DEPARTMENT OF STATISTICS
The Pennsylvania State University
University Park, PA 16802 U.S.A.

TECHNICAL REPORTS AND PREPRINTS

Number 10-05: April 2010

Estimating Average Treatment Effects When the Treatment is a Latent Class

Joseph Kang[1] and Joseph L. Schafer[2]

Suggested Citation:
Schafer, J. L., & Kang, J. (2010). *Estimating average treatment effects when the treatment is a latent class* (Technical Report 10-05). Department of Statistics, Penn State.

[1]**Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL**
[2]**Department of Statistics, The Pennsylvania State University, University Park, PA**

Estimating Average Treatment Effects When the Treatment is a Latent Class

Joseph Kang and Joseph L. Schafer

April 28, 2010

Joseph Kang is Assistant Professor, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, 680 N. Lake Shore Drive, Suite 1102, Chicago, IL 60611. Joseph L. Schafer is Associate Professor, Department of Statistics, The Pennsylvania State University, University Park, PA 16802. Authors' names appear in alphabetical order. This research was supported by the National Institute on Drug Abuse 1-P50-DA10075, by National Institute of Child Health and Human Development 1-R03-HD060659, and by National Institute of Diabetes and Digestive and Kidney Diseases 1-R21-DK082858. This research uses data from Add Health, designed by J.R. Udry, P.S. Bearman, and K.N. Harris and supported National Institute of Child Health and Human Development, P01-HD31921.

ABSTRACT

In the potential-outcomes framework for causal inference, treatment effects are defined as differences among outcomes that would be realized if different treatments were applied to the same individual. In typical applications of the potential-outcomes approach, the treatment is characterized as a binary variable observed without error. We present a new model that accounts for error in measuring the treatment status. Our model combines a latent-class regression for the treatment with linear regressions for the potential outcomes. Maximum-likelihood estimates of model parameters are computed by an EM algorithm. To estimate average treatment effects, we average over covariates nonparametrically using expected estimating equations. As a motivating example, we analyze the effects of naturalistic weight-control strategies on body-mass index (BMI) among adolescent girls. Using data from a large national survey, we identify latent classes of weight-control behavior and estimate effects of hypothetical changes in behavior on BMI five years later, taking into account the complex sample design. *Supplemental documents and software are available online.*

KEY WORDS: Causal inference; Dieting; EM algorithm; Expected estimating functions; Nonrandomized studies; Propensity scores.

1 INTRODUCTION

Causal inference requires care when treatments are not randomized. In the potential-outcomes framework, each individual has an outcome under each treatment that could have been received (Rubin, 1974a). Only one outcome is seen for any individual, so the treatment effect, usually defined as a difference between potential outcomes, is unobservable. Nevertheless, by making assumptions about the treatment mechanism or the distribution of potential outcomes, it becomes possible to estimate an average treatment effect (ATE). Overviews of this framework are provided by Gelman and Meng (2004), Rosenbaum (2002), Rubin (2005), and Schafer and Kang (2008).

In most applications of the potential-outcomes model, the treatment has been characterized as a binary variable that is always observed. For example, D'Agostino (1998) examined the effects of post-term birth on outcomes in childhood; the status (1=post-term, 0=not) of each child was ascertained from birth records. Dehejia and Wahba (1999) studied the effects of participation in a job-training program on subsequent earnings, and participation (1=in program, 0=not) was known.

However, there are many important research questions for which the treatment is not well characterized by an observed variable. Consider the effect of dieting on body weight among adolescent girls. Dieting is known to predict weight gain, possibly through its association with binge eating (Stice et al., 1999; Neumark-Sztainer et al., 2006). Dieting can be difficult to measure, however, because notions of what constitutes dieting vary widely (Neumark-Sztainer and Story, 1998). Moreover, dieting does happen in a vacuum; girls who diet may do so in conjuction with or in lieu of other strategies such as exercising or using diet pills. The effect of a change from non-dieting to dieting or vice-versa should be understood in the context of accompa-

nying changes in other behaviors. Popular methods for estimating average treatment effects ignore uncertainty and bias that arise when the putative cause is measured imperfectly. Failure to account for measurement error in the treatment is one reason why the potential-outcomes framework is still met with resistance by some in the social and behavioral sciences (Bollen, 2007).

Some have recently begun to address this problem. Lewbel (2007) developed a correction for treatment misclassification by assuming that, given covariates, classification errors are independent of the treatment received. Imai and Yamamoto (2008) investigated the impact of differential classification error on nonparametric indentification of treatment effects. In contrast, we propose a model that combines potential outcomes with latent-class analysis (Goodman, 1974).

In Section 2, we define our model and present an EM algorithm for computing maximum-likelihood (ML) estimates. Because our model conditions on covariates, ATE's do not appear as model parameters. In Section 3, we define estimating equations for average potential outcomes and replace unseen outcomes with model predictions. This technique, which has been called expected estimating equations (Wang et al., 2008), enables us to average over covariates without specifying their distribution. In Section 4, we extend the procedure to surveys with complex designs, showing how to compute estimates and standard errors under the class of with-replacement designs. In Section 5, we apply these methods to data from the National Longitudinal Study of Adolecent Health (Udry, 2003) to estimate the effects of weight-control behaviors on body-mass index.

2 THE MODEL

2.1 Modeling the Latent Treatment

Latent-class analysis (Goodman, 1974) explains relationships among a set of observed categorical variables by supposing that they are conditionally independent given an unobserved categorical variable. For each individual $i = 1, \ldots, N$, define a latent treatment variable T_i which takes possible values $c = 1, \ldots, C$. The treatment is measured by manifest items $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{iM})$, where U_{im} takes possible values $r = 1, \ldots, r_m$. The realized value of $\boldsymbol{U}_i$ is denoted by $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iM})$. We allow an arbitrary subset of these items to be missing at random (Rubin, 1976), and we partition the items as $\boldsymbol{U}_i = (\boldsymbol{U}_{i,obs}, \boldsymbol{U}_{i,mis})$, where $\boldsymbol{U}_{i,obs}$ is observed and $\boldsymbol{U}_{i,mis}$ is missing. Similarly, we partition $\boldsymbol{u}_i$ as $(\boldsymbol{u}_{i,obs}, \boldsymbol{u}_{i,mis})$. We assume that the U_{im}'s are conditionally independent given T_i, with item-response probabilities

$$\rho_{mr|c} = \Pr(U_{im} = r \,|\, T_i = c).$$

Following Dayton and Macready (1988), we add covariates $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ to the latent-class model through a baseline-category logistic regression (Agresti, 2002). The realized value of $\boldsymbol{X}_i$ is $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$. Define $\gamma_{ic} = \Pr(T_i = c \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i)$, and assume that each γ_{ic} is bounded away from zero. We suppose that

$$\gamma_{ic} = \frac{\exp\left(\boldsymbol{x}_i{}^T \boldsymbol{\alpha}_c\right)}{\sum_{c'=1}^{C} \exp\left(\boldsymbol{x}_i{}^T \boldsymbol{\alpha}_{c'}\right)}, \tag{1}$$

where $\boldsymbol{\alpha}_c = (\alpha_{1c}, \ldots, \alpha_{pc})^T$, $c = 1, \ldots, C$ are coefficients to be estimated. Coefficients for one class, called the baseline or reference class, will be fixed at zero ($\boldsymbol{\alpha}_c = \mathbf{0}$).

Model (1) is analogous to the models commonly used to estimate propensity scores (Rosenbaum and Rubin, 1983). The vector $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iC})^T$ is a multivari-

ate balancing score in the sense that individuals in different treatment groups with identical γ_i's have identical distributions for $\boldsymbol{X}_i$ (Imai and van Dyk, 2004).

