ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Rios-Avila, Fernando

Working Paper

A semi-parametric approach to the Oaxaca-Blinder decomposition with continuous group variable and self-selection

Working Paper, No. 930

Provided in Cooperation with:
Levy Economics Institute of Bard College

Suggested Citation: Rios-Avila, Fernando (2019) : A semi-parametric approach to the Oaxaca-Blinder decomposition with continuous group variable and self-selection, Working Paper, No. 930, Levy Economics Institute of Bard College, Annandale-on-Hudson, NY

This Version is available at:
https://hdl.handle.net/10419/209173

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Working Paper No. 930

A Semi-Parametric Approach to the Oaxaca-Blinder Decomposition with Continuous Group Variable and Self-Selection

by

Fernando Rios-Avila
Levy Economics Institute of Bard College

May 2019

The Levy Economics Institute Working Paper Collection presents research in progress by Levy Institute scholars and conference participants. The purpose of the series is to disseminate ideas to and elicit comments from academics and professionals.

Levy Economics Institute of Bard College, founded in 1986, is a nonprofit, nonpartisan, independently funded research organization devoted to public service. Through scholarship and economic research it generates viable, effective public policy responses to important economic problems that profoundly affect the quality of life in the United States and abroad.

Levy Economics Institute
P.O. Box 5000
Annandale-on-Hudson, NY 12504-5000
http://www.levyinstitute.org

Copyright © Levy Economics Institute 2019 All rights reserved

ISSN 1547-366X

ABSTRACT

This paper describes the application of a semiparametric approach, known as a varying coefficients model (Hastie and Tibshirani 1993), to implement a Oaxaca-Blinder type of decomposition in the presence of self-selection into treatment groups for a continuum of comparison groups. The flexibility of this methodology may allow for detecting heterogeneity of the role of endowment and coefficient effects when analyzing endogenous dose treatments. The methodology is then used to revisit the impact of obesity on wages (Cawley 2004), using body mass index (BMI) as the continuous group variable. The results suggest that body weight does have a negative impact on wages for white women, but the impact decreases for higher BMI levels. For white men, the impact is also negative and significant, but positive for low levels of BMI, which explains why they are not significant in the linear instrumental variables approach.

KEYWORDS: Oaxaca-Blinder Decomposition; Heckman Selection; Heckit; Semiparametric; Kernel; Nonlinear; Body Mass Index; Weight; Wages

JEL CLASSIFICATIONS: C14; I19; J31; J71

1. INTRODUCTION

Since the seminal papers from Blinder (1973) and Oaxaca (1973), many studies have used what is known as the Oaxaca-Blinder (OB) decomposition for analyzing outcomes differences between two well-defined groups. Such differences are characterized as functions of differences in characteristics (composition effect) and differences in coefficients associated with those characteristics (*wage* structure effect). Subsequent research provided refinements that extended the OB decomposition analysis to nonlinear functions and distributional statistics other than the mean, as well as strategies to identify the model when some of the underlying assumptions do not hold (see Fortin, Lemieux, and Firpo [2011] for a review of other methodological extensions).

While the OB decomposition can be directly applied to scenarios with naturally discreet groups (i.e., union and nonunion workers, men and women, whites and nonwhites), the application of OB-type decompositions in cases with a continuum of comparison groups is not standard. Ñopo (2008) and Ulrick (2012) have proposed extensions to the standard OB decomposition allowing for a continuous group variable, using ad hoc parametric approximations. Neither strategy, however, deals with the scenario where the assumption of conditional independence does not hold, as is the case when there is self-selection of individuals into groups based on unobservables (endogenous membership).

The purpose of this paper is to propose a strategy to extend the OB decomposition to a continuous group variable using a semiparamatric approach known as varying coefficient models (Hastie and Tabshiran 1993). To account for endogenous selection, I abstract from a generalization of the Heckman selection model (Heckman 1979; Lee 1978; Li and Racine 2007; Vella 1998). This strategy can be useful for analyzing heterogeneous dose-treatment effects when endogeneity in terms of self-selection is expected. For example, in the context of labor market outcomes, the methodology can be used for analyzing the impact of smoking and smoking intensity on wages (Hotchkiss and Pitts 2013), obesity and body mass index (BMI) on wages (Cawley 2004), or training duration on employment probabilities (Kluve et al. 2011).

The rest of the paper is structured as follows. Section 2 describes the basic OB decomposition analysis in the presence of self-selection/endogenous membership. Section 3 introduces the use of a generalized selection term, here called the generalized inverse Mills ratio (GIMR), when individuals self-select into more than two *ordered* groups. Section 4 describes the use of varying coefficient models in the implementation of an OB-type decomposition. Section 5 provides an example of the implementation of the methodology by revisiting the wage penalty of obesity based on the research of Cawley (2004). Section 6 concludes.

2. THE OB DECOMPOSITION WITH SELECTION: BASICS

In the standard OB approach, the goal is to analyze how differences in observed characteristics and returns to these characteristics contribute to the average differences in the outcomes between two groups. For the appropriate identification of the OB decomposition, the strategy requires that the potential outcomes can be estimated using two well-specified linear models with exogenous membership into each group. This ensures that the distribution of the errors is orthogonal to the group membership.

In many instances, however, the assumption of membership exogeneity is likely to be violated if individuals self-select to be part of a specific group (i.e., part of the treated group).[1] When this happens, the conditional distribution of the errors is no longer independent of the group membership, ruling out the identification strategy of the standard decomposition approach.

As described in Heckman (1979), this endogenous selection can be considered an omitted variable problem that can be corrected by modeling the selection process and using this information to add a correction term in the model specification.[2] This strategy requires the estimation of a three-equation model that is described as follows:

[1] Fortin, Lemiux, and Firpo (2011) provide other scenarios where the conditional independence assumption might be violated.
[2] This strategy has been used in the framework of the OB decomposition in terms of a switching regression model with unknown selection. See, for example, Lee (1978).

$$y_i = X_i\beta_A + \mu_{A,i}\ if\ D_i^* < 0 \text{ or } \varepsilon_i < -Z_i\gamma \quad (1a)$$

$$y_i = X_i\beta_B + \mu_{B,i}\ if\ D_i^* \geq 0 \text{ or } \varepsilon_i \geq -Z_i\gamma \quad (1b)$$

$$D_i^* = Z_i\gamma + \varepsilon_i \quad (1c)$$

Where D_i^* is the latent propensity of an individual (i) to be part of group B, and Z_i is a vector of variables related to individuals' membership that may include variables not included in X.[3] If we assume that $(\mu_{A,i}, \mu_{B,i}, \varepsilon_i)$ are distributed jointly normal:

$$\mu_{A,i}, \mu_{B,i}, \varepsilon_i \sim N\begin{pmatrix} 0 & \sigma_{A\mu}^2 & . & \rho_A\sigma_{A\mu} \\ 0, & . & \sigma_{B\mu}^2 & \rho_B\sigma_{B\mu} \\ 0 & \rho_A\sigma_{A\mu} & \rho_B\sigma_{B\mu} & 1 \end{pmatrix} \quad (2)$$

the model can be estimated using a full information maximum likelihood (FIML) or a two-step procedure (heckit). The latter involves including estimates for the inverse Mills ratio (IMR), also known as the selection correction term, in the main outcome model based on the information coming from the selection equation. In specific, for this setup, the IMR (λ) would be defined as follows:

$$E\left(\mu_{k,i}|Z_i, D\right) \propto \lambda_i = \frac{-\phi(Z_i\gamma)}{\Phi(-Z_i\gamma)} * 1(i \in A) + \frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} * 1(i \in B) \quad (3)$$

where $\phi(.)$ stands for the normal density function and $\Phi(.)$ for the normal cumulative density function.

The parameters (γ) can be obtained by estimating equation (1c) using a probit model, while unbiased estimations for equations (1a) and (1b) can be obtained using ordinary least squares (OLS) by including the corresponding IMR as explanatory variables:

$$y_i = X_i\beta_A + \delta_A\lambda_i + e_i^A\ if\ i \in A \quad (4a)$$

[3] While identification of the Heckman selection model can be obtained based on nonlinearity alone, having an instrumental variable is recommended for better identification of the model.

$$y_i = X_i\beta_B + \delta_B\lambda_i + e_i^B \; if \; i \in B \tag{4b}$$

In this setting, an estimation of the adjusted outcome gap after controlling for selection can be written as follows:

$$E(y_i \,|i \in B) - E(y_i \,|i \in A) = \Delta\bar{y} = \bar{x}_B\hat{\beta}_B + \hat{\delta}_B\bar{\lambda}_B - \bar{x}_A\hat{\beta}_A + \hat{\delta}_A\bar{\lambda}_A \tag{5}$$

$$\Delta\bar{y} - \left(\hat{\delta}_B\bar{\lambda}_B - \hat{\delta}_A\bar{\lambda}_A\right) = \Delta\bar{y}_S = \bar{x}_B\hat{\beta}_B - \bar{x}_A\hat{\beta}_A \tag{6}$$

and can be used to implement any of the standard OB decompositions based on assumptions of the counterfactual wage structure.[4] As described in Fortin, Lemieux, and Firpo (2011), outcome differences accounted for by differences in the coefficients (structure effect) can be interpreted as the treatment effect of membership, after adjusting for differences in observed characteristics and endogenous selection.

