SMOOTH QUANTILE RATIO ESTIMATION WITH REGRESSION: ESTIMATING MEDICAL EXPENDITURES FOR SMOKING ATTRIBUTABLE DISEASES

Francesca Dominici and Scott L. Zeger

December 27, 2004

Abstract

The methodological development of this paper is motivated by a common problem in econometrics where we are interested in estimating the difference in the average expenditures between two populations, say with and without a disease, as a function of the covariates. For example, let Y_1 and Y_2 be two non negative random variables denoting the health expenditures for cases and controls. Smooth Quantile Ratio Estimation (SQUARE) is a novel approach for estimating $\Delta = E[Y_1] - E[Y_2]$ by smoothing across percentiles the log-transformed ratio of the two quantile functions. Dominici et al. (2004) have shown that SQUARE: defines a large class of estimators of Δ, is more efficient than common parametric and non-parametric estimators of Δ, and is consistent and asymptotically normal.

However in applications it is often desirable to estimate $\Delta(\boldsymbol{x}) = E[Y_1 \mid \boldsymbol{x}] - E[Y_2 \mid \boldsymbol{x}]$, that is the difference in means as a function of $\boldsymbol{x}$. In this paper we extend SQUARE to a regression model and we introduce a two-part regression SQUARE for estimating $\Delta(\boldsymbol{x})$ as a function of $\boldsymbol{x}$. We use the first part of the model to estimate the probability of incurring any costs, and the second part of the model to estimate the mean difference in health expenditures, given that a non-zero cost is observed. In the second part of the model, we apply the basic definition of SQUARE for positive

costs to compare expenditures for the cases and controls having "similar" covariate profiles. We determine strata of cases and control with "similar" covariate profiles by use of propensity score matching.

We then apply two-part regression SQUARE to the 1987 National Medicare Expenditure Survey to estimate the difference $\Delta(\boldsymbol{x})$ between persons suffering from smoking attributable diseases and persons without these diseases as a function of the propensity of getting the disease. Using a simulation study, we compare frequentist properties of two-part regression SQUARE with maximum likelihood estimators for the log-transformed expenditures.

KEYWORDS: Comparing means, skewed distributions, log-normal, regression splines, quantile regression, Q-Q plots, smoking, health expenditures, propensity scores. *Francesca Dominici, Scott L. Zeger, Department of Biostatistics at the Johns Hopkins University Bloomberg School of Public Health. 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. phone: 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 NIHES grant R01ES012054. We thank Timothy Wyant for providing data on the National Medical Expenditures Survey; and Elizabeth Johnson for assistance in database development and software.

1 Introduction

This paper is motivated by a common problem in health economics of estimating the difference in mean or total health expenditures between diseased and otherwise similar non-diseased persons as a function of covariates. In our motivating application, we study people affected by major smoking attributable diseases: lung cancer and chronic obstructive pulmonary diseases (COPD). The non-diseased group comprises people not affected by any of the diseases above nor by other major smoking caused illness such as cardiovascular diseases.

Let Y_1 and Y_2 be two non negative random variables representing health expenditures for the cases and controls, and let $\boldsymbol{x}$ be a vector of covariates, such as smoking, age, race, gender, and socio-economic factors. We seek to estimate the difference $\Delta(\boldsymbol{x}) = E[Y_1 \mid \boldsymbol{x}] - E[Y_2 \mid \boldsymbol{x}]$ as a function of the covariates.

Estimation of $\Delta(\boldsymbol{x})$ is challenging because health expenditures are very skewed toward high values, tend to have a high proportion of zeros, and the number of cases tends to be much smaller than the number of controls. Nevertheless $\Delta(\boldsymbol{x})$ is an important target for inference 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; Lipscomb et al., 1999). Econometric approaches for analyses of health expenditure have been discussed extensively. Among the most common approaches are linear regression models for log-transformed dependent variables and generalized linear models (GLM) with a logarithm link function (Duan, 1983; Jones, 2000; Manning, 1998; Mullahy, 1998; Blough et al., 1999). GLM estimate $\log E[Y \mid \boldsymbol{x}]$ directly, whereas the linear regression model for the log-transformed costs estimate $E[\log(Y) \mid \boldsymbol{x}]$ which can be converted into an estimate of $E[Y \mid \boldsymbol{x}]$ by a suitable transformation that involves higher moments of the distribution of $\log Y$ (Duan, 1983). See Manning and Mullahy (2001) for a simulation-based comparison of suitable estimators of $E[Y \mid \boldsymbol{x}]$ under the parametric approaches described above.

Dominici et al. (2004) have recently introduced a novel estimator of the mean difference of two highly skewed distributions $\Delta = E[Y_1] - E[Y_2]$ called Smoothed Quantile Ratio Estimation or SQUARE. Note that the most obvious non-parametric estimator of Δ is the sample mean difference $\bar{y}_1 - \bar{y}_2 = \int \widehat{Q}_1(p)dp - \int \widehat{Q}_2(p)dp$ which here is defined as a function of the empirical quantiles $\widehat{Q}_1(p), \widehat{Q}_2(p)$. The basic idea of SQUARE is to augment the empirical quantiles $\widehat{Q}_1(p)$ and $\widehat{Q}_2(p)$ with smoother and less variable versions obtained by smoothing the log-transformed ratio of the two quantile functions across percentiles, that is $\log(Q_1(p)/Q_2(p)) = s(p)$.

SQUARE encompasses a large class of estimators of Δ including the class of L-estimates (Serfling, 1980). For example if $s(p)$ interpolates the log ratios of the order statistics, then SQUARE reduces to the sample mean difference. If $s(p)$ is very smooth, then SQUARE reduces to the maximum likelihood estimate of Δ under a log-normal sampling distribution for Y_1 and Y_2 (Dominici et al., 2004; Cope, 2003). Broadly speaking, SQUARE is a semi-parametric estimate of Δ which interpolates between parametric estimates (such as maximum likelihood estimates), and non-parametric estimates (such as the sample mean difference) with weights depending on the degrees of smoothness of $s(p)$.

Simulation studies (Dominici et al., 2004; Cope, 2003) have shown that SQUARE outperforms common estimators of Δ, such as sample mean difference and log-normal estimators commonly used for the analysis of skewed data (Aitchison and Shen, 1980; Zellner, 1971; Zhou et al., 1997;

Zhou and Gao, 1997; Land, 1971; Angus, 1994; Duan et al., 1983; Zhou and Melfi, 1997; Lipscomb et al., 1999; Andersen et al., 2000). Theoretical developments of SQUARE including proofs of consistency, and asymptotic normality are detailed in Cope (2003).