2.2 Modeling the Potential Outcomes

Let $\boldsymbol{Y}_i = (Y_i(1), Y_i(2), \ldots, Y_i(C))^T$ denote numeric potential outcomes, where $Y_i(c)$ is the outcome that would be realized if $T_i = c$. The observed outcome for individual i is $Y_{i,obs} = Y_i(T_i)$, and its realized value is $y_{i,obs}$. We suppose that

$$\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i = \boldsymbol{x}_i \;\sim\; N(\,\boldsymbol{\beta}^T \boldsymbol{x}_i,\, \boldsymbol{\Sigma}\,), \tag{2}$$

where $\boldsymbol{\beta}$ is a $p \times C$ matrix of coefficients to be estimated, and Σ is a $C \times C$ covariance matrix. The cth column of $\boldsymbol{\beta}$ will be noted by $\boldsymbol{\beta}_c = (\beta_{1c}, \beta_{2c}, \ldots, \beta_{pc})^T$. The diagonal elements of $\boldsymbol{\Sigma}$ are σ_c^2 for $c = 1, \ldots, C$, and the off-diagonal elements are $\sigma_{cc'}$. Due to the pattern of missing values, the correlations $r_{cc'} = \sigma_{cc'}/(\sigma_c \sigma_{c'})$ are strictly inestimable (Rubin, 1974b), and we set them to zero. However, inferences about ATE's are insensitive to these correlations (Frangakis, Rubin and Zhou, 2002), and our estimates and standard errors do not change if different correlations are used.

To keep the notation simple, we have supposed that the same covariates used to predict T_i are also used to predict $\boldsymbol{Y}_i$. Covariates in (1) and (2) should be drawn from the same pool of potential confounders and prognostic variables (Rubin and Thomas, 2000). However, different subsets of predictors, transformations and interactions may appear in the two models, and thus the two versions of $\boldsymbol{x}_i$ need not be identical. To allow for this possibility in the sections below, individual covariates in models (1) and (2) will be denoted by $x_{ij}^{(\alpha)}$ and $x_{ij}^{(\beta)}$.

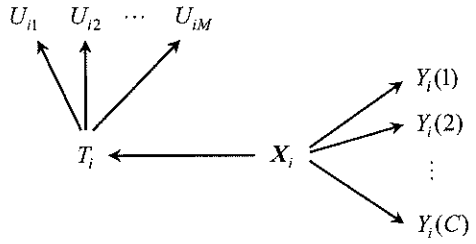

Figure 1: Causal model with latent treatment.

2.3 The Loglikelihood Function

By combining (1) with (2), we make three key assumptions. The first is *unconfounded treatment assignment:* T_i is independent of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$ (Rosenbaum and Rubin, 1983). The second is *unconfounded measurement:* $\boldsymbol{U}_i$ is independent of $\boldsymbol{Y}_i$ given T_i. The third is *local independence:* $U_{i1}, \ldots, U_{iM}$ are mutually independent given T_i. A graph of these relationships is shown in Figure 1. In this graph, an arrow from one variable to another, as in $A \rightarrow B$, indicates that the relationship between them is parameterized as the conditional distribution of B given A.

In many applications, $Y_{i,obs}$ will be measured later than $\boldsymbol{U}_i$, and an individual may drop out before $Y_{i,obs}$ can be seen. If so, we will suppose that $Y_{i,obs}$ is missing at random, and we will still make use of the information in $\boldsymbol{U}_{i,obs}$ to estimate the parameters of the treatment model. Denote the model parameters by $\boldsymbol{\theta} = (\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ and the loglikelihood function given the observed data by $l(\boldsymbol{\theta}) = \sum_{i=1}^{N} l_i(\boldsymbol{\theta})$. If individual i remains in the study long enough for $Y_{i,obs} = y_{i,obs}$ to be seen, then

$$\begin{aligned} l_i(\boldsymbol{\theta}) &= \log \sum_{c=1}^{C} \left\{ \frac{\exp\left(\boldsymbol{x}_i{}^T \boldsymbol{\alpha}_c\right)}{\sum_{c'=1}^{C} \exp\left(\boldsymbol{x}_i{}^T \boldsymbol{\alpha}_{c'}\right)} \right\} \left\{ \prod_{m \in obs_i} \prod_{r=1}^{r_m} \rho_{mr|c}^{I(u_{im}=r)} \right\} \\ &\quad \times \left(2\pi\sigma_c^2\right)^{-1/2} \exp\left\{ -\frac{1}{2\sigma_c^2} \left(y_{i,obs} - \boldsymbol{x}_i{}^T \boldsymbol{\beta}_c\right)^2 \right\}, \end{aligned} \tag{3}$$

where obs_i denotes the subset of $\{1, \ldots, M\}$ corresponding to the items that are

observed for individual i. If the individual drops out prior to realization of $Y_{i,obs}$, then

$$l_i(\boldsymbol{\theta}) = \log \sum_{c=1}^{C} \left\{ \frac{\exp\left(\boldsymbol{x}_i{}^T \boldsymbol{\alpha}_c\right)}{\sum_{c'=1}^{C} \exp\left(\boldsymbol{x}_i{}^T \boldsymbol{\alpha}_{c'}\right)} \right\} \left\{ \prod_{m \in obs_i} \prod_{r=1}^{r_m} \rho_{mr|c}^{I(u_{im}=r)} \right\}. \tag{4}$$

To streamline the notation, we rewrite (3)– (4) as $l_i(\boldsymbol{\theta}) = \log \sum_{c=1}^{C} \gamma_{ic}\, \mathcal{P}_{ic}\, g_{ic}$, where

$$\mathcal{P}_{ic} = \prod_{m \in obs_i} \prod_{r=1}^{r_m} \rho_{mr|c}^{I(u_{im}=r)}$$

and

$$g_{ic} = \exp\left\{ -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma_c^2 - \frac{1}{2\sigma_c^2}\left(y_{i,obs} - \boldsymbol{x}_i{}^T \boldsymbol{\beta}_c\right)^2 \right\}$$

if $Y_{i,obs}$ is seen; if $Y_{i,obs}$ is unseen, define $g_{ic} = 1$. Given the observed data, the posterior probability that individual i belongs to class $T_i = c$ is then

$$\eta_{ic} = \frac{\gamma_{ic}\, \mathcal{P}_{ic}\, g_{ic}}{\sum_{c'=1}^{C} \gamma_{ic'}\, \mathcal{P}_{ic'}\, g_{ic'}}. \tag{5}$$

2.4 An EM Algorithm

To maximize $l(\boldsymbol{\theta})$, we apply an EM algorithm that augments the observed data with assumed values for $T_1, \ldots, T_N$. Define $\mathcal{U}_m$ to be the subset of $\{1, \ldots, N\}$ corresponding to the individuals for whom U_{im} is seen. Similarly, define $\mathcal{Y}$ as the subset of $\{1, \ldots, N\}$ corresponding to the individuals for whom $Y_{i,obs}$ is seen. The augmented-data loglikelihood can be written as the sum of three distinct terms,

$$\begin{aligned} l^*(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \sum_{c=1}^{C} I(T_i = c)\, \log \gamma_{ic} \\ &\quad + \sum_{c=1}^{C} \sum_{m=1}^{M} \sum_{i \in \mathcal{U}_m} \sum_{r=1}^{r_m} I(T_i = c)\, I(u_{im} = r)\, \log \rho_{mr|c} \\ &\quad + \sum_{c=1}^{C} \sum_{i \in \mathcal{Y}} I(T_i = c) \left\{ -\frac{1}{2}\log(2\pi\sigma_c^2) - \frac{1}{2\sigma_c^2}\left(y_{i,obs} - \boldsymbol{x}_i{}^T \boldsymbol{\beta}_c\right)^2 \right\}. \end{aligned}$$

