 0 \).

Under a true model, \( \pi_0, \pi_1, b_0, b_1, c_0 \) and \( c_1 \) are known quantities, whereas \( u_0, u_1, v_0 \) and \( v_1 \) are not known because \( g = g(\epsilon, x) \) and \( a = a(x) \) are not specified. Further, to obtain \( u_0, u_1, v_0 \) and \( v_1 \), we rewrite

\[ u\mid a(X)\kappa(X, Y) = E(S - g \mid \epsilon, X) - v_0 f_D \mid X, Y(0, X, Y) - v_1 f_D \mid X, Y(1, X, Y) \]

as

\[ g(\epsilon, x) = E(S \mid \epsilon, x) - u\mid a(x)\kappa(X, Y) - v_0 f_D \mid X, Y(0, X, Y) - v_1 f_D \mid X, Y(1, X, Y). \quad (8) \]

Since \( v_0 = E(S - g \mid D = 0) \), we obtain

\[ v_0 = E(S \mid D = 0) - E\{E(S \mid \epsilon, X) - u\mid a(X)\kappa(X, Y) - v_0 f_D \mid X, Y(0, X, Y) - v_1 f_D \mid X, Y(1, X, Y) \mid D = 0\} \]
\[ = c_0 + u_0 + v_0 (1 - b_0) + v_1 b_0. \]

Thus, we have \( b_0 v_0 - b_0 v_1 - u_0 = c_0 \). Similarly, from \( v_1 = E(S - g \mid D = 1) \), we obtain
\[v_1 = E(S \mid D = 1) - E(S \mid \epsilon_{\tau} x) - \epsilon_{\tau} a(x) - v_0 f D \mid X, y(0, x, Y) \mid D = 1 \]
\[= \epsilon_1 + u_1 + v_0 b_1 + v_1 (1 - b_1).\]

Thus, we have \(-b_1 v_0 + b_1 v_1 = c_1\).

Since \(E(u_\tau a(x) \kappa(x, Y)) = 0\), we have \(u_0 v_0 + u_1 v_1 = 0, h_0 v_0 = h_1 v_1 = 0\) and \(-b_1 v_0 + b_1 v_1 = c_1\). The last two equations are equivalent so one is redundant. Using these relationships, we can rewrite \(u_1, v_0\) and \(v_1\) as a function of \(u_0\):

\[
\begin{align*}
u_1 &= -\left(\frac{n_0}{n_1}\right) u_0, \\
v_0 &= \left(\frac{\pi_1}{b_0}\right)(u_0 + c_0), \\
v_1 &= -\left(\frac{\pi_0}{b_0}\right)(u_0 + c_0).
\end{align*}
\]

We cannot obtain a more explicit expression for \(u_0\) at this stage, but we can further obtain \(a(x)\) as a function of \(u_0\). Using equation (8) and since \(E_{true}(u_\tau g(x)) = 0\), we have

\[E_{true}[u_\tau E(S \mid \epsilon_{\tau} x) \mid x] - E_{true}[u_\tau^2 a(x, Y) \mid x] a(x) - v_0 E_{true}[u_\tau f D \mid X, y(0, x, Y) \mid x] - v_1 E_{true}[u_\tau f D \mid X, y(1, x, Y) \mid x] = 0.
\]

Hence,

\[
a(x) = E_{true}[u_\tau^2 a(x, Y) \mid x]^{-1} [E_{true}[u_\tau E(S \mid \epsilon_{\tau} x) \mid x] - v_0 E_{true}[u_\tau f D \mid X, y(0, x, Y) \mid x] - v_1 E_{true}[u_\tau f D \mid X, y(1, x, Y) \mid x]]
\]

To simplify the notation further, denote
\[ t_1(x) = E_{\text{true}}\{ u_\tau^2 \kappa(x, Y) \mid x \}^{-1}, \]
\[ t_2(x) = E_{\text{true}}\{ u_\tau E(S \mid e, x) \mid x \} - (\pi_1/b_0)E_{\text{true}}\{ u_\tau f_{D \mid X, Y}(0, x, Y) \mid x \} + (\pi_0/b_0)c_0E_{\text{true}}\{ u_\tau f_{D \mid X, Y}(1, x, Y) \mid x \} \]
\[ t_3(x) = -\frac{\pi_1}{b_0}E_{\text{true}}\{ u_\tau f_{D \mid X, Y}(0, x, Y) \mid x \} + \frac{\pi_0}{b_0}c_0E_{\text{true}}\{ u_\tau f_{D \mid X, Y}(1, x, Y) \mid x \}. \]

