CHAPTER 6

Identification of Models of the Labor Market[☆]

Eric French[*], **Christopher Taber**[**]

[*] Federal Reserve Bank of Chicago

[**] Department of Economics, University of Wisconsin-Madison and NBER

Contents

[☆] We thank Pedro Carneiro, Bruce Hansen, John Kennan, Salvador Navarro, Jim Walker, and students in Taber's 2010 Economics 751 class for comments. The opinions and conclusions are solely those of the authors, and should not be construed as representing the opinions of the Federal Reserve System. We thank Zach Seeskin and David Benson for excellent research assistance.

Handbook of Labor Economics, Volume 4a
ISSN 0169-7218, DOI 10.1016/S0169-7218(11)00412-6

Abstract

This chapter discusses identification of common selection models of the labor market. We start with the classic Roy model and show how it can be identified with exclusion restrictions. We then extend the argument to the generalized Roy model, treatment effect models, duration models, search models, and dynamic discrete choice models. In all cases, key ingredients for identification are exclusion restrictions and support conditions.

JEL classification: C14; C51; J22; J24

Keywords: Identification; Roy model; Discrete choice; Selection; Treatment effects

1. INTRODUCTION

This chapter discusses identification of common selection models of the labor market. We are primarily concerned with nonparametric identification. We view nonparametric identification as important for the following reasons.

First, recent advances in computer power, more widespread use of large data sets, and better methods mean that estimation of increasingly flexible functional forms is possible. Flexible functional forms should be encouraged. The functional form and distributional assumptions used in much applied work rarely come from the theory. Instead, they come from convenience. Furthermore, they are often not innocuous.[1]

Second, the process of thinking about nonparametric identification is useful input into applied work. It is helpful to an applied researcher both in informing her about which type of data would be ideal and which aspects of the model she might have some hope of estimating. If a feature of the model is not nonparametrically identified, then one knows it cannot be identified directly from the data. Some additional type of functional form assumption must be made. As a result, readers of empirical papers are often skeptical of the results in cases in which the model is not nonparametrically identified.

Third, identification is an important part of a proof of consistency of a nonparametric estimator.

However, we acknowledge the following limitation of focusing on nonparametric identification. With any finite data set, an empirical researcher can almost never be completely nonparametric. Some aspects of the data that might be formally identified could never be estimated with any reasonable level of precision. Instead, estimators are usually only nonparametric in the sense that one allows the flexibility of the model to

[1] A classic reference on this is Lalonde (1986) who shows that parametric models cannot replicate the results of an experiment. Below we present an example on Catholic schools from Altonji et al. (2005a) suggesting that parametric assumptions drive the empirical estimates.

grow with the sample size. A nice example of this is Sieve estimators in which one estimates finite parameter models but the number of parameters gets large with the data set. An example would be approximating a function by a polynomial and letting the degree of the polynomial get large as the sample size increases. However, in that case one still must verify that the model is nonparametrically identified in order to show that the model is consistent. One must also construct standard errors appropriately. In this chapter we do not consider the purely statistical aspects of nonparametric estimation, such as calculation of standard errors. This is a very large topic within econometrics.[2]

The key issue in identification of most models of the labor market is the selection problem. For example, individuals are typically not randomly assigned to jobs. With this general goal in mind we begin with the simplest and most fundamental selection model in labor economics, the Roy (1951) model. We go into some detail to explain Heckman and Honoré's (1990) results on identification of this model. A nice aspect of identification of the Roy model is that the basic methodology used in this case can be extended to show identification of other labor models. We spend the rest of the chapter showing how this basic intuition can be used in a wide variety of labor market models. Specifically we cover identification in the generalized Roy model, treatment effect models, the competing risk model, search models, and forward looking dynamic models. While we are clearly not covering all models in labor economics, we hope the ideas are presented in a way that the similarities in the basic models can be seen and can be extended by the reader to alternative frameworks.

The plan of this chapter is specifically as follows. Section 2 discusses some econometric preliminaries. We consider the Roy model in Section 3, generalize this to the Generalized Roy model in Section 4, and then use the model to think about identification of treatment effects in Section 5. In Section 6 we consider duration models and search models and then consider estimation of dynamic discrete choice models in Section 7. Finally in Section 8 we offer some concluding thoughts.

2. ECONOMETRIC PRELIMINARIES

2.1. Notation

Throughout this chapter we use capital letters with i subscripts to denote random variables and small letters without i subscripts to denote possible outcomes of that random variable. We will also try to be explicit throughout this chapter in denoting conditioning. Thus, for example, we will use the notation

$$E(Y_i \mid X_i = x)$$

to denote the expected value of outcome Y_i conditional on the regressor variable X_i being equal to some realization x.

[2] See Chen (2007) for discussion of Sieve estimators, including standard error calculation.

2.2. Identification

The word "identification" has come to mean different things to different labor economists. Here, we use a formal econometrics definition of identification. Consider two different models that lead to two data generating processes. If the data generated by these two models have exactly the same distribution then the two models are not separately identified from each other. However, if any two different model specifications lead to different data distributions, the two specifications are separately identified. We give a more precise definition below. Our definition of identification is based on some of the notation and set up of Matzkin's (2007) following an exposition based on Shaikh (2010).

Let P denote the true distribution of the observed data X. An econometric model defines a data generating process. We assume that the model is specified up to an unknown vector θ of parameters, functions and distribution functions. This is known to lie in space Θ. Within the class of models, the element $\theta \in \Theta$ determines the distribution of the data that is observable to the researcher P_θ. Notice that identification is fundamentally data dependent. With a richer data set, the distribution P_θ would be a different object.

Let $\mathcal{P}$ be the set of all possible distributions that could be generated by the class of models we consider (i.e. $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$). We assume that the model is correctly specified, which means that $P \in \mathcal{P}$. The identified set is defined as

$$\Theta(P) \equiv \{\theta \in \Theta : P_\theta = P\}.$$

This is the set of possible θ that could have generated data that has distribution P. By assuming that $P \in \mathcal{P}$ we have assumed that our model is correctly specified so this set is not empty. We say that θ *is identified if* $\Theta(P)$ *is a singleton for all* $P \in \mathcal{P}$.

The question we seek to answer here is under what conditions is it possible to learn about θ (or some feature of θ) from the distribution of the observed data P. Our interest is not always to identify the full data generating process. Often we are interested in only a subset of the model, or a particular outcome from it. Specifically, our goal may be to identify

$$\psi = \Psi(\theta),$$

where Ψ is a known function. For example in a regression model $Y_i = X_i'\beta + u_i$, the feature of interest is typically the regression coefficients. In this case Ψ would take the trivial form

$$\Psi(\theta) = \beta.$$

However, this notation allows for more general cases in which we might be interested in identifying specific aspects of the model. For example, if our interest is in identifying the covariance between X and Y in the case of the linear regression model, we do not need to know θ per se, but rather a transformation of these parameters. That is we could be interested in

$$\Psi(\theta) = \operatorname{cov}(X_i, Y_i).$$

We could also be interested in a forecast of the model such as

$$\Psi(\theta) = x'\beta$$

for some specific x. The distinction between identification of features of the model as opposed to the full model is important, as in many cases the full model is not identified but the key feature of interest is identified.

To think about identification of ψ we define

$$\Psi(\Theta(P)) = \{\Psi(\theta) : \theta \in \Theta(P)\}.$$

That is, it is the set of possible values of ψ that are consistent with the data distribution P. We say that ψ *is identified if* $\Psi(\Theta(P))$ *is a singleton.*

As an example consider the standard regression model with two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{2.1}$$

with $E(\varepsilon_i \mid X_i = x) = 0$ for any value x (where $X_i = (X_{1i}, X_{2i})$). In this case $\theta = (\beta, F_{X,\varepsilon})$, where $F_{X,\varepsilon}$ is the joint distribution of $(X_{1i}, X_{2i}, \varepsilon_i)$ and $\beta = (\beta_0, \beta_1, \beta_2)$. One would write Θ as $\mathcal{B} \times \mathcal{F}_{X,\varepsilon}$, where $\mathcal{B}$ is the parameter space for β and $\mathcal{F}_{X,\varepsilon}$ is the space of joint distributions between X_i and ε_i that satisfy $E(\varepsilon_i \mid X_i = x) = 0$ for all x. Since the data here is represented by (X_{1i}, X_{2i}, Y_i), P_θ represents the joint distribution of (X_{1i}, X_{2i}, Y_i). Given knowledge of β and $F_{X,\varepsilon}$ we know the data generating process and thus we know P_θ.

To focus ideas suppose we are interested in identifying β (i.e. $\Psi(\beta, F_{X,\varepsilon}) = \beta$) in regression model (2.1) above. Let the true value of the data generating process be $\theta^* = (\beta^*, F^*_{X,\varepsilon})$ so that by definition $P_{\theta^*} = P$. In this case $\Theta(P) \equiv \left\{(\beta, F_{X,\varepsilon}) \in \mathcal{B} \times \mathcal{F}_{X,\varepsilon} : P_{\beta, F_{x,\varepsilon}} = P\right\}$, that is it is the set of $(\beta, F_{X,\varepsilon})$ that would lead our data (X_i, Y_i) to have distribution P. In this case $\Psi(\Theta(P))$ is the set of values of β in this set (i.e. $\Psi(\Theta(P)) = \{\beta : (\beta, F_{X,\varepsilon}) \in \Theta(P)$ for some $F_{X,\varepsilon} \in \mathcal{F}_{X,\varepsilon}\})$.

In the case of 2 covariates, we know the model is identified as long as X_{1i} and X_{2i} are not degenerate and not collinear. To see how this definition of identification applies to this model, note that for any $\beta^* \neq \beta$ the lack of perfect multicollinearity means that we can always find values of (x_1, x_2) for which

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \neq \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2.$$

Since $E(Y_i \mid X_i = x)$ is one aspect of the joint distribution of P_θ, it must be the case that when $\beta^* \neq \beta$, $P_\theta \neq P$. Since this is true for any value of $\beta \neq \beta^*$, then $\Psi(\Theta(P))$ must be the singleton β^*.

However, consider the well known case of perfect multicollinearity in which the model is not identified. In particular suppose that

$$X_{1i} + X_{2i} = 1.$$

For the true value of $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)$ consider some other value $\widetilde{\beta} = (\beta_0^* + \beta_2^*, \beta_1^* - \beta_2^*, 0)$. Then for any x,

$$\begin{aligned} E(Y_i \mid X_i = x) &= \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2 \\ &= \beta_0^* + \beta_1^* x_1 + \beta_2^* (1 - x_1) \\ &= \beta_0^* + \beta_2^* + \left(\beta_1^* - \beta_2^*\right) x_1 \\ &= \widetilde{\beta}_0 + \widetilde{\beta}_1 x_1. \end{aligned}$$

If $F_{X,\varepsilon}$ is the same for the two models, then the joint distribution of (Y_i, X_i) is the same in the two cases. Thus the identification condition above is violated because with $\widetilde{\theta} = (\widetilde{\beta}, F_{X,\varepsilon}^*)$, $P_{\widetilde{\theta}} = P$ and thus $\widetilde{\beta} \in \Psi(\Theta(P))$. Since the true value $\beta^* \in \Psi(\Theta(P))$ as well, $\Psi(\Theta(P))$ is not a singleton and thus β is not identified.

2.3. Support

Another important issue is the support of the data. The simplest definition of support is just the range of the data. When data are discrete, this is the set of values that occur with positive probability. Thus a binary variable that is either zero or one would have support $\{0, 1\}$. The result of a die roll has support $\{1, 2, 3, 4, 5, 6\}$. With continuous variables things get somewhat more complicated. One can think of the support of a random variable as the set of values for which the density is positive. For example, the support of a normal random variable would be the full real line (which we will often refer to as "full support"). The support of a uniform variable on $[0, 1]$ is $[0, 1]$. The support of an exponential variable would be the positive real line.

This can be somewhat trickier in dealing with outcomes that occur with measure zero. For example one could think of the support of a uniform variable as $[0, 1]$, $(0, 1]$, $[0, 1)$, or $(0, 1)$. The distinction between these objects will not be important in what we are doing, but to be formal we will use the Davidson (1994) definition of support. He defines the support of a random variable with distribution F

as the set of points at which F is (strictly) increasing.[3] By this definition, the support of a uniform would be $[0, 1]$. We will also use the notation $\operatorname{supp}(Y_i)$ to denote the unconditional support of random variable Y_i and $\operatorname{supp}(Y_i \mid X_i = x)$ to denote the conditional support.

To see the importance of this concept, consider a simple case of the separable regression model

$$Y_i = g(X_i) + u_i$$

with a single continuous X_i variable and $E(u_i \mid X_i = x) = 0$ for $x \in \operatorname{supp}(X_i)$. In this case we know that

$$E(Y_i \mid X_i = x) = g(x).$$

Letting $\mathcal{X}$ be the support of X_i, it is straightforward to see that g is identified on the set $\mathcal{X}$. But g is not identified outside the set $\mathcal{X}$ because the data is completely silent about these values. Thus if $\mathcal{X} = \mathbb{R}$, g is globally identified. However, if $\mathcal{X}$ only covers a subset of the real line it is not. For example, one interesting counterfactual is the change in the expected value of Y_i if X_i were increased by $\delta : E(g(X_i + \delta))$. If $\mathcal{X} = \mathbb{R}$ this is trivially identified, but if the support of X_i were bounded from above, this would no longer be the case. That is, if the supremum of $\mathcal{X}$ is $\bar{x} < \infty$, then for any value of $x > \bar{x} - \delta$, $g(x + \delta)$ is not identified and thus the unconditional expected value of $g(X_i + \delta)$ is not identified either. This is just a restatement of the well known fact that one cannot project out of the data unless one makes functional form assumptions. Our point here is that support assumptions are very important in nonparametric identification results. One can only identify g over the range of plausible values of X_i if X_i has full support. For this reason, we will often make strong support condition assumptions. This also helps illuminate the tradeoff between functional form assumptions and flexibility. In order to project off the support of the data in a simple regression model one needs to use some functional form assumption. The same is true for selection models.

2.4. Continuity

There is one complication that we need to deal with throughout. It is not a terribly important issue, but will shape some of our assumptions. Consider again the separable regression model

$$Y_i = g(X_i) + u_i. \tag{2.2}$$

[3] He defines F (strictly) increasing at point x to mean that for any $\varepsilon > 0, F(x + \varepsilon) > F(x - \varepsilon)$.

As mentioned above $E(Y_i \mid X_i = x) = g(x)$, so it seems trivial to see that g is identified, but that is not quite true. To see the problem, suppose that both X_i and u_i are standard normals. Consider two different models for g,

Model 1:

$$g(x) = \begin{cases} 0 & x < 1.4 \\ 1 & x \geq 1.4 \end{cases}$$

versus

Model 2:

$$g(x) = \begin{cases} 0 & x \leq 1.4 \\ 1 & x > 1.4. \end{cases}$$

These models only differ at the point $x = 1.4$, but since X_i is normal this is a zero probability event and we could never distinguish between these models because they imply the same joint distribution of (X_i, Y_i). For the exact same reason it isn't really a concern (except in very special cases such as if one was evaluating a policy in which we would set $X_i = 1.4$ for everyone). Since this will be an issue throughout this chapter we explain how to deal with it now and use this convention throughout the chapter.

We will make the following assumptions.

Assumption 2.1. X_i can be written as (X_i^c, X_i^d), where the elements of X_i^c are continuously distributed (no point has positive mass), and X_i^d is distributed discretely (all support points have positive mass).

Assumption 2.2. For any $x^d \in \text{supp}(X_i^d)$, $g(x^c, x^d)$ is almost surely continuous across $x^c \in \text{supp}(X_i^c \mid X_i^d = x^d)$.

The first part says that we can partition our observables into continuous and discrete ones. One could easily allow for variables that are partially continuous and partially discrete, but this would just make our results more tedious to exposit. The second assumption states that choosing a value of X at which g is discontinuous (in the continuous variables) is a zero probability event.

Theorem 2.1. *Under Assumptions 2.1 and 2.2 and assuming model (2.2) with* $E(u_i \mid X_i = x) = 0$ *for* $x \in \text{supp}(X_i)$, $g(x)$ *is identified on a set* $\mathcal{X}^*$ *that has measure* 1.

(Proof in Appendix.)

The proof just states that g is identified almost everywhere. More specifically it is identified everywhere that it is continuous.

3. THE ROY MODEL

The classic model of selection in the labor market is the Roy (1951) model. In the Roy model, workers choose one of two possible occupations: hunting and fishing. They cannot pursue both at the same time. The worker's log wage is Y_{fi} if he fishes and Y_{hi} if he hunts. Workers maximize income so they choose the occupation with the higher wage. Thus a worker chooses to fish if $Y_{fi} > Y_{hi}$. The occupation is defined as

$$J_i = \begin{cases} f & \text{if } Y_{fi} > Y_{hi} \\ h & \text{if } Y_{hi} \geq Y_{fi} \end{cases} \tag{3.1}$$

and the log wage is defined as

$$Y_i = \max\{Y_{fi}, Y_{hi}\}. \tag{3.2}$$

Workers face a simple binary choice: choose the job with the highest wage. This simplicity has led the model to be used in one form or another in a number of important labor market contexts. Many discrete choice models share the Roy model's structure. Examples in labor economics include the choice of whether to continue schooling, what school to attend, what occupation to pursue, whether to join a union, whether to migrate, whether to work, whether to obtain training, and whether to marry.

As mentioned in the introduction, we devote considerable attention to identification of this model. In subsequent sections we generalize these results to other models.

The responsiveness of the supply of fishermen to changes in the price of fish depends critically on the joint distribution of $\left(Y_{fi}, Y_{hi}\right)$. Thus we need to know what a fisherman would have made if he had chosen to hunt. However, we do not observe this but must infer its counterfactual distribution from the data at hand. Our focus is on this selection problem. Specifically, much of this chapter is concerned with the following question: **Under what conditions is the joint distribution of $\left(Y_{fi}, Y_{hi}\right)$ identified?** We start by considering estimation in a parametric model and then consider nonparametric identification.

Roy (1951) is concerned with how occupational choice affects the aggregate distribution of earnings and makes a series of claims about this relationship. These claims turn out to be true when the distribution of skills in the two occupations is lognormal.

Heckman and Honoré (1990) consider identification of the Roy model (i.e., the joint distribution of (Y_{fi}, Y_{hi})). They show that there are two methods for identifying the Roy model. The first is through distributional assumptions. The second is through exclusion restrictions.[4]

[4] Heckman and Honoré discuss price variation as separate from exclusion restrictions. However, in our framework price changes can be modeled as just one type of exclusion restriction so we do not explicitly discuss price variation.

In order to focus ideas, we use the following case:

$$Y_{fi} = g_f(X_{fi}, X_{0i}) + \varepsilon_{fi} \tag{3.3}$$

$$Y_{hi} = g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}, \tag{3.4}$$

where the unobservable error terms $(\varepsilon_{fi}, \varepsilon_{hi})$ are independent of the observable variables $X_i = (X_{fi}, X_{hi}, X_{0i})$ and Y_{fi} and Y_{hi} denote log wages in the fishing and hunting sectors respectively. We distinguish between three types of variables. X_{0i} influences productivity in both fishing and hunting, X_{fi} influences fishing only, and X_{hi} influences hunting only. The variables X_{fi} and X_{hi} are "exclusion restrictions," and play a very important role in the identification results below. In the context of the Roy model, an exclusion restriction could be a change in the price of rabbits which increases income from hunting, but not from fishing. The notation is general enough to incorporate a model without exclusion restrictions (in which case one or more of the X_{ji} would be empty).

