Warsaw School of Economics
Collegium of Economic Analysis

Linear Models in the Analysis of Heterogeneous Treatment Effects

Tymon Słoczyński

A summary of the
doctoral dissertation

Supervisor:
Prof. dr hab. Marek Góra

Co-supervisor:
Dr Michał Myck

In this dissertation I study the applicability of basic linear models, namely the linear regression model and the Oaxaca–Blinder decomposition, in settings with heterogeneous treatment effects. A large body of recent research in the econometrics of programme evaluation has allowed for general heterogeneity in treatment effects (see, e.g., Blundell and Costa Dias 2009; Imbens and Wooldridge 2009). Quite understandably, researchers have typically considered semiparametric and nonparametric estimators such as inverse probability weighting, methods based on the propensity score, and matching on covariates. Little research has been devoted to understanding what is being identified in the linear regression model in the presence of heterogeneous treatment effects and whether there exist alternative linear models which would allow for treatment effect heterogeneity in a satisfactory way. In this dissertation I attempt to fill this gap in the recent literature.

In Chapters 1 and 2, I provide an introduction to this dissertation as well as describe its background, namely the treatment effects literature and the decomposition literature. Until recently, these two frameworks have been developed independently of each other, even though they are strikingly similar and often ask related questions. Recent research of Barsky et al. (2002), Black et al. (2006, 2008), Melly (2006), Fortin et al. (2011), and Kline (2011) has allowed for some degree of convergence of these disciplines, and this dissertation can be seen as a further step in this process.

Chapters 3–5 contain the main contributions of this dissertation. Chapter 3 provides a new interpretation of the linear regression estimand in the presence of heterogeneous treatment effects. I study the implications of treatment effect heterogeneity for least squares estimation when the effects are inappropriately assumed to be homogeneous. I prove that under a set of benchmark assumptions linear regression provides a consistent estimator of the population average treatment effect on the treated (PATT) times the population proportion of the nontreated individuals plus the population average treatment effect on the nontreated (PATN) times the population proportion of the treated individuals. Consequently, in many empirical applications the linear regression estimates might not be close to any of the standard average treatment effects of interest. This result stands in stark contrast to the previous interpretations in Angrist (1998) and Humphreys (2009), and calls into question some of the recommendations in Angrist and Pischke (2009).

Chapter 4 studies various versions of the Oaxaca–Blinder decomposition, a popular method used in empirical labour economics to study differentials in mean wages. I develop a consistent estimator of the population average treatment effect (PATE) which is based on a nonstandard version of the Oaxaca–Blinder decomposition. I also reinterpret

other versions of this method, namely the Reimers (1983), Cotton (1988), and Fortin (2008) decompositions. As a result, I extend the recent literature which has utilised the treatment effects framework to reinterpret this technique, and propose an alternative solution to its fundamental problem of comparison group choice. I also use the Oaxaca–Blinder decomposition and its semiparametric extension to decompose gender wage differentials with the UK Labour Force Survey (LFS) data, while providing separate estimates of the average gender effect on men, women, and the whole population.

Chapter 5 uses data from the National Supported Work (NSW) Demonstration (see, e.g., LaLonde 1986; Dehejia and Wahba 1999; Smith and Todd 2005) to examine the finite-sample performance of the Oaxaca–Blinder decomposition as an estimator of the population average treatment effect on the treated. Precisely, I follow sample and variable selections from Dehejia and Wahba (1999), and conclude that Oaxaca–Blinder performs better than any of the estimators in this influential paper, provided that overlap is imposed. As a robustness check, I consider alternative sample (Smith and Todd 2005) and variable (Abadie and Imbens 2011) selections, and present an "empirical Monte Carlo study" (Huber et al. 2013) which is also based on the NSW data.

In other words, I provide a fairly negative result on the applicability of the linear regression model in the presence of heterogeneous treatment effects in Chapter 3. Next, in Chapter 4, I derive a simple solution to this problem which also constitutes a previously unknown bridge between the treatment effects literature and the decomposition literature. In Chapter 5, I examine the finite-sample performance of a related estimator, and conclude that it performs remarkably well. Finally, I summarise and discuss my findings in Chapter 6.

