Propensity Score and Heckman Adjustments for Treatment Selection Bias in Database Studies

Robert L. Obenchain and Catherine A. Melfi
Health Services and Policy Research, Eli Lilly and Company

KEY WORDS Preliminary probit or logit model for treatment selection; Cochran-Rosenbaum-Rubin propensity score binning; Murnane-Newstead-Olsen generalization of Heckman model; significance of a linear combination; Scheffe′ maximum F-projection.

1. Introduction

We describe and compare two statistical methods, useful in retrospective studies, that adjust for non-randomization of experimental units (patients) to treatment cohorts as well as for any differences in observed covariate values. Our analyses start with a preliminary probit (or logit) model that predicts the probability of treatment selection for any given vector of covariates, x. Second stage analyses then use either the "propensity-score binning" approach of Rosenbaum and Rubin(1984) or else a parametric modelling approach derived from seminal work by Heckman(1979) in labor economics.

Here, we will keep our descriptions of the statistical models and their estimation rather brief, but a great deal of material is covered. Literature references are also given in case readers wish to examine more complete and/or more philosophical discussions of fundamental distinctions between these two approaches.

To compare and contrast the two methods, we use an example that illustrates the sorts of complexity that analysts typically have to cope with when applying each approach.

2. Variables in Retrospective Studies

Observed response variables will be denoted here by the symbol y. Typical response variables are, say, measures of health care costs per year or indicators of pharmacotherapy effectiveness. Effectiveness indicators may be discrete (6 consecutive months on pharmacotherapy? : 1=yes, 0=no) or continuous (% cholesterol reduction.)

Observed covariates will be denoted by the symbol x. Typical covariates are patient age, gender, indicators of type and/or total number of comorbidities, and other measures of demand (historical or potential) for health-care services.

Unobserved (or unobservable) "latent" variables will be denoted by the symbol z. Latent variables typically play a key role in the theory of models for retrospective studies because important clinical measures, e.g. severity of illness, are not routinely recorded in current U.S. administrative databases.

The observed "treatment selection" indicator variable will be denoted here by the hybrid symbol, z^*. (The "star" superscript distinguishes this indicator from the unobserved z variables that actually determine treatment choice by the patient and/or doctor.) This indicator is typically discrete (taking on only 2 levels in the example considered here), and its parameters (determining the expected value of its probability distribution within each level) are to be modelled using linear functions of only the observed covariates, x.

3. Preliminary Probit/Logit Model

The first step is to fit a model to the observed z^* that estimates the conditional probability of treatment selection given x [SAS(1989),Greene(1993,1995)]. Specifically, a probit model with **linear predictor** $\zeta_i = x_i'\theta$ uses the relationship $\Phi(\zeta_i) =$ [probability that a normally distributed random variable with mean $x_i'\theta$ and variance 1 is non-negative]. Figure 1 illustrates this fundamental identity which connects the numerical sign ($<$ or ≥ 0) of some unobserved z variable with the resulting treatment selection, z^*. In other words, $\Phi(\zeta_i)$ represents the probability that the i-th patient receives the new treatment ($z_i^* = 1$) rather than the standard treatment ($z_i^* = 0$), where $\Phi()$ denotes the cumulative distribution function of the standardized normal distribution (mean zero and variance 1.) See section 5 for more detail on the econometric "truncation" interpretation of the first stage model.

4. Propensity-Score "Binning"

Rosenbaum and Rubin(1983,1984) proposed a relatively simple and easy to appreciate method of reducing potential bias due to lack of randomization that is based on some early

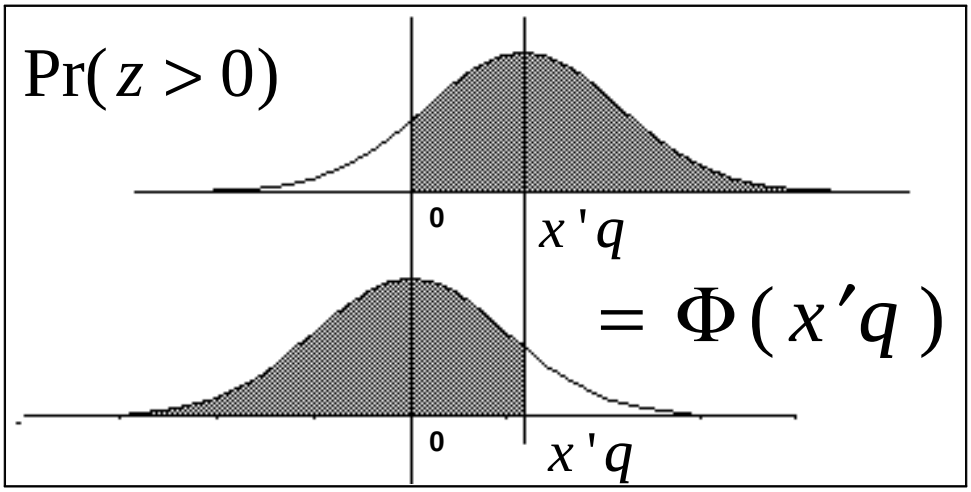

Figure 1. Here we see why the probability that a latent z variable with mean $x^0\mu$ is positive (and hence that $z^¤ = 1$ rather than $z^¤ = 0$) equals the probability that a standardized normal deviate (mean zero and variance one) is less than $x^0\mu$. Technically, the …gure illustrates only the case where $x^0\mu$ is positive (and © > 0:5), but the corresponding illustration when $x^0\mu$ is negative is quite similar.