(10)

Then
\[ a(x) = t_1(x)(t_2(x) + t_3(x))u_0, \quad (11) \]
and the definition of \( u_0 \) yields
\[ u_0 = E(u_{\tau_1}(x)(t_2(x) + t_3(x))\kappa(x, Y) \mid D = 0) \]
\[ = E(u_{\tau_1}(x)t_2(x)\kappa(x, Y) \mid D = 0) + E(u_{\tau_1}(x)t_3(x)\kappa(x, Y) \mid D = 0)u_0. \]

As a consequence,
\[ u_0 = [1 - E\{ u_\tau t_1(x)t_3(x)\kappa(x, Y) \mid D = 0 \}]^{-1}E\{ u_\tau f_{D \mid X, Y}(x)\kappa(x, Y) \mid D = 0 \}. \]

(12)

Combining these results, we have obtained the analytic form of \( S_{\text{eff}} = S - g - E(S - g \mid D=0) \), where \( g \) is given in equation (8), \( a(x) \) is given in equation (11), \( v_0 \) and \( v_1 \) are given in equation (9), \( u_0 \) is given in equation (12) and the functions \( t_1, t_2 \) and \( t_3 \) are given in equation (10).

\[ \sum_{i=1}^n [S(X_i, Y_i, D_i) - g(Y_i - X_i^T \beta \mid X_i)] - n_0v_0 - n_1v_1 = 0. \]

Besides, we can write the estimating equation as

A.3. Proof of theorem 1

For simplicity of proof, we split the \( n \) observations randomly into two sets. The first set contains \( n-n_1\delta \) observations and the second set contains \( n_1\delta \) observations, where

\[ J R Stat Soc Series B Stat Methodol. Author manuscript; available in PMC 2018 October 16. \]
$0 < \delta < \frac{1}{2} - dw$ is a small positive number. We form and solve the estimating equation by using data in the first set, while calculating all the ‘hatted’ quantities that are described in the algorithm in Section 4 by using data in the second set. We use this only as a technical device, although in our simulations and empirical example we used all the data.

In the algorithm, the approximations involve either replacing expectation with averaging or standard kernel regression estimation; hence the differences between the quantities with hats and without the hat expression have either mean 0 and standard deviation $O(h^{(1-\delta)/2})$, or mean $O(h)$ and standard deviation $O((n^{1-\delta}h^d)^{-1/2})$. In particular, $\hat{S}_{\text{eff}}^* (D_i, X_i, Y_i, \theta_0) - S^*_{\text{eff}} (D_i, X_i, Y_i, \theta_0)$ has bias $O(h)$ and standard deviation $O((n^{1-\delta}h^d)^{-1/2})$.

Thus,

$$0 = (n - n_1 - n - n^{1-\delta})^{-1/2} \sum_{i=1}^{n} \hat{S}_{\text{eff}}^* (D_i, X_i, Y_i, \theta)$$

$$= (n - n^{1-\delta})^{-1/2} \sum_{i=1}^{n} \hat{S}_{\text{eff}}^* (D_i, X_i, Y_i, \theta_0) + (n - n^{1-\delta})^{-1/2} \sum_{i=1}^{n} \{\hat{S}_{\text{eff}}^* (D_i, X_i, Y_i, \theta_0) - \hat{S}_{\text{eff}}^* (D_i, X_i, Y_i, \theta_0)\} + o_p(1).$$

We see that $\hat{S}_{\text{eff}}^* (D_i, X_i, Y_i, \theta_0)$ differs from $S^*_{\text{eff}} (D_i, X_i, Y_i, \theta_0)$ in that all the unknown quantities, except $S^*$, are estimated. This is equivalent to estimating the unknown functions $\eta_1(x)$ and $\eta_2(\varepsilon, x)$ in equation (3) and using the estimates $\hat{\eta}_1(x)$ and $\hat{\eta}_2(\varepsilon, x)$ in calculating $S^*_{\text{eff}}$ from the posited $S^*$. Thus, denoting $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2)$, we can approximate
\[(n-n^1-\delta)^{-1/2}\sum_{i=1}^{n-n^1-\delta} \{\mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0) - \mathbf{S}_{\text{eff}}^*(D_i, X_i, \theta_0)\} \]

\[= (n-n^1-\delta)^{-1/2}\sum_{i=1}^{n-n^1-\delta} \{\mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0) - \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0)\} \]

