RIETI Discussion Paper Series 14-E-061

Decomposition of Gender or Racial Inequality with Endogenous Intervening Covariates: An extension of the DiNardo-Fortin-Lemieux method

YAMAGUCHI Kazuo
RIETI

RIETI
Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
http://www.rieti.go.jp/en/

RIETI Discussion Paper Series 14-E-061
October 2014

Decomposition of Gender or Racial Inequality with Endogenous Intervening Covariates: An extension of the DiNardo-Fortin-Lemieux method[1]

YAMAGUCHI Kazuo
University of Chicago / RIETI

Abstract

This paper first clarifies that, unlike propensity-score weighting in Rubin's causal model where confounding covariates can be endogenous, propensity-score weighting in the DiNardo-Fortin-Lemieux (DFL) decomposition analysis may generate biased estimates for the decomposition of inequality into "direct" and "indirect" components when intervening variables are endogenous. The paper also clarifies that the Blinder-Oaxaca method confounds the modeling of two distinct counterfactual situations: one where the *covariate effects* of the first group become equal to those of the second group, and the other where the *covariate distribution* of the second group becomes equal to that of the first group. The paper shows that the DFL method requires a distinct condition to provide an unbiased decomposition of inequality that remains under each counterfactual situation. The paper then introduces a combination of the DFL method with Heckman's two-step method as a way of testing and eliminating bias in the DFL estimate when some intervening covariates are endogenous. The paper also intends to bring gender and race back into the center of statistical causal analysis. An application focuses on the decomposition of gender inequality in earned income among white-collar regular employees in Japan.

Keywords: DiNardo-Fortin-Lemieux method, Decomposition of equality, Propensity-score weighting, Causal analysis, Endogeneity, Gender inequality in income
JEL classification: C14, J71

[1] This paper is a part of research result of "Women's economic empowerment, low fertility rate, and work-life balance" study undertaken at Research Institute of Economy, Trade and Industry (RIETI).

DECOMPOSITION OF GENDER OR RACIAL INEQUALITY WITH ENDOGENOUS INTERVENING COVARIATES: AN EXTENSION OF THE DINARDO-FORTIN-LEMIEUX METHOD

1. INTRODUCTION

The aim of this paper is to reconceptualize, and reformulate, from the point-of-view of Rubin's causal model (RCM) (Rosenbaum and Rubin 1983 1984, Rubin 1985), the DiNardo-Fortin-Lemieux (DFL) method (DiNardo et al. 1996) for the decomposition of inequality based on propensity-score weighting and to extend the method for handling cases when covariates include intervening variables that are endogenous. Decomposition analysis is typically concerned with dividing the effects of a dichotomous group variable *X*, such as the distinction between men and women, on the outcome into two components, a component explained by the difference in covariates **V** between the groups, and a component not explained (See Fortin et al. (2011) for a comprehensive review of decomposition analysis).

The decomposition method of analyzing inequality that has been most frequently used is the Blinder-Oaxaca method (Blinder 1973; Oaxaca 1973). Further extensions or modifications of it to solve its "identification issue" have also been discussed (e.g., Jones and Kelley 1984; Oaxaca and Ransom 1999; Yun 2008; Kim 2013). Extending it to decomposition for hazard rate was also proposed (Powers and Yun 2009). For decomposition analysis based on the Blinder-Oaxaca (BO) method, the two counterfactual situations described below are always treated as identical. The method assumes a pair of regression equations, such as $y^M = \boldsymbol{\beta}'^M \mathbf{V}^M + \varepsilon$ for men and $y^W = \boldsymbol{\beta}'^W \mathbf{V}^W + \varepsilon$ for women. Then, under the assumed independence of the error terms from covariates, we obtain,

$$\overline{y}^W - \overline{y}^M = \left[\boldsymbol{\beta}^{W}{}'(\overline{\mathbf{V}}^W - \overline{\mathbf{V}}^M)\right] + \left[(\boldsymbol{\beta}^{W}{}' - \boldsymbol{\beta}^{M}{}')\overline{\mathbf{V}}^M\right]. \quad (1)$$

Since $\boldsymbol{\beta}^{W}{}'\overline{\mathbf{V}}^{M}$ represents the mean of *Y* for *a counterfactual situation where the covariate distribution is that of men while the covariate effects are those of women*, the first component of the equation can be interpreted as the sex difference in the mean of *Y* which would be eliminated *if women had men's covariate distribution*, and the second component can be interpreted as the sex difference in the mean of *Y* that would be eliminated *if men had women's covariate effects*. Considering either of those two counterfactual situations leads to the same decomposition result in the BO method. The first portion is also called the "explained component" of the inequality because it is attributable to sex differences in covariate values. The second component, called the "unexplained component," reflects inequality in society that results from different "treatments" of men and women with the same covariates. In their review paper on decomposition analysis, Fortin et al. (2011) discussed that the latter component of the BO method can be interpreted as the *average treatment effect for the treated* (Morgan and Winship 2007) when the group variable can be regarded as a treatment variable. As explained later in this section, this component becomes equivalent with *the average treatment effect for the untreated* for the case where the group variable, such as the sex dummy variable, is time-constant. Generally, the two counterfactual situations described above differ when the error term is endogenous, as explained below. This point has not been mentioned in the previous literature.

The DFL method makes a weaker assumption for the outcome than the BO method. In the original formulation, the pair of equations assumed for men and women are expressed as $y^{M} = \phi(\mathbf{V}^{M}, \boldsymbol{\theta}^{M}) + \varepsilon$ for men and $y^{W} = \phi(\mathbf{V}^{W}, \boldsymbol{\theta}^{W}) + \varepsilon$ for women, where ϕ is an unspecified function, and $\boldsymbol{\theta}^{M}$ and $\boldsymbol{\theta}^{W}$ are parameters that indicate the covariate effects on the outcome for men and women respectively. Under the assumed independence of the error term from the covariates – that is, if the covariates are exogenous – we obtain the following decomposition,

$$\overline{y}^{W} - \overline{y}^{M} = \left[\overline{\phi}(\mathbf{V}^{W}, \boldsymbol{\theta}^{W}) - \overline{\phi}(\mathbf{V}^{M}, \boldsymbol{\theta}^{W}))\right] + \left[\overline{\phi}(\mathbf{V}^{M}, \boldsymbol{\theta}^{W}) - \overline{\phi}(\mathbf{V}^{M}, \boldsymbol{\theta}^{M}))\right]. \quad (2)$$

Similar to equation (1) of the BO method, the first component of equation (2) reflects inequality that would be eliminated if women had men’s covariate distribution, and the second component reflects inequality that would be eliminated if men were treated like women. Since the estimates of $\bar{\phi}(\mathbf{V}^W, \boldsymbol{\theta}^W)$ and $\bar{\phi}(\mathbf{V}^W, \boldsymbol{\theta}^W)$ are simply sample means, we only need an estimate of $\bar{\phi}(\mathbf{V}^M, \boldsymbol{\theta}^W)$ obtainable from the following equation for a dichotomous variable *X* that takes a value of 0 for men and 1 for women:

$$\begin{aligned}\bar{\phi}(\mathbf{V}^M, \boldsymbol{\theta}^W) &\equiv \int_{\mathbf{v}} \phi(\mathbf{v}, \boldsymbol{\theta}^W) f(\mathbf{v} \mid X = 0) d\mathbf{v} = \int_{\mathbf{v}} E(Y^W \mid \mathbf{v}) f(\mathbf{v} \mid X = 0) d\mathbf{v} \\ &= \int_{\mathbf{v}} \omega(\mathbf{v}) E(Y^W \mid \mathbf{v}) f(\mathbf{v} \mid X = 1) d\mathbf{v} = E_{\omega}(Y^W) \qquad (3)\end{aligned}$$

where $f(\mathbf{v} \mid X)$ indicates the conditional probability density of **V**, and E_{ω} indicates the weighted mean with weights:

$$\omega(\mathbf{v}) \equiv \frac{f(\mathbf{v} \mid X = 0)}{f(\mathbf{v} \mid X = 1)} = \frac{p(X = 0 \mid \mathbf{v}) f(\mathbf{v}) / p(X = 0)}{p(X = 1 \mid \mathbf{v}) f(\mathbf{v}) / p(X = 1)} = \frac{p(X = 1) p(X = 0 \mid \mathbf{v})}{p(X = 0) p(X = 1 \mid \mathbf{v})}. \quad (4)$$

A consistent estimate of the weight can be obtained by a consistent estimate of $p(X \mid \mathbf{v})$ via logit or probit regression.

Hence, the DFL method generalizes the BO method by weakening the assumption for the outcome equation by semiparametric modeling. As Barskey et al. (2002) point out, a major limitation of the BO method is that it requires a linear relationship between the dependent variable and its covariates. In contrast, the DFL method does not assume such a relationship. In particular, the DFL method can be applied to the decomposition of difference in proportion, as in the decomposition analysis of gender difference in the proportion of managers, while the BO method may not be applied to such an analysis. This is because linear probability models for regression analysis usually fail to yield consistent parameter estimates, with the exception of saturated models with categorical covariates where estimates of outcome probabilities always lie between 0 and 1.

The BO and DFL methods also differ where some covariates are endogenous. For the BO method, a typical method of handling an endogenous covariate is the instrumental variable (IV) method for linear regression. The application of the BO method using least-squares or maximum likelihood estimation is invalid when covariates are endogenous because the regression coefficient estimates become inconsistent.

The situation differs for the DFL method as described in this paper. Note that when the covariates are endogenous, $E(\varepsilon \mid \mathbf{v}) \neq 0$, and therefore a change in the the error term distribution may affect the mean outcome. In particular, two counterfactual situations become different. For the set of parameters, covariates, and the error term $\{\boldsymbol{\theta}, \mathbf{V}, \boldsymbol{\varepsilon}\}$ that affect the outcome, the counterfactual situation where women have men's covariate distribution implies a combination of women's $\{\boldsymbol{\theta}, \boldsymbol{\varepsilon}\}$ and men's **V**. On the other hand, the counterfactual situation where men have women's covariate effects implies a combination of men's $\{\mathbf{V}, \boldsymbol{\varepsilon}\}$ and women's $\boldsymbol{\theta}$. Hence, the error distribution differs for these two counterfactual situations. This paper shows that the DFL method can reflect each of these counterfactual situations under different conditions. In particular, as explained later, the direct effect of *X* on *Y* not through **V** will be equated with the counterfactual effect that would be eliminated when people in one group with covariates **V** were treated the same in the society as members of the other group with the same **V**. That is, the direct effect will be equated with the counterfactual effect that would be eliminated if men had women's *covariate effects*, and not equated with the other counterfactual effect that would be eliminated if women had men's *covariate* distributions.

I also need to discuss the misconception of some researchers that time-constant variables such as gender or race cannot be the "treatment variables" in causal analysis. The argument is based on the fact that if we employ the person-specific fixed-effect model, or the difference-in-difference (DID) estimation, with panel survey data, the effect of race or gender is wiped out because it is completely collinear with, or is explained by, unobserved population heterogeneity.

