Journal of Data Science **4**(2006), 67-91

A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data

Elaine L. Zanutto
University of Pennsylvania,

Abstract: We extend propensity score methodology to incorporate survey weights from complex survey data and compare the use of multiple linear regression and propensity score analysis to estimate treatment effects in observational data from a complex survey. For illustration, we use these two methods to estimate the effect of gender on information technology (IT) salaries. In our analysis, both methods agree on the size and statistical significance of the overall gender salary gaps in the United States in four different IT occupations after controlling for educational and job-related covariates. Each method, however, has its own advantages which are discussed. We also show that it is important to incorporate the survey design in both linear regression and propensity score analysis. Ignoring the survey weights affects the estimates of population-level effects substantially in our analysis.

Key words: Complex survey data, information technology careers, multiple linear regression, propensity scores, salary, gender gap, SESTAT.

1. Introduction

We compare the use of multiple linear regression and propensity score analysis to estimate treatment effects in observational data arising from a complex survey. To do this, we extend propensity score methodology to incorporate survey weights from complex survey data. Multiple linear regression is a commonly used technique for estimating treatment effects in observational data, however, the statistical literature suggests that propensity score analysis has several advantages over multiple linear regression (Hill, Reiter, and Zanutto, 2004; Perkins, Tu, Underhill, Zhou, and Murray, 2000; Rubin, 1997) and is becoming more prevalent, for example, in public policy and epidemiologic research (e.g., D'Agostino, 1998; Dehejia and Wahba, 1999; Hornik *et al.*, 2002; Perkins et al., 2000; Rosenbaum, 1986; Rubin, 1997). Propensity score analysis techniques use observational data to create groups of treated and control units that have similar covariate values so that subsequent comparisons, made within these matched groups, are not confounded by differences in covariate distributions. These groups are formed by

matching on the estimated propensity score, which is the estimated probability of receiving treatment given background covariates

For illustration, we use these two methods to estimate the effect of gender on information technology (IT) salaries. Although we may not consider the effect of gender on salary to be a treatment effect in the causal sense, because we cannot manipulate gender (Holland, 1986), both propensity score and linear regression methods can be used to make descriptive comparisons of the salaries of similar men and women. We estimate gender gaps in IT salaries using data from the U.S. National Science Foundation's 1997 SESTAT (Scientists and Engineers Statistical Data System) database (NSF 99-337). Because SESTAT data is obtained using a complex sampling design, we extend propensity score methodology to incorporate survey weights from complex survey data.

The outline of the remainder of this paper follows. Multiple linear regression and propensity score methodologies are summarized in Sections 2 and 3, with a discussion of the necessary modifications to both methods to accommodate complex survey data in Section 4. The results of our data analysis are described in Section 5, with a discussion of the relative advantages of each of the methods in Section 6. Section 7 concludes with an overall discussion.

2. Multiple Linear Regression

Multiple linear regression can be used to estimate treatment effects in observational data by regressing the outcome on the covariates, including an indicator variable for treatment status and interactions between the treatment variable and each of the covariates. A statistically significant coefficient of treatment or statistically significant coefficient of an interaction involving the treatment variable indicates a treatment effect. This is the most common method, for example, for estimating gender salary gaps after controlling for important covariates such as education, experience, job responsibilities and other market factors such as region of the country (Finkelstein and Levin, 2001; Gastwirth, 1993; Gray, 1993).

3. Propensity Score Methodology

As an alternative to multiple linear regression, a propensity score analysis of observational data (Rosenbaum and Rubin, 1983, 1984; Rubin, 1997) can be used to create groups of treated and control units that have similar characteristics so that comparisons can be made within these matched groups. The propensity score is defined as the conditional probability of receiving treatment given a set of observed covariates. The propensity score is a balancing score, meaning that conditional on the propensity score the distributions of the observed covariates are independent of the binary treatment assignment (Rosenbaum and Rubin, 1983;

1984). As a result, subclassifying or matching on the propensity score makes it possible to estimate treatment effects, controlling for covariates, because within subclasses that are homogeneous in the propensity score, the distributions of the covariates are the same for treated and control units (e.g., are "balanced"). In particular, for a specific value of the propensity score, the difference between the treated and control means for all units with that value of the propensity score is an unbiased estimate of the average treatment effect at that propensity score, assuming the conditional independence between treatment assignment and potential outcomes given the observed covariates ("strongly ignorable treatment assignment" assumption) (Rosenbaum and Rubin, 1983). In other words, unbiased treatment effect estimates are obtained when we have controlled for all relevant covariates, which is similar to the assumption of no omitted-variable bias in linear regression.

Unlike other propensity score applications (D'Agostino, 1998; Rosenbaum and Rubin, 1984; Rubin, 1997), when estimating the effect of gender on salary we cannot imagine that given similar background characteristics the treatment (gender) was randomly assigned. Nevertheless, we can use the propensity score framework to create groups of men and women who share similar background characteristics to facilitate descriptive comparisons.

The estimated propensity scores can be used to subclassify the sample into strata according to propensity score quantiles, usually quintiles (Rosenbaum and Rubin, 1984). Strata boundaries can be based on the values of the propensity scores for both groups combined or for the treated or control group alone (D'Agostino, 1998). To estimate gender salary gaps in IT, since we are interested in estimating gender salary gaps for women and since there are many fewer women than men, we create strata based on the estimated propensity scores for women, so that each stratum contains an equal number of women. This ensures an adequate number of women in each stratum. As an alternative to subclassification, individual men and women can be matched using estimated propensity scores (Rosenbaum, 2002, chapter 10) however, it is less clear in this case how to incorporate the survey weights from a complex survey design and so we do not use this approach here.

