10-28-2003

Smooth Quantile Ratio Estimation

Francesca Dominici
The Johns Hopkins Bloomberg School of Public Health, fdominic@jhsph.edu

Leslie Cope
The Johns Hopkins University, Lcope1@jhmi.edu

Daniel Q. Naiman
The Johns Hopkins University, daniel.naiman@jhu.edu

Scott L. Zeger
The Johns Hopkins Bloomberg School of Public Health, szeger@jhsph.edu

SMOOTH QUANTILE RATIO ESTIMATION

Francesca Dominici, Leslie Cope, Daniel Q. Naiman, and Scott L. Zeger

October 28, 2003

Abstract

In a study of health care expenditures attributable to smoking, we seek to compare the distribution of medical costs for persons with lung cancer or chronic obstructive pulmonary disease (cases) to those without (controls) using a national survey which includes hundreds of cases and thousands of controls. The distribution of costs is highly skewed toward larger values, making estimates of the mean from the smaller sample dependent on a small fraction of the biggest values. One approach to deal with the smaller sample is to rely on a simple parametric model such as the log-normal, but this makes the undesirable assumption that the distribution of the log-expenditures is symmetric.

We propose a novel approach to estimate the mean difference of two highly skewed distributions (Δ), which we call Smooth Quantile Ratio Estimation (SQUARE). SQUARE is obtained by smoothing, over percentiles, the ratio of the cost quantiles of the cases and controls. SQUARE defines a large class of estimators of Δ including: 1) the sample mean difference, 2) the maximum likelihood estimate under log-normal samples, and 3) L-estimates. We detail asymptotic properties of SQUARE such as consistency and asymptotic normality, and also provide a closed form expression for the asymptotic variance.

Through a simulation study, we show that SQUARE has lower mean squared error than several competitors including the sample mean difference, and log-normal parametric estimates in several realistic situations. We apply SQUARE to the 1987 National Medicare Expenditure Survey to estimate the difference in medical expenditures between persons suffering from the smoking attributable diseases, lung cancer and chronic obstructive pulmonary disease, and persons without these diseases. Software in R (Ihaka and Gentleman, 1996) for the implementation of SQUARE and of all its special cases, and the cost data used in this paper are available at `http://biostat.jhsph.edu/~fdominic/square.html`.

KEYWORDS: Comparing means, skewed distributions, order statistics, log-normal, regression splines, Q-Q plots, smoking, health expenditures.

CONTACT INFORMATION: *Francesca Dominici, Scott L. Zeger, Department of Biostatistics at the Bloomberg School of Public Health, Johns Hopkins University. Leslie Cope, Daniel Q. Naiman, Department of Mathematical Sciences at the Johns Hopkins University. Correspondence may be addressed to Dr. Francesca Dominici, Department of Biostatistics, Bloomberg School of Public Health, 615 N. Wolfe Street, The Johns Hopkins University, Baltimore, MD 21205-3179, USA. : 410-614-5107, fax: 410-955-0958, e-mail: fdominic@jhsph.edu.*

ACKNOWLEDGMENTS: Funding for Scott L. Zeger was provided from NIMH grant R01 MH56639. Funding for Francesca Dominici was provided by a grant from the Health Effect Institute (Walter A. Rosenblith New Investigator Award). Funding for Daniel Q. Naiman was provided in part by NSF Grant #DMI-0087032. We thank Timothy Wyant for providing data on the National Medical Expenditures Survey; Mark van der Laan, Giovanni Parmigiani, Michael Griswold, and Tom Louis for comments and suggestions on the paper; and Elizabeth Johnson for assistance in data base development and software.

1 Introduction

This paper is motivated by the question of how to estimate smokers' medical expenditures attributable to their having lung cancer, chronic obstructive pulmonary disease (COPD) or other diseases predominantly caused by smoking. As a component of our analysis, we compare medical expenditures between persons with lung cancer or COPD (cases) and persons without a major smoking attributable disease (controls) in a given year. That is, we seek to estimate the difference $\Delta = E[Y_1] - E[Y_2]$ where Y_1 and Y_2 are random variables representing the expenditures for a case and control groups, respectively. We estimate Δ using the 1987 National Medical Expenditure Survey (National Center For Health Services Research, 1987), one data set on annual medical expenditures and disease status for a representative sample of U.S. non-institutionalized adults.

This statistical problem is made interesting by two facts. First, the distribution of the non-zero medical expenditures is highly skewed to large values. Figure 1 shows histograms of non-zero medical expenditures in NMES with and without a logarithmic transformation. Ninety percent of the total expenditures is contributed by only forty percent of the people. Second, we have a much smaller sample of disease cases than controls. Among persons 40 years and older with non-zero expenditures in NMES, only 118 persons have lung cancer or COPD, while 2262 persons are without a major smoking attributable disease. The problem addressed in this paper is how to reliably estimate the difference in means from two right-skewed distributions given two independent samples, one being substantially smaller than the other.

The problem introduced above is one of a set of problems that arises in studying expenditure data. These include a significant fraction of zero expenditures, right censoring, and lack of independence among observations within clusters (Lipscomb et al., 1999). The general problem of comparing costs among two or more groups is important in econometrics, statistics, and other disciplines (Duan, 1983; O'Brien, 1988; Fenn et al., 1996; Lin et al., 1997; Hlatky et al., 1997; Lin, 2000; Tu and Zhou, 1999).

To motivate our approach, let $y_{11}, \ldots, y_{1n_1}$ and $y_{21}, \ldots, y_{2n_2}$ be the observed non-zero costs in the case and control groups. An obvious estimator of Δ is the difference in sample means $\bar{y_1} - \bar{y}_2$ where $\bar{y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{gi},\ g = 1, 2$. Because we anticipate a highly skewed distribution and one of the samples to be much smaller than the other ($n_1 < n_2$), this unbiased estimator may be more variable than alternatives.

One such alternative is the log-normal model, in which the logarithms of the expenditures are assumed to follow normal distributions: $\log y_{gi} \sim N(\nu_g, \sigma_g^2),\ i = 1, \ldots, n_g,\ g = 1, 2$. Under the log-normal assumption, the difference in mean expenditures for the two populations is by $\Delta = \exp(\nu_1 + \sigma_1^2/2) - \exp(\nu_2 + \sigma_2^2/2)$ (Aitchison and Shen, 1980). The maximum likelihood estimate of Δ is biased (Zellner, 1971), but has reduced variability relative to the sample mean difference because it reduces the degree of dependence on the few largest observations. Zhou et al. (1997) and Zhou and Gao (1997) have studied methods for testing the null hypothesis that $\Delta = 0$ under the log-normal model. Many authors have used the log-normal model for inferences about the mean of a non-zero random variable (Land, 1971; Angus, 1994; Duan et al., 1983; Zhou and Melfi, 1997; Lipscomb et al., 1999; Andersen et al., 2000).

An important limitation of the log-normal model for estimating total or mean costs results from the symmetry inherent in the normal distribution for the logarithms of the expenditures. When the mean expenditure is the scientific focus, the right tail of the distribution contributes most to the mean; the smaller values in the left tail have less influence. Under the symmetry assumption for the log expenditures, we assume that the right and left tails have the same shape on the logarithm scale so that the very smallest expenditures in the sample can be viewed as providing information about the largest ones. In most applications, including the lung cancer and COPD expenditures problem that motivates this work, this symmetry assumption is not based upon any meaningful mechanism and is not likely to be realistic.

One way to view the limitation of the log-normal model to address skewness, is in terms of the quantile-quantile or Q-Q plot. The quantile estimates of the two distributions are plotted against each other (Wilk and Gnanadesikan, 1968; Doksum and Sievers, 1976; Parzen, 1979; Nair, 1982; Wilcox, 1995). Under the log-normal model, the logarithms of the quantiles from each distribution satisfy the linear equation:

$$\log Q_1(p) = \left(\nu_2 - \frac{\sigma_2}{\sigma_1}\nu_1\right) + \frac{\sigma_2}{\sigma_1}\log Q_2(p) \tag{1}$$

where $Q_1(p)$ and $Q_2(p)$ are the quantile functions of the random variables Y_1 and Y_2 representing the non-zero expenditures for the case and control groups. If we use the mean and variance of the log-transformed data to estimate the intercept and slope from the Q-Q plot, then the smallest observations have as much influence on the intercept and slope as the largest observations. This is clearly undesirable when the goal is to estimate the difference in population means, especially when evidence exists in the Q-Q plot against the linearity assumed. Figure 2 displays the Q-Q plot of the log expenditures for the cases versus those for the controls, as well as the straight line corresponding to the maximum likelihood estimates of the log-normal parameters for each sample.

If evidence exists in the Q-Q plot against the linearity assumed under the log-normal model, we might assume that $Q_1(p)$ is an arbitrary function of $Q_2(p)$, that is $Q_1(p) = g\left(Q_2(p)\right)$ or equivalently $F_1(y) = F_2\left(h(y)\right)$ where $F_g(y),\ g = 1, 2$ are the cumulative distribution functions of Y_1 and Y_2. Doksum and Sievers (1976) define $h(.)$ as the amount of "shift" needed to bring Y_1s up to the Y_2s in distribution. For example, we might assume that $Q_1(p)$ is a smooth function of $Q_2(p)$ with λ degrees of freedom, $Q_1(p) = s(Q_2(p), \lambda)$, where s is a parametric or a non-parametric smoother.

There are three possible limitations of the shift model for application to the estimation of Δ. First, one might estimate Q_1 and Q_2 at a given set of percentiles and regress $\widehat{Q}_1$ on $\widehat{Q}_2$. This regression approach leads to conditioning on Q_2 rather than treating the two quantile functions symmetrically, as would be natural when the target for inference is Δ. Second, the smooth function s would take arguments on the positive real line making choice of λ critical. Third, if we then use the fitted values from the smoother to calculate $\hat{\Delta}$, this estimate is simply the difference in the sample means.