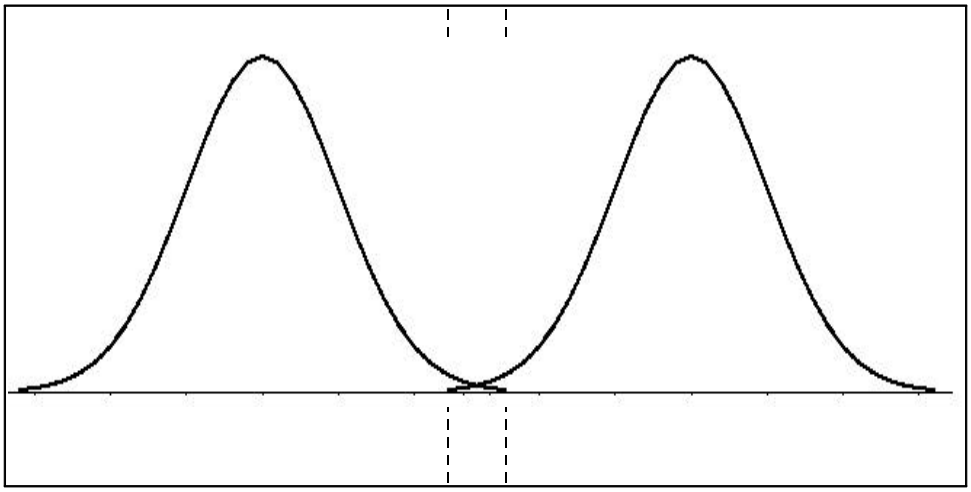

Figure 2. In propensity binning analysis, one's preliminary model for prediction of treatment selection can actually end up being "too good!" If insu¢cent overlap between patients selecting the di¤erent treatments is observed, then these two populations are really not compable.

work by Cochran(1968). Apparently, the basic "statistical theory" underlying this methodology is contained in the following "observation" about the joint distribution of x and z^* conditional upon a given "propensity score" defined to be $e = e(x) = Pr(\, z^* = 1 \,|\, x\,)$. This reasoning argues that

$$\begin{aligned} Pr(\, x,\, z^* \,|\, e\,) &= Pr(\, x \,|\, e\,)\, Pr(\, z^* \,|\, x,\, e\,) \\ &= Pr(\, x \,|\, e\,)\, Pr(\, z^* \,|\, x\,) \\ &= Pr(\, x \,|\, e\,) \text{ times } e \text{ or } (1-e) \\ &= Pr(\, x \,|\, e\,)\, Pr(\, z^* \,|\, e\,) \qquad (1) \end{aligned}$$

In other words, x (observed patient characteristics) and z^* (observed treatment selection) have **conditionally independent** distributions given their computed numerical value on propensity score, e.

A patient's estimated propensity score is simply his/her predicted value from the above sort of preliminary probit model, $\widehat{e} = \Phi\left(\widehat{\zeta}\right)$. No two patients may have the exact same numerical propensity score, but patients who are "nearest-neighbors" in terms of propensity scores are indeed well matched in the following sense. These neighbors will either have quite similar observed characteristics (x vectors) or else their characteristics will vary only in ways that do not lead to big differences in z^* prediction.

Steps in Propensity Score Binning

1. Model the probability of $z^* = 1$ or 0 given x (logit or probit.)

2. Sort all observations on estimated $e(x) = Pr(\, z^* = 1 \,|\, x)$.

3. Form 5 (or more) relatively homogeneous subgroups of patients with similar estimated propensity scores with bin boundaries at order statistic quintiles (20%, 40%, 60% and 80%) ...or at finer divisions.

4. Calculate average differences in response [$\overline{Y}$ for $z^* = 1$ patients minus $\overline{Y}$ for $z^* = 0$ patients] within each of the 5 (or more) bins.

5. Compute an overall **weighted** average response difference, using weights inversely proportional to the observed variances of within bin differences.

Attempts to perform this sort of analysis can fail, primarily, in step 4 above. Specifically, one's preliminary z^* model (step 1) may be "too good," as is illustrated in Figure 2. Here the linear predictor, $\zeta_i = x_i'\theta$, has done such a good job of separating the patient populations receiving the two different treatments that there is virtually no overlap between them (for the sorting and binning operations in steps 2 and 3.) After all, no within-bin average cost difference can be calculated in step 4 when a bin contains patients of only one type! Here, the proper conclusion is that the patient populations represented by the two treatment cohorts really are not comparable. They may well have different expected costs, but there is no way to adjust these cost differences for the vast observed (and unobserved) differences between these patients on their other (x and z) characteristics.

Instead, what you really hope to find in propensity scoring is a preliminary z^* model that detects only modest separation between treatment cohorts, as is illustrated in Figure 3. You may have to set aside the data for a very few patients with outlying ζ scores, but the vast majority of patients can be grouped into 5 (to 8) bins, each of which contains a good mix of patients of both treatment types.

