Matching

James J. Heckman

Econ 374, Spring 2011
This draft, May 3, 2011
Introduction

The assumption most commonly made to circumvent problems with randomization is that even though \( D \) is not random with respect to potential outcomes, the analyst has access to variables \( X \) that effectively produce a randomization of \( D \) with respect to \((Y_0, Y_1)\) given \( X \).
• This is the method of matching, which is based on the conditional independence assumption

\[ (Y_0, Y_1) \perp \!\!\!\!\!\!\!\!\!\perp D \mid X. \] (M-1)

• Conditioning on \( X \) randomizes \( D \) with respect to \( (Y_0, Y_1) \).

• (M-1) assumes that any selective sampling of \( (Y_0, Y_1) \) can be adjusted by conditioning on observed variables.

• Randomization and (M-1) are different assumptions and neither implies the other.
In a linear equations model, assumption (M-1) that $D$ is independent from $(Y_0, Y_1)$ given $X$ justifies application of least squares on $D$ to eliminate selection bias in mean outcome parameters.

For means, matching is just nonparametric regression.

In order to be able to compare $X$-comparable people in the treatment regime one must assume

$$0 < \Pr(D = 1 \mid X = x) < 1.$$  \hfill (M-2)
Assumptions (M-1) and (M-2) justify matching.

Assumption (M-2) is required for any evaluation estimator that compares treated and untreated persons.

It is produced by random assignment if the randomization is conducted for all $X = x$ and there is full compliance.
Observe that from (M-1) and (M-2), it is possible to identify $F_1(Y_1 | X = x)$ from the observed data $F_1(Y_1 | D = 1, X = x)$, since we observe the left hand side of

$$F_1(Y_1 | D = 1, X = x) = F_1(Y_1 | X = x) = F_1(Y_1 | D = 0, X = x).$$

The first equality is a consequence of conditional independence assumption (M-1).

The second equality comes from (M-1) and (M-2).
By a similar argument, we observe the left hand side of

$$F_0(Y_0 \mid D = 0, X = x) = F_0(Y_0 \mid X = x)$$

$$= F_0(Y_0 \mid D = 1, X = x).$$

The equalities are a consequence of (M-1) and (M-2).

Since the pair of outcomes \((Y_0, Y_1)\) is not identified for anyone, as in the case of data from randomized trials, the joint distributions of \((Y_0, Y_1)\) given \(X\) or of \(Y_1 - Y_0\) given \(X\) are not identified without further information.

This is a problem that plagues all selection estimators.
From the data on $Y_1$ given $X$ and $D = 1$ and the data on $Y_0$ given $X$ and $D = 0$ it follows that

\[
E(Y_1 | D = 1, X = x) = E(Y_1 | X = x)
\]

\[
= E(Y_1 | D = 0, X = x)
\]

and

\[
E(Y_0 | D = 0, X = x) = E(Y_0 | X = x)
\]

\[
= E(Y_0 | D = 1, X = x).
\]
Thus,

\[ E(Y_1 - Y_0 \mid X = x) = E(Y_1 - Y_0 \mid D = 1, X = x) \]

\[ = E(Y_1 - Y_0 \mid D = 0, X = x). \]

Effectively, we have a randomization for the subset of the support of \( X \) satisfying (M-2).
At values of $X$ that fail to satisfy M-2, there is no variation in $D$ given $X$. One can define the residual variation in $D$ not accounted for by $X$ as

$$
\mathcal{E}(x) = D - E(D \mid X = x) = D - \Pr(D = 1 \mid X = x).
$$
If the variance of $\mathcal{E}(x)$ is zero, it is not possible to construct contrasts in outcomes by treatment status for those $X$ values and (M-2) is violated.

To see the consequences of this violation in a regression setting, use $Y = Y_0 + D(Y_1 - Y_0)$ and take conditional expectations, under (M-1), to obtain

$$E(Y \mid X, D) = E(Y_0 \mid X) + D[E(Y_1 - Y_0 \mid X)].$$

If $\text{Var}(\mathcal{E}(x)) > 0$ for all $x$ in the support of $X$, one can use nonparametric least squares to identify

$$E(Y_1 - Y_0 \mid X = x) = \text{ATE}(x)$$

by regressing $Y$ on $D$ and $X$. 
The function identified from the coefficient on $D$ is the average treatment effect.

If $\text{Var}(\mathcal{E}(x)) = 0$, $\text{ATE}(x)$ is not identified at that $x$ value because there is no variation in $D$ that is not fully explained by $X$. 

A special case of matching is linear least squares where one can write

\[ Y_0 = X\alpha + U \quad Y_1 = X\alpha + \beta + U. \]

\( U_0 = U_1 = U \), and hence under (M-1)

\[ E(Y \mid X, D) = X\alpha + \beta D + E(U \mid X). \]
If $D$ is perfectly predictable by $X$, one cannot identify $\beta$ because of a multicollinearity problem.

(M-2) rules out perfect collinearity.

Matching is a nonparametric version of least squares that does not impose functional form assumptions on outcome equations, and that imposes support condition (M-2).

It identifies $\beta$ but not necessarily $\alpha$.

(look at the term $E(U \mid X)$).
Conventional econometric choice models make a distinction between variables that appear in outcome equations \((X)\) and variables that appear in choice equations \((Z)\).

The same variables may be in \((X)\) and \((Z)\), but more typically there are some variables not in common.
• Matching makes no distinction between the $X$ and the $Z$.

• It does not rely on exclusion restrictions.

• The conditioning variables used to achieve conditional independence can in principle be a set of variables $Q$ distinct from the $X$ variables (covariates for outcomes) or the $Z$ variables (covariates for choices).

• I use $X$ solely to simplify the notation.
The key identifying assumption is the assumed existence of a random variable $X$ with the properties satisfying (M-1) and (M-2).

Conditioning on a larger vector ($X$ augmented with additional variables) or a smaller vector ($X$ with some components removed) may or may not produce suitably modified versions of (M-1) and (M-2).

Without invoking further assumptions there is no objective principle for determining what conditioning variables produce (M-1).
Assumption (M-1) is strong.

Many economists do not have enough faith in their data to invoke it.

Assumption (M-2) is testable and requires no act of faith.

To justify (M-1), it is necessary to appeal to the quality of the data.
Below, I show the biases that can result in matching when standard econometric model selection criteria are applied to pick the $X$ that are used to satisfy (M-1).

Conditional independence condition (M-1) cannot be tested without maintaining other assumptions.

Choice of the appropriate conditioning variables is a problem that plagues *all* econometric estimators.
Matching on $P = p$ (which is equivalent to nonparametric regression given $P = p$) produces a biased estimator of $\Delta TT(p)$. 

Matching assumes a flat MTE (average return equals marginal return) as we develop below.

Therefore it is systematically biased for $\Delta^{TT}(p)$ in a model with essential heterogeneity, where $\beta \perp \!\!\!\!\perp D$. 
Figure 1: Treatment Parameters and OLS/Matching as a function of $P(Z) = p$

Parameter Definition Under Assumptions (*)

Marginal Treatment Effect

$$E[Y_1 - Y_0 | D^* = 0, P(Z) = p] = \bar{\beta} + \sigma U_1 - U_0 \Phi^{-1}(1 - p)$$

Average Treatment Effect

$$E[Y_1 - Y_0 | P(Z) = p] = \beta$$

Treatment on the Treated

$$E[Y_1 - Y_0 | D^* > 0, P(Z) = p] = \bar{\beta} + \sigma U_1 - U_0 \phi(\Phi^{-1}(1 - p))$$

Treatment on the Untreated

$$E[Y_1 - Y_0 | D^* \leq 0, P(Z) = p] = \bar{\beta} - \sigma U_1 - U_0 \phi(\Phi^{-1}(1 - p))$$

Note: $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively. $\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$.

(*): The model in this case is the same as the one presented below Figure 5.

Source: Heckman, Urzua and Vytlacil (2006)
Parameter | Definition | Under Assumptions (*)
--- | --- | ---
Marginal Treatment Effect | $E [Y_1 - Y_0 | D^* = 0, P(Z) = p]$ | $\bar{\beta} + \sigma_{U_1 - U_0} \Phi^{-1}(1 - p)$
Average Treatment Effect | $E [Y_1 - Y_0 | P(Z) = p]$ | $\bar{\beta}$
Treatment on the Treated | $E [Y_1 - Y_0 | D^* > 0, P(Z) = p]$ | $\bar{\beta} + \sigma_{U_1 - U_0} \frac{\phi(\Phi^{-1}(1-p))}{\Phi^{-1}(1-p)}$
Treatment on the Untreated | $E [Y_1 - Y_0 | D^* \leq 0, P(Z) = p]$ | $\bar{\beta} - \sigma_{U_1 - U_0} \frac{\phi(\Phi^{-1}(1-p))}{1-p}$
OLS/Matching on $P(Z)$ | $E [Y_1 | D^* > 0, P(Z) = p] - E [Y_0 | D^* \leq 0, P(Z) = p]$ | $\bar{\beta} + \left( \frac{\sigma_{U_1 - U_0}^2}{\sqrt{\sigma_{U_1 - U_0}^2}} \right) \left( \frac{1-2p}{p(1-p)} \right) \phi(\Phi^{-1}(1-p))$

Note: $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively. $\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$.

Source: Heckman, Urzua and Vytlačil (2006)
Making observables alike makes the unobservables dissimilar.

Holding $p$ constant across treatment and control groups understates $TT(p)$ for low values of $p$ and overstates it for high values of $p$.

I develop this point further in the next section.
Method of Matching

- The method of matching assumes selection of treatment based on potential outcomes

\[(Y_0, Y_1) \perp \!
\!\!\perp D,\]

so \(\Pr (D = 1 \mid Y_0, Y_1)\) depends on \(Y_0, Y_1\).

- It assumes access to variables \(Q\) such that conditioning on \(Q\) removes the dependence:

\[(Y_0, Y_1) \perp \!
\!\!\perp D \mid Q.\]  \hspace{1cm} (Q-1)

- Thus,

\[\Pr (D = 1 \mid Q, Y_0, Y_1) = \Pr (D = 1 \mid Q).\]
Comparisons between treated and untreated can be made at all points in the support of $Q$ such that

$$0 < \Pr(D = 1 \mid Q) < 1.$$  \hspace{1cm} (Q-2)

The method does not explicitly model choices of treatment or the subjective evaluations of participants, nor is there any distinction between the variables in the outcome equations ($X$) and the variables in the choice equations ($Z$) that is central to the IV method and the method of control functions.

In principle, condition (Q-1) can be satisfied using a set of variables $Q$ distinct from all or some of the components of $X$ and $Z$.

The conditioning variables do not have to be exogenous.
From condition (Q-1) one recovers the distributions of $Y_0$ and $Y_1$ given $Q$.

$$\Pr (Y_0 \leq y_0 \mid Q = q) = F_0 (y_0 \mid Q = q).$$

And $$\Pr (Y_1 \leq y_1 \mid Q = q) = F_1 (y_1 \mid Q = q).$$

But not the joint distribution $F_{0,1} (y_0, y_1 \mid Q = q)$.