As an alternative, we assume that the log-quantile ratio is a smooth function of the percentile p with λ degrees of freedom:

$$\log \frac{Q_1(p)}{Q_2(p)} = s(p, \lambda),\ 0 < p < 1. \tag{2}$$

This is the basic idea of Smooth Quantile Ratio Estimation (SQUARE). Differently from the shift estimator (Doksum and Sievers, 1976), SQUARE "spends" its degrees of freedom λ over the interval (0,1) rather than over the real line, and hence imposes stronger smoothness constraints in the tails where little information is available in our smaller sample. As a result, SQUARE produces an estimator of Δ that tends to be less variable than the difference in sample means. For different distributional assumptions, shapes of $s(p, \lambda)$ and choices of λ, SQUARE encompasses rich class of estimators including the sample mean difference, the maximum likelihood estimate under log-normal samples, and L-estimates. SQUARE borrows strength across neighboring percentiles of the distribution to reduce the variability in the estimated mean difference rather than relying on symmetry assumptions inherent in the log-normal model. In summary, the development of SQUARE is motivated by the needs to: 1) address skewness without making fully parametric assumptions on the underlying distributions for the two samples; 2) treat the two samples symmetrically; 3) and expand the class of estimators to include existing ones and others that may be very efficient in some circumstances.

In section 2, we introduce SQUARE as a semi-parametric method using a parametric model for the log-quantile ratio and non-parametric estimate of the quantile functions. In section 3, we show that, under certain bounding conditions for the log-quantiles, SQUARE is consistent and asymptotically normal. Here we also provide an explicit expression for the asymptotic variance of SQUARE, and examples where our asymptotic results apply. In Section 4, we present a simulation study that compares bias and variance properties of SQUARE with the log-normal maximum likelihood estimate and the sample mean difference. Here we also illustrate SQUARE with an analysis of the NMES data shown in Figure 1. In this section we also illustrate a cross-validation method for estimating the number of degrees of freedom λ. Section 5 is a discussion of opportunities for further development of this idea. Proofs of the asymptotic results are detailed in the Appendix.

2 Smooth Quantile Ratio Estimation (SQUARE)

2.1 Definition

Let Y_1 and Y_2 be two positive random variables. For example in the motivating application, these are the non-zero expenditures for the cases and controls, respectively. We consider the two cumulative distribution functions F_1 and F_2, and define Q_1 and Q_2 to be the corresponding quantile functions so that $Q_g(p) = F_g^{-1}(p)$ and $F_g(Q_g(p)) = \Pr\{Y_g \leq Q_g(p)\} = p,\ g = 1, 2$ and $0 \leq p \leq 1$. Our goal is to estimate the difference:

$$\Delta = E[Y_1] - E[Y_2] = \int_0^1 \{Q_1(p) - Q_2(p)\}dp\ , \tag{3}$$

assuming that the ratio of the quantiles is a smooth function of the percentiles with λ degrees of freedom:

$$\log \frac{Q_1(p)}{Q_2(p)} = s(p, \lambda),\ 0 < p < 1. \tag{4}$$

Then Equations (3) and (4), lead to:

$$\Delta = \int_0^1 Q_1(p)\left[1 - \exp\left(-s(p,\lambda)\right)\right] dp = \int_0^1 Q_2(p)\left[\exp\left(s(p,\lambda)\right) - 1\right] dp. \tag{5}$$

2.2 Estimation Approach

Let $\boldsymbol{y}_1 = (y_{11}, y_{12}, \cdots, y_{1n_1})$ be an iid sample of size n_1 from F_1, and $\boldsymbol{y}_2 = (y_{21}, y_{22}, \cdots, y_{2n_2})$ be an iid sample of size n_2 from F_2. We define $\boldsymbol{y}_{(g)} = (y_{g(1)}, y_{g(2)}, \cdots, y_{g(n_g)})$ to be the order statistics for the sample from F_g. We first estimate Δ for the case $n_1 = n_2 = n$, and then extend our definition to the more common situation $n_1 << n_2$.

The estimation approach can be described in two steps. First, we define a regression model for $s(p,\lambda)$ and we use it to smooth the observed log-ratio $\log(\boldsymbol{y}_{(1)}/\boldsymbol{y}_{(2)})$ across percentiles (parametric part). Second, we estimate Δ by using the smoothed quantile ratios and non-parametric estimates of F_1 and F_2 (non-parametric part). The two steps are detailed below.

Parametric step: we impose a smoothness assumption for $s(p,\lambda)$ by assuming a regression model:

$$\log \frac{y_{1(i)}}{y_{2(i)}} = s(p_i, \boldsymbol{\beta}) + \epsilon_i,\ i = 1, \ldots, n \tag{6}$$

where: $s(p_i, \boldsymbol{\beta}) = \sum_{j=0}^{\lambda} B_j(p_i)\beta_j$, $p_i = i/(n+1)$, and $B_j(p)$ are orthonormal basis functions, with $B_0(p) = 1$. We estimate $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_\lambda)$ by ordinary least squares, although alternative and more efficient methods could be substituted.

Non-parametric step: we define $\boldsymbol{u}_1 = (\boldsymbol{y}_{(1)}, \boldsymbol{y}^\star_{(1)})$ and $\boldsymbol{u}_2 = (\boldsymbol{y}_{(2)}, \boldsymbol{y}^\star_{(2)})$ to be two samples of size $2n$ where $y^\star_{1(i)} = y_{2(i)} \exp\left(s(p_i, \widehat{\boldsymbol{\beta}})\right)$, and $y^\star_{2(i)} = y_{1(i)} \exp\left(-s(p_i, \widehat{\boldsymbol{\beta}})\right)$, and $s(p_i, \widehat{\boldsymbol{\beta}})$ be the fitted values from the regression model (6). We estimate Δ by:

$$\begin{aligned} \widehat{SQ}(\lambda) &= \bar{u}_1 - \bar{u}_2 \\ &= \tfrac{1}{2n}\textstyle\sum_{i=1}^{n} y_{1(i)}\left[1 - \exp\left(-s(p_i, \widehat{\boldsymbol{\beta}})\right)\right] + \tfrac{1}{2n}\sum_{i=1}^{n} y_{2(i)}\left[\exp\left(s(p_i, \widehat{\boldsymbol{\beta}})\right) - 1\right]. \end{aligned} \tag{7}$$

$\widehat{SQ}(\lambda)$ is then the sample mean difference of the two "extended samples" $\boldsymbol{u}_g,\ g = 1, 2$, by which we mean the vector of actual observations $\boldsymbol{y}_{(g)}$ augmented with the transformed values from the other sample $\boldsymbol{y}^\star_{(g)}$. Therefore, it has the desirable property of being symmetric, that is $\widehat{SQ}(\boldsymbol{u}_1, \boldsymbol{u}_2, \lambda) = -\widehat{SQ}(\boldsymbol{u}_2, \boldsymbol{u}_1, \lambda)$ which is not necessarily shared by shift estimators.

Furthermore, $\widehat{SQ}(\lambda)$ can also be viewed as a linear combination of order statistics, but with weights estimated from the data, and thus it is related to L-estimation (Huber, 1996; Serfling, 1980).

The motivating application for SQUARE is $n_1 < n_2$, that is, one sample is much smaller than the other. Here we calculate $\widehat{SQ}(\lambda)$ by replacing $\boldsymbol{y}_2$ by $\boldsymbol{q}_2$, the linear interpolation of the order statistics $y_{2(i)}$ to the grid of points $p_{1i} = i/(n_1 + 1),\ i = 1, \dots, n_1$. Similarly if $n_1 > n_2$, then we replace $\boldsymbol{y}_1$ by $\boldsymbol{q}_1$, the linear interpolation of the order statistics $y_{1(i)}$ to the grid of points $p_{2i} = i/(n_2 + 1),\ i = 1, \dots, n_2$. This definition of SQUARE still maintains the property of symmetry.

This paper focus on non-zero random variables, but a common difficulty in the statistical analysis of expenditure data is the presence of a significant percentage of zero-cost observations. For example, in our application, the total number of cases and controls are $N_1 = 188$ and $N_2 = 9228$, respectively. Among these only $n_1 = 118$ and $n_2 = 2262$ have non-zero expenditures, the remaining $N_1 - n_1 = 70$ and $N_2 - n_2 = 6966$ have observations with zero costs. If we let $\pi_1 = P(Y_1 > 0)$ and $\pi_2 = P(Y_2 > 0)$ be the probabilities of non-zero expenditure the disease and control groups, respectively, and let $\mu_1 = E[Y_1 \mid Y_1 > 0]$ and $\mu_2 = E[Y_2 \mid Y_2 > 0]$ be the mean of the non-zero values in the disease and control groups, then we seek to estimate $\Delta = \pi_1\mu_1 - \pi_2\mu_2$. It is appropriate to revise Equation (7) as $\widehat{SQ}(\lambda) = \hat{\pi}_1\bar{u}_1 - \hat{\pi}_2\bar{u}_2$ where $\hat{\pi}_j$ is the fraction of non-zero responses for population j.

2.3 Special Cases of SQUARE

In the previous section we have illustrated how to estimate Δ with SQUARE, that is by using a semi-parametric procedure where: 1) we first estimate $s(p, \boldsymbol{\beta})$ by taking the fitted values from the regression model (6); 2) given the estimated $s(p, \widehat{\boldsymbol{\beta}})$, we estimate Δ non parametrically. For different shapes of $s(p, \boldsymbol{\beta})$, choices of the basis functions $B_j(p)$, and specifications of parametric cdf, SQUARE encompasses a very large class of estimators. Below are detailed some special cases.