In this paper we generalize SQUARE to a two-part regression model, and present a detailed example of its use in the important public health problem of estimating the difference in medical expenditures between people with and without smoking-related disease taking covariates into account. In the first part of the model, we estimate the probability of incurring any costs among the cases and the controls, $P(Y_1 > 0 \mid \boldsymbol{x})$ and $P(Y_2 > 0 \mid \boldsymbol{x})$. In the second part, we estimate the mean difference of the positive expenditures for the cases and the controls. In summary we produce an estimate of the following parameter:

$$\Delta(\boldsymbol{x}) = P(Y_1 > 0 \mid \boldsymbol{x}) \times E[Y_1 \mid Y_1 > 0, \boldsymbol{x}] - P(Y_2 > 0 \mid \boldsymbol{x}) \times E[Y_2 \mid Y_2 > 0, \boldsymbol{x}].$$

In the second part of the model we use SQUARE to compare the positive expenditures for the cases and controls having "similar" covariate profiles. We identify these homogeneous covariate groups by using propensity score matching (Rosenbaum and Rubin, 1983). The propensity score, here denoted by $e(\boldsymbol{x})$, is the probability of having a smoking related disease given the covariates: smoking dose, age, race and socio-economic factors.

For our analyses, we use the National Medical Expenditure Survey (NMES) (National Center For Health Services Research, 1987) supplemented by the Adult Self-Administred Questionnarie Household Survey (ASAQS). NMES and ASAQS provide data on annual medical expenditures, disease status, age, race, socio-economic factors, and critical information on health risk behaviors such as smoking, for a representative sample of U.S. non-institutionalized adults. A key component of our analysis is to estimate $\widehat{\Delta}(\boldsymbol{x})$ as a function of $e(\boldsymbol{x})$ and to illustrate how differences in medical expenditures might vary with respect to the propensity of having the disease.

Because SQUARE is a new idea, we compare it in a simulation study to a more standard econometric approach: two-part linear regression model for log-transformed cost. We illustrate under which sampling mechanisms two-part regression SQUARE provides a more efficient estimate of $\Delta(\boldsymbol{x})$ than parametric alternatives commonly used in analysis of health cost data.

2 Smooth quantile ratio estimation (SQUARE)

In this section we briefly review the definition of SQUARE and its estimation approach. Details are in Dominici et al. (2004) and asympotic properties and examples are in Cope (2003). Let Y_1 and Y_2 be the non negative expenditures for the cases and controls, and let Q_1 and Q_2 be the quantile functions of the random variables $Y_1 \mid Y_1 > 0$ and $Y_2 \mid Y_2 > 0$, respectively. Our goal is to estimate the difference:

$$\Delta = E[Y_1] - E[Y_2] = P(Y_1 > 0) \int_0^1 Q_1(p)dp - P(Y_2 > 0) \int_0^1 Q_2(p)dp. \tag{1}$$

The basic idea of SQUARE is to replace the empirical quantiles $\widehat{Q}_1(p)$ and $\widehat{Q}_2(p)$ with smoother and less variable versions obtained by smoothing the log-transformed ratio of the two quantile functions across percentiles. That is SQUARE estimates Δ by smoothing across percentiles the log ratio of the quantile functions:

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

By borrowing strength across the two samples to learn about the shape of the distribution, SQUARE produces an estimator of Δ that tends to be less variable than the sample mean difference but with small bias.

More specifically, let $\widehat{Q}_1, \widehat{Q}_2$ be the empirical quantile functions and let $\boldsymbol{y}_1 = (y_{1(1)}, y_{1(2)}, \cdots, y_{1(n_1)})$ and $\boldsymbol{y}_2 = (y_{2(1)}, y_{2(2)}, \cdots, y_{2(n_2)})$ be the order statistics of the positive medical expenditures for the cases and the controls respectively. If $n_1 = n_2 = n$, then SQUARE estimates Δ by the use of "smoothed" quantile functions $\widetilde{Q}_1 = \widehat{Q}_2 \exp(\widehat{s}(p, \lambda))$ and $\widetilde{Q}_2 = \widehat{Q}_1 \exp(-\widehat{s}(p, \lambda))$, where $\widehat{s}(p, \lambda)$ is obtained by fitting the model

$$\log \frac{y_{1(i)}}{y_{2(i)}} = s(p_i, \lambda) + \epsilon_i, \; i = 1, \ldots, n \tag{3}$$

with $s(p_i, \lambda) = \sum_{j=0}^{\lambda} B_j(p_i)\beta_j$, $p_i = i/(n+1)$, and where $B_j(p)$ are orthonormal basis functions, with $B_0(p) = 1$. If $n = n_1 < n_2$ as in our real application, then we replace $\boldsymbol{y}_2$ by $\boldsymbol{q}_2$, the linear interpolation of the order statistics $y_{2(i)}$ to the grid of points $p_{1i} = i/(n_1 + 1), \; i = 1, \ldots, n_1$.

Notice that 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 for the cases and controls, and let $E[Y_1 \mid Y_1 > 0]$ and $E[Y_2 \mid Y_2 > 0]$ be the corresponding averages of the non-zero values, then the SQUARE estimate of $\Delta = P(Y_1 > 0)E[Y_1 \mid Y_1 > 0] - P(Y_2 > 0)E[Y_2 \mid Y_2 > 0]$ is defined as:

$$\begin{aligned} \widehat{SQ}(\lambda) &= \widehat{\pi}_1 \times \tfrac{1}{2} \int \left[\widehat{Q}_1(p) + \widetilde{Q}_1(p)\right] dp \quad - \quad \widehat{\pi}_2 \times \tfrac{1}{2} \int \left[\widehat{Q}_2(p) + \widetilde{Q}_2(p)\right] dp \\ &= \widehat{\pi}_1 \bar{u}_1 - \widehat{\pi}_2 \bar{u}_2 \end{aligned} \tag{4}$$

where $\widehat{\pi}_1$ and $\widehat{\pi}_2$ are the proportions of non-zero costs among the cases and the controls, and $\boldsymbol{u}_1 = (\boldsymbol{y}_{(1)}, \boldsymbol{y}^\star_{(1)})$ and $\boldsymbol{u}_2 = (\boldsymbol{y}_{(2)}, \boldsymbol{y}^\star_{(2)})$ are two samples of size $2n$ where $y^\star_{1(i)} = y_{2(i)} e^{\widehat{s}_i}$, and $y^\star_{2(i)} = y_{1(i)} e^{-\widehat{s}_i}$.

Dominici et al. (2004) has shown that: 1) SQUARE is asymptotically normal; 2) it defines a large class of estimators of Δ including: the sample mean difference, the maximum likelihood estimate under log-normal samples, and L-estimates; and 3) it has lower mean squared error than several competitors including the sample mean difference, and log-normal parametric estimates in several realistic situations.