In fact, the argument has two conceptual confusions. One of them is a confusion for whether gender or race can be conceived as a variable for treatment or can be conceived as a variable for treatment assignment. Causal analysis based on the use of propensity scores (Rosenbaum and Rubin 1983, 1984, Rubin 1985, Robins 1998), called Rubin's causal models (RCM), distinguishes between treatment and treatment assignment. In the standard conception, the treatment assignment distinguishes between the treatment group and the control group. Subjects are assigned to one of those two groups, and only subjects in the treatment group get treated. The causal analysis, however, conceives for each subject both the outcome under treatment and the outcome under no treatment, one of which is unobserved and counterfactual, and therefore the difference between the two is always unobservable. The observed variable that distinguishes between two sexual or racial groups is not a variable that indicates a treatment but can be considered a variable that indicates a treatment assignment, but unlike a distinction in the standard conception, it is not a distinction between the treatment group and the control group, but a distinction between two different control groups. The treatment itself is a hypothetical change in sex or race that nobody in the sample experienced. Hence, no sample subjects are in the treatment group, and as a result, *the average treatment effect for the treated* (Morgan and Winship2007) cannot be estimated. Although there can be such cases as transgender persons, we are concerned here with the effect of a hypothetical change in the value of a time-constant variable. We can conceive, however, *the average treatment effect for the untreated*, the effect of treatment in the outcome obtained by comparing the observed outcome under each control state of the variable with the counterfactual outcome under treatment among those whose observation is made for just one of the control states, and can estimate the treatment effect under certain assumptions described later.

The other confusion is between the question of whether the treatment effect of change in sex or race can be conceived and that of whether the treatment effect of such a change is estimable. The argument about the fixed-effect model or the DID indicates only that if we assume unconstrained time-constant unobserved population heterogeneity in determining the outcome, we cannot estimate the treatment effect of a time-constant variable. On the other hand, if we make some different assumptions, the effect of a hypothetical change in sex or race becomes estimable. A major problem, however, is that the standard ignorability assumption that the RCM makes is not likely to hold when intervening variables that the decomposition analysis employs as covariates are endogenous, as explained below.

Another argument against the conception of causal analysis about the effect of sex or race is substantive. The argument goes that if women are discriminated against, for example, in employment, the cause of discrimination is employer's attitude, and the effect of gender would disappear if such employers' discrimination against women became absent, and therefore, gender is not a true cause. It is true that gender is not a cause of discrimination. However, gender discrimination has been concerned with whether men and women with the same set of attributes other than sex are treated unequally in the society. The analysis of racial or gender discrimination based on an experimental audit study (Pager 2003; Correll et al. 2007) is concerned with an estimation of the effect of race or gender on the outcome through a differential treatment of people based on gender or race in the society.
We are interested in such a "treatment effect" of gender, and although it is methodologically different from the experimental audit study, what we refer to as causal analysis in this paper is a method of estimating the average treatment effect as defined initially by Rosenbaum and Rubin (1983) based on the counterfactual conception of treatment.

An important issue for a causal modeling of the decomposition analysis is the endogeneity of covariates. As shown below, the ignorability assumption that the method makes may not hold when some intervening covariates are endogenous or if the population is defined by a state of an endogenous intervening variable. Such a situation will occur when an intervening covariate includes, for example, educational attainment, and uncontrolled selectivity into higher education affects the outcome. It may also occur when the analysis is applied only to the population of employees, and uncontrolled selectivity into and exit from employment affects the outcome. In other words, both selection bias in the state of an intervening variable and sample selection that results from using a population that correspond to a particular state of an intervening variable may generate bias in the DFL decomposition as a causal analysis.

Since the DFL method relies on propensity-score weighting (Rubin 1985, Robins 1998) and makes an ignorability assumption that appears to be similar to that of the RCM for the analysis of cross-sectional survey data, one may consider that the method can handle endogenous covariates because in the standard use of propensity-score weighting, the endogeneity of confounding variables is allowed, and that is one of the major merits of the RCM.   The situation, however, differs between the use of propensity-score weighting for the RCM, and the use of it in the DFL method for decomposition analysis. Let me clarify this by using simplified causal diagrams.

Figure 1: Two causal diagrams with endogenous covariates

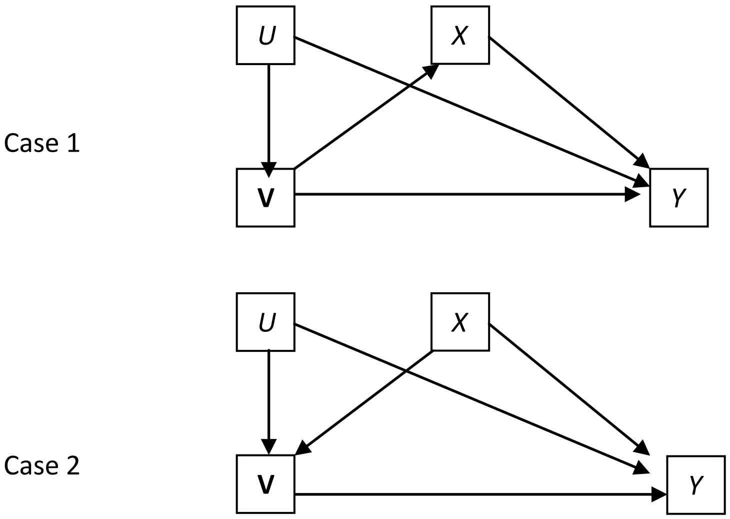


Figure 1 consists of two causal diagrams with observed endogenous covariates **V.** They are endogenous because an unobserved confounder *U* that affects both **V** and *Y* exists. In case 1, covariates **V** are confounders of treatment variable *X* and outcome variable *Y*, and conditional independence of *X* and *U*, $X \perp U \mid \mathbf{V}$, holds. In case 2, however, covariates **V** are intervening variables, and even though *X* is assumed to be independent of *U*, the control for covariates **V**, which are common causal descendants of *X* and *U*, induces non-independence between *U* and *X* (Morgan and Winship 2007, Pearl 2009). As an illustration, assume that latent ability *U* is independent of gender *X*, and both ability and gender affect education *V* such that people with higher ability are more likely to attend college, and due to gender inequality in educational opportunity, men are more likely than women to attend college. Assume further for simplicity of discussion that the top 50% of men in ability and the top 25% of women in ability go to college. Then, among college attendants, average ability is higher for women (because it is the average of the top 25% for women while it is the average of the top 50% for men). Similarly, among people who do not go to college, average ability is also higher for women than for men (because it is the average of the bottom 75% for women while it is the average of the bottom 50% for men). Hence, correlation between latent ability and gender emerges when education is held constant.

Generally, $X \perp U \mid \mathbf{V}$ does not hold in case 2. Case 1 is the situation that the RCM assumes for causal analysis with cross-sectional data, and the condition $X \perp U \mid \mathbf{V}$ is equivalent to the ignorability assumption described later. On the other hand, case 2 is the typical situation we have for decomposition analysis with endogenous covariates. Even though the independence of *X* and *U* may exist without controlling for **V**, by which the *total treatment effect* of *X* on *Y* becomes estimable, we are interested in the *direct treatment effect* of *X* on *Y* not through **V** in decomposition analysis. With the control for **V**, however, we cannot make the ignorability assumption $X \perp U \mid \mathbf{V}$ unless all intervening covariates **V** are exogenous and not subject to selection bias by *U*.

Although the DFL method may thus yield a biased estimate for the direct treatment effect of *X* not through **V** when covariates include endogenous intervening variables, the method also enables an extension that can eliminate this bias, and this paper introduces such an extension for the DFL method. Note that the assumption of exogenous intervening variables is a very strong one, and the development of a method that can handle the issue of endogenous covariates for decomposition analysis will be a very useful advancement.

This paper also performs a decomposition analysis of gender inequality in earned income among white-collar regular employees in Japan to demonstrate the usefulness of the extension of the DFL method.

2. METHOD

2.1 **The Causal Conception of Sex or Race, and Conditions and Results for the Case Where the Ignorability Assumption Holds**

Below, I (1) describe the standard set of assumptions for the RCM model, (2) modify the assumptions for analysis of the decomposition of inequality between sexual or racial groups, and (3) clarify the condition where the ignorability assumption holds and the DFL estimate for the decomposition is unbiased as a causal analysis. A formal expression of bias in the DFL method that exists when the ignorability assumption does not hold and a method that eliminates this bias will be introduced in the following sections.

The RCM for cross-sectional data analysis makes the following set of assumptions (Rosenbaum and Rubin 1983 1984; Morgan and Winship 2007).

(A1) The Stable Unit Treatment Value Assignment (SUTVA) is assumed. This assumption justifies conceiving the treatment at the individual level to be independent of the results of treatment assignment. The SUTVA assumes that the counterfactual treatment effect, defined as $Y_{1i} - Y_{0i}$, -- where Y_{1i} indicates either the observed outcome or the potential outcome under the treatment and Y_{0i} indicates either the observed outcome or the potential outcome under no treatment for subject *i* --, does not depend on who in the population is assigned to the treatment group.

(A2) The treatment effects are heterogeneous and vary with persons. This leads to the dependence of the average treatment effect in the RCM on the specification of the population, such as the total population, the population of the treated, or the population of the untreated.

(A3) The ignorability assumption, $(Y_{1,}Y_0) \perp X \mid \mathbf{V}$, applies, such that if observed *confounding variables* **V** that affect both the outcome and the treatment assignment *X* are controlled, the potential outcomes and the treatment assignment are conditionally independent. This assumption implies that no *unobserved* confounding variables exist.

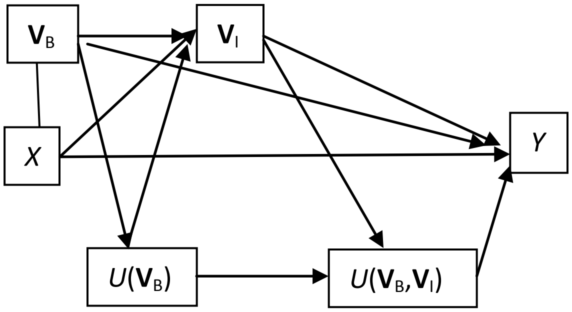
(A4) The confounding covariates can be endogenous; that is, they may not be independent of the unobserved determinant of the outcome.

Note that only A1 and A3 are constraining assumptions. Assumptions A2 and A4 imply the absence of the constraints imposed by many other methods, such as regression-based methods.

In the case of the treatment effect of sex or race, we retain assumption A2 on the heterogeneity of treatment effects. Assumption A1, the SUTVA assumption, becomes irrelevant, because nobody in the sample has been treated and the "treatment," a change in gender or race, is only imaginary. We need a notational change for Y_0 and Y_1 , however, because the distinction of two sexual groups or two racial groups is not a distinction between the treatment group and the control group, but a distinction between two control groups. Unlike the standard notation, where Y_1 and Y_0 refer to the outcome under treatment and under no treatment, respectively, we use notations Y_1 and Y_0, in conceiving the treatment effect of gender or race, to indicate the observed or potential outcome that is generated when the subject is treated as a member of group 1 and group 0, respectively. Therefore, for example, for men for whom $X = 0$, Y_0 is the observed outcome, and Y_1 is the counterfactual potential outcome with a change of sex, and for women for whom $X = 1$, Y_0 is the counterfactual potential outcome with a change of sex, and Y_1 is the observed outcome. Y_1-Y_0 indicates the effect of being treated as a woman versus being treated as a man, and is the treatment effect for the untreated for both men and women. The observed outcome can be expressed as $Y_{obs} = XY_1 + (1-X)Y_0$ as in the case of the standard RCM.