5. Econometric Models

In this approach, second stage analysis requires a parametric model, related to that of Heckman(1979), that contains adjustments for treatment differences, covariates and treatment selection bias. This sort of model is commonly written

$$y_i = \mu + z^*\alpha + x_i'\beta + \lambda(\zeta_i, z_i^*)\gamma + \epsilon_i \qquad (2)$$

where y_i is the observed response (effectiveness or cost) for the i-th patient; z_i^* is again the zero-one treatment indicator variable [$z_i^* = 0$ for the "standard" treatment, $z_i^* = 1$ for the "new" treatment]; x_i' is a $(1 \times p)$ vector of (non-constant) covariate values for the i-th patient; $\lambda(\zeta_i, z_i^*)$ is Heckman's normal-theory "inverse Mill's ratio" term that is a non-linear function of the first-stage linear predictor $\zeta_i = x_i'\theta$ and the observed treatment selection z_i^* of the form $\lambda(\zeta, z^*) = z^*\varphi(\zeta)/[1 - \Phi(\zeta)] - (1 - z^*)\varphi(\zeta)/\Phi(\zeta)$; and ϵ_i is the error term. The linear model coefficients to be estimated are μ, α, β and γ.

Technically, inclusion of the $\lambda(\zeta, z^*)$ term in the above model is due to a normal-theory "incidental truncation" argument. In fact, this assumption dictates that $\gamma = \rho\sigma_y$, where σ_y is the response standard deviation and ρ is the

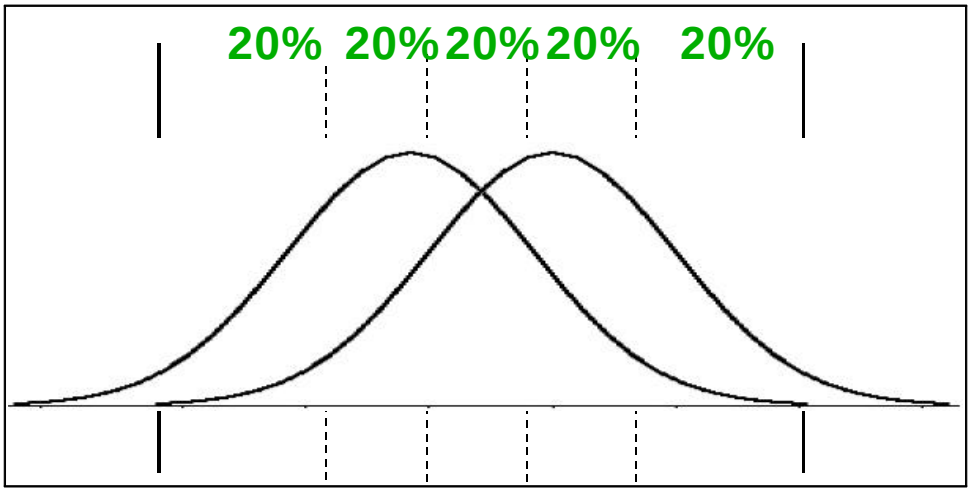

Figure 3. As illustrated here, considerable overlap between patient populations selecting the di¤erent treatments is actually highly desirable for propensity score binning. One needs to divide the "shared" region into approximately 5 parts, each with about the same total number of patients.

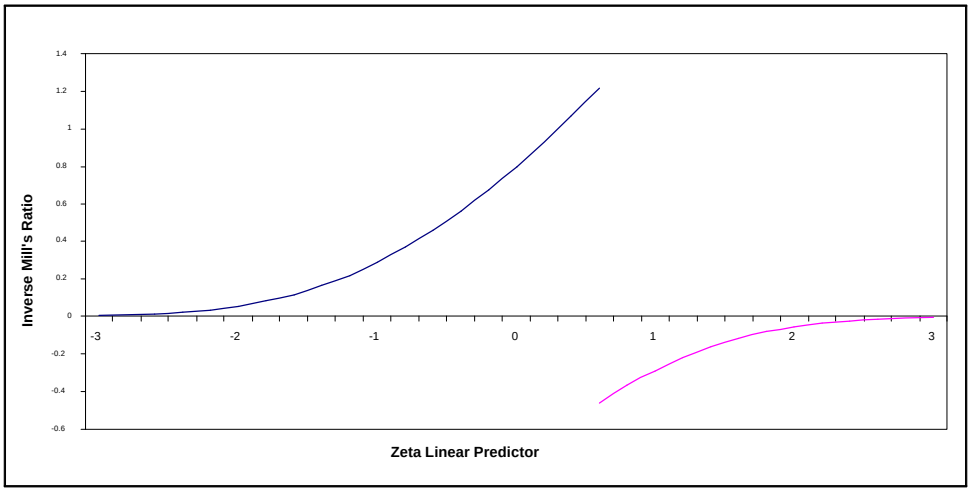

Figure 4. In parametric modelling, the inverse Mill's ratio "instrument," ¸, takes on a particularly simple form in the extreme case where prediction of treatment selection is exact. Here we see this limiting case when all patients to the left of $^{3}_{0}$ = 0:6 are in one treatment cohort while all patients to the right of $^{3}_{0}$ = 0:6 are in the other cohort.

correlation between the response and an unobservable, normal-theory **latent** variable z with mean $\zeta = x'\theta$ and variance 1; see Greene(1993), Theorem 22.4, page 707. Again, this latent z determines treatment cohort membership via the rule: $z^* = 0$ when $z < 0$, and $z^* = 1$ and when $z \geq 0$. Obviously, z must be correlated with x in order for the first-stage model to estimate the mean of z and, hence, the probability that $z^* = 1$. But y must also be correlated with z to make the λ term useful in predicting the mean of y in the second-stage model.