To estimate the average difference in outcomes between treated and control units, using propensity score subclassification, we calculate the average difference in outcomes within each propensity score stratum and then average these differences across all five strata. In the case of estimating average IT salary differences, this is summarized by the following formula:

$$\Delta_1 = \sum_{k=1}^{5} \frac{n_{Fk}}{N_F}(\bar{y}_{Mk} - \bar{y}_{Fk}) \tag{3.1}$$

where Δ_1 is the estimated overall gender difference in salaries, k indexes the propensity score stratum, n_{Fk} is the number of women (treated units) in propensity score stratum k (the total sample size in stratum k is used here if quintiles are based on the treated and control units combined), $N_{Fk} = \sum_k n_{Fk}$, and $\bar{y}_{Mk}$ and $\bar{y}_{Fk}$, respectively, are the average salary for men (control units) and women (treated units) within propensity score stratum k. The estimated standard error of this estimated difference is commonly calculated as (Benjamin, 2003; Larsen, 1999; Perkins *et al.* 2000)

$$\hat{s}(\Delta_1) = \sqrt{\sum_{k=1}^{5} \frac{n_{Fk}^2}{N_F^2} \left(\frac{s_{Mk}^2}{n_{Mk}} + \frac{s_{Fk}^2}{n_{Fk}} \right)} \tag{3.2}$$

where n_{Mk} and n_{Fk} are the number of men and women, respectively, in stratum k, and s_{Mk}^2 and s_{Fk}^2 are the sample variances of salary for men and women, respectively, in stratum k. This standard error estimate is only approximate for several reasons (Du, 1998). It does not account for the fact that since the subclassification is based on propensity scores estimated from the data, the responses within each stratum and between the strata are not independent. Also, the stratum boundary cut-points are sample-dependent and so are the subsequent sample sizes, n_{Mk} and n_{Fk}. However, previous studies (Agodini and Dynarski, 2001; Benjamin, 2003) have found this standard error estimate to be a reasonable approximation.

Simple diagnostic tests can be used to assess the degree of covariate balance achieved by the propensity subclassification (Rosenbaum and Rubin, 1984). If differences between the two groups remain after subclassification, the propensity score model should be re-estimated including interaction or quadratic terms of variables that remain out of balance. If differences remain after repeated modeling attempts, regression adjustments can be used at the final stage to adjust for remaining covariate differences (Dehejia and Wahba, 1999; Rosenbaum, 1986). In this case, the regression-adjusted propensity score estimate of the average gender salary gap is:

$$\Delta_2 = \sum_{k=1}^{5} \frac{n_{Fk}}{N_F} \hat{\beta}_{k,male} \tag{3.3}$$

where $\hat{\beta}_{k,male}$ is the coefficient of the indicator variable for male (1=male, 0=female) in the linear regression model fit in propensity stratum k that predicts salary (outcome) from the indicator variable for male (treatment indicator) and any other variables that are out of balance after propensity score subclassification.

A standard error estimate is given by

$$\hat{s}(\Delta_2) = \sqrt{\sum_{k=1}^{5} \frac{n_{Fk}^2}{N_F^2} \left(s.e.(\hat{\beta}_{k,male}) \right)}$$

where $s.e.(\hat{\beta}_{k,male})$ is the usual estimate of the standard error of $\hat{\beta}_{k,male}$. Again, this estimate is only approximate due to the sample-dependent aspects of the propensity score subclassification.

3.1 Propensity score example

To briefly illustrate the propensity score subclassification method, we use the following simple example. We generated 1000 observations with two covariates, X_1 and X_2, both distributed as uniform(0, 2). Each observation was randomly assigned to either the treatment or control group. The probability of being assigned to the treatment group was given by $p = (1 + \exp(3 - X_1 - X_2))^{-1}$, resulting in 30% of the sample being assigned to the treatment group (roughly comparable to the proportion of women in the gender salary data). These treatment assignment probabilities are such that observations with large $X_1 + X_2$ were likely to be assigned to treatment and those with small values were likely to be assigned to control. This created a dataset in which there were relatively few controls with large propensity score values and relatively few treated units with small propensity score values, a pattern often observed in practice. The outcome was generated as $Y = 3Z + 2X_1 + 2X_2 + \epsilon$, where ϵ is $N(0, 1)$ and $Z = 1$ for treated units and $Z = 0$ for control units, so that the treatment effect is 3. The unadjusted estimate of the treatment effect in the raw data, calculated simply as the difference in average outcomes for treated and control units, is 4.16 ($s.e. = 0.12$), with treated outcomes larger than control outcomes, which overestimates the treatment effect. However this estimate is clearly confounded by differences in the values of the covariates between the two groups. The average difference between the treated and control units for X_1 is 0.24 ($s.e.$ =0.04) and for X_2 is 0.36 ($s.e.$ =0.04), with covariate values larger in the treated group.

Using the propensity score subclassification method to estimate the average treatment effect, controlling for covariate differences, we estimated the propensity scores using a logistic regression model with X_1 and X_2 as covariates. Then we subclassified the data into five strata based on the quintiles of the estimated propensity scores for the treated units. The resulting estimates of stratum-specific and overall treatment effects and covariate differences and corresponding standard errors (s.e.) are presented in Table 1. Table 1 shows that, within each stratum, the average values of X_1 and X_2 are comparable for treated and control units. A two-way ANOVA with X_1 as the dependent variable and treatment

indicator (Z) and propensity score stratum index as the independent variables yields a nonsignificant main effect of treatment and a nonsignificant interaction of treatment and propensity score stratum index, confirming that X_1 is balanced across treated and control groups within strata. Similar results are obtained for X_2. As a result, within each stratum, estimates of the treatment effect, calculated as the difference between the treated and control mean outcomes ($\bar{Y}_T - \bar{Y}_C$), are not confounded by differences in the covariates. As Table 1 shows, the treatment effect estimate is close to 3 within each stratum. The overall treatment effect estimate, calculated using formulas (3.1) and (3.2) is 2.97 ($s.e.$ =0.09) which is very close to the true value. Because propensity score subclassification balances both X_1 and X_2, no further regression adjustments are necessary.

Table 1: Example propensity score analysis (T = treatment, C = control)

	$\bar{Y}_T - \bar{Y}_c$		$\bar{X}_{1,T} - \bar{X}_{1,C}$		$\bar{X}_{2,T} - \bar{X}_{2,C}$		Sample Size	
Stratum	mean	s.e.	mean	s.e.	mean	s.e.	treated	control
1	3.33	0.19	0.07	0.05	0.02	0.08	60	337
2	3.17	0.16	0.04	0.08	−0.02	0.08	60	169
3	2.81	0.18	−0.10	0.09	0.06	0.08	60	104
4	2.95	0.21	0.04	0.08	0.00	0.08	60	56
5	2.60	0.24	0.02	0.06	0.01	0.08	60	34
overall treatment effect estimate	2.97***	0.09						

*** indicates p-value $< .01$, ** $.01 \leq p$-value $< .05$, * $.05 \leq p$-value $< .10$.

4. Complex Survey Design Considerations

Both linear regression and propensity score analyses are further complicated when the data have been collected using a complex sampling design, as is the case with the SESTAT data. In complex surveys, each sample unit is assigned a survey weight, which in the simplest case is the inverse of the probability of selection, but is often modified to adjust for nonresponse and poststratification. These survey weights indicate the number of people that each sampled person represents in the population. A common strategy to incorporate survey weights into linear regression modeling is to fit the regression model using both ordinary least squares and a survey-weighted least squares (e.g. Lohr, 1999, chapter 11). Large differences between the two analyses suggest model misspecification (DuMouchel and Duncan, 1983; Lohr and Liu, 1994; Winship and Radbill, 1994). If these differences cannot be resolved by modifying the model (e.g., including more

covariates related to the survey weights), then the weighted analysis should be used since the weights may contain information that is not available in the covariates. Survey-weighted linear regression and the associated linearization variance estimates can be computed by statistical analysis software such as Stata[1] and SAS (An and Watts, 1998).