We need to discuss the modification of assumptions A3 and A4 together. I will illustrate, though it is simplified because of the omission of various interaction effects and heterogeneity in the effects of variables on their causal descendants, by the causal diagram of Figure 2.

Figure 2: A general causal structure



Variable X in Figure 2 is a time-constant treatment variable, such as gender or race. Note first that since variable X is by assumption not affected by any other variables, no confounders that affect both X and Y exist for the total population. However, an analysis based on a sample of a particular population, such as an analysis of gender inequality in wages among employees, may make unobserved as well as observed covariates of Y correlated with X because men and women may become employed and remain employed according to gender-specific selection processes. In Figure 2, such a sample selection bias is equated with a control for an intervening variable V_I that indicates the presence versus absence of employment. Even if there are no confounders that affect both X and Y, there can be exogenous causes of Y that are neither causal antecedents nor causal descendants of X but are correlated with X. We call such variables *exogenous correlated causes.* Various family background characteristics as the correlates of race, such as parents' education, occupation, and income, and family intactness at subject's birth, are examples. In Figure 2, variables $\mathbf{V}_B$ indicate exogenous correlated causes, and the line without arrowheads that connects X and $\mathbf{V}_B$ indicates a correlation without causation. Variable $U(\mathbf{V}_B)$ is the unobserved determinant of Y that remains when we control for $\mathbf{V}_B$. We assume that although correlated causes $\mathbf{V}_B$ affect $U(\mathbf{V}_B)$, as family background affects subject's latent trait $U(\mathbf{V}_B)$, and the affected $U(\mathbf{V}_B)$ in turn affects the outcome, $U(\mathbf{V}_B)$ is conditionally independent of X, controlling for $\mathbf{V}_B$, that is, $X \perp U(\mathbf{V}_B) \mid \mathbf{V}_B$. This is the ignorability assumption we will make in this section. However, we will also introduce later a method for the case where this ignorability assumption does not hold.

The variables $\mathbf{V}_\text{I}$ indicate the intervening variables that can be subject to selection bias by an unobserved variable and can affect the unobserved variable, as in the case of selection bias in higher education by latent ability and the effect of educational attainment on latent ability. In Figure 2, the path from $U(\mathbf{V}_\text{B})$ to $\mathbf{V}_\text{I}$, and the path from $\mathbf{V}_\text{I}$ to $U(\mathbf{V}_\text{B},\mathbf{V}_\text{I})$ indicate those effects, where variable $U(\mathbf{V}_\text{B},\mathbf{V}_\text{I})$ indicates the unobserved determinant of *Y* that remains when we control for both $\mathbf{V}_\text{B}$ and $\mathbf{V}_\text{I}$. In accordance with the discussion of case 2 of Figure 1 in the previous section, even when $X \perp U(\mathbf{V}_B) \mid \mathbf{V}_B$ holds, $X \perp U(\mathbf{V}_B, \mathbf{V}_I) \mid \mathbf{V}_B, \mathbf{V}_I$ may not hold unless all intervening variables are exogenous and are not subject to selection bias by $U(\mathbf{V}_\text{B})$. In summary, in replacing assumption A3, we assume in this section,

Modified A3: $X \perp U(\mathbf{V}_B) \mid \mathbf{V}_B$ for exogenous correlated causes $\mathbf{V}_B$ of *Y*.

It follows that:

Corollary from the modified A3: $X \perp U(\mathbf{V}_B, \mathbf{V}_I) \mid \mathbf{V}_B, \mathbf{V}_I$ does not hold generally unless intervening variables $\mathbf{V}_I$ are all exogenous, and, therefore, independent of $U(\mathbf{V}_B)$.

We do not make assumption A4. Instead, we distinguish below between the case where all intervening variables are exogenous, and, therefore, $X \perp U(\mathbf{V}_B, \mathbf{V}_I) \mid \mathbf{V}_B, \mathbf{V}_I$ holds, and the case where some intervening variables are endogenous, and, therefore, $X \perp U(\mathbf{V}_B, \mathbf{V}_I) \mid \mathbf{V}_B, \mathbf{V}_I$ may not hold.

For the rest of this section, I will show that when $X \perp U(\mathbf{V}) \mid \mathbf{V}$ holds, where **V** denotes covariates taken into account in the decomposition analysis, the DFL estimate of the decomposition provides an unbiased estimate of the direct treatment effect of *X* on *Y* not through **V**. I derive this result by using both the standard RCM expression and an alternative expression that is more useful in identifying bias in the DFL estimate. In the next section, I will formally express the bias in the DFL estimate when $X \perp U(\mathbf{V}) \mid \mathbf{V}$ does not hold, and in the section after that, I introduce a method that tests and eliminates this bias.

Generally, the average direct treatment effect of *X* on *Y* not through **V** is concerned with the estimation of $E(Y_1 - Y_0 \mid \mathbf{V})$ and its average in the total population $E(Y_1 - Y_0) \equiv \int_{\mathbf{v}} E(Y_1 - Y_0 \mid \mathbf{v}) f(\mathbf{v}) d\mathbf{v}$, where $f(\mathbf{v})$ is the probability density function of **V** in the total population, its average among the population of group 0, $E(Y_1 - Y_0 \mid X = 0) \equiv \int_{\mathbf{v}} E(Y_1 - Y_0 \mid \mathbf{v}) f(\mathbf{v} \mid X = 0) d\mathbf{v}$, or its average among the population of group 1, $E(Y_1 - Y_0 \mid X = 1) \equiv \int_{\mathbf{v}} E(Y_1 - Y_0 \mid \mathbf{v}) f(\mathbf{v} \mid X = 1) d\mathbf{v}$.

Suppose that we are interested in estimating $E(Y_1 - Y_0 \mid X = 0)$. Since the estimate of $E(Y_0 \mid X = 0)$ is simply the sample mean of *Y* for subjects in group *X* = 0, we only need an estimate of the counterfactual mean $E(Y_1 \mid X = 0)$. Suppose that an ignorability assumption, $(Y_1, Y_0) \perp X \mid \mathbf{V}$, holds. This assumption is equivalent to assuming $X \perp U(\mathbf{V}) \mid \mathbf{V}$, as explained later. Then $E(Y_1 \mid X = 0)$ can be expressed as follows.

$$\begin{aligned}
E(Y_1 \mid X = 0) &= \int_{\mathbf{v}} E(Y_1 \mid X = 0, \mathbf{v}) f(\mathbf{v} \mid X = 0) d\mathbf{v} \\
&= \int_{\mathbf{v}} E(Y_1 \mid X = 1, \mathbf{v}) f(\mathbf{v} \mid X = 0) d\mathbf{v} \quad \text{(by the ignorability assumption)} \\
&= \int_{\mathbf{v}} E(Y_{obs} \mid X = 1, \mathbf{v}) f(\mathbf{v} \mid X = 0) d\mathbf{v} \\
&= \int_{\mathbf{v}} \omega(\mathbf{v}) E(Y_{obs} \mid X = 1, \mathbf{v}) f(\mathbf{v} \mid X = 1) d\mathbf{v} \\
&= E_{\omega}(Y_{obs} \mid X = 1), \qquad (5)
\end{aligned}$$

where $\omega(\mathbf{v})$ is a weight that is the same as that defined in equation (4), that is, $\omega(\mathbf{v}) = \left[p(X = 1) p(X = 0 \mid \mathbf{v}) \right] / \left[p(X = 0) p(X = 1 \mid \mathbf{v}) \right]$.

and E_ω denotes the weighted mean with the weight $\omega(\mathbf{v})$. Note, however, that unlike the derivation of equations (3) and (4) for the original DFL method described in the introduction, this derivation relies on the ignorability assumption. Hence, a consistent estimator of the counterfactual mean $E(Y_1 | X = 0)$ is given as the weighted mean of observed *Y* among subjects in group *X* = 1 with weights that can be calculated using the consistent estimate of $p(X = 1 | \mathbf{v})$ by logit or probit regression. However, weight estimates $\hat{\omega}_{1i}$ may not have the average of 1 for sample subjects in group *X* = 1, and therefore, it is better to employ the following adjusted weight for the ratio estimator because it is known to be more efficient than the unadjusted estimate:

$$\hat{\omega}_i^* \equiv N_1 \hat{\omega}_i \Big/ \left(\sum_{i|x_i=1} \hat{\omega}_i \right) \quad , \qquad (6)$$

where N_1 is the number of sample subjects in group 1. Then we obtain the following decomposition of difference in the mean of *Y* between the two groups, and this is the same as the DFL decomposition:

$$\begin{aligned} E(Y | X = 1) - E(Y | X = 0) &= E(Y_1 | X = 1) - E(Y_0 | X = 0) \\ &= E(Y_1 - Y_0 | X = 0) + \{E(Y_1 | X = 1) - E(Y_1 | X = 0)\}. \end{aligned} \qquad (7)$$

Suppose that *X* indicates the distinction between women (*X* = 1) and men (*X* = 0). Then $E(Y_1 - Y_0 | X = 0)$ indicates the average direct treatment effect of gender on the outcome among men, that is, the average effect on men of being treated as a woman. The second component, $E(Y_1 | X = 1) - E(Y_1 | X = 0)$, represents the difference in the mean of *Y* that would be eliminated if women had men's covariate distribution, because

$$\begin{aligned} E(Y_1 | X = 1) - E(Y_1 | X = 0) &= \int_v \left\{ E(Y_1 | X = 1, \mathbf{v}) p(\mathbf{v} | X = 1) - E(Y_1 | X = 0, \mathbf{v}) p(\mathbf{v} | X = 0) \right\} d\mathbf{v} \\ &= \int_v \left\{ E(Y_1 | X = 1, \mathbf{v}) (p(\mathbf{v} | X = 1) - p(\mathbf{v} | X = 0)) \right\} d\mathbf{v} \qquad (8) \end{aligned}$$

holds. Note that $E(Y_1 \mid X = 1) - E(Y_1 \mid X = 0)$ becomes equal to the amount explained by the difference in the covariate distribution, because $E(Y_1 \mid X = 0, \mathbf{v}) = E(Y_1 \mid X = 1, \mathbf{v})$ holds from the ignorability assumption, and without this assumption such an interpretation cannot be made for $E(Y_1 \mid X = 1) - E(Y_1 \mid X = 0)$.

Now we consider a slightly different derivation of the average direct treatment effect. This alternative derivation becomes the basis of the extension of the DFL method for causal analysis described in the following sections. We keep the assumptions made above for the modification of the RCM model and, in addition, we make the following additional assumption A5, which may appear to make the model slightly more specific than the typical RCM model – though generality is actually not lost. The assumption is the following:

A5: The unobserved determinant of outcome *Y*, including its possible interaction effects with **V**, is functionally linear in affecting the outcomes.