3. GENERALIZED SAMPLE SELECTION

In the model described above, we assume that the only information known about the selection process is that individuals are members of one of two groups (A or B). As discussed in Vella (1998), the grouping variable may contain additional information, such as intensity, that can be used to obtain a better approximation of the selection correction term, even if the interest remains in analyzing differences between two groups.

As before, consider a model where we observe the continuous characteristic (D_i) for each individual, which can be used to broadly classify them into groups A and B (dichotomization of the groups). This characteristic could be the number of hours worked per week, number of cigarettes smoked in a month, or weeks of training before reentering into the labor force, among others. The selection process and outcome equations can be described as follows:

[4] For example, assuming counterfactual wages are given by the wage structure observed in group B, the components of the decomposition would be given by $\Delta\bar{y}_S = (\bar{x}_B - \bar{x}_A)\beta_B + \bar{x}_A(\beta_B - \beta_A)$, where $\bar{x}_A(\beta_B - \beta_A)$ can be interpreted as a treatment effect under the conditional independence assumption.

$$y_i = X_i\beta_A + \mu_{A,i}\ if\ D_i \leq c \quad (7a)$$

$$y_i = X_i\beta_B + \mu_{B,i}\ if\ D_i > c \quad (7b)$$

$$D_i = Z_i\gamma + \varepsilon_i \quad (7c)$$

with $\mu_{A,i}, \mu_{B,i}, \varepsilon_i$ following a joint normal distribution as defined before, with some arbitrary threshold (c) to define membership, and with equation (7c) representing the equation (or equations) that describe the data-selection process. It is easy to see that this model reverts to the standard switching regression model if a dichotomous transformation $1(D_i > c)$ is used for equation (7c).

Many authors have proposed various alternatives for the estimation of these types of selection models, using both parametric and semiparametric strategies (see Li and Racine [2007, sec 10.3] and Vella [1998]). In general, following the approach proposed by Heckman (1979), these methodologies suggest that to obtain consistent estimators for the parameters (β), one should include an approximation of the selection bias term as a control in the main regression model. In this paper I concentrate on three methodologies that assume the overall distribution of D is observed, with extensions to scenarios where D is partially observed.

Vella (1998) discusses the estimation of models such as the one described above and suggests that a feasible strategy is to estimate the selection process (equation [7c]) as a tobit model if D has a censored distribution. Without loss of generality, assuming D is censored at zero, the selection correction term or IMR is defined as:[5]

$$E\left(\mu_{k,i}\middle|D_i, Z_i\right) \propto \lambda_i^* = -\frac{1}{\sigma_e}\frac{\phi\left(\frac{Z_i\gamma}{\sigma_e}\right)}{\Phi\left(-\frac{Z_i\gamma}{\sigma_e}\right)}1(D_i = 0) + \frac{1}{\sigma_e}\frac{D_i - Z_i\gamma}{\sigma_e} * 1(D_i > 0) \quad (8)$$

These are often called generalized residuals. It should be noticed that when D is not censored, equation (7c) can be estimated using standard OLS and the IMRs are simply the OLS residuals.

[5] It should be noticed that if the whole distribution of D is observed (i.e., there is no censored information in D), the procedure can also be done using a simple OLS model. This would be equivalent to a control function approach that includes the errors from the first step to the outcome model.

Including the residuals to the main models would be equivalent to the control function for endogeneity (Wooldridge 2015).

As Vella (1998) and Li and Racine (2007) describe, using this correction term provides estimations that are more stable and efficient than using the standard IMR (which assumes dichotomous grouping). However, similar to the analysis of endogenous variables, an instrumental variable is required to identify the coefficients of the selection correction term and the treatment intensity (D).

An alternative method described in Vella (1998) is one where the selection process corresponds to a setting with discreet but ordered selection rules. If we assume that $\widetilde{D}$ is a discretized transformation or classification of D (i.e., $\widetilde{D} = K \; if \; c \in \{ll_k, ul_k\}$), and that $\widetilde{D}_k^*$ is the latent propensity of an individual to be part of group $\widetilde{D} = K$, then the selection equation process can be written as:

$$\widetilde{D}_{k,i}^* = Z_i\gamma_k + \varepsilon_i \tag{9a}$$

$$\widetilde{D}_i = \begin{cases} 0 & if \quad \widetilde{D}_{1,i}^* < 0 & \rightarrow \varepsilon_i < -Z_i\gamma_1 \\ 1 & if \quad \widetilde{D}_{1,i}^* > 0 \;\&\; \widetilde{D}_{2,i}^* < 0 & \rightarrow -Z_i\gamma_1 \leq \varepsilon_i < -Z_i\gamma_2 \\ \vdots & & \vdots \\ J-1 & if \quad \widetilde{D}_{J-1,i}^* > 0 \;\&\; \widetilde{D}_{J,i}^* < 0 & \rightarrow -Z_i\gamma_{J-1} \leq \varepsilon_i < -Z_i\gamma_J \\ J & if \quad \widetilde{D}_{J,i}^* > 0 & \rightarrow -Z_i\gamma_J \leq \varepsilon_i \end{cases} \tag{9b}$$

Note that equation (9b) is a different way of writing the selection model described in Vella (1998), where all coefficients in γ_k are permitted to vary. Also note that all latent coefficients are affected by the same shock (ε_i). Under the parallel lines assumption (Williams 2016), an ordered probit model (O-probit) can be used to estimate this model, where only the constant is allowed to vary across models.

Similar to the binary-group case, the outcome equations can be consistently estimated using OLS by simply including a selection correction term, which for the selection rule described by equations (9a) and (9b) takes the form:

$$E\left(\mu_{k,i}\middle|\widetilde{D}_i, Z_i\right) \propto \lambda_i^* = \frac{-\phi(Z_i\gamma_1)}{1-\Phi(Z_i\gamma_1)} 1_{\widetilde{D}_i=0} + \sum_{k=1}^{J-1} \frac{\phi(Z_i\gamma_k)-\phi(Z_i\gamma_{k+1})}{\Phi(Z_i\gamma_k)-\Phi(Z_i\gamma_{k+1})} * 1_{\widetilde{D}_i=k} + \frac{\phi\left(Z_i\gamma_J\right)}{\Phi\left(Z_i\gamma_J\right)} * 1_{\widetilde{D}_i=J} \quad (10)$$

where λ_i^* is the GIMR (Vella 1998). Here the term $E\left(\mu_{k,i}\middle|\widetilde{D}_i, Z_i\right)$ is only an approximation of the correction term $E\left(\mu_{k,i}\middle|D_i, Z_i\right)$, as it can be considered as the expected value of the correction term for all values of D_i within the group $\widetilde{D}_i$. However, if more detailed groups are created and larger samples are available, one should expect $E\left(\mu_{k,i}\middle|\widetilde{D}_i, Z_i\right) \to E\left(\mu_{k,i}\middle|D_i, Z_i\right)$. If no instrumental variables are used in the selection equation model, the GIMR will be strongly linear with the estimated latent index, and the estimator will be poorly identified.

As described in Chernozhukov, Fernandez-Val, and Melly (2013), there are more flexible alternatives for the estimation of the selection model, by allowing all parameters in γ_k to vary with D and by estimating all possible models for each threshold in D. This can be done using independent models (Foresi and Perachi 1995), or using simultaneous models such as the generalized O-probit model (Terza 1985). Both alternatives, however, impose great computational burden and may produce unrealistic predicted probabilities in the model as the number of groups (J) increase.[6]

Taking from the literature on distributional regressions (Chernozhukov, Fernandez-Val, and Melly 2013), the last alternative suggested here is to use global distributional regressions to characterize the cumulative distribution of the outcome $F(D|z)$. This can be done using a fractional probit model that takes the form:

$$F(D_i|z) = P(d \leq D_i|z) = \Phi(Z_i\gamma) \quad (11)$$

[6] See Williams (2016) for a brief discussion of this problem in the case of generalized ordered logit models, where the model produces negative probabilities of belonging to a particular group.

Empirically, this model can be estimated by substituting $P(d \leq D_i|x)$ with the sample estimator of the unconditional cumulative distribution $\hat{F}(D_i) = \frac{1}{n}\sum 1(d_i < D_i)$, or some other approximation of it.[7] In this case, the corresponding GIMR takes the form:

$$E\left(\mu_{k,i}\middle|D_i, Z_i\right) \propto \lambda_i^* = \hat{F}(D_i) * \frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} - \left(1 - \hat{F}(D_i)\right)\frac{\phi(Z_i\gamma)}{\Phi(-Z_i\gamma)} \quad (12)$$

Once the corresponding selection correction terms have been estimated, and the average wage gap corrected for the selection term, the OB decomposition can be implemented in the standard way, using equation (6). In this framework, the structure effect can be interpreted as the average treatment between the untreated and treated group.