Theoretical Contributions

The homogeneous linear regression model is often believed to provide a good benchmark to study treatment effects, i.e. partial effects for a binary explanatory variable. A convincing explanation is given in Angrist and Pischke (2009), while many influential studies (e.g., Neal and Johnson 1996; Fryer and Levitt 2004) explicitly rely on linear regression to capture the possibly heterogeneous effects for a binary variable. A major contribution of this dissertation is to provide new evidence on the limitations of such an approach in light of "the pervasiveness of heterogeneity" (Heckman 2001). In particular, what is the appropriate interpretation of the least squares estimand in the homogeneous linear model if treatment effects are actually heterogeneous? In Chapter 3, I provide

a new answer to this question by exploiting the link between linear regression and the Oaxaca–Blinder decomposition (Oaxaca 1973; Blinder 1973) as well as utilising a recent theoretical result in Elder et al. (2010). I prove that under the assumptions of *(i)* a single control variable *(ii)* whose variance is equal in both subpopulations the linear regression estimand is a weighted average of both subpopulation-specific average treatment effects; while weights are equal to the population proportions of both groups, they are inappropriately interchanged between them. Consequently, the ability of linear regression to provide a good benchmark to study treatment effects is heavily data-dependent. Least squares estimation can be preferred on efficiency grounds if there is little heterogeneity in treatment effects or both subsamples are of approximately equal size; in the latter case both weights are more or less equal anyway. However, in other cases linear regression will provide biased estimates of all the standard average treatment effects of interest, even asymptotically. Also, linear regression possesses a highly undesirable property in that it attaches the *greater* weight to the linear estimate of the population average treatment effect on the treated (nontreated), the *smaller* is the sample proportion of the treated (nontreated) individuals.

This negative result on linear regression and treatment effect heterogeneity is taken for granted in Chapter 4. Therefore, the question that remains is whether there exist more flexible linear models which would allow for treatment effect heterogeneity in a satisfactory way. Such models have been discussed by Imbens and Wooldridge (2009) and Wooldridge (2010). The key contribution of Chapter 4 is to demonstrate that these models are equivalent to various versions of the Oaxaca–Blinder decomposition and that treatment effect heterogeneity can help to solve the well-known "comparison group choice problem" in the decomposition literature (see, e.g., Elder et al. 2010). Indeed, the comparison group choice problem, i.e. the question of which wage structure should be used as the counterfactual for observed wages, has constituted one of the major discussions in the Oaxaca–Blinder decomposition literature since the seminal papers of Oaxaca (1973) and Blinder (1973). Both these authors, whose main goal was to study U.S. gender wage gaps, referred to the problem of choosing either male or female wage coefficients as an "index number problem", thus suggesting this choice to be unclear. Subsequent contributions have established a tendency to regard the comparison coefficients as the "nondiscriminatory" or "competitive" wage structure (Reimers 1983; Cotton 1988; Neumark 1988; Oaxaca and Ransom 1994; Fortin 2008). It is only recently that Fortin et al. (2011) have distinguished between comparison wage structures based on the assumption of "simple counterfactual treatment" (Oaxaca 1973; Blinder 1973) and

alternative structures which "represent the appropriate counterfactual for the way women would be paid in the absence of labour market discrimination" (Fortin et al. 2011). In that case, the latter structures are assumed to capture what would happen in general equilibrium if discrimination ceased to exist.

In Chapter 4, I challenge this common practice of interpreting the propositions of Reimers (1983), Cotton (1988), and Fortin (2008) as reflecting "the world without discrimination". First, neither of these propositions has been based on a theoretical model of the labour market, so they can hardly address such general equilibrium considerations. Second, I use the treatment effects framework to show that these decompositions are easily interpretable within it, and they estimate some generally uninteresting weighted averages of subpopulation-specific average treatment effects.