1. $\widehat{SQ}(\mathrm{U}nif, \lambda = 0)$**:** $s(p, \boldsymbol{\beta})$ **is constant and** $Y_g,\ g = 1, 2$ **are uniform r.v.** We assume $Y_g \sim U[0, \theta_g]$, then $Q_1(p)/Q_2(p) = \theta_1/\theta_2$ and $\Delta = (\theta_1 - \theta_2)/2$. The SQUARE estimate of Δ, denoted as $\widehat{SQ}(\mathrm{U}nif, \lambda = 0)$ is obtained by: 1) fitting the regression model (6) with $B_0(p) = 1$ and $B_1(p) = 0$, and 2) using $s(p_i, \widehat{\boldsymbol{\beta}}) = \hat{\beta}_0 = \bar{ly}_1 - \bar{ly}_2$ where $ly = \log(y)$

in equation (7). This leads to $\widehat{SQ}(\mathrm{U}nif, \lambda = 0) = \frac{1}{2}\left[\bar{y}_1(1 - \exp(-\widehat{\beta}_0)) - \bar{y}_2(1 - \exp(\widehat{\beta}_0))\right]$. Note that $\widehat{SQ}(\mathrm{U}nif, \lambda = 0)$ is not the MLE of Δ which is equal to $(y_{1(n)} - y_{2(n)})/2$.

2. $\widehat{SQ}(LN, \lambda = 1)$**:** $s(p, \boldsymbol{\beta})$ **is linear in** $\Phi^{-1}(p)$ **and** Y_g, $g = 1, 2$ **are log-normal r.v.** We assume $Y_g \sim LN(\nu_g, \sigma_g)$, then $\log(Q_1(p)/Q_2(p)) = \beta_0 + \beta_1\Phi^{-1}(p)$ where $\Phi^{-1}(p)$ is the quantile function of the Normal r.v., $\beta_0 = (\nu_1 - \nu_2)$, $\beta_1 = (\sigma_1 - \sigma_2)$, and $\Delta = \exp(\nu_1 + \sigma_1^2/2) - \exp(\nu_2 + \sigma_2^2/2)$. The SQUARE estimate of Δ, denoted as $\widehat{SQ}(LN, \lambda = 1)$, is obtained by: 1) fitting the regression model (6) with $B_0(p) = 1$ and $B_1(p) = \Phi^{-1}(p)$, and 2) using $s(p_i, \widehat{\boldsymbol{\beta}}) = \hat{\beta}_0 + \hat{\beta}_1\Phi^{-1}(p_i)$ in equation (7). Note that $\widehat{SQ}(LN, 1)$ is not the MLE of Δ, which instead is defined as $LN = \exp(\bar{ly}_1 + s_1^2/2) - \exp(\bar{ly}_2 + s_2^2/2)$, where $ly = \log y$ and s is the standard deviation of the log-transformed data. Also note that, if $\sigma_1 = \sigma_2$, then $s(p, \lambda)$ is constant in p and equal to β_0.

3. **The sample mean difference:** $s(p, \boldsymbol{\beta})$ **interpolates the log-quantile ratios.** Here $n_1 = n_2 = n = \lambda$ and the basis functions in (6) can be chosen so that $s(p, \widehat{\boldsymbol{\beta}})$ interpolates the values $\log\left(\frac{y_{1(i)}}{y_{2(i)}}\right)$. In this case, we treat the two samples as independent and we do not borrow strength from one distribution to the other in estimating Δ. Here SQUARE reduces to the difference in sample means $\bar{y}_1 - \bar{y}_2$.

4. **L-estimates:** $s(p, \boldsymbol{\beta})$ **is known but with unspecified shape.** Equation (7) shows that, if $s(p, \boldsymbol{\beta})$ is known then SQUARE is the average of two L-estimates (Huber, 1981).

3 Asymptotic Properties of SQUARE

In this section we show that the random coefficients $\widehat{\boldsymbol{\beta}}$ and SQUARE itself are both asymptotically normal. We give explicit expressions for the variance of each.

Theorem 1 Consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}$ *Assume* $n_1, n_2 \to \infty$ *and there exist* M, b_1, b_2 *and* $\delta > 0$ *such that*

A) $|\log F_g^{-1}(p)| \leq Mp^{-1/2+b_1+\delta}(1-p)^{-1/2+b_2+\delta}$, *for* $g = 1, 2$, *B) the basis functions $|B_j(p)|$ are continuously differentiable on $(0,1)$ and $|B_j(p)| \leq Mp^{-b_1}(1-p)^{-b_2}$*

then $\widehat{\boldsymbol{\beta}}_j$ is strongly consistent for $\boldsymbol{\beta}_j$, $j = 0, 1, \ldots, \lambda$. In addition, if we assume that

C) the limit $\lim_{n_1,n_2 \to \infty} n_1/(n_1+n_2)$ exists and is in the interval $(0,1)$

then $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ has an asymptotic multivariate normal distribution with mean 0 and covariance matrix $\Sigma = (\sigma_{ij})$ where

$$\sigma_{ij} = \left\{ \frac{1}{n_1} + \frac{1}{n_2} \right\} \int_0^1 \int_0^1 (\min\{p,q\} - pq)\, B_i(p) B_j(q) dpdq.$$

Remark 1 *For consistency alone, the first condition can be replaced by the following relaxed condition*

A′) $\int |\log Y_g|^r dF_g(x) < \infty$, $|\log F_g^{-1}(p)| \leq Mt^{-1+b_1+\delta}(1-t)^{-1+b_2+\delta}$, for $g = 1, 2$.

Proof: The consistency and asymptotic normality of individual coefficients $\widehat{\boldsymbol{\beta}}_j$ is an immediate corollary to the L-statistic results of Shorack Shorack (1972) and Wellner Wellner (1977). The Cramer-Wold device is applied to show that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ has an asymptotic multivariate normal distribution. See, for example, Billingsly **?** for details of Cramer and Wold's method.

Asymptotic Normality of SQUARE The principal result of this section is the following proof that $\widehat{\Delta} - \Delta$ has an asymptotic normal distribution. Our general approach will be the the differentiable statistical functional method (functional δ-method) developed by von Mises and described in (Serfling, 1980). To use this method, we establish an asymptotic equivalence between the SQUARE estimator and the functional below, to which we can adapt the von Mises framework. Details of this result, including a proof of the equivalence, are found at Cope (2003). Our functional takes the form:

$$T(F_1, F_2) = \frac{1}{2} \int_0^1 F_1^{-1}(p)\, (1 - \exp\{-s(p, \boldsymbol{\beta})\})\, dp + \frac{1}{2} \int_0^1 F_2^{-1}(p)\, (\exp\{s(p, \boldsymbol{\beta}) - 1\})\, dp$$

where

$$s(p, \boldsymbol{\beta}) = \sum_{j=0}^{\lambda} \beta_v j B_j(p),$$

and

$$\boldsymbol{\beta}_j = \int_0^1 B_j(p) \left[\log\left(F_1^{-1}(p)\right) - \log\left(F_2^{-1}(p)\right)\right] dp.$$

so that the functional version of the estimator is given by $T(\widehat{F}_1, \widehat{F}_2)$. To prove asymptotic normality, we expand the functional in a one term Taylor series. The directional derivative of the functional at the point (F_1, F_2) converges to a Gaussian distribution. If the remainder converges in probability to zero, then the estimator, like the derivative, has a Gaussian limiting distribution.

The assumptions are bounding conditions very similar to those required to prove that L-estimators are asymptotically normal.

Theorem 2 *Assume that* $n_1, n_2 \to \infty$ *and that* $\lambda_g = \lim_{n_1,n_2\to\infty} n_1/(n_1+n_2)$ *exists and lies in* $(0,1)$. *Suppose there exist* $M, b,$ *and* $\delta > 0$ *such that the following conditions hold for* $g = 1, 2,\ j = 1, 2, \ldots, k$ *and all* $p \in (0,1)$

A) $F_g^{-1} \leq M(p(1-p))^{-b+\delta}$ *and* $|\log F_g^{-1}| \leq M(p(1-p))^{-1/2+\delta}$,

B) $\exp\left\{(-1)^g \sum_{j=1}^{\lambda} \boldsymbol{\beta}_j B_j(p)\right\} \leq M(p(1-p))^{-1/2+b}$,

C) $|B_j(p)| \leq M(p(1-p))^{-\delta/(\delta+2)}$.

Then $\sqrt{n}\left(\widehat{\Delta} - \Delta\right)$ *has an asymptotic normal distribution with mean 0 and variance* σ^2 *where*

$$\sigma^2 = \int_{p=0}^1 \int_{q=0}^1 (min(p,q) - pq)\left(\lambda_1 \eta_1(p)\eta_1(q) + \lambda_2 \eta_2(p)\eta_2(q)\right) dp dq, \tag{8}$$

and

$$\eta_g(p) = \frac{F_g^{-1}(p) + \frac{1}{2}\left(F_1^{-1}(p) + F_2^{-1}(p) - \int_{q=0}^1 \sum_{j=1}^K B_j(q)\left(F_1^{-1}(q) + F_2^{-1}(q)\right) dq\right)}{(-1)^g F_g^{-1}(p) f_g\left(F_g^{-1}(p)\right)}.$$

A sketch of the proof is found in the appendix. Weak consistency is obtained as a corollary to asymptotic normality.

We conclude with a couple of examples to which our asymptotic results apply.

Example 1 *If both samples are drawn from lognormal distributions, then SQUARE is consistent and asymptotically normal. In this case,* $F_g = \Phi(\frac{\log x - \mu_g}{\sigma_g})$, $F_g^{-1} = \exp\{\mu_g + \sigma_g \Phi^{-1}(p)\}$, *and* $\log F_g^{-1} = \mu_g + \sigma_g \Phi^{-1}(p)$, *where* Φ *is the standard normal distribution function. The log-quantile ratio is a linear function of* $\Phi^{-1}(p)$ *so a natural orthonormal basis is* $B_0 \equiv 1$ *and* $B_1 = \Phi^{-1}(p)$.