The form of the nonlinear $\lambda(\zeta, z^*)$ term is of particular interest in **two limiting, special cases**. For example, when the best linear predictor, $\zeta_i = x_i'\theta$, in the preliminary probit or logit model turns out to be $\widehat{\zeta}_i = 0$ ($\widehat{\theta} = 0$), then prediction of treatment selection has totally failed, and $\lambda(0, z^*) = \pm\varphi(0)/[\Phi(0)]$ takes on only two different values, $\pm\sqrt{2/\pi} = \pm 0.7979$, because $\Phi(0) = 1 - \Phi(0) = 1/2$. In other words, choice of treatment then appears as if it had occurred simply by coin-flip, and the $\lambda(\zeta_i, z_i^*)\gamma$ term in the model is equivalent to a simple "treatment effect" step function like $(z^* - 0.5)\gamma$. Estimation obviously becomes difficult in this limit because the $\lambda(\zeta_i, z_i^*)\gamma$ and $z^*\alpha$ terms then become highly colinear.

At the opposite extreme, which was depicted in Figure 2 for the propensity score approach, $\zeta = x'\theta$ leads to extremely accurate predictions of treatment selection. Again, there is some "cut-point" numerical value for ζ such that all "new" treatment assignments, $z^* = 1$, tend to occur when (say) $\zeta \leq \zeta_0$ while all "standard" treatment assignments, $z^* = 0$, tend to occur when $\zeta > \zeta_0$. In other words, to the left of ζ_0, $\lambda(\zeta, z^*)$ is almost always $\lambda(\zeta, 1) = \varphi(\zeta)/[1 - \Phi(\zeta)]$ and, to the right of ζ_0, $\lambda(\zeta, z^*)$ is almost always $\lambda(\zeta, 0) = -\varphi(\zeta)/\Phi(\zeta)$. This situation is depicted in Figure 4, which looks somewhat like a "tilted" step function. The $\lambda(\zeta_i, z_i^*)\gamma$ and $z^*\alpha$ terms are also somewhat colinear in this limit.

In the simple model of equation (2), evidence for true treatment differences after adjustment for selection bias is most clearly indicated when the α estimate is significantly different from zero. After all, predictions that "eliminate" treatment selection bias can be viewed as resulting from setting $\lambda = 0$ in equation (2). But some economists apparently recommend making predictions using the fitted values for the λ instrumental variable, in which case the fitted γ coefficient impacts the inference. When the γ coefficient is either insignificant or else increases estimated separation between cohort means, the α coefficient is still the key estimate. On the other hand, evidence for only "weak" true treatment differences after adjustment for selection bias is provided when the α and γ coefficients are both significant but tend to cancel each other out!

5.1 Time Sequencing of Information

Econometric approaches also specifically recognize and incorporate time sequencing information by placing restrictions on use of covariates. Specifically, only covariates, x_1, that measure patient characteristics prior to treatment selection (e.g. previous comorbidities, demography, geography) may be used in the preliminary probit or logit model. Covariates, x_2, that measure patient characteristics that develop during the study period (e.g. usage patterns and current comorbidities) are used only in the second stage, along with the λ instrumental variable (nonlinear function of x_1) and, perhaps, some of the original x_1 covariates.

5.2 Generalized Heckman Models

We now consider the Murnane, Newstead and Olsen(1985) generalization of the econometric model of equation (2). This generalization allows the μ and β regression coefficients for covariate adjustment to differ between treatment cohorts. Specifically, the $p+1$ regression parameters $(\mu, \beta_1, \ldots, \beta_p)'$ will be replaced by $2(p+1)$ parameters of the form $z_i^*\mu_1 + (1 - z_i^*)\mu_0$ and $z_i^*\beta_1 + (1 - z_i^*)\beta_0$ where, again, $z_i^* = 0$ for the "standard" treatment and $z_i^* = 1$ for the "new" treatment. In other words, this generalization contains simultaneous equations of two types:

$$y_{1i} = \mu_1 + x_i'\beta_1 + \lambda(\zeta_i, 1)\gamma_1 + \epsilon_{1i} \qquad (3)$$

when $z_i^* = 1$ or

$$y_{0i} = \mu_0 + x_i'\beta_0 + \lambda(\zeta_i, 0)\gamma_0 + \epsilon_{0i} \qquad (4)$$

when $z_i^* = 0$. The two γ coefficients are, ideally, constrained to have the same numerical value, $\gamma_1 = \gamma_0 = \gamma$; this equality is dictated by the normal-theory "incidental truncation" assumption described above. However, $\gamma_0 \neq \gamma_1$ is another generalization discussed

(without much motivation) in econometric literature; see Dolton and Makepeace(1987).