This is because the analyst does not observe the same persons in the treated and untreated states.

This is a standard evaluation problem common to all econometric estimators.

Methods for determining which variables belong in $Q$ rely on untested exogeneity assumptions which we discuss in this section.
• OLS is a special case of matching that focuses on the identification of conditional means.

• In OLS linear functional forms are maintained as exact representations or valid approximations.

• Considering a common coefficient model, OLS writes

\[ Y = \pi Q + D\alpha + U, \quad (Q-3) \]

where \( \alpha \) is the treatment effect and

\[ E (U \mid Q, D) = 0. \quad (Q-4) \]

• With a rank condition, this identifies \( \pi \) and \( \alpha \).
Matching just assumes

\[ E(U \mid Q, D) = E(U \mid Q) \]

(weaker than OLS).

Just identifies \( \alpha \).
The assumption is made that the variance-covariance matrix of \((Q, D)\) is of full rank:

\[
\text{Var} (Q, D) \text{ full rank.} \quad (Q-5)
\]
Under these conditions, one can identify $\alpha$ even though $D$ and $U$ are dependent: $D \not \perp \perp U$.

Controlling for the observable $Q$ eliminates any spurious mean dependence between $D$ and $U$: $E(U \mid D) \neq 0$ but $E(U \mid D, Q) = 0$.

Matching weakens it to $E(U \mid D, Q) = E(U \mid Q)$. 
(Q-3) is the linear regression counterpart to (Q-1).

(Q-5) is the linear regression counterpart to (Q-2).

Failure of (Q-5) would mean that using a nonparametric estimator one might perfectly predict $D$ given $Q$, and that $\Pr(D = 1 \mid Q = q) = 1$ or 0.

This condition might be met only at certain values of $Q = q$. 
Matching can be implemented as a nonparametric method.

When this is done, the procedure does not require specification of the functional form of the outcome equations.

It enforces the requirement that (Q-2) be satisfied by estimating functions pointwise in the support of $Q$. 
Assume that $Q = (X, Z)$ and that $X$ and $Z$ are the same except where otherwise noted.

Thus I invoke assumptions M-1 and M-2, even though in principle one can use a more general conditioning set.

Note that matching is a weakened version of OLS.
Assumptions M-1 and M-2 or (Q-1) and (Q-2) rule out the possibility that after conditioning on $X$ (or $Q$), agents possess more information about their choices than econometricians, and that the unobserved information helps to predict the potential outcomes.

Put another way, the method allows for potential outcomes to affect choices but only through the observed variables, $Q$, that predict outcomes.

This is the reason why Heckman and Robb (1985, 1986a) call the method selection on observables.
Matching: A Further Analysis

This section establishes the following points.

- Matching assumptions M-1 and M-2 generically imply a flat MTE in $u_D$,
  
  i.e. they assume that $E(Y_1 - Y_0 \mid X = x, U_D = u_D)$ does not depend on $u_D$.

- Thus the unobservables central to the Roy model and its extensions and the unobservables central to the modern IV literature are assumed to be absent once the analyst conditions on $X$.

- M-1 implies that all mean treatment parameters are the same.
Matching implies that conditional on $X$, the marginal return is assumed to be the same as the average return (marginal = average).

This is a strong behavioral assumption implicit in statistical conditional independence assumption M-1.

It says that the marginal participant has the same return as the average participant.
Matching Assumption M-1 Implies a Flat MTE

- An immediate consequence of M-1 is that the MTE does not depend on $U_D$.
- This is so because $(Y_0, Y_1) \perp \perp D \mid X$ implies that $(Y_0, Y_1) \perp \perp U_D \mid X$ and hence that

$$
\Delta^{MTE}(x, u_D) = E(Y_1 - Y_0 \mid X = x, U_D = u_D)
= E(Y_1 - Y_0 \mid X = x).
$$

- This, in turn, implies that $\Delta^{MTE}$ conditional on $X$ is flat in $u_D$. 
Under the stated assumptions for the generalized Roy model, it assumes that \( E(Y \mid P(Z) = p) \) is linear in \( p \).

Thus the method of matching assumes that mean marginal returns and average returns are the same and all mean treatment effects are the same given \( X \).

However, one can still distinguish marginal from average effects of the observables (\( X \)) using matching.

It is sometimes said that the matching assumptions are “for free” (See, e.g., Gill and Robins, 2001) because one can always replace unobserved \( F_1(Y_1 \mid X = x, D = 0) \) with observed \( F_1(Y_1 \mid X = x, D = 1) \) and unobserved \( F_0(Y_0 \mid X = x, D = 1) \) with observed \( F_0(Y_0 \mid X = x, D = 0) \).

Such replacements do not contradict any observed data.
While the claim is true, it ignores the counterfactual states generated under the matching assumptions.

The assumed absence of selection on unobservables is not a “for free” assumption, and produces fundamentally different counterfactual states for the same model under matching and selection assumptions.

To explore these issues in depth, consider a nonparametric regression model more general than the linear regression model (Q-3).
If there is no selection on unobserved variables conditional on covariates, \( U_D \independent (Y_0, Y_1) \mid X \), then
\[
E(U_1 \mid X, U_D) = E(U_1 \mid X) \quad \text{and} \quad E(U_0 \mid X, U_D) = E(U_0 \mid X)
\]
so that the OLS weights are unity and OLS identifies both ATE and the parameter treatment on the treated (TT), which are the same under this assumption.

This condition is an implication of matching condition M-1.

Given the assumed conditional independence in terms of \( X \), we can identify ATE and TT without use of any instrument \( Z \) satisfying assumptions (A-1) – (A-2).
If there is such a $Z$, the conditional independence condition implies under \((A-1) - (A-5)\) that $E(Y \mid X, P(Z) = p)$ is linear in $p$.

The conditional independence assumption invoked in the method of matching has come into widespread use for much the same reason that OLS has come into widespread use.

It is easy to implement with modern software and makes little demands of the data because it assumes the existence of $X$ variables that satisfy the conditional independence assumptions.
The crucial conditional independence assumption is not testable.

As I note below, additional assumptions on the $X$ are required to test the validity of the matching assumptions.
If the sole interest is to identify treatment on the treated, $\Delta^{TT}$, it is apparent from the representation

$$\Delta^{OLS}(X) = E(Y_1 | X, D = 1) - E(Y_0 | X, D = 0) = E(Y_1 - Y_0 | X, D = 1) + \{E(Y_0 | X, D = 1) - E(Y_0 | X, D = 0)\}.$$  

that one can weaken M-1 to

$$(M-1)' \quad Y_0 \perp \perp D | X.$$  

This is possible because we know $E(Y_1 | X, D = 1)$ from data on outcomes of the treated and only need to construct $E(Y_0 | X, D = 1)$. 

44 / 169
In this case MTE is not restricted to be flat in $u_D$ and all treatment parameters are not the same.

A straightforward implication of $(M-1)'$ in the Roy model, where selection is made solely on the gain, is that persons must sort into treatment status positively in terms of levels of $Y_1$.

I now consider more generally the implications of assuming mean independence of the errors rather than full independence.
Implementing the Method of Matching

- To operationalize the method of matching, I assume two samples: “t” for treatment and “c” for comparison group.

- Treatment group members have $D = 1$ and control group members have $D = 0$.

- Unless otherwise noted, I assume that observations are statistically independent within and across groups.

- Simple matching methods are based on the following idea.
For each person $i$ in the treatment group, I find some group of “comparable” persons.

The same individual may be in both treated and control groups if that person is treated at one time and untreated at another.

I denote outcomes for person $i$ in the treatment group by $Y_{i}^{t}$ and I match these outcomes to the outcomes of a subsample of persons in the comparison group to estimate a treatment effect.

In principle, I can use a different subsample as a comparison group for each person.
In practice, one can construct matches on the basis of a neighborhood $\xi(X_i)$, where $X_i$ is a vector of characteristics for person $i$.

Neighbors to treated person $i$ are persons in the companion sample whose characteristics are in neighborhood $\xi(X_i)$.

Suppose that there are $N_c$ persons in the comparison sample and $N_t$ in the treatment sample.
Thus the persons in the comparison sample who are neighbors to \( i \), are persons \( j \) for whom \( X_j \in \xi(X_i) \), i.e., the set of persons \( A_i = \{ j \mid X_j \in \xi(X_i) \} \).

Let \( W(i, j) \) be the weight placed on observation \( j \) in forming a comparison with observation \( i \) and further assume that the weights sum to one, \( \sum_{j=1}^{N_c} W(i, j) = 1 \), and that \( 0 \leq W(i, j) \leq 1 \).

Form a weighted comparison group mean for person \( i \), given by

\[
\bar{Y}_i^c = \sum_{j=1}^{N_c} W(i, j) Y_j^c.
\]
The estimated treatment effect for person $i$ is $Y_i - \bar{Y}_i^c$.

This selects a set of comparison group members associated with $i$ and the mean of their outcomes.

Unlike IV or the control function approach, the method of matching identifies counterfactuals for each treated member.
Heckman, Ichimura and Todd (1997) and Heckman, LaLonde and Smith (1999) survey a variety of alternative matching schemes proposed in the literature.

Todd (2006a,b) provides a comprehensive survey.

I briefly consider two widely-used methods.

The nearest-neighbor matching estimator defines $A_i$ such that only one $j$ is selected so that it is closest to $X_i$ in some metric:

$$A_i = \{ j \mid \min_{j\in\{1,...,N_c\}} \| X_i - X_j \| \},$$

where “$\| \|$” is a metric measuring distance in the $X$ characteristics space.
The Mahalanobis metric is one widely used metric for implementing the nearest neighbor matching estimator.

This metric defines neighborhoods for $i$ as

$$
\| \| = (X_i - X_j)' \sum_c^{-1} (X_i - X_j),
$$

where $\sum_c$ is the covariance matrix in the comparison sample.

The weighting scheme for the nearest neighbor matching estimator is

$$
W(i, j) = \begin{cases} 
1 & \text{if } j \in A_i, \\
0 & \text{otherwise.}
\end{cases}
$$
The nearest neighbor in the metric \( \| \cdot \| \) is used in the match.

A version of nearest-neighbor matching, called “caliper” matching (Cochran and Rubin, 1973), makes matches to person \( i \) only if

\[
\| X_i - X_j \| < \varepsilon,
\]

where \( \varepsilon \) is a pre-specified tolerance.

Otherwise person \( i \) is bypassed and no match is made to him or her.
• Kernel matching uses the entire comparison sample, so that 
\( A_i = \{1, \ldots, N_c\} \), and sets 

\[
W(i, j) = \frac{K(X_j - X_i)}{\sum_{j=1}^{N_c} K(X_j - X_i)},
\]

where \( K \) is a kernel.

• See, e.g., Härdle (1990) or Ichimura and Todd (2007) for a discussion of kernels and choices of bandwidths.
Kernels are typically a standard distribution function such as the normal cumulative distribution function.