Example 2 *If both samples are drawn from Pareto distributions with cdf* $F(x) = 1 - b^a x^{-a}$ *and* $a > 2$, *then SQUARE is strongly consistent and asymptotically normal. The Pareto distribution makes an interesting example for SQUARE because it is very heavy tailed, and has a finite* k^{th} *moment only if the shape parameter* $a \geq k$. *Its density function is* $f(x) = ab^a x^{-a-1}$, *where* $x \geq 1$ *and* $a, b > 0$. *The log-quantile function is given by* $\log F^{-1}(p) = \log b - \log(1-p)/a$ *leading to the two basis functions* $B_0(p) \equiv 1$ *and* $B_1(p) = \log(1-p)$. *Note that the orthonormalized version of* B_1 *is equal to* $(\log(1-p)+1)/\sqrt(3)$.

4 Simulations and Data Analysis

This section demonstrates that $\widehat{SQ}(\lambda)$ (for moderate λ, and for λ estimated from the data) has substantially lower mean squared error and bias than common used estimators of Δ, such as the maximum likelihood estimator for log-normal populations and the sample mean difference.

More specifically, we estimate λ by use of a B-fold cross-validation method (Efron, 1983; Breiman and Spector, 1992; Efron and Tibshirani, 1993; Shao and Tu, 1995) which minimizes

$$CV(\lambda) = \sum_{b=1}^{B} \left[(\bar{y}_1^{(b)} - \bar{y}_2^{(b)}) - \widehat{SQ}_{\lambda}^{(-b)} \right]^2 \tag{9}$$

where $(\bar{y}_1^{(b)} - \bar{y}_2^{(b)})$ is the sample mean difference applied to the two b-th random sub-vectors for the cases and the control (the training sets) a $\widehat{SQ}_{\lambda}^{(-b)}$ is the SQUARE estimate obtained with the rest of the data. We choose $B = 10$ and we minimizes $CV(\lambda)$ for $\lambda = 1, 2, 4, 6, 8$.

Data are generated under 4 scenarios, A, B, C, and D. Under each scenario, we compare bias and variance properties of the following six estimators of Δ: 1) $\widehat{SQ}(\widehat{\lambda})$ where $\widehat{\lambda}$ is estimated by minimizing $CV(\lambda)$ in equation (9); 2)$\widehat{SQ}(\lambda = 2)$; 3) $\widehat{SQ}(\lambda = 4)$; 4) SQUARE under the assumption that the two populations are log-normal $\widehat{SQ}(LN, 1)$; 5) the maximum likelihood estimator under the log-normal model LN; and 6) the sample mean difference $\bar{y}_1 - \bar{y}_2$.

Table 1 and Figure 3 summarize the four scenarios studied. The first three scenarios A,B and C, are theoretical distributions in which population 2 is log-normal with mean $\nu_2 = 7$ and standard error $\sigma_2 = 1.5$. These parameters were chosen to roughly approximate the sample statistics from the medical expenditures datasets for non-diseased subjects. In scenario A, population 1 is is also log-normal with a higher mean $\nu_1 = 7.5$ and a higher standard error $\sigma_1 = 1.75$. In scenarios B and C, population 1 differs from 2 by the functions $s(p)$ shown in figure 3, chosen to represent a range of plausible shapes. We also studied $\widehat{SQ}(\lambda)$'s performance for the empirical expenditure distribution drawn from the NMES data. Scenario D, whose log-quantile functions are pictured in Figure 3 with dark solid lines, contrasts the distributions of non-zero medicare expenditures for 118 lung cancer or COPD patients to 2262 controls.

In our simulations, we generated 500 data sets for each scenario, and we compared estimators with equal sample sizes $n_1 = n_2 = 100$ and unequal samples with $n_1 = 100,\ n_2 = 1000$. The results were qualitatively similar and hence we report only the unequal case. For each generated data set, we estimate SQUARE for $\widehat{\lambda}$ and for $\lambda = 2, 4$ by use of natural cubic splines as basis functions.

These results show that $\widehat{SQ}$ outperforms both $\bar{y}_1 - \bar{y}_2$ and the log-normal estimators in terms of MSE. Table 2 presents the relative mean square error (MSE) as a percent of the MSE for $\bar{y}_1 - \bar{y}_2$, e.g. $\left(mse(\bar{y}_1 - \bar{y}_2) - mse(\widehat{\Delta})\right)/mse(\bar{y}_1 - \bar{y}_2)$. Negative values imply that $\bar{y}_1 - \bar{y}_2$ is preferred, positive percents favor the comparator $\widehat{\Delta}$.

In scenario A, when both populations are log-normal, $\widehat{SQ}(\widehat{\lambda})$ and $\widehat{SQ}(\lambda)$ for $\lambda = 2, 4$ are $52, 49$ and 40 percent better than $\bar{y}_1 - \bar{y}_2$. Note that the SQUARE estimates perform better even than the log-normal MLE (LN) which in this case is asymptotically efficient. This is because the relatively small sample size of the case group ($n_1 = 100$) leads to a maximum likelihood estimate of $E[Y_1]$ that is less efficient than the sample mean of the extended sample $\bar{u}_1$. By borrowing strength from the distribution in the control group, SQUARE gains efficiency in estimating $E[Y_1]$.

In scenario B, the six estimators have comparable performance and they are all superior than the sample mean difference. In scenarios C and D, the estimates $\widehat{SQ}(LN, 1)$ and LN performs very poorly due to the substantial non-linearity of $s(p)$ whereas the SQUARE estimates are $51, 54$ and 44 percent better than the sample mean difference. Finally, for the empirical scenario D, the SQUARE estimates are 20 percent better than the sample mean difference, and again the log-normal estimators performs very poorly.

Table 3 summarizes bias of $\widehat{\Delta}$ as a percent of the true Δ, e.g. $\left(E(\widehat{\Delta}) - \Delta\right)/\Delta$, and shows that $\widehat{SQ}(\lambda)$ has small percent bias in the cases considered. In most cases the bias of the SQUARE estimates is comparable to the bias of the sample mean difference, which is unbiased in large samples. As expected, the bias of $\widehat{SQ}(LN, 1)$ is small only when $s(p)$ is almost constant. Finally, except in scenario A when the two populations are log-normal, the LN is badly biased.

We have also performed a sensitivity analysis of SQUARE to the choice of the basis functions. For each of the 500 data sets, and for each scenario, we estimate $s(p)$ by using, in addition to the natural cubic splines, smoothing splines and polynomials. These estimates are all close to each other and to the true $s(p)$ (results not shown) .

Finally, we analyze the NMES data represented in Figures 1 and 2, also used as scenario D, to estimate the mean difference between annual Medicare expenditures for persons with lung cancer or COPD (cases), diseases caused largely by smoking, and otherwise similar persons without these two smoking-attributable diseases (controls). In addition to estimating the overall mean difference in expenditures for persons with and without disease caused by smoking, a second question is whether this difference is smaller for smokers who perhaps seek or receive fewer services. That is, does smoking status modify the difference in medical expenditures between the cases and the controls? Table 4 shows the number of disease cases and controls for smokers (current or former), and for the non-smokers (never). The numbers within parentheses represent the percentage of people in that cell with non-zero expenditures. The percentage of cases with non-zero expenditures is more than twice as large as for the controls (65% and 25%); this is consistent with our expectation that people with disease receive more services. These proportions are similar for smokers and non-smokers.

We apply the two-part analogues of the: 1) SQUARE estimates, 2) maximum likelihood estimator under log-normal population; and 3) weighted difference in the sample means to the NMES data base, and to the subset of the NMES data for smokers only. Figure 4 summarizes boxplots of 100 bootstrap estimates of Δ for everyone and for the smokers only. Simulation study results suggest that the SQUARE estimates are far more efficient with respect to the selected competitors. In addition the non-linearity of the estimated $s(p_i, \boldsymbol{\beta})$ (bottom right panel of Figure 3) also suggests that and that the maximum likelihood estimator (LN) is likely to be highly biased. As expected the estimator $\widehat{SQ}(\widehat{\lambda})$ is more variable than the SQUARE estimates with λ known, because it also takes into account of the model uncertainty. Estimates for the smokers are slightly larger than for everyone.

5 Discussion

In this paper, we have proposed a novel class of estimators of the difference of the expected values of two skewed distributions that encompasses most of the current approaches. Our innovative approach model the log-ratio of the two quantile functions as a smooth function of the percentiles where the degree of smoothness can be estimated from the data. By smoothing across percentiles, we borrow strength across the two samples and produce an estimator that is more efficient than the difference in sample means and log-normal estimators in the cases relevant to the motivating approach. In summary, SQUARE is a semi-parametric method using a parametric model for smoothing the log-quantile ratios across percentiles, and a non-parametric estimator of the two quantile functions. The software for implementing SQUARE and the data for reproducing all the analyses reported in this paper are available at `http://biostat.jhsph.edu/~fdominic/square.html`.

The idea of linking two samples in a semi-parametric model is obviously not new. Perhaps the most famous and influential example is the Cox proportional hazards model (Cox, 1972) where the target is the hazard ratio. A second example is the density ratio model of Qin and Zhang (1997). Here, the ratio of densities $f(x)/g(x)$is assumed to be a smooth function of x. This model would lead to an estimator of the mean difference that is analogous to ours but where a smooth function of the unordered data is used in Equation 7 rather than a smooth function of the order statistics. There is a fundamental difference in the two approaches. We smoothly map the random variables themselves to one another; they map their probabilities.

SQUARE encompasses and generalizes a large class of estimators including L-estimates (Huber, 1996; Serfling, 1980). If $s(p)$ is known, then SQUARE is a linear combination of order statistics. If $s(p)$ is unknown, then SQUARE is still a linear combination of order statistics, but with weights estimated from data. This fact has been exploited to develop the asymptotic theory presented in this paper, which owes much to the L-estimation results of Shorack (1972), Wellner (1977), Boos (1979), and Serfling (1980).

Under certain bounding conditions for the quantile functions, we showed that SQUARE is consistent and asymptotically normal. We provided an explicit expression for the asymptotic variance of SQUARE, and examples where our asymptotic results apply. Although the bounding conditions included in the theorems may appear awkward, their purpose is quite straightforward: these bounds are sufficient to ensure that all integrals are bounded in probability. The bounds are tight and cannot be relaxed, but other combinations of conditions will suffice as well.