4. THE OB DECOMPOSITION WITH A CONTINUUM OF GROUPS WITH SELECTION

4.1. Varying Coefficients Model and Heterogeneity of the Treatment Effect

The previous section described the construction of sample selection correction terms that use the information on the intensity of the treatment/selection variable to obtain the GIMR, which can be used to implement an OB decomposition comparing any two groups. A simple generalization of the OB structure that accounts for a continuum of groups can be written as:

$$y_i = X_i\boldsymbol{\beta_D} + \mu_i \quad (13)$$

where $\boldsymbol{\beta_D}$ is a vector of parameters that vary with the grouping variable (D). Without loss of generality, including the GIMR term into the model to obtain unbiased estimates through OLS provides a model that can be written as:

$$y_i = X_i\boldsymbol{\beta_D} + \boldsymbol{\delta_D}\lambda_i^* + e_i \quad (14)$$

[7] This can be done, for example, using the kernel cumulative density estimation of D.

where X_i is a vector that includes the constant and explanatory variables, and λ_i^* is the estimate of the GIMR for person i.[8]

In principle, as stated in Ulrick (2012), with enough information it is possible to estimate all the parameters in the above equation for any value of D by simply estimating models with constrained data. However, in most applications, the number of observations with a specific value for D may be insufficient to provide an appropriate estimation of coefficients $B(D) = \{\beta_D, \delta_D\}$. Borrowing from the literature on nonparametric econometrics, feasible estimations can be obtained for the parameters $B(D)$ using an extension of local regression estimations, known as varying coefficient models (Hastie and Tibshirani 1993; Li and Racine 2007).[9] Using this strategy, one imposes no restrictions on the coefficients $B(D)$ other than them being smooth and differentiable in D.

This method expands on the use of kernel local smoothing regressions, allowing for a flexible parameterization of the outcome model in equation (14), modeling the conditional mean $E(y_i|D = d)$ as a linear function of explanatory variables and a selection term in the neighbor of $D = d$. This would, in principle, allow us to obtain estimates of the coefficients $B(D)$ for every point of interest (*d*):

$$E(y_i|D = d) = \widehat{m}_y(d) = E(W_i B(d)|d) = E(W_i|d)B(d) = \widehat{m}_w(d)B(d) \qquad (15)$$

with $W_i = [1, X_i, \lambda_i^*]$ and the function $\widehat{m}_z(d)$ representing the conditional mean of any variable Z in the neighborhood *d*. This model can be estimated by minimizing the objective function:

$$Min_{B(d)}\, L = \sum \left(y_i - W_i B(d)\right)^2 K\left(\frac{D_i - d}{h}\right) \qquad (16)$$

which is equivalent to minimizing the weighted squares errors of the model, with weights given by the kernel function, $K(.)$, and the bandwidth, h. As discussed in Hastie and Tibshirani (1993),

[8] Notice that λ_i^* does not vary with respect to the grouping variable (D), but rather the individual realization (D_i).
[9] See Cameron and Trivedi (2005, ch. 9) for details on kernel regression estimators.

to reduce problems with boundary bias, the recommendation is to use a local constant approximation for $B(d) \cong \mathrm{B}^0(d) + \mathrm{B}^1(d)(D_i - d)$. The constant component of these coefficients, $B^0(d) = [\beta_D^0, \delta_D^0]$, represents the local effect that a variable (X) has on the outcome (y) in the neighborhood of $D = d$. This can be used to implement the OB decomposition for the selectivity-corrected outcome between any two particular groups, depending on assumptions regarding the reference group (Fortin, Lemieux, and Firpo 2011).

4.2. Bandwidth Selection and Standard Errors

An important aspect of the estimation of varying coefficients is the choice of bandwidth (h). Larger bandwidths help reduce the variance of the estimated parameters, but increase the bias. In contrast, smaller bandwidths can reduce the bias at a cost of higher variance.[10] While there are a few suggestions in the literature regarding to the choice of bandwidths (see, for example, Hoover et al. [1998]), a leave-one-out crossvalidation procedure using a single smoothing parameter (h) for smoothing all explanatory variables is used here. This implies choosing h so that it minimizes the following expression:

$$CV_{loo}(h) = \sum_{i \in B} \omega(D_i)\left(y_i - X_i \beta_{D_i}^{-i}(h) - \delta_{D_i}^{-i}(h) * \lambda_i^*\right)^2 \qquad (17)$$

where $\beta_{D_i}^{-i}(h)$ and $\delta_{D_i}^{-i}(h)$ are the i^{th} leave-one-out estimated coefficients for a given bandwidth (h) and a point of interest (D); $\omega(D_i)$ is a weight function that serves to avoid difficulties of slow convergence caused by the sparse distribution of D. Because the bandwidth does not affect the calculation of the GIMR, the parameter λ_i^* will be considered exogenous for the estimation of the crossvalidation criteria.

In the present context, the analytical estimation of the standard error of varying coefficient models with selection can be considerably cumbersome to implement. Under the assumption that the selection term is fixed and exogenous, Li and Racine (2007) provide expressions for the asymptotic distribution of the standard errors for the local linear estimator of varying coefficient

[10] See Li and Racine (2007, sec. 9.3.2)

models.[11] However, because the model described above is based on a two-step estimation process, the estimation of the standard errors needs additional adjustments (Heckman 1979).

Because of the added complexity, a more feasible method, albeit computationally intensive, is using pair-wise bootstrapped standard errors. The benefits of this strategy have been discussed in Yatchew (2003) and Keele (2008), and, more recently, its application has been formally discussed in Cattaneo and Jansson (2018) in the framework of kernel regressions. The procedure can be described as follows:

Step 1. Obtain a random paired bootstrap sample (S_1) from the original sample.
Step 2. Estimate the selection correction term (λ^*_{S1}) using any of the methods presented in section 3.
Step 3. Estimate the coefficient for the outcome models for all points of interest d, based on the bootstrap sample (S_1) using local kernel regressions.
Step 4. Estimate the decomposition components for the group(s) of interest.
Step 5. Repeat steps 1 through 4 B-times to obtain the empirical distributions' aggregated and detailed decomposition components.

In the next section, an application of this semiparametric strategy is presented, revising the main results from Cawley (2004), where BMI will be used as the continuum of groups for the decomposition analysis.

5. APPLICATION: REVISING THE IMPACT OF OBESITY ON WAGES

Several studies have found that body weight is negatively correlated to wages, in particular for white women (Cawley 2004; Sabia and Rees 2012; Averett 2011; Fikkan and Rothblum 2012). The most common explanations for this negative correlation are: obesity lowers wages by reducing productivity and increasing discrimination; low wages may cause obesity due to unhealthy eating habits caused by lower income; or that unobserved factors simultaneously cause

[11] See Li and Racine (2007, sec 9.3.2) for further details.

higher body weights and lower wages. In his review of the literature, Cawley (2004) criticizes the robustness of various strategies that have been followed in the literature to analyze the relationship between body weight and wages, and suggests the application of an instrumental variable approach to better capture the causal relationship between BMI and wages.

Using data from the National Longitudinal Survey of the Youth (NLSY) for the years 1981 to 2000, Cawley (2004) provides estimations for the impact of BMI and weight on wages, using siblings' BMI, sex, and age as instruments for own BMI.[12] Correcting for reporting errors on weight and height, the evidence of his preferred model suggests that the negative effect of higher BMI on wages is only statistically significant for white women, with no statistically significant effect for other groups.

For the illustration of the proposed methodology, BMI will be considered the continuous group variable that is used to analyze the wage gaps in relation to body weight. Due to the higher demands that the methodology imposes on the data, some changes on the data definitions and model specifications are introduced. To compare the results with the instrumental variable approach used in Cawley (2004), I first replicate the original results and present various estimations showing how sensitive the results are to changes in the variables' definitions and model specifications. Second, I briefly describe the specific OB decomposition approach used for the present example, given that no natural comparison group exists. Finally, I provide estimations of the semiparametric decomposition approach under the preferred assumptions.

5.1. Replication and Variable Definition Changes

Cawley (2004) estimates instrumental variable models for six different demographic groups based on gender and race, using measures for BMI that are corrected for self-reporting error[13] as the main explanatory variable, and using siblings' BMI, age, and sex as instrumental variables. Making use of clustered standard errors at the individual level and using sampling weights, he reports that BMI has a negative impact on wages for all groups and races, but is only statistically

[12] The author implements a larger set of regression analysis using methodologies previously used in the literature. However, for the purpose of this paper, I will concentrate only on the instrumental variable approach. Further details on the data construction can be found in Cawley (2004).
[13] See Cawley (2004, 454) for a complete description of the data and model specification.

significant for white women. Replications of these results are provided in table 1, pooling together blacks and Hispanics into nonwhites. According to this result, an increase in BMI of one point would translate into a 1.5 percent reduction in wages for white women, with no statistical impact for other groups.