A natural question in this context is whether the population average treatment effect (PATE) can be consistently estimated with some new version of the Oaxaca–Blinder decomposition. I derive such a new estimator which uses a linear combination of the regression coefficients for both subpopulations (treated and nontreated, men and women, union and nonunion workers, etc.) as the comparison wage structure. However, these coefficients are weighted in a nonstandard way, namely the sample proportion of group one is used to weight the coefficients for group two, and vice versa. Although such a weighting procedure may at first look counterintuitive, the treatment effects framework provides a clear rationale for this approach. Precisely, the role of each group's wage structure is to serve as counterfactual for the other group, so we need more weight to be put on the coefficients for the smaller group in order to consistently estimate the PATE.

Empirical Applications

In this dissertation I also provide empirical applications of my theoretical results, and analyse data from two well-known microdata sets: the UK Labour Force Survey (LFS) and the U.S. National Supported Work (NSW) Demonstration. Both these analyses provide an illustration of my theoretical contributions.

In Chapter 3, I analyse the NSW data to illustrate my theoretical result on linear least squares regression and treatment effect heterogeneity. I show empirically that my proposition continues to provide a good approximation to the behaviour of linear regression estimates even when the assumptions of this proposition are not satisfied. I also carry out a simple simulation exercise in which I demonstrate that the *larger* the sample proportion of a given group (treated or nontreated), the *more distant* are linear

regression estimates from the population average treatment effect on this group.

In Chapter 4, I provide a further empirical example which uses the new version of the Oaxaca–Blinder decomposition as well as its other versions to study gender wage differentials with the LFS data, for each year from 2002 to 2010. I also use normalised reweighting and a combination of stratification and different versions of the Oaxaca–Blinder decomposition to account for possible nonlinearities in the existing wage structures, and provide separate estimates of three parameters which I refer to as the population average gender effect (PAGE), the population average gender effect on men (PAGM), and the population average gender effect on women (PAGW). This is the first piece of work to clarify the distinction between these parameters and provide separate estimates for each of them. The major empirical finding of this study is that men gain typically more in comparison with similar women than women lose in comparison with similar men (the PAGM is consistently larger than the PAGW). This phenomenon is explained by the fact that average gender effects tend to increase with wages and this is indeed the case in the UK labour market.

In Chapter 5, I study the finite-sample performance of the Oaxaca–Blinder decomposition as an estimator of the population average treatment effect on the treated – in a further application to the NSW data. In this case, however, I closely follow Dehejia and Wahba (1999) in their sample and variable selections, so that I can reassess their influential claim that methods based on the propensity score compare favourably with other estimators. When overlap is imposed, the Oaxaca–Blinder decomposition is shown to perform superior compared to any of the estimators in Dehejia and Wahba (1999) and to additional methods such as inverse probability weighting, kernel matching, matching on covariates, and bias-corrected matching. To assess the robustness of this result, I consider alternative variable (Abadie and Imbens 2011) and sample (Smith and Todd 2005) selections, and present an "empirical Monte Carlo study" (Huber et al. 2013) which is also based on the NSW data. Generally, the Oaxaca–Blinder decomposition always performs very well, and never significantly worse than any other method. At first, this might be seen as surprising, given the simplicity of this estimator. Note, however, that at least two recent papers, Khwaja et al. (2011) and Huber et al. (2013), have presented simulation studies which are suggestive of very good finite-sample performance of flexible OLS. In both cases the authors have actually applied an estimator which is either equivalent or very similar to Oaxaca–Blinder, although have referred to this method in a different way. In Chapter 5, I complement these previous analyses by exploring the connection with the decomposition literature, and focus on the NSW data.

