Economics Bulletin

Volume 33, Issue 3

Combining information from Heckman and matching estimators: testing and controlling for hidden bias

Gerry H. Makepeace
Cardiff Business School, Cardiff University and IZA, Bonn

Michael J. Peel
Cardiff Business School, Cardiff University

Abstract

We demonstrate how the Heckman methodology can be applied to the Rosenbaum sensitivity model and the Rubin matched difference estimator. We develop a statistical test of the conditional independence assumption (CIA), based on Heckit for matched pairs. If the CIA is rejected, the method facilitates the estimation of matched treatment effects adjusted for hidden bias. We illustrate this methodology empirically for the full-time/part-time pay gap for British women. The proposed method has clear utility in establishing whether propensity score matched treatment estimates are prone to unobserved selection bias and for controlling for such bias

We are grateful to two anonymous Referees for helpful comments. Any errors are solely those of the authors.
Citation: Gerry H. Makepeace and Michael J. Peel, (2013) "Combining information from Heckman and matching estimators: testing and controlling for hidden bias", *Economics Bulletin*, Vol. 33 No. 3 pp. 2422-2436.
Contact: Gerry H. Makepeace - makepeace@cardiff.ac.uk, Michael J. Peel - peel@cardiff.ac.uk.
Submitted: April 26, 2013.   **Published:** September 19, 2013.

1. Introduction and motivation

Propensity Score Matching (PSM) estimators (Rosenbaum and Rubin, 1983) are widely employed in observational studies to estimate treatment effects in economics (e.g., Black and Smith, 2004; Blanchflower and Bryson, 2004; Gibson, 2009; Lin and Lue, 2010). For instance, Gemici et al. (2012) comment (p. 220) that, in economic and other studies, PSM 'has become particularly prominent internationally'. Furthermore, a search of Google Scholar revealed that around 23,400 papers dated between 2002 and 2012 contained the phrase propensity score matching. Though the Rosenbaum bounds[1] (RB) method enables an assessment to be made of how large the impact of a potential (simulated) unobserved confounding variable must be to challenge PSM treatment inferences, it does not test or control for selection bias. Hence a principal limitation of PSM treatment estimates is that the conditional independence assumption (CIA) is assumed to hold; specifically, that the PSM estimate of the average treatment effect on the treated (ATT) is not subject to unobserved selection bias (the 'ignorability' assumption). The development of techniques to control for endogenous PSM treatment effects is therefore of high import and utility.

To our knowledge, this paper is the first to specify a method for testing and controlling for unobserved bias for matched ATT estimates. We augment the RB sensitivity approach (Rosenbaum, 2002; 2005; 2010), which simulates the combined impact of a potential hidden variable on selection into treatment and on the outcome variable, to formulate a Heckit control function PSM treatment model (e.g., Greene, 2006, p. 787). Following from this, we modify Rubin's (1973; 1979) matched difference specification to incorporate Heckit selection terms for treated and untreated observations. Our method provides a statistical test for the potential failure of the CIA of no hidden bias for matched treatment estimates. If the CIA is rejected, bias adjusted treatment effects can be estimated with a modification to Rubin's matched difference model. The method is applied in an examination of the British part-time (PT) women's pay gap.

It is important to stress at the outset that our specification shares the limitations of the standard Heckman model. In this paper we incorporate the Heckman parametric control function assumptions, particularly that of the joint normality of error terms. As with the Heckit model, these assumptions are relatively restrictive. In addition, although selection effects are formally identified from distributional error assumptions, it is desirable to employ

[1] See Caliendo and Kopeinig (2007) and Keele (2010) for RB specifications for binary and continuous outcomes respectively.

an additional instrumental variable[2] (which may be difficult to obtain), not least to mitigate multicollinearity concerns (Puhani, 2000).

As with other estimators which control for omitted variable bias, assumptions are more exacting than conventional methods. There is a trade-off between employing standard techniques and more exacting methods that aim to control for such bias. Given this, in Section 4 we outline potential extensions to our modelling approach. In particular, more flexible semiparametric and nonparametric estimators have been developed in the econometrics literature. To date, however, control function methods have focused on endogenously determined continuous variables (Wooldridge, 2012). For the sample selection problem, Huber (2013) develops a semiparametric IV-type estimator which employs inverse propensity scores (selection probabilities) to control for bias. Klein and Vella (2009) formulate an IV estimator to control for endogenous binary response variables for continuous outcomes, employing a semi-parametric selection model to estimate first-step fitted values. As described in Section 4, these more flexible methods have the potential to estimate control functions as applied in our modelling approach (Wooldridge, 2012, pp. 22-28).