Because of the multiple steps involved in the semiparametric methodology proposed here, the original model specification required some adjustments.[14] First, sampling weights are excluded from the analysis, so that clustered bootstraped standard errors can be applied directly. Second, in the original replication files, Cawley (2004) kept sample observations with missing data in the model specification. He did so by replacing missing values with zeros and adding dummy variables indicating if a variable has missing observations. To reduce the number of explanatory variables in the model, data with missing information in the general intelligence score, highest grade attained, job tenure, and county employment rate are excluded from the sample so that the dummy indicators are dropped from the specification as well. Instead of including both father's and mother's highest degree of education, both variables are combined to a single variable (parents' highest degree of education). Observations with missing data on both parents are also excluded from the sample. Finally, observations with a BMI below 14 and above 60 are also excluded from the sample. This reduces the total sample from 44,026 observations to 40,087 observations.

Reestimating the results using the same specifications used in Cawley (2004) incorporating the changes described above, shows that the conclusions are robust to the model and sample specification changes, with small changes in the point estimates (see table 1). From here forward, the replication will focus on the estimates for white males and females only, since the results are small and not statistically significant for nonwhite groups.

[14] See the appendix for a complete set of results and intermediate steps for the data and model specification changes.

Table 1. Replication and Modified Specification Results

Replication of Cawley (2004)				
Ln(wage per hr)	White		Nonwhite	
	Male	Female	Male	Female
BMI	-0.0131	-0.0168***	-0.00369	-0.00515
	[0.00831]	[0.00496]	[0.00508]	[0.00544]
N	13355	10800	11185	8686
Replication with changes in model specification and sample				
Ln(wage per hr)	White		Nonwhite	
	Male	Female	Male	Female
BMI	-0.0127	-0.0154***	-0.00425	-0.00735
	[0.00804]	[0.00493]	[0.00504]	[0.00545]
N	12184	10101	9844	7958

Note: Clustered standard errors at the individual level are in parenthesis. * p<0.1, ** p<0.05, *** p<0.01

Since the methodology proposed here uses various options for the estimation of the GIMR (a control function approach), I test the sensitivity of the results from Cawley (2004) in the restricted sample by reestimating the model including the GIMR in the specification. This is equivalent to adjusting for endogeneity using a control function approach (Wooldridge 2015). For the estimation of the O-probit model, two options for dependent variables are used: one that categorizes BMI data in 10 groups of equal size (OP1), and one that categorizes BMI data in 10 groups with the same range (OP2). For the fractional probit model (FP), the empirical cumulative distribution of BMI is used as the dependent variable. In addition to using the siblings' data as instruments, second-order interactions are also included as instruments to account for further nonlinear effects. The results with clustered bootstrapped standard errors are shown in table 2.

Table 2. Replication with Restricted Data: Control Function Approach

Instruments:	Siblings: BMI, age, and sex		BMI, age, sex + quadratic terms and interactions	
	White		Nonwhite	
	Male	Female	Male	Female
OLS GIMR	-0.0127	-0.0154***	-0.012	-0.0150***
	[0.00833]	[0.00527]	[0.00748]	[0.00530]
OP1 GIMR	-0.0149***	-0.0135***	-0.0143***	-0.0131***
	[0.00419]	[0.00280]	[0.00423]	[0.00280]
OP2 GIMR	-0.0171***	-0.0149***	-0.0162***	-0.0143***
	[0.00457]	[0.00297]	[0.00464]	[0.00296]
FP GIMR	-0.0128***	-0.0130***	-0.0122***	-0.0127***
	[0.00412]	[0.00273]	[0.00418]	[0.00272]
N	12184	10101	12184	10101

Note: Clustered bootstrapped standard errors at the individual level in parenthesis using 250 repetitions. OP1 uses a categorical variable that divides the sample in 10 groups of equal size; OP2 uses a categorical variable that divides the sample in 10 groups of equal range. * p<0.1, ** p<0.05, *** p<0.01

As expected, the results using the OLS residuals are identical to the standard instrumental variable approach, showing marginal changes when interactions are added as instruments. For the rest of the models, however, the results are somewhat different. Using alternative methods for the estimation of the GIMR shows that the impact of BMI is statistically significant for both white men and women at conventional levels. Furthermore, the impact of BMI is slightly larger for men, while it declines somewhat for women. It will be shown later that BMI does have a negative and significant impact on wages for men, but only over a restricted segment of the BMI distribution. For the rest of the paper, linear and quadratic terms of the instrumental variables will be used to account for nonlinear effects and identification, and decomposition will be implemented using OLS GIMR.

5.2. Semiparametric Oaxaca Decomposition

Oaxaca Decomposition Approach and Implementation

In order to implement an OB decomposition in the present framework, it is necessary to define an appropriate comparison/baseline/untreated group to analyze wage gaps across BMI. A common approach is to use individuals with a "healthy" BMI level as the baseline group, and compare the results against all other groups (over- and underweight). Following this premise, people with a BMI considered healthy (between 18.5 and 25) are used as the comparison group.

They represent approximately 48 percent of white men and 62 percent of white women. Using this reference group, the OB decomposition is obtained by estimating the following equations:

$$\ln(wage_i) = X_i\beta_h + \delta_1(BMI_i - \overline{BMI_i}) + \delta_2 GIMR_i + e \; if \; BMI_i \in (18.5,25) \quad (18a)$$

$$\ln(wage_i) = X_i\beta_t(d) + \delta_1(d) * (BMI_i - d) + \delta_2(d) * GIMR_i + e \; \forall d \in BMI_i \quad (18b)$$

The first equation is estimated using a sample of the comparison group only, whereas the second is estimating using kernel local linear regressions, as described in section 4.1. Notice that both equations include the GIMR variable to adjust for sample selection, and that BMI is also included in equation (18a) to control for any impact it may have on wages even within the healthy weight group.[15] This variable is centered at the mean, so it uses the average BMI as the reference point for estimating the constant.

For the implementation of the OB decomposition, a threefold decomposition is used on the selectivity-corrected wage gap using the following formulas:

Composition effect: $$\Delta X(d) = \big(\hat{m}_x(d) - E(X|h)\big)\beta_h \quad (19a)$$

Wage structure effect: $$\Delta\beta \; = E(X|h)(\beta^0(d) - \beta_h) \quad (19b)$$

Interaction: $$\Delta X(d)\Delta\beta = \big(\hat{m}_x(d) - E(X|h)\big)(\beta^0(d) - \beta_h) \quad (19c)$$

where $\hat{m}_x(d)$ is the local linear predicted mean of the variable X, $E(X|h)$ is the average characteristics for people with a healthy BMI, and β_h and $\beta^0(d)$ are the coefficients corresponding to the comparison group and for people with a BMI around d.

As described in section 3.2, the bandwidth for the kernel regressions is selected separately for white men and white women using a crossvalidation procedure. The specification in equation (18b) and the OLS GIMR are used as a benchmark for the estimation of the optimal bandwidths, which are used for all models, even if the GIMR is estimated through other methods. To reduce the impact of sparse areas in the distribution of BMI on the bandwidth selection, two approaches

[15] This follows the critique raised by Cain (1986) in regards to the use of a pooled sample as a comparison or nondiscriminatory group, where he suggests including the group variable as a control in the pooled regression.

were taken. The first is to set $\omega(D_i) = 0$ for observations that fall in the top and bottom 1 percent of the distribution. The second is to use a strictly monotonic transformation of BMI—specifically the cumulative distribution G(BMI)—as the running variable for the local linear regressions, avoiding the sparse distribution problem.[16] Using this transformation is similar to using of a varying bandwidth, since more information will be used in areas that are more sparsely distributed than others, but can also be compared to the use of k-nearest-neighbors estimators. All models are estimated using Gaussian kernel functions. Table 3 provides the optimal bandwidths obtained from the crossvalidation criteria for both men and women in the sample.

Table 3 Cross-validated Optimal Bandwidths

Variable of Reference	Men	CV criterion	Women	CV criterion
BMI	3.2900	-1.40814	4.8540	-1.56378
G(BMI)	0.1769	-1.40852	0.2241	-1.54543

Note: CV=Crossvalidation log of mean squared leave-one-out error.

5.3. Aggregate Decomposition Results

Figure 1 plots the total selectivity-corrected wage gap across the BMI for men and women, comparing people at all points of the BMI distribution with those in the comparison group. The panels on the right provide the estimates that use the original BMI variable for the semiparametric regression, while the panels on the left show the estimates using the transformed variable, G(BMI). The darker and lighter regions show the 90 percent and 95 percent confidence intervals constructed using a clustered bootstrap procedure with 1,000 repetitions. For men and women, the displayed gaps are provided for the relevant range of BMI, which excludes the top and bottom 1 percent of the distribution.