With this additional assumption, we can express the observed and potential outcomes by excluding the random error term whose mean for each individual is zero without loss of generality as:

$$Y_{0i} = \phi(\boldsymbol{\theta}_0, \mathbf{v}_i) + \theta_{0u}(\mathbf{v}_i)u_i \text{ , and}$$

$$Y_{1i} = \phi(\boldsymbol{\theta}_1, \mathbf{v}_i) + \theta_{1u}(\mathbf{v}_i)u_i \qquad (9)$$

Here, ϕ is an unknown function, $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are parameters that characterize the effects of covariates **V** on Y_0 and Y_1, respectively, *U* is the unobserved determinant of the outcome, and $\theta_{0u}(\mathbf{v})$ and $\theta_{1u}(\mathbf{v})$, respectively, characterize the effect of *U* on Y_0 and Y_1 and its possible dependence on **V** due to possible interaction effects of **V** and *U* on Y_0 and Y_1. The RCM model, without assumption 5, corresponds to a more general expression such that $Y_{0i} = \phi(\boldsymbol{\theta}_0, \mathbf{v}_i, u_i)$ and $Y_{1i} = \phi(\boldsymbol{\theta}_1, \mathbf{v}_i, u_i)$. However, since *U* is an unobserved variable that is not constrained in form, and since equation (9) can reflect the unconstrained interaction effects of *U* and **V** on the outcome, the expression of equation (9) does not lose the generality of the RCM regarding the unconstrained heterogeneity of the treatment effects.

The ignorability assumption can now be expressed as $X \perp U \mid \mathbf{V}$, because as assumed in equation (9), Y_0 and Y_1 are functions of variables **V** and *U*. Then, $E(Y_1 \mid X = 0)$ is given as follows:

$$\begin{aligned}
E(Y_1 \mid X = 0) &= \int_{\mathbf{v}}\int_{u} (\phi(\boldsymbol{\theta}_1, \mathbf{v}) + \theta_{1u}(\mathbf{v})u) f(\mathbf{v}, u \mid X = 0) du d\mathbf{v} \\
&= \int_{\mathbf{v}} f(\mathbf{v} \mid X = 0) \left[\int_{u} (\phi(\boldsymbol{\theta}_1, \mathbf{v}) + \theta_{1u}(\mathbf{v})u) f(u \mid \mathbf{v}, X = 0) du \right] d\mathbf{v} \\
, \qquad &= \int_{\mathbf{v}} f(\mathbf{v} \mid X = 0) \left[\int_{u} (\phi(\boldsymbol{\theta}_1, \mathbf{v}) + \theta_{1u}(\mathbf{v})u) f(u \mid \mathbf{v}, X = 1) du \right] d\mathbf{v} \;\; (\text{because } U \perp X \mid \mathbf{V}) \\
&= \int_{\mathbf{v}} E(Y_{obs} \mid \mathbf{v}, X = 1) f(\mathbf{v} \mid X = 0) d\mathbf{v} \\
&= \int_{\mathbf{v}} \omega(\mathbf{v}) E(Y_{obs} \mid \mathbf{v}, X = 1) f(\mathbf{v} \mid X = 1) d\mathbf{v} \\
&= E_{\omega}(Y_{obs} \mid X = 1), \qquad\qquad (10)
\end{aligned}$$

where $\omega(\mathbf{v})$ is the same as that defined by equation (2). Naturally, we obtain the same results for $E(Y_1 \mid X = 0)$ as equation (1).

2.2 **The Case Where Endogenous Intervening Variables Exist**

In this section, I assume that the ignorability assumption $X \perp U \mid \mathbf{V}$ may not hold for the pair of equations (9) given in the previous section, either (a) because the covariates of the decomposition analysis include some endogenous intervening variables or (b) because the ignorability assumption $X \perp U(\mathbf{V}_B) \mid \mathbf{V}_B$ does not hold even before the control for the intervening variables.

When *X* indicates a gender distinction as before, the mean of the counterfactual outcome where men had women's covariate effects is now given as:

$$\begin{aligned}
E(Y_1 \mid X = 0) &= \int_{\mathbf{v}}\int_{u} (\phi(\boldsymbol{\theta}_1, \mathbf{v}) + \theta_{1u}(\mathbf{v})u) f(\mathbf{v}, u \mid X = 0) du d\mathbf{v} \\
&= \int_{\mathbf{v}} \phi(\boldsymbol{\theta}_1, \mathbf{v}) f(\mathbf{v} \mid X = 0) d\mathbf{v} + E(\theta_{1u}(\mathbf{v})u \mid X = 0). \qquad (11)
\end{aligned}$$

On the other hand, the weighted mean of observed outcomes among women with weights

$$\omega(\mathbf{v}) \equiv \frac{f(\mathbf{v} \mid X=0)}{f(\mathbf{v} \mid X=1)} = \frac{p(X=1)\,p(X=0 \mid \mathbf{v})}{p(X=0)\,p(X=1 \mid \mathbf{v})}$$

defined before in equation (2) is given as:

$$\begin{aligned}
E_\omega(Y_{obs} \mid X=1) &= \int_{\mathbf{v}} \omega(\mathbf{v}) \Big[\int_u (\phi(\boldsymbol{\theta}_1, \mathbf{v}) + \theta_{1u}(\mathbf{v})u)\, f(\mathbf{v}, u \mid X=1)\,du \Big] d\mathbf{v} \\
&= \int_{\mathbf{v}} \frac{f(\mathbf{v} \mid X=0)}{f(\mathbf{v} \mid X=1)} \Big[\int_u (\phi(\boldsymbol{\theta}_1, \mathbf{v})\, f(\mathbf{v} \mid X=1)\, f(u \mid \mathbf{v}, X=1)\,du \Big] d\mathbf{v} + E_\omega(\theta_{1u}(\mathbf{v})u \mid X=1)) \\
&= \int_{\mathbf{v}} \phi(\boldsymbol{\theta}_1, \mathbf{v})\, f(\mathbf{v} \mid X=0)\,d\mathbf{v} + E_\omega(\theta_{1u}(\mathbf{v})u \mid X=1)). \qquad (12)
\end{aligned}$$

Then from equations (11) and (12), we obtain

$$E(Y_1 \mid X=0) = E_\omega(Y_{obs} \mid X=1) + \{E(\theta_{1u}(\mathbf{v})u \mid X=0) - E_\omega(\theta_{1u}(\mathbf{v})u \mid X=1)\}. \quad (13)$$

The DFL estimator of the “unexplained” inequality with covariates **V** is given as $E_\omega(Y_{obs} \mid X=1) - E(Y_{obs} \mid X=0)$, and, therefore

$$\begin{aligned}
E(Y_1 - Y_0 \mid X=0) = [&\text{DFL estimator for unexplained inequality}] \\
&+\{E(\theta_{1u}(\mathbf{v})u \mid X=0) - E_\omega(\theta_{1u}(\mathbf{v})u \mid X=1)\}. \qquad (14)
\end{aligned}$$

Hence, the DFL estimator is an unbiased estimator of $E(Y_1 - Y_0 \mid X=0)$ only if *U* is independent of *X* in the weighted population with weights $\omega(\mathbf{v})$ for *X* = 1 and constant weights of 1 for *X* = 0. However, if *U* affects intervening variables **V**, this does not hold because the DFL estimator has a bias of $E(\theta_{1u}(\mathbf{v})u \mid X=0) - E_\omega(\theta_{1u}(\mathbf{v})u \mid X=1)$. In the next section, I introduce a method to eliminate this bias.

It is worth examining whether the DFL estimator for explained inequality can be interpreted as inequality that would be eliminated under another counterfactual situation: where women had men’s covariate distribution but retained their own value of *u*. Recall that for such an interpretation, we have already seen in equation (8) that $E(Y_1 \mid X=1) - E(Y_1 \mid X=0)$ requires the ignorability assumption $X \perp U \mid \mathbf{V}$ to hold. However, we are concerned here with the meaning of $E(Y_{obs} \mid X=1) - E_\omega(Y_{obs} \mid X=1)$, which can be expressed as

$$\begin{aligned} E(Y_{obs} \mid X=1) - E_{\omega}(Y_{obs} \mid X=1) &= \int_v E(Y_{obs} \mid X=1, \mathbf{v}, u) f(\mathbf{v}, u \mid X=1) d\mathbf{v} du \\ &- \int_v E(Y_{obs} \mid X=1, \mathbf{v}, u) \omega(\mathbf{v}) f(\mathbf{v}, u \mid X=1) d\mathbf{v} du \\ &= \int_v E(Y_{obs} \mid X=1, \mathbf{v}, u)(f(\mathbf{v} \mid X=1) - f(\mathbf{v} \mid X=0)) f(u \mid X=1, \mathbf{v}) d\mathbf{v} du \; . \\ &= \int_v E(Y_{obs} \mid X=1, \mathbf{v})(f(\mathbf{v} \mid X=1) - f(\mathbf{v} \mid X=0)) d\mathbf{v}. \quad (15) \end{aligned}$$

Hence, the interpretation holds: the DFL estimate for the “explained” component of inequality can still be interpreted as the extent of inequality that would be eliminated if women had men’s covariate distribution.

However, there is an important caveat. There is an inherent ambiguity in conceiving the counterfactual situation at the individual level for the case where women have men’s covariate distribution. This is unlike the other counterfactual situation where men have women’s covariate effects, which can be specified at the individual level as a change of parameters from the set of $\{\boldsymbol{\theta}_0, \theta_{0u}\}$ to the set of $\{\boldsymbol{\theta}_1, \theta_{1u}\}$. The derivation above is based on an implicit assumption that $f(u \mid X=1, \mathbf{v})$ remains the same when the macro-social distribution of **V** changes. Such a result will be realized under the counterfactual situation where for each group of women for a given **v**, their proportion in the population changes from $f(\mathbf{v} \mid X=1)$ to $f(\mathbf{v} \mid X=0)$. Then, $f(u \mid X=1, \mathbf{v})$ will remain the same because we have the same set of women for each value of **v**. As an example, this implies that for a counterfactual situation where the proportion of college graduates increases from a factual 40% to a hypothetical 50%, we are assuming implicitly that the 40% of women who actually attained college graduation would occupy the 50% of women. This contrasts with counterfactual situations where women’s proportion of college graduation reached 50% via other mechanisms: for example, where an additional 10% of women, among those who actually did not attain college education, newly attain college graduation. In this case, the $f(u \mid X=1, \mathbf{v})$ may not remain the same because we do not know the joint distribution of *U* and **V** for those additional 10%. Hence, if we do not make the assumption that $f(u \mid X=1, \mathbf{v})$ will remain the same under macro-social change in the distribution of **V**, we cannot interpret the DFL component of "explained inequality" as the extent of inequality that would be eliminated when women had men's covariate distribution. Note that if we make a stronger assumption that the covariates are exogenous, and therefore $f(u \mid X=1, \mathbf{v}) = f(u \mid X=1)$, then the above-mentioned qualification for the counterfactual situation when women have men's covariate distribution becomes unnecessary. This is the situation that the original DFL method as well as the BO method assumes.

As a conclusion, the DFL method permits an interpretation of the "explained" component of inequality – even when covariates are endogenous – as the extent of inequality that would be eliminated under the counterfactual situation where each group of women who attained a distinct set of covariate values occupies the women's population as much as the group of men of the same covariate set does for the men's population. However, the "unexplained" component may not be interpreted as the extent of inequality that would be eliminated if men were treated like women.