Kernel matching is a smooth method that reuses and weights the comparison group sample observations differently for each person $i$ in the treatment group with a different $X_i$.

Kernel matching can be defined pointwise at each sample point $X_i$ or for broader intervals.
For example, the impact of treatment on the treated can be estimated by forming the mean difference across the $i$:

\[
\hat{\Delta}^{TT} = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i^t - \bar{Y}_i^c) = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i^t - \sum_{j=1}^{N_c} W(i,j)Y_j^c). \quad (4)
\]

One can define this mean for various subsets of the treatment sample defined in various ways.

Regression-adjusted matching, proposed by Rubin (1979) and clarified in Heckman, Ichimura and Todd (1997, 1998), uses regression-adjusted $Y_i$, denoted by $\tau(Y_i) = Y_i - X_i \beta$, in place of $Y_i$ in the preceding calculations.

See the cited papers for the econometric details of the procedure.
Matching assumes that conditioning on $X$ eliminates selection bias.

The method requires no functional form assumptions for outcome equations.

If, however, a functional form assumption is maintained, as in the econometric procedure proposed by Barnow, Cain and Goldberger (1980), it is possible to implement the matching assumption using standard regression analysis.
Suppose, for example, that \( Y_0 \) is linearly related to observables \( X \) and an unobservable \( U_0 \), so that

\[
E(Y_0 \mid X, D = 0) = X\beta + E(U_0 \mid X, D = 0),
\]

and

\[
E(U_0 \mid X, D = 0) = E(U_0 \mid X)
\]

is linear in \( X \) (\( E(U_0 \mid X) = \varphi X \)).

Under these assumptions, controlling for \( X \) via linear regression allows one to identify \( E(Y_0 \mid X, D = 1) \) from the data on nonparticipants.

Setting \( X = Q \), this approach justifies OLS equation (Q-3).

In equation (Q-3), this approach shows that \( \pi \) combines the estimate \( U_0 - Q \) effect with the causal effect of \( Q \) on \( Y \).
Such functional form assumptions are not strictly required to implement the method of matching.

Moreover, in practice, users of the method of Barnow, Cain and Goldberger (1980) do not impose the common support condition M-2 for the distribution of $X$ when generating estimates of the treatment effect.

The distribution of $X$ may be very different in the treatment group ($D = 1$) and comparison group ($D = 0$) samples, so that comparability is only achieved by imposing linearity in the parameters and extrapolating over different regions.
One advantage of the method of Barnow, Cain and Goldberger (1980) is that it uses data parsimoniously.

If the $X$ are high dimensional, the number of observations in each cell when matching can get very small.
Another solution to this problem that reduces the dimension of the matching problem without imposing arbitrary linearity assumptions is based on the probability of participation or the “propensity score,” \( P(X) = \Pr(D = 1 \mid X) \).

Rosenbaum and Rubin (1983) demonstrate that under assumptions M-1 and M-2,

\[
(Y_0, Y_1) \perp\!
\perp D \mid P(X) \text{ for } X \in \chi_c, \tag{5}
\]

for some set \( \chi_c \), where it is assumed that M-2 holds in the set.
Conditioning either on $P(X)$ or on $X$ produces conditional independence.

Their analysis is generalized to a multiple treatment setting in Lechner (2001) and Imbens (2003).
Conditioning on $P(X)$ reduces the dimension of the matching problem down to matching on the scalar $P(X)$.

The analysis of Rosenbaum and Rubin (1983) assumes that $P(X)$ is known rather than estimated.

Heckman, Ichimura and Todd (1998), Hahn (1998), and Hirano, Imbens and Ridder (2003) present the asymptotic distribution theory for the kernel matching estimator in the cases in which $P(X)$ is known and in which it is estimated both parametrically and nonparametrically.
Conditioning on $P$ identifies all treatment parameters but as has been shown, it imposes the assumption of a flat MTE.

Marginal returns and average returns are the same.

A consequence of (5) is that

$$E(Y_1|D = 0, P(X)) = E(Y_1|D = 1, P(X)) = E(Y_1|P(X)),$$

$$E(Y_0|D = 1, P(X)) = E(Y_0|D = 0, P(X)) = E(Y_0|P(X)).$$
Support condition M-2 has the unattractive feature that if the analyst has too much information about the decision of who takes treatment, so that \( P(X) = 1 \) or \( 0 \), the method breaks down at such values of \( X \) because people cannot be compared at a common \( X \).

The method of matching assumes that, given \( X \), some unspecified randomization in the economic environment allocates people to treatment.

This produces assumption (Q-5) in the OLS example.

The fact that the cases \( P(X) = 1 \) and \( P(X) = 0 \) must be eliminated suggests that methods for choosing \( X \) based on the fit of the model to data on \( D \) are potentially problematic, as I discuss below.
Offsetting these disadvantages, the method of matching with a known conditioning set that produces condition M-2 does not require separability of outcome or choice equations, exogeneity of conditioning variables, exclusion restrictions, or adoption of specific functional forms of outcome equations.

Such features are commonly used in conventional selection (control function) methods and conventional applications of IV although recent work in semiparametric estimation relaxes these assumptions.
As noted previously, the method of matching does not strictly require M-1.

One can get by with weaker mean independence assumptions (M-3) in the place of the stronger conditions M-1.

However, if (M-3) is invoked, the assumption that one can replace $X$ by $P(X)$ does not follow from the analysis of Rosenbaum and Rubin (1983), and is an additional new assumption.
Methods for implementing matching are provided in Heckman, Ichimura, Smith and Todd (1998) and are discussed extensively in Heckman, LaLonde and Smith (1999).

See Todd (1999, 2006a,b) for software and extensive discussion of the mechanics of matching.

I now contrast the identifying assumptions used in the method of control functions with those used in matching.
Comparing Matching and Control Functions Approaches

- The method of matching eliminates the dependence between $(Y_0, Y_1)$ and $D$, $(Y_0, Y_1) \not\perp\!\!\!\perp D$, by assuming access to conditioning variables $X$ such that M-1 is satisfied: $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$.

- By conditioning on observables, one can identify the distributions of $Y_0$ and $Y_1$ over the support of $X$ satisfying M-2.
Other methods model the dependence that gives rise to the spurious relationship and in this way attempt to eliminate it.

IV involves exclusion and a different type of conditional independence, \((Y_0, Y_1) \perp \perp Z \mid X\), as well as a rank condition \((\Pr (D = 1 \mid X, Z) \text{ depends on } Z)\).

The instrument \(Z\) plays the role of the implicit randomization used in matching by allocating people to treatment status in a way that does not depend on \((Y_0, Y_1)\).
I have already established that matching and IV make very different assumptions.

Thus, in general, a matching assumption that

\[(Y_0, Y_1) \perp \perp D \mid X, Z\]

neither implies nor is implied by

\[(Y_0, Y_1) \perp \perp Z \mid X.\]

One special case where they are equivalent is when treatment status is assigned by randomization with full compliance (letting \(\xi = 1\) denote assignment to treatment, \(\xi = 1 \Rightarrow A = 1\) and \(\xi = 0 \Rightarrow A = 0\) and \(Z = \xi\), so that the instrument is the assignment mechanism.

\(A = 1\) if the person actually receives treatment, and \(A = 0\) otherwise.
The method of control functions explicitly models the dependence between \((Y_0, Y_1)\) and \(D\) and attempts to eliminate it.
Equivalently, the method of control functions does not require

\[(U_0, U_1) \perp \perp V \mid (X, Z), \quad \text{or that} \quad (U_0, U_1) \perp \perp V \mid X\]

whereas matching does and typically equates \(X\) and \(Z\).

Thus matching assumes access to a richer set of conditioning variables than is assumed in the method of control functions.
The method of control functions allows for outcome unobservables to be dependent on $D$ even after conditioning on $(X, Z)$, and it models this dependence.

The method of matching assumes no such $D$ dependence.

Thus in this regard, and maintaining all of the assumptions invoked for control functions in this section, matching is a special case of the method of control functions in which under assumptions M-1 and M-2,

$$E (U_1|X, D = 1) = E (U_1|X)$$
$$E (U_0|X, D = 0) = E (U_0|X).$$
Comparing Matching and Classical Control Function Methods for a Generalized Roy Model

- Figure 1 shows the contrast between the shape of the MTE and the OLS matching estimand as a function of $p$ for the extended Roy model.

- The MTE($p$) shows its typical declining shape associated with diminishing returns, and the assumptions justifying matching are violated.

- Matching attempts to impose a flat MTE($p$) and therefore flattens the estimated MTE($p$) compared to its true value.

- It understates marginal returns at low levels of $p$ (associated with unobservables that make it likely to participate in treatment) and overstates marginal returns at high levels of $p$. 
Figure 1: Treatment Parameters and OLS/Matching as a function of $P(Z) = p$

Parameter Definition Under Assumptions (*):

- Marginal Treatment Effect
  \[ E[Y_1 - Y_0 | D^* = 0, P(Z) = p] = \bar{\beta} + \sigma U_1 - U_0 \Phi^{-1}(1 - p) \]

- Average Treatment Effect
  \[ E[Y_1 - Y_0 | P(Z) = p] = \bar{\beta} \]

- Treatment on the Treated
  \[ E[Y_1 - Y_0 | D^* > 0, P(Z) = p] = \bar{\beta} + \sigma U_1 - U_0 \varphi(\Phi^{-1}(1 - p)) \]

- Treatment on the Untreated
  \[ E[Y_1 - Y_0 | D^* \leq 0, P(Z) = p] = \bar{\beta} - \sigma U_1 - U_0 \varphi(\Phi^{-1}(1 - p)) \]

Note: $\Phi(\cdot)$ and $\varphi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively. $\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$.

(*): The model in this case is the same as the one presented below Figure 5.

Source: Heckman, Urzua and Vytlacil (2006)
### Parameter Definitions

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
<th>Under Assumptions (*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marginal Treatment Effect</td>
<td>$E [Y_1 - Y_0</td>
<td>D^* = 0, P(Z) = p]$</td>
</tr>
<tr>
<td>Average Treatment Effect</td>
<td>$E [Y_1 - Y_0</td>
<td>P(Z) = p]$</td>
</tr>
<tr>
<td>Treatment on the Treated</td>
<td>$E [Y_1 - Y_0</td>
<td>D^* &gt; 0, P(Z) = p]$</td>
</tr>
<tr>
<td>Treatment on the Untreated</td>
<td>$E [Y_1 - Y_0</td>
<td>D^* \leq 0, P(Z) = p]$</td>
</tr>
<tr>
<td>OLS/Matching on $P(Z)$</td>
<td>$E [Y_1</td>
<td>D^* &gt; 0, P(Z) = p] - E [Y_0</td>
</tr>
</tbody>
</table>

Note: $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively. $\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$. 

---

Figure 1: Treatment Parameters and OLS/Matching as a function of $P(Z) = p$

Source: Heckman, Urzua and Vytlacil (2006)
To further illustrate the bias in matching and how the control function eliminates it, I perform sensitivity analyses under different assumptions about the parameters of the underlying selection model.