\[= (n-n^1-\delta)^{-1/2}\sum_{i=1}^{n-n^1-\delta} \partial \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0, \eta_0)/\partial \eta \} (\tilde{\eta} - \eta_0) + O_p \]

\[= (n-n^1-\delta)^{-1/2}\sum_{i=1}^{n-n^1-\delta} \partial \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0, \eta_0)/\partial \eta \} (\tilde{\eta} - \eta_0) + O_p \]

where \(\partial \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0, \eta_0)/\partial \eta \) is a pathwise derivative. However, \(\mathbf{S}_{\text{eff}}^* \) is the projection of \(\mathbf{S}^* \) to \(\Lambda^\perp \) so \(\mathbf{S}_{\text{eff}}^* \in \Lambda^\perp \). Thus, for any parametric submodel of \(\eta \) involving parameter \(\gamma \), we have

\[
\mathbb{E}\left\{\partial \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0, \gamma) / \partial \gamma \right\} = \int \partial \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0, \gamma) / \partial \gamma d\mathbf{X}, Y_i, D_i(x, y, d) \mu(x) \mu(d) \]

\[
= -\int \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0, \gamma) / \partial \gamma d\mathbf{X}, Y_i, D_i(x, y, d) \mu(x) \mu(d) \]

\[
= -\mathbb{E}\{\partial \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0, \gamma) / \partial \gamma \} = 0. \]

The last equality is because by definition \(\mathbf{S}_\gamma \in \Lambda \) which is orthogonal to \(\Lambda^\perp \) and \(\mathbf{S}_{\text{eff}}^* \in \Lambda^\perp \).

Here, \(f_{\mathbf{X}, Y_i, D_i}(x, y, d) \) is defined in equation (3). Because \(\gamma \) is a parameter of any arbitrary submodel of \(\eta \), we actually have obtained

\[
\mathbb{E}\{\partial \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0, \eta_0) / \partial \eta \} = -\mathbb{E}\{\mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0, \eta_0) \mathbf{S}_{\eta}^T \} = 0, \]

where \(\mathbf{S}_\eta \) is the nuisance score function along the arbitrarily chosen specific path of the pathwise derivative. Thus, the first term of equation (13) is of order \(o_p(1)\). In contrast, \(O_p(n-n^1-\delta)^{1/2}(\tilde{\eta} - \eta_0)^2) = O_p(n^{1/2} h^2 + n^{1/2}(n^1-\delta) h^{-1}) = O_p(n^{1/2} h^2 + n^{1/2} d^2) = O_p(1) \). We therefore obtain

\[0 = (n-n^1-\delta)^{-1/2}\sum_{i=1}^{n-n^1-\delta} \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0) + \mathbb{E}\left\{\partial \mathbf{S}_{\text{eff}}^*(D_i, X_i, Y_i, \theta_0) / \partial \theta \right\} (n-n^1-\delta)^{-1/2}(\tilde{\theta} - \theta_0) + o_p(1). \]

This yields \((n-n^1-\delta)^{1/2}(\tilde{\theta} - \theta_0) \rightarrow N(0, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^T)\), and hence
\[ n^{1/2} (\hat{\theta} - \theta_0) \rightarrow N(0, A^{-1}B(A^{-1})^T), \]

when \( n \rightarrow \infty \).

Lastly, we prove that \( A = -B \) when \( \eta^* = \eta \). We write \( S = S - S_{\text{eff}} + S_{\text{eff}} \). Because \( S - S_{\text{eff}} \in A \) and \( S_{\text{eff}} \in A^\perp \), we have \( E[(S - S_{\text{eff}})S_{\text{eff}}^T] = 0 \). Besides, we have

\[
0 = E(S_{\text{eff}}) = \sum_d \int S_{\text{eff}}(D, X, Y; \theta) f_{X, Y, D}(x, y, d; \theta, \eta_1, \eta_2) d\mu(x) d\mu(y).
\]

for any \( \theta \). Taking the derivative of this equation with respect to \( \theta \), we obtain