If observed data and parameters are regarded as fixed, l^* becomes a linear function of the indicators $I(T_i = c)$. Therefore, in the E-step of EM, we replace each $I(T_i = c)$ by η_{ic}, where the latter is computed under the current estimate of $\boldsymbol{\theta}$. The M-step separates into three steps corresponding to three terms in $l^*(\boldsymbol{\theta})$. The term

$$\sum_{c=1}^{C} \sum_{m=1}^{M} \sum_{i \in \mathcal{U}_m} \sum_{r=1}^{r_m} \eta_{ic}\, I(u_{im} = r)\, \log \rho_{mr|c}$$

is maximized at

$$\hat{\rho}_{mr|c} = \frac{\sum_{i \in \mathcal{U}_m} \eta_{ic}\, I(u_{im} = r)}{\sum_{i \in \mathcal{U}_m} \eta_{ic}}$$

for $r = 1, \ldots, r_m$, $m = 1, \ldots, M$ and $c = 1, \ldots, C$. The term

$$\sum_{c=1}^{C} \sum_{i \in \mathcal{Y}} \eta_{ic} \left\{ -\frac{1}{2} \log(2\pi\sigma_c^2) - \frac{1}{2\sigma_c^2} (y_{i,obs} - \boldsymbol{x}_i{}^T \boldsymbol{\beta}_c)^2 \right\}$$

is maximized at

$$\begin{aligned} \hat{\boldsymbol{\beta}}_c &= \left(\sum_{i \in \mathcal{Y}} \eta_{ic} \boldsymbol{x}_i \boldsymbol{x}_i{}^T \right)^{-1} \left(\sum_{i \in \mathcal{Y}} \eta_{ic} \boldsymbol{x}_i y_{i,obs} \right), \\ \hat{\sigma}_c^2 &= \left(\sum_{i \in \mathcal{Y}} \eta_{ic} \right)^{-1} \sum_{i \in \mathcal{Y}} \eta_{ic} (y_{i,obs} - \boldsymbol{x}_i{}^T \hat{\boldsymbol{\beta}}_c)^2 \end{aligned}$$

for $c = 1, \ldots, C$.

The maximizer of the term involving the α's cannot in general be written in closed form, but it may be computed by a Newton-Raphson procedure. Let $\boldsymbol{\alpha}$ denote the vector of coefficients α_{jc} for all classes $c = 1, \ldots, C$ except the baseline class. The function to be maximized is $Q_\alpha(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \eta_{ic} \log \gamma_{ic}$, where the γ_{ic}'s are given by (1) and the η_{ic}'s are regarded as fixed. One cycle of Newton-Raphson is

$$\boldsymbol{\alpha}^{(new)} = \boldsymbol{\alpha}^{(old)} + \left[-Q''_\alpha(\boldsymbol{\alpha}^{(old)}) \right]^{-1} Q'_\alpha(\boldsymbol{\alpha}^{(old)}),$$

where $Q'_\alpha(\boldsymbol{\alpha})$ is the vector of first derivatives of $Q_\alpha(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, and $Q''_\alpha(\boldsymbol{\alpha})$ is the matrix of second derivatives. The elements of $Q'_\alpha(\boldsymbol{\alpha})$ are

$$\frac{\partial}{\partial \alpha_{jc}} Q_\alpha = \sum_{i=1}^{N} (\eta_{ic} - \gamma_{ic})\, x_{ij}^{(\alpha)},$$

and the elements of $Q''_\alpha(\boldsymbol{\alpha})$ are

$$\frac{\partial^2}{\partial\alpha_{jc}\partial\alpha_{j'c'}} Q_\alpha = -\sum_{i=1}^{N} \gamma_{ic} \left[I(c = c') - \gamma_{ic'} \right] x_{ij}^{(\alpha)} x_{ij'}^{(\alpha)}.$$

This algorithm has been implemented by the authors in combination of Fortran 95 and R. Procedures have been documented and bundled as an R package for Windows. A beta release of this package, called *LCCA Version 1* (built under R Version 2.10.1), has been made available online as a supplement to this article.

2.5 Standard Errors

EM provides the ML estimates, but it does not automatically give standard errors. We estimate the covariance matrix for $\hat{\boldsymbol{\theta}}$ by

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \left[-\sum_{i=1}^{N} l_i''(\hat{\boldsymbol{\theta}}) \right]^{-1},$$

where $l_i''(\boldsymbol{\theta})$ denotes the Hessian of $l_i(\boldsymbol{\theta})$. Expressions for loglikelihood derivatives are provided in Section 1 of the online supplemental document.

2.6 Starting Values and Boundary Solutions

Depending on the starting values, EM may converge to any one of $C!$ equivalent modes in which the class labels $1, \ldots, C$ have been permuted (Titterington, Smith and Makov, 1985). EM may also converge to local minor modes. Users are advised to repeat the estimation procedure from multiple starting values for the ρ's and compare the loglikelihoods at the solutions to determine if minor modes are present. Starting values for ρ's may be randomly generated from uniform distributions and normalized to satisfy the sum-to-one constraints.

With latent-class models, it is common for some estimated ρ's to approach zero, and boundary solutions make Hessian-based standard errors untenable. When this happens, we apply a flattening constant k to each set $\{\rho_{mr|c} : r = 1, \ldots, r_m\}$ which adds information equivalent to k prior observations spread equally across the categories. A small positive value such as $k = 1$ is often sufficient to nudge the solution away from the boundary. When $k > 0$, EM maximizes a function equal to the loglikelihood plus a penalty term which may be regarded as a Bayesian log-prior density.

2.7 Model Construction

It is helpful to contruct this model in stages. In the first stage, we ignore covariates and outcomes and investigate latent-class models for $\boldsymbol{U}_i$. Models with different values for C should not be compared by standard likelihood ratio tests (Titterington, Smith and Makov, 1985), but statistics such as AIC and BIC may provide some guidance.

In the second stage, we incorporate a rich set confounders and prognostic variables into the model for T_i. Overfitting a propensity model can make estimates of ATE's more efficient (Lunceford and Davidian, 2004). To avoid post-treatment selection bias (Robins and Greenland, 1992), covariates influenced by T_i should not be used. After fitting this model, we generate a single random imputation of T_i from the estimated posterior probabilities (5) (Bandeen-Roche et al., 1997). Distributions of estimated propensities γ_{ic} may then be compared across the imputed treatment groups to assess overlap and imbalance (Gelman and Hill, 2007).

At the third stage, we incorporate predictors for $\boldsymbol{Y}_i$. To mitigate bias due to misspecification of the outcomes model, we have found it helpful to include functions of estimated logit-propensities saved from Stage 2 as predictors for $\boldsymbol{Y}_i$, such as

piecewise-constant terms or a spline bases (Kang and Schafer, 2007).