In particular, I assume that the data are generated by the model

\[
Y_1 = \mu_1(X) + U_1 \\
Y_0 = \mu_0(X) + U_0 \\
D = 1(Z\gamma + V \geq 0),
\]

where \( \mu_D(Z) = Z\gamma, \mu_0(X) = \mu_0, \mu_1(X) = \mu_1, \) and

\[
(U_1, U_0, V) \sim N(0, \Sigma) \\
\text{corr}(U_j, V) = \rho_{jV} \\
\text{Var}(U_j) = \sigma_j^2; \quad j = \{0,1\}.
\]
I assume in this section that $D = 1 \left[ \mu_D(Z) + V > 0 \right]$, in conformity with the examples presented in Heckman and Navarro (2004), on which we build.

This reformulation of choice model

$$D^* = \mu_D(Z) - V; \quad D = 1 \quad \text{if} \quad D^* \geq 0; \quad D = 0 \quad \text{otherwise};$$

(6)

entails a simple change in the sign of $V$.

I assume that $Z \perp \perp (U_1, U_0, V)$. 
Using the selection formulae derived in the appendix, we can write the biases conditional on $P(Z) = p$ as

$$
\text{Bias}_{TT} (Z = z) = \text{Bias } TT (P(Z) = p) = \sigma_0 \rho_0 \nu M(p)
$$

$$
\text{Bias } ATE (Z = z) = \text{Bias } ATE (P(Z) = p) = M(p) [\sigma_1 \rho_1 \nu (1 - p) + \sigma_0 \rho_0 \nu p]
$$

where $M(p) = \frac{\phi(\Phi^{-1}(1-p))}{p(1-p)}$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of a standard normal random variable and the propensity score $P(z)$ is evaluated at $P(z) = p$.

I assume that $\mu_1 = \mu_0$ so that the true average treatment effect is zero.
I simulate the mean bias for TT (table 1) and ATE (table 2) for different values of the $\rho_{jV}$ and $\sigma_j$.

The results in the tables show that, as one lets the variances of the outcome equations grow, the value of the mean bias that we obtain can become substantial.

With larger correlations between the outcomes and the unobservables generating choices, come larger biases.

These tables demonstrate the greater generality of the control function approach, which models the bias rather than assuming it away by conditioning.
Even if the correlation between the observables and the unobservables ($\rho_{jV}$) is small, so that one might think that selection on unobservables is relatively unimportant, we still obtain substantial biases if we do not control for relevant omitted conditioning variables.

Only for special values of the parameters can one avoid bias by matching.

These examples also demonstrate that sensitivity analyses can be conducted for analysis based on control function methods even when they are not fully identified.

Vijverberg (1993) provides an example.
Table 1: Mean Bias for Treatment on the Treated

<table>
<thead>
<tr>
<th>$\rho_0V$</th>
<th>Average Bias ($\sigma_0 = 1$)</th>
<th>Average Bias ($\sigma_0 = 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.00</td>
<td>-1.7920</td>
<td>-3.5839</td>
</tr>
<tr>
<td>-0.75</td>
<td>-1.3440</td>
<td>-2.6879</td>
</tr>
<tr>
<td>-0.50</td>
<td>-0.8960</td>
<td>-1.7920</td>
</tr>
<tr>
<td>-0.25</td>
<td>-0.4480</td>
<td>-0.8960</td>
</tr>
<tr>
<td>0.00</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>0.25</td>
<td>0.4480</td>
<td>0.8960</td>
</tr>
<tr>
<td>0.50</td>
<td>0.8960</td>
<td>1.7920</td>
</tr>
<tr>
<td>0.75</td>
<td>1.3440</td>
<td>2.6879</td>
</tr>
<tr>
<td>1.00</td>
<td>1.7920</td>
<td>3.5839</td>
</tr>
</tbody>
</table>

$\text{BIAS TT} = \rho_0V \times \sigma_0 \times M(p)$

$M(p) = \frac{\phi(\Phi^{-1}(p))}{[p^*(1-p)\Phi^{-1}(p)]}$

## Table 2: Mean Bias for Average Treatment Effect

\[
\begin{array}{cccccccccc}
| \rho_0 \sigma_0 | & -1.00 & -0.75 & -0.50 & -0.25 & 0 & 0.25 & 0.50 & 0.75 & 1.00 \\
\hline
\rho_1 \sigma_1 = 1 & & & & & & & & & \\
-1.00 & -1.7920 & -1.5680 & -1.3440 & -1.1200 & -0.8960 & -0.6720 & -0.4480 & -0.2240 & 0 \\
-0.75 & -1.5680 & -1.3440 & -1.1200 & -0.8960 & -0.6720 & -0.4480 & -0.2240 & 0 & 0.2240 \\
-0.50 & -1.3440 & -1.1200 & -0.8960 & -0.6720 & -0.4480 & -0.2240 & 0 & 0.2240 & 0.4480 \\
-0.25 & -1.1200 & -0.8960 & -0.6720 & -0.4480 & -0.2240 & 0 & 0.2240 & 0.4480 & 0.6720 \\
0 & -0.8960 & -0.6720 & -0.4480 & -0.2240 & 0 & 0.2240 & 0.4480 & 0.6720 & 0.8960 \\
0.25 & -0.6720 & -0.4480 & -0.2240 & 0 & 0.2240 & 0.4480 & 0.6720 & 0.8960 & 1.1200 \\
0.50 & -0.4480 & -0.2240 & 0 & 0.2240 & 0.4480 & 0.6720 & 0.8960 & 1.1200 & 1.3440 \\
0.75 & -0.2240 & 0 & 0.2240 & 0.4480 & 0.6720 & 0.8960 & 1.1200 & 1.3440 & 1.5680 \\
1.00 & 0 & 0.2240 & 0.4480 & 0.6720 & 0.8960 & 1.1200 & 1.3440 & 1.5680 & 1.7920 \\
\hline
\rho_1 \sigma_1 = 2 & & & & & & & & & \\
-1.00 & -2.6879 & -2.2399 & -1.7920 & -1.3440 & -0.8960 & -0.4480 & 0 & 0.4480 & 0.8960 \\
-0.75 & -2.4639 & -2.0159 & -1.5680 & -1.1200 & -0.6720 & -0.2240 & 0.2240 & 0.6720 & 1.1200 \\
-0.50 & -2.2399 & -1.7920 & -1.3440 & -0.8960 & -0.4480 & 0 & 0.4480 & 0.8960 & 1.3440 \\
-0.25 & -2.0159 & -1.5680 & -1.1200 & -0.6720 & -0.2240 & 0.2240 & 0.6720 & 1.1200 & 1.5680 \\
0 & -1.7920 & -1.3440 & -0.8960 & -0.4480 & 0 & 0.4480 & 0.8960 & 1.3440 & 1.7920 \\
0.25 & -1.5680 & -1.1200 & -0.6720 & -0.2240 & 0.2240 & 0.6720 & 1.1200 & 1.5680 & 2.0159 \\
0.50 & -1.3440 & -0.8960 & -0.4480 & 0 & 0.4480 & 0.8960 & 1.3440 & 1.7920 & 2.2399 \\
0.75 & -1.1200 & -0.6720 & -0.2240 & 0.2240 & 0.6720 & 1.1200 & 1.5680 & 2.0159 & 2.4639 \\
1.00 & -0.8960 & -0.4480 & 0 & 0.4480 & 0.8960 & 1.3440 & 1.7920 & 2.2399 & 2.6879 \\
\end{array}
\]

\[
\text{BIAS ATE} = \rho_1 \sigma_1 M_1(p) - \rho_0 \sigma_0 M_0(p)
\]

\[
M_1(p) = \frac{\phi(\Phi^{-1}(p))}{p}
\]

\[
M_0(p) = -\frac{\phi(\Phi^{-1}(p))}{1-p}
\]

The Informational Requirements of Matching and the Bias When They Are Not Satisfied

- In this section I present some examples of when matching “works” and when it breaks down.

- This section extends the analysis of Heckman and Navarro (2004).

- In particular, I show how matching on some of the relevant information but not all can make the bias using matching worse for standard treatment parameters.
The method of matching assumes that the econometrician has access to and uses all of the relevant information in the precise sense defined there.

That means that the $X$ that guarantees conditional independence $M-1$ is available and is used.

The concept of relevant information is a delicate one and it is difficult to find the true conditioning set.
Assume that the economic model generating the data is a generalized Roy model of the form

\[ D^* = Z\gamma + V \]

where

\[ Z \perp \perp V \]

and

\[ V = \alpha V_1 f_1 + \alpha V_2 f_2 + \varepsilon V \]

\[ D = \begin{cases} 1 & \text{if } D^* \geq 0 \\ 0 & \text{otherwise} \end{cases} \]

and

\[ Y_1 = \mu_1 + U_1 \quad \text{where } U_1 = \alpha_{11} f_1 + \alpha_{12} f_2 + \varepsilon_1, \]

\[ Y_0 = \mu_0 + U_0 \quad \text{where } U_0 = \alpha_{01} f_1 + \alpha_{02} f_2 + \varepsilon_0. \]
Observe that contrary to the analysis throughout this paper we add $V$ and do not subtract it in the decision equation.

This is the familiar representation.

By a change in sign in $V$ one can go back and forth between the specification used in this section and the specification used in other sections of the paper.
In this specification, \((f_1, f_2, \varepsilon_V, \varepsilon_1, \varepsilon_0)\) are assumed to be mean zero random variables that are mutually independent of each other and \(Z\) so that all the correlation among the elements of \((U_0, U_1, V)\) is captured by \(f = (f_1, f_2)\).

Models that take this form are known as factor models and have been applied in the context of selection models by Aakvik, Heckman and Vytlacil (2005), Carneiro, Hansen and Heckman (2001, 2003), and Hansen, Heckman and Mullen (2004), among others.

I keep implicit any dependence on \(X\) which may be general.
Generically, the minimal relevant information for this model when the factor loadings are not zero \((\alpha_{ij} \neq 0)\) is, for general values of the factor loadings,

\[ I_R = \{ f_1, f_2 \}. \]

Notice that for a fixed set of \(\alpha_{ij}\), the minimal information set is

\[(\alpha_{11} - \alpha_{01}) f_1 + (\alpha_{12} - \alpha_{02}) f_2, \]

which captures the dependence between \(D\) and \((Y_1, Y_0)\).

Recall that I assume independence between \(Z\) and all error terms.
If the econometrician has access to $I_R$ and uses it, M-1 is satisfied conditional on $I_R$.

Note that $I_R$ plays the role of $\theta$ in

$$(Y_0, Y_1) \perp D \mid X, Z, \theta.$$ 

In the case where the economist knows $I_R$, the economist's information set $\sigma(I_E)$ contains the relevant information ($\sigma(I_E) \supseteq \sigma(I_R)$).
The agent’s information set may include different variables.