Our version of the Roy framework imposes two strong assumptions. First, that Y_{ji} is separable in $g_j(X_{ji}, X_{0i})$ and ε_{ji} for $j \in \{f, h\}$. Second, we assume that $g_j(X_{ji}, X_{0i})$ and ε_{ji} are independent of one another. Note that independence implies homoskedasticity: the variance of ε_{ji} cannot depend on X_{ji}. There is a large literature looking at various other more flexible specifications and this is discussed thoroughly in Matzkin (2007). It is also trivial to extend this model to allow for a general relationship between X_{0i} and $(\varepsilon_{fi}, \varepsilon_{hi})$, as we discuss in Section 3.3 below.

We focus on the separable independent model for two reasons. First, the assumptions of separability and independence have bite beyond a completely general nonparametric relationship. That is, to the extent that they are true, identification is facilitated by these assumptions. Presumably because researchers think these assumptions are approximately true, virtually all empirical research uses these assumptions. Second, despite these strong assumptions, they are obviously much weaker than the standard assumptions that g is linear (i.e. $g_f(X_{fi}, X_{0i}) = X'_{fi}\gamma_{ff} + X'_{0i}\gamma_{0f}$) and that ε_{fi} is normally distributed. One approach to writing this chapter would have been to go through all of the many specifications and alternative assumptions. We choose to focus on a single base specification for expositional simplicity.

Heckman and Honoré (1990) first discuss identification of the joint distribution of $\left(Y_{fi}, Y_{hi}\right)$ using distributional assumptions. They show that when one can observe the distribution of wages in both sectors, and assuming $\left(Y_{fi}, Y_{hi}\right)$ is joint normally distributed, then the joint distribution of $\left(Y_{fi}, Y_{hi}\right)$ is identified from a single cross section even without any exclusion restrictions or regressors. To see why, write equations (3.3) and (3.4) without regressors (so $g_f = \mu_f$, the mean of Y_{fi}):

$$Y_{fi} = \mu_f + \varepsilon_{fi}$$

$$Y_{hi} = \mu_h + \varepsilon_{hi}$$

where

$$\begin{bmatrix} \varepsilon_{fi} \\ \varepsilon_{hi} \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_f^2 & \sigma_{fh} \\ \sigma_{fh} & \sigma_h^2 \end{bmatrix}\right).$$

Letting

$$\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$$

(with ϕ and Φ the pdf and cdf of a standard normal),

$$c = \frac{\mu_f - \mu_h}{\sqrt{\sigma_f^2 + \sigma_h^2 - 2\sigma_{fh}}},$$

and for each $j \in \{h, f\}$,

$$\tau_j = \frac{\sigma_j^2 - \sigma_{fh}}{\sqrt{\sigma_f^2 + \sigma_h^2 - 2\sigma_{fh}}}.$$

One can derive the following conditions from properties of normal random variables found in Heckman and Honoré (1990):

$$\begin{aligned}
\Pr(J_i = f) &= \Phi(c) \\
E(Y_i \mid J_i = f) &= \mu_f + \tau_f \lambda(c) \\
E(Y_i \mid J_i = h) &= \mu_h + \tau_h \lambda(-c) \\
\operatorname{var}(Y_i \mid J_i = f) &= \sigma_f^2 + \tau_f^2(-\lambda(c)c - \lambda^2(c)) \\
\operatorname{var}(Y_i \mid J_i = h) &= \sigma_h^2 + \tau_h^2(\lambda(-c)c - \lambda^2(-c)) \\
E([Y_i - E(Y_i \mid J_i = f)]^3 \mid J_i = f) &= \tau_f^3 \lambda(c)[2\lambda^2(c) + 3c\lambda(c) + c^2 - 1] \\
E([Y_i - E(Y_i \mid J_i = h)]^3 \mid J_i = h) &= \tau_h^3 \lambda(-c)[2\lambda^2(-c) - 3c\lambda(-c) + c^2 - 1].
\end{aligned}$$

This gives us seven equations in the five unknowns $\mu_f, \mu_h, \sigma_f^2, \sigma_h^2$, and σ_{fh}. It is straightforward to show that the five parameters can be identified from this system of equations.

However, Theorems 7 and 8 of Heckman and Honoré (1990) show that when one relaxes the log normality assumption, without exclusion restrictions in the outcome equation, the model is no longer identified. This is true despite the strong assumption of agent income maximization. This result is not particularly surprising in the sense that our goal is to estimate a full joint distribution of a two dimensional object (Y_{fi}, Y_{hi}), but all we can observe is two one dimensional distributions (wages conditional on job choice). Since there is no information in the data about the wage that a fisherman may have received as a hunter, one cannot identify this joint distribution. In fact, Theorem 7 of Heckman and Honoré (1990) states that we can never distinguish the actual model from an alternative model in which skills are independent of each other.

3.1. Estimation of the normal linear labor supply model

It is often the case that we only observe wages in one sector. For example, when estimating models of participation in the labor force, the wage is observed only if the individual works. We can map this into our model by associating working with "fishing" and not working with "hunting." That is, we let Y_{fi} denote income if working and let Y_{hi} denote the value of not working.[5]

But there are other examples in which we observe the wage in only one sector. For example, in many data sets we do not observe wages of workers in the black market sector. Another example is return immigration in which we know when a worker leaves the data to return to their home country, but we do not observe that wage.

In Section 3.2 we discuss identification of the nonparametric version of the model. However, it turns out that identification of the more complicated model is quite similar to estimation of the model with normally distributed errors. Thus we review this in detail before discussing the nonparametric model. We also remark that providing a consistent estimator also provides a constructive proof of identification, so one can also interpret these results as (informally) showing identification in the normal model. The model is similar to Willis and Rosen's (1979) Roy Model of educational choices or Lee's (1978) model of union status and the empirical approach is analogous. We assume that

$$\begin{aligned} Y_{fi} &= X'_{fi}\gamma_{ff} + X'_{0i}\gamma_{0f} + \varepsilon_{fi} \\ Y_{hi} &= X'_{hi}\gamma_{hh} + X'_{0i}\gamma_{0h} + \varepsilon_{hi} \\ \begin{bmatrix} \varepsilon_{fi} \\ \varepsilon_{hi} \end{bmatrix} &= N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_f^2 & \sigma_{fh} \\ \sigma_{fh} & \sigma_h^2 \end{bmatrix} \right). \end{aligned}$$

In a labor supply model where f represents market work, Y_{fi} is the market wage which will be observed for workers only. Y_{hi}, the pecuniary value of not working, is never

[5] There are two common participation models. The first is the home production model in which the individual chooses between home and market production. The second is the labor supply model in which the individual chooses between market production and leisure. In practice the two types of models tend to be similar and some might argue the distinction is semantic. In a model of home production, Y_{hi} is the (unobserved) gain from home production. In a model of labor supply, Y_{hi} is the leisure value of not working.

observed in the data. Keane et al. (2011) example of the static model of a married woman's labor force participation is similar.

One could simply estimate this model by maximum likelihood. However we discuss a more traditional four step method to illustrate how the parametric model is identified. This four step process will be analogous to the more complicated nonparametric identification below. Step 1 is a "reduced form probit" of occupational choices as a function of all covariates in the model. Step 2 estimates the wage equations by controlling for selection as in the second step of a Heckman Two step (Heckman, 1979). Step 3 uses the coefficients of the wage equations and plugs these back into a probit equation to estimate a "structural probit." Step 4 shows identification of the remaining elements of the variance-covariance matrix of the residuals.

Step 1: Estimation of choice model

The probability of choosing fishing (i.e., work) is:

$$
\begin{aligned}
\Pr\left(J_i = f \mid X_i = x\right) &= \Pr\left(Y_{fi} > Y_{hi} \mid X_i = x\right) \\
&= \Pr\left(x_f'\gamma_{ff} + x_0'\gamma_{0f} + \varepsilon_{fi} > x_0'\gamma_{0h} + x_h'\gamma_{hh} + \varepsilon_{hi}\right) \\
&= \Pr\left(x_f'\gamma_{ff} - x_h'\gamma_{hh} + x_0'\left(\gamma_{0f} - \gamma_{0h}\right) > \varepsilon_{hi} - \varepsilon_{fi}\right) \\
&= \Phi\left(\frac{x_f'\gamma_{ff} - x_h'\gamma_{hh} + x_0'\left(\gamma_{0f} - \gamma_{0h}\right)}{\sigma^*}\right) \\
&= \Phi\left(x'\gamma^*\right) \qquad (3.5)
\end{aligned}
$$

where Φ is the cdf of a standard normal, σ^* is the standard deviation of $\left(\varepsilon_{hi} - \varepsilon_{fi}\right)$, and

$$
\gamma^* \equiv \left(\frac{\gamma_{ff}}{\sigma^*}, \frac{-\gamma_{hh}}{\sigma^*}, \frac{\gamma_{0f} - \gamma_{0h}}{\sigma^*}\right).
$$

This is referred to as the "reduced form model" as it is a reduced form in the classical sense: the parameters are a known function of the underlying structural parameters. It can be estimated by maximum likelihood as a probit model. Let $\widehat{\gamma^*}$ represent the estimated parameter vector. This is all that can be learned from the choice data alone. We need further information to identify σ^* and to separate γ_{0f} from γ_{0h}.

Step 2: Estimating the wage equation

This is essentially the second stage of a Heckman (1979) two step. To review the idea behind it, let

$$
\varepsilon_i^* = \frac{\varepsilon_{hi} - \varepsilon_{fi}}{\sigma^*}.
$$

Then consider the regression

$$\varepsilon_{fi} = \tau \varepsilon_i^* + \zeta_i$$

where $\operatorname{cov}\left(\varepsilon_i^*, \zeta_i\right) = 0$ (by definition of regression) and thus:

$$\begin{aligned}\tau &= \frac{\operatorname{cov}\left(\varepsilon_{fi}, \varepsilon_i^*\right)}{\operatorname{var}\left(\varepsilon_i^*\right)} \\ &= E\left[\varepsilon_{fi}\left(\frac{\varepsilon_{hi} - \varepsilon_{fi}}{\sigma^*}\right)\right] \\ &= \frac{\sigma_{fh} - \sigma_f^2}{\sigma^*}.\end{aligned}$$

The wage of those who choose to work is

$$\begin{aligned}E\left(Y_{fi} \mid J_i = f, X_i = x\right) &= x_f'\gamma_{ff} + x_0'\gamma_{0f} + E\left(\varepsilon_{fi} \mid J_i = f, X_i = x\right) \\ &= x_f'\gamma_{ff} + x_0'\gamma_{0f} + E\left(\tau\varepsilon_i^* + \zeta_i \mid \varepsilon_i^* \leq x'\gamma^*\right) \\ &= x_f'\gamma_{ff} + x_0'\gamma_{0f} + \tau E\left(\varepsilon_i^* \mid \varepsilon_i^* \leq x'\gamma^*\right) \\ &= x_f'\gamma_{ff} + x_0'\gamma_{0f} - \tau\lambda\left(x'\gamma^*\right). \qquad (3.6)\end{aligned}$$

Showing that $E\left(\varepsilon_i^* \mid \varepsilon_i^* \leq x'\gamma^*\right) = -\lambda\left(x'\gamma^*\right)$ is a fairly straightforward integration problem and is well known. Because Eq. (3.6) is a conditional expectation function, OLS regression of Y_i on X_{0i}, X_{fi}, and $\lambda\left(X_i'\widehat{\gamma^*}\right)$ gives consistent estimates of γ_{ff}, γ_{0f}, and τ. $\widehat{\gamma^*}$ is the value of γ^* estimated in Eq. (3.5).

Note that we do not require an exclusion restriction. Since λ is a nonlinear function, but g_f is linear, this model is identified. However, without an exclusion restriction, identification is purely through functional form. When we consider a nonparametric version of the model below, exclusion restrictions are necessary. We discuss this issue in Section 3.2.

Step 3: The structural probit

Our next goal is to estimate γ_{0h} and γ_{hh}. In Step 1 we obtained consistent estimates of $\gamma^* \equiv \left(\frac{\gamma_{0f}-\gamma_{0h}}{\sigma^*}, \frac{\gamma_{ff}}{\sigma^*}, \frac{-\gamma_{hh}}{\sigma^*}\right)$ and in Step 2 we obtained consistent estimates of γ_{0f} and γ_{ff}.

When there is only one exclusion restriction (i.e. γ_{ff} is a scalar), identification proceeds as follows. Because we identified γ_{ff} in Step 2 and γ_{ff}/σ^* in Step 1, we can identify σ^*. Once σ^* is identified, it is easy to see how to identify γ_{hh} (because $\frac{-\gamma_{hh}}{\sigma^*}$ is identified) and γ_{0h} (because $\frac{\gamma_{0f}-\gamma_{0h}}{\sigma^*}$ and γ_{0f} are identified).

In terms of estimation of these objects, if there is more than one exclusion restriction the model is over-identified. If we have two exclusion restrictions, γ_{ff} and γ_{ff}/σ^* are both 2×1 vectors, and thus we wind up with 2 consistent estimates of σ^*. The most standard way of solving this model is by estimating the "structural probit:"

$$\Pr(J_i = f \mid X_i = x) = \Phi\left(\frac{1}{\sigma^*}\left(x'_f\widehat{\gamma_{ff}} + x'_0\widehat{\gamma_{0f}}\right) - x'_h\frac{\gamma_{hh}}{\sigma^*} - x'_0\frac{\gamma_{0h}}{\sigma^*}\right). \quad (3.7)$$

That is, one just runs a probit of J_i on $(X'_{fi}\widehat{\gamma_{ff}} + X'_{0i}\widehat{\gamma_{0f}})$, X_{0i}, and X_{hi} where $\widehat{\gamma_{ff}}$ and $\widehat{\gamma_{0f}}$ are our estimates of γ_{ff} and γ_{0f}.

Step 3 is essential if our goal is to estimate the labor supply equation. If we are only interested in controlling for selection to obtain consistent estimates of the wage equation, we do not need to worry about the structural probit. However, notice that

$$\frac{\partial \Pr(J_i = f \mid X_i = x)}{\partial Y_{fi}} = \frac{1}{\sigma^*}\phi\left(x'\gamma^*\right).$$

and thus the labor supply elasticity is:

$$\begin{aligned}\frac{\partial \log[\Pr(J_i = f \mid X_i = x)]}{\partial Y_{fi}} &= \frac{\partial \Pr(J_i = f \mid X_i = x)}{\partial Y_{fi}}\frac{1}{\Pr(J_i = f \mid X_i = x)}\\ &= \frac{1}{\sigma^*}\frac{\phi\left(x'\gamma^*\right)}{\Phi\left(x'\gamma^*\right)},\end{aligned}$$

where, as before, Y_{fi} is the log of income if working. Thus knowledge of σ^* is essential for identifying the effects of wages on participation.

One could not estimate the structural probit without the exclusion restriction X_{fi} as the first two components of the probit in Eq. (3.7) would be perfectly collinear. For any $\sigma^* > 0$ we could find a value of γ_{0h} and γ_{hh} that delivers the same choice probabilities. Furthermore, if these parameters were not identified, the elasticity of labor supply with respect to wages would not be identified either.

Step 4: Estimation of the variance matrix of the residuals

Lastly, we identify all the components of Σ, $(\sigma_f^2, \sigma_h^2, \sigma_{fh})$ as follows. We have described how to obtain consistent estimates of $\sigma^* = \sqrt{\sigma_f^2 + \sigma_h^2 - 2\sigma_{fh}}$ and $\tau = \frac{\sigma_{fh} - \sigma_f^2}{\sigma^*}$. This gives us two equations in three parameters. We can obtain the final equation by using the variance of the residual in the selection model since

$$\text{var}(\varepsilon_{fi} \mid J_i = f, X_i = x) = \sigma_f^2 + \tau^2\left[-\lambda(x'\gamma^*)x'\gamma^* - \lambda^2(x'\gamma^*)\right].$$

Let $i = 1, \ldots, N_f$ index the set of individuals who choose $J_i = f$ and $\widehat{\varepsilon_{fi}}$ is the residual $Y_{fi} - X'_{fi}\widehat{\gamma_{ff}} - X'_{0i}\widehat{\gamma_{0f}}$ for individuals who choose $J_i = f$. Using "hats" to denote estimators we can estimate the model as

$$\widehat{\sigma_f^2} = \frac{1}{N_f}\sum_{i=1}^{N_f}\left(\widehat{\varepsilon_{fi}} + \tau\lambda\left(X_i'\widehat{\gamma^*}\right)\right)^2 - \widehat{\tau}^2\left(-\lambda\left(X_i'\widehat{\gamma^*}\right)X_i'\widehat{\gamma^*} - \lambda^2\left(X_i'\widehat{\gamma^*}\right)\right)$$

$$\widehat{\sigma_{fh}} = \widehat{\sigma_f^2} - \widehat{\tau}\widehat{\sigma^*}$$

$$\widehat{\sigma_h}^2 = \widehat{\sigma^*}^2 - \widehat{\sigma_f}^2 + 2\widehat{\sigma_{fh}}.$$

3.2. Identification of the Roy model: the non-parametric approach

Although the parametric case with exclusion restrictions is more commonly known, the model in the previous section is still identified non-parametrically if the researcher is willing to impose stronger support conditions on the observable variables. Heckman and Honoré (1990, Theorem 12) provide conditions under which one can identify the model nonparametrically using exclusion restrictions. We present this case below.

Assumption 3.1. $(\varepsilon_{fi}, \varepsilon_{hi})$ is continuously distributed with distribution function G, support $\mathbb{R}^2$, and is independent of X_i. The marginal distributions of ε_{fi} and $\varepsilon_{fi} - \varepsilon_{hi}$ have medians equal to zero.

Assumption 3.2. $\text{supp}(g_f(X_{fi}, x_0), g_h(X_{hi}, x_0)) = \mathbb{R}^2$ for all $x_0 \in \text{supp}(X_{0i})$.

Assumption 3.2 is crucial for identification. It states that for any value of $g_h(x_h, x_0)$, $g_f(X_{fi}, x_0)$ varies across the full real line and for any value of $g_f(x_f, x_0)$, $g_h(X_{hi}, x_0)$ varies across the full real line. This means that we can condition on a set of variables for which the probability of being a hunter (i.e. $\Pr(J_i = h | X_i = x)$) is arbitrarily close to 1. This is clearly a very strong assumption that we will discuss further.

We need the following two assumptions for the reasons discussed in Section 2.4.

Assumption 3.3. $X_i = (X_{fi}, X_{hi}, X_{0i})$ can be written as $(X^c_{fi}, X^d_{fi}, X^c_{hi}, X^d_{hi}, X^c_{0i}, X^d_{0i})$ where the elements of $(X^c_{fi}, X^c_{hi}, X^c_{0i})$ are continuously distributed (no point has positive mass), and $(X^d_{fi}, X^d_{hi}, X^d_{0i})$ is distributed discretely (all support points have positive mass).