Although the implications of complex survey design on propensity score estimates of treatment effects have not been discussed in the statistical literature, similar advice of performing the analysis with and without survey weights should apply. Since the propensity score model is used only to match treated and control units with similar background characteristics together in the sample and not to make inferences about the population-level propensity score model, it is not necessary to use survey-weighted estimation for the propensity score model. However, to estimate a population-level treatment effect, it is necessary to consider the use of survey weights in equations (3.1) and (3.3). A survey-weighted version of (3.1) is:

$$\Delta_{w1} = \sum_{k=1}^{5} \left(\frac{\sum_{i \in S_{Fk}} w_i}{\sum_{k=1}^{5} \sum_{i \in S_{Fk}} w_i} \right) \left(\frac{\sum_{i \in S_{Mk}} w_i y_i}{\sum_{i \in S_{Mk}} w_i} - \frac{\sum_{i \in S_{Fk}} w_i y_i}{\sum_{i \in S_{Fk}} w_i} \right) \tag{4.1}$$

where w_i denotes the survey weight for unit i, and S_{Fk} and S_{Mk} denote, respectively, the set of females in propensity score stratum k and the set of males in propensity score stratum k. This formula allows for potential differences in distributions between the sample and the population both within and between sample strata. Within a propensity score stratum, some types of people in the sample may be over- or underrepresented relative to other types of people. The use of the weighted averages within each stratum ensures that these averages reflect the distribution of people in the population. This formula also weights each stratum by the estimated population proportion of women in each stratum ensuring that our calculations reflect the population distribution of women across the five sample quintiles.

Noting that (4.1) is a linear combination of subdomain (ratio) estimators, assuming unequal probability sampling without replacement with overall inclusion probabilities $1/w_i$, an approximate standard error estimate that is analogous to (3.2) is (Lohr, 1999, p. 68)[2]

$$\hat{s}(\Delta_{w1}) = \sqrt{\sum_{k=1}^{5} \left(\frac{\sum_{i \in S_{Fk}} w_i}{\sum_{k=1}^{5} \sum_{i \in S_{Fk}} \omega_i} \right)^2 (s_{Mk}^2 + s_{Fk}^2)}$$

[1] see StataCorp (2003). Stata Statistical Software: Release 8.0. College Station, TX: Stata Corporation.

[2] Also see Stata Press (2003). Stata Survey Data Reference Manual Release 8.0. College Stataion, TX: Stata Corporation, p.66.

where

$$s^2_{Mk} = \frac{n}{n-1}\sum_{i=1}^{n}\left(z_{ik} - \frac{1}{n}\sum_{j=1}^{n} z_{jk}\right)^2,$$

and

$$\begin{aligned} z_{ik} &= \frac{w_i}{\sum_{i\in S_{Mk}}\omega_i}\left(y_i - \frac{\sum_{i\in S_{Mk}} w_i y_i}{\sum_{i\in S_{Mk}} w_i}\right) && i \in S_{Mk} \\ &= 0 && i \notin S_{Mk} \end{aligned}$$

where n is the total sample size. A similar formula for s^2_{Fk} applies for women. As in the simple random sampling case, this standard error estimate is only approximate because we are not accounting for the sample-dependent aspects of the propensity score subclassification. We are also not accounting for any extra variability due to sample-based nonresponse or poststratification adjustments to the survey weights. Replication methods can be used to account for this extra source of variability (Canty and Davison, 1999; Korn and Graubard, 1999, chapter 2.5; Wolter, 1985, chapter 2), however this issue is beyond the scope of this paper.

Extensions of these formulas to include regression adjustments within propensity score strata to adjust for remaining covariate imbalance is straightforward. In this case, the vector of estimated regression coefficients in a survey-weighted linear regression model fit in propensity stratum k that predicts salary (outcome) from the indicator variable for male (treatment indicator) and any covariates that remain out of balance after subclassification on the propensity score, is given by

$$\hat{\boldsymbol{\beta}}_k^{\,w} = (X_k^T W_k X_k)^{-1} X_k^T W_k \mathbf{y}_k$$

where X_k is the matrix of explanatory variables, W_k is a diagonal matrix of the sample weights, and y_k is the vector of responses in propensity score stratum k. The usual linearization variance estimate of $\hat{\beta}_k^w$ is given by (Binder, 1983; Shah, Holt, and Folsom, 1977)

$$\hat{V}(\hat{\boldsymbol{\beta}}_k^{\,w}) = (X_k^T W_k X_k)^{-1} \hat{V}\left(\sum_{i\in S_k} w_i \mathbf{q}_{ik}\right)(X_k^T W_k X_k)^{-1} \tag{4.2}$$

where S_k denotes the set of sample units in propensity score stratum k, and

$$\mathbf{q}_{ik} = \mathbf{x}_{ik}^T(y_{ik} - \mathbf{x}_{ik}^T\hat{\boldsymbol{\beta}}_k^{\,w})$$

where x_{ik}^T is the i-th row of X_k and y_{ik} is the i-th element of y_k. The variance estimate in the middle of (4.2) depends on the sample design. For example, for

unequal probability sampling without replacement, with overall inclusion probabilities $1/w_i$, we can use the following approximation for the (j,ℓ)-th element of the variance-covariance matrix (Sarndal, Swensson, and Wretman, 1992, p.99)[3]

$$\hat{V}\left(\sum_{i\in S_k} w_i \mathbf{q}_{ik}\right)_{jl} = \frac{n}{n-1}\left(\sum_{i=1}^{n} w_i\right)^2 \times \sum_{i=1}^{n}\left(d_{ijk} - \frac{1}{n}\sum_{i=1}^{n} d_{ijk}\right)\left(d_{ilk} - \frac{1}{n}\sum_{i=1}^{n} d_{ilk}\right) \quad (4.3)$$

where

$$d_{ijk} = \frac{w_i}{\sum_{i=1}^{n}\omega_i}\left(u_{ijk} - \frac{\sum_{i=1}^{n} w_i u_{ijk}}{\sum_{i=1}^{n} w_i}\right)$$

and $u_{ijk} = q_{ijk}$ if unit i is in propensity score stratum k and zero otherwise, where q_{ijk} is the j-th element of $\mathbf{q}_{ik}$.