If one assumes that $\varepsilon_0, \varepsilon_1$ are shocks to outcomes not known to the agent at the time treatment decisions are made, but the agent knows all other aspects of the model, the agent’s information is

$$I_A = \{f_1, f_2, Z, \varepsilon_V\}.$$

Under perfect certainty, the agent’s information set includes $\varepsilon_1$ and $\varepsilon_0$:

$$I_A = \{f_1, f_2, Z, \varepsilon_V, \varepsilon_1, \varepsilon_0\}.$$

In either case, all of the information available to the agent is not required to satisfy conditional independence M-1.

All three information sets guarantee conditional independence, but only the first is minimal relevant.
In the previous notation, the observing economist may know some variables not in $I_A, I_{R^*}$ or $I_R$ but may not know all of the variables in $I_R$.

In the following sections, I study what happens when the matching assumption that $\sigma(I_E) \supseteq \sigma(I_R)$ does not hold.

That is, I analyze what happens to the bias from matching as the amount of information used by the econometrician is changed.
In order to get closed form expressions for the biases of the treatment parameters I make the additional assumption that

\[(f_1, f_2, \varepsilon_V, \varepsilon_1, \varepsilon_0) \sim N(0, \Sigma),\]

where \(\Sigma\) is a matrix with \((\sigma_{f_1}^2, \sigma_{f_2}^2, \sigma_{\varepsilon_V}^2, \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_0}^2)\) on the diagonal and zero in all the non-diagonal elements.

This assumption links matching models to conventional normal selection models.

However, the examples based on this specification illustrate more general principles.

I now analyze various commonly encountered cases.
The Economist Uses the Minimal Relevant Information:
\[ \sigma (I_R) \subseteq \sigma (I_E) \]

- I begin by analyzing the case in which the information used by the economist is \( I_E = \{ Z, f_1, f_2 \} \), so that the econometrician has access to a relevant information set and it is larger than the minimal relevant information set.

- In this case, it is straightforward to show that matching identifies all of the mean treatment parameters with no bias.
The matching estimator has population mean

\[ E(Y_1|D = 1, I_E) - E(Y_0|D = 0, I_E) = \mu_1 - \mu_0 + (\alpha_{11} - \alpha_{01}) f_1 + (\alpha_{12} - \alpha_{02}) f_2, \]

and all of the mean treatment parameters collapse to this same expression since, conditional on knowing \( f_1 \) and \( f_2 \), there is no selection because \((\varepsilon_1, \varepsilon_0) \perp \perp V\).

Recall that \( I_R = \{f_1, f_2\} \) and the economist needs less information to achieve M-1 than is contained in \( I_E \) for arbitrary choices of \( \alpha_{11}, \alpha_{01}, \alpha_{12}, \alpha_{02} \).
In this case, the analysis of Rosenbaum and Rubin (1983) tells us that knowledge of \((Z, f_1, f_2)\) and knowledge of \(P(Z, f_1, f_2)\) are equally useful in identifying all of the treatment parameters conditional on \(P\).

If we write the propensity score as

\[
P(I_E) = \Pr \left( \frac{\varepsilon_V}{\sigma_{\varepsilon_V}} > \frac{-Z\gamma - \alpha_{V1}f_1 - \alpha_{V2}f_2}{\sigma_{\varepsilon_V}} \right)
\]

\[
= 1 - \Phi \left( \frac{-Z\gamma - \alpha_{V1}f_1 - \alpha_{V2}f_2}{\sigma_{\varepsilon_V}} \right) = p,
\]

the event \(\left( D^* \leq 0, \text{ given } f = \tilde{f} \text{ and } Z = z \right)\) can be written as

\[
\frac{\varepsilon_V}{\sigma_{\varepsilon_V}} \leq \Phi^{-1} \left( 1 - P(z, \tilde{f}) \right),
\]

where \(\Phi\) is the cdf of a standard normal random variable and \(f = (f_1, f_2)\).
I abuse notation slightly by using $z$ as the realized fixed value of $Z$ and $\tilde{f}$ as the realized value of $f$.

The population matching condition M-1 implies that

$$E \left( Y_1 | D = 1, P(l_E) = P(z, \tilde{f}) \right) - E \left( Y_0 | D = 0, P(l_E) = P(z, \tilde{f}) \right)$$

$$= \mu_1 - \mu_0$$

$$+ E \left( U_1 | D = 1, P(l_E) = P(z, \tilde{f}) \right)$$

$$- E \left( U_0 | D = 0, P(l_E) = P(z, \tilde{f}) \right)$$

$$= \mu_1 - \mu_0$$

$$+ E \left( U_1 | \frac{\epsilon V}{\sigma_{\epsilon V}} > \Phi^{-1} \left( 1 - P(z, \tilde{f}) \right) \right) - E \left( U_0 | \frac{\epsilon V}{\sigma_{\epsilon V}} \leq \Phi^{-1} \left( 1 - P(z, \tilde{f}) \right) \right)$$

$$= \mu_1 - \mu_0.$$
Under these conditions, this expression is all treatment parameters discussed in this paper, since

$$E \left( U_1 \left| \frac{\varepsilon_V}{\sigma_{\varepsilon_V}} > \Phi^{-1} \left( 1 - P(z, \tilde{f}) \right) \right. \right) = \frac{\text{Cov} \left( U_1, \varepsilon_V \right)}{\sigma_{\varepsilon_V}} M_1 \left( P(z, \tilde{f}) \right)$$

and

$$E \left( U_0 \left| \frac{\varepsilon_V}{\sigma_{\varepsilon_V}} \leq \Phi^{-1} \left( 1 - P(z, \tilde{f}) \right) \right. \right) = \frac{\text{Cov} \left( U_0, \varepsilon_V \right)}{\sigma_{\varepsilon_V}} M_0 \left( P(z, \tilde{f}) \right),$$

where

$$M_1(P(z, \tilde{f})) = \frac{\phi(\Phi^{-1}(1 - P(z, \tilde{f})))}{P(z, \tilde{f})}$$

$$M_0(P(z, \tilde{f})) = -\frac{\phi(\Phi^{-1}(1 - P(z, \tilde{f})))}{1 - P(z, \tilde{f})},$$

where $\phi$ is the density of a standard normal random variable.
As a consequence of the assumptions about mutual independence of the errors

\[ \text{Cov} \left( U_i, \varepsilon_V \right) = \text{Cov} \left( \alpha_i f_1 + \alpha_i f_2 + \varepsilon_i, \varepsilon_V \right) = 0, \quad i = 0, 1. \]
In the context of the generalized Roy model, the case considered in this section is the one matching is designed to solve.

Even though a selection model generates the data, the fact that the information used by the econometrician includes the minimal relevant information makes matching a correct solution to the selection problem.

One can estimate the treatment parameters with no bias since, as a consequence of the assumptions, \((U_1, U_0) \perp \perp D \mid (f, Z)\), which is exactly what matching requires.

The minimal relevant information set is even smaller.
For arbitrary factor loadings, one only needs to know \((f_1, f_2)\) to secure conditional independence.

One can define the propensity score solely in terms of \(f_1\) and \(f_2\), and the Rosenbaum-Rubin result still goes through.

The analysis in this section focuses on treatment parameters conditional on particular values of \(P(Z, f) = P(z, \tilde{f})\), i.e., for fixed values of \(p\), but we could condition more finely.

Conditioning on \(P(z, \tilde{f})\) defines the treatment parameters more coarsely.

One can use either fine or coarse conditioning to construct the unconditional treatment effects.
In this example, using more information than what is in the relevant information set (i.e. using $Z$) is harmless.

But this is not generally true.

If $Z \not\perp (U_0, U_1, V)$, adding $Z$ to the conditioning set can violate conditional independence assumption M-1:

$$(Y_0, Y_1) \perp D \mid (f_1, f_2),$$

but

$$(Y_0, Y_1) \not\perp D \mid f_1, f_2, Z.$$

Adding extra variables can destroy the crucial conditional independence property of matching.
I present an example of this point below.

I first consider a case where $Z \perp (U_0, U_1, V)$ but the analyst conditions on $Z$ and not $(f_1, f_2)$.

In this case, there is selection on the unobservables that are not conditioned on.
The Economist Does Not Use All of the Minimal Relevant Information

Next, suppose that the information used by the econometrician is

\[ I_E = \{Z\}, \]

and there is selection on the unobservable (to the analyst) \( f_1 \) and \( f_2 \), i.e. the factor loadings \( \alpha_{ij} \) are all non zero.

Recall that I assume that \( Z \) and the \( f \) are independent.

In this case the event \( \left(D^* \leq 0, Z = z\right) \) is characterized by

\[
\frac{\alpha_V f_1 + \alpha V_2 f_2 + \varepsilon_V}{\sqrt{\alpha^2 V_1 \sigma^2 f_1 + \alpha^2 V_2 \sigma^2 f_2 + \sigma^2 \varepsilon}} \leq \Phi^{-1} \left(1 - P(z)\right).
\]
Using the analysis presented in the appendix, the bias for the different treatment parameters is given by

$$
\text{Bias } TT (Z = z) = \text{Bias } TT (P(Z) = P(z)) \\
= \eta_0 M(P(z)),
$$

(7)

where $M(p) = M_1(P(z)) - M_0(P(z))$.

$$
\text{Bias } ATE (Z = z) = \text{Bias } ATE (P(Z) = P(z)) \\
= M(p(z))\{\eta_1[1 - P(z)] + \eta_0 P(z)\},
$$

(8)

where

\[
\eta_1 = \frac{\alpha V_1 \alpha_{11} \sigma^2_{f_1} + \alpha V_2 \alpha_{12} \sigma^2_{f_2}}{\sqrt{\alpha^2 V_1 \sigma^2_{f_1} + \alpha^2 V_2 \sigma^2_{f_2} + \sigma^2_{\varepsilon V}}} \]

\[
\eta_0 = \frac{\alpha V_1 \alpha_{01} \sigma^2_{f_1} + \alpha V_2 \alpha_{02} \sigma^2_{f_2}}{\sqrt{\alpha^2 V_1 \sigma^2_{f_1} + \alpha^2 V_2 \sigma^2_{f_2} + \sigma^2_{\varepsilon V}}}. 
\]
It is not surprising that matching on sets of variables that exclude the relevant conditioning variables produces bias for the conditional (on $P(z)$) treatment parameters.

The advantage of working with a closed form expression for the bias is that it allows me to answer questions about the magnitude of this bias under different assumptions about the information available to the analyst, and to present some simple examples.

I next use expressions (7) and (8) as benchmarks against which to compare the relative size of the bias when I enlarge the econometrician’s information set beyond $Z$. 
Adding Information to the Econometrician’s Information Set $I_E$: Using Some but not All the Information from the Minimal Relevant Information Set $I_R$

- Suppose that the econometrician uses more information but not all of the information in the minimal relevant information set.

- He still reports values of the parameters conditional on specific $p$ values but now the model for $p$ has different conditioning variables.

- For example, the data set assumed in the preceding section might be augmented or else the econometrician decides to use information previously available.