2.3 A Method for Eliminating Bias in the Decomposition Analysis

A good thing about the results of equation (14) is that in order to eliminate bias in the DFL estimate of unexplained inequality as an estimate of the direct treatment effect of *X* on *Y*, we do not need to eliminate selection bias in all the endogenous intervening variables, but only need to eliminate the effect of covariance between *X* and *U* that may exist when covariates **V** are held constant. This is because we already have some methods for handling such a problem. Another merit of equation (11) is that it shows that bias in the DFL estimate arises from a significant covariance between *X* and *U* independent of the cause of the covariance between *X* and *U*. The covariance may be introduced by holding intervening covariates $\mathbf{V}_I$ constant or may have existed prior to a control for intervening variables because the ignorabillity assumption $X \perp U(\mathbf{V}_B) \mid \mathbf{V}_B$ does not hold. Regardless of the alternative cause of the covariance between *X* and *U*, we can obtain an unbiased estimate of the direct treatment effect of *X* if we eliminate the effect of this covariance.

When the methodological issue is the endogeneity of variable *X*, then a standard econometric method includes the econometric instrumental-variable (IV) method for linear regression. However, we will not use that method. The method makes a much stronger assumption than the RCM, and, in particular, assumes that the effect of *X* on *Y* is homogenous and is represented by a single regression coefficient. Although the reformulation of the IV method from the RCM was made by Angrist et al. (1996) as the local average treatment estimator (LATE), this alternative method requires that the single dichotomous instrumental variable causally precedes and affects the treatment variable *X.* Such an instrumental variable assumed by the LATE, however, does not exist when the treatment variable indicates a variable such as sex or race that is not affected by any other variable. A person-specific fixed-effect model with panel survey data can be employed for the BO decomposition method to handle the issue of endogenous covariates if the treatment variable is time-varying (Fortin et. al. 2011). A a model with person-specific fixed effects, however, cannot be employed in the present case where the treatment-assignment variable is gender or race where nobody in the sample experienced a treatment, that is, an observed change in race or gender.

Hence, this paper considers a combination of the method introduced in the preceding section with Heckman's two-step estimation method (Heckman 1979; Heckman and Robb 1986) for handling selection bias in the state of a dichotomous categorical state by modeling correlated errors, because although it makes additional assumptions about a parametric characterization of the error terms, and about a condition described below for instrumental variables to satisfy , we can retain other assumptions described in the preceding sections. The original Heckman method, however, was developed for linear regression models for the outcome. We therefore need a modified derivation of the method with a semiparametric model for the outcome without assuming independence between covariates and the unobserved determinant of the outcome.

Since this method relies on a probit regression that assumes an underlying continuous variable *X**, described below, one may argue that it is not adequate for predicting *X*, such as sex or gender, which is a nominal categorical variable. If we assume the error to be subject to unobserved random variability beyond the binomially distributed random error associated with a dichotomous outcome of a latent probability, probit regression with the normally distributed error seems an acceptable option. However, researchers who are hesitant about this option may employ a variation of the Heckman method based on the logistic selection equation (Dubin and Rivers 1989), though such an alternative does not permit the two-step procedure described below.

We consider the weighted population characterized by weights $\omega(\mathbf{v})$, which are defined by equation (2) for $X = 1$ and are constant at 1 for $X = 0$. It is because we are concerned with the elimination of non-independence between *X* and *U* that may exist after this weighting. For the weighted population, we assume $X \perp \mathbf{V}$ because both group 0 and group 1 have $f(\mathbf{v} \mid X = 0)$ when observations in the group with $X = 1$ are weighted by $\omega(\mathbf{v}) = p(\mathbf{v} \mid X = 0) / p(\mathbf{v} \mid X = 1)$. We assume that the sample weighted by the estimate of $\omega(\mathbf{v})$ also satisfies $X \perp \mathbf{V}$ statistically – though this condition actually needs to be satisfied with the propensity scores estimated in the analysis. The model for the observed and potential outcome at the individual level is the same as equation (6), as before, although we denote the unobserved determinant of *Y* by U_y below.

We assume that for the weighted population, the set of covariates **W** predicts variable *X** defined below. We do not include covariates **V** as predictors of *X* because **V** should be statistically independent of *X* for the weighted sample. Note that not unlike the estimation of the propensity score $p(X = 1 \mid \mathbf{v})$ in the DFL method, the use of the probit model in predicting *X* does not imply that the variables **W** affect *X*, because we are concerned only with the modeling of covariance of U_x and U_y, where U_x indicates the error term of *X** defined in equation (17) given below, to eliminate bias generated by covariance between *X* and U_y . However, variables **W** should be instrumental variables for which the following condition must hold:

$$Y_{obs} \perp \mathbf{W} \mid \mathbf{V}, X. \qquad (16)$$

It is a condition that can actually be tested with data for the candidates of instrumental variables. It follows that the parameters of equations (9) are not affected by **W**. Note that since $Y_1 = Y_{obs}$ when *X* = 1 and $Y_0 = Y_{obs}$ when *X* = 0, Y_1 and Y_0 are both are functions of **V** and U_y . It follows that the condition of equation (16) also implies that $U \perp \mathbf{W} \mid \mathbf{V}, X$ holds. In other words, although we do not make the ignorability assumption of $U \perp X \mid \mathbf{V}$, we make a distinct ignorability assumption $U \perp \mathbf{W} \mid \mathbf{V}, X$ for the instrumental variables.

For such **W**, we also make the following assumption:

A6: Variable *X* takes a value of 1 if *X** > 0 is satisfied for the latent continuous variable *X**, and *X** satisfies the following model:

$$X_i{}^* = \gamma(\mathbf{w}_i) + u_{x,i}, \qquad (17)$$

where $E(u_{x,i}) = 0,\ V(u_{x,i}) = 1,$ and $\gamma(\mathbf{w})$ is a parametric function of **W** that needs to be specified. Finally, we also make another assumption about the bivariate distribution of U_x and U_y :

A7 For given values of **V**, Variable U_y has conditional mean $E(u_y \mid \mathbf{v})$ and variance $V(u_y \mid \mathbf{v}) = \sigma_y{}^2(\mathbf{v}),$ and {U_x, U_y} are bivariate normally distributed with covariance $COV(u_x, u_y \mid \mathbf{v}) = \sigma_{xy}(\mathbf{v}).$

Hence, we allow the mean and the variance of U_y and the covariance between U_x and U_y to depend on **V**. This seems natural, because the covariance between *X* and U_y can be induced by holding covariates **V** constant.

According to Johnson and Kotz (1972; see also Heckman 1979, Maddala 1983), we generally obtain for a bivariate normal distribution $N(u_x, u_y, \sigma_x, \sigma_y, \sigma_{xy})$ with $E(u_x) = E(u_y) = 0$,

$$E(u_y \mid u_x > c) = (\sigma_{xy} / \sigma_x)\frac{\phi(Z)}{1-\Phi(Z)},$$

where $Z = c/\sigma_x$, and $\phi(Z)$ and $\Phi(Z)$ are, respectively, the probability density function and the cumulative distribution function of the standard normal distribution.

Since $E(u_y \mid u_x > c)p(u_x > c) + E(u_y \mid u_x < c)p(u_x < c) = E(u_y) = 0,$

we obtain

$$\begin{aligned} E(u_y \mid u_x < c) &= -E(u_y \mid u_x > c)p(u_x > c))/p(u_x < c)) \\ &= -(\sigma_{xy}/\sigma_x)\frac{\phi(Z)}{1-\Phi(Z)}\frac{1-\Phi(Z)}{\Phi(Z)} = -(\sigma_{xy}/\sigma_x)\frac{\phi(Z)}{\Phi(Z)}. \end{aligned}$$

For the present method, the mean of u_y may depend on covariates **V** because of their possible endogeneity. However, since the above characteristics hold for $u_y{*} \equiv u_y - E(u_y \mid \mathbf{v})$ that satisfies $E(u_y{*} \mid \mathbf{v}) = 0$ and $COV(u_x, u_y{*} \mid \mathbf{v}) = COV(u_x, u_y \mid \mathbf{v})$, we obtain from $U_y \perp \mathbf{W} \mid \mathbf{V}, X$ and $\sigma_x = 1$ that

$$\begin{aligned} E(u_y \mid \mathbf{v}, \mathbf{w}, X = 0) &= E(u_y \mid \mathbf{v}, X = 0) = E(u_y \mid \mathbf{v}, u_x < -\gamma(\mathbf{w})) \\ &= E(u_y \mid \mathbf{v}) - \sigma_{xy}(\mathbf{v})\frac{\phi(-\gamma(\mathbf{w}))}{\Phi(-\gamma(\mathbf{w}))} \\ &= E(u_y \mid \mathbf{v}) - \sigma_{xy}(\mathbf{v})\frac{\phi(\gamma(\mathbf{w}))}{1-\Phi(\gamma(\mathbf{w}))}, \qquad (18) \end{aligned}$$

and

$$\begin{aligned} E(u_y \mid \mathbf{v}, \mathbf{w}, X=1) &= E(u_y \mid \mathbf{v}, X=1) = E(u_y \mid \mathbf{v}, u_x > -\gamma(\mathbf{w})) \\ &= E(u_y \mid \mathbf{v}) + \sigma_{xy}(\mathbf{v}) \frac{\phi(-\gamma(\mathbf{w}))}{1-\Phi(-\gamma(\mathbf{w}))} \\ &= E(u_y \mid \mathbf{v}) + \sigma_{xy}(\mathbf{v}) \frac{\phi(\gamma(\mathbf{w}))}{\Phi(\gamma(\mathbf{w}))}. \qquad (19) \end{aligned}$$

Now we can reevaluate bias in the DFL estimate, which is given in equation (14) as $E(\theta_{1u}(\mathbf{v})u \mid X=0) - E_\omega(\theta_{1u}(\mathbf{v})u \mid X=1)$. We first derive the case where neither $\theta_{1u}(\mathbf{v})$ nor $\sigma_{xy}(\mathbf{v})$ depends on **V.** Then, we obtain from equations (18) and (19),

$$\begin{aligned} & E(\theta_{1u}u_y \mid X=0) - E_\omega(\theta_{1u}u_y \mid X=1) \\ &= \theta_{1u}\left\{\int_{\mathbf{v}}\int_{\mathbf{w}} E(u_y \mid \mathbf{v}, \mathbf{w}, X=0) f(\mathbf{w}, \mathbf{v} \mid X=0) d\mathbf{w} d\mathbf{v}\right\} \\ &- \theta_{1u}\left\{\int_{\mathbf{v}}\int_{\mathbf{w}} \omega(\mathbf{v}) E(u_y \mid \mathbf{v}, \mathbf{w}, X=1) f(\mathbf{w}, \mathbf{v} \mid X=1) d\mathbf{w} d\mathbf{v}\right\} \\ &= \theta_{1u}\left\{\int_{\mathbf{v}} E(u_y \mid \mathbf{v}) f(\mathbf{v} \mid X=0) d\mathbf{v} - \int_{\mathbf{v}}\int_{\mathbf{w}} \sigma_{xy} \frac{\phi(\gamma(\mathbf{w}))}{1-\Phi(\gamma(\mathbf{w}))} f(\mathbf{w}, \mathbf{v} \mid X=0) d\mathbf{w} d\mathbf{v}\right\} \\ &- \theta_{1u}\left\{\int_{v} \omega(\mathbf{v}) E(u_y \mid \mathbf{v}) f(\mathbf{v} \mid X=1) d\mathbf{v} + \int_{\mathbf{v}}\int_{\mathbf{w}} \omega(\mathbf{v}) \sigma_{xy} \frac{\phi(\gamma(\mathbf{w}))}{\Phi(\gamma(\mathbf{w}))} f(\mathbf{w}, \mathbf{v} \mid X=1) d\mathbf{w} d\mathbf{v}\right\} \\ &= -\theta_{1u}\sigma_{xy}\left\{E\left(\frac{\phi(\gamma(\mathbf{w}))}{1-\Phi(\gamma(\mathbf{w}))}\middle| X=0\right) + E_\omega\left(\frac{\phi(\gamma(\mathbf{w}))}{\Phi(\gamma(\mathbf{w}))}\middle| X=1\right)\right\}. \qquad (20) \end{aligned}$$

Note that the first component of $E(\theta_{1u}u_y \mid X=0)$ and that of $E_\omega(\theta_{1u}u_y \mid X=1)$ that depend on $E(u_y \mid \mathbf{v})$ are canceled out because $\omega(\mathbf{v}) f(\mathbf{v} \mid X=1) = f(\mathbf{v} \mid X=0)$.