Assumption 3.4. For any $(x^d_f, x^d_h, x^d_0) \in \text{supp}(X^d_{fi}, X^d_{hi}, X^d_{0i})$, $g_f(x^c_f, x^d_f, x^c_0, x^d_0)$ and $g_h(x^c_h, x^d_h, x^c_0, x^d_0)$ are almost surely continuous across $x^c \in \text{supp}(X^c_i \mid X^d_i = x^d)$.

Under these assumptions we can prove the theorem following Heckman and Honoré (1990).

Theorem 3.1. *If $(J_i \in \{f, h\}$, Y_{fi} if $J_i = f$, $X_i)$ are all observed and generated under model (3.1)–(3.4), then under Assumptions 3.1–3.4, g_f, g_h, and G are identified on a set $\mathcal{X}^*$ that has measure 1.*

(Proof in Appendix.)

A key theme of this chapter is that the basic structure of identification in this model is similar to identification of more general selection models, so we explain this result in much detail. The basic structure of the proof we present below is similar to Heckman and Honoré's proof of their Theorems 10 and 12. We modify the proof to allow for the case where Y_{hi} is not observed.

The proof in the Appendix is more precise, but in the text we present the basic ideas. We follow a structure analogous to the parametric empirical approach when the residuals are normally distributed as presented in Section 3.1. First we consider identification of the occupational choice given only observable covariates and the choice model. This is the nonparametric analogue of the reduced form probit. Second we estimate g_f given the data on Y_{fi}, which is the analogue of the second stage of the Heckman two step, and is more broadly the nonparametric version of the classical selection model. In the third step we consider the nonparametric analogue of identification of the structural probit. Since we will have already established identification of g_f, identification of this part of the model boils down to identification of g_h. Finally in the fourth step we consider identification of G (the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$). We discuss each of these steps in order.

To map the Roy model into our formal definition of identification presented in Section 2.2, the model is determined by $\theta = (g_f, g_h, G, F_x)$, where F_x is the joint distribution of (X_{fi}, X_{hi}, X_{0i}). The observable data here is $(X_{fi}, X_{hi}, X_{0i}, J_i, 1(J_i = f)Y_{fi})$. Thus P is the joint distribution of this observable data and $\Theta(P)$ represents the possible data generating processes consistent with P.

Step 1: Identification of choice model

The nonparametric identification of this model is established in Matzkin (1992). We can write the model as

$$\begin{aligned}\Pr(J_i = f \mid X_i = x) &= \Pr(\varepsilon_{hi} - \varepsilon_{fi} < g_f(x_f, x_0) - g_h(x_h, x_0)) \\ &= G_{h-f}(g_f(x_f, x_0) - g_h(x_h, x_0)),\end{aligned}$$

where G_{h-f} is the distribution function of $\varepsilon_{hi} - \varepsilon_{fi}$.

Using data only on choices, this model is only identified up to a monotonic transformation. To see why, note that we can write $J_i = f$ when

$$g_f(x_f, x_0) - g_h(x_h, x_0) > \varepsilon_{hi} - \varepsilon_{fi} \tag{3.8}$$

but this is equivalent to the condition

$$M(g_f(x_f, x_0) - g_h(x_h, x_0)) > M(\varepsilon_{hi} - \varepsilon_{fi}) \tag{3.9}$$

where $M(.)$ is any strictly increasing function. Clearly the model in Eq. (3.8) cannot be distinguished from an alternative model in Eq. (3.9). This is the nonparametric analog of the problem that the scale (i.e., the variance of $\varepsilon_{hi} - \varepsilon_{fi}$) and location (only the difference between $g_f(x_f, x_0)$ and $g_h(x_h, x_0)$ but not the level of either) of the parametric binary choice model are not identified. Without loss of generality we can normalize the model up to a monotonic transformation. There are many ways to do this. A very convenient normalization is to choose the transformation $M(\cdot) = G_{h-f}(\cdot)$ because $G_{h-f}\left(\varepsilon_{hi} - \varepsilon_{fi}\right)$ has a uniform distribution.[6] So we define

$$\begin{aligned}\varepsilon_i &\equiv G_{h-f}(\varepsilon_{hi} - \varepsilon_{fi}) \\ g(x) &\equiv G_{h-f}(g_f(x_f, x_0) - g_h(x_h, x_0)).\end{aligned}$$

Then

$$\begin{aligned}\Pr(J_i = f \mid X_i = x) &= \Pr(g_f(x_f, x_0) - g_h(x_h, x_0) > \varepsilon_{hi} - \varepsilon_{fi}) \\ &= \Pr(G_{h-f}(g_f(x_f, x_0) - g_h(x_h, x_0)) > G_{h-f}(\varepsilon_{hi} - \varepsilon_{fi})) \\ &= \Pr(\varepsilon_i < g(x)) \\ &= g(x).\end{aligned}$$

Thus we have established that we can (i) write the model as $J_i = f$ if and only if $g(X_i) > \varepsilon_i$ where ε_i is uniform [0, 1] and (ii) that g is identified.

This argument can be mapped into our formal definition of identification from Section 2.2 above. The goal here is identification of g, so we define $\Psi(\theta) = g$. Note that even though g is not part of θ, it is a known function of the components of θ. The key set now is $\Psi(\Theta(P))$, which is now defined as the set of possible values g that could have generated the joint distribution of $(X_{fi}, X_{hi}, X_{0i}, J_i, 1(J_i = f)Y_{fi})$. Since $\Pr(J_i = f \mid X_i = x) = g(x)$, no other possible value of g could generate the data. Thus $\Psi(\Theta(P))$ only contains the true value and is thus a singleton.

Step 2: Identification of the wage equation g_f

Next consider identification of $g_{f.}$ Median regression identifies

$$\text{Med}(Y_i \mid X_i = x, J_i = f) = g_f(x_f, x_0) + \text{Med}(\varepsilon_{fi} \mid X_i = x, \varepsilon_i < g(x)).$$

The goal is to identify $g_f(x_f, x_0)$. The problem is that when we vary (x_f, x_0) we also typically vary $\text{Med}(\varepsilon_{fi} \mid X_i = x, g(x) > \varepsilon_i)$. This is the standard selection problem. Because we can add any constant to g_f and subtract it from ε_{fi} without changing the model, a normalization that allows us to pin down the location of g_f is that $\text{Med}(\varepsilon_{fi}) = 0$. The problem is that this is the unconditional median rather than

[6] To see why note that for any x,$\Pr(G_{h-f}\left(\varepsilon_{hi} - \varepsilon_{fi}\right) < x) = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq G^{-1}_{h-f}(x)) = G_{h-f}(G^{-1}_{h-f}(x)) = x$.

the conditional one. The solution here is what is often referred to as identification at infinity (e.g. Chamberlain, 1986, or Heckman, 1990). For some value $\left(x_f, x_0\right)$ suppose we can find a value of x_h to send $\Pr(\varepsilon_i < g(x))$ arbitrarily close to one. It is referred to as identification at infinity because if g_h were linear in the exclusion restriction x_h this could be achieved by sending $x_h \rightarrow -\infty$. In our fishing/hunting example, this could be sending the price of rabbits to zero which in turn sends log income from hunting to $-\infty$. Then notice that[7]

$$\begin{aligned}\lim_{g(x)\rightarrow 1} \operatorname{Med}(Y_i \mid X_i = x, J_i = f) &= g_f(x_f, x_0) + \lim_{g(x)\rightarrow 1} \operatorname{Med}(\varepsilon_{fi} \mid \varepsilon_i \leq g(x)) \\ &= g_f(x_f, x_0) + \operatorname{Med}(\varepsilon_{fi} \mid \varepsilon_i \leq 1) \\ &= g_f(x_f, x_0) + \operatorname{Med}(\varepsilon_{fi}) \\ &= g_f(x_f, x_0).\end{aligned}$$

Thus g_f is identified.

Conditioning on x so that $\Pr(J_i = 1 \mid X_i = x)$ is arbitrarily close to one is essentially conditioning on a group of individuals for whom there is no selection, and thus there is no selection problem. Thus we are essentially saying that if we can condition on a group of people for whom there is no selection we can solve the selection bias problem.

While this may seem like cheating, without strong functional form assumptions it is necessary for identification. To see why, suppose there is some upper bound of $\operatorname{supp}[g(X_i)]$ equal to $g^u < 1$ which would prevent us from using this type of argument. Consider any potential worker with a value of $\varepsilon_i > g^u$. For those individuals it must be the case that

$$\varepsilon_i > g(X_i)$$

so they must always be a hunter. As a result, the data is completely uninformative about the distribution of ε_{fi} for these individuals. For this reason the unconditional median of ε_{fi} would not be identified. We will discuss approaches to dealing with this potential problem in the Treatment Effect section below.

To relate this to the framework from Section 2.2 above now we define $\Psi(\theta) = g_f$, so $\Psi(\Theta(P))$ contains the values of g_f consistent with P. However since

$$\lim_{g(x)\rightarrow \infty} \operatorname{Med}(Y_f \mid X_i = x, J_i = f) = g_f(x_f, x_0),$$

g_f is the only element of $\Psi(\Theta(P))$, thus it is identified.

[7] We are using loose notation here. What we mean by $lim_{g(x)\rightarrow 1}$ is to hold (x_f, x_0) fixed, but take a sequence of values of x_h so that $g(x) \rightarrow 1$.

Identification of the slope only without “identification at infinity”

If one is only interested in identifying the “slope” of g_f and not the intercept, one can avoid using an identification at infinity argument. That is, for any two points (x_f, x_0) and $(\tilde{x}_f, \tilde{x}_0)$, consider identifying the difference $g_f(x_f, x_0) - g_f(\tilde{x}_f, \tilde{x}_0)$. The key to identification is the existence of the exclusion restriction X_{hi}. For these two points, suppose we can find values x_h and $\tilde{x}_h$ so that

$$g(x_f, x_h, x_0) = g(\tilde{x}_f, \tilde{x}_h, \tilde{x}_0).$$

There may be many pairs of $(x_h, \tilde{x}_h)$ that satisfy this equality and we could choose any of them. Define $\tilde{x} \equiv (\tilde{x}_f, \tilde{x}_h, \tilde{x}_0)$. The key aspect of this is that since $g(x) = g(\tilde{x})$, and thus the probability of being a fisherman is the same given the two sets of points, then the bias terms are also the same: $\text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(x)) = \text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(\tilde{x}))$.

This allows us to write

$$\begin{aligned}
&\text{Med}(Y_i \mid X_i = x, J_i = f) - \text{Med}(Y_i \mid X_i = \tilde{x}, J_i = f) \\
&\quad = g_f(x_f, x_0) + \text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(x)) \\
&\qquad - [g_f(\tilde{x}_f, \tilde{x}_0) + \text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(\tilde{x}))] \\
&\quad = g_f(x_f, x_0) - g_f(\tilde{x}_f, \tilde{x}_0).
\end{aligned}$$

As long as we have sufficient variation in X_{hi} we can do this everywhere and identify g_f up to location.

Step 3: Identification of g_h

In terms of identifying g_h, the exclusion restriction that influences wages as a fisherman but not as a hunter (i.e. X_{fi}) will be crucial. Consider identifying $g_h(x_h, x_0)$ for any particular value (x_h, x_0). The key here is finding a value of x_f so that

$$\Pr(J_i = f \mid X_i = (x_f, x_h, x_0)) = 0.5. \tag{3.10}$$

Assumption 3.2 guarantees that we can do this. To see why Eq. (3.10) is useful, note that it must be that for this value of (x_f, x_h, x_0)

$$0.5 = \Pr\left(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(x_f, x_0) - g_h(x_h, x_0)\right). \tag{3.11}$$

But the fact that $\varepsilon_{hi} - \varepsilon_{fi}$ has median zero implies that

$$g_h(x_h, x_0) = g_f(x_f, x_0).$$

Since g_f is identified, g_h is identified from this expression.[8]

Again to relate this to the framework in Section 2.2 above, now $\Psi(\theta) = g_h$ and $\Psi(\Theta(p))$ is the set of functions g_h that are consistent with P. Above we showed that if $\Pr(J_i = f \mid X_i = x) = 0.5$, then $g_h(x_h, x_0) = g_f(x_f, x_0)$. Thus since we already showed that g_f is identified, g_h is the only element of $\Psi(\Theta(p))$.

Step 4: Identification of G

Next consider identification of G given g_f and g_h. We will show how to identify the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$ closely following the exposition of Heckman and Taber (2008). Note that from the data one can observe

$$\begin{aligned}
&\Pr(J_i = f, Y_{fi} < s \mid X_i = x) \\
&\quad = \Pr(g_h(x_h, x_0) + \varepsilon_{hi} \leq g_f(x_f, x_0) + \varepsilon_{fi}, g_f(x_f, x_0) + \varepsilon_{fi} \leq s) \\
&\quad = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(x_f, x_0) - g_h(x_h, x_0), \varepsilon_{fi} \leq s - g_f(x_f, x_0)) \quad (3.12)
\end{aligned}$$

which is the cumulative distribution function of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$ evaluated at the point $(g_f(x_f, x_0) - g_h(x_h, x_0), s - g_f(x_f, x_0))$. By varying the point of evaluation one can identify the joint distribution of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$ from which one can derive the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$.

Finally in terms of the identification conditions in Section 2.2 above, now $\Psi(\theta) = G$ and $\Psi(\Theta(P))$ is the set of distributions G consistent with P. Since G is uniquely defined by the expression (3.12) and since everything else in this expression is identified, G is the only element of $\Psi(\Theta(P))$.

3.3. Relaxing independence between observables and unobservables

For expositional purposes we focus on the case in which the observables are independent of the unobservables, but relaxing these assumptions is easy to do. The simplest case is to allow for a general relationship between X_{0i} and $(\varepsilon_{fi}, \varepsilon_{hi})$. To see how easy this is, consider a case in which X_{0i} is just binary, for example denoting men and women. Independence seems like a very strong assumption in this case. For example, the distribution of unobserved preferences might be different for women and men, leading to different selection patterns. In order to allow for this, we could identify and estimate the Roy model separately for men and for women. Expanding from binary X_{0i} to finite support X_{0i} is trivial, and going beyond that to continuous X_{0i} is straightforward. Thus one can

[8] Note that Heckman and Honoré (1990) choose a different normalization. Rather than normalizing the median of $\varepsilon_{hi} - \varepsilon_{fi}$ to zero (which is convenient in the case in which Y_{hi} is not observed) they normalize the median of ε_{hi} to zero (which is more convenient in their case). Since this is just a normalization, it is innocuous. After identifying the model under our normalization we could go back to redefine the model in terms of theirs.

relax the independence assumption easily. But for expositional purposes we prefer our specification.

The distinction between X_{fi} and X_{0i} was not important in steps 1 and 2 of our discussion above. When one is only interested in the outcome equation $Y_{fi} = g_f(X_{fi}, X_{0i}) + \varepsilon_{fi}$, relaxing the independence assumption between X_{fi} and $(\varepsilon_{fi}, \varepsilon_{hi})$ can be done as well. However, in step 3 this distinction is important in identifying g_h and the independence assumption is not easy to relax.

If we allow for general dependence between X_{0i} and $(\varepsilon_{fi}, \varepsilon_{hi})$, the "identification at infinity" argument becomes more important as the argument about "Identification of the Slope Only without Identification at Infinity" no longer goes through. In that case the crucial feature of the model was that $\text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(x)) = \text{Med}(\varepsilon_{fi} \mid \varepsilon_i < g(\tilde{x}))$. However, without independence this is no longer generally true because $\text{Med}(\varepsilon_{fi} \mid X_i = x, J_i = f) = \text{Med}(\varepsilon_{fi} \mid X_{0i} = x_0, \varepsilon_i < g(x))$. Thus even if $g(x) = g(\tilde{x})$, when $x_0 \neq \tilde{x}_0$, in general $\text{Med}(\varepsilon_{fi} \mid X_{0i} = x_0, \varepsilon_i < g(x)) \neq \text{Med}(\varepsilon_{fi} \mid X_{0i} = \tilde{x}_0, \varepsilon_i < g(\tilde{x}))$.

3.4. The importance of exclusion restrictions

We now show that the model is not identified in general without an exclusion restriction.[9] Consider a simplified version of the model,

$$J_i = \begin{cases} f & \text{if } g(X_i) - \varepsilon_i \geq 0 \\ h & \text{otherwise} \end{cases}$$
$$Y_{fi} = g_f(X_i) + \varepsilon_{fi}$$

where ε_i is uniform (0,1) and $(\varepsilon_i, \varepsilon_{fi})$ is independent of X_i with distribution G and we use the location normalization $\text{Med}(\varepsilon_{fi} \mid X_i) = 0$. As in Section 3.2, we observe X_i, whether $J_i = f$ or h, and if $J_i = f$ then we observe Y_{fi}.

We can think about estimating the model from the median regression

$$\begin{aligned} \text{Med}[Y_{fi} | X_i = x] &= g_f(X_i) + \text{Med}[\varepsilon_{fi} | X_i = x] \\ &= g_f(X_i) + \text{Med}[\varepsilon_{fi} | g(X_i) > \varepsilon_i] \\ &= g_f(X_i) + h(g(X_i)). \end{aligned} \tag{3.13}$$

Under the assumption that $\text{Med}(\varepsilon_{fi} \mid X_i) = 0$ it must be the case that $h(1) = 0$, but this is our only restriction on h and g. Thus the model above has the same conditional

[9] An exception is Buera (2006), who allows for general functional forms and does not need an exclusion restriction. Assuming wages are observed in both sectors, and making stronger use of the independence assumption between the observables and the unobservables, he shows that the model can be identified without exclusion restrictions.

median as an alternative model

$$\text{Med}[Y_{fi}|X_i = x] = \widetilde{g}_f(X_i) + \widetilde{h}(g(X_i)) \tag{3.14}$$

where $\widetilde{g}_f(X_i) = g_f(X_i) + k(g(X_i))$ and $\widetilde{h}(g(X_i)) = h(g(X_i)) - k(g(X_i))$. Equations (3.13) and (3.14) are observationally equivalent. Without an exclusion restriction, it is impossible to tell if observed income from working varies with X_i because it varies with g_f or because it varies with the labor force participation rate and thus the extent of selection. Thus the models in Eqs (3.13) and (3.14) are not distinguishable using conditional medians.

To show the two models are indistinguishable using the full joint distribution of the data, consider an alternative data generating model with the same first stage, but now Y_{fi} is determined by

$$Y_{fi} = \widetilde{g}_f(X_i) + \widetilde{\varepsilon}_{fi}$$

where $\widetilde{\varepsilon}_{fi}$ is independent of X_i with $\text{Med}(\widetilde{\varepsilon}_{fi} \mid X_i) = 0$. Let $\widetilde{G}(\varepsilon_i, \widetilde{\varepsilon}_{fi})$ be the joint distribution of $(\varepsilon_i, \widetilde{\varepsilon}_{fi})$ in the alternative model. We will continue to assume that in the alternative model $\widetilde{g}_f(X_i) = g_f(X_i) + k(g(X_i))$. The question is whether the alternative model is able to generate the same data distribution.