For an arbitrary pair of distributions, it is not guaranteed that the quantile ratio falls within the span of any pre-defined and finite basis. Therefore it is important to allow the basis to grow with the sample. Asymptotic normality can generally be extended to this case if the size of the basis is no greater than $\log n$. We do require some regularity conditions on the basis for this result to hold, but for square-integrable, log quantile functions, we can guarantee the existence of a suitable basis. An extensive discussion of the asymptotic properties of SQUARE and its relationships with L-estimation is reported in Cope (2003).

In the simulation study, we showed that SQUARE out-performs the sample mean difference and the log-normal MLE estimator for moderate values of the smoothing parameter λ and for λ estimated by use of cross-validation methods. We also performed a simulation study where we sample Y_1 and Y_2 from exponential distributions. Here the maximum likelihood estimator of the mean difference of two exponential random variables is the sample mean difference, and SQUARE's performance is similar to the MLE. As an alternative to SQUARE, we could assume that $Q_{\log Y_1}(p) = s(Q_{\log Y_2}(p), \lambda)$, and estimate Δ by using the fitted values of the QQ-plot. In our simulation study we compared this estimator with SQUARE. We found that, although this is certainly a reasonable estimation approach, it is not as efficient as SQUARE.

The development of SQUARE was motivated by the estimation of smoking attributable medical expenditures. A key component is to estimate the mean medical expenditures between persons with smoking attributable-diseases (e.g. lung cancer or COPD) and otherwise similar persons without such diseases. Our analysis of expenditures allows smoking status to modify the effect of disease on expenditures. We examine this effect modification first by stratifying the cases and the controls with respect to their smoking status, and then by estimating SQUARE separately for smokers and non-smokers, within each stratum.

However, a more desirable goal would be to compare medical expenditures for cases and controls taking into account individual-level characteristics $\boldsymbol{x}$. In this case SQUARE can be extended to the regression case by assuming

$$\log Q_1(p; \boldsymbol{x}) = \log Q_2(p; \boldsymbol{x}) + s(p; \boldsymbol{x}). \tag{10}$$

To control for systematic differences in covariates between two populations, a common strategy is to group units into subclasses based on covariate values, for example using propensity score matching (Cochran and Rubin, 1973; Rubin, 1973), and then estimate SQUARE within strata of propensity scores. The extension of SQUARE to the regression case, and a comparison between regression SQUARE and common econometric models such as a two-part log-linear regression models (Duan, 1983; Mullahy, 1998; Mullahy and Manning, 1995) is exploited in Dominici and Zeger (2003).

The potential applications of SQUARE are numerous. For example, in clinical trials our approach can be used to estimate treatment effects that vary smoothly with respect to the percentiles of the health outcome. If Y has a more nearly symmetric distribution, rather than smoothing the ratio of the quantiles, we can smooth their difference; that is, we can assume $Q_1(p) - Q_2(p) = s(p)$. Under this model, we estimate the treatment effect, Δ, by $\int s(p)dp$. The plot of the estimated $s(p)$ versus p is also informative for identifying the outcome percentiles where the treatment is mostly effective.

6 Appendix

In using the differentiable statistical functional approach, the largest task is to demonstrate that the remainder

$$R_1 = \sqrt{n}\left(T(\widehat{F}) - T(F) - d_1\left(T, F; \sqrt{n}(\widehat{F} - F)\right)\right)$$

converges to zero in distribution. If the remainder vanishes in the limit, then $\sqrt{n}\left(T(\widehat{F}) - T(F)\right)$ is equivalent to $d_1\left(T, F; \sqrt{n}(\widehat{F} - F)\right)$ and the asymptotic properties of the former can be derived from the latter. In the case of SQUARE, the derivative (**??**) is given by

$$\begin{aligned} d_1\left(T, F; \sqrt{n}(\widehat{F} - F)\right) &= \sqrt{n\lambda_1}\int_{p=0}^{1}\left(F_1(\widehat{F}_1^{-1}(p)) - p\right)\eta_1(p)dp \\ &\quad + \sqrt{n\lambda_2}\int_{p=0}^{1}\left(F_2(\widehat{F}_2^{-1}(p)) - p\right)\eta_2(p)dp. \end{aligned} \tag{11}$$

Both $\sqrt{n\lambda_1}\left(F_1(\widehat{F}_1^{-1}(p)) - p\right)$ and $\sqrt{n\lambda_2}\left(F_2(\widehat{F}_2^{-1}(p)) - p\right)$ converge to Brownian bridges, so the derivative has a normal asymptotic distribution with variance σ^2 as defined above.

Sketch of proof that the remainder converges to zero. At points in this proof it is necessary to evaluate expressions like $\int_{p=0}^{1} \widehat{F}^{-1}(p)J_n(p)dp$ where $\widehat{F}^{-1}$ is an empirical quantile function and $J_n(p)$ may also be data dependent. In order to simplify treatment of these expressions, the following lemma establishes conditions under which the range of integration can be truncated.

Lemma 1 *Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample. Suppose that $\widehat{F}^{-1}$ is the empirical quantile function corresponding to this data, and let $J_n : (0,1) \to \Re$ be a possibly random function. Assume that there exist positive constants M, b and δ such that*

A) the quantile function $F^{-1}(p) \le M(p(1-p))^{-b+\delta}$, and

B) the random function $|J_n(x)| \le \left(M(p(1-p))^{-1/2+b}\right)^{1+\epsilon_n}$ where $\epsilon_n \xrightarrow{p} 0$.

Then

$$T_n = \sqrt{n}\left(\int_0^{k/n} \widehat{F}^{-1}(p)J_n(p)dp + \int_{(n-k)/n}^{1} \widehat{F}^{-1}J_n(p)dp\right) \xrightarrow{p} 0.$$

The proof is not included here.

We break the remainder up into several pieces and prove convergence separately for each piece. Here $R_{1,n}$ can be written as

$$\left.\begin{array}{rl} & \frac{1}{2}\int_{p=0}^{1}\sqrt{n}\left(\frac{\widehat{F}_1^{-1}(p)}{2}-\frac{F_1^{-1}(p)}{2}-\frac{\left[\widehat{F}_1(F_1^{-1}(p))-p\right]}{2f_1(F_1^{-1}(p))}\right)(1-\exp\left(-s(p,\boldsymbol{\beta})\right))\,dp, \\ - & \frac{1}{2}\int_{q=0}^{1}\frac{\sqrt{n}}{2}F_1^{-1}(q)\left[\exp\left(-s(q,\widehat{\boldsymbol{\beta}})\right)-\exp\left(-s(q,\boldsymbol{\beta})\right)\right. \\ - & \left.\exp\left(-s(q,\boldsymbol{\beta})\right)\sum_j B_j(q)\int_{p=0}^{1}B_j(p)\left[\frac{\left[\widehat{F}_1(F_1^{-1}(p))-p\right]}{F_1^{-1}(p)f_1(F_1^{-1}(p))}-\frac{\left[\widehat{F}_2(F_2^{-1}(p))-p\right]}{F_2^{-1}(p)f_2(F_2^{-1}(p))}\right]\right]dpdq, \\ + & \frac{1}{2}\int_{p=0}^{1}\left[\exp\left(-s(p,\widehat{\boldsymbol{\beta}})\right)-\exp\left(-s(p,\boldsymbol{\beta})\right)\right]\left[\widehat{F}_1^{-1}(p)-F_1^{-1}(p)\right]dp \end{array}\right\} \quad (12)$$

$$\left.\begin{array}{rl} + & \frac{1}{2}\int_{p=0}^{1}\sqrt{n}\left(\frac{\widehat{F}_2^{-1}(p)}{2}-\frac{F_2^{-1}(p)}{2}-\frac{\left[\widehat{F}_2(F_2^{-1}(p))-p\right]}{2f_2(F_2^{-1}(p))}\right)(\exp(s(p,\boldsymbol{\beta}))-1)\,dp, \\ + & \frac{1}{2}\int_{q=0}^{1}\frac{\sqrt{n}}{2}F_1^{-1}(q)\left[\exp\left(s(q,\widehat{\boldsymbol{\beta}}\right)-\exp(s(q,\boldsymbol{\beta}))\right. \\ - & \left.\exp(s(q,\boldsymbol{\beta}))\sum_j B_j(q)\int_{p=0}^{1}B_j(p)\left[\frac{\left[\widehat{F}_1(F_1^{-1}(p))-p\right]}{F_1^{-1}(p)f_1(F_1^{-1}(p))}-\frac{\left[\widehat{F}_2(F_2^{-1}(p))-p\right]}{F_2^{-1}(p)f_2(F_2^{-1}(p))}\right]\right]dpdq, \\ + & \frac{1}{2}\int_{p=0}^{1}\left[\exp\left(s(p,\widehat{\boldsymbol{\beta}}\right)-\exp(s(p,\boldsymbol{\beta}))\right]\left[\widehat{F}_2^{-1}(p)-F_2^{-1}(p)\right]dp. \end{array}\right\} \quad (13)$$

Only (12) is examined here since the treatment of (13) is identical.