3 AVERAGE TREATMENT EFFECTS

3.1 Defining the Effects

None of the parameters in $\boldsymbol{\theta}$ are average treatment effects. The coefficients $\boldsymbol{\beta}$ describe the means of each potential outcome *given* the covariates, but ATE's are contrasts among the means of the potential outcomes *averaged over* the covariates. Let $\mu(c) = E(Y_i(c))$ denote the marginal mean of $Y_i(c)$ in the population. The average effect of treatment $T_i = c$ versus $T_i = c'$ is $\mu(c) - \mu(c')$. Similarly, let $\mu(c \mid d) = E(Y_i(c) \mid T_i = d)$ denote the mean of $Y_i(c)$ among individuals receiving $T_i = d$. The average effect of treatment $T_i = c$ versus $T_i = c'$ among those with $T_i = d$ is $\mu(c \mid d) - \mu(c' \mid d)$. In general, $\mu(c) - \mu(c')$ and $\mu(c \mid d) - \mu(c' \mid d)$ are not the same, because in a nonrandomized study the treatment groups are not drawn from the same population.

3.2 Expected Estimating Functions

We now describe a method for estimating $\boldsymbol{\mu} = E(\boldsymbol{Y}_i)$ and $\mu_{(d)} = E(\boldsymbol{Y}_i \mid T_i = d)$ that does not require a model for $\boldsymbol{X}_i$ but averages over it nonparametrically.

If $Y_i(1), \ldots, Y_i(C)$ and T_i were seen for every individual, we could consistently estimate $\mu(c)$ and $\mu(c \mid d)$ with minimal assumptions by

$$\frac{1}{N}\sum_{i=1}^{N} Y_i(c) \quad \text{and} \quad \frac{\sum_{i=1}^{N} I(T_i = d)\, Y_i(c)}{\sum_{i=1}^{N} I(T_i = d)}.$$

These may be regarded as the solutions to the estimating equations

$$\sum_{i=1}^{N} \left(Y_i(c) - \mu(c) \right) = 0 \tag{6}$$

and

$$\sum_{i=1}^{N} I(T_i = d) \left(Y_i(c) - \mu(c|d) \right) = 0. \tag{7}$$

Because $Y_i(c)$ and T_i are unknown, we replace the expressions on the left-hand sides of (6)–(7) by their expected values given the observed data under the model from Section 2. Replacing score functions by their conditional expectations has been described by Wang et al. (2008). In a parametric model, a solution to expected estimating equations is an ML estimate (Wang and Pepe, 2000). Our method is not ML, because we have not specified a distribution for $\boldsymbol{X}_i$. Nevertheless, the estimates will be consistent and asymptotically normal if the model from Section 2 is correct.

To obtain the expected estimating equations, we replace $Y_i(c)$ in (6), and $I(T_i = d)$ and $I(T_i = d)\, Y_i(c)$ in (7), by their expected values given $\boldsymbol{X}_i = \boldsymbol{x}_i$, $\boldsymbol{U}_{i,obs} = \boldsymbol{u}_{i,obs}$, and $Y_{i,obs} = y_{i,obs}$ if the latter is seen. The expected value of $Y_i(c)$ given that $T_i = d$ is

$$\hat{y}_i(c \,|\, d) = \begin{cases} y_{i,obs} & \text{if } Y_{i,obs} \text{ is seen and } c = d, \\ \boldsymbol{x}_i{}^T \boldsymbol{\beta}_c + \left(\dfrac{\sigma_{cd}}{\sigma_d^2} \right) \left(y_{i,obs} - \boldsymbol{x}_i{}^T \boldsymbol{\beta}_d \right) & \text{if } Y_{i,obs} \text{ is seen and } c \neq d, \text{ and} \\ \boldsymbol{x}_i{}^T \boldsymbol{\beta}_c & \text{if } Y_{i,obs} \text{ is unseen.} \end{cases}$$

The expectations of $Y_i(c)$, $I(T_i = d)$, and $I(T_i = d)\, Y_i(c)$ given the observed data are then $\sum_{c'=1}^{C} \eta_{ic'}\, \hat{y}_i(c \,|\, c')$, η_{id} and $\eta_{id}\, \hat{y}_i(c \,|\, d)$, respectively. Plugging these expressions into (6)–(7) and solving the equations gives

$$\hat{\mu}(c) = \frac{1}{N} \sum_{i=1}^{N} \sum_{c'=1}^{C} \eta_{ic'}\, \hat{y}_i(c \,|\, c')$$

and

$$\hat{\mu}(c \,|\, d) = \frac{\sum_{i=1}^{N} \eta_{id}\, \hat{y}_i(c \,|\, d)}{\sum_{i=1}^{N} \eta_{id}}.$$

3.3 Standard Errors

A covariance matrix for $\hat{\boldsymbol{\mu}} = (\hat{\mu}(1), \ldots, \hat{\mu}(C))^T$ may be estimated as follows. Define $\boldsymbol{\omega}_i = (\omega_i(1), \ldots, \omega_i(C))^T$, where

$$\omega_i(c) = \sum_{c'=1}^{C} \eta_{ic'} \hat{y}_i(c \,|\, c') - \mu(c)$$

is the contribution of individual i to the expected estimating function for $\mu(c)$. Define

$$\boldsymbol{S}_i = l_i'(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta})$$

as the vector of derivatives of the loglikelihood defined in Section 2. The estimate $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\mu}}^T)^T$ can be regarded as the solution to stacked estimating equations $\sum_{i=1}^{N} \boldsymbol{\psi}_i = \mathbf{0}$, where $\boldsymbol{\psi}_i = (\boldsymbol{S}_i^T, \boldsymbol{\omega}_i^T)^T$. Under mild regularity conditions, we have $\sqrt{N}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \rightarrow N(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = A^{-1}BA^{-1T}$, $A = -E(\partial \boldsymbol{\psi}_i / \partial \boldsymbol{\phi}^T)$ and $B = E(\boldsymbol{\psi}_i \boldsymbol{\psi}_i^T)$ (Newey and McFadden, 1994). An estimated covariance matrix for $\hat{\boldsymbol{\phi}}$ is

$$\hat{V}(\hat{\boldsymbol{\phi}}) = \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\psi}_i}{\partial \boldsymbol{\phi}^T} \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{\psi}_i \boldsymbol{\psi}_i^T \right) \left(\sum_{i=1}^{N} \frac{\partial \boldsymbol{\psi}_i}{\partial \boldsymbol{\phi}^T} \right)^{-1T}, \tag{8}$$

where all functions on the right-hand side of (8) are evaluated at $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$. The matrix $\partial \boldsymbol{\psi}_i / \partial \boldsymbol{\phi}^T$ has the form

$$\frac{\partial \boldsymbol{\psi}_i}{\partial \boldsymbol{\phi}^T} = \left[\begin{array}{c|c} l_i''(\boldsymbol{\theta}) & \mathbf{0} \\ \hline \dfrac{\partial \boldsymbol{\omega}_i}{\partial \boldsymbol{\theta}^T} & \dfrac{\partial \boldsymbol{\omega}_i}{\partial \boldsymbol{\mu}^T} \end{array} \right].$$

The covariance matrix for $\hat{\boldsymbol{\mu}}_{(d)} = (\hat{\mu}(1|d), \ldots, \hat{\mu}(C|d))^T$ may be estimated in a similar fashion. Regard $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\mu}}_{(d)}^T)^T$ as the solution to stacked estimating equations $\sum_{i=1}^{N} \boldsymbol{\psi}_i = \mathbf{0}$, where $\boldsymbol{\psi} = (\boldsymbol{S}_i^T, \boldsymbol{\omega}_{i(d)}^T)^T$, and $\boldsymbol{\omega}_{i(d)}$ is the vector with elements

$$\omega_i(c \,|\, d) = \eta_{id} \left[\hat{y}_i(c \,|\, d) - \mu(c \,|\, d) \right]$$

for $c = 1, \ldots, C$. The estimated covariance matrix is given by (8), with

$$\frac{\partial \boldsymbol{\psi}_i}{\partial \boldsymbol{\phi}^T} = \left[\begin{array}{c|c} l_i''(\boldsymbol{\theta}) & \mathbf{0} \\ \hline \dfrac{\partial \boldsymbol{\omega}_{i(d)}}{\partial \boldsymbol{\theta}^T} & \dfrac{\partial \boldsymbol{\omega}_{i(d)}}{\partial \boldsymbol{\mu}_{(d)}^T} \end{array}\right].$$

Expressions for the derivatives of $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_{i(d)}$ are provided in Section 2 of the the online supplemental document.