Although we demonstrate that our specification is consistent with PSM, RB and the Heckman treatment effect model - and our empirical results support expectations - our findings should be viewed as provisional. As discussed in Section 4, further simulation research is warranted to assess the properties of the proposed estimator, including robustness to misspecification, not least to the violation of the joint normality assumption. However, in defence of our parametric specification, Heckman and Vytlacil (2007), stress (p. 4783) that there has been slow progress in applying semiparametric and nonparametric methods to empirical problems and that they appear sensitive to parameter choices (including smoothing and trimming ones). They state that:

'Most of this literature is based on Monte Carlo analysis or worst case analyses on artificial samples. The empirical evidence on nonrobustness of conventional parametric models is

[2] Where an instrumental variable is available, Imbens and Angrist (1994) have developed an IV method which identifies the local average treatment effect (LATE); which is (p. 467) 'the average treatment effect for individuals whose treatment status is influenced by changing an exogenous regressor that satisfies an exclusion restriction.' Based on instrumental variables, recent working papers have proposed tests for whether the CIA is valid. Given a binary instrument, and employing inverse probability weighted estimators, Donald et al. (2011) propose a Durbin-Wu-Hausman formulation to test for unconfoundness conditional on observable variables. De Luna and Johansson (2012) specify a non-parametric test of the CIA employing a quasi-instrument; that is (p. 2), an instrument which is permitted 'to be confounded by unobserved variables'. Based on the Heckman approach, our methodology provides both a test of whether the CIA holds for matched treatment effects and a means of controlling for omitted variable bias.

mixed … The evidence in Heckman and Sedlacek (1985) and Blundell, Reed and Stoker (2003) shows that normality is an accurate approximation to log earnings data in economic models of self-selection. The analysis of Todd (1996) shows that parametric probit analysis is accurate for even extreme departures from normality.'

In our empirical application, we employ both the log of earnings and a probit selection specification. Specifically we examine the PT women's pay gap (penalty) relative to their female full-time (FT) counterparts. The pay gap between men and women has been declining steadily over time (e.g., Pike, 2011; Joshi et al., 2007). This contrasts with the PT/FT pay gap for women. For example, Connolly and Gregory (2009) report unadjusted gaps of 15% in 1975, 21% in 1985 and 29% in 2001; while Pike (2011) demonstrates that the differential scarcely changes over the period to 1997 to 2010. As well having import *per se*, these penalties have a substantial influence on overall gender differentials because the incidence of part-time work is much higher for women than for men. About 42-44% of women were working part-time in our data and a substantial number more will work part-time at some point in their career (Connelly and Gregory, 2009).

The full PT/FT women's pay gap is therefore an important issue, not least for policy purposes, with a number of recent studies having identified a strong and robust unexplained part-time penalty (e.g., Connolly and Gregory, 2009; Manning and Petrongolo, 2008; Mumford and Smith, 2009; Neuberger et al., 2011). Our empirical application builds on the research of Manning and Petrongolo (2008) who report that the pay gap reduces substantially when occupation controls are included in their model. We examine selection effects for models with and without occupation controls.

2. Econometric model

2.1 Rosenbaum bounds methodology

As discussed below, conceptually our methodology follows that of RB. An outcome variable (Y) varies between the treated group (D=1) and the untreated group (D=0). Let Ω be the set of treated and (pair) matched untreated observations in the common support. If Z is a vector of control variables, the propensity score is $p(Z_i)=\Pr(D=1|Z)$. The PSM estimated ATT is defined as the difference in the sample means of the treated (Y_1) and untreated observations (Y_0) in Ω:

$$ATT=\left[\frac{1}{N_\Omega}\sum_{i\in\Omega}Y_1\right]-\left[\frac{1}{N_\Omega}\sum_{i\in\Omega}Y_0\right]=\overline{Y_{1\Omega}}-\overline{Y_{0\Omega}} \qquad (1)$$

where the subscript Ω indicates that the mean is for observations in the matched sample.