Our first step is to rewrite (12) as $R_1+R_2+R_3$ where we define

$$R_1=\frac{\sqrt{n}}{2}\int_{p=0}^{1}\left(\widehat{F}_1^{-1}(p)-F_1^{-1}(p)-\frac{\left[\widehat{F}_1(F_1^{-1}(p))-p\right]}{f_1(F_1^{-1}(p))}\right)(1-\exp\left(-s(p,\boldsymbol{\beta})\right))\,dp,$$

$$\begin{aligned} R_2= & -\int_{q=0}^{1}\sqrt{n}F_1^{-1}(q)\left[\frac{\exp\left(-s(q,\widehat{\boldsymbol{\beta}})\right)-\exp\left(-s(q,\boldsymbol{\beta})\right)}{2}-\exp\left(-s(q,\boldsymbol{\beta})\right)\right. \\ & \left.\times\sum_{i=0}^{\lambda}B_i(q)\int_{p=0}^{1}B_i(p)\left[\frac{\left[\widehat{F}_1(F_1^{-1}(p))-p\right]}{2F_1^{-1}(p)f_1(F_1^{-1}(p))}-\frac{\left[\widehat{F}_2(F_2^{-1}(p))-p\right]}{2F_2^{-1}(p)f_2(F_2^{-1}(p))}\right]\right]dpdq, \end{aligned}$$

and

$$R_3=\frac{\sqrt{n}}{2}\int_{p=0}^{1}\left[\exp\left(s(p,\widehat{\boldsymbol{\beta}}\right)-\exp(s(p,\boldsymbol{\beta}))\right]\left[\widehat{F}_2^{-1}(p)-F_2^{-1}(p)\right]dp.$$

R_1 is the remainder from the differentiable statistical functional representation of an L-statistic and converges in probability to zero.

After a little bit of algebra, R_2 can likewise largely be expressed in terms of L-statistic remainders and demonstrated to converge in probability to zero. With some manipulation, R_2 can be written as

$$\begin{aligned} R_2 &= \frac{\sqrt{n}}{2}\int_{p=0}^{1}\left(\log\widehat{F}_1^{-1}-\log F_1^{-1}-\frac{\left[\widehat{F}_1(F_1^{-1}(p))-p\right]}{F_1^{-1}(p)f_1(F_1^{-1}(p))}\right) \\ &\quad\times\sum_{j=1}^{\lambda}B_j(p)\int_{q=0}^{1}F_2^{-1}B_j(q)dq, \end{aligned} \quad (14)$$

$$+ \quad \frac{\sqrt{n}}{2} \int_{p=0}^{1} \left(\log \widehat{F}_2^{-1} - \log F_2^{-1} - \frac{\left[\widehat{F}_2(F_2^{-1}(p)) - p \right]}{F_2^{-1}(p) f_2(F_2^{-1}(p))} \right)$$

$$\times \sum_{j=1}^{\lambda} B_j(p) \int_{q=0}^{1} F_2^{-1} B_j(q) dq, \tag{15}$$

$$+ \quad \frac{\sqrt{n}}{2} \int_{p=0}^{1} F_2^{-1}(p) \exp\left(- \sum_j \xi_j B_j(p) \right) \sum_j \left((\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) B_j(p) \right)^2 dp. \tag{16}$$

Each of the first two terms, (14) and (15), is the remainder from the differentiable statistical functional form of an L-statistic with functional, and so converges in probability to zero.

It remains to deal with (16), as well as R_3. We have

$$\exp\left(- \sum_j \xi_j B_j(p) \right) \quad \leq \quad \exp\left(\left| \sum_j \boldsymbol{\beta}_j B_j(p) \right| \right)^{1+\Gamma_n},$$

and thus

$$\frac{\sqrt{n}}{2} \int_{p=0}^{1} F_2^{-1}(p) \exp\left(- \sum_j \xi_j B_j(p) \right) \left(\sum_j (\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) B_j(p) \right)^2 dp$$

$$\leq \quad n^{1/2} M_2^{1+\Gamma_n} \max_i \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i \right)^2 \int_{p=0}^{1} \left((p(1-p))^{-1/2+\delta^2/(\delta+2)} \right)^{1+\Gamma_n} dp. \tag{17}$$

Setting $\widehat{F}^{-1}(p) \equiv 1$, Lemma 1 can be applied so that (17) is bounded from above by

$$M_3 \max_i \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i \right)^2 n^{1/2+(1/2-\delta^2/(\delta+2))(1+\Gamma_n)}.$$

When n is sufficiently large (and so Γ_n is sufficiently small), then the exponent on n is less than 1, ensuring convergence in probability.

The final term, R_3 is handled in very similar fashion. We first apply the mean value theorem to represent $\exp\left(s(p, \widehat{\boldsymbol{\beta}} \right) - \exp(s(p, \boldsymbol{\beta}))$ in the form $\exp\left(- \sum_{j=1}^{\lambda} \xi_j B_j(p) \right) \sum_j B_j(p) (\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$, where the ξ_j are strictly between $\widehat{\boldsymbol{\beta}}_j$ and $\boldsymbol{\beta}_j$. Then, defining $\mathcal{F}_2^{-1} \doteq \widehat{F}_2^{-1}(p) - F_2^{-1}(p)$, we have

$$\begin{aligned} R_3 &= \frac{\sqrt{n}}{2} \int_{p=0}^{1} \exp\left(- \sum_{j=1}^{\lambda} \xi_j B_j(p) \right) \sum_j B_j(p) (\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) \mathcal{F}_2^{-1} dp \\ &\leq \frac{\sqrt{n}}{2} \sum_j |\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j| \int_{p=0}^{1} \exp\left(\left| \sum_{j=1}^{\lambda} \boldsymbol{\beta}_j B_j(p) \right| \right)^{1+\Gamma_n} \max_j |B_j(p)| \mathcal{F}_2^{-1} dp \\ &\leq \frac{\sqrt{n}}{2} \sum_j |\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j| \int_{p=0}^{1} \left[\exp\left(\left| \sum_{j=1}^{\lambda} \boldsymbol{\beta}_j B_j(p) \right| \right) \left(\max_j |B_j(p)| + 1 \right) \right]^{1+\Gamma_n} \mathcal{F}_2^{-1} dp. \end{aligned}$$

(18)

Truncating the integrals and applying the bounding functions, we can see that (18) is asymptotically equivalent to

$$\begin{aligned}
&\frac{\sqrt{n}}{2}\sum_j |\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j| \int_{1/(n+1)}^{n/(n+1)} \left[\exp\left(\left|\sum_{j=1}^{\lambda} \boldsymbol{\beta}_j B_j(p)\right|\right)\left(\max_j |B_j(p)| + 1\right)\right]^{1+\Gamma_n} \mathcal{F}_2^{-1} dp \\
\leq \quad & M_2^{1+\Gamma_n} n^{(1-b)(1+\Gamma_n)} \sum_j |\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j||\overline{X}_2 - \mu_2|.
\end{aligned}$$

The random exponent Γ_n converges to zero at the same $\sqrt{n}$-rate as does $\sum_j |\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j||$. Also $|\overline{X}_2 - \mu_2|$ converges to zero at a $\sqrt{n}$-rate, and so R_3 converges in probability to zero.

References

Aitchison, J. and Shen, S. M. (1980). "Logistic normal Distributions: Some Properties and Uses." *Biometrika*, 67, 261–272.

Andersen, C. K., Andersen, K., and Kragh-Sorensen, P. (2000). "Cost Function Estimation: The Choice of a Model to Apply to Dementia." *Health Economics*, 9, 397–409.

Angus, J. E. (1994). "Bootstrap One-sided Confidence Intervals for the Log-normal Mean." *The Statistician*, 43, 395–401.

Boos, D. (1979). "A differential for L-statistics." *Annals of Statististics*, 7, 955–959.

Breiman, L. and Spector, P. (1992). "Submodel selection and evaluation in regression: The X-random case." *nternational Statistical Review*, 60, 291–319.

Cochran, W. G. and Rubin, D. B. (1973). "Controlling Bias in Observational Studies: A Review." *Sankhyā, Series A, Indian Journal of Statistics*, 35, 417–446.

Cope, L. (2003). "Some Asymptotic Properties of Smooth Quantile Ratio Estimation." Ph.D. thesis, Department of Applied Mathematics Johns Hopkins University, Baltimore, MD.

Cox, D. R. (1972). "Regression models and life tables." *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Doksum, K. A. and Sievers, G. L. (1976). "Plotting With Confidence: Graphical Comparisons of Two Populations." *Biometrika*, 63, 421–434.

Dominici, F. and Zeger, S. (2003). "Smooth Quantile Ratio Estimation with Regression: An Analysis of Medical Expenditures for Smoking Attributable Disease." Technical report, Department of Biostatistis Johns Hopkins University.

Duan, N. (1983). "Smearing Estimate: A Nonparametric Retransformation Method." *Journal of the American Statistical Association*, 78, 605–610.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983). "A comparison of Alternative Models for the Demand for Medical Care." *Journal of Business and Economic Statistics*, 1, 115–125.

Efron, B. (1983). "Estimating the error rate of a prediction rule: Improvement on cross-validation." *Journal of the American Statistical Association*, 78, 316–331.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Fenn, P., McGuire, A., Backhouse, M., and Jones, D. (1996). "Modelling programme costs in economic evaluation." *Journal of Health Economics*, 15, 115–125.

Hlatky, M., Rogers, W., Johnstone, I., et al. (1997). "Medical care costs and quality of life after randomization to coronary angioplasty and coronary bypass surgery." *New England Journal of Medicine*, 336, 92–99.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

— (1996). "Robust Statistical Procedures (2nd ed.)." In *CBMS-NSF Regional Conference Series in Applied Mathematics, Number 68*. Soc. Industr. Appl. Math., Philadelphia, Pennsylvania.

Ihaka, R. and Gentleman, R. (1996). "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*, 5, 299–314.

Land, C. E. (1971). "Confidence Intervals for Linear Functions of the Normal Mean and Variance." *The Annals of Mathematical Statistics*, 42, 1187–1205.

Lin, D. (2000). "Linear regression analysis of censored medical costs." *Biostatistics*, 1, 35–47.

Lin, D. Y., Feuer, E. J., Etzioni, R., and Wax, Y. (1997). "Estimating Medical Costs From Incomplete Follow-up Data." *Biometrics*, 53, 419–434.

Lipscomb, J., Ancukiewicz, M., Parmigiani, G., Hasselblad, V., Samsa, G., and Matchar, D. (1999). "Predicting the Cost of Illness: A comparison of Alternative Models applied to Stroke." *Medical Decision Making*, 18, S39–S56.