Two important implications of the assumptions that yield the $\lambda(\zeta, z)$ terms in the above models are that (i) observations are heteroscedastic and (ii) common regression estimation procedures, such as ordinary least squares, may not be unbiased or consistent. To account for these problems, we use the LIMDEP maximum likelihood algorithm of Greene(1995) in our analyses of econometric models for the case study in section 7. LIMDEP allows $\gamma_0 \neq \gamma_1$in equations (3) and (4).

Our experience is that the Murnane, Newstead and Olsen(1985) generalization of the Heckman-type model of equation (2) is quite useful in actual practice. Specifically, equation (2) dictates a potentially unrealistic sort of "parallelism" between health care outcomes in the two treatment cohorts. After all, except for the z^* step-function and nonlinear $\lambda(\zeta_i, z_i)$ terms, equation (2) dictates that outcomes for the two cohorts have identical dependence on x. The question is: "What will now constitute evidence for true treatment differences after adjustment for selection bias?"

6. Testing for True Treatment Differences

To simplify notation, let x and β now denote $(p+1) \times 1$ vectors that include initial intercept elements, 1 or μ, as well as the original p covariates or coefficients. The difference in predicted response due to treatment at a given set of patient characteristics, x, can then be written in the form

$$\widehat{y}_1 - \widehat{y}_0 = x' \left(\widehat{\beta}_1 - \widehat{\beta}_0 \right) \tag{5}$$

with estimated variance

$$Var\left[\widehat{y}_1 - \widehat{y}_0\right] = x' \left(V_1 + V_0\right) x \ , \tag{6}$$

where V_1 and V_0 are the $(p+1) \times (p+1)$ estimated variance-covariance matrices for $\widehat{\beta}_1$and $\widehat{\beta}_0$, respectively.

Murnane, Newstead and Olsen(1985) suggest that evidence for treatment differences, after adjustment for covariates and selection bias via the above generalized model, is provided by testing the statistical significance of the $\widehat{y}_1 - \widehat{y}_0$ difference at either $x = \overline{x}_1$ or $x = \overline{x}_0$, which are the observed covariate centroids of the two treatment cohorts. Unfortunately, these two choices generally lead to different p-values! This raises questions such as, "Which value should one report?" and even "How influential is one's choice of x on the resulting statistical inference?"

Here, we assume that $(\widehat{\beta}_1 - \widehat{\beta}_0) \neq 0$; otherwise, no evidence for treatment differences emerges by estimating the above models. Similarly, note that an estimated $\widehat{y}_1 - \widehat{y}_0$ difference of zero also results when x is any vector orthogonal to $\widehat{\beta}_1 - \widehat{\beta}_0$. [Technically, the first element of the x yielding $x'(\widehat{\beta}_1 - \widehat{\beta}_0) = 0$ must initially be nonzero so that it can be made to equal $+1$ when each element of x is then divided by this first element.]

Here, we really wish to answer questions like "What choice for x maximizes the t-statistic for the predicted treatment difference?" and "How does one test the significance of this maximum-t statistic?" These are the topics of the next two subsections.

6.1 Derivation of Maximum-t

Consider the Lagrange equation

$$\begin{aligned} \psi(x) &= x' \left(\widehat{\beta}_1 - \widehat{\beta}_0 \right) - \\ &\quad \eta \left[x' \left(V_1 + V_0\right) x - \rho^2 + \nu^2 \right] \ , \end{aligned} \tag{7}$$

with multiplier variable η, slack variable ν, and a strictly positive constant ρ^2. Then $\partial\psi/\partial\nu = 2\nu\eta = 0$ implies

$$\eta = 0 \text{ or } \nu = 0 \ , \tag{8}$$

while $\partial\psi/\partial\eta = 0$ implies

$$x' \left(V_1 + V_0\right) x = \rho^2 - \nu^2 \leq \rho^2. \tag{9}$$

Next, note that

$$\partial\psi/\partial x = \left(\widehat{\beta}_1 - \widehat{\beta}_0 \right) - 2\eta \left(V_1 + V_0\right) x \tag{10}$$

and that

$$\partial^2\psi/\partial x^2 = -2\eta \left(V_1 + V_0\right), \tag{11}$$

which is negative definite when $\eta > 0$. As a result, $\partial\psi/\partial x = 0$ implies that the treatment difference with **maximum** potential t-statistic occurs when x is proportional to

$$x^{\max} = \left[V_1 + V_0\right]^{-1} \left[\widehat{\beta}_1 - \widehat{\beta}_0 \right] \ , \tag{12}$$

assuming, again, that the first element of $x^{\max}$ is nonzero. In other words, $x = k \times x^{\max}$,

where k is any non-zero constant, implies an observed difference of

$$k \times \left[\widehat{\beta}_1 - \widehat{\beta}_0\right]' [V_1 + V_0]^{-1} \left[\widehat{\beta}_1 - \widehat{\beta}_0\right] \ , \quad (13)$$

and a corresponding standard deviation of

$$|k| \times \sqrt{\left[\widehat{\beta}_1 - \widehat{\beta}_0\right]' [V_1 + V_0]^{-1} \left[\widehat{\beta}_1 - \widehat{\beta}_0\right]} \ . \quad (14)$$

Thus the resulting maximum t-ratio (estimated difference divided by its standard deviation) via choice of x is

$$\pm\sqrt{\left[\widehat{\beta}_1 - \widehat{\beta}_0\right]' [V_1 + V_0]^{-1} \left[\widehat{\beta}_1 - \widehat{\beta}_0\right]} \ . \quad (15)$$

When the first element of $x^{\max}$ is estimated to be zero, one could drop intercept terms from the maximum-t computation, so that both $\widehat{\beta}_1$ and $\widehat{\beta}_0$ have only p elements, and then recalculate $x^{\max}$ using equation (12).