- In particular, assume that the econometrician’s information set is

$$I'_E = \{Z, f_2\},$$

and that he uses this information set.
Under conditions 1 and 2 presented below, the biases for the treatment parameters conditional on values of $P = p$ are reduced in absolute value relative to their values in the previous section by changing the conditioning set in this way.

But these conditions are not generally satisfied, so that adding extra information does not necessarily reduce bias and may actually increase it.

To show how this happens in the model considered here, I define expressions comparable to $\eta_1$ and $\eta_0$ for this case:

$$
\eta_1' = \frac{\alpha V_1 \alpha_{11} \sigma^2_{f_1}}{\sqrt{\alpha^2 V_1 \sigma^2_{f_1} + \sigma^2_{\varepsilon V}}}
$$

$$
\eta_0' = \frac{\alpha V_1 \alpha_{01} \sigma^2_{f_1}}{\sqrt{\alpha^2 V_1 \sigma^2_{f_1} + \sigma^2_{\varepsilon V}}}
$$
I compare the biases under the two cases using formulas (7)–(8), suitably modified, but keeping $p$ fixed at a specific value even though this implies different conditioning sets in terms of $(z, \tilde{f})$. 
• **Condition 1** The bias produced by using matching to estimate TT is smaller in absolute value for any given p when the new information set $\sigma(I'_E)$ is used if

$$|\eta_0| > |\eta'_0|.$$ 

• There is a similar result for ATE:

**Condition 2** The bias produced by using matching to estimate ATE is smaller in absolute value for any given p when the new information set $\sigma(I'_E)$ is used if

$$|\eta_1 (1 - p) + \eta_0 p| > |\eta'_1 (1 - p) + \eta'_0 p|.$$
Proof.

These conditions are a direct consequence of formulas (7) and (8), modified to allow for the different covariance structure produced by the information structure assumed in this section (replacing $\eta_0$ with $\eta'_0$, $\eta_1$ with $\eta'_1$).

- It is important to notice that I condition on the same value of $p$ in deriving these expressions although the variables in $P$ are different across different specifications of the model.
- Recall that propensity-score matching defines them conditional on $P = p$. 
These conditions do not always hold.

In general, whether or not the bias will be reduced by adding additional conditioning variables depends on the relative importance of the additional information in both the outcome equations and on the signs of the terms inside the absolute value.
Consider whether Condition (1) is satisfied in general.

Assume $\eta_0 > 0$ for all $\alpha_{02}, \alpha_{V2}$.

Then $\eta_0 > \eta'_0$ if

$$
\eta_0 = \frac{\alpha_{V1}\alpha_{01}\sigma_{f_1}^2 + (\alpha_{V2}^2) \left( \frac{\alpha_{02}}{\alpha_{V2}} \right) \sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2 \sigma_{f_1}^2 + \alpha_{V2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon_{V}}^2}} > \frac{\alpha_{V1}\alpha_{11}\sigma_{f_1}^2}{\sqrt{\alpha_{V1}^2 \sigma_{f_1}^2 + \sigma_{\varepsilon_{V}}^2}} = \eta'_0.
$$

When $\frac{\alpha_{02}}{\alpha_{V2}} = 0$, clearly $\eta_0 < \eta'_0$.

Adding information to the conditioning set increases bias.

I can vary $\left( \frac{\alpha_{02}}{\alpha_{V2}} \right)$ holding all of the other parameters constant and hence can make the left hand side arbitrarily large.
A direct computation shows that

\[
\frac{\partial \eta_0}{\partial \left( \frac{\alpha_{02}}{\alpha V_2} \right)} = \frac{\alpha^2 V_2 \sigma_f^2}{\sqrt{\alpha^2 V_1 \sigma_f^2 + \alpha^2 V_2 \sigma_f^2 + \sigma_{\epsilon V}^2}} > 0.
\]

As \( \alpha_{02} \) increases, there is some critical value \( \alpha_{02}^* \) beyond which \( \eta_0 > \eta_0' \).

If I assumed that \( \eta_0 < 0 \), however, the opposite conclusion would hold, and the conditions for reduction in bias would be harder to meet, as the relative importance of the new information is increased.

Similar expressions can be derived for ATE and MTE, in which the direction of the effect depends on the signs of the terms in the absolute value.
Figures 2a and 2b illustrate the point that adding some but not all information from the minimal relevant set might increase the point-wise bias and the unconditional or average bias for ATE and TT, respectively.

Heckman and Navarro (2004) show comparable plots for MTE.

Values of the parameters of the model are presented at the base of the figures.
Note: Using proxy $\tilde{Z}$ for $f_2$ increases the bias. Correlation ($\tilde{Z}, f_2$) = 0.5.

Model:

$V = Z + f_1 + f_2 + \epsilon_\nu$;

$Y_1 = 2f_1 + 0.1f_2 + \epsilon_1$;

$Y_0 = f_1 + 0.1f_2 + \epsilon_0$;

$\epsilon_\nu \sim N(0, 1)$;

$\epsilon_1 \sim N(0, 1)$;

$\epsilon_0 \sim N(0, 1)$

$f_1 \sim N(0, 1)$;

$f_2 \sim N(0, 1)$

Note: Using proxy $\tilde{Z}$ for $f_2$ increases the bias. Correlation $(\tilde{Z}, f_2) = 0.5$.

Model:

$$V = Z + f_1 + f_2 + \varepsilon \nu; \quad Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1; \quad Y_0 = f_1 + 0.1f_2 + \varepsilon_0$$

$$\varepsilon \nu \sim N(0, 1); \quad \varepsilon_1 \sim N(0, 1); \quad \varepsilon_0 \sim N(0, 1)$$

$$f_1 \sim N(0, 1); \quad f_2 \sim N(0, 1)$$

The fact that the point-wise (and overall) bias might increase when adding some but not all information from $I_R$ is a feature that is not shared by the method of control functions.

Because the method of control functions models the stochastic dependence of the unobservables in the outcome equations on the observables, changing the variables observed by the econometrician to include $f_2$ does not generate bias.

It only changes the control function used.
That is, by adding $f_2$ one changes the control function from

$$K_1 (P(Z) = P(z)) = \eta_1 M_1(P(z))$$
$$K_0 (P(Z) = P(z)) = \eta_0 M_0(P(z))$$

to

$$K'_1 \left( P(Z, f_2) = P(z, \tilde{f}_2) \right) = \eta'_1 M_1(P(z, \tilde{f}_2))$$
$$K'_0 \left( P(Z, f_2) = P(z, \tilde{f}_2) \right) = \eta'_0 M_0(P(z, \tilde{f}_2))$$

but do not generate any bias by using the control function estimator.

This is a major advantage of this method.
It controls for the bias of the omitted conditioning variables by modeling it.

Of course, if the model for the bias term is not valid, neither is the correction for the bias.

Semiparametric selection estimators are designed to protect the analyst against model misspecification.
Matching evades this problem by assuming that the analyst always knows the correct conditioning variables and that they satisfy M-1.

In actual empirical settings, agents rarely know the relevant information set.

Instead they use proxies.
Adding Information to the Econometrician’s Information Set: Using Proxies for the Relevant Information

- Suppose that instead of knowing some part of the minimal relevant information set, such as $f_2$, the analyst has access to a proxy for it.

- For example, the returns-to-schooling literature often uses different test scores, like AFQT or IQ, to proxy for missing ability variables.

- These proxy, replacement function, methods are discussed in Heckman and Vytlacil (2007b, section 11) and in Abbring and Heckman (2007).

- In particular, assume that he has access to a variable $\tilde{Z}$ that is correlated with $f_2$ but that is not the full minimal relevant information set.
That is, define the econometrician’s information to be

\[ \widetilde{I}_{E^*} = \{ Z, \widetilde{Z} \} , \]

and suppose that he uses it so \( I_E = \widetilde{I}_{E^*} \).

In order to obtain closed-form expressions for the biases \( I \) assume that

\[ \widetilde{Z} \sim N \left( 0, \sigma^2_{\widetilde{Z}} \right) \]

\[ \text{corr} \left( \widetilde{Z}, f_2 \right) = \rho, \text{ and } \widetilde{Z} \perp \perp (\varepsilon_0, \varepsilon_1, \varepsilon_V, f_1) . \]
Define expressions comparable to $\eta$ and $\eta'$:

$$\tilde{\eta}_1 = \frac{\alpha_{11} \alpha V_1 \sigma_{f_1}^2 + \alpha_{12} \alpha V_2 (1 - \rho^2) \sigma_{f_2}^2}{\sqrt{\alpha^2 V_1 \sigma_{f_1}^2 + \alpha^2 V_2 \sigma_{f_2}^2 (1 - \rho^2) + \sigma_{\varepsilon V}^2}}$$

$$\tilde{\eta}_0 = \frac{\alpha_{01} \alpha V_1 \sigma_{f_1}^2 + \alpha_{02} \alpha V_2 (1 - \rho^2) \sigma_{f_2}^2}{\sqrt{\alpha^2 V_1 \sigma_{f_1}^2 + \alpha^2 V_2 \sigma_{f_2}^2 (1 - \rho^2) + \sigma_{\varepsilon V}^2}}.$$
By substituting for $I_E'$ by $\tilde{I}_E$ and $\eta_j'$ by $\tilde{\eta}_j$ ($j = 0, 1$) in Conditions (1) and (2) of the previous section, I can obtain results for the bias in this case.

Whether $\tilde{I}_E$ will be bias-reducing depends on how well it spans $I_R$ and on the signs of the terms in the absolute values in those conditions in.
In this case, however, there is another parameter to consider: the correlation $\rho$ between $\tilde{Z}$ and $f_2$, $\rho$.

If $|\rho| = 1$ we are back to the case of $\tilde{l}_E = l'_E$ because $\tilde{Z}$ is a perfect proxy for $f_2$.

If $\rho = 0$, we are essentially back to the case previously analyzed.
Because we know that the bias at a particular value of $p$ might either increase or decrease when $f_2$ is used as a conditioning variable but $f_1$ is not, we know that it is not possible to determine whether the bias increases or decreases as we change the correlation between $f_2$ and $\tilde{Z}$.

That is, we know that going from $\rho = 0$ to $|\rho| = 1$ might change the bias in any direction.

Use of a better proxy in this correlational sense may produce a more biased estimate.
From our previous analysis, it is straightforward to derive conditions under which the bias generated when the econometrician’s information is \( \tilde{I}_E \) is smaller than when it is \( I'_E \).

That is, it can be the case that knowing the proxy variable \( \tilde{Z} \) is better than knowing the actual variable \( f_2 \).

Take again the analysis of treatment on the treated as an example (i.e., Condition (1)).

The bias in absolute value (at a fixed value of \( p \)) is reduced when \( \tilde{Z} \) is used instead of \( f_2 \) if

\[
\left| \frac{\alpha_{01} \alpha_{V1} \sigma^2_{f1} + \alpha_{02} \alpha_{V2} (1 - \rho^2) \sigma^2_{f2}}{\sqrt{\alpha^2_{V1} \sigma^2_{f1} + \alpha^2_{V2} \sigma^2_{f2} (1 - \rho^2) + \sigma^2_{\varepsilon V}}} \right| < \frac{\alpha_{01} \alpha_{V1} \sigma^2_{f1}}{\sqrt{\alpha^2_{V1} \sigma^2_{f1} + \sigma^2_{\varepsilon V}}}. \]
Figures 3a and 3b, use the same true model as used in the previous section to illustrate the two points being made here.