4 COMPLEX SURVEY DESIGNS

4.1 Survey weights

Data from surveys with complex sampling designs are usually accompanied by weights. A weight w_i may be regarded as the number of population individuals represented by individual i. If individuals were sampled with unequal probabilities (i.e., if some groups were oversampled), modeling procedures that ignore the weights can lead to biased estimates of population parameters (Lohr, 1999).

When parametric models are applied to survey data, a standard practice is to compute pseudo-maximum likelihood (PML) estimates (Skinner, 1989). PML maximizes the likelihood for a pseudo-population in which individual i has been "cloned" w_i times. The pseudo-loglikelihood for the model of Section 2 is $l^P(\boldsymbol{\theta}) = \sum_{i=1}^N w_i l_i(\boldsymbol{\theta})$, and this function can be maximized with trivial changes to the EM algorithm of Section 2.4. The E-step is unchanged. The M-steps for $\boldsymbol{\rho}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ become

$$\begin{aligned} \hat{\rho}_{mr|c} &= \frac{\sum_{i \in \mathcal{U}_m} w_i \, \eta_{ic} \, I(u_{im} = r)}{\sum_{i \in \mathcal{U}_m} w_i \, \eta_{ic}}, \\ \hat{\boldsymbol{\beta}}_c &= \left(\sum_{i \in \mathcal{Y}} w_i \eta_{ic} \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \left(\sum_{i \in \mathcal{Y}} w_i \eta_{ic} \boldsymbol{x}_i y_{i,obs} \right), \end{aligned}$$

$$\hat{\sigma}_c^2 = \left(\sum_{i \in \mathcal{Y}} w_i \eta_{ic} \right)^{-1} \sum_{i \in \mathcal{Y}} w_i \eta_{ic} \, (y_{i,obs} - \boldsymbol{x}_i{}^T \hat{\boldsymbol{\beta}}_c)^2,$$

and the M-step for $\boldsymbol{\alpha}$ is performed by Newton-Raphson with derivatives

$$\begin{aligned} \frac{\partial}{\partial \alpha_{jc}} Q_\alpha &= \sum_{i=1}^{N} w_i \, (\eta_{ic} - \gamma_{ic}) \; x_{ij}^{(\alpha)}, \\ \frac{\partial^2}{\partial \alpha_{jc} \partial \alpha_{j'c'}} Q_\alpha &= -\sum_{i=1}^{N} w_i \, \gamma_{ic} \, [\, I(c = c') \; - \; \gamma_{ic'} \,] \; x_{ij}^{(\alpha)} \, x_{ij'}^{(\alpha)}. \end{aligned}$$

The estimated average potential outcomes from Section 3.2 become

$$\begin{aligned} \hat{\mu}(c) &= \frac{\sum_{i=1}^{N} w_i \sum_{c'=1}^{C} \eta_{ic'} \, \hat{y}_i(c \,|\, c')}{\sum_{i=1}^{N} w_i}, \\ \hat{\mu}(c \,|\, d) &= \frac{\sum_{i=1}^{N} w_i \eta_{id} \, \hat{y}_i(c \,|\, d)}{\sum_{i=1}^{N} w_i \eta_{id}}. \end{aligned}$$

4.2 Modeling a Subpopulation

Suppose we want to model a subset of the population (e.g., males 12–18 years old). With a simple random sample, we may discard individuals who do not fit this description. With a complex design this may not be appropriate, because the design may not scale down to the subpopulation. To model a subpopulation, define h_i equal to 1 if individual i is in the subpopulation and 0 otherwise, and in each formula in Section 4.1, replace w_i by $h_i w_i$.

4.3 With-Replacement Designs

Many popular designs can be viewed, at least approximately, as special cases of the following class. The population is divided into $S \geq 1$ sampling strata indexed by $s = 1, \ldots, S$. Within stratum s, primary clusters $c = 1, \ldots, C_s$ are selected with replacement. Within primary cluster c in stratum s, individuals $i = 1, \ldots, N_{cs}$

are sampled by any method, possibly in multiple stages, so that the total sample size is $N = \sum_{s=1}^{S} \sum_{c=1}^{C_s} N_{cs}$. This is known as the "with replacement" (WR) class, and many popular software packages for analyzing survey data assume it by default. Design information is conveyed by three user-supplied variables: the individual's survey weight, the cluster identifier (if $n_{cs} > 1$), and the stratum identifier (if $S > 1$). In most surveys, sampling is done without replacement (WOR) to insure that no cluster or individual is selected twice. When sampling is WOR, standard errors computed under a WR assumption tend to be conservative (Wolter, 2007).

4.4 Standard Errors for WR Designs

With WR designs, it is convenient to index sampled individuals by the three subscripts i, c and s. For example, the survey weight w_i becomes w_{ics}.

Variance estimates for WR designs may be obtained as follows. Define the pseudo-score vector for individual i in cluster c and stratum s by

$$\boldsymbol{S}_{ics}(\boldsymbol{\theta}) = h_{ics}\, w_{ics}\, \frac{\partial}{\partial \boldsymbol{\theta}}\, l_{ics}(\boldsymbol{\theta}),$$

where $l_{ics}(\boldsymbol{\theta})$ is the individual's contribution to the loglikelihood function. Let $\boldsymbol{\omega}_{ics}$ denote the vector of expected estimating functions for $\boldsymbol{\mu} = E(\boldsymbol{Y}_i)$, which is defined as in Section 3.3 except that each function is now multiplied by $h_{ics}w_{ics}$. Stacking the estimating functions as $\boldsymbol{\psi}_{ics} = (\boldsymbol{S}_{ics}^T, \boldsymbol{\omega}_{ics}^T)^T$, the joint estimate $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\mu}}^T)^T$ can be regarded as the solution to $\sum_{c,s,i} \boldsymbol{\psi}_{ics}(\boldsymbol{\theta}) = \mathbf{0}$. The estimated covariance matrix for $\hat{\boldsymbol{\phi}}$ comes from a modified sandwich formula,

$$V(\hat{\boldsymbol{\theta}}) = \left(-\sum_{s,c,i} \boldsymbol{\psi}'_{ics}\right)^{-1} \left(\sum_{s,c} \left(\boldsymbol{\psi}_{cs} - \bar{\boldsymbol{\psi}}_s\right)\left(\boldsymbol{\psi}_{cs} - \bar{\boldsymbol{\psi}}_s\right)^T\right) \left[\left(-\sum_{s,c,i} \boldsymbol{\psi}'_{ics}\right)^{-1}\right]^T,$$

where $\boldsymbol{\psi}_{cs} = \sum_i \boldsymbol{\psi}_{ics}$ is the total within cluster c in stratum s, $\bar{\boldsymbol{\psi}}_s = C_s^{-1} \sum_{c=1}^{C_s} \boldsymbol{\psi}_{cs}$ is the average of the cluster totals within stratum s (Wolter, 2007). An analogous procedure gives variance estimates for $\boldsymbol{\mu}_{(d)} = E(\boldsymbol{Y}_i \,|\, T_i = d)$.