In the true model

$$\begin{aligned}\Pr(\varepsilon_i \leq g(x), Y_{fi} < y) &= \Pr(\varepsilon_i \leq g(x), g_f(x) + \varepsilon_{fi} \leq y) \\ &= G(g(x), y - g_f(x)).\end{aligned}$$

In the alternative model

$$\begin{aligned}\Pr(\varepsilon_i \leq g(x), Y_{fi} < y) &= \Pr(\varepsilon_i \leq g(x), \widetilde{g}_f(x) + \widetilde{\varepsilon}_{fi} \leq y) \\ &= \widetilde{G}(g(x), y - \widetilde{g_f}(x)).\end{aligned}$$

Thus these two models generate exactly the same joint distribution of data and cannot be separately identified as long as we define $\widetilde{G}$ so that[10]

$$\begin{aligned}\widetilde{G}(g(x), y - \widetilde{g}_f(x)) &= G(g(x), y - g_f(x)) \\ &= G(g(x), y - \widetilde{g}_f(x) + k(g(x))).\end{aligned}$$

[10] One cannot do this with complete freedom as one needs $\widetilde{G}$ to be a legitimate cdf. That is, it must be nondecreasing in both of its arguments. However, there will typically be many examples of k for which $\widetilde{G}$ is a cdf and the model is not identified. For example, if k is a nondecreasing function, $\widetilde{G}$ will be a legitimate cdf.

4. THE GENERALIZED ROY MODEL

We next consider the "Generalized Roy Model" (as defined in e.g. (Heckman and Vytlacil, 2007a). The basic Roy model assumes that workers only care about their income. The Generalized Roy Model allows workers to care about non-pecuniary aspects of the job as well. Let U_{fi} and U_{hi} be the utility that individual i would receive from being a fisherman or a hunter respectively, where for $j \in \{f, h\}$,

$$U_{ji} = Y_{ji} + \varphi_j(Z_i, X_{0i}) + \nu_{ji}. \tag{4.1}$$

where $\varphi_j(Z_i, X_{0i})$ represents the non-pecuniary utility gain from observables Z_i and X_{0i}. The variable Z_i allows for the fact that there may be other variables that affect the taste for hunting versus fishing directly, but do not affect wages in either sector.[11] Note that we are imposing separability between Y_{ji} and φ_j. In general we can provide conditions in which the results presented here will go through if we relax this assumption, but we impose it for expositional simplicity. The occupation is now defined as

$$J_i = \begin{cases} f & \text{if } U_{fi} > U_{hi} \\ h & \text{if } U_{fi} \leq U_{hi}. \end{cases} \tag{4.2}$$

We continue to assume that

$$\begin{aligned} Y_{fi} &= g_f(X_{fi}, X_{0i}) + \varepsilon_{fi} \\ Y_{hi} &= g_h(X_{hi}, X_{0i}) + \varepsilon_{hi} \end{aligned} \tag{4.3}$$

$$Y_i = \begin{cases} Y_{fi} & \text{if } J_i = f \\ Y_{hi} & \text{if } J_i = h. \end{cases} \tag{4.4}$$

It will be useful to define a reduced form version of this model. Note that people fish when

$$\begin{aligned} 0 &< U_{fi} - U_{hi} \\ &= (Y_{fi} + \varphi_f(Z_i, X_{0i}) + \nu_{fi}) - (Y_{hi} + \varphi_h(Z_i, X_{0i}) + \nu_{hi}) \\ &= g_f(X_{fi}, X_{0i}) + \varphi_f(Z_i, X_{0i}) - g_h(X_{hi}, X_{0i}) - \varphi_h(Z_i, X_{0i}) \\ &\quad + \varepsilon_{fi} + \nu_{fi} - \varepsilon_{hi} - \nu_{hi}. \end{aligned}$$

In the previous section we described how the choice model can only be identified up to a monotonic transform and that assuming the error term is uniform is a convenient

[11] In principle some of the elements of Z_i may affect φ_f and others may affect φ_h, but this distinction will not be important here, so we use the most general notation.

normalization. We do the same thing here. Let F^* be the distribution function of $\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}$. Then we define

$$\nu_i \equiv F^*\left(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}\right) \tag{4.5}$$

$$\varphi(Z_i, X_i) \equiv F^*(g_f(X_{fi}, X_{0i}) + \varphi_f(Z_i, X_{0i}) - g_h(X_{hi}, X_{0i}) - \varphi_h(Z_i, X_{0i})). \tag{4.6}$$

As above, this normalization is convenient because it is straightforward to show that

$$J_i = f \quad \text{when } \varphi(Z_i, X_i) > \nu_i$$

and that ν_i is uniformly distributed on the unit interval.

We assume that the econometrician can observe the occupations of the workers and the wages that they receive in their chosen occupations as well as (X_i, Z_i).

4.1. Identification

It turns out that the basic assumptions that allow us to identify the Roy model also allow us to identify the generalized Roy model.

We start with the reduced form model in which we need two more assumptions.

Assumption 4.1. $(\nu_i, \varepsilon_{fi}, \varepsilon_{hi})$ is continuously distributed and is independent of (Z_i, X_i). Furthermore, ν_i is distributed uniform on the unit interval and the medians of both ε_{fi} and ε_{hi} are zero.

Assumption 4.2. The support of $\varphi(Z_i, x)$ is $[0, 1]$ for all $x \in \text{supp}(X_i)$.

We also slightly extend the restrictions on the functions to include φ_f and φ_h.

Assumption 4.3. $(Z_i, X_i) = (Z_i, X_{fi}, X_{hi}, X_{0i})$ can be written as $(Z_i^c, Z_i^d, X_{fi}^c, X_{fi}^d, X_{hi}^c, X_{hi}^d, X_{0i}^c, X_{0i}^d)$ where the elements of $(Z_i^c, X_{fi}^c, X_{hi}^c, X_{0i}^c)$ are continuously distributed (no point has positive mass), and $(Z_i^d, X_{fi}^d, X_{hi}^d, X_{0i}^d)$ are distributed discretely (all support points have positive mass).

Assumption 4.4. For any $(z^d, x_f^d, x_h^d, x_0^d) \in \text{supp}(Z_i^d, X_{fi}^d, X_{hi}^d, X_{0i}^d)$, $g_f(x_f^c, x_f^d, x_0^c, x_0^d)$, $g_h(x_h^c, x_h^d, x_0^c, x_0^d)$, $\varphi_f(z^c, z^d, x_0^c, x_0^d)$ and $\varphi_h(z^c, z^d, x_0^c, x_0^d)$ are almost surely continuous across

$$(z^c, x^c) \in \text{supp}(Z_i^c, X_i^c \mid (Z_i^d, X_i^d) = (z^d, x^d)).$$

Theorem 4.1. *Under Assumptions 4.1–4.4, φ, g_f, g_h and the joint distribution of $(\nu_i, \varepsilon_{fi})$ and of $(\nu_i, \varepsilon_{hi})$ are identified from the joint distribution of (J_i, Y_i) on a set $\mathcal{X}^*$ that has measure 1 where (J_i, Y_i) are generated by model (4.1)–(4.4).*

(Proof in Appendix.)

The intuition for identification follows directly from the intuition given for the basic Roy model. We show this in 3 steps:

1. Identification of φ is like the "Step 1: identification of choice model" section. We can only identify φ up to a monotonic transformation for exactly the same reason given in that section. We impose the normalization that ν_i is uniform in Assumption 4.2. Given that assumption
$$\Pr(J_i = f \mid Z_i = z, X_i = x) = \varphi(z, x)$$
so identification of φ from $\Pr(J_i = f \mid Z_i = z, X_i = x)$ comes directly.
2. Identification of g_f and g_h are completely analogous to "Step 2: identification of g_f" in Section 3.2. That is
$$\begin{aligned}&\lim_{\varphi(z,x)\to 1} \text{Med}(Y_i \mid Z_i = z, X_i = x, J_i = f)\\ &= g_f(x_f, x_0) + \lim_{\varphi(z,x)\to 1} \text{Med}(\varepsilon_{fi} \mid Z_i = z, X_i = x, J_i = f)\\ &= g_f(x_f, x_0) + \lim_{\varphi(z,x)\to 1} \text{Med}(\varepsilon_{fi} \mid \nu_i \le \varphi(z, x))\\ &= g_f(x_f, x_0) + \text{Med}(\varepsilon_{fi})\\ &= g_f(x_f, x_0).\end{aligned}$$
The analogous argument works for g_h when we send $\varphi(z, x) \to 0$.
3. Identification of the joint distribution of $(\nu_i, \varepsilon_{fi})$ and of $(\nu_i, \varepsilon_{hi})$ are analogous to the "Step 4: identification of G" discussion in the Roy model. That is if we let G_{ν,ε_f} represent the joint distribution of $(\nu_i, \varepsilon_{fi})$ then
$$\begin{aligned}&\Pr(J_i = f, Y_{fi} \le y \mid (Z_i, X_i) = (z, x))\\ &\quad = \Pr(\nu_i \le \varphi(z, x), g_f(x_f, x_0) + \varepsilon_{fi} \le y)\\ &\quad = G_{\nu,\varepsilon_f}(\varphi(z, x), y - g_f(x_f, x_0)).\end{aligned}$$
The analogous argument works for the joint distribution of $(\nu_i, \varepsilon_{hi})$.

Note that not all parameters are identified such as the non-pecuniary gain from fishing $\varphi_f - \varphi_h$. To identify the "structural" generalized Roy model we make two additional assumptions:

Assumption 4.5. The median of $\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}$ is zero.

Assumption 4.6. For any value of $(z, x_0) \in \text{supp}(Z_i, X_{0i})$, $g_f(X_{fi}, x_0) - g_h(X_{hi}, x_0)$ has full support (i.e. the whole real line).

Theorem 4.2. *Under Assumptions 4.1–4.6, $\varphi_f - \varphi_h$, the distribution of $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{fi})$, and the distribution of $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{hi})$ are identified.*

(Proof in Appendix.)

Note that Theorem 4.1 gives the joint distribution of $(\nu_i, \varepsilon_{fi})$ while Theorem 4.2 gives the joint distribution of $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{fi})$. Since $\nu_i = F^*(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi})$, this really just amounts to saying that F^* is identified.

Furthermore, whereas g_f and g_h are identified in Theorem 4.1, $\varphi_f - \varphi_h$ is identified in Theorem 4.2. Recall $\varphi_f - \varphi_h$ is the added utility (measured in money) of being a fisherman relative to a hunter. The exclusion restrictions X_{fi} and X_{hi} help us identify this. These exclusion restrictions allow us to vary the pecuniary gains of the two sectors, holding preferences $\varphi_f - \varphi_h$ constant. Identification is analogous to the "Step 3: identification of g_h" in the standard Roy model. To see where identification comes from, for every (z, x_0) think about the following conditional median

$$\begin{aligned} 0.5 &= \Pr(J_i = f \mid Z_i = z, X_i = x) \\ &= \Pr(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi} \leq g_f(x_f, x_0) + \varphi_f(z, x_0) - g_h(x_h, x_0) - \varphi_h(z, x_0)). \end{aligned}$$

Since the median of $\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}$ is zero, this means that

$$g_f(x_f, x_0) + \varphi_f(z, x_0) - g_h(x_h, x_0) - \varphi_h(z, x_0) = 0,$$

and thus

$$\varphi_f(z, x_0) - \varphi_h(z, x_0) = g_h(x_h, x_0) - g_f(x_f, x_0).$$

Because g_f and g_h is identified, $\varphi_f - \varphi_h$ is identified also. The argument above shows that we do not need both X_{fi} and X_{hi}, we only need X_{fi} or X_{hi}.

Suppose there is no variable that affects earnings in one sector but not preferences (X_{fi} or X_{hi}). An alternative way to identify $\varphi_f - \varphi_h$ is to use a cost measured in dollars. Consider the linear version of the model with normal errors and without exclusion restrictions (X_{hi}, X_{fi}) so that

$$\begin{aligned} g_h(x_0) &= x_{0i}'\gamma_h \\ g_f(x_0) &= x_{0i}'\gamma_f \\ \varphi_f(z, x_0) - \varphi_h(z, x_0) &= x_0'\beta_0 + z'\beta_z. \end{aligned}$$

The reduced form probit is:

$$\Pr(J_i = f \mid Z_i = z, X_i = x) = \Phi\left(x_{0i}' \frac{\gamma_f - \gamma_h + \beta_0}{\sigma} + z_i' \frac{\beta_z}{\sigma}\right)$$

where σ is the standard deviation of $\varepsilon_{hi}+\nu_{hi}-\varepsilon_{fi}-\nu_{fi}$. Theorem 4.1 above establishes that the functions g_f and g_h (i.e., γ_f and γ_h) as well as the variance of ε_{hi} and ε_{fi} are identified. We still need to identify β_0, β_z and σ. Thus we are able to identify

$$\frac{\gamma_f-\gamma_h+\beta_0}{\sigma} \quad \text{and} \quad \frac{\beta_z}{\sigma}.$$

If β_0 and β_z are scalars we still have three parameters $(\beta_0, \beta_z, \sigma)$ and two restrictions $(\frac{\gamma_f-\gamma_h+\beta_0}{\sigma}, \frac{\beta_z}{\sigma})$. If they are not scalars, we still have one more parameter than restriction. However suppose that one of the exclusion restrictions represents a cost variable that is measured in the same units as $Y_{fi}-Y_{hi}$. For example in a schooling case suppose that Y_{fi} represents the present value of earnings as a college graduate, Y_{hi} represents the present value of high school graduate as a college graduate, and the exclusion restriction, Z_i, represents the present value of college tuition. In this case $\beta_z=-1$ the coefficient on Z_i is $-1/\sigma$, so σ is identified. Given σ it is very easy to show that the rest of the parameters are identified as well. Heckman et al. (1998) provide an example of this argument using tuition as in the style above. In Section 7.3 we discuss Heckman and Navarro (2007) who use this approach as well.

4.2. Lack of identification of the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$

In pointing out what is identified in the model it is also important to point out what is not identified. Most importantly in the generalized Roy model we were able to identify the joint distribution between the error terms in the selection equation and each of the outcomes, but not the joint distribution of the variables in the outcome equation. In particular the joint distribution between the error terms $(\varepsilon_{fi}, \varepsilon_{hi})$ is not identified. Even strong functional form assumptions will not solve this problem. Fir example, it is easy to show that in the joint normal model the covariance of $(\varepsilon_{fi}, \varepsilon_{hi})$ is not identified.

4.3. Are functional forms innocuous? Evidence from Catholic schools

As the theorems above make clear, nonparametric identification requires exclusion restrictions. However, completely parametric models typically do not require exclusion restrictions. In specific empirical examples, identification could primarily be coming from the exclusion restriction or identification could be coming primarily from the functional form assumptions (or some combination between the two). When researchers use exclusion restrictions in data, it is important to be careful about which assumptions are important.

We describe one example from Altonji et al. (2005b). Based on Evans and Schwab (1995), Neal (1997), and Neal and Grogger (2000) they consider a bivariate probit model of Catholic schooling and college attendance.

$$CH_i = 1(X_i'\beta + \lambda Z_i + u_i > 0) \tag{4.7}$$

$$Y_i = 1(\alpha CH_i + X_i'\gamma + \varepsilon_i > 0), \tag{4.8}$$

where $1(\cdot)$ is the indicator function taking the value one if its argument is true and zero otherwise, CH_i is a dummy variable indicating attendance at a Catholic school, and Y_i is a dummy variable indicating college attendance. Identification of the effect of Catholic schooling on college attendance (or high school graduation) is the primary focus of these studies. The question at hand is in practice whether the assumed functional forms for u_i and ε_i are important for identifying the α coefficient and thus the effect of Catholic schools on college attendance.

The model in Eqs (4.7) and (4.8) is a minor extension of the generalized Roy model. The first key difference is that the outcome variable in Eq. (4.8) is binary (attend college or not), whereas in the case of the Generalized Roy model the outcomes were continuous (earnings in either sector). The second key difference is that the outcome equation for Catholic versus Non-Catholic school only differs in the intercept (α). The error term (ε_i) and the slope coefficients (γ) are restricted to be the same. Nevertheless, the machinery to prove non-parametric identification of the Generalized Roy model can be applied to this framework.[12]

Using data from the National Longitudinal Survey of 1972, Altonji et al. (2005b) consider an array of instruments and different specifications for Eqs (4.7) and (4.8). In Table 1 we present a subset of their results. We show four different models. The "Single Equation Model" gives results in which selection into Catholic school is not accounted for. The first column gives results from a probit model (with point estimates, standard errors, and marginal effects). The second column give results from a Linear Probability model. Next we present the estimates of α from a Bivariate Probit models with alternative exclusion restrictions. The final row presents the results with no exclusion restrictions. Finally we also present results from an instrumental variable linear probability model with the same set of exclusion restrictions.

One can see that the marginal effect from the single equation probit is very similar to the OLS estimate. It indicates that college attendance rates are approximately 23.9 percentage points higher for Catholic high school graduates than for public high school graduates. The rest of the table presents results from three bivariate probit models and two instrumental variables models using alternative exclusion restrictions. The problem is clearest when the interaction between the student coming from a Catholic school and distance to the nearest Catholic school is used as an instrument. The 2SLS gives nonsensical results: a coefficient of 2.572 with an enormous standard error. This indicates that the instrument has little power. However, the bivariate probit result is more reasonable. It suggests that the true marginal causal effect is around 0.478 and the point

[12] Following Matzkin (1992), we need a monotonic normalization on the outcome model (such as assuming the error term is uniform). Once we have done this, proving identification of this model is almost identical to the generalized Roy model and is easily done with an exclusion restriction with sufficient support.

Table 1 Estimated effects of Catholic schools on college attendance from linear and nonlinear models.

Single equation models

	Probit	OLS
	0.239	0.239
	[0.640]	
	(0.198)	(0.070)

Two equation models

Excluded variable	Bivariate probit	2SLS
Catholic	0.285	−0.093
	[0.761]	
	(0.543)	(0.324)
Catholic × Distance	0.478	2.572
	[1.333]	
	(0.516)	(2.442)
None	0.446	
	[1.224]	
	(0.542)	

Urban Non-Whites from NLS-72.
The first set of results come from simple probits and from OLS.
The further results come from Bivariate Probits and from two stage least squares.
We present the marginal effect of Catholic high school attendance on college attendance.
[Point Estimate from Probit in Brackets.]
(Standard Errors in Parentheses.)
Source: Altonji et al. (2005b).

estimate is statistically significant. This seems inconsistent with the 2SLS results which indicated that this exclusion restriction had very little power. However it is clear what is going on when we compare this result to the model at the bottom of the table without an exclusion restriction. The estimate is very similar with a similar standard error. The linearity and normality assumptions drive the results.