6.2 Statistical Significance of Maximum-t

The statistical significance of the above maximum-t is most appropriately tested using a Scheffe′ projection argument, as outlined in section 2.2 of Miller(1980). The resulting critical value is thus of the form $\sqrt{(p+1)\,F}$ instead of the "usual" Student's $t-$statistic critical value, where $t = \sqrt{F}$ when $p = 0$ (i.e. intercept only; no non-constant covariates.) Here we write $F = F(\alpha; p+1, n-p-1)$ to denote the upper $100(1-\alpha)\%$ point of Snedecor's $F-$distribution with $p+1$ degrees-of-freedom in the numerator and $(n-p-1)$ degrees-of-freedom in the denominator. Similarly, $t = t(\alpha/2; n-p-1)$ denotes the upper $100(1-\alpha/2)\%$ point of Student's $t-$statistic with $(n-p-1)$ degrees-of-freedom.

Here are some numerical values for $\alpha = 0.05$ critical points in cases where the degrees-of-freedom-for-error are either $(n-p-1) = 120$ or else very large, $(n-p-1) = \infty$. When $p = 0$, $F^{0.05}_{(1,120)} = 3.92$ while $t^{0.025} = \sqrt{F} = 1.98$ is simply its square root. But when $p = 4$, $F^{0.05}_{(5,120)} = 2.29$ while $\sqrt{5 \cdot F} = 3.38$ is a somewhat larger value. Similarly, in the very large degrees-of-freedom limit, $F^{0.05}_{(5,\infty)} = 2.21$ while $\sqrt{5 \cdot F} = 3.32$. When $p = 29$, $F^{0.05}_{(30,120)} = 1.55$ while $\sqrt{30 \cdot F} = 6.82$ is a much larger value. Similarly, in the very large degrees-of-freedom limit, $F^{0.05}_{(30,\infty)} = 1.46$ while $\sqrt{30 \cdot F} = 6.62$.

7. Case Study Example

We analyze data from a retrospective study of the cost and effectiveness of antidepressant pharmacotherapy; Crown(1998) and Hylan(1998) give much more background information on studies using the MarketScan® database. Here we use data from 3443 patients who suffered a episode of major depression within the years 1990-1994. All of these patients were continuously enrolled for at least 18 months, including both a 6 month antidepressant free prior period and a 12 month minimum follow-up period. The treatment cohorts compared here consist of 2,410 patients whose initial antidepressant pharmacotherapy was a selective serotonin reuptake inhibitor (SSRI: fluoxetine, paroxetine or sertraline) versus 1033 patients whose initial drug came from a relatively inexpensive, older antidepressant class called tricyclics (TCAs.)

Five patients were eliminated during preliminary analyses because their follow-up health care costs (ranging from \$249K to \$662K per year) were such wild outliers that they were skewing the cost distribution even on the log(\$/year) scale. All 5 of these eliminated patients were from the SSRI treatment cohort, but 4 of them had incurred \$13K to \$90K in total charges during the 6 months just before the decision to treat their depression with an SSRI was made. These patients apparently suffered from chronic diseases (such as cancer, severe cardiovascular disease, and diabetes with complications) were comorbid depression is common; our findings will not apply directly to this extremely ill population. Thirteen of the remaining 3,438 patients (6 on SSRIs and 7 on TCAs) incurred follow-up costs ranging from \$100K to \$137K per year. These patients were retained in our analyses because their costs helped make the log(\$/year) distribution more nearly symmetric.

The observed average yearly health care costs for the remaining 2,405 patients who initiated pharmacotherapy with an SSRI was $\$6,828$/year, while the corresponding average for the 1033 patients who initiated pharmacotherapy with a TCA was $\$8,062$/year. The resulting raw (unadjusted) difference in yearly health care costs was thus $-\$1,234$/year, with SSRI patients experiencing lower average costs than TCA patients. (The corresponding median and 90% points were $\$3,773$/year

and $14,676/year for SSRI patients versus $4,212/year and $17,429/year for TCA patients; SSRI patients incurred consistently lower follow-up costs than TCA patients.)

Here we wish to adjust observed cost differences for any between cohort differences on a relatively large battery of observed covariates that measure patient demography, geography, and comorbidity. In all, we wish to adjust for observed differences on 47 x variables as well as for any latent z variables correlated with these 47 covariates.

Our preliminary probit model was highly significant, statistically, but patient reclassification rules of the form [predict SSRI when (estimated probability of SSRI) $> \pi$; TCA, otherwise] misclassify many (at least 1000) patients. For example, the $\pi = 0.5$ rule misclassifies 60 of 2,405 SSRI patients and 948 of 1033 TCA patients; $\pi = 0.6645$ misclassifies 615 SSRI patients and 615 TCA patients. In other words, our first-stage model is not a very accurate predictor of treatment selection.