3 Regression SQUARE

In our case study we are interested in estimating the difference in medical expenditures between the cases and the controls as a function of their covariates, that is we seek to estimate

$$\Delta(\boldsymbol{x}) = E[Y_1 \mid \boldsymbol{x}] - E[Y_2 \mid, \boldsymbol{x}] = \pi_1(\boldsymbol{x}) E[Y_1 \mid Y_1 > 0, \boldsymbol{x}] - \pi_2(\boldsymbol{x}) E[Y_2 \mid Y_2 > 0, \boldsymbol{x}].$$

To extend SQUARE to the regression case we assume that the log-ratio of the quantile functions is a smooth function of the percentiles given the covariates $\boldsymbol{x}$, that is:

$$\log Q_1(p; \boldsymbol{x}) = \log Q_2(p; \boldsymbol{x}) + s(p, \lambda; \boldsymbol{x}). \tag{5}$$

To control for systematic differences in covariates between the two populations, a common strategy is to group units into sub-classes based on covariate values, and then to compare medical expenditures only for the cases and controls units who fall in the same sub-class. However, as the number of covariates increases, the number of sub-classes grows exponentially (Cochran, 1965). This problem can be overcame by matching with respect to the propensity scores (Cochran and Rubin, 1973; Rubin, 1973). The propensity score in this case can be defined as the conditional probability that an individual with vector $\boldsymbol{x}_i$ of observed covariates has the disease, $e_i(\boldsymbol{x}_i) = P(d_i = 1 \mid \boldsymbol{x}_i)$. Rosenbaum and Rubin (1983) showed that sub-classifications on the population propensity score will balance $\boldsymbol{x}$, in other words, population subgroups of cases and controls that have "similar" propensity scores, will have a similar distribution of all their covariates.

We use the propensity score matching in the definition of regression SQUARE as follows:

1. for each case $i = 1, \ldots, N_1$, we construct a stratum $[i]$ of m_1 cases and m_2 controls both with propensity scores as similar to the case i as possible, by using a nearest-neighbor matching algorithm detailed at the end of this section;

2. within each propensity score stratum $[i]$, we estimate:

   - the fractions of non-zero expenditures $\widehat{\pi}_1^{[i]}$ and $\widehat{\pi}_2^{[i]}$; and
   - the difference in average medical expenditures between the cases and the controls by applying the definition of SQUARE (Equation 4) to the m_1 cases and m_2 controls that belong to the i-th stratum, that is: $T_1(\boldsymbol{x}_i) = \widehat{SQ}(\boldsymbol{x}_i) = \widehat{\pi}_1^{[i]}\bar{u}_1^{[i]} - \widehat{\pi}_2^{[i]}\bar{u}_2^{[i]}$;

3. we estimate Δ by averaging the SQUARE estimates across the N_1 strata, that is

$$T_1 = \widehat{SQ} \quad = \quad \frac{1}{N_1}\sum_{i=1}^{N_1} \widehat{\Delta}^{[i]} \quad = \quad \frac{1}{N_1}\sum_{i=1}^{N_1}\left(\widehat{\pi}_1^{[i]}\bar{u}_1^{[i]} - \widehat{\pi}_2^{[i]}\bar{u}_2^{[i]}\right) \tag{6}$$

Matching was performed by using a modification of the nearest-neighbor matching algorithm (Rubin and Thomas, 2000), beginning with the case with lowest propensity score and proceeding to the case with highest propensity score. More specifically, let $\boldsymbol{e}_1 = (e_1(\boldsymbol{x}_1), \ldots, e_{N_1}(\boldsymbol{x}_{N_1})$ be the ordered vector of propensity scores for the cases. Then, for each case i:

1. we select the m_1 closest cases to the i-th propensity score and we identify their m_1 propensity scores;

2. we divide the m_1 propensity scores into S strata,

3. within each stratum, we sample with replacement H matched controls, thus obtaining a total of $S \times H = m_2$ matched controls.

4 Analysis of Medical Expenditures

In this section, we use two-part regression SQUARE to estimate the mean difference between annual Medicare expenditures for persons with lung cancer (LC) or chronic obstructive pulmonary disease (COPD) (cases, $d = 1$), diseases caused largely by smoking, and otherwise similar persons without these two smoking-attributable diseases and without cardiovascular disease (controls, $d = 0$).

In our problem, the propensity score $e_i(\boldsymbol{x}_i)$, is an estimate of the probability that a person i has lung cancer or COPD given his/her covariate profile $\boldsymbol{x}_i$. We estimate this risk by using the following logistic regression model (Johnson et al., 2003):

$$\begin{aligned}
\text{logit} P(d_i = 1 \mid \boldsymbol{x}_i) &= \texttt{male}_i + \texttt{afro-american}_i + \texttt{recent quit}_i + \\
&+ \texttt{poverty}_i + \texttt{marital status}_i + \texttt{census region}_i + \\
&+ \texttt{education}_i + \texttt{seat belt use}_i + \\
&+ ns(\texttt{age}_i, 3) + ns(\texttt{age}_i, 3) \times \texttt{male}_i + ns(\texttt{smoking}_i, 3)
\end{aligned} \tag{7}$$

where `male`, `afro-american`, and `recentquit` are indicators for being male, being African American, have ever smoked, and having quit smoking within one year; `poverty`, `marital status`, `education`, `census region`, and `seat belt use` are categorical variables indicating socio-economic status, place of residence, and propensity of an individual to take risks. The variable `smoking` indicates self-reported total smoking exposure (packs of cigarettes over the lifetime). We model age and smoking as natural cubic splines with 3 degrees of freedom. The full set of variables included in the model are listed in Table 1. Details of this modelling approach and results for the NMES data are given by Johnson et al. (2003).

We match the propensity scores on the logistic scale (Rubin and Thomas, 2000), with $m_1 = 50, S = 5$ and $H = 25$ leading to a $50 : 125$ matching scheme. The sensitivity of the results to the matching scheme is summarized at the end of this section.

Figure 1 shows the average logit propensity scores for cases versus the average for controls within each matched set. The proximity of the points to the diagonal line indicates excellent performance of the matching algorithm. Some deviation occurs among the highest risk subjects where the cases are at slightly higher risk than the controls. To further assess the relative success of the propensity score model for creating balanced matched samples, Table 1 compares the observed proportions for categorical covariates, and the sample means for continuous covariates between cases and controls for the matched samples. The matching appears to have performed well.