The case in which Catholic religion by itself is used as an instrument is less problematic. The IV result suggests a strong amount of positive selection but still yields a large standard error. The bivariate probit model suggests a marginal effect that is a bit larger than the OLS effect. However, note that the standard errors for the model with and without an exclusion restriction are quite similar, which seems inconsistent with the idea that the exclusion restriction is providing a lot of identifying information. Further note that the IV result suggests a strong positive selection bias while the bivariate probit without exclusion restrictions suggests a strong negative bias. The bivariate probit in which Catholic is excluded is somewhere between the two. This suggests that both functional form and exclusion restrictions are important in this case. We should emphasize the "suggests" part of this sentence as none of this is a formal test. It does, however, make one wonder how much trust to put in the bivariate probit results by themselves.

Another paper documenting the importance of functional form assumptions is Das et al. (2003), who estimate the return to education for young Australian women. They estimate equations for years of education, the probability of working, and wages. When estimating the wage equation they address both the endogeneity of years of education and also selection caused because we only observe wages for workers. They allow for flexibility in the returns to education (where the return depends on years of education) and also in the distribution of the residuals. They find that when they assume normality of the error terms, the return to education is approximately 12%, regardless of years of education. However, once they allow for more flexible functional forms for the error terms, they find that the returns to education decline sharply with years of education. For example, they find that at 10 years of education, the return to education is over 15%. However, at 14 years, the return to education is only about 5%.

5. TREATMENT EFFECTS

There is a very large literature on the estimation of treatment effects. For more complete summaries see Heckman and Robb (1986), Heckman et al. (1999), Heckman and Vytlacil (2007a,b), Abbring and Heckman (2007), or Imbens and Wooldridge (2009).[13] DiNardo and Lee (2011) provide a discussion that is complementary to ours. Our goal in this section is not to survey the whole literature but provide a brief summary and to put it into the context of identification of the Generalized Roy Model.

The goal of this literature is to estimate the value of receiving a treatment defined as:

$$\pi_i = Y_{fi} - Y_{hi}. \tag{5.1}$$

In the context of the Roy model, π_i is the income gain from moving from hunting to fishing. This income gain potentially varies across individuals in the population. Thus for people who choose to be fishermen, π_i is positive and for people who choose to be hunters, π_i is negative.

Estimation of treatment effects is of great interest in many literatures. The term "treatment effect"makes the most sense in the context of the medical literature. Choice f could represent taking a medical treatment (such as an experimental drug) while h could represent no treatment. In that case Y_{fi} and Y_{hi} would represent some measure of health status for individual i with and without the treatment. Thus the treatment effect π_i is the effect of the drug on the health outcome for individual i.

[13] There is also a substantial literature on the tradeoffs between different empirical approaches. Key papers include Leamer (1983), Heckman (1979, 1999, 2000), Angrist and Imbens (1999), Rosenzweig and Wolpin (2000), Deaton (2009), Heckman and Urzúa (2010), Imbens (2009), Angrist and Pischke (2010) and Sims (2010).

The classic example in labor economics is job training. In that case, Y_{fi} would represent a labor market outcome for individuals who received training and Y_{hi} would represents the outcome in the absence of training.

In both the case of drug treatment and job training, empirical researchers have exploited randomized trials. Medical patients are often randomly assigned either a treatment or a placebo (i.e., a sugar pill that should have no effect on health). Likewise, many job training programs are randomly assigned. For example, in the case of the Job Training Partnership Act, a large number of unemployed individuals applied for job training (see e.g. Bloom et al., 1997). Of those who applied for training, some were assigned training and some were assigned no training.

Because assignment is random and affects the level of treatment, one can treat assignment as an exclusion restriction that is correlated with treatment (i.e., the probability that $J_i = f$) but is uncorrelated with preferences or ability because it is random. In this sense, random assignment solves the selection problem that is the focus of the Roy model. As we show below, exogenous variation provided by experiments allows the researcher to cleanly identify some properties of the distribution of Y_{fi} and Y_{hi} under relatively weak assumptions. Furthermore, the methods for estimating these objects are simple, which adds to their appeal.

The treatment effect framework is also widely used for evaluating quasi-experimental data as well. By quasi-experimental data, we mean data that are not experimental, but exploit variation that is "almost as good as" random assignment.

5.1. Treatment effects and the generalized Roy model

Within the context of the generalized Roy model note that in general

$$\pi_i = g_f(X_{fi}, X_{0i}) - g_h(X_{hi}, X_{0i}) + \varepsilon_{fi} - \varepsilon_{hi}.$$

An important special case of the treatment effect defined in Eq. (5.1) is when

$$g_f(X_{fi}, X_{0i}) = g_h(X_{hi}, X_{0i}) + \pi_0 \tag{5.2}$$

$$\varepsilon_{fi} = \varepsilon_{hi}. \tag{5.3}$$

In this case, the treatment effect $\pi_i = Y_{fi} - Y_{hi} = \pi_0$ is a constant across individuals. Identification of this parameter is relatively straightforward. However, there is a substantial literature that studies identification of heterogeneous treatment effects. As we point out above, treatment effects are positive for some people and negative for others in the context of the Roy model. Furthermore, there is ample empirical evidence that the returns to job training are not constant, but instead vary across the population (Heckman et al., 1999).

In Section 4.2 we explain why the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$ is not identified. This means that the distribution of π_i is not identified and even relatively simple summary statistics like the median of this distribution is not identified in general. The key problem is that even when assignment is random, we do not observe the same people in both occupations.

Since the full generalized Roy model is complicated, hard to describe, and very demanding in terms of data, researchers often focus on a summary statistic to summarize the result. The most common in this literature is the Average Treatment Effect (ATE) defined as

$$\begin{aligned}\text{ATE} &\equiv E(\pi_i)\\ &= E(Y_{fi}) - E(Y_{hi}).\end{aligned}$$

From Theorem 4.1 we know that (under the assumptions of that theorem) the distribution of Y_{fi} and Y_{hi} are identified. Thus, their expected values are also identified under the one additional assumption that these expected values exist.

Assumption 5.1. The expected values of Y_{fi} and Y_{hi} are finite.

Theorem 5.1. *Under the assumptions of Theorem 4.1 and Assumption 5.1, the Average Treatment effect is identified.*

(Proof in Appendix.)

To see where identification of this object comes from, abstract from X_i so that the only observable is Z_i, which affects the non-pecuniary gain in utility from occupation across occupations. With experimental data, Z_i could be randomly generated assignments to occupation. Notice that

$$\begin{aligned}&\lim_{\varphi(z)\to 1} E(Y_{fi} \mid Z_i = z, J_i = f) - \lim_{\varphi(z)\to 0} E(Y_{hi} \mid Z_i = z, J_i = h)\\ &\quad = \lim_{\varphi(z)\to 1} E(Y_{fi} \mid \nu_i \le \varphi(z)) - \lim_{\varphi(z)\to 0} E(Y_{hi} \mid \nu_i > \varphi(z))\\ &\quad = E(Y_{fi}) - E(Y_{hi}).\end{aligned}$$

Thus the exclusion restriction is the key to identification. Note also that we need groups of individuals where $\varphi(Z_i) \approx 1$ (who are always fishermen) and $\varphi(Z_i) \approx 0$ (who are always hunters); thus "identification at infinity" is essential as well. For the reasons discussed in the nonparametric Roy model above, if $\varphi(Z_i)$ were never higher than some $\varphi(z^u) < 1$ then $E(Y_{fi})$ would not be identified. Similarly if $\varphi(Z_i)$ were never lower than some $\varphi(z^\ell) > 0$, then $E(Y_{hi})$ would not be identified.

While one could directly estimate the ATE using "identification at infinity", as described above, this is not the common practice and not something we would advocate.

The standard approach would be to estimate the full Generalized Roy Model and then use it to simulate the various treatment effects. This is often done using a completely parametric approach as in, for example, the classic paper by Willis and Rosen (1979). However, there are quite a few nonparametric alternatives as well, including construction of the Marginal Treatment effects as discussed in Sections 5.3 and 5.4 below.

As it turns out, even with experimental data, it is rarely the case that $\varphi(Z_i)$ is identically one or zero with positive probability. In the case of medicine, some people assigned the treatment do not take the treatment. In the training example, many people who are offered subsidized training decide not to undergo the training. Thus, when compliance with assignment is less than 100%, we cannot recover the ATE. In Section 5.2 we discuss more precisely what we do recover when there is less than 100% compliance.

It is also instructive to relate the ATE to instrumental variables estimation. Let Y_i be the outcome of interest

$$Y_i = \begin{cases} Y_{fi} & \text{if } J_i = f \\ Y_{hi} & \text{if } J_i = h, \end{cases}$$

and let D_{fi} be a dummy variable indicating whether $J_i = f$. Consider estimating the model

$$Y_i = \beta_0 + \beta_1 D_{fi} + u_i \tag{5.4}$$

using instrumental variables with Z_i as an instrument for D_{fi}. Assume that Z_i is correlated with D_{fi} but not with Y_{fi} or Y_{hi}. Consider first the constant treatment effect model described in Eqs (5.2) and (5.3) so that $\pi_i = \pi_0$ for everyone in the population. In that case

$$\begin{aligned} Y_i &= Y_{fi} D_{fi} + Y_{hi}(1 - D_{fi}) \\ &= Y_{hi} + D_{fi}(Y_{fi} - Y_{hi}) \\ &= Y_{hi} + D_{fi}\pi_0. \end{aligned}$$

Then two stage least squares on the model above yields

$$\begin{aligned} \operatorname{plim} \widehat{\beta}_1 &= \frac{\operatorname{cov}(Z_i, Y_i)}{\operatorname{cov}(Z_i, D_{fi})} \\ &= \frac{\operatorname{cov}(Z_i, Y_{hi} + D_{fi}\pi_0)}{\operatorname{cov}(Z_i, D_{fi})} \\ &= \frac{\operatorname{cov}(Z_i, Y_{hi})}{\operatorname{cov}(Z_i, D_{fi})} + \frac{\operatorname{cov}(Z_i, \pi_0 D_{fi})}{\operatorname{cov}(Z_i, D_{fi})} \\ &= \pi_0. \end{aligned}$$

Thus in the constant treatment effect model, instrumental variables provide a consistent estimate of the treatment effect. However, this result does not carry over to heterogeneous treatment effects or the average treatment effects as Heckman (1997) shows. Following the expression above we get

$$\begin{aligned}\operatorname{plim} \widehat{\beta}_1 &= \frac{\operatorname{cov}(Z_i, Y_{hi} + D_{fi}\pi_i)}{\operatorname{cov}(Z_i, D_{fi})}\\ &= \frac{\operatorname{cov}(Z_i, D_{fi}\pi_i)}{\operatorname{cov}(Z_i, D_{fi})}\\ &\neq \text{ATE}\end{aligned} \tag{5.5}$$

in general. In Sections 5.2 and 5.3 below, we describe what instrumental variables identify.

In practice there are two potential problems with the assumptions behind Theorem 5.1 above

- The researcher may not have a valid exclusion restriction. We discuss some of the options for this case in Sections 5.5–5.7.
- Even if they do, the variable may not have full support. By this we mean that the instrumental variable Z_i may not vary enough, so that for some observed values of Z_i everyone is always a fisherman and for other observed values of Z_i everyone is always a hunter. We discuss what can be identified using exclusion restrictions with limited support in Sections 5.2–5.4 and 5.6.

We discuss a number of different approaches, some of which assume an exclusion restriction but relax the support conditions and others that do not require exclusion restrictions.

5.2. Local average treatment effects

Imbens and Angrist (1994) and Angrist et al. (1996) consider identification when the support of Z_i takes on a finite number of points. They show that when varying the instrument over this range, they can identify what they call a Local Average Treatment Effect. Furthermore, they show how instrumental variables can be used to estimate it. It is again easiest to think about this problem after abstracting from X_i, as it is straightforward to condition on these variables (see Imbens and Angrist, 1994, for details). For simplicity's sake, consider the case in which the instrument Z_i is binary and takes on the values $\{0, 1\}$. In many cases not only is the instrument discrete, but it is also binary. For example, in randomized medical trials, $Z_i = 1$ represents assignment to treatment, whereas $Z_i = 0$ represents assignment to the placebo. In job training programs, $Z_i = 1$ represents assignment to the training program, whereas $Z_i = 0$ represents no assigned training.

It is important to point out that not all patients assigned treatment actually receive the treatment. Thus $J_i = f$ if the patient actually takes the drug and $J_i = h$ if the individual does not take the drug. Likewise, not all individuals who are assigned training actually receive the training, so $J_i = f$ if the individual goes to training and $J_i = h$ if she does not. The literature on Local Average Treatment Effects handles this case as well as many others. However, we do require that the instrument of assignment has power: $\Pr(J_i = f \mid Z_i = 1) \neq \Pr(J_i = f \mid Z_i = 0)$. Without loss of generality we will assume that $\Pr(J_i = f \mid Z_i = 1) > \Pr(J_i = f \mid Z_i = 0)$.

Using the reduced form version of the generalized Roy model the choice problem is

$$J_i = f \quad \text{if}\, \varphi(Z_i) > \nu_i \tag{5.6}$$

where ν_i is uniformly distributed.

The following six objects can be learned directly from the data:

$$\begin{aligned}
&\Pr(J_i = f|Z_i = 0) = \Pr(\nu_i \leq \varphi(0))\\
&\Pr(J_i = f|Z_i = 1) = \Pr(\nu_i \leq \varphi(1))\\
&E(Y_{fi} \mid Z_i = 0, J_i = f) = E(Y_{fi} \mid \nu_i \leq \varphi(0))\\
&E(Y_{hi} \mid Z_i = 0, J_i = h) = E(Y_{hi} \mid \nu_i > \varphi(0))\\
&E(Y_{fi} \mid Z_i = 1, J_i = f) = E(Y_{fi} \mid \nu_i \leq \varphi(1))\\
&E(Y_{hi} \mid Z_i = 1, J_i = h) = E(Y_{hi} \mid \nu_i > \varphi(1)).
\end{aligned}$$

The above equations show that our earlier assumption that $\Pr(J_i = f|Z_i = 1) > \Pr(J_i = f|Z_i = 0)$ implies $\Pr(\nu_i \leq \varphi(1)) > \Pr(\nu_i \leq \varphi(0))$. This, combined with the structure embedded in Eq. (5.6) means that

$$\Pr(\nu_i \leq \varphi(1)|\nu_i \leq \varphi(0)) = 1, \tag{5.7}$$

so then an individual who is a fisherman when $Z_i = 0$ is also a fisherman when $Z_i = 1$. Similar reasoning implies $\Pr(\nu_i \leq \varphi(1)|\varphi(0) < \nu_i \leq \varphi(1)) = 1$. Using this and Bayes rule yields

$$\begin{aligned}
\Pr(\nu_i \leq \varphi(0) \mid \nu_i \leq \varphi(1)) &= \frac{\Pr(\nu_i \leq \varphi(1) \mid \nu_i \leq \varphi(0)) \Pr(\nu_i \leq \varphi(0))}{\Pr(\nu_i \leq \varphi(1))}\\
&= \frac{\Pr(\nu_i \leq \varphi(0))}{\Pr(\nu_i \leq \varphi(1))},
\end{aligned} \tag{5.8}$$

$$
\begin{aligned}
&\Pr(\varphi(0) < \nu_i \leq \varphi(1) \mid \nu_i \leq \varphi(1)) \\
&\quad = \frac{\Pr(\nu_i \leq \varphi(1) \mid \varphi(0) < \nu_i \leq \varphi(1)) \Pr(\varphi(0) < \nu_i \leq \varphi(1))}{\Pr(\nu_i \leq \varphi(1))} \\
&\quad = \frac{\Pr(\varphi(0) < \nu_i \leq \varphi(1))}{\Pr(\nu_i \leq \varphi(1))}.
\end{aligned} \tag{5.9}
$$

Using the fact that $\Pr(\nu_i \leq \varphi(1)) = \Pr(\nu_i \leq \varphi(0)) + \Pr(\varphi(0) < \nu_i \leq \varphi(1))$, one can show that

$$
\begin{aligned}
E(Y_{fi} \mid \nu_i \leq \varphi(1)) &= E(Y_{fi} \mid \nu_i \leq \varphi(0)) \Pr(\nu_i \leq \varphi(0) \mid \nu_i \leq \varphi(1)) \\
&\quad + E(Y_{fi} \mid \varphi(0) < \nu_i \leq \varphi(1)) \Pr(\varphi(0) < \nu_i \leq \varphi(1) \mid \nu_i \leq \varphi(1)).
\end{aligned} \tag{5.10}
$$

Combining Eq. (5.10) with Eqs (5.8) and (5.9) yields

$$
\begin{aligned}
E(Y_{fi} \mid \nu_i \leq \varphi(1)) &= \frac{E(Y_{fi} \mid \nu_i \leq \varphi(0)) \Pr(\nu_i \leq \varphi(0))}{\Pr(\nu_i \leq \varphi(1))} \\
&\quad + \frac{E(Y_{fi} \mid \varphi(0) < \nu_i \leq \varphi(1)) \Pr(\varphi(0) < \nu_i \leq \varphi(1))}{\Pr(\nu_i \leq \varphi(1))}.
\end{aligned} \tag{5.11}
$$

Rearranging Eq. (5.11) shows that we can identify

$$
\begin{aligned}
&E(Y_{fi} \mid \varphi\,(0) \leq \nu_i < \varphi(1)) \\
&= \frac{E(Y_{fi} \mid Z_i = 1, J_i = f) \Pr(J_i = f \mid Z_i = 1) - E(Y_{fi} \mid Z_i = 0, J_i = f) \Pr(J_i = f \mid Z_i = 0)}{\Pr(J_i = f \mid Z_i = 1) - \Pr(J_i = f \mid Z_i = 0)}
\end{aligned} \tag{5.12}
$$

since everything on the right hand side is directly identified from the data.

Using the analogous argument one can show that

$$
\begin{aligned}
&E(Y_{hi} \mid \varphi\,(0) \leq \nu_i < \varphi(1)) \\
&= \frac{E(Y_{hi} \mid Z_i = 0, J_i = h) \Pr(J_i = h \mid Z_i = 0) - E(Y_{hi} \mid Z_i = 1, J_i = h) \Pr(J_i = h \mid Z_i = 1)}{\Pr(J_i = f \mid Z_i = 1) - \Pr(J_i = f \mid Z_i = 0)}
\end{aligned}
$$

is identified. But this means that we can identify

$$
E(\pi_i \mid \varphi\,(0) \leq \nu_i < \varphi(1)) = E(Y_{fi} - Y_{hi} \mid \varphi\,(0) \leq \nu_i < \varphi(1)) \tag{5.13}
$$

which Imbens and Angrist (1994) define as the Local Average Treatment Effect. This is the average treatment effect for that group of individuals who would alter their treatment status if their value of Z_i changed. Given the variation in Z_i, this is the only group for whom we can identify a treatment effect. Any individual in the data with $\nu_i > \varphi(1)$

would never choose $J_i = f$, so the data are silent about $E(Y_{fi} \mid \nu_i > \varphi(1))$. Similarly the data is silent about $E(Y_{hi} \mid \nu_i \leq \varphi(0))$.