Let *H* for each person *i* with $\mathbf{w}_i$ be defined as

$$H_i \equiv \left\{\begin{array}{ll} \dfrac{\phi(\gamma(\mathbf{w}_i))}{1-\Phi(\gamma(\mathbf{w}_i))} & \text{when } X_i = 0 \\ \dfrac{\phi(\gamma(\mathbf{w}_i))}{\Phi(\gamma(\mathbf{w}_i))} & \text{when } X_i = 1 \end{array}\right\}. \qquad (21)$$

Then, by combining this with equation (14), we obtain:

$$E(Y_1 - Y_0 \mid X = 0) = E_\omega(Y_{obs} \mid X = 1) - E(Y_{obs} \mid X = 0)$$
$$- \theta_{1u}\sigma_{xy}\left\{E\left(H\middle| X = 0\right) + E_\omega\left(H \mid X = 1\right)\right\}. \qquad (22)$$

Since the estimates of $E\left(H\middle| X = 0\right)$ and $E_\omega\left(H \mid X = 1\right)$ can be obtained from weighted probit regression with *X* as the dependent variable and **W** as covariates, the remaining issue is the estimation of $\theta_{1u}\sigma_{xy}$. The expected value of the observed outcome at the individual level for person *i* with $X_i = 1$, is given from equations (9) and (19) as:

$$E_u(y_{obs,i|x_i=1}) = \phi(\mathbf{v}_i, \boldsymbol{\theta}_1) + \theta_{1u} E(u_y \mid \mathbf{v}_i, X_i = 1)$$
$$= \phi(\mathbf{v}_i, \boldsymbol{\theta}_1) + \theta_{1u} E(u_y \mid \mathbf{v}_i) + \theta_{1u}\sigma_{xy}\frac{\phi(\gamma(\mathbf{w}_i))}{\Phi(\gamma(\mathbf{w}_i))}. \qquad (23)$$

Hence, an estimate of $\theta_{1u}\sigma_{xy}$ is given by the coefficient of $\phi(\gamma(\mathbf{w}_i))/\Phi(\gamma(\mathbf{w}_i))$, which is Heckman's inverse Mill's ratio. Note that $E(u_y \mid \mathbf{v}_i)$ is not a function of unknown individual $u_{y,i}$ but is a function of $\mathbf{v}_i$ and the parameters of the conditional distribution of u_y, $f(u_y \mid \mathbf{v})$. Since $\phi(\mathbf{v}, \boldsymbol{\theta}_1)$ and $E(u_y \mid \mathbf{v}_i)$ are both functionally unspecified, we have two options here in estimating $\theta_{1u}\sigma_{xy}$.

Suppose that we have an effective single categorical instrumental variable *W*, and let

$$\omega_W(\mathbf{v}) \equiv p(W \mid X = 1) / p(W \mid X = 1, \mathbf{V}) \ . \qquad (24)$$

Then, it follows that

$$\omega_W(\mathbf{v}) f(\mathbf{v}, w, u_y \mid X = 1) = \frac{p(w \mid X = 1)}{p(w \mid X = 1, \mathbf{v})} p(\mathbf{v} \mid X = 1) p(w \mid X = 1, \mathbf{v}) p(u_y \mid X = 1, \mathbf{v}, w)$$
$$= p(\mathbf{v} \mid X = 1) p(u_y \mid X = 1, \mathbf{v}) p(w \mid X = 1) \text{ since } U_y \perp W \mid X, \mathbf{V}$$
$$= p(\mathbf{v}, u_y \mid X = 1) p(w \mid X = 1). \qquad (25)$$

Equation (25) shows that the data of *X* = 1 weighted by $\omega_W(\mathbf{v})$ preserve the joint distribution of **V** and *U* and the distribution of *W* becomes independent . This independence of *W* from the weighted joint distribution of **V** and *U* holds when the data of *X* =1 is weighted further by $\omega(\mathbf{v}) = p(\mathbf{v} \mid X = 0) / p(\mathbf{v} \mid X = 1)$. When *W* is independent of (**V**,*U*$_y$) for *X* = 1, then $H_i = \phi(\gamma(\mathbf{w}_i)) / \Phi(\gamma(\mathbf{w}_i))$ in equation (23) is independent of both $\phi(\mathbf{v}, \boldsymbol{\theta}_1)$ and $E(u_y \mid \mathbf{v}_i)$. Then we can apply a regression equation $y = a + bH + \varepsilon$ to the data of *X* = 1 weighted by the estimate of $\omega(\mathbf{v})\omega_W(\mathbf{v})$, because the effects of $\phi(\mathbf{v}, \boldsymbol{\theta}_1)$ and $E(u_y \mid \mathbf{v}_i)$ are independent of *H* in the weighted sample and can therefore be ignored, and we obtain an estimate of $\theta_{1u}\sigma_{xy}$ as the estimate for coefficient *b.* This strategy retains the semiparametric characteristic for the outcome.

When an effective single categorical instrumental variable is not available, however, we have to control for the effects of covariates **V** in linearizing the covariate effects in equation (19) in order to obtain an estimate of $\theta_{1u}\sigma_{xy}$. Since we wish this estimate to depend little on the modeling of $\phi(\mathbf{v}_i, \boldsymbol{\theta}_1) + E(u_y \mid \mathbf{v}_i)$ as a function of $\mathbf{v}_i$, it is desirable to employ the categorical expression of covariates, and all significant two-way and higher-order interaction effects among the covariates should be included in the model.

When either $\theta_{1u}(\mathbf{v})$ or $\sigma_{xy}(\mathbf{v})$ or both depend on **V**, equation (20) holds for each given set of covariates **V**, and therefore, we obtain

$$\begin{aligned} E(Y_1 - Y_0 \mid X = 0) = E_\omega(Y_{obs} \mid X = 1) - E(Y_{obs} \mid X = 0) \\ - E\left\{\theta_{1u}(\mathbf{v})\sigma_{xy}(\mathbf{v})\left(E\left(H \middle| \mathbf{v}, X = 0\right) + E_\omega\left(H \mid \mathbf{v}, X = 1\right)\right)\right\}. \end{aligned} \quad (26)$$

We can model and estimate $\theta_{1u}(\mathbf{v})\sigma_{xy}(\mathbf{v})$ by including the interaction effect of covariates **V** and the control variable *H*. Two options regarding the unbiased estimation of such interaction effects are the same as before.

We need caveats for interpreting the interaction effects of *X* and another variable, say V_1 , on the outcome, that is, for the estimate of $E(Y_1 - Y_0 \mid X = 0, V_1 = a) - E(Y_1 - Y_0 \mid X = 0, V_1 = b)$. Interpreting them requires caveats because although we can obtain an unbiased treatment effect of *X* by following the procedure described above, we do not make any control for possible selection bias in the state of V_1 , and therefore, the interaction effect $E(Y_1 - Y_0 \mid X = 0, V_1 = a) - E(Y_1 - Y_0 \mid X = 0, V_1 = b)$ cannot be interpreted as the unbiased effect of V_1 on change in the direct treatment effect of *X* on *Y*. On the other hand, each of $E(Y_1 - Y_0 \mid X = 0, V_1 = a)$ and $E(Y_1 - Y_0 \mid X = 0, V_1 = b)$ permits an interpretation respectively as the direct treatment effect of change in *X* among people with *X* = 0 and $V_1 = a$ and that among people with *X* = 0 and $V_1 = b$, but the interaction effect should be interpreted simply as a difference in the treatment effect of *X* on *Y* between the two groups that differ in the state of V_1 , and this difference is not independent of the selection bias in the state of V_1 .

The application should be done in the following steps.

Step 1. We employ either a logit or probit regression to obtain a consistent estimate of $p(X = 1 \mid \mathbf{v})$. We need to conduct a diagnosis for an adequate construction of propensity scores to generate weighted sample for which statistical independence between *X* and **V** should hold.

Step 2. We create weights $\omega(\mathbf{v})$ for the group with *X* = 1, and constant weights of 1 for the group with *X* = 0. We need to re-create $\omega(\mathbf{v})$ when the diagnosis indicates inadequacy.

Step 3. We apply the probit model using the instrumental variable **W** with data weighted by $\omega(\mathbf{v})$.

Step 4. We create variable $H_i = \dfrac{\phi(\gamma(\mathbf{w}_i))}{\Phi(\gamma(\mathbf{w}))}$ for each sample individual *i* in group *X* = 1, and assign a value of $H_i = \dfrac{\phi(\gamma(\mathbf{w}_i))}{1-\Phi(\gamma(\mathbf{w}_i))}$ for each sample individual *i* in group *X* = 0.

Step 5. If an effective single categorical instrumental variable *W* was available and employed in step 3, we need to create a second set of propensity-score weights, $\omega_W(\mathbf{v})$ by obtaining the estimate of $p(W \mid X = 1, \mathbf{v})$ with the data of *X* = 1 by logistic regression or probit regression, or by their multinomial versions. We need to conduct a diagnosis for an adequate construction of the weight that should attain statistical independence between *W* and **V**.

Step 6. With sample data of *X* = 1 weighted by $\omega(\mathbf{v})\omega_W(\mathbf{v})$, we run the outcome regression with only *H* as the predictor to obtain an estimate of $\theta_{1h}\sigma_{xy}$. as the regression coefficient of *H* .

Step 7. We should elaborate step 6 by trying various possible interaction effects of *H* and a subset of **V** to identify the best fitting model.

Step 8. If the control variable *H* shows a significant effect in step 6 or 7, then we calculate the adjusted estimate of $E(Y_1 - Y_0 \mid X = 0)$ by using equation (22) and the estimate of $\theta_{1h}\sigma_{xy}$ when no interaction effect of *H* and **V** exist, and by using equation (26) and the estimate of $\theta_{1u}(\mathbf{v})\sigma_{xy}(\mathbf{v})$ when some interaction effects of *H* and **V** exist.

The following step 5* should be substituted for steps 5 and 6 when we do not have an effective single categorical instrumental variable, and should conduct step 7 as an elaboration of step 5*.