Our choice of propensity scoring resulted in 5 bins, each containing 687 or 688 patients. As is clear from Figure 5, this binning exercise went very well, much like the ideal case depicted in Figure 3. The corresponding within bin comparisons of average yearly health care costs between the SSRI and TCA cohorts are displayed in Figure 6. The figure clearly shows that average yearly costs tend to decrease from left to right across the 5 bins; MarketScan® patients with relatively high costs were also relatively more likely to have initiated pharmacotherapy for depression with a TCA in the years 1990-1994.

On the other hand, Figure 6 also clearly shows that SSRI patients incur lower yearly costs than TCA patients within 3 of the 5 bins. What is not clear from Figure 6 is that **variability** in yearly health care costs also tends to dramatically decrease from left to right across the 5 bins. In any case, the overall, weighted difference in average yearly costs (SSRI minus TCA; $z^* = 1$ minus $z^* = 0$) is "only" $-$239/year when adjusted for between cohort differences on all 47 covariates, which is almost $1,000/year less than the raw (unadjusted) difference of $-$1,234/year.

Our attempts at parametric econometric modelling with these same data produced much less clear-cut results. Due to low R-squares like 0.4, predictions from log-cost models display drastic regression towards their mean; this makes it essential to use substantial multiplicative "smearing" factors, Duan(1983), when re-expressing results back on the $/year scale.

Yet another complication is that, compared with the two-stage approach outlined above, we got strikingly different numerical θ, β and γ estimates when we ran the maximum likelihood algorithm of LIMDEP that simultaneously estimates both the treatment-selection probit (with 28 x_1 covariates) and the log-cost equation (with 31 x_2 covariates.) Simultaneous estimation does cause the observed significance levels for α ($p = 0.00037$) and γ ($p = 0.00053$) in equation (2) to decrease to $p = 0.00002$. Interestingly, simultaneous estimation causes the $z^*\alpha$ and $\lambda(\zeta_i, z_i^*)\gamma$ terms to switch from cancelling each other out to re-enforcing each other, yielding "curious" average predictions (median like, before smearing) of $3,849/year for SSRI patients and $1,986/year for TCA patients. This implies a quite "misleading" cost difference estimate of +$1,863/year before smearing!

The +$1,986/year prediction for the TCA cohort is unreasonably low because the treatment "parallelism" restriction of equation (2) is quite unrealistic for these data. In fact, we find in our Murnane, Newstead and Olsen generalized models, (3) and (4), that the (μ_1, β_1) and (μ_0, β_0) cohort estimates are distinctly different and the γ_1 and γ_0 estimates have opposite signs! Furthermore, numerical differences between two-stage and simultaneous equations estimates are even more striking. Simultaneous estimation again decreases significance levels for the γ terms in equations (3) and (4), from $p = 0.008$ to $p = 0.0000$ within the SSRI cohort and from $p = 0.03$ to $p = 0.01$ within the TCA cohort. Similarly, simultaneous estimation increases the maximum-t of equation (15) from 5.9 to 36.3, where 6.6 is the appropriate Scheffe′ projection $p = 0.05$ significance value. In other words, these more reasonable econometric models definitely suggest that cost differences between the SSRI and TCA cohorts do remain after covariate adjustment.

Unfortunately, there are many different ways to make econometric predictions using fitted values for equations like (3) and (4). For example, one straight-forward approach is to make cost predictions at both treatment cohort centroids. These predictions (using zeroed-out λ instruments) suggest that $254/year savings

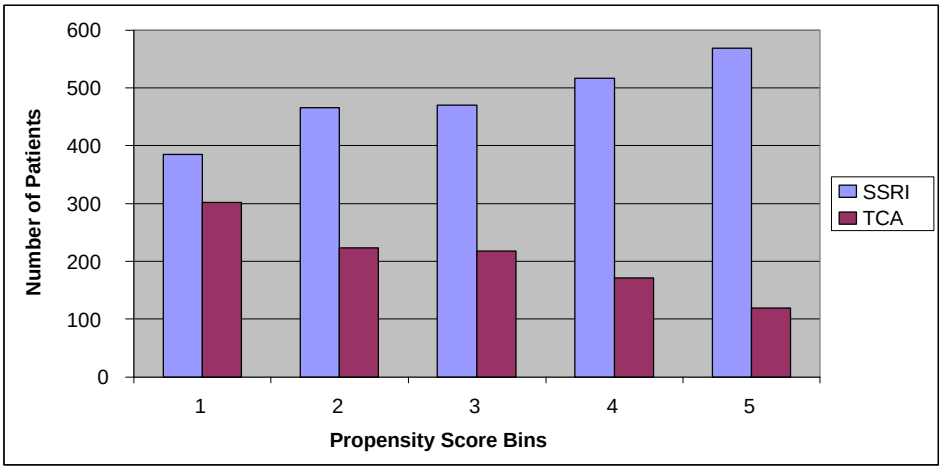

Figure 5. Each propensity score bin contains 687 or 688 patients. Almost 70% of these patients with a "new" depression episode in the years 1990-1994 started pharmacotherapy on a SSRI rather than on a TCA. However, note that the likelihood of starting with a TCA decreases dramatically from left to right across the 5 bins.