5 ANALYZING THE EFFECTS OF WEIGHT-CONTROL BEHAVIORS

5.1 Motivation

Dieting, defined as a voluntary and temporary reduction of caloric intake, has been shown to predict weight gain in cross-sectional and longitudinal studies (Field et al., 2003; Stice et al., 1999, 2005; Neumark-Sztainer et al., 2006). It also predicts anxiety and depression (Kovacs, Obrosky and Sherrill, 2003), decreased cognitive performance (Green and Rogers, 1998), eating disorders (Patton et al., 1999) and emotional distress (Neumark-Sztainer and Hannan, 2000). Yet the effects of dieting are difficult to ascertain for the following reasons. First, dieting is self-selected. Dieters and non-dieters differ in ways that may confound relationships between dieting and outcomes. Second, dieting is self-defined. Behaviors identified as dieting vary across individuals, peers groups and cultural contexts. Third, dieting may supplement or supplant other weight-control strategies (e.g., exercise), making it difficult to separate the effects of dieting from those of concurrent behaviors.

Using data from a large survey of adolescents from the United States, we identified patterns of weight-control behavior and investigated effects of these behaviors on body weight five years later.

5.2 Data

The National Longitudinal Study of Adolescent Health (Add Health) (Udry, 2003) is a nationally representative study of youth risk behaviors. Researchers sampled 132 high schools within strata defined by region, urbanicity, school size, school type and percent minority enrollment. Within each school, a core sample of students was selected with probabilities depending on school size and student characteristics. Students were interviewed in grades 7–12 (Wave I, 1994–95) with reinterviews one year later (Wave II, 1996) and at age 18–26 (Wave III, 2001-02). Items at each wave covered a broad range of health-related attitudes and behaviors.

The treatments in this analysis are weight-control strategies inferred from Wave II. The outcome is log body-mass index (LOGBMI) at Wave III based on self-reported height and weight. Sixty-one baseline measures thought to be related to the treatments or the outcome were considered as covariates. To reduce the possibility that we might inadvertently adjust for intermediate outcomes along the causal pathway, all covariates were drawn from Wave I. Two key covariates were LOGBMI at baseline, which is strongly related to the outcome, and self-perceived weight relative to peers, which is strongly predictive of the treatments. Our analyses are based on $N = 6,679$ girls interviewed at Wave II who had nonmissing values for these two covariates.

5.3 Treatment Classes

Weight-control strategies were measured by eight items. The first item, TRYWEIGHT, asked whether the girl was trying to 1=lose weight, 2=gain weight, 3=stay the same or 4=not trying to do anything about weight. Remaining items (DIETED, EXERCISED, VOMITED, DIETPILLS, LAXATIVES, OTHER, NONE) asked what she

Table 1: Fit statistics for latent-class models of naturalistic weight-control strategies: number of parameters, loglikelihood, AIC, and BIC.

Classes	Params	-2 Loglik	AIC	BIC
$C = 1$	17	86,864	86,898	87,013
$C = 2$	35	38,717	38,787	39,025
$C = 3$	53	32,910	33,016	33,377
$C = 4$	71	31,684	31,826	32,309
$C = 5$	89	31,195	31,373	31,979
$C = 6$	107	31,023	31,237	31,966

did to lose or maintain weight during the last seven days. Those items were skipped if TRYWEIGHT=2 or 4, so we coded them as 1=yes, 2=no, 3=legitimate skip.

Ignoring the covariates, we examined latent-class models with 1–6 classes. We did not account for the sampling design at this stage, because statistics traditionally used to compare latent-class models are not defined for PML. Summaries of fit (number of parameters, loglikelihood, AIC and BIC) are reported in Table 1. The fit appears to improve dramatically as each new class is added, which is typical behavior for a large sample. We fit each model 100 times using different random starting values. For the models with 1–3 classes, all 100 runs converged to equivalent solutions. With four classes, a minor mode appeared in one solution. With five and six classes, the majority of solutions corresponded to numerous minor modes. Comparing major modes for the four- and five-class models, the class prevalences and item-response probabilities looked similar; the only major difference was that, in the five-class model, a small class (estimated at 3.5% of the population) emerged consisting of girls who tended to respond OTHER=1. Because this class was rare and difficult to interpret, we decided to proceed using a model with four classes.

We refit the four-class model by PML using the Add Health cluster and stratum

Table 2: Estimated prevalences and item-response probabilities for Classes 1-3 in the four-class model.

	Class 1	Class 2	Class 3
Prevalence	0.286	0.324	0.191
TRYWEIGHT=1	0.903	0.484	0.182
DIETED=1	0.723	0.015	0.000
EXERCISED=1	0.704	1.000	0.000
VOMITED=1	0.022	0.000	0.000
DIETPILLS=1	0.059	0.000	0.000
LAXATIVES=1	0.014	0.000	0.000
OTHER=1	0.114	0.016	0.000
NONE=1	0.000	0.000	1.000

identifiers and Wave II sampling weights. Parameter estimates changed little, and standard errors for the class prevalences became 30–80% larger. The members of Class 4, estimated to be 19.9% of the population, answered TRYWEIGHT=2 or 4 and skipped the remaining items with probability 1; this class represents girls who were not trying to lose or maintain weight. The remaining classes described girls who were applying various weight-control strategies. Estimated prevalences and item-response probabilities for Classes 1–3 are reported in Table 2. Class 1 (28.6%) contains girls who were likely to be trying to lose weight by one or more strategies, but those strategies varied. Essentially all girls who dieted are in this class, but many of them were also exercising or doing other things. Class 2 (32.4%) was trying to lose or maintain weight through exercise alone. Class 3 (19.1%) consists of girls who said were trying to lose weight or stay the same (mostly the latter) but were not doing anything (NONE=1). Based on these interpretations, we will refer to Classes 1–3 as "Try Something," "Exercise Only," and "Do Nothing," respectively.

5.4 Predicting the Treatments

Based on a review of the dieting literature, we created a pool of 61 baseline measures for predicting the treatment. Known correlates of dieting include race, ethnicity, body weight, body image, pubertal timing, female physical development, self-esteem, academic performance, parental and peer relationships, disinhibited eating, externalizing behaviors, emotional distress and use of licit and illicit substances. Variables from Add Health Wave I related to these were included in the pool. We also included variables that seemed useful for predicting subsequent BMI (e.g., parental obesity and diabetes). Occasional missing values in the baseline variables (usually 2% or less) were handled by mean imputation, by assignment to the modal class (for highly skewed binary variables) or by creating dummy indicators for missingness. Such ad hoc missing-data procedures are not generally recommended for regression analyses (Little and Rubin, 2002), but it is reasonable to use them in propensity-score modeling, because the purpose of a propensity model is prediction, not interpretation (D'Agostino and Rubin 2000), and when the ATE's are estimated, these variables are averaged out. Descriptions of these variables, and the Add Health items from which they were derived, are given in Section 3 of the online supplement.