According to the estimations, the selectivity-corrected wage gap for men and women exhibits an inverse-U shape with respect to their BMI. For women, I estimate a negative but not statistically significant wage gap for all points of the BMI distribution. Based on the semiparametric estimation that relies on the transformed BMI data, women at the top of the BMI distribution

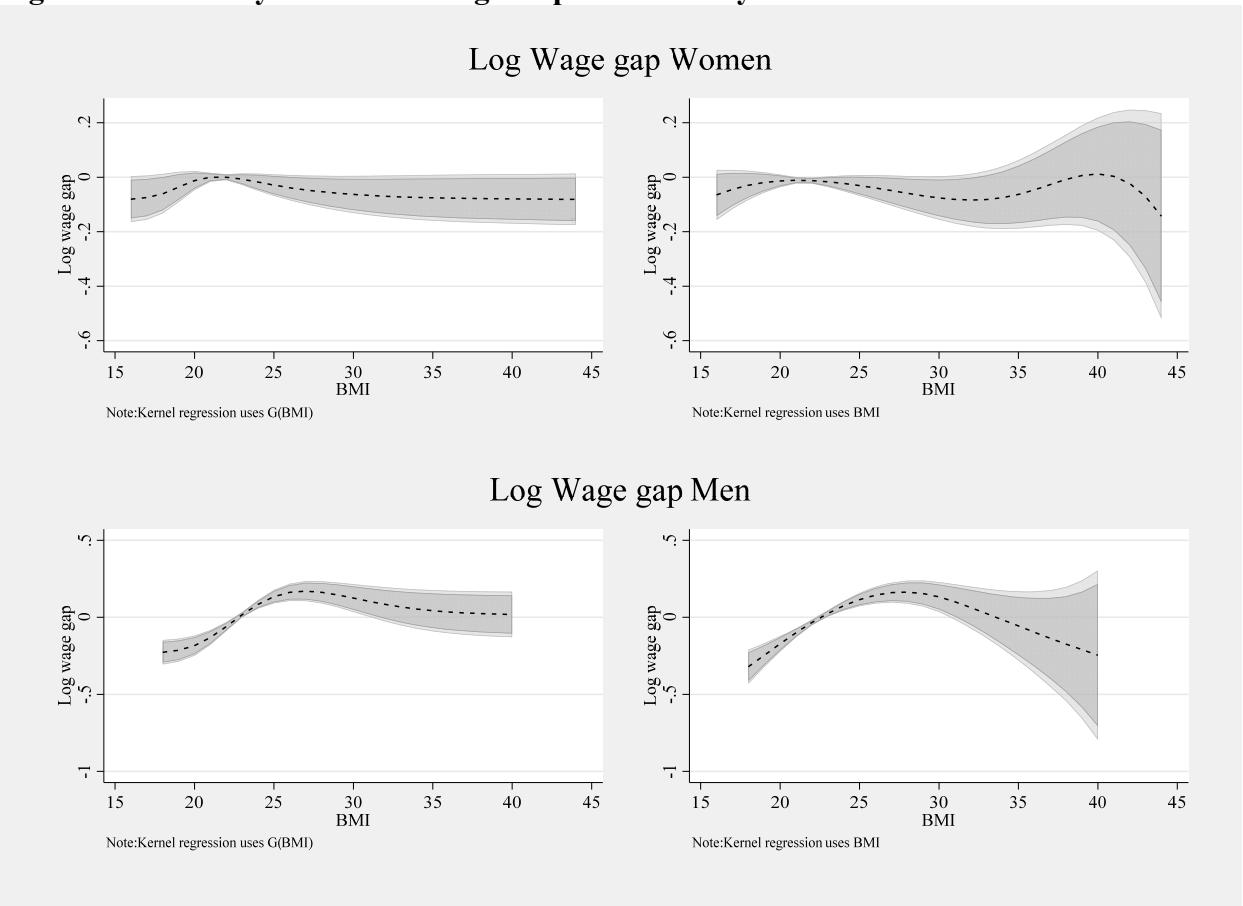
[16] In principle, this transformation should have no effect on the estimation of the semiparametric model. If $z = g(x)$, and $g()$ is a strictly monotone transformation, then $E(y|X = x) = E\big(y\big|g(X) = g(x)\big) = E(y|X = x)$.

earn, on average, 8 percent less than a woman with an average BMI, which is significant only at the 10 percent level. The results based on kernel regressions with the original distribution of BMI provide qualitatively similar results but with estimations with lower precision, especially when looking at women with high and low BMI scores.

In the case of men, the results suggest those with a BMI above 23 exhibit a positive and statistically significant wage gap compared to the average. The largest positive gap (17 percent) is observed for men with a BMI around 27, but this declines steadily for men with higher BMI, and turns statistically not significant for men with a BMI above 31. Men with a BMI below 22 show a negative wage gap as large as 32 percent (based on the original variable distribution). Similar to the results for women, the estimates for men at the top of the BMI distribution are less precise when using the original BMI for the semiparametric regression. Because the results using the transformed variable are more precise than the alternative, the rest of the paper will center on these estimations alone.[17]

[17] Figures in the appendix provide various robustness checks including: sensitivity to alternative GIMRs, results based on kernel regressions with original BMI distribution, and differences in the bandwidth estimation.

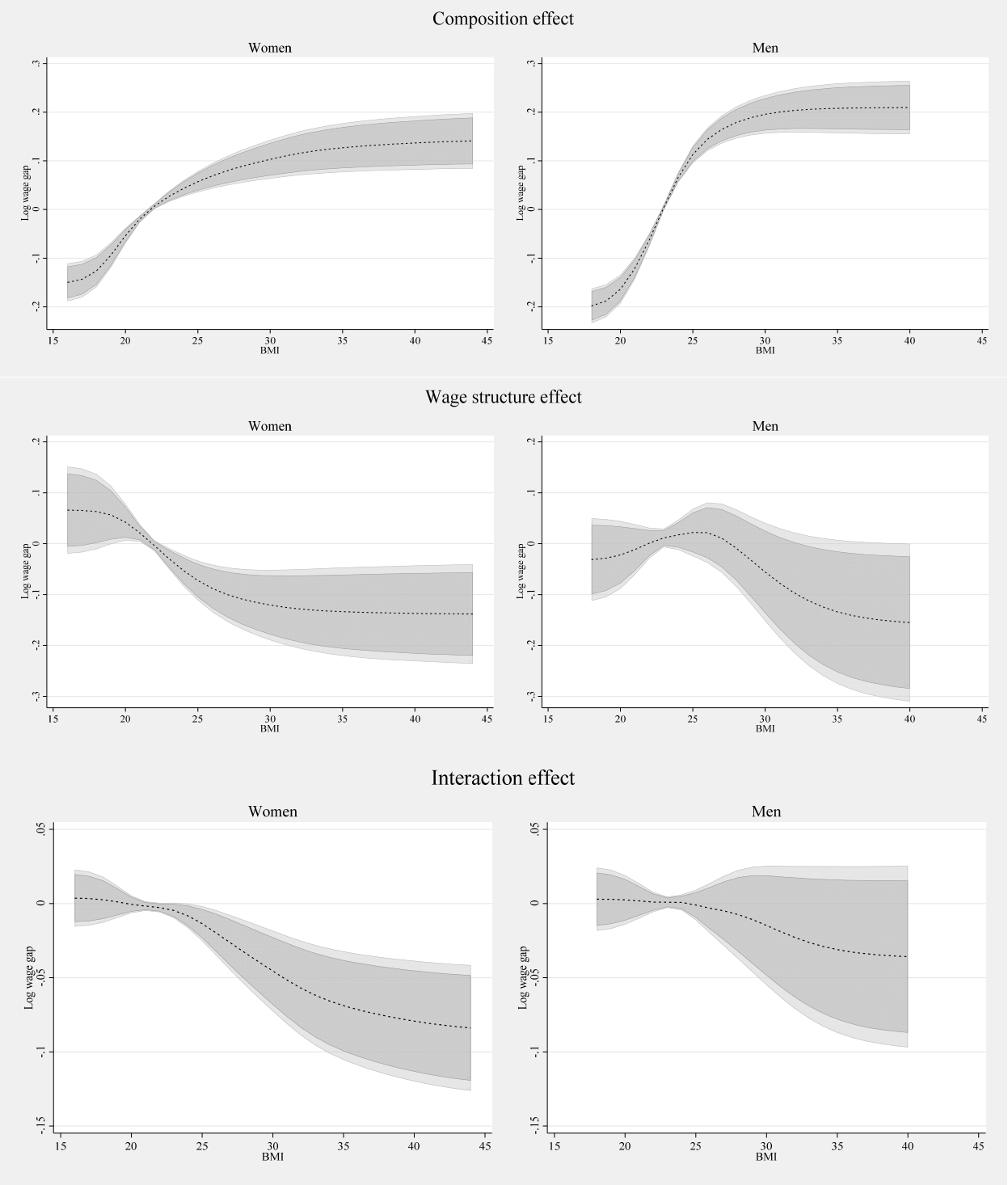
Figure 1. Selectivity Corrected Wage Gap over BMI by Gender



Note: Darker and lighter areas correspond to the 90 percent and 95 percent confidence intervals. Confidence intervals constructed based on bootstrapped standard errors with 1,000 repetitions clustered at the individual level.

Similar to the standard OB analysis, the total wage gap reported in figure 1 is not an adequate measure of the wage gap driven by differences in BMI because it is driven by differences in characteristics (composition effect), coefficients (wage structure effect), or a combination of both. In figure 2, I provide the semiparametric estimations for these three components for men and women, using the kernel regression based on the transformed data and using the OLS GIMR.