\[
0 = \sum_d \int \frac{\partial S_{\text{eff}}(D, X, Y; \theta)}{\partial \theta} f_{X, Y, D}(x, y, d; \theta, \eta_1, \eta_2) d\mu(x) d\mu(y)
+ \sum_d \int S_{\text{eff}}(D, X, Y; \theta) \frac{\partial f_{X, Y, D}(x, y, d; \theta, \eta_1, \eta_2)}{\partial \theta} d\mu(x) d\mu(y)
= E[\frac{\partial S_{\text{eff}}(D, X, Y; \theta)}{\partial \theta}] + E[S_{\text{eff}}(D, X, Y; \theta) S_{\text{eff}}(D, X, Y; \theta)]
= E[\frac{\partial S_{\text{eff}}(D, X, Y; \theta)}{\partial \theta}] + E[S_{\text{eff}}(D, X, Y; \theta) S_{\text{eff}}(D, X, Y; \theta)]
= A + B.
\]
Fig. 1.
Power of the test as a function of the distance $d$ between $\beta_{\tau,1}$ and $\beta_{\tau,0}$ for the location-scale model in Section 6.2 (power(0) yields the type I error rate) ($\tau = 0.1$; $\ldots$ $\tau = 0.25$; $\ldots$ $\tau = 0.5$; $\ldots$ $\tau = 0.75$; $\ldots$ $\tau = 0.9$): (a), (b) heteroscedastic normal error case; (c), (d) heteroscedastic gamma error case; (a), (c) misspecified model used; (b), (d) correct model used.
Simulation study for the location–scale model in Section 6.2

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Parameter</th>
<th>Results for the following values of $\tau$:</th>
<th>$\tau = 0.1$</th>
<th>$\tau = 0.25$</th>
<th>$\tau = 0.5$</th>
<th>$\tau = 0.75$</th>
<th>$\tau = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$\beta_{x}$</td>
<td></td>
<td>0.744</td>
<td>0.709</td>
<td>0.717</td>
<td>0.720</td>
<td>0.746</td>
</tr>
<tr>
<td>Truth</td>
<td>$\beta_{x}$</td>
<td></td>
<td>0.865</td>
<td>0.828</td>
<td>0.835</td>
<td>0.848</td>
<td>0.864</td>
</tr>
<tr>
<td>Controls</td>
<td>mean</td>
<td></td>
<td>1.135</td>
<td>1.074</td>
<td>1.086</td>
<td>1.097</td>
<td>1.125</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.209</td>
<td>0.151</td>
<td>0.167</td>
<td>0.209</td>
<td>0.202</td>
</tr>
<tr>
<td>SICO rare</td>
<td>mean</td>
<td></td>
<td>1.256</td>
<td>1.183</td>
<td>1.197</td>
<td>1.209</td>
<td>1.244</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.202</td>
<td>0.146</td>
<td>0.162</td>
<td>0.202</td>
<td>0.202</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td></td>
<td>1.066</td>
<td>1.079</td>
<td>1.097</td>
<td>1.103</td>
<td>1.107</td>
</tr>
<tr>
<td>SICO true</td>
<td>mean</td>
<td></td>
<td>1.244</td>
<td>1.244</td>
<td>1.244</td>
<td>1.293</td>
<td>1.317</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.196</td>
<td>0.142</td>
<td>0.156</td>
<td>0.192</td>
<td>0.192</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td></td>
<td>1.144</td>
<td>1.176</td>
<td>1.250</td>
<td>1.317</td>
<td>1.317</td>
</tr>
<tr>
<td>Semi miss</td>
<td>mean</td>
<td></td>
<td>1.241</td>
<td>1.241</td>
<td>1.241</td>
<td>1.293</td>
<td>1.317</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.189</td>