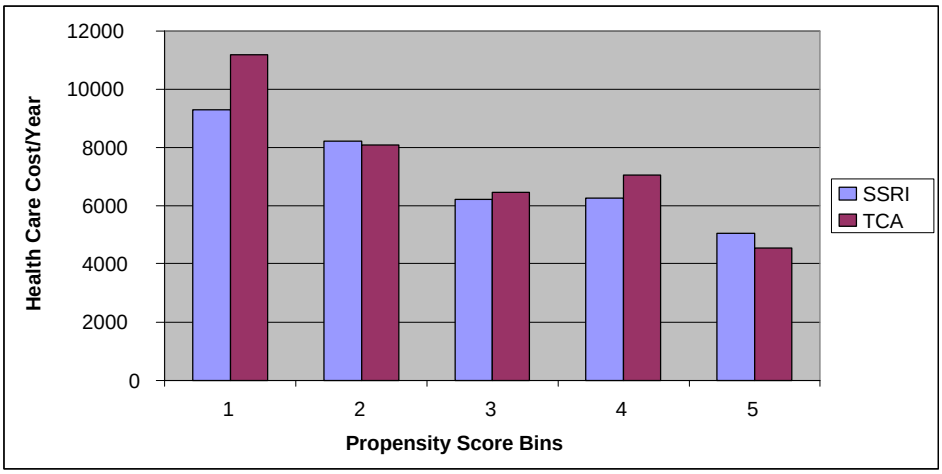

Figure 6. Note that average yearly health care costs tend to decrease from left to right across the 5 propensity score bins. But average costs for SSRI patients are lower than those of TCA patients in 3 of the 5 bins. Thus the overall weighted average di¤erence in costs (SSRI minus TCA) is negative, ¡ $239=year.

would result from switching the most typical (high cost) TCA patients to SSRIs. On the other hand, a savings of $13/year would result from switching the most typical (low cost) SSRI patients to TCAs.

8. Summary

We generally prefer to use propensity-score methods because they are much easier to understand and to explain than econometric methods. The results of a propensity-score analysis can be summarized quite effectively using only a pair of histograms, such as those in Figures 5 and 6.

Our experience is that econometric models using simultaneous equations and/or instrumental variable analyses quickly become not only quite complex but also frustratingly sensitive to the validity/reasonableness of underlying model assumptions. Reporting the detail necessary to describe such an econometric model typically requires large tables of little real interest to results-oriented readers. However, the maximum-t testing procedure proposed here does appear to be more reasonable (intuitively and theoretically) than previous methods of testing for true treatment differences after adjustment for both covariates and potential (unknown) selection bias.

REFERENCES

Angrist, J. D. (1997). "Conditional independence in sample selection models." **Economic Letters** 54, 103-112.

Cochran, W. G. (1968). "The effectiveness of adjustment by subclassification in removing bias in observational studies." **Biometrics** 24: 205-213.

Crown W., Lair T. J., Engelhart L., et. al. (1998). "The application of sample selection models to outcomes research: the case of evaluating the effects of antidepressant therapy on resource utilization." to appear in **Statistics in Medicine.**

Dolton P. J. and Makepeace G. H. (1987). "Interpreting sample selection effects." **Econometric Letters** 24, 373-379.

Duan, N. (1983). "Smearing estimate: a nonparametric retransformation method." **Journal of the American Statistical Association** 78, 605-610.

Greene, W. H. (1993). **Economic Analysis, Second Edition.** (Chapter 22, Limited Dependent Variable and Duration Models.) Englewood Cliffs, NJ: Prentice Hall.

Greene, W. H. (1995). **LIMDEP User's Manual, Version 7.** Bellport, NY: Econometric Software, Inc.

Heckman, J. J. (1979). "Sample selection bias as a specification error." **Econometrica** 47: 153-161.

Heckman, J. J. and Robb, R. (1986). "Alternative methods for solving the problem of selection bias in evaluating the impact of treatment on outcomes." In H. Wainer (ed.) **Drawing Inferences from Self-Selected Samples**. Berlin: Springer-Verlag, pp. 63-107.

Hylan, T. R., Crown, W. H., Meneades, L.,et al. (1998). "TCA and SSRI antidepressant selection and health care costs in the naturalistic setting: a multivariate analysis." to appear in **Journal of Affective Disorders**.

Manning, W. G., Duan, N. and Rogers, W. H. (1987). "Monte Carlo evidence on the choice between sample selection and two-part models." **Journal of Econometrics** 35: 59-82.

Miller, R. G. Jr. (1980). **Simultaneous Statistical Inference**, Second Edition. New York, NY: Springer-Verlag.

Murnane, R. J., Newstead, S. and Olsen, R. J. (1985). "Comparing public and private schools: the puzzling role of selectivity bias." **Journal of Business and Economic Statistics** 3: 23-35.

Rosenbaum, P. R. and Rubin, D. B. (1983). "The central role of the 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. B. (1996). "Causal inferences using observational data." Presentation at **Sixth Biennial Regenstrief Institute Conference** on "Measuring Quality, Outcomes, and Cost of Care Using Large Databases." Indiana University Medical Center: Regenstrief Health Center.

SAS Institute Inc. (1989). **SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2.** [Chapter 35: The PROBIT Procedure, and Chapter 36: The REG Procedure]. Cary, NC: SAS Institute Inc.