In addition to estimating the mean difference in expenditures for persons with and without disease caused by smoking, a second question is whether this difference is smaller for smokers than for non-smokers perhaps because one group has a tendency to seek or receive fewer services. That is, does smoking status modify the difference in medical expenditures between the cases and the controls? Table 2 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 (63% and 25%); this is consistent with our expectation that people with disease receive more services. These proportions are similar for smokers and non-smokers. Because of the very small number of cases among the non-smokers, we report the results for everyone in the sample and for the smokers.

We apply two-part regression SQUARE with $\lambda = 2$ to the NMES data base, and to the subset of the NMES data for smokers only. We choose $\lambda = 2$, because previous applications of SQUARE to the NMES data base (Dominici et al., 2004) have shown that $\lambda = 2$ minimizes a 10-fold cross-validation method (Efron, 1983; Breiman and Spector, 1992; Efron and Tibshirani, 1993; Shao and Tu, 1995).

Table 3 summarizes the estimated mean differences in annual Medicare expenditures for the cases and controls, with and without covariate adjustment, for everyone in the sample and for the smokers alone. We estimate Δ by using four approaches. The first is a two-part regression SQUARE (T_1) as defined in Equation 6. The second is the weighted sample mean difference within each stratum, $T_2 = \frac{1}{N_1} \sum_{i=1}^{N_1} \left(\widehat{\pi}_1^{[i]} \bar{y}_1^{[i]} - \widehat{\pi}_2^{[i]} \bar{y}_2^{[i]} \right)$. The third and the fourth are two MLE estimators of Δ under a log-normal distribution. More specifically the third, called MLE with propensity score, calculates the MLE of $\Delta(\boldsymbol{x}_i)$ within each propensity score stratum and it is defined as:

$$T_3(\boldsymbol{x}_i) = \widehat{\pi}_1^{[i]} \exp\left[\bar{ly}_1^{[i]} + \frac{1}{2}\text{var}(ly_1^{[i]})\right] - \widehat{\pi}_2^{[i]} \exp\left[\bar{ly}_2^{[i]} + \frac{1}{2}\text{var}(ly_2^{[i]})\right] \quad (8)$$

where $(\bar{ly}_1^{[i]}, \bar{ly}_2^{[i]})$ and $\text{var}(ly_1^{[i]}), \text{var}(ly_2^{[i]})$ are the sample means and variances of the log-expenditures for the m_1 cases and the m_2 controls, respectively. We then estimate Δ by averaging the estimates across the N_1 strata, that is $T_3 = \frac{1}{N_1}\sum_{i=1}^{N_1} T_3(\boldsymbol{x}_i)$.

Finally the fourth, called MLE with two-part log-linear model, provides an estimate of Δ by fitting the following model:

$$\begin{aligned} \text{logit}P(Y_i > 0 \mid d_i, \boldsymbol{x}_i) &= \beta_0 d_i + \boldsymbol{\gamma}_0 \boldsymbol{x}_i,\ i = 1, \ldots, N \\ Y_i \mid Y_i > 0, d_i, \boldsymbol{x}_i &\sim N(\beta d_i + \boldsymbol{\gamma}\boldsymbol{x}_i, \sigma^2)\ i = 1, \ldots, N - n \end{aligned} \quad (9)$$

where $N = N_1 + N_2$ and $n = n_1 + n_2$, and $\boldsymbol{x}_i$ is the i-th row of the design matrix including all the covariates specified in the propensity score model (7). Under model (9), the MLE estimate of Δ is defined as:

$$T_4 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\exp(\hat{\beta}_0 + \hat{\gamma}_0 \boldsymbol{x}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\gamma}_0 \boldsymbol{x}_i)} \exp(\hat{\beta} + \hat{\boldsymbol{\gamma}}\boldsymbol{x}_i + \frac{1}{2}\hat{\sigma}^2) - \frac{\exp(\hat{\gamma}_0 \boldsymbol{x}_i)}{1 + \exp(\hat{\gamma}_0 \boldsymbol{x}_i)} \exp(\hat{\boldsymbol{\gamma}}\boldsymbol{x}_i + \frac{1}{2}\hat{\sigma}^2)\right]. \quad (10)$$

Note that T_4 takes into account the covariates by use of a regression model instead of propensity score matching, and therefore it does not provide an estimate of $\Delta(\boldsymbol{x}_i)$ as a function of the propensity score.

When we don't adjust for the covariates, SQUARE (T_1) and the weighted sample mean difference (T_2) provides smaller estimates than the MLEs under a log-normal model (T_3 and T_4). When we adjust for the covariates, T_1, T_2 and T_4 are similar, but T_4 has a smaller bootstrap standard error suggesting greater efficiency of the MLE two-part regression model for estimating Δ. T_3 provides much larger estimate of Δ than the other methods. Finally, estimates for the smokers are larger. Frequentist properties of these estimators are studied more carefully in a simulation study presented in Section 4.1.

Figure 2 shows estimated probabilities of any cost (first row), estimated means of non-zero costs (second row), and estimated mean costs (third row) for the cases and controls plotted against propensity scores. The darker lines are the estimates for the smokers only. The grey polygon represents the 95% bootstrap confidence intervals. At the far right, we display the pooled estimates averaged across propensity scores with their 95% bootstrap confidence intervals.

We found that the estimated probabilities of any expenditure smoothly increase as the risk of disease increases. The probabilities of any cost are consistently higher for the cases than for the controls across propensity scores. In addition, at low propensity scores and for both the cases and the controls, the probability of any cost for the smokers is slightly smaller than for everyone. This may indicate that healthy smokers are more reluctant to seek for services that the rest of the population.

Average positive expenditures are larger for the cases than for the controls. At low propensity scores and for the cases, the average positive costs for the smokers are larger than everyone. This indicates that, although the smokers with low propensity of disease are more reluctant to seek for services that the rest of the population, if they do use any service, they tend to have larger medical expenditures than the rest of the population.

Figure 3 (top) shows the estimated mean differences plotted against propensity scores. As in Figure

2, the darker lines are the estimates for the smokers only. At the far right are plotted the pooled estimates across propensity scores with their 95% bootstrap confidence intervals also reported in Table 3. The shape of the distribution of the estimated mean differences is driven by the estimates of mean costs for the cases (Figure 2). We found that: 1) at the very low propensity scores, the estimated mean differences are roughly constant at approximately $3000; 2) at the moderate values of the propensity scores, the estimated mean differences are larger reaching about $9000; and 3) at the very high propensity scores the estimated mean differences drop to $4000. By examining the covariates for the cases within low, medium and high propensity score strata, we found that cases with high risk of disease tend to be older, poorer and less educated than the other cases, raising the possibility that they have poorer access to services.