Imbens and Angrist (1994) also show that the standard linear Instrumental Variables estimator yield consistent estimates of Local Average Treatment Effects. Consider the instrumental variables estimator of Eq. (5.4)

$$Y_i = \beta_0 + \beta_1 D_{fi} + u_i.$$

In Eq. (5.5) we showed that

$$\begin{aligned}\widehat{\beta}_1 &\xrightarrow{p} \frac{\operatorname{cov}(Z_i, D_{fi}\pi_i)}{\operatorname{cov}(Z_i, D_{fi})} \\ &= \frac{E(\pi_i D_{fi} Z_i) - E\left(\pi_i D_{fi}\right) E\left(Z_i\right)}{E(D_{fi} Z_i) - E\left(D_{fi}\right) E\left(Z_i\right)}.\end{aligned}$$

Let P_z denote the probability that $Z_i = 1$. The numerator of the above expression is

$$\begin{aligned}&E(\pi_i D_{fi} Z_i) - E(\pi_i D_{fi}) E\left(Z_i\right) \\ &\quad = P_z E(\pi_i D_{fi} \mid Z_i = 1) - E\left(\pi_i D_{fi}\right) P_z \\ &\quad = P_z E(\pi_i D_{fi} \mid Z_i = 1) \\ &\qquad - \left[P_z E(\pi_i D_{fi} \mid Z_i = 1) + (1 - P_z)\, E(\pi_i, D_{fi} \mid Z_i = 0)\right] P_z \\ &\quad = P_z(1 - P_z)\left[E(\pi_i D_{fi} \mid Z_i = 1) - E(\pi_i D_{fi} \mid Z_i = 0)\right] \\ &\quad = P_z(1 - P_z) E(\pi_i \mid \varphi(0) < \nu_i \leq \varphi\left(1\right)) \Pr(\varphi\left(0\right) < \nu_i \leq \varphi\left(1\right))\end{aligned}$$

where the key simplification comes from the fact that

$$\begin{aligned}E(\pi_i D_{fi} \mid Z_i = 1) &= E\left(\pi_i 1\left(\nu_i \leq \varphi(1)\right)\right) \\ &= E\left(\pi_i \left[1\left(\nu_i \leq \varphi(0)\right) + 1\left(\varphi(0) < \nu_i \leq \varphi(1)\right)\right]\right) \\ &= E(\pi_i D_{fi} \mid Z_i = 0) \\ &\quad + E(\pi_i \mid \varphi(0) < \nu_i \leq \varphi\left(1\right)) \Pr(\varphi\left(0\right) < \nu_i \leq \varphi\left(1\right)).\end{aligned}$$

Next consider the denominator

$$\begin{aligned}&E(D_{fi} Z_i) - E(D_{fi}) E\left(Z_i\right) \\ &\quad = P_z E(D_{fi} \mid Z_i = 1) - E(D_{fi}) P_z \\ &\quad = P_z E(D_{fi} \mid Z_i = 1) - \left[P_z E(D_{fi} \mid Z_i = 1) + (1 - P_z)\, E(D_{fi} \mid Z_i = 0)\right] P_z\end{aligned}$$

$$= P_z(1 - P_z)\left[E(D_{fi} \mid Z_i = 1) - E(D_{fi} \mid Z_i = 0)\right]$$
$$= P_z(1 - P_z)\Pr(\varphi(0) < \nu_i \le \varphi(1)).$$

Thus

$$\begin{aligned}\widehat{\beta}_1 &\xrightarrow{p} \frac{E(\pi_i D_{fi} Z_i) - E(\pi_i D_{fi})E(Z_i)}{E(D_{fi} Z_i) - E(D_{fi})E(Z_i)} \\ &= \frac{P_z(1 - P_z)E(\pi_i \mid \varphi(0) < \nu_i \le \varphi(1))\Pr(\varphi(0) < \nu_i \le \varphi(1))}{P_z(1 - P_z)\Pr(\varphi(0) < \nu_i \le \varphi(1))} \\ &= E(\pi_i \mid \varphi(0) < \nu_i \le \varphi(1)).\end{aligned}$$

Imbens and Angrist never explicitly use the generalized Roy model or the latent index framework. Instead, they write their problem only in terms of the choice probabilities. However, in order to do this they must make one additional assumption. Specifically, they assume that if $J_i = f$ when $Z_i = 0$ then $J_i = f$ when $Z_i = 1$. Thus changing $Z_i = 0$ to $Z_i = 1$ never causes some people to switch from fishing to hunting. It only causes people to switch from hunting to fishing. They refer to this as a monotonicity assumption. Vytlacil (2002) points out that this is implied by the latent index model when the index $\varphi(Z_i)$ is separable from ν_i, as we assumed in Eq. (5.6). As is implied by Eq. (5.7), increasing the index $\varphi(Z_i)$ will cause some people to switch from hunting to fishing, but not the reverse.[14]

Throughout, we use the latent index framework that is embedded in the Generalized Roy model, for three reasons. First, we can appeal to the identification results of the Generalized Roy model. Second, the latent index can be interpreted as the added utility from making a decision. Thus we can use the estimated model for welfare analysis. Third, placing the choice in an optimizing framework allows us to test the restrictions on choice that come from the theory of optimization.

As we have pointed out, not everyone offered training actually takes the training. For example, in the case of the JTPA, only 60% of those offered the training actually received it (Bloom et al., 1997). Presumably, those who took the training are those who stood the most to gain from the training. For example, the reason that many people do not take training is that they receive a job offer before training begins. For these people, the training may have been of relatively little value. Furthermore, 2% of those who applied for and were not assigned training program wind up receiving the training (Bloom et al., 1997). Angrist et al. (1996) refer to those who were assigned training, but did not take the training as *never-takers*. Those who receive the training whether or not

[14] However, he points out that the non-separable model $D_{fi} = 1(f(Z_i, \nu_i) > 0)$ does not necessarily give rise to monotonicity. All other differences between the latent variable framework and the LATE framework are extremely technical and minor.

they are assigned are *always-takers*. Those who receive the training only when assigned the training are *compliers*. In terms of the latent index framework, the never-takers are those for whom ($\nu_i \geq \varphi(1)$), the compliers are those for whom ($\varphi(0) \leq \nu_i < \varphi(1)$), and the always-takers are those for whom ($\nu_i < \varphi(0)$).

The monotonicity assumption embedded in the latent index framework rules out the existence of a final group: the *defiers*. In the context of training, this would be an individual who receives training when not assigned training but would not receive training when assigned. At least in the context of training programs (and many other contexts) it seems safe to assume that there are no defiers.

5.3. Marginal treatment effects

Heckman and Vytlacil (1999, 2001, 2005, 2007b) develop a framework that is useful for constructing many types of treatment effects. They focus on the marginal treatment effect (MTE) defined in our context as

$$\Delta^{\text{MTE}}(x, \nu) \equiv E(\pi_i \mid X_i = x, \nu_i = \nu).$$

They show formally how to identify this object. We present their methodology using our notation.

Note that if we allow for regressors X_i, let the exclusion restriction Z_i to take on values beyond zero and one, then if (z^ℓ, x) and (z^h, x) are in the support of the data, then Eq. (5.12) can be rewritten as

$$\begin{aligned} & E(Y_{fi} \mid \varphi(z^\ell, x) \leq \nu_i < \varphi(z^h, x), X_i = x) \\ & \quad = \frac{E(Y_{fi} \mid (Z_i, X_i) = (z^h, x), J_i = f)\Pr(J_i = f \mid (Z_i, X_i) = (z^h, x))}{\Pr(J_i = f \mid (Z_i, X_i) = (z^h, x)) - \Pr(J_i = f \mid (Z_i, X_i) = (z^\ell, x))} \\ & \qquad - \frac{E(Y_{fi} \mid (Z_i, X_i) = (z^\ell, x), J_i = f)\Pr(J_i = f \mid (Z_i, X_i) = (z^\ell, x))}{\Pr(J_i = f \mid (Z_i, X_i) = (z^h, x)) - \Pr(J_i = f \mid (Z_i, X_i) = (z^\ell, x))} \end{aligned} \tag{5.14}$$

for $\varphi\left(z^\ell, x\right) < \varphi(z^h, x)$. Now notice that for any ν,

$$\begin{aligned} & \lim_{\varphi(z^\ell, x)\uparrow \nu, \varphi(z^h, x)\downarrow \nu} E(Y_{fi} \mid \varphi(z^\ell, x) \leq \nu_i < \varphi(z^h, x), X_i = x) \\ & \quad = E(Y_{fi} \mid \nu_i = \nu, X_i = x). \end{aligned}$$

Thus if (x, ν) is in the support of $(X_i, \varphi(Z_i, X_i))$, then $E(Y_{fi} \mid \nu_i = \nu, X_i = x)$ is identified. Since the model is symmetric, under similar conditions $E(Y_{hi} \mid \nu_i = \nu,$

$X_i = x)$ is identified as well. Finally since

$$\begin{aligned}\Delta^{\mathrm{MTE}}(x, \nu) &= E(\pi_i \mid X_i = x, \nu_i = \nu) \\ &= E(Y_{fi} \mid \nu_i = \nu, X_i = x) - E(Y_{hi} \mid \nu_i = \nu, X_i = x), \qquad (5.15)\end{aligned}$$

the marginal treatment effect is identified.

The marginal treatment effect is interesting in its own right. It is the value of the treatment for any individual with $X_i = x$ and $\nu_i = \nu$. In addition, it is also useful because the different types of treatment effects can be defined in terms of the marginal treatment effect. For example

$$\mathrm{ATE} = \int \int_0^1 \Delta^{\mathrm{MTE}}(x, \nu) \mathrm{d}\nu \mathrm{d}G(x).$$

One can see from this expression that without full support this will not be identified because $\Delta^{\mathrm{MTE}}(x, \nu)$ will not be identified everywhere.

Heckman and Vytlacil (2005) also show that the instrumental variables estimator defined in Eq. (5.5) (conditional on x) is

$$\int_0^1 \Delta^{\mathrm{MTE}}(x, \nu) h_{IV}(x, \nu) \mathrm{d}\nu$$

where they give an explicit functional form for h_{IV}. It is complicated enough that we do not repeat it here but it can be found in Heckman and Vytlacil (2005).

This framework is also useful for seeing what is not identified. In particular if $\varphi(Z_i, x)$ does not have full support so that it is bounded above or below, the average treatment effect will not be identified. However, many other interesting treatment effects can be identified. For example, the Local Average Treatment Effect in a model with no regressors (x) is

$$\mathrm{LATE} = \frac{\int_{\varphi(0)}^{\varphi(1)} \Delta^{\mathrm{MTE}}(\nu) \mathrm{d}\nu}{\varphi(1) - \varphi(0)}. \qquad (5.16)$$

More generally, in this series of papers, Heckman and Vytlacil show that the marginal treatment effect can also be used to organize many ideas in the literature. One interesting case is policy effects. They define the policy relevant treatment effect as the treatment resulting from a particular policy. They show that if the relationship between the policy and the observable covariates is known, the policy relevant treatment effect can be identified from the marginal treatment effects.

5.4. Applications of the marginal treatment effects approach

Heckman and Vytlacil (1999, 2001, 2005) suggest procedures to estimate the marginal treatment effect. They suggest what they call "local instrumental variables." Using our notation for the generalized Roy model in which $J_i = f$ when $\varphi(X_i, Z_i) - \nu_i > 0$, where ν_i is uniformly distributed, they show that

$$\Delta^{\text{MTE}}(x, \nu) = \frac{\partial E(Y_i \mid X_i = x, \varphi(X_i, Z_i) = \nu)}{\partial \nu}.$$

To see why this is the same definition of MTE as in Eq. (5.15)), note that

$$\begin{aligned}
&\frac{\partial E(Y_i \mid X_i = x, \varphi(X_i, Z_i) = \nu)}{\partial \nu} \\
&\quad = \frac{\partial \left[E(Y_{fi} \mid X_i = x, \nu_i \leq \nu) \Pr(\nu_i \leq \nu) + E(Y_{hi} \mid X_i = x, \nu_i > \nu) \Pr(\nu_i > \nu)\right]}{\partial \nu} \\
&\quad = \frac{\partial \left[\int_0^\nu E(Y_{fi} \mid \nu_i = \omega, X_i = x) \mathrm{d}\omega + \int_\nu^1 E(Y_{hi} \mid \nu_i = \omega, X_i = x) \mathrm{d}\omega\right]}{\partial \nu} \\
&\quad = E(Y_{fi} \mid \nu_i = \nu, X_i = x) - E(Y_{hi} \mid \nu_i = \nu, X_i = x) \\
&\quad = \Delta^{\text{MTE}}(x, \nu).
\end{aligned}$$

Thus one can estimate the marginal treatment effect in three steps. First estimate φ, second estimate $E(Y_i \mid X_i = x, \varphi(X_i, Z_i) = \nu)$ using some type of nonparametric regression approach, and third take the derivative.

Because as a normalization ν_i is uniformly distributed

$$\begin{aligned}
\varphi(x, z) &= \Pr(\nu_i \leq \varphi(X_i, Z_i) \mid X_i = x, Z_i = z) \\
&= \Pr(J_i = f \mid X_i = x, Z_i = z) \\
&= E(D_{fi} \mid X_i = x, Z_i = z).
\end{aligned}$$

Thus we can estimate $\varphi(x, z)$ from a nonparametric regression of D_{fi} on (X_i, Z_i).

A very simple way to do this is to use a linear probability model of D_{fi} regressed on a polynomial of Z_i. By letting the terms in the polynomial get large with the sample size, this can be considered a nonparametric estimator. For the second stage we regress the outcome Y_i on a polynomial of our estimate of $\varphi(Z_i)$. To see how this works consider the case in which both polynomials are quadratics. We would use the following two stage least squares procedure:

$$D_{fi} = \gamma_0 + \gamma_1 Z_i + \gamma_2 Z_i^2 + \gamma_x X_i + e_i, \tag{5.17}$$

$$Y_i = \beta_0 + \beta_1 \widehat{D_{fi}} + \beta_2 \widehat{D_{fi}}^2 + \beta_x X_i + u_i, \tag{5.18}$$

where $\widehat{D_{fi}} = \widehat{\gamma}_0 + \widehat{\gamma}_1 Z_i + \widehat{\gamma}_2 Z_i^2 + \widehat{\gamma}_x X_i$ is the predicted value from the first stage. The β_2 coefficient may not be 0 because as we change $\widehat{D_{fi}}$ the instrument affects different groups of people. The MTE is the effect of changing $\widehat{D_{fi}}$ on Y_i. For the case above the MTE is:

$$\frac{\partial Y_i}{\partial \widehat{D_{fi}}} = \beta_1 + 2\beta_2 \widehat{D_{fi}}. \tag{5.19}$$

Although the polynomial procedure above is transparent, the most common technique used to estimate the MTE is local linear regression.

French and Song (2010) estimate the labor supply response to Disability Insurance (DI) receipt for DI applicants. Individuals are deemed eligible for DI benefits if they are "unable to engage in substantial gainful activity"—i.e., if they are unable to work. Beneficiaries receive, on average $12,000 per year, plus Medicare health insurance. Thus, there are strong incentives to apply for benefits. They continue to receive these benefits only if they earn less than a certain amount per year ($10,800 in 2007). For this reason, the DI system likely has strong labor supply disincentives. A healthy DI recipient is unlikely to work if that causes the loss of DI and health insurance benefits.

The DI system attempts to allow benefits only to those who are truly disabled. Many DI applicants have their case heard by a judge who determines those who are truly disabled. Some applicants appear more disabled than others. The most disabled applicants are unable to work, and thus will not work whether or not they get the benefit. For less serious cases, the applicant will work, but only if she is denied benefits. The question, then, is what is the optimal threshold level for the amount of observed disability before the individual is allowed benefits? Given the definition of disability, this threshold should depend on the probability that an individual does not work, even when denied the benefit. Furthermore, optimal taxation arguments suggest that benefits should be given to groups whose labor supply is insensitive to benefit allowance. Thus the effect of DI allowance on labor supply is of great interest to policy makers.

OLS is likely to be inconsistent because those who are allowed benefits are likely to be less healthy than those who are denied. Those allowed benefits would have had low earnings even if they did not receive benefits. French and Song propose an IV estimator using the process of assignment of cases to judges. Cases are assigned to judges on a rotational basis within each hearing office, which means that for all practical purposes, judges are randomly assigned to cases conditional on the hearing office and the day. Some judges are much more lenient than others. For example, the least lenient 5% of all judges allow benefits to less than 45% of the cases they hear, whereas the most lenient 5% of all judges allow benefits to 80% of all the cases they hear. Although some of those

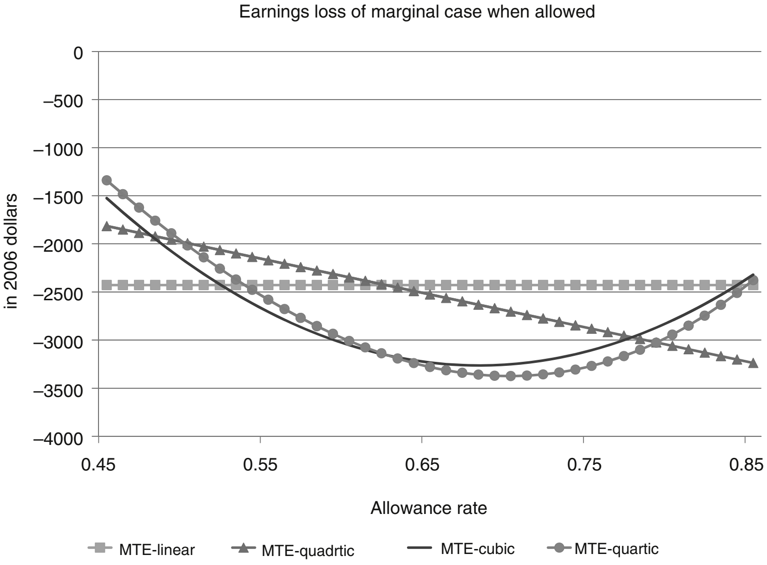


Figure 1 ***Marginal treatment effect.***

who are denied benefits appeal and get benefits later, most do not. If assignment of cases to judges is random then the instrument of judge assignment is a plausibly exogenous instrument. Furthermore, and as long as judges vary in terms of leniency and not ability to detect individuals who are disabled,[15] the instrument can identify a MTE.

French and Song use a two stage procedure. In the first stage they estimate the probability that an individual is allowed benefits, conditional on the average judge specific allowance rate. They estimate a version of Eq. (5.17) where D_{fi} is an indicator equal to 1 if case i was allowed benefits and Z_i is the average allowance rate of the judge who heard case i. In the second stage they estimate earnings conditional on whether the individual was allowed benefits (as predicted by the judge specific allowance rate). They estimate a version of Eq. (5.18) where Y_i is annual earnings 5 years after assignment to a judge. Figure 1 shows the estimated MTE (using the formula in Eq. (5.19)) using several different specifications of polynomial in the first and second stage equations. Assuming that the treatment effect is constant (i.e., $\beta_2 = 0$), they find that annual earnings 5 years after assignment to a judge are \$1500 for those allowed benefits and \$3900 for those denied benefits, so the estimated treatment effect is \$2400. This is the MTE-linear case in Fig. 1. However, this masks considerable heterogeneity in the treatment effects. They find that when allowance rates rise, the labor supply response of the marginal case also rises. When allowing for the quadratic term β_2 to be non-zero, they find that less lenient

[15] If judges vary in terms of ability to detect disability, then a case that is allowed by a low allowance judge might be denied by a high allowance judge. This would violate the monotonicity assumption shown in Eq. (5.7).

judges (who allow 45% of all cases) have a MTE of a $1800 decline in earnings. More lenient judges (who allow 80% of all cases) have a MTE of $3200 decline in earnings. Figure 1 also shows results when allowing for cubic and quartic terms in the polynomials in the first and second stage equations. This result is consistent with the notion that as allowance rates rise, more healthy individuals are allowed benefits. These healthier individuals are more likely to work when not receiving DI benefits, and thus their labor supply response to DI receipt is greater.