The RB method (see Rosenbaum, 2002; DiPrete and Gangl, 2004, for full derivations) provides an estimate of the impact a potential (simulated) unobserved confounding variable must exert to render ATT PSM treatment effects statistically insignificant via its dual effect on selection and outcome (a control function approach). If Z is a vector of control variables determining the probability of treatment ($\Pr(D=1|Z)$), the ratio of the odds ratios for two individuals, j and k with $Z_j=Z_k$, are specified as:

$$\frac{1}{\Gamma} \leq \frac{Odds(Z_j)}{Odds(Z_k)} \leq \Gamma \qquad \text{for } j \neq k,\ Z_j = Z_k \qquad (2)$$

This expression places bounds on the odds ratios. Γ is equivalent to the odds ratio associated with the coefficient of an unobserved variable in a logit selection model into treatment. If $\Gamma=1$, subjects with the same attributes have the same probabilities and odds of selection into the treatment group (the CIA is assumed to hold) and matched treatment effects are bias free.

Since Z contains all the observable information, deviations from $\Gamma=1$ are attributed to unobserved confounding variables. For example, $\Gamma=2$ indicates an unobserved covariate would double the odds of selection into treatment. Rosenbaum (2002; 2005) derives bounds on the confidence intervals for matched ATT estimates as Γ varies, thus defining a critical value of Γ at which the ATT is statistically insignificant.

In the next Section we extend the PSM approach by formulating a Heckit treatment version of the RB model. Our methodology provides a specific test of the CIA for matched treatment estimates together with bias adjusted treatment effects when the CIA is rejected. The two approaches are very similar conceptually and consistent with each other. However, whereas the RB technique simulates the impact of a potential unobserved confounding variable (via its dual effect in a logit selection model into treatment and on the outcome variable), our method employs inverse Mills ratios (IMRs) from a probit selection model into treatment to correct (or test) for hidden bias for PSM treatment estimates.

In this respect, the IMRs can be considered as a proxy for actual unobservables[3]. Whilst the RB technique simulates the impact of a potential confounding (omitted) variable, it does not test or control for omitted variable bias.

[3] In accord with the Heckman treatment effect model (and the standard assumptions as specified in the note), the IMRs (selection model errors) are surrogates for one or more unobservable (confounding) variables. As stated, conceptually, this is similar to the RB simulation method. Note also that, though the employment of an instrument is preferred, identification on distributional grounds is obtained in the absence of instruments in the presence of one or more confounders.

For instance the RB methodology may suggest that a large selection effect is required to negate a PSM treatment estimate; whereas our method for the same data might indicate unobserved bias is a significant factor (the CIA is rejected). Such findings (or other variations) are not inconsistent.

2.2 Heckit estimation of the Rosenbaum model

As described below, the premise of our method is that we can estimate selection terms from the full sample, match on covariates in the full sample and then employ the selection terms for matched pairs when estimating the ATT. Importantly, this is consistent with the standard matching framework (Rubin 1973, 1979; Ho et al, 2007; Rubin and Thomas, 2000).

The Heckit model for an endogenous treatment variable (Greene, 2006) applicable to the Rosenbaum additive treatment model is:

Outcome equation $$Y_i = \delta D_i + \mu(X_i) + \varepsilon_{1i} \qquad i=1, .. ,N \tag{3}$$

Selection equation $$D_i^* = Z_i\theta + \varepsilon_{2i} \qquad i=1, .. ,N \tag{4}$$

D=1 when $D^*>0$, D=0 when $D^*\leq 0$

$$\begin{bmatrix}\varepsilon_1 \\ \varepsilon_2\end{bmatrix} \sim N.I.D.\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & 1\end{pmatrix}\right] \tag{5}$$

where X and Z are vectors of regressors.

This specification gives the following two step model.

$$Y_i = \delta D_i + \mu(X_i) + \sigma_{12}\lambda_i + \upsilon_i \qquad i=1, .. ,N \tag{6}$$

$$\lambda_i = \lambda_{1i} = \frac{\phi(Z_i\theta)}{\Phi(Z_i\theta)} \text{ if } D_i=1 \text{ and } \lambda_{0i} = -\frac{\phi(Z_i\theta)}{1-\Phi(Z_i\theta)} \text{ if } D_i=0 \tag{7}$$

where ϕ and Φ are the normal density and distribution functions.