Letting $\hat{\beta}^{w}_{k,male}$ denote the coefficient of the indicator variable for male in the survey-weighted linear regression model in propensity score stratum k, we have the following estimate of gender salary gap after regression adjustment within propensity score strata

$$\Delta_{w2} = \sum_{k=1}^{5}\left(\frac{\sum_{i\in S_{Fk}} w_i}{\sum_{k=1}^{5}\sum_{i\in S_{Fk}} w_i}\right)\hat{\beta}^{w}_{k,male} \quad (4.4)$$

with an estimated standard error of

$$\hat{s}(\Delta_{w2}) = \sqrt{\sum_{k=1}^{5}\left(\frac{\sum_{i\in S_{Fk}} w_i}{\sum_{k=1}^{5}\sum_{i\in S_{Fk}} w_i}\right)^2 \hat{V}(\hat{\beta}^{w}_{k,male})}\ . \quad (4.5)$$

5. Data Analysis

The field of Information Technology (IT) has experienced a dramatic growth in jobs in the United States, but there are concerns about women being underpaid in IT occupations (AAUW, 2000; Council of Economic Advisers, 2000; Gearan, 2000a, 2000b). To address this issue it is necessary to have an accurate estimate of the gender salary gap.

[3]Also see Stata Press (2003). Stata Survey Data Reference Manual Release 8.0. College Stataion, TX: Stata Corporation, p.66.

5.1 The data

We analyze data from the 1997 U.S. SESTAT database. This database contains information from several national surveys of people with at least a bachelor's degree in science or engineering or at least a bachelor's degree in a non-science and engineering field but working in science and engineering. For a detailed description of the coverage limitations see NSF 99-337. Our analysis focuses on 2035 computer systems analysts (1497 men, 538 women), 1081 computer programmers (817 men, 264 women), 2495 software engineers (2096 men, 399 women), and 839 information systems scientists (609 men, 230 women) who were working full-time in the United States in 1997 and responded to the U.S. National Survey of College Graduates or the U.S. Survey of Doctoral Recipients. A total of 13 workers with professional degrees (e.g., doctor of medicine (M.D.), doctor of dental sugery (D.D.S.), juris doctor (J.D.)) were excluded from the analysis since this was too small a sample to draw conclusions about workers with professional degrees. Also one extreme outlier was excluded from the sample of information systems scientists.

The sample designs for the component surveys making up the SESTAT database used unequal probability sampling. Although each survey has a different design, generally more of the sample is allocated to women, underrepresented minorities, the disabled, and individuals in the early part of their career, so that these groups of people are overrepresented in the database. Survey weights that adjust for these differential selection probabilities and also for nonresponse and poststratification adjustments are present in the database. We use these weights in the survey-weighted linear regression and propensity analyses in Sections 5.3 and 5.4 to illustrate calculations for an unequal probability sampling design. Refinements to the standard error estimates are possible if additional information about stratification, poststratification, or nonresponse adjustments is available, but that is beyond the scope of this illustration.

A comparison of the weighted and unweighted linear regression and propensity score analyses yielded substantially different results that could not be resolved by modifying the models. Because the survey weights are correlated with salary it is important to incorporate the survey weights into the analysis to accurately estimate the gender salary gap in these populations. Differences in the weighted and unweighted gender gap estimates seem to be related to the differential underrepresentation of lower paid men and women in these samples. We return to this issue in Section 5.5.

Table 2 presents survey-weighted unadjusted average differences in salaries for men and women in the four occupations. On average, women earn 7% to 12% less than men in the same occupation in this population. Similar results have been

reported for IT salaries (AAUW, 2000) and engineering salaries (NSF 99-352). Revised estimates of the gender differences, that control for relevant background characteristics, are presented in Sections 5.4 and 5.5.

Table 2: Unadjusted average gender differences in salary (survey weighted)

Occupation	Annual salary Men[a]	Women[a]	Difference[b]
Computer Systems	58,788	54,278	4,510***
Analyst	(680)	(986)	(7.7%)
Computer Programmer	58,303	54,209	4,094**
	(972)	(1,406)	(7.0%)
Software Engineer	67,906	63,407	4,499***
	(604)	(1,748)	(6.6%)
Information Systems	60,902	53,305	7,597***
Scientists	(1,039)	(1,747)	(12.5%)

[a]Standard errors in parentheses, [b]Percentage of average salary for men in parentheses.
*** p-value $< .01$, ** $.01 \leq p$-value $< .05$, * $.05 \leq p$-value $< .10$.

5.2 Confounding variables

To estimate gender differences in salary, it is necessary to control for educational and job-related characteristics. We control for the confounding variables listed in Table 3. Similar covariates have been used in other studies of gender gaps in careers (e.g., Kirchmeyer, 1998; Marini and Fan, 1997; Marini 1989; Schneer and Reitman, 1990; Long, Allison, and McGinnis, 1993; Stanley and Jarrell, 1998; Hull and Nelson, 2000).

We comment here on a few of the variables for clarification. The work activities variables represent whether each activity represents at least 10% of the employee's time during a typical workweek (1=yes, 0=no). The supervisory work variable represents whether the employee's job involves supervising the work of others (1=yes, 0=no). Employer size is measured on a scale of 1-7 (1=under 10 employees, 2=10-24 employees, 3=25-99 employees, 4=100-499 employees, 5=500-999 employees, 6=1000-4999 employees, 7=5000 or more employees). We treat this as a quantitative variable in the regression since larger values are associated with larger employers. Finally, the regression models contain quadratic terms for years since most recent degree and years in current job, since the rate of growth of salaries may slow as employees acquire more experience (Gray, 1993).

To avoid multicollinearity, these variables have been mean-centered before squaring.

Table 3: Survey weighted regression results (Y = Annual Salary).