Figure 3 (bottom) shows the estimated mean differences plotted against propensity scores under four alternative propensity score matching methods. These scenarios were selected after having assessed the balance on observed covariates in the matched samples, and only scenarios that assured a reasonable balance were examined in the sensitivity analysis. The scenario 50 : 125 is our baseline the other three scenarios represent more or less coarse matching samples and they were 25 : 125, 25 : 50, and 50 : 50. Pooled estimates averaged across propensity scores are very similar under the four scenarios. As expected, case-specific estimates are somewhat sensitive to the selection of the number of cases leading to less smooth curves under the scenarios 25 : 125 and 25 : 50 than under the scenarios 50 : 125 and 50 : 50. However, these differences are small and all within the case-specific confidence intervals of the baseline estimates.

4.1 Model Comparisons

In this section, we implement a simulation study where we compare frequentist properties of two-part-regression SQUARE (T_1) to the three alternative estimators of Δ used in the data analysis: 1) the weighted sample mean difference (T_2); 2) MLE with propensity score matching (T_3); and 3) MLE with two-part log-linear model regression model for the log-transform costs (T_4) (Duan, 1983; Mullahy, 1998; Mullahy and Manning, 1995).

We generate cost data under non-parametric and parametric sampling mechanisms:

A. **Sampling from the empirical distribution of the cost data:** we divide the propensity scores for the cases estimated under model (7) into 25 strata. Within each strata, first we identify the matched cases and the matched controls, and second we sample with replacement observations from the corresponding empirical distributions of the observed costs. Here we assume that the true value of Δ is equal to the weighted sample mean difference $\frac{1}{25}\sum_{j=1}^{25}\left(\widehat{\pi}_1^{[j]}\bar{y}_1^{[j]} - \widehat{\pi}_2^{[j]}\bar{y}_2^{[j]}\right)$ averaged across 1000 bootstrap samples.

B. **Sampling from a two-part linear regression model of the log-transformed costs:** we generate cost data from the same two-part log-linear model used in the data analysis and defined in Equation 9. Under this data-generating mechanism we assume that the "true" estimate of Δ is the MLE which is equal to T_4.

Note that under scenario B, the presence of heteroscedasticity implies that the log-scale prediction $E[\exp(\epsilon_i)]\exp(\beta d_i + \boldsymbol{\gamma}\boldsymbol{x}_i)$ provides a biased estimate of $E[Y_i|d_i, \boldsymbol{x}_i]$ and the bias depends on the covariates $(d_i, \boldsymbol{x}_i)$. This bias can be reduced by including an estimate of $E[\exp(\epsilon_i)|d_i, \boldsymbol{x}_i]$, called the smearing coefficient (Duan, 1983; Parmigiani et al., 1997; Andersen et al., 2000). This would lead to another estimate of Δ so defined:

$$T_5 = \text{sme} \times \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\exp(\hat{\beta}_0 + \hat{\gamma}_0 \boldsymbol{x}_i)}{1+\exp(\hat{\beta}_0 + \hat{\gamma}_0 \boldsymbol{x}_i)}\exp(\hat{\beta} + \hat{\boldsymbol{\gamma}}\boldsymbol{x}_i) - \frac{\exp(\hat{\gamma}_0 \boldsymbol{x}_i)}{1+\exp(\hat{\gamma}_0 \boldsymbol{x}_i)}\exp(\hat{\boldsymbol{\gamma}}\boldsymbol{x}_i)\right],$$

where $\text{sme} = \frac{1}{N}\sum_{i=1}^{N}\exp(r_i)$ is the so-called smearing coefficients (Duan, 1983) calculated as functions of the residuals $r_i,\ i = 1, \ldots, N$.

In summary two-part regression SQUARE and the weighted sample mean difference (T_1, T_2) are non-parametric estimates of $\Delta(\boldsymbol{x}_i)$ within each propensity score stratum, whereas T_3 uses a MLE of $\Delta(\boldsymbol{x}_i)$ also within each propensity score stratum. The last two estimators T_4, T_5 are not based on propensity score matching and therefore they cannot estimate $\Delta(\boldsymbol{x}_i)$ as a function of the propensity score. On the other end they can estimate Δ already marginalized with respect to the covariates.

Scenario A differs substantially from scenarios B: scenario A favors propensity score matching and non-parametric estimation methods, scenario B favors model-based estimation approaches. The results are summarized in Table 4. In Scenario A, if the goal is to estimate $\Delta(\boldsymbol{x}_i)$ then T_1 and T_2 outperform T_3. This result suggests that MLE with propensity score matching is terribly inefficient when the data are not log-normal and the sample size is small. If the goal is to estimate Δ, then T_1, T_2, T_4 and T_5 have similar performance.

In Scenario B, if the goal is to estimate $\Delta(\boldsymbol{x}_i)$, then two-part regression SQUARE (T_1) outperforms both the weighted sample mean difference T_2 and the MLE with propensity score matching T_3. It is interesting to note that T_1 is better than T_3 in estimating $\Delta(\boldsymbol{x}_i)$ even when the data are log-normal. If the goal is to estimate Δ, then by definition T_4 is the best and it is similar to T_5. However T_1 is the second best.

In summary these simulation studies suggest that two-part regression SQUARE produces the most efficient estimate of $\Delta(\boldsymbol{x}_i)$ with respect to non-parametric and parametric alternatives. If the goal if to estimate Δ marginalized with respect to the covariates then SQUARE can be more variable than the MLE obtained under a two-part log-linear model.

5 Discussion

In this paper, we have extended SQUARE, a novel estimator of the difference in means for two right-skewed distributions, to the regression case. The premise of SQUARE is to model the log ratio of the two quantile functions as a smooth function of the percentiles producing an estimator that is less variable than the difference in sample means and nearly unbiased in many practical situations. SQUARE is a novel approach that uses a non-parametric estimate of the quantile function from the larger sample, and a parametric model for the log quantile ratio $s(p)$. Additional details on the theoretical development of SQUARE with its software implementation are available at in (Dominici et al., 2004).

The development of SQUARE and its extension to the regression case was motivated by the estimation of smoking attributable expenditures. A key component of this analysis is the estimation of the mean difference of Medicare-financed medical expenditures between persons with smoking attributable-diseases (lung cancer or COPD) and otherwise similar persons without these diseases. To address this substantive question we created an estimator of the difference of means of two highly skewed distributions that borrows strength across the two samples. This idea has applications in a variety of setting. For example SQUARE can be applied to estimate the mean of a single sample by borrowing strength from a theoretical distribution such as the log-normal. In addition, SQUARE can be used to compare multiple groups, where each group borrows strength from a “referent” sample, which can be one of the samples or an average of all samples.