Figure 2. Aggregated Semiparametric Decomposition Components



Note: Darker and lighter areas correspond to the 90 percent and 95 percent confidence intervals. Confidence intervals constructed based on bootstrapped standard errors with 1,000 repetitions clustered at the individual level.

According to the estimations, the composition (or endowment) effect has a large and statistically significant impact when explaining the wage gaps based on BMI. Its magnitude, which is larger for men than women, shows a monotonically increasing trend with respect to BMI, but at a decreasing rate. Across the distribution of BMI, differences in characteristics explain a wage gap that ranges between -20 percent to 21 percent for men and -12.6 percent to 13.7 percent for women when looking at people with BMI of 18 and 40, and compared to people with a healthy BMI. This implies that white men and women with higher BMI have in average better endowments, which translates into higher wages.

The most important component of the decomposition is the wage structure effect. This effect can be interpreted as the treatment effect of BMI on wages after controlling for observed characteristics and endogenous selection. The first thing to notice, consistent with Cawley (2004), is that the wage structure effect for women shows a monotonically decreasing trend with respect to BMI across the whole distribution. However, the results suggest that BMI has a negative and nonlinear impact on wages. The estimations show that there is a steady decline in the wage structure component among women with a BMI between 18 to 30, with a wage gap that goes from 6.3 percent for women with a BMI of 18, to -12 percent for women with a BMI of 30. In comparison, only marginal changes in the wage gaps are observed above and below these thresholds.

For men, the effect of BMI on wages shows a different pattern. On the one hand, the results are less precise and no statistically significant differences across BMI levels are observed. Setting aside the low precision of the estimates, the wage structure effect for men shows an inverse-u shape with respect to BMI. Compared to men with a BMI of 25, for whom a point estimate of 2.2 percent wage gap is estimated, the wage premium declines at lower and higher ends of the BMI distribution. Men at the top of the BMI distribution are estimated to have a wage gap of -15 percent, while men at the bottom face a wage gap of -3 percent. This may explain why the instrumental variable estimates for men (see table 1 and 2) are negative but not statistically significant.

The last component of the decomposition is the interaction effect, which accounts for the fact that average wages are different because both coefficients and characteristics differ across groups. For men and women, the interaction effect grows negative with a higher BMI. In the case of men, the interaction effect is never statistically significant, whereas for women it is statistically significant at conventional levels and accounts for up to -9 percentage points of the total wage gap for women with high BMI.

5.4. Revisiting Cawley (2004): Partial Effect of BMI on Wages

One of the conclusions in Cawley (2004) is that a one standard-deviation increase in body weight (roughly 32 pounds), or equivalently a 5.5 point increase in BMI, is associated with a 9 percent drop in wages.[18] This is a linear extrapolation of the estimates of Cawley's preferred model, which suggests that a one-point increase in BMI is associated with a wage reduction of 1.7 percent.

While the results provided above cannot be directly compared to these findings, the delta method can be used to obtain partial effects that can be directly compared to Cawley's results. Figure 3 provides the estimations of the change of the wage structure effect as a function of BMI, and compares them to the effect based on the instrumental variable approach.[19]

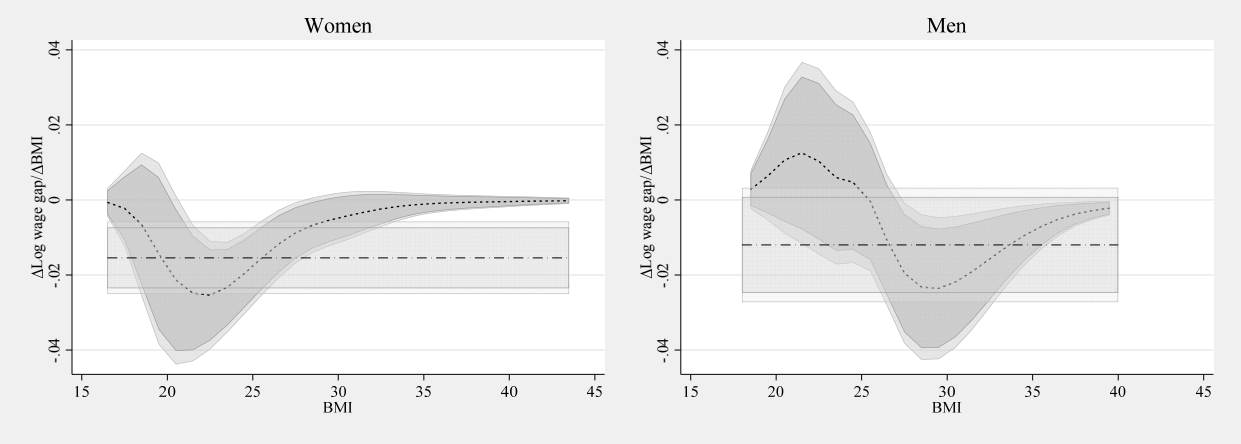
As described in table 2, the instrumental variable estimations suggested that BMI has a negative impact on wages, where a one-point increase in BMI is associated with 1.5 percent lower wages for women and 1.2 percent lower wages (not statistically significant) for men. Looking at the partial effects estimated with the semiparametric OB decomposition (Figure 3) suggests that the effect is negative, nonlinear, and statistically significant for men and women.

The marginal effect of BMI on the wage structure for women with a BMI between 20 to 25 is larger than that based on the linear instrumental variable estimate. The largest estimated partial effect indicates that an increase in BMI of one point for a woman with a starting BMI score of

[18] Cawley (2004, 465) stated that a two standard-deviation change in weight is associated with a 9 percent change in wages, when in fact this estimate reflects the impact of a one standard-deviation change in weight.
[19] For internal consistency, the instrumental variable estimations include the quadratic terms and interactions as instruments.

22.5 relates to a wage decline of 2.5 percent—an almost 65 percent greater effect than the instrumental variable estimate of 1.5 percent. The negative impact of a higher BMI is not statistically significant for women with a BMI below 20 or above 29, and the impact is below 0.5 percent for women with a BMI below 18 or above 30. Men with a BMI below 25 seem to enjoy a small positive wage gain associated with increasing BMI, although it is not statistically significant. The wage penalty due to a higher BMI is statistically significant above 27, with the largest wage decline measured at 2.3 percent (at a BMI of 29.5), almost twice as large as the instrumental variable estimates. While the partial effect on wages decrease as BMI increases, it remains statistically significant through the rest of the BMI distribution.

Figure 3. Partial Effect of BMI on the Wage Structure Effect



Note: Darker and lighter areas correspond to the 90 percent and 95 percent confidence intervals. Confidence intervals are constructed using the delta method and are based on bootstrapped standard errors with 1,000 repetitions clustered at the individual level.

6. CONCLUSIONS

In this paper, I have presented a methodology for the implementation of OB decomposition when the grouping variable is continuous in the presence of potential endogenous selection into groups. This methodology uses a semiparametric approach, known as varying coefficient models (Hastie and Tabshiran 1993), which has the advantage of providing a more flexible specification on the parameterization of the coefficients. The use of the GIMR, also known as generalized residuals, allows for a feasible strategy to control for endogenous selection based on the

continuous grouping variable. This methodology may prove useful for the analysis of endogenous treatment effects with varying treatment intensity, especially when heterogeneous effects are present.

In the application example, I revise the results from Cawley (2004) to evaluate the causal effect of BMI on wages. Using BMI as the endogenous but continuous grouping variable, I apply the proposed methodology, using siblings' BMI, age, and sex and their interactions as instruments for the estimation of the selection correction terms (GIMR) that should correct for the endogeneity of body weight and BMI. Similar to Cawley (2004), the application of the strategy does not account for possible self-selection into the labor force driven by body weight.

The application of the semiparametric OB decomposition shows that the association between BMI and wages is nonlinear, and that the negative impact of BMI on wages may be larger for women than that described in Cawley's (2004) original paper for women with a healthy BMI, but much smaller for women at the top and bottom of the BMI distribution. Furthermore, it showed that for men, BMI also has a statistically significant and negative association with wages, which was not captured previously because of the weak but positive impact that BMI has on wages for men with a low BMI.

REFERENCES

Averett, S. 2011. "Labor market consequences: employment, wages, disability, and absenteeism." In J. Cawley (ed.), *The Oxford Handbook of the Social Science of Obesity.* New York: Oxford University Press.

Blinder, A. S. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *The Journal of Human Resources* 8(4): 436–55.

Cain, G. G. 1986. "The Economic Analysis of Labor Market Discrimination: A Survey." In O. Ashenfelter and R. Laynard (eds.), *Handbook of Labor Economics.* Amsterdam: Elsevier Science Publishers.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* New York: Cambridge University Press.

Cattaneo, M., and M. Jansson. 2018. "Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency." *Econometrica* 86(3): 955–95.

Cawley, J. 2004. "The Impact of Obesity on Wages." *The Journal of Human Resources* 39(2): 451–74.