With exact matching, where $X_{k1}=X_{k0}$ and $Z_{k1}=Z_{k0}$, the standard two-step specification in (6) and (7) for matched pairs can be reformulated as:

$$\Delta Y = Y_1 - Y_0 = \delta + \sigma_{12}\lambda_\Delta + \upsilon \tag{8}$$

$$\lambda_\Delta = \lambda_1 - \lambda_0 = \frac{\phi(\sum\theta_m Z_m)}{\Phi(\sum\theta_m Z_m)} + \frac{\phi(\sum\theta_m Z_m)}{1-\Phi(\sum\theta_m Z_m)} = \frac{\phi(\sum\theta_m Z_m)}{\Phi(\sum\theta_m Z_m)[1-\Phi(\sum\theta_m Z_m)]} \tag{9}$$

Equation 8 can be estimated as a simple regression model. However, due to the dimensionality problem, exact covariate matching would be highly unusual in economic and other studies; with PSM employed to circumvent this problem. Furthermore, due to imperfect matching with PSM, it is accepted practice to implement parametric adjustments to PSM treatment effects by including the covariates in standard regression models in the matched samples to estimate the ATT, though to date without controlling for bias via $\sigma_{S1}\lambda_{\Delta}$ (see Ho et al., 2007 for a detailed discussion).

Importantly, Rubin (1973; 1979) demonstrates that, where a linear approximation to $\mu(X_i)$ is used, adjustments of δ derived from paired covariate difference regression models for matched samples exhibit less observed bias than standard matched estimates; or those for OLS models applied to matched (but not differenced) samples. More recently, employing PSM and applying covariate difference linear regression adjustments to control for remaining bias in the matched samples, Rubin and Thomas (2000) confirm Rubin's earlier (1973; 1979) research. They conclude (p. 581) that 'the superior performance of regression adjustments applied to matched samples is consistent with previous theoretical and simulation studies'.

Applying Rubin's methodology, the difference in outcomes for PSM matched pairs is:

$$\Delta Y_m = Y_{1m} - Y_{0m} = \delta + (X_{1m} - X_{0m})\beta + \sigma_{12}(\lambda_{1m} - \lambda_{0m}) + \upsilon_m \qquad m \in \Omega \tag{10}$$

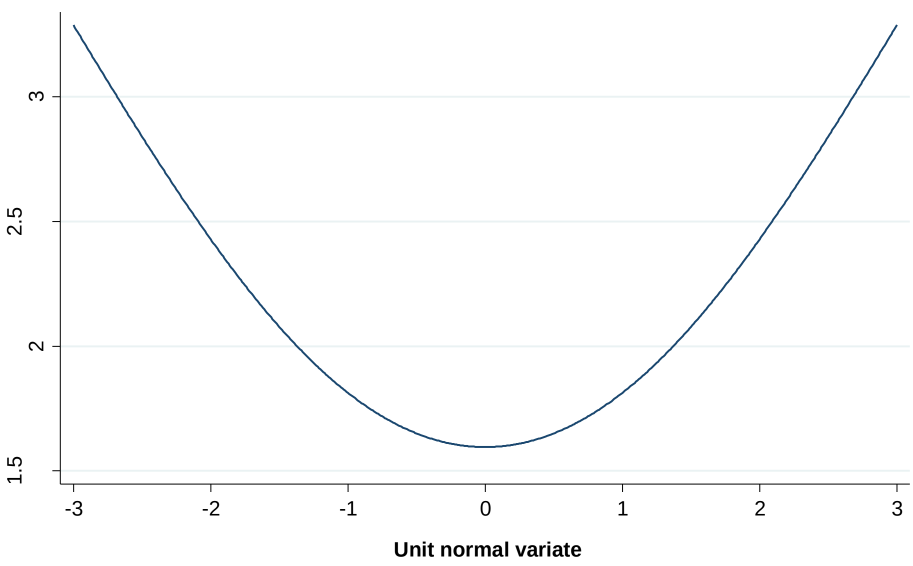
where the λs are the inverse Mills ratios (IMR) in (4), obtained from a probit model for the unmatched sample but applied to the matched sample. Hence, our proposed method is to add the term $\sigma_{12}(\lambda_{1m}-\lambda_{0m})$ to the Rubin matched difference model, producing a sequence of nested models. If σ_{12}=0 (the CIA holds), the model reduces to a PSM model incorporating differences in X to adjust for possible matching imperfection as per Rubin (1973; 1979). If β=0 and σ_{12}=0, then the regression gives the familiar (standard) PSM estimate.