To control for possible imbalances in the observed covariates, we proposed an extension of SQUARE to the regression case. Here we use a variation of propensity score matching (Rosenbaum and Rubin, 1983, 1984) and estimate differences in mean expenditures for the cases and controls within strata of propensity scores. Our analysis of Medicare expenditures allows smoking status to modify the effect of disease on expenditures. We examine this effect modification by stratifying the cases and the controls with respect to their smoking status, and then by estimating SQUARE separately for smokers and all subjects. In addition, our plots of the estimated mean differences as a function of the propensity scores allow detection of effect modification by variables that are important predictors of disease. For example the visual inspection of Figure 3 suggests that the estimated mean differences drop from 9000$ to 4000$ for large propensity scores. We found that these individuals tend to be older, poorer and less educated than the others, suggesting the the hypothesis that they have poorer access to services.

In the application we use a fixed smoothing parameter λ that does not vary with the propensity scores. We select $\lambda = 2$ because in previous analyses (Dominici et al., 2004) we found that this choice minimizes a 10-fold cross-validation method (Efron, 1983; Breiman and Spector, 1992; Efron and Tibshirani, 1993; Shao and Tu, 1995). Although simulation studies have shown that the frequentist properties of SQUARE are robust to the choice of λ, for λ small, generalizations of our approach might include methods for estimating a degree of smoothness that also vary with the propensity scores.

Our formulation of SQUARE in the regression case is related to quantile regression (Ruppert and Carroll, 1980; Koenker, 1982; Lifson and Bhattacharyya, 1983). Regression SQUARE is a two step procedure: 1) we first estimate the difference in medical expenditures in a $[0,1] \times [0,1]$ grid of points of the values of propensity scores and percentiles; and 2) we then estimate the parameter of interest $\Delta(\boldsymbol{x})$ at each value of the propensity score by smoothing across percentiles (see Figures 2 and 3). In quantile regression, we estimate the parameter of interest as a function of the covariates for a fixed percentile.

In the simulation study, we showed that under non-parametric and parametric sampling mechanisms, two-part regression SQUARE outperforms the MLE with propensity score matching when estimating $\Delta(\boldsymbol{x}_i)$, and it is comparable to the MLE with a two-part log-linear model when estimating Δ. In summary we found that two-part regression SQUARE is the most suitable approach for estimating $\Delta(\boldsymbol{x}_i)$ as a function of the propensity score. However, if the ultimate goal is to estimate Δ marginalized with respect to the covariates, then standard regression adjustment provides a more efficient estimate of Δ than propensity score matching. As future work, our simulation study can be extended to compare two-part regression SQUARE with respect to the more general GLM framework (McCullagh and Nelder, 1989) with an exponential conditional mean which include Poisson, Gamma, Weibull, and Chi-square structures (Manning and Mullahy, 2001).

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.

Blough, D., Madden, C., and Hornbrook, M. (1999). "Modelling Risk using generalized linear models." *Journal of Health Economics*, 18, 153–171.

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

Cochran, W. G. (1965). "The Planning of Observational Studies of Human Populations (with Discussion)." *Journal of the Royal Statistical Society, Series A, General*, 128, 234–266.

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.

Dominici, F., Cope, L., Naiman, D., and Zeger, S. L. (2004). "Smooth Quantile Ratio Estimation (SQUARE)." *Biometrika*, ?, ?–?

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.

Johnson, E., Dominici, F., Griswold, M., and Zeger, S. L. (2003). "Disease cases and their medical costs attribuitable to smoking: an analysis of the national medical expenditure survey." *Journal of Econometrics*, 112, 135–151.

Jones, A. (2000). *Health Econometrics*. Culyer, A. and Newhouse, J. (Eds): Handbook of Health Economics, Elsevier, Amsterdam.

Koenker, R. (1982). "Robust Methods in Econometrics." *Econometric Reviews*, 1, 213–255.

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

Lifson, D. P. and Bhattacharyya, B. B. (1983). "Quantile Regression Method and Its Application to Estimate the Parameters of Lognormal and Other Distributions." In *Contributions to Statistics: Essays in Honour of Norman L. Johnson*, 313–327. North-Holland/Elsevier (Amsterdam; New York).

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.

Manning, W. (1998). "The logged dependent variable: heteroschedasticity and the transformation problme." *Journal of Health Economics*, 17, 283–295.

Manning, W. G. and Mullahy, J. (2001). "Estimating log models: to transform or not to transform." *Journal of Health Economics*, 20, 461–494.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (Second Edition)*. Chapman & Hall.

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.

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.

Parmigiani, G., Samsa, G., Ancukiewicz, M., Lipscomb, J., Hasselblad, V., and Matchar, D. (1997). "Assessing Uncertainty in Cost-Effectiveness Analyses: Application to a Complex Decision Model." *Medical Decision Making*, 17, 390–401.

Rosenbaum, P. and Rubin, D. (1983). "The Central Role of Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association*, 79, 516–524.

Rubin, D. and Thomas, N. (2000). "Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association*, 95, 573–585.

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

Ruppert, D. and Carroll, R. J. (1980). “Trimmed Least Squares Estimation in the Linear Model.” *Journal of the American Statistical Association*, 75, 828–838.

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

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

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.

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: *Matching variables used in the generalized additive model to estimate the propensity scores. Comparison between the the distributions of each matching variable to check balance in the matched samples of cases and controls.*

Variable	Cases	Controls
gender		
female	0.39	0.41
male	0.61	0.59
race		
Other	0.97	0.94
African American	0.03	0.06
poverty		
Poor	0.14	0.15
Near Poor	0.10	0.09
Low Income	0.21	0.20
Middle Income	0.30	0.28
High Income	0.24	0.28
marital status		
Married	0.63	0.64
Separated	0.24	0.23
Divorced	0.10	0.10
Widowed	0.01	0.03
Never Married	0.01	0.01
census region		
Northeast	0.20	0.21
Midwest	0.28	0.27
South	0.35	0.35
West	0.17	0.17
education		
4+ Years of College	0.07	0.09
1-3 Years of College	0.09	0.09
Some/All High School	0.53	0.50
Less than High School	0.31	0.33
seat belt use		
Seldom/Never	0.30	0.29
Sometimes	0.16	0.17
Nearly Always/Always	0.54	0.52
recent quit		
current smoker	0.93	0.92
former smoker who quit within one year	0.07	0.08
age	69	69
smoking	45	48

Table 2: *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	Total
cases	165 (62%)	23 (70%)	188 (63%)
controls	4682 (21%)	4546 (28%)	9228 (25%)
	4847 (22%)	4569 (28%)	9416 (25%)

Table 3: *Unadjusted and covariate-adjusted 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 ($N_1 = 165,\ N_2 = 4862$). Bootstrap standard errors are in parentheses.*