	Computer system analysts	Computer programmers	Software engineers	Information systems scientists
Intercept	31,571***	17,168**	51,144***	40,080***
Male	2,429**	3,577**	-2,461	4,555**
Years since MRD[a]	631***	866***	502	901***
(Years since MRD[a])^2	-21***	-43***	11	-38***
MRD[a] in computer/math	4,150***	4,481***	2,827***	2,459
Type of MRD[a,c]				
Master's	8,917***	7,871***	7,044***	10,523***
Doctorate	13,863***	13,020***	14,889***	19,752***
College courses after MRD[a]	-1,433	-1,185	-573	-6,740***
Employment Sector[d]				
Government	-5,846***	-9,313***	-8,028***	-11,807***
Education	-13,536***	-12,559***	-11,100***	-16,554***
Hours worked during a typical week	223**	559***	391***	308*
Years in current job	-81	-81	16	93
(Years in current job)^2	31***	19*	12	-5
Work Activities:				
Basic Research[b]	44	-611	-4,217***	-1,217
Applied Research[b]	781	-2,907*	2,820***	883
Computer App.[b]	5,760*	3,149	-10,666**	-10,241*
Development[b]	-1,958	419	1,752	-521
Design[b]	2,397**	-554	2,444**	4,824***
Management/Admin.[b]	2,437	2,274	3,132*	-3,364

Table continues on next page

[a] MRD = most recent degree, [b]response is yes/no, [c]reference category is Bachelor's degree, [d]reference category is business/industry, *** p-value $< .01$, ** $.01 \leq p$-value $< .05$, * $.05 \leq p$-value $< .10$.

Table 3 continued: Survey weighted regression results (Y = Annual Salary)

	Computer system analysts	Computer programmers	Software engineers	Information systems scientists
Supervisory work[b]	3,334**	6,606***	4,015**	7,464***
Attended work related training during past year	-88	-407	69	-15
Employer size	-440	388	-707**	308
Location[c]				
New England	1,376	-12,496***	-6,102***	35
Mid Atlantic	-21	-3,293	-11,094***	-2,590
East North Central	-4,032**	-9,044***	-9,481***	-8,586***
West North Central	-2,776	-11,437***	-12,976***	-12,311***
South Atlantic	-4,352**	-5,868**	-9,182***	-6,784**
East South Central	-7,419***	-16,788***	-32,278***	-15,300***
West South Central	-5,397***	-8,527***	-6,572***	-5,029
Mountain	-6,595***	-10,490***	-10,064***	-9,667**
Male*(years since MRD[a])			643**	
Male*(years since MRD[a])^2			-53**	
Male*(Mid Atlantic)			7,278**	
Male*(E-S Central)			26,078***	
R^2	.18	.25	0.29	0.31
Overall F-statistic	10.64***	11.94***	24.06***	15.81***
Sample size	2035	1081	2495	839

[a]MRD = most recent degree, [b]response is yes/no, [c]reference category is Pacific,
*** p-value $< .01$, ** $.01 \leq p$-value $< .05$, * $.05 \leq p$-value $< .10$.

5.3 Regression results

Table 3 presents the survey-weighted regression results for each of the four IT occupations. To arrive at these final models, first a linear regression model predicting salary from all the covariates in Table 3 along with interactions between all of these covariates and the indicator variable for male was fit. An F-test, using a Wald statistic appropriate for complex survey data (Korn and Graubard, 1990)[4], was used to test whether the coefficients of the interactions with male were all simultaneously zero. Results are presented in Table 4. When this test was statistically significant, as it was for software engineers, a backward selection procedure was used to identify the significant ($p < 0.05$) interactions. Residual plots and other diagnostics for these models were satisfactory. Values of R-squared are comparable to those in similar studies (Schneer and Reitman, 1990; Stroh *et al.*, 1992; Marini and Fan, 1997).

Table 4: Tests of interactions with male in the linear regression models

	Computer system analysts	Computer programmers	Software engineers	Information systems scientists
F-statistic	1.24^a	1.39^b	2.47^c	1.41^d
p-value	0.19	0.09	0.00	0.08

[a]degrees of freedom are (28, 2007), [b]degrees of freedom are (28, 1053), [c]degrees of freedom are (28, 2467), and [d]degrees of freedom are (28, 811).

The results in Table 3 show that after controlling for educational and job-related characteristics, there are significant gender salary gaps in all four occupations. For computer systems analysts, computer programmers and information systems scientists there is a statistically significant shift in the regression equation for men relative for women ($2,429, $3,577, and $4,555 respectively). For male software engineers, there is a shift in the regression equation in the Mid Atlantic ($7,278) and East South Central ($26,078) regions, combined with statistically significant interactions with years since most recent degree ($643) and the quadratic term for years since most recent degree (−$53) suggesting differential rewards for experience for male and female software engineers. Note, however, the gender gap for software engineers in the East South Central region should be interpreted with caution since the sample contains only 40 men and only 6 women in this region.

[4]Also see Stata Press (2003). Stata Survey Data Reference Manual Release 8.0. College Stataion, TX: Stata Corporation, p.97.

Although we are most concerned with the coefficient of the indicator variable for male and coefficients of any interactions involving male, these models generally confirm, as we would expect, an increase in salaries for workers with more experience (with the rate of increase slowing over time), workers with more education, and workers with supervisory responsibilities. These models also show large differences in salaries across geographic regions and employment sectors.

5.4 Propensity Score Results

For the propensity score analysis, an unweighted logistic regression model was fit for each occupation to predict the propensity of being male, including main effects for all the covariates listed in Table 3. Because we are concerned with balancing the distribution of the covariates, and not with obtaining a parsimonious model, we did not discard statistically insignificant predictors (Rubin and Thomas, 1996). All four occupations had very good overlap in the estimated propensity scores for men and women. Since the propensity score can be thought of as a one-number summary of the characteristics of a person, checking for overlap in the propensity scores verifies that there are comparable men and women in the data set. If there is little or no overlap in the propensity score distributions, this is an indication that the men and women in the sample are very different and comparisons between these groups should be made with extreme caution or not at all. This ability to easily check that the data can support comparisons between the two groups is one of the advantages of a propensity score analysis over a regression analysis. The overlap in the propensity scores also indicates the range over which comparisons can be made (Dehejia and Wahba, 1999; Zanutto, Lu, and Hornik, 2005). Samples sizes in the regions of propensity score overlap are shown in Table 5.