Summary and Conclusions

In this dissertation I study the applicability of two linear models – the linear regression model and the Oaxaca–Blinder decomposition – in settings with treatment effect heterogeneity. Recent research in applied microeconometrics (see, e.g., Heckman 2001; Bitler et al. 2006, 2008) has generally confirmed that heterogeneity in human behaviour is a pervasive phenomenon. Consequently, empirical researchers need reliable methods which account for treatment effect heterogeneity in a satisfactory way. In an influential textbook, Angrist and Pischke (2009) have recently claimed that the linear least squares regression provides a good benchmark to study treatment effects, even if the assumption of treatment effect homogeneity is not satisfied (see Angrist 1998 and Humphreys 2009 for further theoretical results). In this dissertation I reach a different conclusion, and I believe there are several lessons to be drawn from my results.

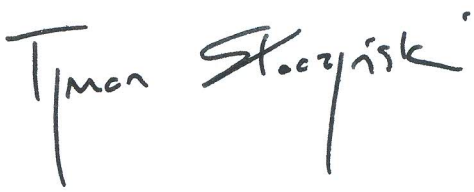
First, empirical researchers are advised *not* to use the linear least squares regression unless both groups of interest (treated and nontreated, men and women, union and nonunion workers, etc.) are of approximately equal size. As demonstrated in Chapter 3, linear regression possesses a previously unknown and highly undesirable property in that it attaches the *greater* weight to the linear estimate of the average effect on a given group, the *smaller* is the sample proportion of this group. In the limit, if group A were about to disappear and group B were about to dominate, linear regression would attach the whole weight to the average effect on group A. Such a result might discourage empirical researchers from using the linear regression model in the standard case of a single cross section of data, but this issue is likely to be especially serious in comparative studies.

Imagine, for example, an empirical study which attempts to capture changes over time in public-private sector wage differentials in Poland. Assume that public sector workers experience larger gains (or smaller losses) from public sector employment than private sector workers and that the researcher has documented a shift in wage structures which has favoured public sector workers. If she has used, however, the linear least squares regression to estimate these effects, such a change might result either from an actual shift in wage structures or from an increase in the proportion of *private* sector workers in the working population. Clearly, a reasonable researcher might want to discriminate between these two hypotheses, but this is not possible as long as she uses the linear least squares regression. A similar criticism is also applicable, of course, to between-country or between-region comparisons of various treatment effects of interest.

Second, even though my results might discourage empirical researchers from using

the linear regression model, it is because of its assumption of treatment effect homogeneity, not because of linearity. In Chapters 4 and 5, I provide new evidence on theoretical properties and finite-sample performance of the Oaxaca–Blinder decomposition – a more flexible linear model which allows for heterogeneity in the response to treatment. Especially, in Chapter 5, I replicate an influential study by Dehejia and Wahba (1999) which was instrumental in popularising methods based on the propensity score in the programme evaluation literature. I demonstrate that the appropriate linear model – a version of the Oaxaca–Blinder decomposition – performs better, on average, than any of the estimators in Dehejia and Wahba (1999). This finding is reconfirmed by several robustness checks as well as an "empirical Monte Carlo study" (Huber et al. 2013), and therefore I provide a strong case in favour of flexible linear models.

Third, in the context of decomposing intergroup wage differentials, my theoretical results suggest that empirical researchers should *not* use several versions of the Oaxaca–Blinder decomposition, namely the Reimers (1983), Cotton (1988), and Fortin (2008) decompositions. As demonstrated in Chapter 4, each of these decompositions is likely to overstate the importance of the smaller group when estimating average gender effects by attaching too large a weight to the effect on this group. Again, in the limit, if the sample proportion of one group goes to zero, the Cotton (1988) and Fortin (2008) decompositions are likely to identify and estimate the average effect on this group. Empirical researchers are generally advised to use either the original decompositions of Oaxaca (1973) and Blinder (1973) or the new decomposition which I derive in Chapter 4. The unexplained component of this decomposition provides an estimator of the population average treatment effect (PATE), and is also equivalent to the flexible OLS estimator in Imbens and Wooldridge (2009) and Wooldridge (2010).



References

[1] Abadie, Alberto, and Guido W. Imbens. 2011. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics* 29:1–11.

[2] Angrist, Joshua D. 1998. Estimating the labor market impact of voluntary military service using Social Security data on military applicants. *Econometrica* 66:249–88.