Heckit estimation often suffers from multicollinearity between λ and the covariates. As well as the advantage of estimating treatment effects in the common support, our proposed matched form offers a further potential benefit. With perfect covariate matching, then:

$$\Delta\lambda_m = \frac{\phi(Z_m\theta)}{\Phi(Z_m\theta)[1-\Phi(Z_m\theta)]} \tag{11}$$

Figure 1 shows that $\Delta\lambda_m$ has a U-shaped function with a minimum at zero. The selection term differs markedly from its traditional shape even if $Z_{1m} \cong Z_{0m}$. This may result in more robust identification than the standard model in the absence of an instrument; and is worthy of further investigation in future research (Section 4). Another practical benefit is that the regressors are less likely to be correlated with the selection term (see below) when the equation is estimated in differences rather than levels.

Figure 1
Difference in IMR with unit normal variate for perfect matching



3. Full and part-time pay differential

In this section we use our augmented PSM model[4] to provide estimates of the pay penalty associated with PT work. These estimates are analogous to the unexplained component in the Oaxaca-Blinder decomposition (e.g., Oostendorp, 2009). We extend the recent research of Manning and Petrongolo (M&P, 2008). Employing LFS data for start of period 2001-2003 and conventional regression methods, they report a pay gap of about 10%. Importantly, M&P highlight the pivotal role of occupation, demonstrating that the pay gap is reduced substantially to about 3% after controlling for occupation differences. Using updated LFS data and our new proposed methodology, we investigate the bias associated with occupation, together with any remaining selection effect. Note that, as with M&P (2008), occupation dummies are assumed to be exogenous[5].

[4] Note that, as in the standard framework of matching and Heckman treatment effect models, our model controls for omitted variable bias rather than bias from any heterogeneity (e.g., as per the Roy model).
[5] Though the occupation controls were found to be key regressors by M&P (2008), as stressed by a Reviewer, they are themselves potentially endogenous; with fields of work or occupational tasks being cognisant (e.g., Firpo et al., 2011). Although PSM is now employed in a large variety of treatment applications, as noted by the Reviewer, it is rooted in programme evaluation. The current empirical study was selected with reference to M&P's important finding regarding occupation. Specifically, it facilitates the illustration of our methodology empirically with reference to an important economic issue where there is an ongoing debate about whether a variable should be included or omitted.

Our models are estimated for log of weekly earnings (*Y*), with *D* (FT) = 1 and *D* (PT) = 0, based on an original sample of 207,970 observations for 2001-2010. Explanatory variables are similar (including instruments) to those employed by M&P. As discussed above, although formal identification is achieved via distributional assumptions, we use the same instruments as M&P (see Appendix). Following M&P (2008), models are estimated for two groups of covariates: *base* and *broad*, with the latter including *base* plus 9 broadly defined occupations. Observations are matched by probit using the nearest neighbour method without replacement and a fine calliper (0.0001).

The standard OLS estimates of the pay gap with a dummy variable for FT in our data are similar to those reported by M&P and endorse the key role of occupation, with the premium reducing markedly from the *base* estimate of 13.3% to 4.6% for the *broad* one. Table 1 summarises our matching results for the FT wage premium. In the current example, the standard PSM estimates (M1) are similar to the OLS ones at 11.9% and 4.0% respectively. Also of note, for this data, for both *broad* and *basic*, the standard matched treatment estimates (M1) are similar to the ones with differenced regressors (M2), at 12.4% and 4.6%. Of course, this degree of congruence may not hold for other studies.