Table 5: Propensity score strata sample sizes

	Computer analysts		Computer programmers		Software engineers		Information systems scientists	
	women	men	women	men	women	men	women	men
Stratum 1	107	184	53	72	79	166	45	58
Stratum 2	107	208	53	124	80	230	46	98
Stratum 3	107	264	53	94	80	360	46	82
Stratum 4	107	303	53	166	80	602	46	123
Stratum 5	106	536	52	313	79	725	45	198
Total	534	1495	264	769	398	2083	228	559

Table 6: Balance statistics before propensity score subclassification

	Computer analysts	Computer programmers	Software engineers	Information systems scientists
Gender[a]				
p-value $< .05$	5	5	7	2
$.05 \leq p$-value $< .10$	3	3	2	2

[a]main effect of gender

Table 7: Balance statistics after propensity score subclassification

	Computer analysts	Computer programmers	Software engineers	Information systems scientists
Gender[a]				
p-value $< .05$	1	0	2	0
$.05 \leq p$-value $< .10$	0	0	0	0
Interactions[b]				
p-value $< .05$	3	0	4	1
$.05 \leq p$-value $< .10$	2	1	0	3

[a] main effect of gender,
[b]interactions between gender and propensity score stratum index.

Data in the region of propensity score overlap were subclassfied into five strata based on the quintiles of the estimated propensity scores for women. As a check of the adequacy of the propensity score model, a series of analyses was conducted to assess the covariate balance in the five groups of matched men and women. For each continuous covariate, we fit a survey-weighted two-way ANOVA where the dependent variable was the covariate and the two factors were gender and propensity score stratum index. For each binary covariate, we fit an analogous survey-weighted logistic regression with the covariate as the dependent variable and gender and propensity score stratum index and their interaction as predictors. In these analyses, nonsignificant main effects of gender and nonsignificant effects of the interaction between propensity score stratum index and gender indicate that men and women within the five propensity score strata have balanced covariates. A summary of the number of these gender and gender interaction effects that are statistically significant before and after propensity score subclassification are shown in Tables 6 and 7. Before subclassification, using survey-weighted one-way ANOVAs for continuous covariates and analogous survey-weighted logistic

regressions for binary covariates, we found more covariates to be out of balance (as indicated by a statistically significant gender main effect) than we would expect by chance alone. After subclassification, the balance statistics (summarized by the p-values of gender main effects and gender by propensity score-stratum interactions) are much closer to what we would expect in a completely randomized experiment. Regression adjustments were used to adjust for remaining imbalances. Specifically, within each propensity score stratum, a survey-weighted linear regression model predicting salary from the indicator for male and any covariates that were out of balance was fit and equations (4.2), (4.3), (4.4), and (4.5) were used to estimate the gender salary gap and its standard error.

The survey-weighted regression-adjusted propensity score estimates of the gender gaps are shown in Table 8. After controlling for educational and job-related covariates, the propensity score analyses show significant gender salary gaps for all four occupations. These results are similar to the results from the linear regression analysis. Note that when comparing the propensity score and linear regression analysis results for software engineers, the linear regression model predicts an overall average gap of \$3,690 ($s.e.$ =1,472) when averaging over the women in this population, which is similar to the gap of \$4,016 ($s.e.$ =1,627) estimated from the propensity score analysis.

Table 8: Survey weighted propensity score estimates of average gender salary gaps

	Computer analysts		Computer programmers		Software engineers		Information systems scientists	
	gap	s.e.	gap	s.e.	gap	s.e.	gap	s.e.
Stratum 1	4,271	1,856	6,586	3,651	2,754	4,451	2,126	3,489
Stratum 2	-1,285	3,007	1,063	3,582	7,167	3,157	6,694	4,328
Stratum 3	1,182	2,004	2,919	3,419	3,715	3,881	9,183	3,578
Stratum 4	4,972	2,150	10,876	3,230	2,503	3,486	3,088	4,476
Stratum 5	4,486	3,315	-4,648	3,995	3,830	3,119	2,740	6,749
Overall	2,691**	1,129	3,192**	1,611	4,016**	1,627	4,770**	1,985
Sample Size	2,029		1,033		2,481		787	

*** p-value $< .01$, ** $.01 \leq p$-value $< .05$, * $.05 \leq p$-value $< .10$.
p-values displayed only for overall means.

5.5 Comparison of Weighted and Unweighted Analysis

To illustrate the effect of ignoring the complex survey design, we compare the results from survey-weighted and unweighted analysis. Summaries of these analyses are presented in Tables 9 and 10. The weighted and unweighted results differ quite substantially in terms of the size of the estimated gender salary gaps and in terms of which interactions with male are significant in the linear regression models. The discrepancies between the weighted and unweighted analyses seem to be related to the differential underrepresentation of lower paid men and women. In particular, unweighted estimates of the salary gap are larger than the weighted estimates for computer programmers and software engineers, where lower paid men are more underrepresented than lower paid women (as seen by a larger negative correlation between the survey weights and salary for men in Table 11). In contrast, unweighted estimates of the salary gap are smaller than the weighted estimates for information systems scientists where lower paid women are more underrepresented than lower paid men.

Table 9: Comparison of weighted and unweighted propensity score results

	Computer analysts		Computer programmers		Software engineers		Information systems scientists	
	gap	s.e.	gap	s.e.	gap	s.e.	gap	s.e.
Weighted	2,691**	1,129	3,192**	1,611	4,016**	1,627	4,770**	1,985
Unweighted	2,597***	921	5,555***	1,438	4,418***	1,109	3,341*	1,730

*** p-value $< .01$, ** $.01 \leq p$-value $< .05$, * $.05 \leq p$-value $< .10$.

Table 10: Comparison of weighted and unweighted regression results

	Computer analysts		Computer programmers		Software engineers		Information systems scientists	
	$\hat{\beta}_{male}$	s.e.	$\hat{\beta}_{male}$	s.e.	$\hat{\beta}_{male}$	s.e.	$\hat{\beta}_{male}$	s.e.
Weighted	2,429**	1,081	3,577**	1,385	-2,461[a]	3,556	4,555**	1,832
Unweighted	2,256**	959	5,181***	1,895	4,375***	988	3,084*	1,646

[a]Some interactions with male are also significant in this model (see Table 2). This model predicts an average salary gap of \$3,690** (s.e.=1,472) when averaging over all the women in this population.
*** p-value $< .01$, ** $.01 \leq p$-value $< .05$, * $.05 \leq p$-value $< .10$.