- Namely, *using a proxy* for an unobserved relevant variable *might increase the bias*.

- On the other hand, it *might be better* in terms of bias to use a proxy than to use the actual variable, $f_2$.

- However, as figures 4a and 4b show, by changing $\alpha_{02}$ from 0.1 to 1, using a proxy might increase the bias versus using the actual variable $f_2$.

- Notice that the bias need not be universally negative or positive but depends on $p$. 
Note: Using proxy $\tilde{Z}$ for $f_2$ increases the bias. Correlation $(\tilde{Z}, f_2) = 0.5$.

Model:

\begin{align*}
V &= Z + f_1 + f_2 + \varepsilon_\nu; \quad Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1; \quad Y_0 = f_1 + 0.1f_2 + \varepsilon_0 \\
\varepsilon_\nu &\sim N(0, 1); \quad \varepsilon_1 \sim N(0, 1); \quad \varepsilon_0 \sim N(0, 1) \\
f_1 &\sim N(0, 1); \quad f_2 \sim N(0, 1)
\end{align*}

Note: Using proxy $\tilde{Z}$ for $f_2$ increases the bias. Correlation $(\tilde{Z}, f_2) = 0.5$.

Model:

\[
V = Z + f_1 + f_2 + \epsilon_\nu;
Y_1 = 2f_1 + 0.1f_2 + \epsilon_1;
Y_0 = f_1 + 0.1f_2 + \epsilon_0
\]

$\epsilon_\nu \sim N(0, 1)$; $\epsilon_1 \sim N(0, 1)$; $\epsilon_0 \sim N(0, 1)$

$f_1 \sim N(0, 1)$; $f_2 \sim N(0, 1)$

Note: Using proxy $\tilde{Z}$ for $f_2$ increases the bias. Correlation $(\tilde{Z}, f_2) = 0.5$.

Model:

\[
V = Z + f_1 + f_2 + \varepsilon_{\nu}; \quad Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1; \quad Y_0 = f_1 + 0.1f_2 + \varepsilon_0
\]

$\varepsilon_{\nu} \sim N(0, 1); \quad \varepsilon_1 \sim N(0, 1); \quad \varepsilon_0 \sim N(0, 1)$

$f_1 \sim N(0, 1); \quad f_2 \sim N(0, 1)$

The general point of these examples is that matching makes very knife-edge assumptions.

If the analyst gets the right conditioning set, M-1 is satisfied and there is no bias.

But determining the correct information set is not a trivial task, as I note below.

Having good proxies in the standard usage of that term can create substantial bias in estimating treatment effects.
The Case of a Discrete Outcome Variable

- The examples given in this section do not depend on all of the assumptions we have made to produce simple examples.

- In particular, we require neither normality nor additive separability of the outcome equations.

- The proposition that matching identifies the correct treatment if the econometrician’s information set includes all the minimal relevant information is true more generally, provided that any additional extraneous information used is exogenous in a sense to be defined precisely in the next section.

- Heckman and Navarro (2004) present parallel analyses of discrete treatment working with odds ratios and discrete outcomes that do not rely on either normality or separability of outcome equations.
On the Use of Model Selection Criteria to Choose Matching Variables

- I have previously shown by way of example that adding more variables from the minimal relevant information set, but not all variables in it, may increase bias.

- By a parallel argument, adding additional variables to the relevant conditioning set may make the bias worse.

- Although I have used a prototypical Roy model as the point of departure, clearly the point is more general.
There is no rigorous rule for choosing the conditioning variables that produce M-1.

Adding variables that are statistically significant in the treatment choice equation is not guaranteed to select a set of conditioning variables that satisfies condition M-1.

This is demonstrated by the analysis that shows that adding $f_2$ when it determines $D$ may increase bias at any selected value of $p$. 
The existing literature (e.g., Heckman, Ichimura, Smith and Todd, 1998) proposes criteria based on selecting a set of conditioning variables based on a goodness of fit criterion ($\lambda$), where a higher $\lambda$ means a better fit in the equation predicting $D$.

The intuition behind such criteria is that by using some measure of goodness of fit as a guiding principle one is using information relevant to the decision process.

Knowing $f_2$ improves goodness of fit, so that in general such a rule is deficient if $f_1$ is not known or is not used.
An implicit assumption underlying such procedures is that the added conditioning variables $\mathcal{X}$ are exogenous in the following sense:

$$ (Y_0, Y_1) \perp D | I_{\text{int}}, \mathcal{X} \quad (E-1) $$

where $I_{\text{int}}$ is interpreted as the variables initially used as conditioning variables before $\mathcal{X}$ is added.

Failure of exogeneity is a failure of M-1 for the augmented conditioning set, and matching estimators based on the augmented information set ($I_{\text{int}}, \mathcal{X}$) are biased when the condition is not satisfied.
Exogeneity assumption (E-1) is not usually invoked in the matching literature, which largely focuses on problem P-1, evaluating a program in place, rather than extrapolating to new environments (P-2).

Indeed, the robustness of matching to such exogeneity assumptions is often viewed as an advantage of the method.

In this section, I show some examples which illustrate the general point that standard model selection criteria fail to produce correctly specified conditioning sets unless some version of exogeneity condition (E-1) is satisfied.
In the literature, the use of model selection criteria is justified in two different ways.

Sometimes it is claimed that they provide a relative guide.

Sets of variables with higher $\lambda$ (better goodness of fit) are alleged to be better than sets of variables with lower $\lambda$ in the sense that they generate lower biases.

However, I have already shown that this is not true.

Enlarging the analyst’s information from $I_{int} = \{Z\}$ to $I'_{int} = \{Z, f_2\}$ will improve fit since $f_2$ is also in $I_{int}$ and $I_{int}$.

But, going from $I_{int}$ to $I'_{int}$ might increase the bias.
So it is not true that combinations of variables that increase some measure of fit $\lambda$ necessarily reduce the bias.

I construct a collection of conditioning variables $\tilde{I}_E$ with a better fit for $D$ and a larger conditional on $P = p$ bias than can be obtained from just conditioning on $\{f_1, f_2\}$ with no bias in the estimated treatment effects.

Let

$$\tilde{I}_E = \{Z, S\}$$

where

$$S = V - Z \gamma + \eta$$

$$\eta \sim N(0, \sigma^2_\eta)$$

$$\eta \perp (f_1, f_2, \varepsilon_0, \varepsilon_1, \varepsilon_V).$$
S might be a preference elicitation from a questionnaire that erroneously elicits the true valuation of the relative benefit of taking treatment for different \( Z \) values.

\( \eta \) is the measurement error.

The expressions for the biases are the same as above using \( \tilde{\beta}_j \) (\( j = 0, 1 \)) instead of \( \beta_j \) where:

\[
\begin{align*}
\tilde{\beta}_1 &= \frac{\pi \left( \alpha_{11}\alpha V_1 \sigma_{f_1}^2 + \alpha_{12}\alpha V_2 \sigma_{f_2}^2 \right)}{\sqrt{\alpha_{V_1}^2 \sigma_{f_1}^2 + \alpha_{V_2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon V}^2}} \\
\tilde{\beta}_0 &= \frac{\pi \left( \alpha_{01}\alpha V_1 \sigma_{f_1}^2 + \alpha_{02}\alpha V_2 \sigma_{f_2}^2 \right)}{\sqrt{\alpha_{V_1}^2 \sigma_{f_1}^2 + \alpha_{V_2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon V}^2}} \\
\pi &= \frac{\sigma_\eta}{\sqrt{\alpha_{V_1}^2 \sigma_{f_1}^2 + \alpha_{V_2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon V}^2 + \sigma_\eta^2}}.
\end{align*}
\]
In general, these expressions are not zero so that using propensity-score matching based on $Z$ and $S$ will generate a bias at most values of $p$, whereas conditioning on $f_1, f_2$ produces no bias.

The source of the bias is the measurement error in $S$ for $V$.

Now, to prove that this combination of variables has a better fit, all we need do is arbitrarily reduce $\sigma_\eta^2$.

In particular, when $\sigma_\eta^2 = 0$ we can perfectly predict $D$. 
That is, for

\[ 2\varepsilon > \sigma^2_\eta > \varepsilon > 0 \]

then

\[ \lim_{\varepsilon \to 0} \Pr(D = 1 | V - Z \gamma + \eta, Z) = \begin{cases} 
  1 & \text{for } V > 0, \\
  0 & \text{for } V < 0
\end{cases} \]

However, when the limit is attained, the assumption

\[ 0 < \Pr(D = 1 | W) = P(W) < 1, \]

is violated and matching breaks down.
Making $\sigma_\eta^2$ arbitrarily small, we can predict $D$ arbitrarily well so we can always increase $\lambda$ enough to find a combination of variables with better fit for predicted probabilities than the fit obtained using $f_1$ and $f_2$ and obtain a larger bias than a model that conditions only on the minimal relevant information $f_1$ and $f_2$ when the bias is zero.
Table 3 illustrates this point by generating two such variables \((S_1, S_2)\) and showing that, by reducing \(\sigma^2_\gamma\), we are able to increase either of two goodness-of-fit criteria (the percentage of correct in-sample predictions of \(D\) and the pseudo \(R^2\)) above those of the model with \(I_E = I_R\).

Adding a model based on \(S_2\) and \(Z\) (bottom row) increases the successful prediction rate over the case when the true model is used (the model based on \(\{Z, f_1, f_2\}\)) but it is biased for all parameters and substantially biased for overall (unconditional) \(ATE\) and \(MTE\).
A more general example that illustrates the key idea in the previous example and that does not entail a violation of assumption (M-1) in the limit considers use of a proxy regressor

\[ Q = \alpha_{QZ} Z + \alpha_{Q1} f_1 + \alpha_{Q2} f_2 + \tau + \eta \]

where \( Z \perp (f_1, f_2, \tau, \eta) ; (f_1, f_2, \tau, \eta) \) has mean zero; \( f_1 \perp f_2, \tau \perp \eta \) and \( (f_1, f_2) \perp (\tau, \eta) \).

\( \tau \) is possibly dependent on \( \varepsilon_V \) in the latent variable generating the treatment choice.

\( \eta \) is measurement error.
For different levels of dependence between \( \tau \) and \( \varepsilon_V \), and different weights on \( Z, f_1, f_2 \) and on the scale of measurement error.

- \( Q \) can be a better predictor of \( D \) than \( f_1, f_2 \) or even \( f_1, f_2, Z \).

- However, in general, \((Y_1, Y_0) \not\perp\!\!\!\perp D \mid Q\) because \( Q \) is an imperfect proxy for the combinations of \( f_1 \) and \( f_2 \) entering \( Y_1 \) and \( Y_0 \).