One problem with an instrument such as this is that the instrument lacks full support. Even the most lenient judge does not allow everyone benefits. Even the strictest judge does not deny everyone. However, the current policy debate is whether the thresholds should be changed by only a modest amount. For this reason, the MTE on the support of the data is the effect of interest, whereas the ATE is not.

Doyle (2007) estimates the Marginal Treatment Effect of foster care on future earnings and other outcomes. Foster care likely increases earnings of some children but decreases it for others. For the most serious child abuse cases, foster care will likely help the child. For less serious cases, the child is probably best left at home. The question, then, is at what point should the child abuse investigator remove the child from the household? What is the optimal threshold level for the amount of observed abuse before which the child is removed from the household and placed into foster care?

Only children from the most disadvantaged backgrounds are placed in foster care. They would have had low earnings even if they were not placed in foster care. Thus, OLS estimates are likely inconsistent. To overcome this problem, Doyle uses IV. Case investigators are assigned to cases on a rotational basis, conditional on time and the location of the case. Case investigators are assigned to possible child abuse cases after a complaint of possible child abuse is made (by the child's teacher, for example). Investigators have a great deal of latitude about whether the child should be sent into foster care. Furthermore, some investigators are much more lenient than others. For example, one standard deviation in the case manager removal differential (the difference between his average removal rate and the removal rate of other investigators who handle cases at the same time and place) is 10%. Whether the child is removed from the home is a good predictor of whether the child is sent to foster care. So long as assignment of cases to investigators is random and investigators only vary in terms of leniency (and not ability to detect child abuse) then the instrument of investigator assignment is a useful and plausibly exogenous instrument.

Doyle uses a two stage procedure where in the first stage he estimates the probability that a child is placed in foster care as a function of the investigator removal rate. In the second stage he estimates adult earnings as a function of whether the child was placed in foster care (as predicted by the instrument). He finds that children placed into foster care earn less than those not placed into foster care over most of the range of the data. Two stage least squares estimates reveal that foster care reduces adult quarterly earnings by about \$1000, which is very close to average earnings. Interestingly, he finds that when child foster care placement rates rise, earnings of the marginal case fall. For example, earnings of the marginal child handled by a lenient investigator (who places only 20% of the children in foster care) are unaffected by placement. For less lenient investigators, who place 25% of the cases in foster care, earnings of the marginal case decline by over \$1500.

Carneiro and Lee (2009) estimate the counterfactual marginal distributions of wages for college and high school graduates, and examine who enters college. They find that those with the highest returns are the most likely to attend college. Thus, increases in college cause changes in the distribution of ability among college and high school graduates. For fixed skill prices, they find that a 14% increase in college participation (analogous to the increase observed in the 1980s) reduces the college premium by 12%. Likewise, Carneiro et al. (2010) find that while the conventional IV estimate of the return to schooling (using distance to a college and local labor market conditions as the instruments) is 0.095, the estimated marginal return to a policy that expands each individual's probability of attending college by the same proportion is only 0.015.

5.5. Selection on observables

Perhaps the simplest and most common assumption is that assignment of the treatment is random conditional on observable covariates (sometimes referred to as unconfoundedness). The easiest way to think about this is that the selection error term is independent of the other error terms:

Assumption 5.2.

$$J_i = f \quad \text{when } \varphi(X_i) > \nu_i$$

where ν_i is independent of $(\varepsilon_{fi}, \varepsilon_{hi})$.

We continue to assume that $Y_{fi} = g_f(X_{fi}, X_{0i}) + \varepsilon_{fi}$ and $Y_{hi} = g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}$. Note that we have explicitly dropped Z_i from the model as we consider cases in which we do not have exclusion restrictions. The implication of this assumption is that unobservable factors that determine one's income as a fisherman do not affect the choice to become a fisherman. That is while it allows for selection on observables in a very general way, it does not allow for selection on unobservables.

Interestingly, this is still not enough for us to identify the Average Treatment Effect. If there are values of observable covariates X_i for which $\Pr(J_i = f \mid X_i = x) = 1$ or $\Pr(J_i = f \mid X_i = x) = 0$ the model is not identified. If $\Pr(J_i = f \mid X_i = x) = 1$ then it is straightforward to identify $E(Y_{fi} \mid X_i = x)$, but $E(Y_{hi} \mid X_i = x)$ is not identified. Thus we need the additional assumption

Assumption 5.3. For almost all x in the support of X_i,

$$0 < \Pr(J_i = f \mid X_i = x) < 1.$$

Theorem 5.2. *Under Assumptions* 5.2 and 5.3 *the Average Treatment Effect is identified.*

(Proof in Appendix.)

Estimation in this case is relatively straightforward. One can use matching[16] or regression analysis to estimate the average treatment effect.

5.6. Set identification of treatment effects

In our original discussion of identification we defined $\Psi(\Theta(P))$ as "the set of values of ψ that are consistent with the data distribution P." We said that ψ was identified if this set was a singleton. However, there is another concept of identification we have not discussed until this point; this is set identification. Sometimes we may be interested in a parameter that is not point identified, but this does not mean we cannot say anything about it. In this subsection we consider the case of set identification (i.e. trying to characterize the set $\Psi(\Theta(P))$ focusing on the case in which ψ is the Average Treatment Effect. Suppose that we have some prior knowledge (possibly an exclusion restriction that gives us a LATE). What can we learn about the ATE without making any functional form assumptions? In a series of papers Manski (1989, 1990, 1995, 1997) and Manski and Pepper (2000, 2009) develop procedures to derive set estimators of the Average Treatment Effect and other parameters given weak assumptions. By "set identification" we mean the set of possible Average Treatment Effects given the assumptions placed on the data. Throughout this section we will continue to assume that the structure of the Generalized Roy model holds and we derive results under these assumptions. In many cases the papers we mentioned do not impose this structure and get more general results.

Following Manski (1990) or Manski (1995), notice that

$$E\left(Y_{fi}\right) = E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + E(Y_{fi} \mid J_i = h)\Pr(J_i = h) \quad (5.20)$$

$$E\left(Y_{hi}\right) = E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + E(Y_{hi} \mid J_i = f)\Pr(J_i = f). \quad (5.21)$$

We observe all of the objects in Eqs (5.20) and (5.21) except $E(Y_{fi} \mid J_i = h)$ and $E(Y_{hi} \mid J_i = f)$. The data are completely uninformative about these two objects. However, suppose we have some prior knowledge about the support of Y_{fi} and Y_{hi}. In particular, suppose that the support of Y_{fi} and Y_{hi} are bounded above by y^u and from below by y^ℓ. Thus, by assumption $y^u \geq E(Y_{fi} \mid J_i = h) \geq y^\ell$ and $y^u \geq E(Y_{hi} \mid J_i = f) \geq y^\ell$.

[16] Our focus is on identification rather than estimation. Thus we avoid a discussion of matching estimators. See Heckman et al. (1999), Imbens and Wooldridge (2009), or DiNardo and Lee (2011) for discussion.

Using these assumptions and Eqs (5.20) and (5.21) we can establish that

$$\begin{aligned}
&E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + y^{\ell}\Pr(J_i = h) \\
&\quad \leq E\left(Y_{fi}\right) \leq E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + y^{u}\Pr(J_i = h) \qquad (5.22) \\
&E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + y^{\ell}\Pr(J_i = f) \\
&\quad \leq E\left(Y_{hi}\right) \leq E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + y^{u}\Pr(J_i = f). \qquad (5.23)
\end{aligned}$$

Using these bounds and the definition of the ATE

$$\text{ATE} = E\left(Y_{fi}\right) - E\left(Y_{hi}\right) \qquad (5.24)$$

yields

$$\begin{aligned}
&(E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + y^{\ell}\Pr(J_i = h)) \\
&\quad - (E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + y^{u}\Pr(J_i = f)) \\
&\leq \text{ATE} \\
&\leq (E(Y_{fi} \mid J_i = f)\Pr(J_i = f) + y^{u}\Pr(J_i = h)) \\
&\quad - (E(Y_{hi} \mid J_i = h)\Pr(J_i = h) + y^{\ell}\Pr(J_i = f)).
\end{aligned}$$

In practice the bounds above can yield wide ranges and are often not particularly informative. A number of other assumptions can be used to decrease the size of the identified set.

Manski (1990, 1995) shows that one method of tightening the bounds is with an instrumental variable. We can write the expressions (5.20) and (5.21) conditional on $Z_i = z$ for any $z \in \text{supp}(Z_i)$ as for each $j \in \{f, h\}$,

$$\begin{aligned}
E\left(Y_{ji}|Z_i = z\right) = {} & E(Y_{ji} \mid J_i = f, Z_i = z)\Pr(J_i = f \mid Z_i = z) \\
& + E(Y_{ji} \mid J_i = h, Z_i = z)\Pr(J_i = h \mid Z_i = z). \qquad (5.25)
\end{aligned}$$

Since Z_i is, by assumption, mean independent of Y_{fi} and Y_{hi} (it only affects the probability of choosing one occupation versus the other), then $E\left(Y_{fi}|Z_i = z\right) = E\left(Y_{fi}\right)$ and $E\left(Y_{hi}|Z_i = z\right) = E(Y_{hi})$. Assume there is a binary instrumental variable, Z_i, which equals either 0 or 1. We can then follow exactly the same argument as in Eqs (5.22) and (5.23), but conditioning on Z_i and using Eq. (5.25) yields

$$\begin{aligned}
&E(Y_{fi} \mid J_i = f, Z_i = 1)\Pr(J_i = f \mid Z_i = 1) + y^{\ell}\Pr(J_i = h \mid Z_i = 1) \\
&\quad \leq E\left(Y_{fi}\right) \\
&\quad \leq E(Y_{fi} \mid J_i = f, Z_i = 1)\Pr(J_i = f \mid Z_i = 1) + y^{u}\Pr(J_i = h \mid Z_i = 1) \qquad (5.26)
\end{aligned}$$

$$\begin{aligned}
&E(Y_{hi} \mid J_i = h, Z_i = 0) \Pr(J_i = h \mid Z_i = 0) + y^{\ell} \Pr(J_i = f \mid Z_i = 0) \\
&\quad \leq E\,(Y_{hi}) \\
&\quad \leq E(Y_{hi} \mid J_i = h, Z_i = 0) \Pr(J_i = h \mid Z_i = 0) + y^{u} \Pr(J_i = f \mid Z_i = 0). \quad (5.27)
\end{aligned}$$

Thus we can bound ATE $= E(Y_{fi}) - E(Y_{hi})$ from below by subtracting (5.27) from (5.26):

$$\begin{aligned}
&E(Y_{fi} \mid J_i = f, Z_i = 1) \Pr(J_i = f \mid Z_i = 1) + y^{\ell} \Pr(J_i = h \mid Z_i = 1) \\
&\qquad - E(Y_{hi} \mid J_i = h, Z_i = 0) \Pr(J_i = h \mid Z_i = 0) + y^{u} \Pr(J_i = f \mid Z_i = 0) \\
&\quad \leq \text{ATE} \\
&\quad \leq E(Y_{fi} \mid J_i = f, Z_i = 1) \Pr(J_i = f \mid Z_i = 1) \\
&\qquad + y^{u} \Pr(J_i = h \mid Z_i = 1) - E(Y_{hi} \mid J_i = h, Z = 0) \\
&\qquad \times \Pr(J_i = h \mid Z_i = 0) + y^{\ell} \Pr(J_i = f \mid Z_i = 0). \qquad (5.28)
\end{aligned}$$

Our choice of a binary value of Z_i can be trivially relaxed. In the cases in which Z_i takes on many values one could choose any two values in the support of Z_i to get upper and lower bounds. If our goal is to minimize the size of the set we would choose the values z^{ℓ} and z^{h} to minimize the difference between the upper and lower bounds in (5.28):

$$(y^{u} - y^{\ell})[\Pr(J_i = h \mid Z_i = z^{h}) + \Pr(J_i = f \mid Z_i = z^{\ell})].$$

The importance of support conditions once again becomes apparent from this expression. If we could find values z^{ℓ} and z^{h}such that

$$\begin{aligned}
\Pr(J_i = h \mid Z_i = z^{h}) = 0 \\
\Pr(J_i = f \mid Z_i = z^{\ell}) = 0
\end{aligned}$$

then this expression is zero and we obtain point identification of the ATE. When $\Pr(J_i = h \mid Z_i = z)$ or $\Pr(J_i = f \mid Z_i = z)$ are bounded from below we are only able to obtain set estimates. A nice aspect of this is that it represents a nice middle point between identifying LATE versus claiming the ATE is not identified. If the identification at infinity effect is not exactly true, but approximately true so that one can find values of z^{ℓ} and z^{h} so that $\Pr(J_i = h \mid Z_i = z^{h})$ and $\Pr(J_i = f \mid Z_i = z^{\ell})$ are small, then the bounds will be tight. If one cannot find such values, the bounds will be far apart.

In many cases these bounds may be wide. Wide bounds can be viewed in two ways. One interpretation is that the bounding procedure is not particularly helpful in learning about the true ATE. However, a different interpretation is that it shows that the data, without additional assumptions, is not particularly helpful for learning about the ATE. Below we discuss additional assumptions for tightening the bounds on the ATE, such as Monotone treatment response, Monotone treatment selection, and Monotone instruments. In order to keep matters simple, below we assume that there is no exclusion restriction. However, if a exclusion restriction is known, this allows us to tighten the bounds.

Next we consider the assumption of Monotone Treatment Response introduced in Manski (1997), which we write as

Assumption 5.4. Monotone Treatment Response

$$Y_{fi} \geq Y_{hi}$$

with probability one.

In the fishing/hunting example this is not a particularly natural assumption, but for many applications in labor economics it is. Suppose we are interested in knowing the returns to a college degree, and Y_{fi} is income for individual i if a college graduate whereas Y_{hi} is income if a high school graduate. It is reasonable to believe that the causal effect of school or training cannot be negative. That is, one could reasonably assume that receiving more education can't causally lower your wage. Thus, Monotone Treatment Response seems like a reasonable assumption in this case. This can lower the bounds above quite a bit because now we know that

$$E(Y_{fi} \mid J_i = h) \geq E(Y_{hi} \mid J_i = h) \tag{5.29}$$

$$E(Y_{hi} \mid J_i = f) \leq E(Y_{fi} \mid J_i = f). \tag{5.30}$$

From this Manski (1997) shows that

$$0 \leq \text{ATE}.$$

Another interesting assumption that can also help tighten the bounds is the Monotone Treatment Selection assumption introduced in Manski and Pepper (2000). In our framework this can be written as

Assumption 5.5. Monotone Treatment Selection: for $j = f$ or h,

$$E(Y_{ji} \mid J_i = f) \geq E(Y_{ji} \mid J_i = h).$$

Again this might not be completely natural for the fishing/hunting example, but may be plausible in many other cases. For example it seems like a reasonable assumption in schooling if we believe that there is positive sorting into schooling. Put differently, suppose the average college graduate is a more able person than the average high school graduate and would earn higher income, even if she did not have the college degree. If this is true, then the average difference in earnings between college and high school graduates overstates the true causal effect of college on earnings. This also helps to further tighten the bounds as this implies that

$$\text{ATE} \leq E(Y_{fi} \mid J_i = f) - E(Y_{hi} \mid J_i = h).$$

Note that by combining the MTR and MTS assumption, one can get the tighter bounds:

$$0 \leq \text{ATE} \leq E(Y_{fi} \mid J_i = f) - E(Y_{hi} \mid J = h).$$

Manski and Pepper (2000) also develop the idea of a monotone instrumental variable. An instrumental variable is defined as one for which for any two values of the instrument z_a and z_b,

$$E(Y_{ji} \mid Z_i = z_a) = E(Y_{ji} \mid Z_i = z_b).$$

In words, the assumption is that the instrument does not directly affect the outcome variable Y_{ji}. It only affects one's choices. Using somewhat different notation, but their exact wording, they define a monotone instrumental variable in the following way

Assumption 5.6. Let $\mathcal{Z}$ be an ordered set. Covariate Z_i is a monotone instrumental variable in the sense of mean-monotonicity if, for $j \in \{f, h\}$,each value of x, and all $(z_b, z_a) \in (\mathcal{Z} \times \mathcal{Z})$ such that $z_b \geq z_a$,

$$E(Y_{ji} \mid X_i = x, Z_i = z_b) \geq E(Y_{ji} \mid X_i = x, Z_i = z_a).$$

This is a straight generalization of the instrumental variable assumption, but imposes much weaker requirements for an instrument. It does not require that the instrument be uncorrelated with the outcome, but simply that the outcome monotonically increase with the instrument. An example is that parental income has often been used as an instrument for education. Richer parents are better able to afford a college degree for their child. However, it seems likely that the children of rich parents would have had high earnings, even in the absence of a college degree.

They show that this implies that

$$\sum_{z \in \mathcal{Z}} \Pr(Z_i = z) \left\{ \sup_{z_a \leq z} \left[E\left(Y_i \mid Z_i = z_a, J_i = f\right) \Pr\left(J_i = f \mid Z_i = z_a\right) \right.\right.$$
$$\left.\left. + \; y^{\ell} \Pr\left(J_i = h \mid Z_i = z_a\right) \right] \right\}$$

$$
\begin{aligned}
&- \sum_{z \in \mathcal{Z}} \Pr(Z_i = z) \left\{ \inf_{z_b \geq z} \left[E\left(Y_i \mid Z_i = z_b, J_i = h\right) \Pr\left(J_i = h \mid Z_i = z_b\right) \right. \right. \\
&\left. \left. + \; y^u \Pr\left(J_i = f \mid Z_i = z_b\right) \right] \right\} \\
\leq\ & \text{ATE} \\
\leq\ & \sum_{z \in \mathcal{Z}} \Pr(Z_i = z) \left\{ \inf_{z_b \geq z} \left[E\left(Y_i \mid Z_i = z_b, J_i = f\right) \Pr\left(J_i = f \mid Z_i = z_b\right) \right. \right. \\
&\left. \left. + \; y^u \Pr\left(J_i = h \mid Z_i = z_b\right) \right] \right\} \\
&- \sum_{z \in \mathcal{Z}} \Pr(Z_i = z) \left\{ \sup_{z_a \leq z} \left[E\left(Y_i \mid Z_i = z_a, J_i = h\right) \Pr\left(J_i = h \mid Z_i = z_a\right) \right. \right. \\
&\left. \left. + \; y^\ell \Pr\left(J_i = f \mid Z_i = z_a\right) \right] \right\}.
\end{aligned}
$$

One can obtain tighter bounds by combining the Monotone Instrumental Variable assumption with the Monotone Treatment Response assumption but we do not explicitly present this result.