Using Class 4 as the reference group, we regressed the latent treatment on each covariate, one at a time. Nearly all were significantly related to the treatment as judged by likelihood-ratio tests, so we decided to use all of them in the treatment model. Estimated prevalances and item-response probabilities changed slightly when the covariates were introduced, but they exhibited the same pattern as in Table 2. After fitting this model, we saved the estimated linear predictors $\boldsymbol{x}_i{}^T\hat{\boldsymbol{\alpha}}_c$ and posterior probabilities $\hat{\eta}_{ic}$. We drew a single random imputation of T_i from the $\hat{\eta}_{ic}$'s to compare

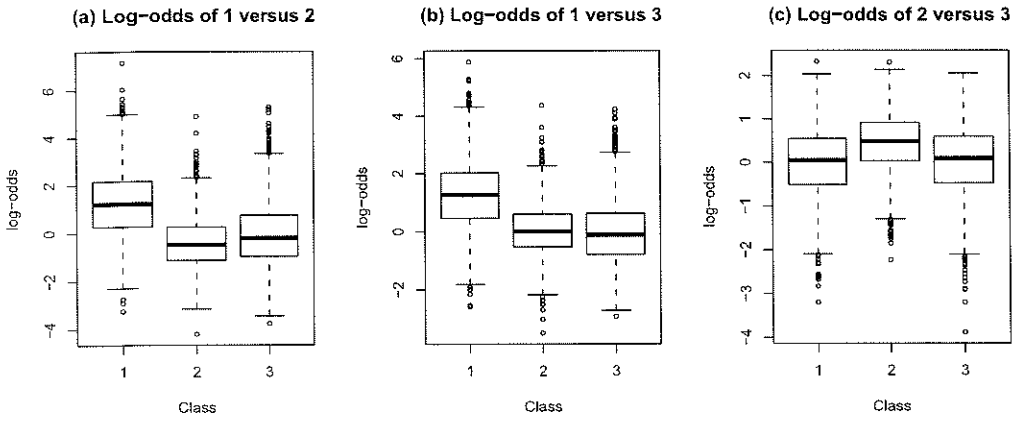

Figure 2: Boxplots of estimated log-odds by imputed treatment class.

propensity distributions across classes. Because we are interested in comparisons among Classes 1–3, we created boxplots of $\log(\hat{\gamma}_{ic}/\hat{\gamma}_{c'}) = \boldsymbol{x}_i{}^T\hat{\boldsymbol{\alpha}}_c - \boldsymbol{x}_i{}^T\hat{\boldsymbol{\alpha}}_{ic'}$ for each pair c and c' in Classes 1–3 (Figure 2). These distributions overlap well, but some of the comparisons are quite unbalanced. For example, in Figure 2 (a), the lower quartile for Class 1 coincides with the upper quartile for Class 2. As imbalance increases, causal inferences become increasingly sensitive to misspecification of the model for the outcomes. This bias can be mitigated by enlarging the model for $\boldsymbol{Y}_i$ to allow outcomes to vary with propensities in a flexible way (Little and An, 2004).

5.5 Predicting the Potential Outcomes

All 61 covariates from the treatment model were used to predict $\boldsymbol{Y}_i$. We also included a linear spline basis for each of the three logits displayed in Figure 2 with knots at the sample quintiles. Similar strategies were recommended by Little and An (2004) and Kang and Schafer (2007) to protect against biases that may arise from a misspecified model. Because each logit is an exact linear combination of the 61 covariates, the logits themselves were omitted from the spline bases to avoid redundancy.

Table 3: Estimated means of LOGBMI under Treatments 1, 2, and 3 for the full population and within each treatment class, with standard errors

Domain	*Treatment*	Est.	SE
Full Population	Try Something	3.231	0.008
	Exercise Only	3.171	0.014
	Do Nothing	3.202	0.008
Try Something	Try Something	3.330	0.009
	Exercise Only	3.232	0.028
	Do Nothing	3.302	0.013
Exercise Only	Try Something	3.179	0.013
	Exercise Only	3.134	0.009
	Do Nothing	3.147	0.011
Do Nothing	Try Something	3.193	0.010
	Exercise Only	3.151	0.012
	Do Nothing	3.164	0.009

5.6 Results

The PML estimation algorithm for the combined treatment and outcomes model converged in 109 iterations. The procedure took approximately 400 seconds on a 2.60 GHz dual-CPU Windows computer. Parameter estimates and standard errors are provided in Section 4 of the online supplement.

The estimated marginal means $\mu(c) = E(Y_i(c))$ and $\mu(c \mid d) = E(Y_i = c \mid T_i = d)$ for treatment classes 1, 2 and 3 are displayed in Table 3. The estimated mean of LOGBMI is highest under Treatment 1 (Try Something), followed by Treatment 3 (Do Nothing), followed by Treatment 2 (Exercise). This same pattern appears in the full population and in the subpopulations defined by the treatments.

Estimated average treatment effects $100(\mu(c) - \mu(c'))$ and $100(\mu(c \mid d) - \mu(c' \mid d))$ are shown in Table 4. Because BMI is expressed on the log scale, these effects may be interpreted as approximate percent changes in body weight. The largest effect

Table 4: Estimated (average treatment effects $\times 100$) comparing Treatments 1, 2, and 3 for the full population and within each treatment class, with standard errors and p-values

Domain	*Comparison*	Est.	SE	p
Full Population	Exercise Only vs. Try Something	−5.98	1.50	.000
	Do Nothing vs. Try Something	−2.87	0.82	.001
	Do Nothing vs. Exercise	3.11	1.54	.044
Try Something	Exercise Only vs. Try Something	−9.73	2.82	.000
	Do Nothing vs. Try Something	−2.74	1.08	.012
	Do Nothing vs. Exercise	7.00	2.98	.019
Exercise Only	Exercise Only vs. Try Something	−4.44	1.25	.000
	Do Nothing vs. Try Something	−3.18	1.18	.007
	Do Nothing vs. Exercise	1.26	1.03	.219
Do Nothing	Exercise Only vs. Try Something	−4.15	1.35	.002
	Do Nothing vs. Try Something	−2.93	0.87	.001
	Do Nothing vs. Exercise	1.22	1.25	.329

(−9.73) is the comparison of Exercise Only versus Try Something for girls in the Try Something class. It suggests that, if the girls who engaged in Try-Something behavior had instead chosen to Exercise Only, their body weight after five years would have been about 10% lower than it was. If these same girls had chosen to Do Nothing, their weight would have been about 3% lower than it was. Within each treatment group, Try Something produces significant increases in weight relative to Do Nothing and Exercise Only. Among girls who chose to Exercise Only, their weight is estimated to be about 4% lower than if they had chosen to Try Something, and about 3% lower than if they had chosen to Do Nothing. In the Do Nothing group, the decision to Do Nothing rather than Try Something appears to be beneficial, but if they had instead chosen to Exercise Only, there would have been little change.

6 DISCUSSION

Because most girls who dieted belonged to the Try-Something class, our results seem consistent with previous findings that dieting leads to weight gain. However, these data do not allow us to estimate a pure effect of dieting, because dieting often appears alongside other weight-control strategies. Observational studies of naturalistic behaviors do not always produce neat treatment groups for testing *a priori* causal hypotheses. The hypotheses testable from these data are comparisons among exercising, doing nothing, and a state in which girls attempt to lose weight by various means but perhaps without firm commitment. We speculate that this Try-Something state may act as a mental substitute for strategies that would be effective (e.g., sustained restrained eating or consistent exercise). In fact, this state appears to be decidedly worse than doing nothing. Our results suggest that behavioral interventions designed to move and keep girls out of this Try-Something state might be effective.

Although we have adjusted for many baseline covariates, the possibility remains that the effects in Table 4 are distorted by unmeasured confounders. Methods for assessing sensitivity to unmeasured confounders (Rosenbaum, 2002) in this latent-treatment setting are an important topic for future research.