Step 5*. With the sample data of *X* = 1 weighted by $\omega(\mathbf{v})$ we run the outcome regression with *X*, covariates **V**, and *H* as predictors to obtain an estimate of $\theta_{1u}\sigma_{xy}$.

Note that if we wish the outcome to reflect an alternative counterfactual situation where women for whom $X = 1$ were treated like men (for whom $X = 0$), we only need to replace the above-described procedure by switching the categories of the dummy variable X.

3. APPLICATION

3.1 Data and Alternative Hypotheses on the Nature of Endogeneity

The application presented here employs data from the Comparative Survey of Work Life Balance conducted in 2009 by the Research Institute of Economy, Trade and Industry (RIETI) in Japan for the population of employees in four countries. The following analysis uses the sample of employed men and women of ages 23-59 in Japan. The survey for Japan includes a nationally representative random sample of white-collar regular employees in firms with 100 or more employees. Excluding 692 samples with missing values of annual wage/salary, or 7.3% of the total, reduces the sample size to 8,811; 6,145 men and 2,666 women. While about 41% of people in the labor force were women in Japan, regular employment was underrepresented by women, leading to a smaller proportion of female samples than that in the labor force.

The analysis focuses on the decomposition of gender inequality in earned annual income into components that are not explained by the following covariates. Two nested models for covariates are employed. The first model employs age (7 categories), education (4 categories), employment duration for the current employment (8 categories), and marital status (2 categories). The second model adds to model 1 the average hours of work per week (4 categories). The two models are used to clarify how much of the proportion of gender inequality in earned annual income is explained by gender difference in the first four covariates and how much more can be explained by taking into account gender difference in hours of work as an additional covariate. Unlike employees in the United States, the majority of white-collar employees in Japan, other than those in the managerial position and certain professionals, are not exempt from payment for overtime work, and therefore, hours of work strongly affect earned annual income among white-collar regular employees.

For the gender effect, exogenous correlated causes may not exist, because men and women are equally likely to be born into various families. Age, education, and employment duration are major human-capital variables in Japan, where the age-based or duration-based *nenko* wage system is still quite prevalent. Hours of work differ greatly depending on the combination of gender and marital status, because married women work fewer hours than married men even among regular employees. The instrumental variables are (1) whether the subjects ever had a childcare leave, and (2) the extent of agreement about a statement that “men and women should both aim at a balance between family and work.” For item (2), two dummy variables were constructed for the responses “I strongly agree” and “I somewhat agree” as contrasted with the remaining three categories, which did not indicate agreement. As shown later, those three variables, one for childcare leave and two for work-family balance, have strong correlations with gender, and at the same time no unique effects on earned annual income when the effects of the four covariates of model 1 and their significant interactions are controlled for.

Marital status is a strong indicator of gender-specific sample selection because about 60%-70% of women quit their jobs during periods of childrearing, and because many Japanese firms give priority for regular employment to those who have graduated from schools in the recent past, many women reentering the labor force after they quitted their jobs get employed as irregular employees. Hence, a strong interaction effect of gender and marital status on exit from and reentry into regular employment exists. Hence, sample selection bias in regular employment by the unobserved determinants of income may also exist for the combination of gender and marital status, in addition to the possible induced correlation between latent ability and gender when education is held constant under gender inequality in educational attainment in Japan. The gender-specific selection mechanism associated with marital status may generate an association between the unobserved determinant of income and gender in two ways that predict the opposite direction of association.

The opportunity costs hypothesis. Women whose opportunity costs of leaving their jobs are relatively higher because of their better prospects for future income are more likely to remain as regular employees during the periods of childrearing and are more likely to reenter the labor force. This tendency will generate a positive correlation of marital status (the married versus the single) with the unobserved variable that positively affects income among women.

The adverse selection hypothesis. Statistical discrimination against women -- which is believed to exist strongly in Japan, especially for married women, because of many employers' reinforcement of the traditional division of household labor-- will make women with higher income potential than others more likely to quit during the periods of childrearing and less likely to reenter the labor force. This tendency will generate a negative correlation of marital status (the married versus the single) with the unobserved variable that positively affects income among women.

The theory of the adverse selection mechanism associated with discrimination was introduced by Schwab (1986), who applied the theory of information asymmetry and consequent adverse selection initially theorized by Akerlof (1970) in the context of the commodity market rather than in the context of the labor market. Since selection bias in the unobserved variable occurs mainly for women after marriage, the two hypotheses will yield a significant correlation of gender and the unobserved variable among the married, but in the opposite direction.

3.2 Preliminary Analyses

Two kinds of preliminary analyses were conducted. One was concerned with the diagnosis of adequacy in the construction of propensity scores. While the statistical independence of the joint distribution of covariates from gender by weighted data cannot be tested completely, the paper tests (1) the statistical independence of each covariate from gender after weighting by $\omega(\mathbf{v})$ and (2) the absence of significance of all covariates combined for the logistic regression model employed in estimating propensity scores after the weighting.

The second preliminary analysis is concerned with testing the adequacy of the instrumental variables.

3.2.1. Propensity Score Estimation and Its Diagnosis

A logistic regression is employed in the estimation of propensity scores in predicting the gender distinction. For model 1 with four covariates, in addition to the main effects of the covariates, significant category-by-category interaction effects between education and age, and marital status and age, were found by pairwise tests of all possible interaction effects and are therefore included. For model 2, various interaction effects involving hours of work as well as its main effects were found to be significant, and they were all necessary to attain statistical independence between gender and the five covariates. Those interaction effects included in model 2, in their most parsimonious form, are (a) category-by-category interaction effects between hours of work and marital status, (b) interaction effects between working 49 hours or more and education, (c) interaction effects between working 49 hours or more and linear age (based on the distinction of 7 categories), (d) three-factor interaction effects between working 49 hours or more, marital status, and education, (e) three-factor interaction effects between working 49 hours or more, marital status, and linear age.   While factor (c) was not statistically significant, it was included because higher-order interaction effects, factor (e), need to be included.

Table 1 presents for each model the test of independence between each covariate and gender, and the significance test of the effects of all covariates, including their interaction effects, in the logistic regression before and after weighting by the propensity-score weights. The results in Table 1 show that although the covariates are strongly associated with gender individually as well as jointly, those associations almost completely disappear for data weighted by propensity-score weights defined by equation (2).

(Table 1 About Here)

3.2.2 Examinations of Instrumental Variables

We need to confirm two facts for the instrumental variables to be effective: (1) condition $Y_{obs} \perp \mathbf{W} \mid \mathbf{V}, X$ holds, and (2) the instrumental variables are strongly associated with gender.

Table 2 presents the results from a linear regression model on income with gender, the four covariates, and all significant category-by-category interaction effects between gender and each covariate as control variables, along with instrumental variables. The coefficients for the control variables are omitted. Results show that none of the instrumental variables affect income significantly. An additional analysis with five covariates, though the results are omitted here, led to the conclusion.

(Table 2 About Here)

Table 3 shows, for data weighted by propensity scores obtained for each of the two models for predicting propensity scores, the results of a probit regression with instrumental variables. The results indicate that, controlling for covariates **V**, the effects of instrumental variables in predicting the proportion of women are all significant at the 0.1% level for the results from each model and that women are more likely to take a childcare leave and are more likely to agree with the statement on the importance of balancing family and work for both men and women.

(Table 3 About Here)

3.3 Main Analyses

First, I tested whether gender was correlated with the unobserved determinant of the outcome. Since no effective single categorical instrumental variable was available, I employed the second procedure for the estimation of $\theta_{1h}\sigma_{xy}$ by using the sample of women as specified by equation (19). The control variables included in modeling equation (19) to estimate $\theta_{1h}\sigma_{xy}$ for the model with four covariates are categorical effects of all the four covariates and category-by-category interaction effects of age and education that were found to be significant. For the model with the five covariates, the main categorical effects of hours of work and category-by-category interaction effects of age and hours of work were also added. The effect of the inverse Mills ratio was not significant at the 5% significance level, and neither were its interaction effects with each of the covariates for each of the two models. Hence, neither the opportunity costs hypothesis nor the adverse selection hypothesis was supported – though it is possible that both hypotheses held but their effects canceled out each other. Hence, in the present analysis, the DFL estimate can be interpreted as the decomposition of gender inequality in income into the direct treatment effect of gender and the effect explained by gender difference in covariate distributions.

Table 4 shows the results from (1) the unweighted regression without covariates (model 0), (2) the weighted regression without covariates (model 1A) and with covariates (model 1B) with weights based on the propensity scores of model 1 with four covariates, and (3) the weighted regression without covariates (model 2A) and with covariates (model 2B) with weights based on the propensity scores of model 2 with five covariates. Models 1A and 2A give the standard DFL estimates, and models 1B and 2B give doubly robust estimates of them (Bang and Robins 2005). Models 0, 1A and 2A without covariates do not assume any regression equation, because their coefficients are just the average income for men for the intercept and the difference in the average income between women and men for the effect of gender. Agreements in the gender effect between models 1A and 1B and between models 2A and 2B are expected from the statistical independence between gender and covariates in the weighted sample and are confirmed here. A comparison of models 0 and 1A indicates that the gender difference in earned annual income explained by the gender difference in the four covariates (age, education, employment duration, and marital status) is about 34% [=(189.7-125.9)/189.7=0.336]. Similarly, a comparison of models 0 and 2A indicates that the gender difference explained by five covariates (the four covariates and hours of work) is about 43% [=(189.7-108.4)/189.7=0.429]. Hence, by adding hours of work as an additional covariate, the proportion of explained gender inequality increases by about 9%.

Table 5 shows the results from (1) the unweighted regression without covariates, and (2) the two DFL estimates, but adds the main effects of marital status and the interaction effect of gender and marital status. Similar to models 0, 1A, and 2A in Table 4, those three models without covariates do not assume any regression equation, because their results are based only on the average income for the cross-classification of gender and marital status without weights and with weights. It should be noted first that the propensity-score weighting we employed makes only gender distribution statistically independent of the covariates; it does not make marital-status distribution statistically independent of the covariates. Hence, as described in the section "Method," the strong interaction effect of gender and marital status, which indicates that gender inequality in income is much greater for married women than for single women, does not indicate the average treatment effect of marital status. Since the RCM assumes heterogeneity of treatment effects, which are gender effects here, the greater gender effect on income among the married than among the single can reflect both (a) the treatment effect of marital status on income, and (b) difference between single and married persons in the composition of heterogeneous gender effects on income. On the other hand, we can still interpret the gender effect after the weighting by propensity-score weights as the average treatment effect of gender for each category of marital status. The results of Table 5 for those gender effects are summarized in Table 6.

The decomposition in Table 6 is based, for example, on the calculation that the extent of gender inequality in annual income explained by the gender difference in age, education, and employment duration is 38% [=(101.7-63.1)/101.7] for the single and 30% for the married [=(101.3+113.2-63.1-86.4)/(101.3+113.2)] . In particular, the results of Table 5 indicate that the additional consideration of gender difference in hours of work makes only a small contribution to the explained proportion for the single, but it makes a much larger contribution to the explained proportion for the married, increasing the proportion from 30% to 41%. Hence, gender differences in income come not only from gender differences in human capital but also from gender differences in hours of work among married women, which seem to be caused by the greater share of family roles for married women than for married men. However, for both the single and the married, nearly 60% of the gender inequality income cannot be explained by the covariates included in the model, suggesting some consequences of discrimination against women.