Mullahy, J. (1998). "Much ado about two: reconsidering retransformation and the two-part model in health econometrics." *Journal of Health Economics*, 17, 247–281.

Mullahy, J. and Manning, W. (1995). "Statistical issues in cost-effectiveness analysis." In *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceutical and Other Medical Technologies*. New York: Cambridge University Press.

Nair, V. N. (1982). "Q-Q Plots With Confidence Bands for Comparing Several Populations." *Scandinavian Journal of Statistics*, 9, 193–200.

National Center For Health Services Research (1987). *National Medical Expenditure Survey. Methods I I. Questionnaires and data collection methods for the household survey and the Survey of American Indians and Alaska Natives.*. National Center for Health Services Research and Health Technology Assessment.

O'Brien, P. C. (1988). "Comparing Two Samples: Extensions of the t, Rank-sum, and Log-rank Tests." *Journal of the American Statistical Association*, 83, 52–61.

Parzen, E. (1979). "Nonparametric Statistical Data Modeling." *Journal of the American Statistical Association*, 74, 105–121.

Qin, J. and Zhang, A. (1997). "A goodness of fit test for the logistic regression model based on case–control data." *Biometrika*, 84, 609–618.

Rubin, D. B. (1973). "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics*, 29, 185–203.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.

Shao, J. and Tu, D. (1995). New York: Springer-Verlag.

Shorack, G. (1972). "Functions of Order Staistics." *Annals of Mathematical Statistics*, 43, 412–427.

Tu, W. and Zhou, X.-H. (1999). "A Wald Test Comparing Medical Cost Based on Log-Normal Distributions with Zero Valued Costs." *Statistics in Medicine*, 18, 2749–2761.

Wellner, J. (1977). "A Glivenko-Cantelli theorem and strong laws of large numbers for functions of order statistics." *Annal of Statististics*, 5, 473–480.

Wilcox, R. (1995). "Comparing two independent groups via multiple quantiles." *The Statistician*, 44, 91–99.

Wilk, M. B. and Gnanadesikan, R. (1968). "Probability Plotting Methods for the Analysis of Data." *Biometrika*, 55, 1–17.

Zellner, A. (1971). "Bayesian and Non-Bayesian Analysis of the Log-normal Distribution and Log-normal Regression." *Journal of the American Statistical Association*, 66, 327–330.

Zhou, X.-H. and Gao, S. (1997). "Confidence Intervals for the Log-normal Mean." *Statistics in Medicine*, 16, 783–790.

Zhou, X.-H., Gao, S., and Hui, S. L. (1997). "Methods for Comparing the Means of Two Independent Log-normal Samples." *Biometrics*, 53, 1129–1135.

Zhou, X.-H. and Melfi, C.and Hui, S. (1997). "Methods for Comparison of Cost Data." *Biometrics*, 53, 1129–1135.

Table 1: *Description of the sampling mechanisms used under each simulation study scenario.* $\hat{F}_g$, $g = 1, 2$ *are the empirical cdfs of the non-zero Medicare expenditures for patients in the case and control groups.* $g(y) = \exp(\log 7 + \Phi^{-1}(y) \log 1.5)$ *and* Φ *is the cdf of a standard Gaussian variable.*

Scenario	Population 1	Population 2	n_1	n_2
A	$y_1 \sim LN(7.5, 1.75)$	$y_2 \sim LN(7, 1.5)$	100	1000
B	$u \sim \text{Unif}[0,1],\ y_1 = g(u)e^{s_B(u)}$	$y_2 \sim LN(7, 1.5)$	100	1000
C	$u \sim \text{Unif}[0,1],\ y_1 = g(u)e^{s_C(u)}$	$y_2 \sim LN(7, 1.5)$	100	1000
D	$y_1 \sim \hat{F}_1$	$y_2 \sim \hat{F}_2$	118	2262

Table 2: *Mean squared error relative to* $\bar{y}_1 - \bar{y}_2$ *defined by* $\Big((mse(\bar{y}_1 - \bar{y}_2) - mse(\hat{\Delta}))/mse(\bar{y}_1 - \bar{y}_2)\Big) \times 100$ *under the data generation mechanisms described in Section 3. The degrees of freedom* λ *are estimated by the cross-validation approach illustrated in equation (9) for* $B = 10$.

Percent Efficiency				
$\hat{\Delta}$	Scenario A	Scenario B	Scenario C	Scenario D
$\widehat{SQ}(\hat{\lambda})$	52	36	51	20
$\widehat{SQ}(2)$	49	38	54	21
$\widehat{SQ}(4)$	40	30	44	20
$\widehat{SQ}(LN, 1)$	50	40	-316	-196
LN	37	41	-3353	-1311

Table 3: *Percent bias relative to* $\bar{y}_1 - \bar{y}_2$ *defined by* $\Big((E(\hat{\Delta}) - \Delta)/\Delta\Big) \times 100$ *under the data generation mechanisms described in Section 3. The degrees of freedom* λ *are estimated by the cross-validation approach illustrated in equation (9) for* $B = 10$.

Percent bias				
$\hat{\Delta}$	Scenario A	Scenario B	Scenario C	Scenario D
$\widehat{SQ}(\hat{\lambda})$	-11	-21	3	3
$\widehat{SQ}(2)$	-8	-18	4	7
$\widehat{SQ}(4)$	-2	-7	3	3
$\widehat{SQ}(LN, 1)$	-3	-24	39	45
LN	8	-27	119	111
$\bar{y}_1 - \bar{y}_2$	0	-3	-2	1

Table 4: *Disease cases and controls for smokers (current or former) and for non-smokers. Numbers within parentheses represent the percentage of people in that cell with non-zero expenditures.*

	Smokers	Non Smokers	Everyone
cases	165 (64%)	23 (70%)	188 (65%)
controls	4682 (32%)	4546 (28%)	9228 (25%)
	4847 (32%)	4569 (18%)	9416 (25%)

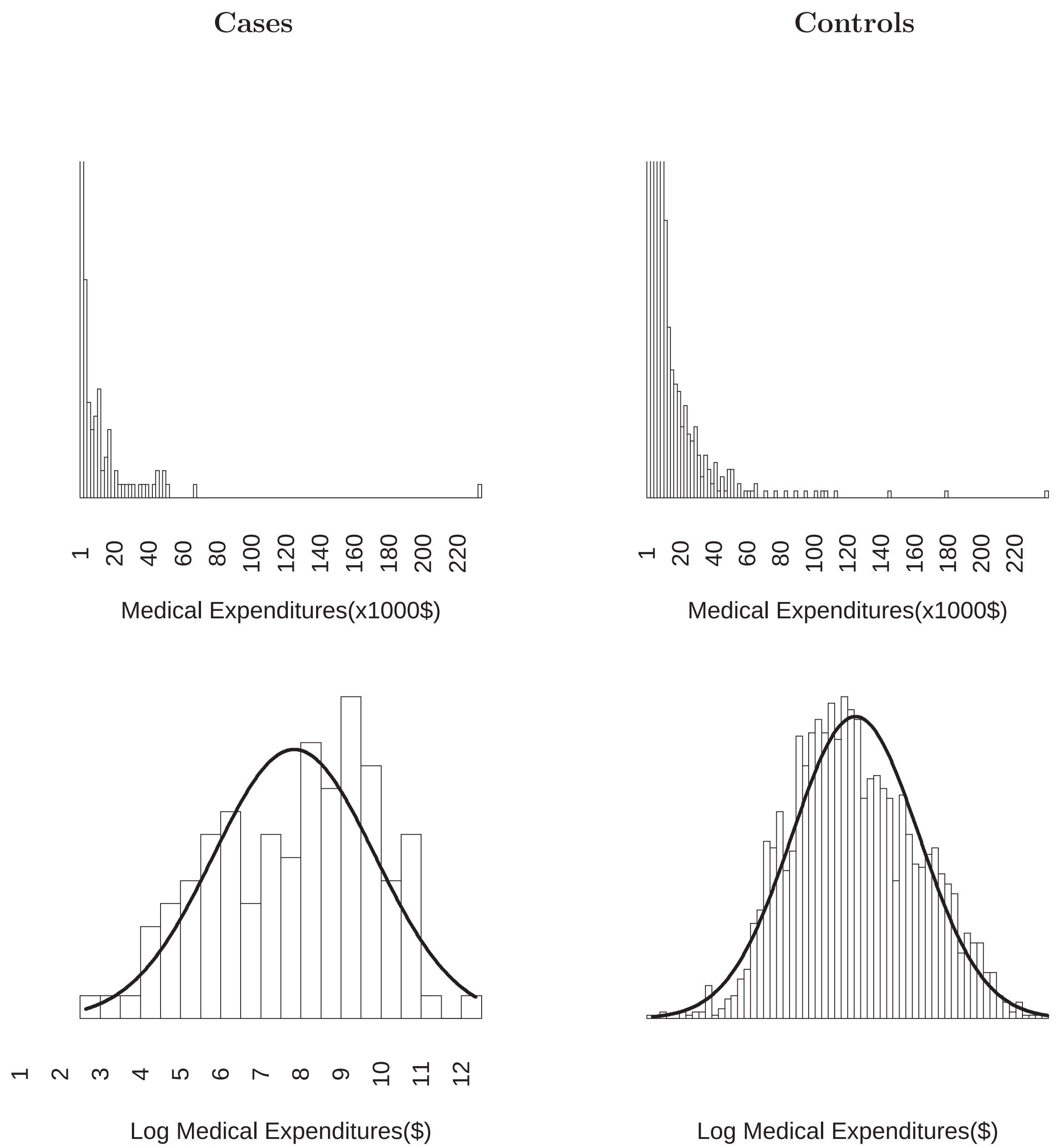

Figure 1: *Histograms of non-zero Medicare medical expenditures for the 1987 National Medical Expenditure Survey (NMES) with and without a logarithm transformation, and for individuals in the case and control groups. For clarity of exposition the histogram of the expenditures has been truncated at the top. On top of the histograms of the log expenditures are density functions from Normal distributions with means* $\hat{\nu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\log y_{1i}$, $\hat{\nu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\log y_{2i}$, *and variances* $\hat{\sigma}_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(\log y_{1i} - \hat{\nu}_1)^2$, $\hat{\sigma}_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(\log y_{2i} - \hat{\nu}_2)^2$, *for the case and control groups, respectively.*