Chernozhukov, V., I. Fernández-Val, and B. Melly. 2013. "Inference on Counterfactual Distributions." *Econometrica* 81(6): 2205–68.

Fikkan, J., and E. Rothblum. 2012. "Is fat a feminist issue? Exploring the gendered nature of weight bias." *Sex Roles* 66(2012): 575–92.

Foresi, S., and F. Peracchi. 1995. "The Conditional Distribution of Excess Returns: an Empirical Analysis." *Journal of the American Statistical Association* 90(430): 451–66.

Fortin, N., T. Lemieux, and S. Firpo. 2011. "Decomposition Methods in Economics." In A. Orley and C. David (eds.), *Handbook of Labor Economics*. Amsterdam: Elsevier Science Publishers.

Hastie, T., and R. Tibshirani. 1993. "Varying-Coefficient Models." *Journal of the Royal Statistical Society* Series B (Methodological) 55(4): 757–96.

Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(l): 53–161.

Hotchkiss, J. L., and M. M. Pitts, 2013. "Even One Is Too Much: The Economic Consequences of Being a Smoker." Federal Reserve Bank of Atlanta Working Paper Series, WP 2013-3. Atlanta: Federal Reserve Bank of Atlanta.

Hoover, D. R., J. A. Rice, C. O. Wu, and L. P. Yang. 1998. “Nonparametric smoothing estimates of time-varying coefficient models in longitudinal data.” *Biometrika* 85(4): 809–22.

Keele, L. J. 2008. *Semiparametric regression for the social sciences*. New York: John Wiley & Sons.

Kluve, J., H. Schneider, A. Uhlendorff, and Z. Zhao. 2012. “Evaluating continuous training programmes by using the generalized propensity score.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175: 587–617.

Lee, L. F. 1978. “Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables.” *International Economic Review* 19(2): 415–33.

Li, Q., and J. S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton, NJ: Princeton University Press.

Lokshin, M., and Z. Sajaia. 2004. “Maximum Likelihood Estimation of Endogenous Switching Regression Models.” *Stata Journal* 4(3): 282–89.

Nadaraya, E. A. 1964. “On Estimating Regression.” *Theory of Probability and its Applications* 9(1): 141–42.

Ñopo, H. 2008. “An extension of the Blinder–Oaxaca decomposition to a continuum of comparison groups.” *Economics Letters* 100(2): 292–96.

Oaxaca, R. 1973. “Male-Female Wage Differentials in Urban Labor Markets.” *International Economic Review* 14(3): 693–709.

Sabia, J. J., and D. I. Rees. 2012. “Body Weight and Wages: Evidence from Add Health.” *Economics & Human Biology* 10(1):14–19.

Terza, J. 1985. “Ordered Probit: A Generalization.” *Communications in Statistics – A. Theory and Methods* 14(1): 1–11.

Ulrick, S. W. 2012. “The Oaxaca decomposition generalized to a continuous group variable.” *Economics Letters* 115(1): 35–37.

Vella, F. 1998. “Estimating Models with Sample Selection Bias: A Survey.” *The Journal of Human Resources* 33(1): 127–69.

Watson, G. S. 1964. “Smooth regression analysis.” *Sankhyā: The Indian Journal of Statistics* Series A. 26(4): 359–72.

Williams, R. 2016. “Understanding and interpreting generalized ordered logit models.” *The Journal of Mathematical Sociology* 40(1): 7–20.

Wooldridge, J. M. 2015. “Control Function Methods in Applied Econometrics.” *Journal of Human Resources* 50(2): 420–55.

Yatchew, A. 2003. *Semiparametric regression for the applied econometrician*. Cambridge, UK: Cambridge University Press.

APPENDIX. SENSITIVITY TO MODEL SPECIFICATIONS AND BANDWIDTH

Table A1. Replication of Cawley (2004) with Model and Sample Modifications

Replication of Cawley (2004)						
	White		Black		Hispanic	
	Male	Female	Male	Female	Male	Female
BMI	-0.0131	-0.0168***	-0.00258	-0.00191	-0.00914	-0.0124
	[0.00831]	[0.00496]	[0.00678]	[0.00600]	[0.00731]	[0.0125]
N	13355	10800	6811	5651	4374	3035
Pooling Black and Hispanic						
	White		Nonwhite			
	Male	Female	Male	Female		
BMI	-0.0131	-0.0168***	-0.00369	-0.00515		
	[0.00831]	[0.00496]	[0.00508]	[0.00544]		
N	13355	10800	11185	8686		
Excluding Sample Weights						
	White		Nonwhite			
	Male	Female	Male	Female		
BMI	-0.0126	-0.0149***	-0.00241	-0.00643		
	[0.00789]	[0.00471]	[0.00472]	[0.00518]		
N	13355	10800	11185	8686		
Dropping if Parents' Education Is Missing						
	White		Nonwhite			
	Male	Female	Male	Female		
BMI	-0.0118	-0.0147***	-0.003	-0.00634		
	[0.00803]	[0.00480]	[0.00481]	[0.00518]		
N	12393	10195	10465	8224		
Modifying Model Specification						
	White		Nonwhite			
	Male	Female	Male	Female		
BMI	-0.0124	-0.0155***	-0.0048	-0.00673		
	[0.00805]	[0.00487]	[0.00492]	[0.00535]		
N	12191	10111	9854	7963		
Dropping Extreme BMI Values (below 16 and above 60)						
	White		Nonwhite			
	Male	Female	Male	Female		
BMI	-0.0127	-0.0154***	-0.00425	-0.00735		
	[0.00804]	[0.00493]	[0.00504]	[0.00545]		
N	12184	10101	9844	7958		

Note: * p<0.1, ** p<0.05, *** p<0.01. Clustered standard errors in parenthesis.