	Unadjusted		Adjusted	
	Everyone	Smokers	Everyone	Smokers
two-part regression SQUARE (T_1)	5863 (1220)	6342 (1411)	5313 (1748)	5695 (1807)
weighted sample mean difference (T_2)	5751 (1341)	6308 (1540)	5631 (1834)	6217 (2011)
MLE with propensity score (T_3)	10735 (2882)	11468 (3061)	11383 (5819)	12065 (5719)
MLE with two-part log-normal model (T_4)	9721 (1956)	9289 (2245)	5763 (1534)	4527 (1193)

Table 4: *Results of the simulation study: average, standard deviation, and mean square error of the estimates across 500 simulated data sets.*

SCENARIO A			
Estimator	Mean	Standard Deviation	MSE/1000
Two-part regression SQUARE (T_1)	5534	38	2.1
Weighted sample mean difference (T_2)	5842	40	2.6
MLE with propensity score (T_3)	9235	198	1553
MLE with two-part log-linear model (T_4)	5295	36	1.8
Common smearing (T_5)	4631	33	2.3
True	5660		
SCENARIO B			
Estimator	Mean	Standard Deviation	MSE/1000
Two-part regression SQUARE (T_1)	7599	61	19
Weighted sample mean difference (T_2)	8917	78	51
MLE with propensity score (T_3)	10691	75	59
MLE with two-part log-linear model (T_4)	5634	37	2
Common smearing (T_5)	5495	39	2.3
True	5316		

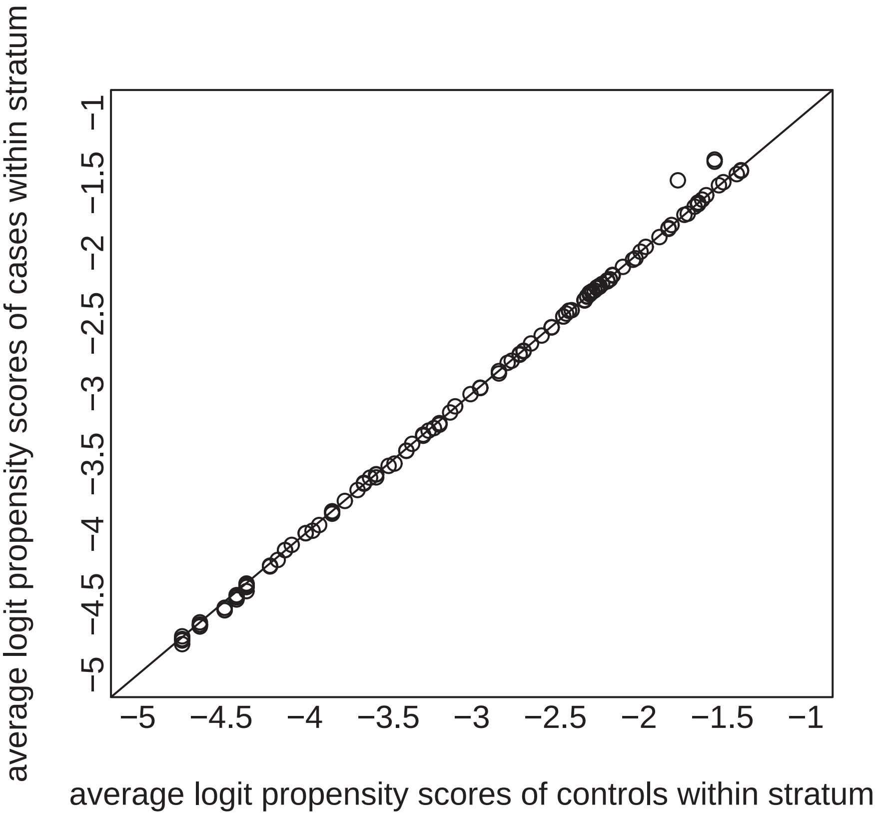


Figure 1: *Propensity score averages of the* $m_1 = 25$ *matching cases plotted against averages of the* 125 *matched controls.*