Table 1
Difference in log hourly earnings: regression estimates for matched pairs

	M1	M2	M3
Controls	None	ΔX	ΔX & $\Delta\lambda$
Base			
Pay gap	0.119**	0.124**	0.062**
Selection term	-	-	0.037*
(p-value)			(0.038)
Γ	1.517		
Broad			
Pay gap	0.040**	0.046**	0.073**
Selection term	-	-	-0.016
(p-value)			(0.304)
Γ	1.162		

The reported results are for nearest neighbour matching without replacement (calliper = 0.0001). Number of matched pairs is 58,866 (base) and 53,812 (broad).
Γ Indicates Rosenbaum bounds critical value of the odds ratio required for unobserved variables to nullify the significance of the pay gap (p=0.05).
∗∗, ∗ Indicates significance at the 1% and 5% levels.

The RB critical Γ parameters for basic (broad) of 1.517 (1.162) indicate that an unobserved variable would have to increase the odds of selection into FT by 52% (16%) to render the pay gap statistically insignificant at p=0.05 for the standard PSM estimates (M1). The decrease in odds associated with the bounds critical parameters correctly reflects the fact that the former (1.517) incorporates the impact of occupation as a confounder, whose omission inflates the treatment effect, so that a larger confounding affect is required to negate the premium. To the unwary, therefore, the large RB critical parameter for the base specification may induce overconfidence in the robustness of the treatment estimate to hidden bias. By the same token, the critical bounds (Γ=1.162) for the broad specification indicates its relative sensitivity to a confounding variable.

However, as M3 (for ΔX and $\Delta\lambda$) reveals, our test for hidden bias suggests that the CIA is rejected for the base specification ($\Delta\lambda$, p=0.038), but that it holds (p= 0.304) for the broad one which controls for occupation differences. In either case, the test is shown to have utility; though, of course, as described above, the significance of $\Delta\lambda$ does not depend on either the magnitude of (simulated) Γ or that of the treatment estimate.

As noted above, in the standard Heckit framework, considerable scepticism has been voiced about the application of selection corrections due to potential multicollinearity between the selection terms and the remaining regressors (e.g., Nawata, 1993). In the current case, the R^2s for regressions of the selection term (λ_{1m}-λ_{0m}) on the remaining regressors are 0.007 (0.008) for basic (broad), indicating that there is no collinearity issue regarding the identification of the selection terms.[6] As described above (Figure 1), this finding is worthy of further investigation.

Importantly, the point estimates for the base specification with the selection term (M3) indicate that a large part of the pay differential can be explained by a selection effect and that the adjusted pay gap estimate (6.2%) is relatively close to the standard (M1) and the adjusted (M2, ΔX) matched pay gap estimates (4.0%, 4.6%) for broad when occupation is included. Although not perfectly capturing the difference in pay due to the confounding effect of occupation in the standard PSM estimates, our differenced model for the base specification which controls for bias (M3) provides a more consistent estimate (6.2%) than the

[6] The results are not sensitive to the use of instruments. The selection estimates are -0.011 (0.044) with p-values of 0.531 (0.037) for M3 when instruments are excluded for the *broad* (*basic*) specifications respectively. The R^2s for the selection terms (λ_{1m}-λ_{0m}) on the remaining regressors increase to 0.234 (0.289), but are still far below the normal point where multicollinearity would be regarded as a problem.

conventional PSM one (11.9%, M1), when compared to the models with occupational controls.

When occupation is included in the model, the results for *broad* reveal that there is no statistically significant selection effect; though the pay gap does increase to 7.3% (M3). However, as with standard Heckit estimates, the statistical insignificance of the selection term (λ_{1m}-λ_{0m}), suggests that the specification without the selection term for broad (M2) is appropriate (note also the absence of multicollinearity as described above). Therefore, our preferred estimate is 4.6% (from M2) because the differenced covariate controls are jointly significant and the CIA is not rejected for this specification.

4. Conclusion, limitations and further research

Rosenbaum and Rubin have developed sensitivity analysis for confounding variables and matched paired difference regression specifications respectively, with the aim of improving the utility and efficacy of PSM causal treatment inferences. Employing standard matching and Heckit assumptions, in this paper we formulate a statistical test for the CIA of no hidden bias (ignorability) for PSM matched treatment estimates. Drawing on the rationale underpinning Rosenbaum bounds, we propose a modification to Rubin's matched difference estimator which nests Heckit model treatment selection terms for propensity score pair-matched treated and untreated observations, providing a statistical test of whether the CIA holds. If not, the differenced Heckit treatment selection term (λ_{1m}-λ_{0m}) can be included in Rubin's matched difference estimator to control for unobserved bias.