<td>0.125</td>
<td>0.132</td>
<td>0.164</td>
<td>0.164</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td></td>
<td>1.225</td>
<td>1.360</td>
<td>1.609</td>
<td>1.794</td>
<td>1.806</td>
</tr>
<tr>
<td>Semi corr</td>
<td>mean</td>
<td></td>
<td>1.246</td>
<td>1.246</td>
<td>1.246</td>
<td>1.293</td>
<td>1.317</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.187</td>
<td>0.119</td>
<td>0.126</td>
<td>0.155</td>
<td>0.155</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td></td>
<td>1.255</td>
<td>1.425</td>
<td>1.763</td>
<td>1.998</td>
<td>2.037</td>
</tr>
<tr>
<td>Gamma</td>
<td>$\beta_{x}$</td>
<td></td>
<td>1.578</td>
<td>1.453</td>
<td>1.286</td>
<td>1.117</td>
<td>1.117</td>
</tr>
<tr>
<td>Truth</td>
<td>$\beta_{x}$</td>
<td></td>
<td>1.163</td>
<td>1.219</td>
<td>1.286</td>
<td>1.344</td>
<td>1.481</td>
</tr>
<tr>
<td>Controls</td>
<td>mean</td>
<td></td>
<td>1.268</td>
<td>1.328</td>
<td>1.481</td>
<td>1.548</td>
<td>1.548</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>1.412</td>
<td>1.344</td>
<td>1.481</td>
<td>1.548</td>
<td>1.548</td>
</tr>
<tr>
<td>SICO rare</td>
<td>mean</td>
<td></td>
<td>1.328</td>
<td>1.344</td>
<td>1.481</td>
<td>1.548</td>
<td>1.548</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>1.453</td>
<td>1.548</td>
<td>1.548</td>
<td>1.548</td>
<td>1.548</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td></td>
<td>1.097</td>
<td>1.096</td>
<td>1.089</td>
<td>1.089</td>
<td>1.096</td>
</tr>
<tr>
<td>SICO true</td>
<td>mean</td>
<td></td>
<td>1.578</td>
<td>1.583</td>
<td>1.583</td>
<td>1.583</td>
<td>1.583</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>1.412</td>
<td>1.403</td>
<td>1.403</td>
<td>1.403</td>
<td>1.403</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td></td>
<td>1.117</td>
<td>1.117</td>
<td>1.117</td>
<td>1.117</td>
<td>1.117</td>
</tr>
<tr>
<td>Semi miss</td>
<td>mean</td>
<td></td>
<td>1.548</td>
<td>1.548</td>
<td>1.548</td>
<td>1.548</td>
<td>1.548</td>
</tr>
</tbody>
</table>
Estimator  | Parameter | Results for the following values of $\tau$: | $\tau = 0.1$ | $\tau = 0.25$ | $\tau = 0.5$ | $\tau = 0.75$ | $\tau = 0.9$ |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.084</td>
<td>0.096</td>
<td>0.116</td>
<td>0.153</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.217</td>
<td>1.263</td>
<td>1.546</td>
<td>1.764</td>
<td>1.750</td>
</tr>
<tr>
<td>Semi corr</td>
<td>mean</td>
<td></td>
<td>1.100</td>
<td>1.163</td>
<td>1.265</td>
<td>1.408</td>
<td>1.561</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.084</td>
<td>0.096</td>
<td>0.115</td>
<td>0.150</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.204</td>
<td>1.278</td>
<td>1.592</td>
<td>1.833</td>
<td>1.891</td>
</tr>
</tbody>
</table>