Table 11: Correlation between survey weights and salary

	Computer analysts	Computer programmers	Software engineers	Information systems scientists
Men	-0.08***	-0.11***	-0.09***	-0.16***
Women	-0.07	0.08	-0.05	-0.21***

*** p-value $< .01$, ** $.01 \leq p$-value $< .05$, * $.05 \leq p$-value $< .10$.

6. Comparison of Methodologies

There are several technical advantages of propensity score analysis over multiple linear regression. In particular, when covariate balance is achieved and no further regression adjustment is necessary, propensity score analysis does not rely on the correct specification of the functional form of the relationship (e.g., linearity or log linearity) between the outcome and the covariates. Although such specific assumptions may not be a problem when the groups have similar covariate distributions, when the covariate distributions in the two groups are very different linear regression models depend on the specific form of the model to extrapolate estimates of gender differences (Dehejia and Wahba, 1999; Drake 1993; Rubin, 1997). When regression adjustment is used to adjust for remaining covariate imbalances, previous research has found that such adjustments are relatively robust against violations of the linear model in matched samples (Rubin 1973, 1979; Rubin and Thomas, 2000). Propensity score analysis depends on the specification of the propensity score model, but the diagnostics for propensity score analysis (checking for balance in the covariates) are much more straightforward than those for regression analysis (residual plots, measures of influence, etc.) and, as explained previously, enable the researcher to easily determine the range over which comparisons can be supported. Furthermore, propensity score analysis can be objective in the sense that propensity score modeling and subclassification can be completed without ever looking at the outcome variables. Complete separation of the modeling and outcome analysis can be guaranteed, for example, by withholding the outcome variables until a final subclassification has been decided upon, after which no modifications to the subclassification are permitted. These two aspects of the analysis are inextricably linked in linear regression analysis.

A nontechnical advantage of propensity score analysis is the intuitive appeal of creating groups of similar treated and control units. This idea may be much easier to explain to a nontechnical audience than linear regression. These groups

formed by subclassifying or matching on the propensity score are also very similar in concept to audit pairs commonly used in labor or housing discrimination experiments (Darity and Mason, 1998; National Research Council, 2002). In an audit pair study of gender discrimination in hiring, for example, one female and one male job candidate would be matched based on relevant characteristics (and possibly given the same resumes) and then would apply for the same jobs to determine whether their success rates are similar.

An advantage of multiple linear regression, however, is that a linear regression model may indicate a difference between the salaries of men and women due to an interaction with other covariates, such as industry or region of the country, as was the case for software engineers. A propensity score analysis estimates the gender gap averaged over the population, possibly obscuring important interactions. Also, in addition to estimating any gender effects, the regression model also describes the effects of other covariates. For example, our regression models show that higher salaries are associated with more experience, more education, and more supervisory responsibilities. In contrast, propensity score analyses are designed only to estimate the overall gender effect. Of course, these interpretations of the linear regression coefficients are only reliable after a careful fitting of the regression model with appropriate diagnostic checks, including a check of whether there is sufficient overlap in the two groups to facilitate comparisons without dangerous extrapolations.

Both multiple linear regression and propensity score analyses are subject to problems of omitted variables, "tainted" variables and mismeasured variables. A tainted variable is a variable like job rank that, for example, may be affected by gender discrimination in the same way that salary is affected (Finkelstein and Levin, 2001; Haignere, 2002). If we control for job rank, in linear regression or propensity score analysis, this may conceal gender differences in salary due to discrimination in promotion. For example, male and female supervisors may be similarly paid, but women may rarely be promoted to supervisory status. Rosenbaum (1984) discusses the possible biasing effect of controlling for a variable that has been affected by the treatment in the propensity score context. Mismeasured variables may also affect the assessment of gender differences. For example, years from an individual's first bachelor's degree or from their most recent degree is often used as a proxy for years of experience (Gray, 1993; NSF 99-352) but this may overstate the experience of anyone who may have temporarily left the workforce since graduating.

Both linear regression and propensity score analysis are also affected by complex survey designs. The survey design must be incorporated into estimates from both these methods to obtain unbiased estimates of population-level effects.

7. Discussion

The results from our linear regression and propensity score analyses agree on the size and statistical significance of the gender salary gaps in these four IT occupations after controlling for educational and job-related covariates. Results from our two different analysis methods may agree so closely in this example because there is good overlap in the distribution of covariates for the men and women in each of the four occupations. More specifically, the propensity score overlap regions used in the propensity score analysis do not differ much from the whole samples used by the regression analysis. An example by Hill et al. (2004) suggests that at least some of the benefit of propensity score methods may result from the restriction of the analysis to a reasonable comparison group. Other research has found statistical modeling to be relatively robust in well-matched samples (Rubin 1973, 1979). These factors may have contributed to the similarity of the results in our analyses. Other studies have found propensity score analysis to more closely estimate known experimental effects than linear regression (Dehejia and Wahba, 1999; Hill *et al.*, 2004).

Our analysis also shows that it is important to incorporate survey weights from the complex survey design into both methodologies. Ignoring the survey weights affects gender salary gap estimates in both the linear regression and propensity score analyses, probably due to the differential underrepresentation of lower paid men and women in these samples.

Finally, the finding of significant gender salary gaps in all four IT occupations agrees with numerous other studies that have shown that gender salary gaps can not usually be fully explained by traditional "human capital" variables such as education, years of experience, job responsibilities (e.g., Bamberger, Admati-Dvir, and Harel, 1995; Jacobs, 1992; Marini, 1989; NSF 99-352; Stanley and Jarrell, 1998). Studies of workers in other fields have estimated similar sized gaps after controlling for covariates similar to the ones used in our study (NSF 99-352, Stanley and Jarrell, 1998). It is possible that the gaps seen in our analysis could be explained by other covariates not available in the SESTAT data, such as quality or diversity of experience, number of years of relevant experience (as opposed to number of years of total experience), job performance, and willingness to move or change employers.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 0089872

References

AAUW Educational Foundation Commission on Technology, Gender, and Teacher Education. (2000). Tech-Savvy: Educating girls in the new computer age, Washington, D.C.: American Association of University Women Educational Foundation.

Agodini, R., and Dynarski, M. (2001). Are experiments the only option? A look at dropout prevention programs. Technical Report, Princeton, NJ: Mathematica Policy Research.

An, A. B., and Watts, D. L. (1998). New SAS procedures for analysis of sample survey data. In (SUGI) *SAS Users Group International Proceedings*, SAS Institute, Cary, NC.

Bamberger, P., Admati-Dvir, M., and Harel, G. (1995). Gender-based wage and promotion discrimination in Israeli high-technology firms: Do unions make a difference? *The Academy of Management Journal* **38**, 1744-1761.