[3] Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton and Oxford: Princeton University Press.

[4] Barsky, Robert, John Bound, Kerwin Kofi Charles, and Joseph P. Lupton. 2002. Accounting for the black-white wealth gap: A nonparametric approach. *Journal of the American Statistical Association* 97:663–73.

[5] Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2006. What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review* 96:988–1012.

[6] Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2008. Distributional impacts of the self-sufficiency project. *Journal of Public Economics* 92:748–65.

[7] Black, Dan, Amelia Haviland, Seth Sanders, and Lowell Taylor. 2006. Why do minority men earn less? A study of wage differentials among the highly educated. *Review of Economics and Statistics* 88:300–13.

[8] Black, Dan A., Amelia M. Haviland, Seth G. Sanders, and Lowell J. Taylor. 2008. Gender wage disparities among the highly educated. *Journal of Human Resources* 43:630–59.

[9] Blinder, Alan S. 1973. Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources* 8:436–55.

[10] Blundell, Richard, and Monica Costa Dias. 2009. Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources* 44:565–640.

[11] Cotton, Jeremiah. 1988. On the decomposition of wage differentials. *Review of Economics and Statistics* 70:236–43.

[12] Dehejia, Rajeev H., and Sadek Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94:1053–62.

[13] Elder, Todd E., John H. Goddeeris, and Steven J. Haider. 2010. Unexplained gaps and Oaxaca–Blinder decompositions. *Labour Economics* 17:284–90.

[14] Fortin, Nicole M. 2008. The gender wage gap among young adults in the United States: The importance of money versus people. *Journal of Human Resources* 43:884–918.

[15] Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. Decomposition methods in economics. In *Handbook of labor economics*, vol. 4A, ed. Orley Ashenfelter and David Card. San Diego and Amsterdam: Elsevier.

[16] Fryer, Roland G., and Steven D. Levitt. 2004. Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics* 86:447–64.

[17] Heckman, James J. 2001. Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture. *Journal of Political Economy* 109:673–748.

[18] Huber, Martin, Michael Lechner, and Conny Wunsch. 2013. The performance of estimators based on the propensity score. *Journal of Econometrics* 175:1–21.

[19] Humphreys, Macartan. 2009. Bounds on least squares estimates of causal effects in the presence of heterogeneous assignment probabilities. Unpublished manuscript, Department of Political Science, Columbia University.

[20] Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47:5–86.

[21] Khwaja, Ahmed, Gabriel Picone, Martin Salm, and Justin G. Trogdon. 2011. A comparison of treatment effects estimators using a structural model of AMI treatment choices and severity of illness information from hospital charts. *Journal of Applied Econometrics* 26:825–53.

[22] Kline, Patrick. 2011. Oaxaca–Blinder as a reweighting estimator. *American Economic Review: Papers & Proceedings* 101:532–37.

[23] LaLonde, Robert J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76:604–20.

[24] Melly, Blaise. 2006. Applied quantile regression. PhD diss., University of St. Gallen.

[25] Neal, Derek A., and William R. Johnson. 1996. The role of premarket factors in black-white wage differences. *Journal of Political Economy* 104:869–95.

[26] Neumark, David. 1988. Employers' discriminatory behavior and the estimation of wage discrimination. *Journal of Human Resources* 23:279–95.

[27] Oaxaca, Ronald. 1973. Male-female wage differentials in urban labor markets. *International Economic Review* 14:693–709.

[28] Oaxaca, Ronald L., and Michael R. Ransom. 1994. On discrimination and the decomposition of wage differentials. *Journal of Econometrics* 61:5–21.

[29] Reimers, Cordelia W. 1983. Labor market discrimination against Hispanic and black men. *Review of Economics and Statistics* 65:570–79.

[30] Smith, Jeffrey A., and Petra E. Todd. 2005. Does matching overcome LaLonde's critique of non-experimental estimators? *Journal of Econometrics* 125:305–53.

[31] Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge and London: MIT Press.