*Displayed are results for the quantile regression slope $\beta_\tau$. The results for the quantile regression intercept are similar but have lower mean-squared error efficiencies. Here $n_1 = 1000$ and $n_0 = 1000$; disease rate 4.5%; 1000 simulated data sets. Top panel: heteroscedastic normal error model. Bottom panel: heteroscedastic gamma error model. The five estimators are ‘Controls’ (analysis using only controls), ‘SICO rare’ (SICO assuming rare disease), ‘SICO true’ (SICO using the true disease rate), ‘Semi miss’ (semiparametric estimator with misspecified posited model for $\eta$) and ‘Semi corr’ (semiparametric estimator with the correct posited model for $\eta$). The sample mean (‘mean’), sample standard deviation (‘sd’) and mean-squared error efficiency across the simulations compared with using only controls (‘MSE Eff’) are given.*
Table 2

Inference results in the simulation study for the location–scale model in Section 6.2 based on 200 bootstrap samples†

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Parameter</th>
<th>Results for the following values of $\tau$:</th>
<th>$\tau = 0.1$</th>
<th>$\tau = 0.25$</th>
<th>$\tau = 0.5$</th>
<th>$\tau = 0.75$</th>
<th>$\tau = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Truth $\beta_s$</td>
<td></td>
<td>0.744</td>
<td>0.865</td>
<td>1.135</td>
<td>1.256</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semi miss mean</td>
<td></td>
<td>0.746</td>
<td>0.864</td>
<td>0.994</td>
<td>1.124</td>
<td>1.241</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.189</td>
<td>0.148</td>
<td>0.125</td>
<td>0.132</td>
<td>0.164</td>
</tr>
<tr>
<td></td>
<td>$\text{sd}$</td>
<td></td>
<td>0.191</td>
<td>0.148</td>
<td>0.131</td>
<td>0.139</td>
<td>0.170</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td></td>
<td>95.1%</td>
<td>94.7%</td>
<td>95.2%</td>
<td>94.8%</td>
<td>95.6%</td>
</tr>
<tr>
<td></td>
<td>Semi corr mean</td>
<td></td>
<td>0.747</td>
<td>0.867</td>
<td>0.996</td>
<td>1.126</td>
<td>1.246</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.187</td>
<td>0.144</td>
<td>0.119</td>
<td>0.126</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td>$\text{sd}$</td>
<td></td>
<td>0.193</td>
<td>0.150</td>
<td>0.133</td>
<td>0.141</td>
<td>0.172</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td></td>
<td>95.3%</td>
<td>95.6%</td>
<td>96.4%</td>
<td>96.3%</td>
<td>97.0%</td>
</tr>
<tr>
<td>Gamma</td>
<td>Truth $\beta_s$</td>
<td></td>
<td>1.097</td>
<td>1.163</td>
<td>1.268</td>
<td>1.412</td>
<td>1.578</td>
</tr>
<tr>
<td></td>
<td>Semi miss mean</td>
<td></td>
<td>1.098</td>
<td>1.159</td>
<td>1.262</td>
<td>1.403</td>
<td>1.548</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.084</td>
<td>0.096</td>
<td>0.116</td>
<td>0.153</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>$\text{sd}$</td>
<td></td>
<td>0.083</td>
<td>0.096</td>
<td>0.120</td>
<td>0.165</td>
<td>0.251</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td></td>
<td>94.1%</td>
<td>94.3%</td>
<td>95.9%</td>
<td>95.7%</td>
<td>95.5%</td>
</tr>
<tr>
<td></td>
<td>Semi corr mean</td>
<td></td>
<td>1.100</td>
<td>1.163</td>
<td>1.265</td>
<td>1.408</td>
<td>1.561</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td></td>
<td>0.084</td>
<td>0.096</td>
<td>0.115</td>
<td>0.150</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td>$\text{sd}$</td>
<td></td>
<td>0.090</td>
<td>0.101</td>
<td>0.127</td>
<td>0.179</td>
<td>0.276</td>
</tr>
<tr>
<td></td>
<td>95%</td>
<td></td>
<td>96.2%</td>
<td>96.2%</td>
<td>97.0%</td>
<td>97.7%</td>
<td>97.1%</td>
</tr>
</tbody>
</table>

†Displayed are results for the quantile regression slope $\beta_s$, $n_1 = 1000$ and $n_0 = 1000$, disease rate 4.5%. 1000 simulated data sets. Top panel: heteroscedastic normal error model. Bottom panel: heteroscedastic gamma error model. The sample mean (‘mean’), sample standard deviation (‘sd’), bootstrap-based estimated standard deviation (‘$\text{sd}$’) and coverage of the 95% confidence interval are given.
Simulation study for the model in Section 6.3

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Parameter</th>
<th>Results for $\tau = 0.25$</th>
<th>Results for $\tau = 0.5$</th>
<th>Results for $\tau = 0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\beta = 0.5$</td>
<td>$\beta = 1.0$</td>
<td>$\beta = 0.5$</td>
</tr>
<tr>
<td>Normal</td>
<td>Controls</td>
<td>mean</td>
<td>0.501</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.063</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>SICO true</td>
<td>mean</td>
<td>0.502</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.062</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.048</td>
<td>1.057</td>
</tr>
<tr>
<td></td>
<td>Semi miss</td>
<td>mean</td>
<td>0.499</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.061</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.190</td>
<td>1.245</td>
</tr>
<tr>
<td></td>
<td>Semi corr</td>
<td>mean</td>
<td>0.501</td>
<td>0.997</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.057</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.221</td>
<td>1.300</td>
</tr>
<tr>
<td>Gamma</td>
<td>Controls</td>
<td>mean</td>
<td>0.501</td>
<td>0.996</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.029</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>SICO true</td>
<td>mean</td>
<td>0.501</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.029</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.037</td>
<td>1.059</td>
</tr>
<tr>
<td></td>
<td>Semi miss</td>
<td>mean</td>
<td>0.497</td>
<td>1.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.028</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.064</td>
<td>1.259</td>
</tr>
<tr>
<td></td>
<td>Semi corr</td>
<td>mean</td>
<td>0.499</td>
<td>1.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.027</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.186</td>
<td>1.338</td>
</tr>
</tbody>
</table>

Here $n_1 = 1000$ and $n_0 = 1000$, disease rate 1%. There were 1000 simulated data sets. Top panel: heteroscedastic normal error model. Bottom panel: heteroscedastic gamma error model. The five estimators are ‘Controls’ (analysis using only controls), ‘SICO rare’ (SICO assuming rare disease), ‘SICO true’ (SICO using the true disease rate), ‘Semi miss’ (semiparametric estimator with misspecified posited
model for $\eta^2$) and 'Semi corr' (semiparametric estimator with correct posited model for $\eta^2$). The sample mean ('mean'), sample standard deviation ('sd') and mean-squared error efficiency across the simulations compared with using only controls ('MSE Eff') are given.
Table 4