Figure A1. Kernel Densities of BMI across Race and Sex

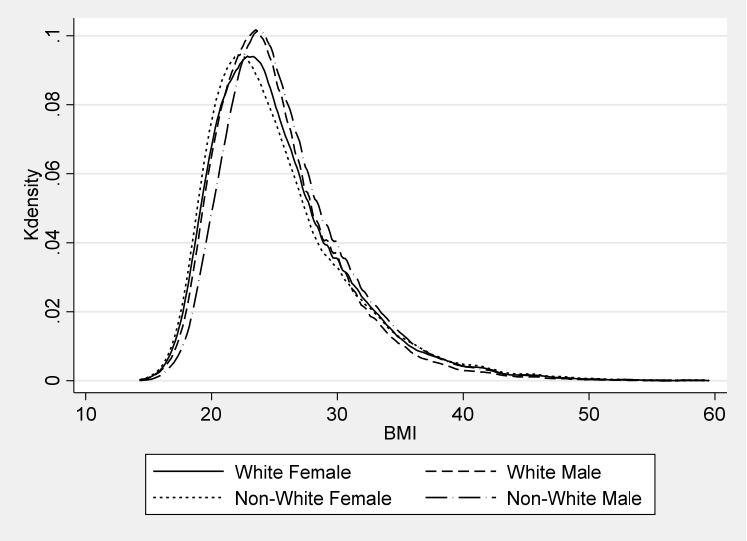


Figure A2. Selectivity Corrected Log Wage Gap by Gender and GIMR Estimation

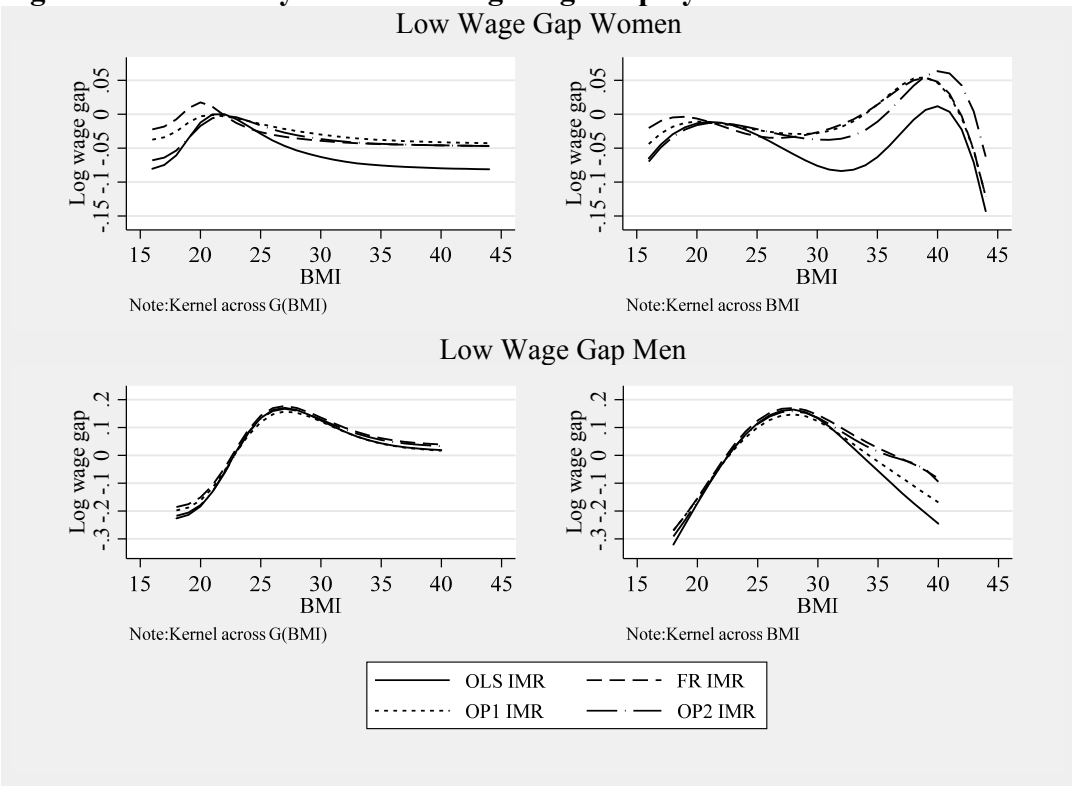
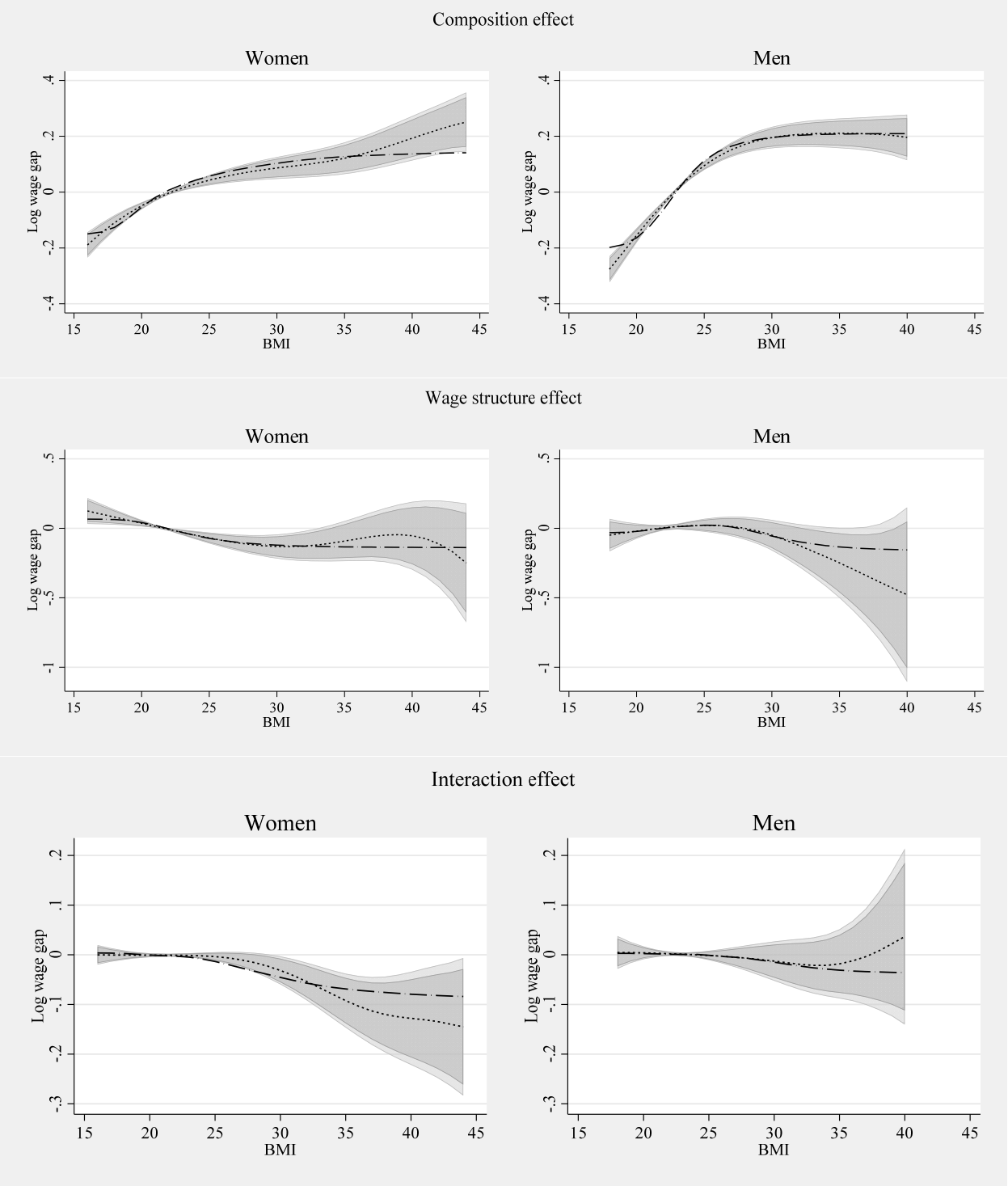


Figure A3. Aggregate Semiparametric Decomposition: OLS-GIMR with Kernel Regression Using BMI



Note: Dashed line is the estimation that uses G(BMI) in the kernel regression.

Figure A4. Aggregate Semiparametric Decomposition with Kernel Regression Using G(BMI) Sensitivity to GIMR Estimation Method

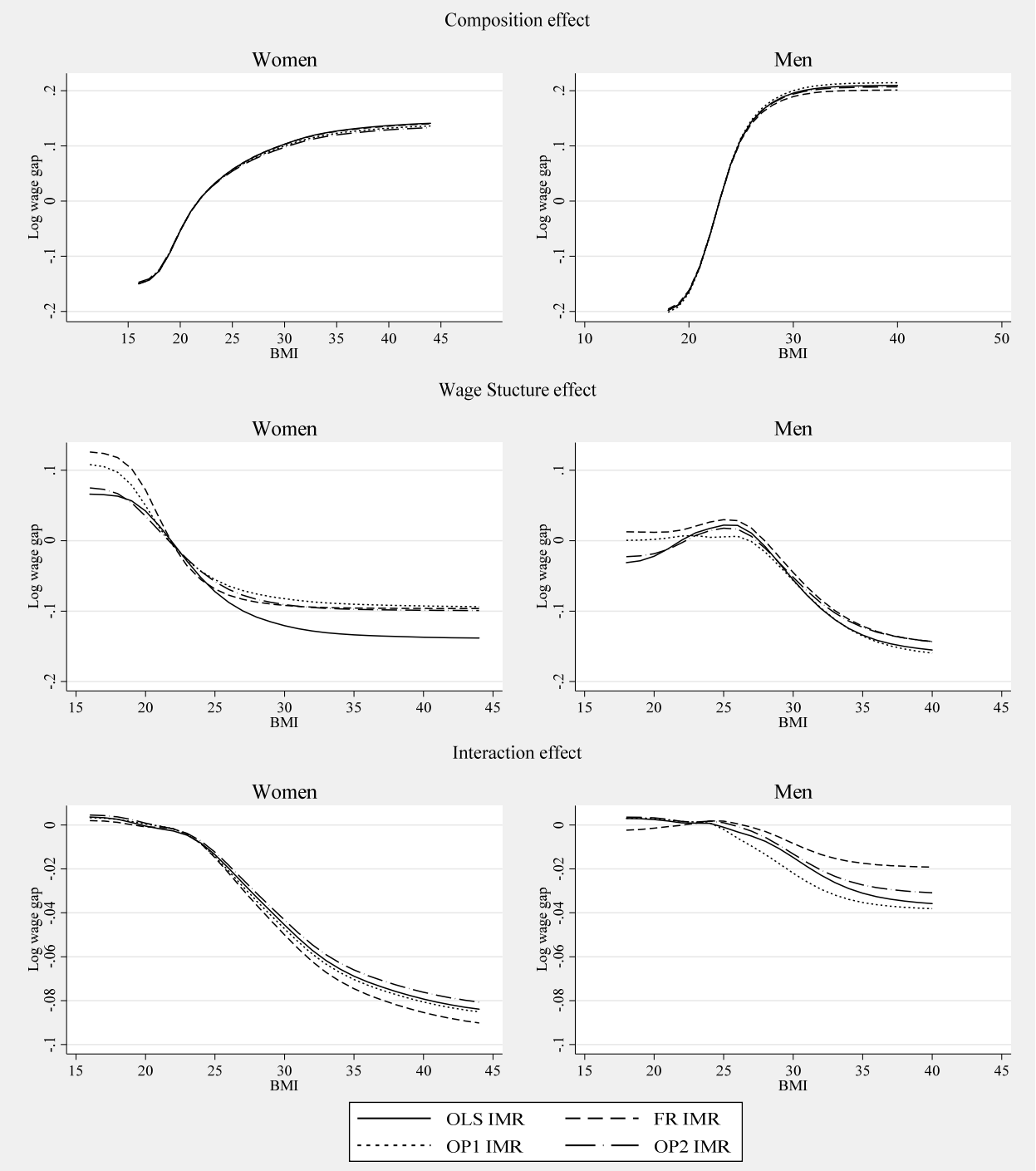


Figure A5. Aggregate Semiparametric Decomposition: Sensitivity to Bandwidth (OLS-GIMR)