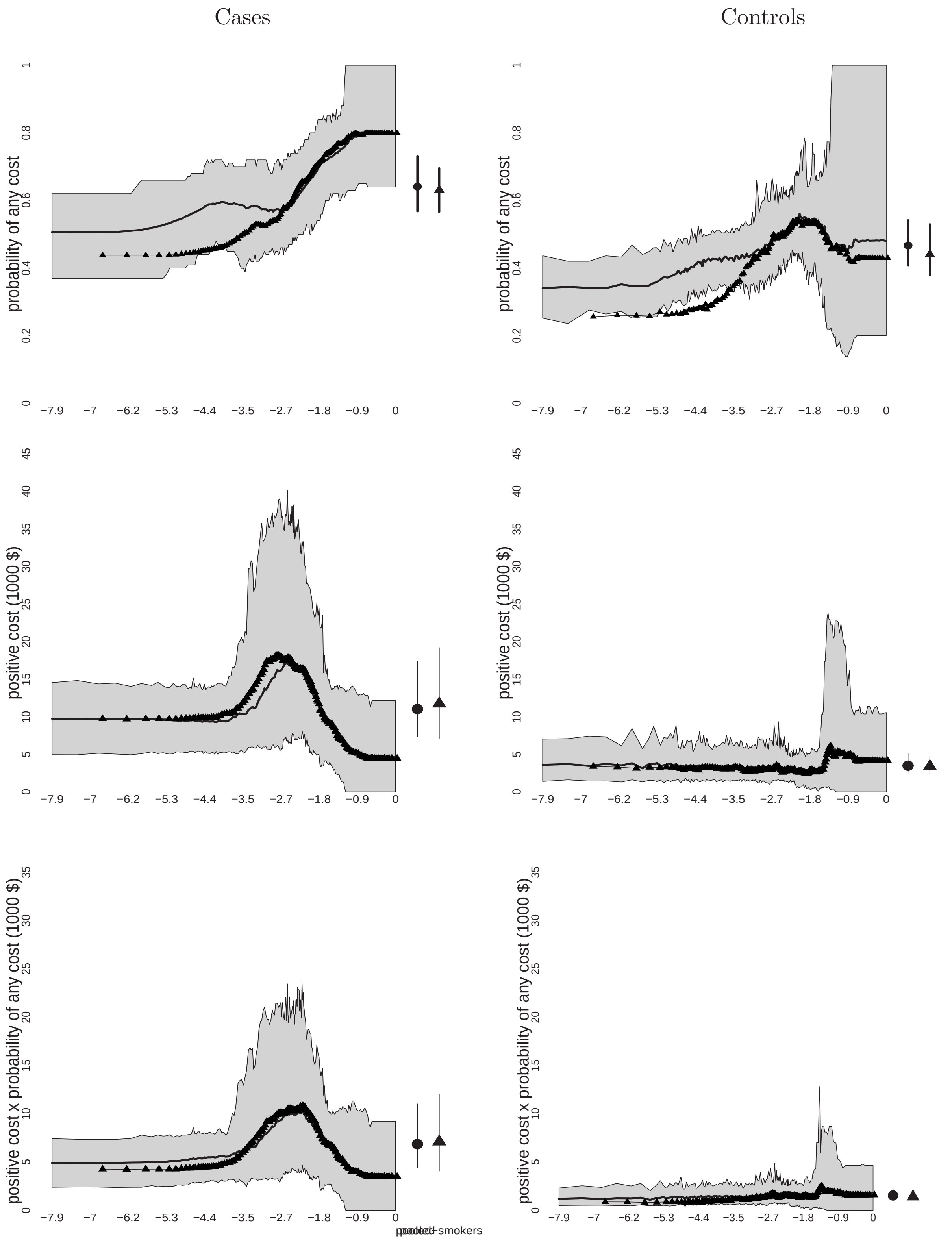
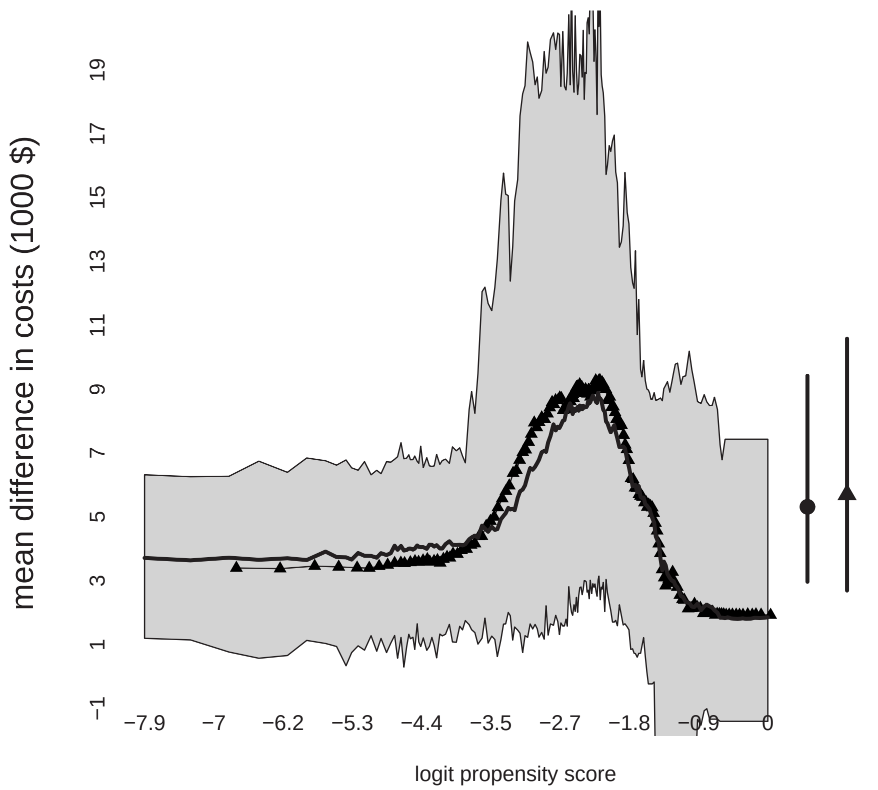


Figure 2: *Estimated probabilities of any Medicare expenditures (first row), estimated mean non-zero expenditures (second row), and estimated mean expenditures (third row) for the cases (left), and controls (right) plotted against propensity scores. The solid and dotted lines are the estimates for everyone and for smokers only, respectively. The polygon represents the* 95% *bootstrap confidence intervals for everyone. At the far right are plotted the estimates pooled across propensity scores with their* 95% *bootstrap confidence intervals.*



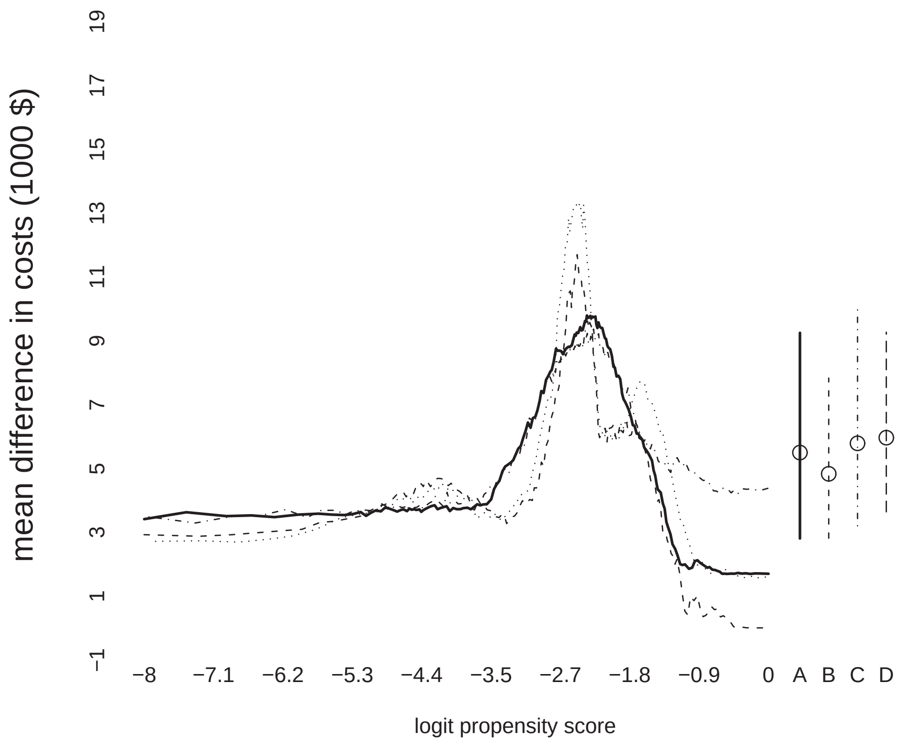


Figure 3: *Top: Two-part regression SQUARE estimates of* $\Delta(\boldsymbol{x}_i)$ *plotted as a function of* $e(\boldsymbol{x}_i)$*. Solid and dotted lines represent the estimates for everyone and for smokers only, respectively. Vertical segments represents the* 95% *bootstrap confidence intervals for everyone. At the far right are shown the SQUARE estimates pooled across propensity scores for everyone and for the smokers only. Bottom: Two-part regression SQUARE estimates of* $\Delta(\boldsymbol{x}_i)$ *as a function of* $e(\boldsymbol{x}_i)$ *under scenarios A-D of strata size selection used in the propensity score matching algorithm corresponding to 50:125, 25:125, 25:50, 50:50.*