Simulation study for the model in Section 6.3

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Parameter</th>
<th>Results for $\tau = 0.25$</th>
<th>Results for $\tau = 0.5$</th>
<th>Results for $\tau = 0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\beta = 0.5$</td>
<td>$\beta = 1.0$</td>
<td>$\beta = 0.5$</td>
</tr>
<tr>
<td>Normal</td>
<td>Controls</td>
<td>mean</td>
<td>0.499</td>
<td>0.976</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.064</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>SICO rare</td>
<td>mean</td>
<td>0.499</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.062</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.043</td>
<td>1.067</td>
</tr>
<tr>
<td></td>
<td>SICO true</td>
<td>mean</td>
<td>0.499</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.061</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.092</td>
<td>1.170</td>
</tr>
<tr>
<td>Semi miss</td>
<td>mean</td>
<td></td>
<td>0.500</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.060</td>
<td>0.106</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.132</td>
<td>1.339</td>
</tr>
<tr>
<td>Semi corr</td>
<td>mean</td>
<td></td>
<td>0.502</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.057</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.228</td>
<td>1.358</td>
</tr>
<tr>
<td>Gamma</td>
<td>Controls</td>
<td>mean</td>
<td>0.502</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.030</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>SICO rare</td>
<td>mean</td>
<td>0.501</td>
<td>0.991</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.030</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.041</td>
<td>1.055</td>
</tr>
<tr>
<td></td>
<td>SICO true</td>
<td>mean</td>
<td>0.501</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.029</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.084</td>
<td>1.146</td>
</tr>
<tr>
<td>Semi miss</td>
<td>mean</td>
<td></td>
<td>0.500</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sd</td>
<td>0.029</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td></td>
<td>1.083</td>
<td>1.305</td>
</tr>
<tr>
<td>Estimator</td>
<td>Parameter</td>
<td>Results for $\tau = 0.25$</td>
<td>Results for $\tau = 0.5$</td>
<td>Results for $\tau = 0.75$</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>----------------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td></td>
<td>$\beta = 0.5$</td>
<td>$\beta = 1.0$</td>
<td>$\beta = 0.5$</td>
<td>$\beta = 1.0$</td>
</tr>
<tr>
<td>Semi corr</td>
<td>mean</td>
<td>0.501</td>
<td>1.001</td>
<td>0.503</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.028</td>
<td>0.052</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td>1.187</td>
<td>1.424</td>
<td>1.290</td>
</tr>
</tbody>
</table>

Here $n_1 = 1000$ and $n_0 = 1000$, disease rate 4.5%. 1000 simulated data sets. Top panel: heteroscedastic normal error model. Bottom panel: heteroscedastic gamma error model. The five estimators are 'Controls' (analysis using only controls), 'SICO rare' (SICO assuming rare disease), 'SICO true' (SICO using the true disease rate), 'Semi miss' (semiparametric estimator with misspecified posited model for $\eta^2$) and 'Semi corr' (semiparametric estimator with correct posited model for $\eta^2$). The sample mean (‘mean’), sample standard deviation (‘sd’) and mean-squared error efficiency across the simulations compared with using only controls (‘MSE Eff’) are given.
Table 5
Simulation study for the model in Section 6.3