The model developed in Section 2 may be extended in various ways. The linear regressions for $\boldsymbol{Y}_i$ could be replaced by logistic or loglinear models for binary outcomes or frequencies. The estimating functions given in Section 3 for estimating $E(\boldsymbol{Y}_i)$ and $E(\boldsymbol{Y}_i \,|\, T_i = c)$ could be replaced by estimating functions for parameters of $E(\boldsymbol{Y}_i \,|\, \boldsymbol{Z}_i)$ and $E(\boldsymbol{Y}_i \,|\, T_i = c, \boldsymbol{Z}_i)$, where $\boldsymbol{Z}_i$ is a vector of covariates that moderate the treatment effects. This would allow us to fit marginal structural models (Robins, Hernan and Brumback, 2000) in situations where the treatment is a latent class.

SUPPLEMENTAL MATERIALS

Document: Supplement to "Estimating Average Causal Effects When the Putative Cause is a Latent Class" (PDF).

Software: *LCCA: Latent-Class Causal Analysis*, software package for Windows R (ZIP archive).

Document: *LCCA Package for R, Version 1 (Beta)* (PDF file)

REFERENCES

Agresti, A. (2002), *Categorical Data Analysis* (2nd Ed.). New York: Wiley.

Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L., and Rathouz, P.R. (1997), "Latent Variable Regression for Multiple Discrete Outcomes," *Journal of the American Statistical Association*, 92, 1375–1386.

Bollen, K.A. (2007), "Causality and Structural Equation Models." Presented at International Meeting of the Psychometric Society, July 11, 2007, Tokyo, Japan.

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.

D'Agostino, R.B., Jr., Rubin, D.B. (2000), "Estimating and Using Propensity Scores with Partially Missing Data," *Journal of the American Statistical Association*, 95, 749-759.

Dayton, C.M. and Macready, G.B. (1988), "Concomitant-Variable Latent-Class Models,' *Journal of the American Statistical Association*, 83, 173-178.

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.

Field, A.E., Austin, S.B., Taylor, C.B., Malspeis, S., Rosner, B., Rockett, H.R., Gillman, M.W., and Colditz, G.A. (2003), "Relation Between Dieting and Weight Change Among Preadolescents and Adolescents," *Pediatrics*, 112, 900–906.

Frangakis, C.E., Rubin, D.B. and Zhou, X.-H. (2002), "Clustered Encouragement Designs with Individual Noncompliance: Bayesian Inference with Randomization, and Application to Advance Directive Forms," *Biostatistics*, 3, 147–164.

Gelman, A., Hill, J. (2007) *Data Analysis Using Regression and Multi- level/Hierarchical Models*, New York: Cambridge University Press.

Gelman, A. and Meng, X.L. (Eds.) (2004), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, New York: Wiley.

Goodman, L. A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika*, 61, 215–231.

Green, M.W. and Rogers, P.J. (1998), "Impairments in Working Memory Associated with Spontaneous Dieting Behaviour," *Psychological Medicine*, 28, 1063–1070.

Imai, K. and van Dyk, D.A. (2004), "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical*

Association, 99, 854–866.

Imai, K. and Yamamoto, T. (2008), "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis," *American Journal of Political Science*, 54, 543-560.

Kang, J.D.Y., Schafer, J.L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating Population Means from Incomplete Data," (with discussion and rejoinder), *Statistical Science*, 26, 523–539.

Kovacs, M., Obrosky, D.S. and Sherrill, J. (2003), "Developmental Changes in the Phenomenology of Depression in Girls Compared to Boys from Childhood Onward," *Journal of Affective Disorders*, 74, 33–48.

Lewbel, A. (2007), "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 75, 537-551.

Little, R.J.A. and An, H. (2004), "Robust Likelihood-Based Analysis of Multivariate Data with Missing Values," *Statistica Sinica*, 14, 949–968.

Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data* (2nd ed.), New York: Wiley.

Lohr, S. (1999), *Sampling: Design and analysis*, Pacific Grove, CA: Duxbury Press.

Lunceford, J.K. and Davidian, M. (2004), "Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study," *Statistics in Medicine*, 23, 2937-2960.

Neumark-Sztainer, D. and Hannan, P.J. (2000), "Weight-Related Behaviors Among

Adolescent Girls and Boys: Results from a National Survey," *Archives of Pediatrics and Adolescent Medicine*, 154, 569–577.

Neumark-Sztainer, D. and Story, M. (1998), "Dieting and Binge Eating Among Adolescents: What Do They Really Mean?" *Journal of the American Dietetic Association*, 98, 446–450.

Neumark-Sztainer, D., Wall, M., Guo, J., Story, M., Haines, J., and Eisenberg, M. (2006), "Obesity, Disordered Eating, and Eating Disorders in a Longitudinal Study of Adolescents: How Do Dieters Fare 5 Years Later?" *Journal of the American Dietetic Association*, 106, 559–568.

Newey, W. K. and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," In Z. Griliches (Ed.), *Handbook of Econometrics*, 4, 2111-2245. Amsterdam: Elsevier.

Patton G.C., Selzer, R., Coffey, C., Carlin, J.B. and Wolfe, R. (1999), "The Onset of Adolescent Eating Disorders: A Population Based Cohort Study Over Three Years," *British Medical Journal*, 318, 765–768.

Robins, J.M. and Greenland, S. (1992), "Identifiability and Exchangeability of Direct and Indirect Effects," *Epidemiology*, 3, 143–155.

Robins, J.M., Hernan, M. and Brumback, B. (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560.

Rosenbaum, P.R. (2002), *Observational Studies* (2nd Ed.), New York: Springer.

Rosenbaum, P.R., Rubin, D.B. (1983), "The Central Role of the Propensity Score in

Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rubin, D.B. (1974a), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Rubin, D.B. (1974b), "Characterizing the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association*, 69, 467–474.

Rubin, D.B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

Rubin, D.B. (2005), "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, 100, 322–331.

Rubin, D. B. and Thomas, N. (2000), "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates," *Journal of the American Statistical Association*, 95, 573585.

Schafer, J.L., Kang, J. (2008), "Average Causal Effects from Non-Randomized Studies: A Practical Guide and Simulated Example," *Psychological Methods*, 13, 279–313.

Skinner, C.J. (1989), "Domain Means, Regression and Multivariate Analysis," In *Analysis of Complex Surveys*, C.J. Skinner, D. Holt and T.F.M. Smith (Eds.), 59–87, Chichester: Wiley.

Stice, E., Cameron, R.P., Killen, J.D., Hayward, C., and Taylor, C.B. (1999), "Naturalistic Weight-Reduction Efforts Prospectively Predict Growth in Relative Weight and Onset of Obesity Among Female Adolescents," *Journal of Consulting and Clinical Psychology*, 67, 967–974.

Stice, E., Presnell, K., Shaw, H. and Rohde, P. (2005), "Psychological and Behavioral

Risk Factors for Obesity Onset in Adolescent Girls: A Prospective Study. *Journal of Consulting and Clinical Psychology*, 73, 195–202.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Udry, J.R. (2003) *The National Longitudinal Study of Adolescent Health (Add Health), Waves I and II, 1994–1996; Wave III, 2001–2002* (machine-readable data file and documentation), Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

Wang, C.Y., Huang, Y., Chao, E.C., and Jeffcoat, M.K. (2008), "Expected Estimating Equations for Missing Data, Measurement Error, and Misclassification, With Application to Longitudinal Nonignorable Missing Data," *Biometrics*, 64, 85–95.

Wang, C.Y. and Pepe, M.S. (2000), "Expected Estimating Equations to Accommodate Covariate Measurement Error," *Journal of the Royal Statistical Society Series B*, 62, 509–524,

Wolter, K.M. (2007), *Introduction to Variance Estimation*, Second Edition, New York: Springer.