4. CONCLUSION

While decomposition analysis for inequality is quite useful, its assumption may be problematic when we evaluate the method from the counterfactual conception of causality despite the fact that the decomposition itself implicitly assumes a counterfactual situation. This paper reformulated the DiNardo-Fortin-Lemieux (DFL) method from the viewpoint of the Rubin's causal model and introduced a method to correct its possible bias by combing the DFL method with Heckman's two-step method for the control of selection bias. I consider the use of the method introduced in this paper a complement to the DFL method for examining the possibility of bias in DFL decomposition analysis and for detecting the direction of bias.

Since the issue in decomposition analysis is the possible non-independence between the unobserved determinant of the outcome and the treatment variable, some other methods, such as the use of a pair of latent-class variables, may be used for the same purpose when the outcome variable is categorical, such as being a manager or not, because the method introduced in this paper by combining the DFL method with a linear regression based on Heckman's method for adjustment is not employable for such a dichotomous outcome. The formal expression of bias by equation (14), however, holds regardless of the choice of a method for eliminating the bias, and the formal results shown in this paper will be useful for such further methodological developments.

This paper also intends to bring gender and race back into the center of statistical causal analysis. As I discussed in the introduction, the use of panel survey data for causal analysis seems to have diminished the importance of gender or race in statistical causal analyses because of a lack of methodological framework to handle such time-constant exogenous variables as the treatment variable in causal analysis. I believe that the discussion and the method presented in this paper will lead to a reconsideration of such trends, and will be complementary to the experimental audit method, since gender and racial inequality is a major substantive research topic in sociology.

Table 1. Diagnostic Tests of Statistical Independence between Gender and Covariates after Weighting

Covariates	L^2	*df*	*P*	L^2	*df*	*P*
I. Before propensity-score weighting						
1. Pairwise test on independence of the covariate from gender						
(1) Age	545.96	6	0.000			
(2) Education	956.63	3	0.000			
(3) Employment duration	247.89	7	0.000			
(4) Marital status	757.98	1	0.000			
(5) Hours of work	804.27	3	0.000			
2. Logistic regression to predict propensity scores						
Model 1 with 4 covariates	2,202.98	41	0.000			
Model 2 with 5 covariates	2,805.20	58	0.000			
II. After propensity-score weighting						
	Model 1			Model 2		
1. Pairwise test on independence of the covariate from gender						
(1) Age	1.82	6	0.935	6.68	6	0.352
(2) Education	1.57	3	0.666	0.49	3	0.922
(3) Employment duration	3.86	7	0.797	6.96	7	0.434
(4) Marital status	0.14	1	0.713	0.10	1	0.750
(5) Hours of work	----	-	-----	6.46	3	0.091
2. Logistic regression to predict propensity score						
	19.52	41	0.998	42.79	58	0.933

Table 2. Insignificance of Instrumental Variables on Income

(in 10,000 yen)

	Coefficient	S.D.	*t*	*P*
I. Intercept + Covariate effects (38 parameters)	Omitted			
II. Instrumental variables.				
1. Ever had a childcare leave	6.385	8.153	.78	.434
2. Importance of balancing family and work for both men and women (versus “can’t agree or disagree”, “somewhat disagree” or “strongly disagree”				
(1) “strongly agree”	-1.753	4.533	.39	.699
(2) “somewhat agree”	-2.164	4.479	.48	.629

Table 3. The effects of Instrumental Variable on Gender for the Weighted Population: Probit Regression

	Model 1				Model 2			
	B	*S.D.*	*Z*	*P*	*B*	*S.D.*	*Z*	*P*
1. Intercept	-1.016	0.032	32.04	.000	-1.017	0.032	31.97	.000
2. Ever had a childcare leave	1.522	0.055	27.92	.000	1.410	0.056	25.11	.000
3. Importance of balancing family and work for both men and women (versus “can’t agree or disagree”, “somewhat disagree” or “strongly disagree”								
“strongly agree”	0.601	0.039	15.51	.000	0.636	0.039	16.42	.000
“somewhat agree”	0.279	0.040	6.97	.000	0.309	0.040	7.73	.000

Table 4. The DFL analysis for the effect of gender on Income (in 10,000 yen)

	Unweighted	DFL-1		DFL-2	
	Model 0	Model 1A	Model 1B	Model 2A	Model 2B
1. Intercept	530.8***	530.8***	342.4***	530.8***	341.7***
2. Gender (versus men)					
Women	-189.7***	-125.9***	-122.8***	-108.4***	-110.7***
3. Age (versus 23-29)					
30—34	-----	-----	45.0***	-----	46.9***
35-49	-----	-----	96.6***	-----	100.7***
40-44	-----	-----	147.7***	-----	153.0***
45-49	-----	-----	187.8***	-----	193.2***
50-54	-----	-----	202.5***	-----	199.5***
55-59	-----	-----	220.8***	-----	228.6***
4. Education (versus four-year college)					
JC/ATS[1]	-----	-----	-59.8***	-----	-62.7***
VS[2]	-----	-----	-73.4***	-----	-63.8***
high school	-----	-----	-99.4***	-----	-103.6***
5. Employment Duration (versus 0-5 years)					
6-10	-----	-----	35.6***	-----	35.9***
11-15	-----	-----	36.1***	-----	34.7***
16-20	-----	-----	59.8***	-----	63.0***
21-25	-----	-----	68.0***	-----	58.3***
26-30	-----	-----	85.4***	-----	90.4***
31+	-----	-----	122.0***	-----	125.5***
Missing	-----	-----	59.0***	-----	61.3***
6. Marital Status (vs. Single)					
Married	-----	-----	35.3***	-----	35.5***
7. Hours of Work per week (vs 40 or less)					
41-48	-----	-----	-----	-----	18.3***
49 or more	-----	-----	-----	-----	36.6***
Missing	-----	-----	-----	-----	2.7

[1] Junior college or advanced technical school (14 years of education)

[2] Vocational School after high school.

***p<.001; **p<0.01; *p<0.05; #p<0.10

Table 5. An Analysis of the Average Gender Effect by Marital Status (in 10,000 yen)

	Unweighted	DFL-1	DFL-2
1. Intercept	434.8***	434.8***	434.8***
2. Gender (versus men)			
Women	-101.7***	-63.1***	-59.5***
3. Interaction: Gender × Married			
	-113.2***	-86.4***	-67.8***
4. Marital Status (vs. Single)			
Married	132.7***	132.7***	132.7***

[1] Junior college or advanced technical school (14 years of education)
[2] Vocational School after high school
[3] Numbers in parentheses are those confounded with the values of *H*.

Table 6. A Summary of Treatment Effects by Marital Status

	Total population	Single	Married
Covariates	Age, education, employment duration, marital status	Age, education, employment duration	
% Explained	33.6	38.0	30.4
% Unexplained	66.4	62.0	69.6
Covariates	Age, education, employment duration, marital status, hours of work	Age, education, employment duration, hours of work	
% Explained	42.9	41.5	40.8
% Unexplained	57.1	58.5	59.2

REFERENCES

Akerlof, G. 1970. “The Market for Lemons: Quality, Uncertainty, and the Market Mechanism.” *Quarterly Journal of Economics.* 84: 488-500.

Angrist, J.,D., G. W. Imbens, and D.B. Rubin, 1996. “Identification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association* 91: 444-55.

Bang, H. and Robins, J.M. 2005. “Doubly Robust Estimation in Missing Data and Causal Inference Models.” *Biometrics* 61: 962-972.

Barsky, R., J. Bound, K.C. Kerwin, and J. P. Lupton. 2002. “Accounting for Black-White Wealh Gap: A Nonparamaric Approach.” *Journal of the American Statistical Association* 93: 663-73.

Blinder, A. 1973. “Wage Discrimination: Reduced Form and Structural Variables.” *Journal of Human Resources* 8: 436-55.

Correll, S.J., S. Benard, and I.P. Source. 2007. “Getting a Job: Is there a Motherhood Penalty?” *American Journal of Sociology* 112: 1297-1339.

DiNardo, J., N. Fortin, and T. Lemieux 1996. “Labor Market Institution and the Distribution of Wages.” *Econometrica* 64:1001-44.

Dubin, J.A. and D. Rivers. 1989. “Selection Bias in Linear Regression, Logit and Probit Models.” *Sociological Methods and Research* 18: 360-90.

Fortin, N., T. Lemieux, and S. Firpo. 2011. “Decomposition Methods in Econometrics.” Pp. 1-102 in *Handbook of Labor Economics*, Volume 4a, edited by O. Ashenfelter and D. Card. New York: Elsevier B.V.

Heckman, J. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica* 47:153-161.

Heckman, J. and Robb, R. 1986. “Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes.” In H. Wainer (Ed.), *Drawing Inferences from Self-Selected Samples* (pp. 63-107). Mahwah, NJ: Lawrence Erlbaum Associates.

Jones, F.L. and J. Kelley. 1984. “Decomposing Differences between Groups: A Cautionary Note on Measuring Discrimination.” *Sociological Methods and Research* 12: 323-43. Johnson, N. and S. Kotz. 1972. *Distribution in Statistics: Continuous Multivariate Distributions.* New York: John Wiley and Sons.

Kim, ChangHwan. 2013. “Detailed Wage Decompositions: Revisiting the Identification Problem.” *Sociological Methodology* 43: 346-63.

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Economics*, New York: Cambridge University Press.

Morgan S. and C. Winship, 2007. *Counterfactuals and Causal Inference*.Cambridge: Cambridge University Press.

Oaxaca, R. 1973. “Male-Female Wage Differentials in Urban labor Markets.” *International Economic Review* 14: 693-709.

Oaxaca, R. and M.R. Ransom. 1999. “Identification in Detailed Wage Decomposition.” *Review of Economics and Statistics* 81: 154-57.

Pager, D. 2003. “The Mark of a Criminal Record.” *American Journal of Sociology* 108: 937-75.

Pearl, J. 2009. *Causality (*2nd edition). Cambridge: Cambridge University Press.

Powers, D. and M. Yun. 2009. “Multivariate Decomposition of Hazard Rate Models. Sociological Methodology 39: 233-63.

Robins, L.M. 1998. “Marginal Struructual Models” 1997*. Proceedings of the American Statistical Association, Section on Bayesian Statistical Science, pp.1-10.*

Rosenbaum, P. R. and D. B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70:41-55.

Rosenbaum, P. R. and D. B. Rubin. 1984. “Reducing Bias in Observational Studies Using Sub-classification on the Propensity Scores.” *Journal of the American Statistical Association* 79: 516-24.

Rubin, D.B. 1985. “The Use of Propensity Scores in Applied Bayesian Inference”. Pp. 463-72 in J. M. Bernardo, M. H. De Groot, D. V. Lindley and A. F. M. Smith (eds.), *Bayesian Statistics,* vol. 2. North-Holland: Elsevier.

Schwab, S. 1986. “Is Statistical Discrimination Efficient?” *The American Economic Review* 76: 228-234.

Yun, Meong-Su. 2008. “Identification Problem and Detailed Oaxaca Decomposition: A General Solution and Inference.” *Journal of Economic and Social Measurement* 33: 27-38.