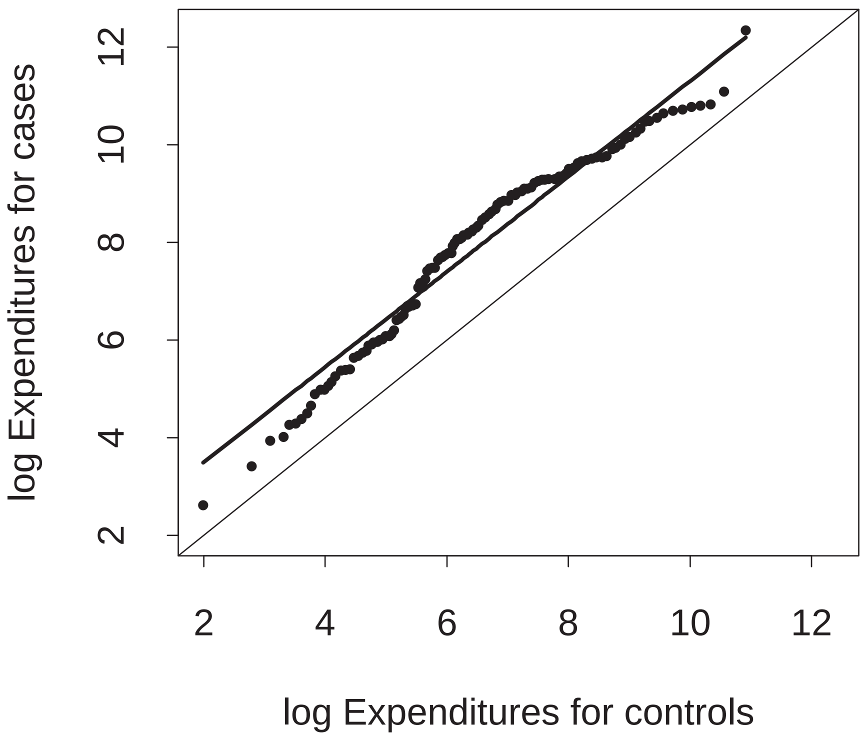

Figure 2: *Quantile-Quantile (Q-Q) plot of log non-zero Medicare expenditures for cases and controls. Solid line is the Q-Q plot assuming each sample is from a log-normal model with parameter values estimated by maximum likelihood.*

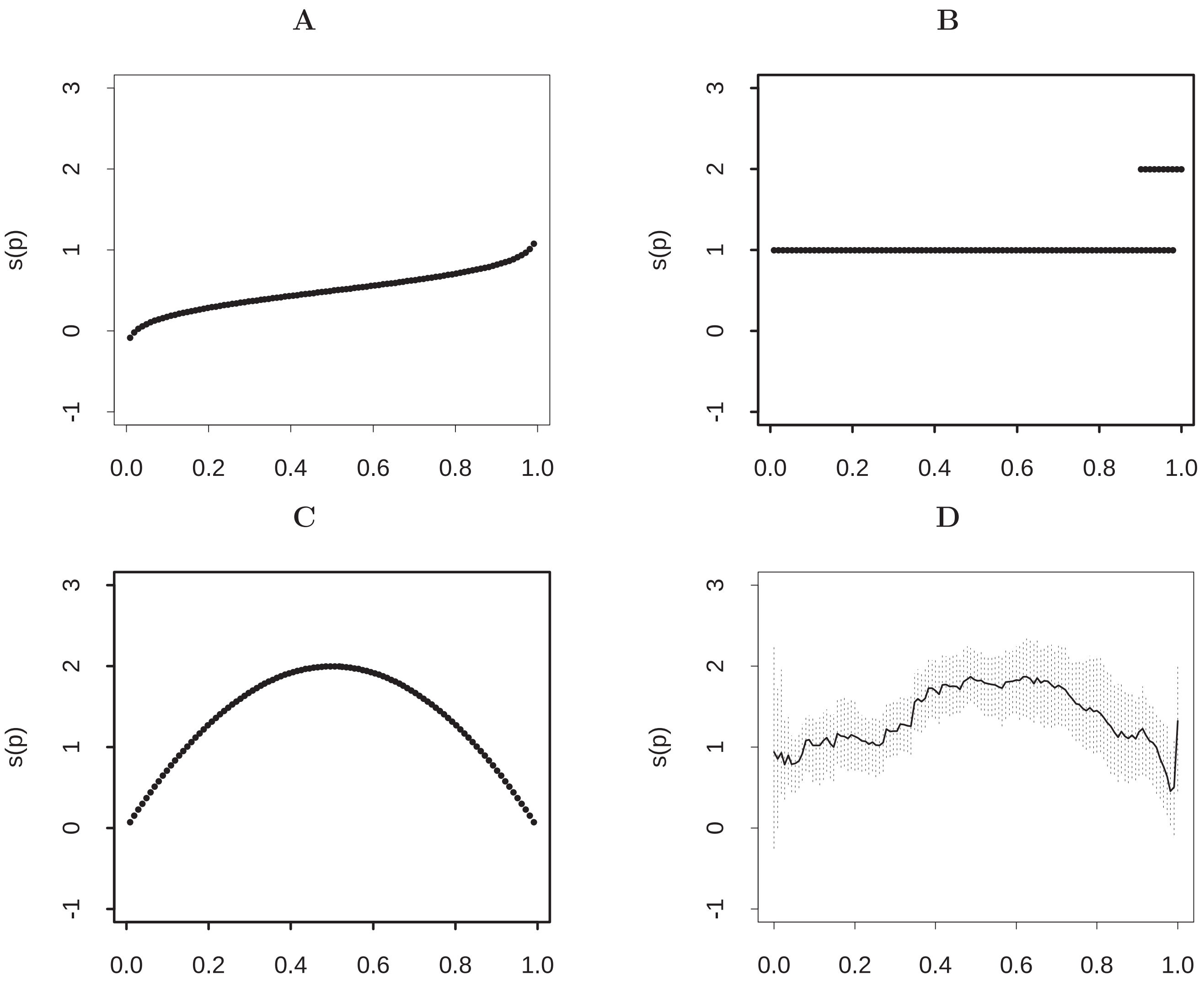

Figure 3: *Theoretical (A-C) and empirical (D) $s(p)$ curves. In scenario D, the solid curve is* $\log\left(\frac{y_{1(i)}}{q_{2(i)}}\right)$ *plotted at the percentiles* $p_{1i} = i/(n_1+1)$, $i = 1, \ldots, n_1$, *where* $q_{2(1)}, \ldots, q_{2(n_1)}$ *are the order statistics of the* $y_{21}, \ldots, y_{2n_2}$ *interpolated at percentiles* p_{1i}. *The vertical segments represent the* 95% *point-wise confidence intervals obtained for a bootstrap.*

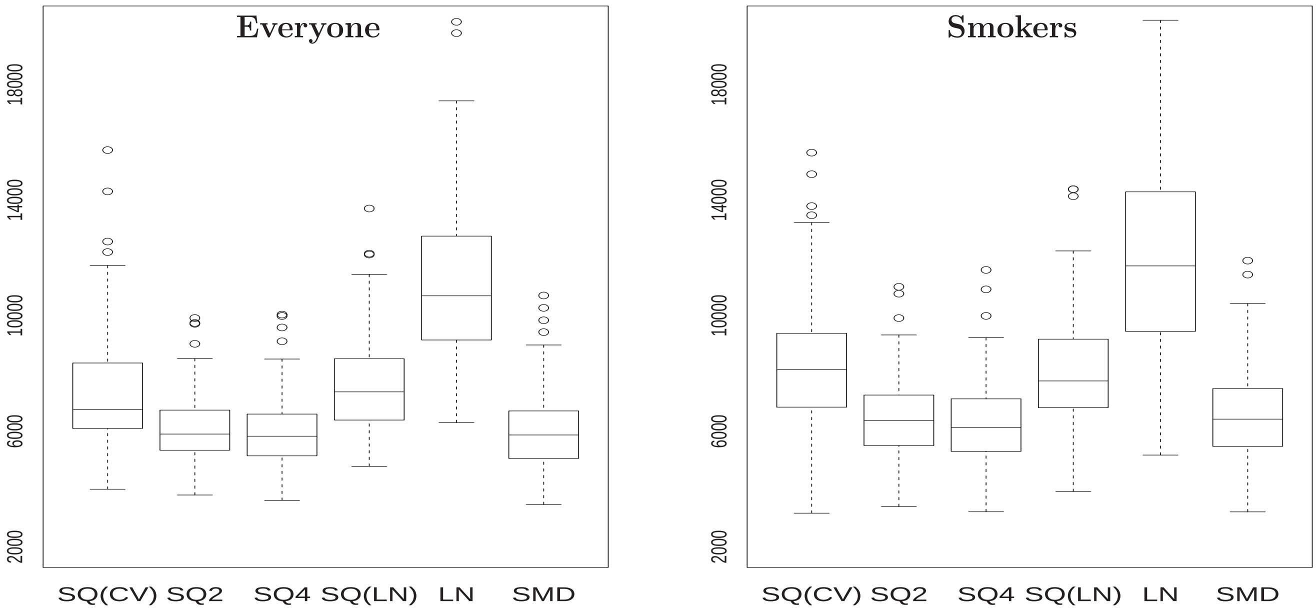

Figure 4: *Boxplots of 100 bootstrap samples of the estimated mean differences of Medicare expenditures for people with and without smoking-attributable diseases. Results are reported for everyone in the sample* ($N_1 = 188,\ N_2 = 9228$) *and for smokers only* ($N_1 = 165,\ N_2 = 4862$).