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Parameter</th>
<th>Results for $\tau = 0.25$</th>
<th>Results for $\tau = 0.5$</th>
<th>Results for $\tau = 0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta = 0.5$</td>
<td>$\beta = 1.0$</td>
<td>$\beta = 0.5$</td>
<td>$\beta = 1.0$</td>
</tr>
<tr>
<td><strong>Normal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>mean</td>
<td>0.501</td>
<td>0.941</td>
<td>0.494</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.062</td>
<td>0.118</td>
<td>0.058</td>
</tr>
<tr>
<td>SICO rare</td>
<td>mean</td>
<td>0.500</td>
<td>0.947</td>
<td>0.494</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.061</td>
<td>0.115</td>
<td>0.057</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td>1.050</td>
<td>1.084</td>
<td>1.038</td>
</tr>
<tr>
<td>SICO true</td>
<td>mean</td>
<td>0.502</td>
<td>0.992</td>
<td>0.497</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.058</td>
<td>0.106</td>
<td>0.053</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td>1.166</td>
<td>1.538</td>
<td>1.183</td>
</tr>
<tr>
<td>Semi miss</td>
<td>mean</td>
<td>0.506</td>
<td>0.993</td>
<td>0.501</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.056</td>
<td>0.098</td>
<td>0.050</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td>1.249</td>
<td>1.816</td>
<td>1.354</td>
</tr>
<tr>
<td>Semi corr</td>
<td>mean</td>
<td>0.507</td>
<td>0.992</td>
<td>0.501</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.054</td>
<td>0.097</td>
<td>0.048</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td>1.326</td>
<td>1.830</td>
<td>1.477</td>
</tr>
<tr>
<td><strong>Gamma</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>mean</td>
<td>0.502</td>
<td>0.972</td>
<td>0.503</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.029</td>
<td>0.059</td>
<td>0.037</td>
</tr>
<tr>
<td>SICO rare</td>
<td>mean</td>
<td>0.502</td>
<td>0.974</td>
<td>0.502</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.029</td>
<td>0.058</td>
<td>0.037</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td>1.042</td>
<td>1.064</td>
<td>1.037</td>
</tr>
<tr>
<td>SICO true</td>
<td>mean</td>
<td>0.501</td>
<td>0.994</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.028</td>
<td>0.054</td>
<td>0.035</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td>1.133</td>
<td>1.442</td>
<td>1.164</td>
</tr>
<tr>
<td>Semi miss</td>
<td>mean</td>
<td>0.504</td>
<td>0.994</td>
<td>0.506</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.027</td>
<td>0.051</td>
<td>0.033</td>
</tr>
<tr>
<td>MSE Eff</td>
<td></td>
<td>1.152</td>
<td>1.655</td>
<td>1.229</td>
</tr>
<tr>
<td>Estimator</td>
<td>Parameter</td>
<td>Results for $\tau = 0.25$</td>
<td>Results for $\tau = 0.5$</td>
<td>Results for $\tau = 0.75$</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td>$\beta = 0.5$</td>
<td>$\beta = 1.0$</td>
<td>$\beta = 0.5$</td>
<td>$\beta = 1.0$</td>
</tr>
<tr>
<td>Semi corr</td>
<td>mean</td>
<td>0.503</td>
<td>0.998</td>
<td>0.503</td>
</tr>
<tr>
<td></td>
<td>sd</td>
<td>0.028</td>
<td>0.050</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>MSE Eff</td>
<td>1.088</td>
<td>1.719</td>
<td>1.281</td>
</tr>
</tbody>
</table>

Here $q_1 = 1000$ and $q_0 = 1000$, disease rate 10%. 1000 simulated data sets. Top panel: heteroscedastic normal error model. Bottom panel: heteroscedastic gamma error model. The five estimators are ‘Controls’ (analysis using only controls), ‘SICO rare’ (SICO assuming rare disease), ‘SICO true’ (SICO using the true disease rate), ‘Semi miss’ (semiparametric estimator with misspecified posited model for $\eta^2$) and ‘Semi corr’ (semiparametric estimator with correct posited model for $\eta^2$). The sample mean (‘mean’), sample standard deviation (‘sd’) and mean-squared error efficiency across the simulations compared with using only controls (‘MSE Eff’) are given.
Table 6

Results of the secondary analysis of the colorectal adenoma data set in Section 7 across 1000 bootstrap samples†

<table>
<thead>
<tr>
<th>Method</th>
<th>β₀</th>
<th>βₓ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>sd</td>
</tr>
<tr>
<td>Controls</td>
<td>−3.607</td>
<td>0.194</td>
</tr>
<tr>
<td>SICO 1%</td>
<td>−3.605</td>
<td>0.185</td>
</tr>
<tr>
<td>SICO 4.5%</td>
<td>−3.595</td>
<td>0.178</td>
</tr>
<tr>
<td>SICO 10%</td>
<td>−3.583</td>
<td>0.170</td>
</tr>
<tr>
<td>Semi</td>
<td>−3.529</td>
<td>0.159</td>
</tr>
</tbody>
</table>

† ‘Controls’ is the quantile regression on only controls, ‘SICO’ is the simulated counterfactual outcomes approach by Wei et al. (2016) and ‘Semi’ is the locally efficient semiparametric approach. The SICO approach is fitted by using three different disease rates, 1%, 4.5% and 10%, whereas the Controls and Semi approaches are fitted under a posited model \( N(-\Phi^{-1}(\tau), 1) \) without specifying the disease rate. Mean (‘mean’), sample standard deviation (‘sd’) and the square of the ratio of the sample standard deviation compared with the controls only approach (‘Eff’) are reported.