We demonstrated the potential utility of the methodology via estimates of the part-time pay gap for British women, an important economic and policy issue. Although not perfectly capturing the difference in the pay penalty attributable to occupation differences, our test did indicate that the CIA was rejected in the specification which omitted occupation as a confounder. When gauged against the model with occupation controls, our matched differenced model which controlled (adjusted) for this bias gave a more consistent estimate of the pay gap than the comparable unadjusted model.

Despite this, our proposed method has potential limitations. As with the standard Heckman control function treatment approach, our model is predicated and estimated under standard parametric assumptions. Further simulation research is required to examine the asymptotic properties of the estimator and the robustness of treatment estimates to violation of the joint normality assumption/misspecification[7]. Similar research is warranted to investigate the reported low degree of correlation between the control variables and the selection term - which if found to generalise, would be an attractive feature - together with the possibility of improved identification (Figure 1).

As discussed in Section 1, there is potential to estimate our model with more flexible techniques. As described by Wooldridge (2012), a number of semiparametric and nonparametric control function methods have been developed for endogenous continuous regressors; but he stresses (p. 9) that solutions are 'much harder' for endogenous discrete variables. He then proposes (pp. 21-28) a number of more flexible estimators which relax parametric assumptions for discrete cases. In addition, the research of Newey et al. (1990) demonstrates how semiparametric methods can be employed to estimate sample selection models. Further research is warranted to investigate the feasibility of employing these methods to estimate our model under less restrictive assumptions.

Appendix: Variables

Outcome variable: real hourly gross pay in constant 2009 prices.

Treatment variable: binary; 1(0) = full-time (part-time).

Base covariates
Education (6 qualification levels), age (8 groups), white or non-white, region (12 standard regions), married/cohabiting dummy, job tenure (5 dummies), dummy for employer size, industry (9 groups), dummies for year and month.

Broad covariates
Base plus dummies for: Managers and senior officials; professional occupations; associate professional and technical; administrative and secretarial; skilled trades occupations; personal service occupations; sales and customer service occupations; process, plant and machine operatives; and elementary occupations (one digit SOC, 2000).

Instruments (excluded from regressions)
Dummies for child aged under 2, 2-4 and 5-15; and number of children aged less than 19.

[7] As stressed by a Reviewer, it is important that the methodology specified in this paper is subject to further investigation; particularly with regard to the finite properties of the model and its robustness to misspecification of the IMRs.

References

Black, D. and Smith, J. (2004), ‘How robust is the evidence on the effects of college quality? evidence from matching’, *Journal of Econometrics*, 121(1), 99–124.

Blanchflower, D. and Bryson A. (2004), *The union wage premium in the US and the UK*, London: Centre for Economic Performance.

Blundell, R., Reed, H. and Stoker, T. (2003), ‘Interpreting aggregate wage growth: The role of labor market participation’, *American Economic Review*, 93(4), 1114–1131.

Caliendo, M. and Kopeinig, S. (2007), ‘Sensitivity analysis for average treatment effects’, *The Stata Journal*, 7(1), 71–83.

Connolly, S. and Gregory, M. (2009), ‘The part-time pay penalty earnings trajectories of British women’, *Oxford Economic Papers*, 61(1), 76–97.

De Luna, X. and Johansson, P. (2012), ‘Testing for nonparametric identification of causal effects in the presence of a quasi-instrument’, IZA DP No. 6692, The Institute for the Study of Labor, Bonn.

DiPrete, T. and Gangl, M. (2004), ‘Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments’, *Sociological Methodology*, 34(1), 271–310.

Donald, S., Hsu, Y. and Lieli, R. (2011), ‘Testing the unconfoundedness assumption via inverse probability weighted estimators of (L)ATT’, Working Paper, University of Texas, Austin.

Firpo, S., Fortin, N. and Lemieux, T. (2011), ‘Occupational tasks and changes in the wage structure’, IZA DP No. 5542, The Institute for the Study of Labor, Bonn.

Gemici, S., Rojewski, J. and Lee, H. (2012), ‘Use of propensity score matching for training research with observational data’, *International Journal of Training Research*, 10(3), 219–232.