Benjamin, D. J. (2003). Does 401(k) eligibility increase saving? Evidence from propensity score subclassification. *Journal of Public Economics* **87**, 1259-1290.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279-282.

Canty, A. J., and Davison, A. C. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician* **48**, 379-391.

Council of Economic Advisers. (2000). Opportunities and gender pay equity in new economy occupations. White Paper, May 11, 2000, Washington, D.C.: Council of Economic Advisors.

D'Agostino, R. B. Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* **17**, 2265-2281.

Darity, W. A. and Mason, P. L. (1998). Evidence on discrimination in employment: Codes of color, codes of gender. *The Journal of Economic Perspectives* **12**, 63-90.

Dehejia, R. H., and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of The American Statistical Association* **94**, 1053-1062.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* **49**, 1231-1236.

Du, J. (1998). Valid inferences after propensity score subclassification using maximum number of subclasses as building blocks. Ph.D. thesis, Harvard University.

DuMouchel, W. H. and Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association* **78**, 535-543.

Finkelstein, M. O., and Levin, B. (2001). *Statistics for Lawyers*, Second Edition. Springer-Verlag.

Gastwirth, J. L. (1993). Comment on 'Can statistics tell us what we do not want to hear? The case of complex salary structures'. *Statistical Science* **8**, 165-171.

Gearan, A. (2000a). Clinton chides tech biz over pay gap. *Associated Press* (May 11, 2000).

Gearan, A. (2000b). President seeks equal pay for women. *Associated Press* (May 11, 2000).

Gray, M. (1993). Can statistics tell us what we do not want to hear? The case of complex salary structures. *Statistical Science* **8**, 144-179.

Haignere, L. (2002). *Paychecks: A Guide to conducting salary-equity studies for higher education faculty.* Washington, D.C.: American Association of University Professors.

Hill, J. L., Reiter, J. P., and Zanutto, E. L. (2004). A comparisons of experimental and observational data analyses. In *Applied Bayesian Modeling and Causal Inference From an Incomplete-Data Perspective* (Edited by Andrew Gelman and Xiao-Li Meng), 44-56. Wiley.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945-960.

Hornik, R. *et al.* (2002). Evaluation of the national youth anti-drug media campaign, fourth semi-annual report of findings. Delivered to National Institute on Drug Abuse, National Institutes of Health, Rockville, MD: Westat.

Hull, K. E., and Nelson, R. L. (2000). Assimilation, choice, or constraint? Testing theories of gender differences in the careers of lawyers. *Social Forces* **79**, 229-264.

Jacobs, J. A. (1992). Women's entry into management: Trends in earnings, authority, and values among salaried managers. *Administrative Sciences Quarterly* **37**, 282-301.

Kirchmeyer, C. (1998). Determinants of managerial career success: Evidence and explanation of male/female differences. *Journal of Management* **24**, 673-692.

Korn, E. L, and Graubard, B. I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics. *The American Statistician* **44**, 270-276.

Korn, E. L., and Graubard, B. I. (1999). *Analysis of Health Surveys.* Wiley.

Larsen, M. D. (1999). An analysis of survey data on smoking using propensity scores. *Sankhya: The Indian Journal of Statistics* **61**, 91-105.

Lohr, S. L. (1999). *Sampling: Design and Analysis.* Duxbury Press.

Lohr, S. L., and Liu, J. (1994). A comparison of weighted and unweighted analyses in the national crime victimization survey. *Journal of Quantitative Criminology* **10**, 343-360.

Long, J. S., Allison, P. D., and McGinnis, R. (1993). Rank advancement in academic careers: Sex differences and the effects of productivity. *American Sociological Review* **58**, 703-722.

Marini, M. M. (1989). Sex differences in earnings in the United States. *Annual Review of Sociology* **15**, 343-380.

Marini, M. M., and Fan. P. -L. (1997). The gender gap in earnings at career entry. *American Sociological Review* **62**, 588-604.

National Research Council (2002). *Measuring housing discrimination in a national study: Report of a workshop, Committee on National Statistics* (Edited by A.W. Foster, F. Mitchell, S.E. Fienberg). Division of Behavioral and Social Sciences and Education. National Academy Press.

National Science Foundation (NSF 99-337). *SESTAT: A Tool for Studying Scientists and Engineers in the United States.* (Authors: Nirmala Kannankutty and R. Keith Wilkinson), Arlington, VA, 1999.

National Science Foundation (NSF 99-352). *How Large is the Gap in Salaries of Male and Female Engineers?* Arlington, VA, 1999.

Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X.-H., and Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety* **9**, 93-101.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* **147**, 656-666.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics* **11**, 207-224.

Rosenbaum, P. R. (2002). *Observational Studies*, second edition. Springer-Verlag.

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 sbclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516-524.

Rosenbaum, P. R., and Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics* **41**, 103-116.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29**, 185-203.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74**, 318-328.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* **127**, 757-763.

Rubin, D. B., and Thomas, N. (1996) Matching using estimated propensity scores: Relating theory to practice. *Biometrics* **52**, 249-264.

Rubin, D. B., and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* **95**, 573-585.

Sarndal, C. -E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Schneer, J. A., and Reitman, F. (1990). Effects of employment gaps on the careers of M.B.A.'s: More damaging for men than for women? *The Academy of Management Journal* **33**, 391-406.

Shah, B. V., Holt, M. M., and Folsom, R. E. (1977). Inference about regression models from complex survey data. *Bulletin of the International Statistical Institute* **47**, 43-57.

Stanley, T. D. and Jarrell, S. B. (1998). Gender wage discrimination bias? A meta-regression analysis. *The Journal of Human Resources* **33**, 947-973.

Stroh, L. K., Brett, J. M., and Reilly, A. H. (1992). All the right stuff: A comparison of female and male managers' career progression. *Journal of Applied Psychology* **77**, 251-260.

Winship, C., and Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods and Research* **23**, 230-257.

Wolter, K.M (1985). *Introduction to Variance Estimation*. Springer-Verlag.

Zanutto, E. L., Lu, B., and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national anti-drug media campaign. *Journal of Educational and Behavioral Statistics* **30**, 59-73.

Received April 16, 2004; accepted September 27, 2004.

Elaine L. Zanutto
Department of Statistics
The Wharton School
University of Pennsylvania
466 J.M.Huntsman Hall,
3730 Walnut St.
Philadelphia, PA 19104, USA
zanutto@wharton.upenn.edu