- Thus conditioning on \( Q \) can produce a better fit for \( D \) but greater bias for the treatment parameters.
Consider the following example where $Y$ is an outcome and $I$ is an index.

Let

$$
Y = \theta + \varepsilon_Y
$$

$$
I = \theta + \varepsilon_I
$$

$\theta \perp \perp (\varepsilon_Y, \varepsilon_I)$, $\varepsilon_Y \perp \perp \varepsilon_I$.

Obviously $Y \perp \perp I \mid \theta$. 
Suppose instead that we have a candidate conditioning variable \( Q = \alpha \theta + \eta + \tau \).

Suppose that all variables are normal with zero mean and are mutually independent.

Then we may write

\[
I = \pi_I Q + \varepsilon_Q
\]

where

\[
\pi_I = \frac{\alpha \theta \sigma^2_{\theta} + \sigma_{\tau,\varepsilon_I}}{\alpha^2 \sigma^2_{\theta} + \sigma^2_\eta + \sigma^2_\tau}.
\]
It is assumed that $\eta$ is independent of all other error components on the right-hand sides of the equations for $Q$, $I$ and $y$.

From normal regression theory we know that conditioning is equivalent to residualizing.

Constructing the residuals we obtain

$$I - \pi_I Q = \theta (1 - \alpha_\theta \pi_I) + \varepsilon_I - \pi_I (\eta + \tau).$$
By a parallel argument

\[ Y - \pi_Y Q = \theta (1 - \alpha \theta \pi_Y) + \varepsilon_Y - \pi_Y (\eta + \tau) \]

\( Y \perp I \mid Q \) requires that \( I - \pi_I Q \) and \( Y - \pi_Y Q \) be uncorrelated, which in general does not happen.

Letting the dependence between \( \tau \) and \( \varepsilon_I \) get large, and setting \( \alpha \theta \) to suitable values, we can predict \( I \) better (in the sense of \( R^2 \)) with \( Q \) than with \( \theta \).

Letting \( D = 1(I > 0) \) produces a simple version of the example in the text because better prediction of \( I \) produces better prediction of \( D \).
The essential feature of these examples is that the selected conditioning variables are endogenous with respect to the outcome equation.

If all candidate conditioning variables were restricted to be exogenous in this sense, our example could not be constructed.

This underscores the importance of the econometric concept of endogeneity which is sometimes viewed as an inessential distinction in selecting the conditioning variables in matching.

Although it is irrelevant for defining parameters, it is essential when using goodness of fit measures for selecting conditioning variables.
Table 3 illustrates this point using the normal example.

Going from row 1 to row 2 (adding $f_2$) improves goodness of fit and increases the unconditional or overall bias for all three treatment parameters, because (E-1) is violated.
Table 3

<table>
<thead>
<tr>
<th>Variables in Probit</th>
<th>Goodness of fit statistics $\lambda$</th>
<th>Average Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct in-sample prediction rate</td>
<td>Pseudo $R^2$</td>
</tr>
<tr>
<td>$Z$</td>
<td>66.88%</td>
<td>0.1284</td>
</tr>
<tr>
<td>$Z, f_2$</td>
<td>75.02%</td>
<td>0.2791</td>
</tr>
<tr>
<td>$Z, f_1, f_2$</td>
<td>83.45%</td>
<td>0.4844</td>
</tr>
<tr>
<td>$Z, S_1$</td>
<td>77.38%</td>
<td>0.3282</td>
</tr>
<tr>
<td>$Z, S_2$</td>
<td>92.25%</td>
<td>0.7498</td>
</tr>
</tbody>
</table>

Model

\[
V = Z + f_1 + f_2 + \varepsilon_V; \quad \varepsilon_V \sim N(0, 1);
\]

\[
Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1; \quad \varepsilon_1 \sim N(0, 1);
\]

\[
Y_0 = f_1 + 0.1f_2 + \varepsilon_0; \quad \varepsilon_0 \sim N(0, 1);
\]

\[
S_1 = V + u_1; \quad u_1 \sim N(0, 4);
\]

\[
S_2 = V + u_2; \quad u_2 \sim N(0, 0.25);
\]

\[
f_1 \sim N(0, 1); \quad f_2 \sim N(0, 1).
\]

The following rule of thumb argument is sometimes invoked as an absolute standard against which to compare alternative models.

In versions of the argument, the analyst asserts that there is a combination of variables $I''$ that satisfy $M-1$ and hence produces zero bias and a value of $\lambda = \lambda''$ larger than that of any other $I$.

In the examples, conditioning on $\{Z, f_1, f_2\}$ generates zero bias.

One can exclude $Z$ and still obtain zero bias.

Because $Z$ is a determinant of $D$, this shows immediately that the best fitting model does not necessarily identify the minimal relevant information set.
In this example including $Z$ is innocuous because there is still zero bias and the added conditioning variables satisfies (E-1) where $I_A = (f_1, f_2)$.

In general, such a rule is not innocuous if $Z$ is not exogenous.

If goodness of fit is used as a rule to choose variables on which to match, there is no guarantee it produces a desirable conditioning set.

If one includes in the conditioning set variables $X$ that violate (E-1), they may improve the fit of predicted probabilities but worsen the bias.
Heckman and Navarro (2004) produce a series of examples that have the following feature.

Variables $S$ (shown at the base of table 3) are added to the information set that improve the prediction of $D$ but are correlated with $(U_0, U_1)$.

Their particular examples use imperfect proxies $(S_1, S_2)$ for $(f_1, f_2)$.

The point is more general.

The $S$ variables fail exogeneity and produce greater bias for TT and ATE but they improve the prediction of $D$ as measured by the correct in-sample prediction rate and the pseudo-$R^2$.

See the bottom two rows of table 3.
### Table 3

<table>
<thead>
<tr>
<th>Variables in Probit</th>
<th>Goodness of fit statistics λ</th>
<th>Average Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct in-sample prediction rate</td>
<td>Pseudo $R^2$</td>
</tr>
<tr>
<td>$Z$</td>
<td>66.88%</td>
<td>0.1284</td>
</tr>
<tr>
<td>$Z, f_2$</td>
<td>75.02%</td>
<td>0.2791</td>
</tr>
<tr>
<td>$Z, f_1, f_2$</td>
<td>83.45%</td>
<td>0.4844</td>
</tr>
<tr>
<td>$Z, S_1$</td>
<td>77.38%</td>
<td>0.3282</td>
</tr>
<tr>
<td>$Z, S_2$</td>
<td>92.25%</td>
<td>0.7498</td>
</tr>
</tbody>
</table>

**Model**

\[
V = Z + f_1 + f_2 + \varepsilon_V; \quad \varepsilon_V \sim N(0, 1);
\]

\[
Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1; \quad \varepsilon_1 \sim N(0, 1);
\]

\[
Y_0 = f_1 + 0.1f_2 + \varepsilon_0; \quad \varepsilon_0 \sim N(0, 1);
\]

\[
S_1 = V + u_1; \quad u_1 \sim N(0, 4);
\]

\[
S_2 = V + u_2; \quad u_2 \sim N(0, 0.25);
\]

\[
f_1 \sim N(0, 1); \quad f_2 \sim N(0, 1).
\]

Summary and Conclusions

This paper exposits the key identifying assumptions of commonly used econometric evaluation estimators.

The emphasis is on the economic content of these assumptions.

Using the prototypical generalized Roy model as the point of departure, I examine the assumptions underlying the method of matching.

In brief, they are:

(a) the analyst knows the relevant information set,

(b) a randomization can be produced by conditioning on regressors, and

(c) conditioning on observables, marginal returns are the same as average returns.
Other econometric estimators do not impose (c) and account for failures of (a) and (b), and in this sense are more general.
Appendix
Selection Formula for the Matching Examples

Consider a generalized Roy model of the form $Y_1 = \mu_1 + U_1$; $Y_0 = \mu_0 + U_0$; $D^* = \mu_D (Z) + V$; $D = 1$ if $D^* \geq 0$, $= 0$ otherwise; and $Y = DY_1 + (1 - D) Y_0$, where

$$(U_1, U_0, V)' \sim N(0, \Sigma); \text{Var}(U_i) = \sigma_i^2 \quad i = 0, 1$$
$$\text{Var}(V) = \sigma_V^2; \text{Cov}(U_1, U_0) = \sigma_{10}$$
$$\text{Cov}(U_1, V) = \sigma_{1V}; \text{Cov}(U_0, V) = \sigma_{0V}.$$ 

Assume $Z \perp \perp (U_0, U_1, V)$.

Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the pdf and the cdf of a standard normal random variable.
Then, the propensity score for this model for $Z = z$ is given by:

$$
\Pr (D^* > 0 | Z = z) = \Pr (V > -\mu_D (z)) = P(z) = \Phi \left( \frac{\mu_D (z)}{\sigma_V} \right).
$$

Thus $\frac{\mu_D (z)}{\sigma_V} = \Phi^{-1} (P(z))$, and

$$
\frac{-\mu_D (z)}{\sigma_V} = \Phi^{-1} (1 - P(z)).
$$

The event $\left( V \gtrless 0, Z = z \right)$ can be written as

$$
\frac{V}{\sigma_V} \gtrless -\frac{\mu_D (z)}{\sigma_V} \iff \frac{V}{\sigma_V} \gtrless \Phi^{-1} (1 - P(z)).
$$

We can write the conditional expectations required to get the biases for the treatment parameters as a function of $P(z) = p$. 
For $U_1$:

$$
E \left( U_1 \mid D^* > 0, Z = z \right) = \frac{\sigma_{1V}}{\sigma_V} E \left( \frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} > \frac{-\mu_D(z)}{\sigma_V} \right) \\
= \frac{\sigma_{1V}}{\sigma_V} E \left( \frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} > \Phi^{-1}(1 - P(z)) \right) \\
= \eta_1 M_1(P(z))
$$

where

$$
\eta_1 = \frac{\sigma_{1V}}{\sigma_V}.
$$
Similarly for $U_0$:

$$E(U_0|D^* > 0, Z = z) = \eta_0 M_1(P(z))$$
$$E(U_0|D^* < 0, Z = z) = \eta_0 M_0(P(z)),$$

where $\eta_0 = \frac{\sigma_{0V}}{\sigma_V}$ and $M_1(P(z)) = \frac{\phi(\Phi^{-1}(1-P(z)))}{P(z)}$ and $M_0(P(z)) = -\frac{\phi(\Phi^{-1}(1-P(z)))}{(1-P(z))}$ are inverse Mills ratio terms.
Substituting these into the expressions for the biases for the treatment parameters conditional on $z$ we obtain

$$
\text{Bias } TT (P(z)) = \eta_0 M_1(P(z)) - \eta_0 M_0(P(z))
= \eta_0 M(P(z)),
$$

$$
\text{Bias } ATE (P(z)) = \eta_1 M_1(P(z)) - \eta_0 M_0(P(z))
= M(P(z)) (\eta_1 (1 - P(z)) + \eta_0 P(z)) .
$$