Gibson, J. (2009), ‘The public sector pay premium, compensating differentials and unions: propensity score matching evidence from Australia, Canada, Great Britain and the United States’, *Economics Bulletin*, 29(3), 2325–2332.

Greene, W. (2005), *Econometric Analysis*, 5th edition, Prentice-Hall International.

Heckman, J. and Sedlacek, G. (1985), ‘Heterogeneity, aggregation and market wage functions: An empirical model of self-selection in the labor market’, *Journal of Political Economy*, 93(6), 1077–1125.

Heckman, J. and Vytlacil, E. (2007), ‘Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation’, in: *Handbook of Econometrics* volume 6b, 4780–4874, Elsevier, North-Holland.

Ho, D., Kosuke, I., King, G. and Stuart, E. (2007), 'Matching as nonparametric preprocessing for improving parametric causal inference', *Political Analysis*, 15(3), 199–236.

Huber, M. (2013), 'Treatment evaluation in the presence of sample selection', *Econometric Reviews* (forthcoming), https://www.alexandria.unisg.ch/export/DL/206865.pdf

Imbens, G. and Angrist, J. (1994), 'Identificationand estimation of local average treatment effects', *Econometrica*, 62(2), 467–75.

Keele, L. (2010), *An R package for Rosenbaum Bounds sensitivity analysis with matched data*', Department of Political Science, Ohio State University.

Klein, R. and Vella, F. (2009), 'A semiparametric model for binary response and continuous outcomes under index heteroscedasticity', *Journal of Applied Econometrics*, 24(5), 735–762.

Lin, E. and Lue, Y-L. (2010), 'The causal effect of the cram schooling timing decision on math scores', *Economics Bulletin*, 30 (3), 2330–2345.

Manning, A. and Petrongolo, B. (2008), 'The part-time pay penalty for women in Britain', *The Economic Journal*, 118 (526), 28–51.

Mumford, K. and Smith, P. (2009) 'What determines part time and gender earnings gaps in Britain: evidence from the workplace', *Oxford Economic Papers*, 61(1), 56–75.

Nawata, K. (1993), 'A note on the estimation of models with sample selection biases, *Economics Letters*, 42 (1), 15–24.

Neuberger J., Joshi, H. and Dex, S. (2011), *Part time working and pay among Millenium Cohort Study mothers*', CLS Cohort Studies Working Paper 2011-2, London: Centre for Longitudinal Studies.

Newey, W., Powell, J. and Walker, J. (1990), 'Semiparametric estimation of selection models: some empirical results', *American Economic Review*, 80 (2), 324–328.

Oostendorp, R. (2009), 'Globalization and the gender wage gap', *World Bank Economic Review*, 23 (1), 141–161.

Pike, R. (2011) 'Patterns of pay: results of the Annual Survey of Hours and Earnings 1997 to 2010', *Economic and Labour Market Review*, March, 14–40.

Puhani, P. (2000), 'The Heckman correction for sample selection and its critique', *Journal of Economic Surveys*, 14 (1), 53–68.

Rosenbaum, P. (2002), *Observational studies*. New York: Springer.

Rosenbaum, P. (2005), 'Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies', *The American Statistician*, 59 (2) 147–152.

Rosenbaum, P. (2010), *Design of observational studies*, Series in Statistics, New York: Springer.

Rosenbaum, P. and Rubin, D. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, 70 (1), 51–5.

Rubin, D. (1973) 'The use of matched sampling and regression adjustment to remove bias in observational studies', *Biometrics*, 29 (1), 185–203.

Rubin, D. (1979),'Using multivariate matched sampling and regression adjustment to control bias in observational studies', *Journal of the American Statistical Association*, 74 (366), 318–32.

Rubin D. and Thomas, N. (2000), 'Combining propensity score matching with additional adjustments for prognostic covariates', *Journal of the American Statistical Association*, 95(450), 573–585.

Todd, P. (1996), *Essays on empirical methods for evaluating the impact education and training*, PhD dissertation. University of Chicago, Chicago.

Wooldridge, J. (2012), *Semiparametric and nonparametric control function methods*, Institute for Fiscal Studies, London. www.ifs.org.uk/docs/wooldridge%20session%203.pdf