Blundell et al. (2007) estimate changes in the distribution of wages in the United Kingdom using bounds to allow for the impact of non-random selection into work. They first document the growth in wage inequality among workers over the 1980s and 1990s. However, they point out that rates of non-participation in the labor force have grown in the UK over the same time period. Nevertheless, they show that selection effects alone cannot explain the rise in inequality observed among workers: the worst case bounds establish that inequality has increased. However, worst case bounds are not sufficiently informative to understand such questions as whether most of the rise in wage inequality is due to increases in wage inequality within education groups versus across education groups. Next, they add an additional assumptions to tighten the bounds. First, they assume the probability of work is higher for those with higher wages, which is essentially the Monotone Treatment Selection assumption shown in Assumption 5.5. Second, they make the Monotone Instrumental Variables assumption shown in Assumption 5.6. They assume that higher values of out of work benefit income are positively associated with wages. They show that both of these assumptions tighten the bounds considerably. They find that when these additional restrictions are made, then they can show that both within group and between group inequality has increased.

5.7. Using selection on observables to infer selection on unobservables

Altonji et al. (2005a) suggest another approach which is to use the amount of selection on observable covariates as a guide to the potential amount of selection on unobservables.

To motivate this approach, consider an experiment in which treatment status is randomly assigned. The key to random assignment is that it imposes that treatment status be independent of the unobservables in the treatment model. Since they are unobservable, one can never explicitly test whether the treatment was truly random. However, if randomization was carried out correctly, treatment should also be uncorrelated with observable covariates. This is testable, and applying this test is standard in experimental approaches.

Researchers use this same argument in non-experimental cases as well. If a researcher wants to argue that his instrument or treatment is approximately randomly assigned, then it should be uncorrelated with observable covariates as well. Even if this is strictly not required for consistent estimates of instrumental variables, readers may be skeptical of the assumption that the instrument is uncorrelated with the unobservables if it is correlated with the observables. Researchers often test for this type of relationship as well.[17] The problem with this approach is that simply testing the null of uncorrelatedness is not that useful. Just because you reject the null does not mean it isn't approximately true. We would not want to throw out an instrument with a tiny bias just because we have a data set large enough to detect a small correlation between it and an observable. Along the same lines, just because you fail to reject the null does not mean it is true. If one has a small data set with little power one could fail to reject the null even though the instrument is poor. To address these issues, Altonji et al. (2005a) design a framework that allows them to describe how large the treatment effect would be if "selection on the unobservables is the same as selection on the observables."

Their key variables are discrete, so they consider a latent variable model in which a dummy variable for graduation from high school can be written as

$$G_i = \begin{cases} 1 & Y_i^* \geq 0 \\ 0 & Y_i^* < 0 \end{cases}$$

where Y_i^* can be written as

$$\begin{aligned} Y_i^* &= \beta_0 + \alpha D_{fi} + \sum_{j=1}^{K} W_{ij}\beta_j \\ &= \beta_0 + \alpha D_{fi} + \sum_{j=1}^{K} S_j W_{ij}\beta_j + \sum_{j=1}^{K}(1 - S_j) W_{ij}\beta_j \\ &= \beta_0 + \alpha D_{fi} + X_i'\beta + \nu_i. \end{aligned}$$

W_{ij} represent all covariates, both those that are observable to the econometrician and those that are unobservable, the variable S_j is a dummy variable representing whether the

[17] Altonji et al. (2005a) discuss a number of studies that do so.

covariate is observable to the empirical researcher, $X_i'\beta = \sum_{j=1}^{K} S_j W_{ij}\beta_j$ represents the observable part of the index, and $\nu_i = \sum_{j=1}^{K}(1 - S_j)W_{ij}\beta_j$ denotes the unobservable part.

Within this framework, one can see that different assumptions about what dictates which observables are chosen (S_j) can be used to identify the model. Their specific goal is to quantify what it means for "selection on the observables to be the same as selection on the unobservables." They argue that the most natural way to formalize this idea is to assume that S_j is randomly assigned so that the unobservables and observables are drawn from the same underlying distribution.

The next question is what this assumption implies on the data that can be useful for identification. They consider the projection:

$$\text{proj}(Z_i \mid X_i'\beta, \nu_i) = \phi_0 + \phi X_i'\beta + \phi_\varepsilon \nu_i$$

where Z_i can be any random variable. They show that if S_j is randomly assigned,

$$\phi \approx \phi_\varepsilon.$$

This restriction is typically sufficient to insure identification of α.[18]

Altonji et al. (2005a,b) argue that for their example this is an extreme assumption and the truth is somewhere in between this assumption and the assumption that Z_i is uncorrelated with the unobservables which would correspond to $\phi_\varepsilon = 0$. They assume that when $\phi > 0$,

$$0 \leq \phi_\varepsilon \leq \phi.$$

There are at least three arguments for why selection on unobservables would be expected to be less severe than selection on observables (as it is measured here). First, some of the variation in the unobservable is likely just measurement in the dependent variable. Second, data collectors likely collect the variables that are likely to be correlated with many things. Third, there is often a time lapse between the time the baseline data is collected (the observables) and when the outcome is realized. If unanticipated events occur in between these two time periods, that would lead to the result.

Notice that if $\phi = 0$ then assuming $\phi_\varepsilon = \phi$ is the same as assuming $\phi_\varepsilon = 0$. However, if ϕ were very large the two estimates would be very different, which would shed doubt on the assumption of random assignment. Since ϕ essentially picks up the relationship between the instrument and the observable covariates, the bounds would be wide when

[18] In some cases it is not point identification, but either 2 or 3 different points.

there is a lot of selection on observables and will be tight when there is little selection on observables.

Altonji, Elder, and Taber consider the case of whether the decision to attend Catholic high school affects outcomes such as test scores and high school graduation rates. Those who attend Catholic schools have higher graduation rates than those who do not attend Catholic schools. However, those who attend Catholic may be very different from those who do not. They find that (on the basis of observables) while this is true in the population, it is not true when one conditions on the individuals who attend Catholic school in eighth grade. To formalize this, they use their approach and estimate the model under the two different assumptions. In their application the projection variable, Z_i, is the latent variable determining whether an individual attends Catholic school. First they estimate a simple probit of high school graduation on Catholic high school attendance as well as many other covariates. This corresponds to the $\phi_\varepsilon = 0$ case. They find a marginal effect of 0.08, meaning that Catholic school raises high school graduation by eight percentage points. Next they estimate a bivariate probit of Catholic high school attendance and high school graduation subject to the constraint that $\phi_\varepsilon = \phi$. In this case they find a Catholic high school effect of 0.05. The closeness of these two estimates strongly suggests that the Catholic high school effect is not simply a product omitted variable bias. The tightness of the two estimates arose both because ϕ was small and because they use a wide array of powerful explanatory variables.

6. DURATION MODELS AND SEARCH MODELS

In this section we relate the previous discussion to the competing risks model and the search model. We show that the competing risk model can be written in a way that is almost identical to the Roy model. We also show how the basic ideas of exclusion restrictions can be used to identify a version of a search model.

6.1. Competing risks model

With duration data a researcher observes the elapsed time until some event occurs. The prototypical example in labor economics is the duration of unemployment and we focus on that example. We explain why identification of this model is almost identical to identification of the Roy model. Let T_i denote the length of an unemployment spell. There are (at least) four different ways to characterize the distribution of T_i. The first is the cumulative distribution function $F(t) \equiv \Pr(t > T_i)$, which in the context of unemployment durations is the probability the individual found a job. The second is the density function f. The third is the survivor function defined as

$$S(t) \equiv \Pr(T_i > t) = 1 - F(t).$$

The fourth is the hazard function, which is the job finding rate at time t, given that the individual was unemployed at time t:

$$\begin{aligned} h(t) &\equiv \lim_{\delta \to 0} \frac{\Pr(T_i \leq t + \delta \mid T_i \geq t)}{\delta} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

The link between the hazard rate and survivor function is:

$$\begin{aligned} h(t) = \frac{f(t)}{S(t)} &= \frac{\mathrm{d}F(t)/\mathrm{d}t}{S(t)} \\ &= \frac{-\mathrm{d}S(t)/\mathrm{d}t}{S(t)} \\ &= \frac{-\mathrm{d}\log S(t)}{\mathrm{d}t}. \end{aligned} \tag{6.1}$$

There is a large literature on identification of duration models. Heckman and Taber (1994), Van den Berg (2001), and Abbring (2010) provide excellent surveys of this literature.[19] Rather than survey the full literature here we relate it to our previous discussion. Given that T_i must be positive, it is natural to model T_i using the basic framework we have been using all along:

$$\log(T_i) = g(X_i) + \varepsilon_i.$$

Clearly if we could observe the distribution of $\log(T_i)$ conditional on X_i, identification of g and the distribution of ε_i would be straightforward.

However, often we cannot observe the full duration of T_i because the spell (or our observation of it) is truncated before the worker is re-employed. For example, the worker may die, be lost from the data, or the survey may end. In the classic medical example we might want to estimate the duration until a patient has a heart attack, but if she dies from cancer we never observe this event. Hence the name "competing risk model." To put this in the context of our Roy model example, suppose an unemployed worker would take the first offer they received and they can get an offer as a fisherman or a hunter. Define the model as

$$\log(T_{fi}) = g_f(X_i) + \varepsilon_{fi} \tag{6.2}$$

$$\log(T_{hi}) = g_h(X_i) + \varepsilon_{hi} \tag{6.3}$$

[19] Key papers include Elbers and Ridder (1982), Heckman and Singer (1984a,b), Ridder (1990), Honoré (1993), and Abbring and Ridder (2009).

where T_{fi} and T_{hi} are the amount of time it would take until the worker received an offer as a fisherman or as a hunter, X_i denotes observable variables that are independent of the unobservables $(\varepsilon_{fi}, \varepsilon_{hi})$.[20] The econometrician can observe whether the worker becomes a fisherman or a hunter and the length of the unemployment spell. However, notice that as Heckman and Honoré (1990) point out, this is just another version of the Roy model. Rather than observe the maximum of Y_{fi} and Y_{hi}, the econometrician observes the minimum of $\log(T_{fi})$ and $\log(T_{hi})$.

The specification (6.2) and (6.3) above is not the way that many researchers choose to model duration data. Often they model the hazard function directly as it is sometimes easier to interpret. Moreover, if the observable covariates change over time, the hazard model is a more reasonable way to model the durations. The most common specification is the mixed proportional hazard model

$$h(t \mid X_i = x) = \xi(t)\phi(x)\omega_i \tag{6.4}$$

where $\xi(t)$ is referred to as the baseline hazard, ω_i is an unobservable variable which is independent of the observables, and X_i denotes observable characteristics. Most studies find that the hazard rate for finding a job tends to decline with the unemployment duration. The model above allows for two possible interpretations of this empirical regularity. First, it could be that as unemployment durations lengthen, skills depreciate, making it harder to find a job. This is captured by $\xi(t)$. Second, it could be that some people are just less able to find a job than others in ways not captured by observables. This is captured in ω_i. Van den Berg (2001) provides a thorough discussion of this model.

Heckman and Honoré (1989) show how to map the hazard specification into a framework that is similar to what we use in our analysis of the Roy model. The transformation is simplest is when $\xi(t) = 1$. In that case one can write the survivor function as

$$\Pr(T_i > t \mid X_i = x, \omega_i = \omega) = \mathrm{e}^{-t\phi(x)\omega}. \tag{6.5}$$

It is straightforward to derive Eq. (6.4) using the survivor function (6.5) and Eq. (6.1). Define $g(\cdot) = -\log(\phi(\cdot))$ and F_ω to be the distribution of ω_i. In order to obtain the cumulative density function of unemployment durations we must integrate over the distribution of unemployed individuals:

$$\Pr(T_i \le t \mid X_i = x) = \int 1 - \mathrm{e}^{-t\phi(x)\omega_i}\mathrm{d}F_\omega$$

[20] We do not need to make use of exclusion restrictions here so we do not distinguish between observables that may enter differently.

$$
\begin{aligned}
&= \int 1 - \exp(-\exp(\log(t) - g(x) + \log(\omega_i)))\mathrm{d}F_\omega \\
&\equiv F_{\widetilde{\omega}}(\log(t) - g(x)) \qquad (6.6)
\end{aligned}
$$

where $F_{\widetilde{\omega}}$ is defined implicitly by this relationship. Note that $F_{\widetilde{\omega}}$ is a legitimate CDF, as it is strictly increasing from 0 to 1.[21] Thus one can think of the data generating process as

$$\log(T_i) = g(X_i) + \widetilde{\omega}_i$$

where $\widetilde{\omega}_i$ is distributed according to $F_{\widetilde{\omega}}$ and is independent of X_i.

In the more general case in which $\xi(t)$ is not constant, it is well known that one can write the survivor function as

$$\mathrm{e}^{-\Xi(t)\phi(X_i)\omega_i} \qquad (6.7)$$

where Ξ is the integrated hazard

$$\Xi(t) \equiv \int_0^t \xi(t)\mathrm{d}t.$$

Equation (6.7) differs from Eq. (6.5) by the term $\Xi(t)$ instead of t. Thus replacing t with $\Xi(t)$ in Eq. (6.6) yields

$$\log(\Xi(T_i)) = g(X_i) + \widetilde{\omega}_i.$$

Heckman and Honoré (1989) use a more general framework to think about the competing risks model in which the probability of not getting a fishing job by time t_f and not getting a hunting job by time t_h, $S(t_f, t_h \mid X_i = x)$, can be written as

$$S(t_f, t_h \mid X_i = x) = K(\exp\{-\Xi_f(t_f)\phi_f(x)\}, \exp\{-\Xi_h(t_h)\phi_h(x)\})$$

where $\phi_j(x) = \exp(-g_j(x))$ for $j = f, h$. This is a generalization of a model in which

$$
\begin{aligned}
\log(\Xi_f(T_{fi})) &= g_f(X_i) + \widetilde{\omega}_{fi} \\
\log(\Xi_h(T_{hi})) &= g_h(X_i) + \widetilde{\omega}_{hi}
\end{aligned}
$$

because

$$
\begin{aligned}
S(t_f, t_h \mid X_i = x) = \Pr[&\log(\Xi_f(T_{fi})) > \log(\Xi_f(t_f)), \log(\Xi_h(T_{hi})) \\
&> \log(\Xi_h(t_h)) \mid X_i = x]
\end{aligned}
$$

[21] It is the distribution of a convolution between $\log(\omega_i)$ and an extreme value.

$$
\begin{aligned}
&= \Pr[g_f(x) + \widetilde{\omega}_{fi} > \log(\Xi_f(t_f)), g_h(x) + \widetilde{\omega}_{hi} > \log(\Xi_h(t_h))] \\
&= \Pr[-\widetilde{\omega}_{fi} < -\log(\Xi_f(t_f)) + g_f(x), -\widetilde{\omega}_{hi} \\
&\quad < -\log(\Xi_h(t_h)) + g_h(x)] \\
&= F_{-\widetilde{\omega}_{fi}-\widetilde{\omega}_{hi}}(-\log(\Xi_f(t_f)) + g_f(x), -\log(\Xi_h(t_h)) + g_h(x)) \\
&\equiv K(\exp\{-\Xi_f(t_f)\phi_f(x)\}, \exp\{-\Xi_h(t_h)\phi_h(x)\}) \qquad (6.8)
\end{aligned}
$$

where $F_{-\widetilde{\omega}_{fi}-\widetilde{\omega}_{hi}}$ is the joint CDF of $(-\widetilde{\omega}^*_{fi}, -\widetilde{\omega}^*_{hi})$, and K is defined implicitly as $K(a, b) = F_{-\widetilde{\omega}_{fi}-\widetilde{\omega}_{hi}}(-\log(-\log(a)), -\log(-\log(b)))$.

Heckman and Honoré (1989), Theorem 1 contains the following result. We reproduce their result, only altering the notation.

Theorem 6.1. *Assume that* (T_{fi}, T_{hi}) *has the joint survivor function as given in* (6.8). *Then* $\Xi_f, \Xi_h, \phi_f, \phi_h$, *and* K *are identified from the identified minimum of* (T_{fi}, T_{hi}) *under the following assumptions*

1. K *is continuously differentiable with partial derivatives* K_1 *and* K_2 *for* $i = 1, 2$, *the limit as* $n \to \infty$ *of* $K_i(\eta_{1n}, \eta_{2n})$ *is finite for all sequences of* η_{1n}, η_{2n} *for which* $\eta_{1n} \to 1$ *and* $\eta_{2n} \to 1$ *for* $n \to \infty$. *We also assume that* K *is strictly increasing in each of its arguments in all of* $[0, 1] \times [0, 1]$.
2. $\Xi_f(1) = 1$, $\Xi_h(1) = 1$, $\phi_f(x_0) = 1$ *and* $\phi_h(x_0) = 1$ *for some fixed point* x_0 *in the support* X.
3. *The support of* $\{\phi_f(x), \phi_h(x)\}$ *is* $(0, \infty) \times (0, \infty)$.
4. Ξ_f *and* Ξ_h *are nonnegative, differentiable, strictly increasing functions, except that we allow them to be* ∞ *for finite* t.

(Proof in Heckman and Honoré (1989).)

Since the model is almost identical to the Roy model, the intuition for identification is very similar so we don't review it here. We do mention a few things about these assumptions. First note that assumption (2) in Theorem 6.1 is just a normalization as one cannot separate the scales of ϕ_f, Ξ_f, and ν_f. The more notable difference between this and the theorem we presented in the Roy model section above is the lack of exclusion restrictions. What is crucial in being able to do this is the assumptions about K in assumption (1). In their proof they show that for any x in the support of X_i,

$$\lim_{t \to 0} \frac{\frac{\partial \Pr(T_{fi} < t, T_{hi} > T_{fi} | X_i = x)}{\partial t}}{\frac{\partial \Pr(T_{fi} < t, T_{hi} > T_{fi} | X_i = x_0)}{\partial t}} = \phi_f(x).$$

One could in principle use this form of identification for the Roy model, but it is somewhat less natural in the Roy framework, as taking the limit as $t \to 0$ corresponds to taking limits as the log of wages become arbitrarily large. It also makes heavy use of the independence assumption, which is not necessary for identification of g_f when one has exclusion restrictions. Finally, the basic approach will not expand to the "labor supply" model in which we only observe wages in one sector and to the generalized Roy model in the same way that exclusion restrictions do.

Abbring and van den Berg (2003) extends Heckman and Honoré's (1989) results on the mixed proportional hazards competing risk models in a few ways, including generalizing the assumptions for identification somewhat and considering identification in the case in which researchers observe multiple spells.

6.2. Search models

Eckstein and van den Berg (2007) present a nice survey of Empirical Search models. We avoid a general discussion, but rather combine the proportional hazard model with a search model. In a well known result Flinn and Heckman (1982) show that the search model is not fully identified. They use the Lippman and McCall (1976) search model in which workers search for jobs until their wage exceeds their reservation wage. In this model, one essentially assumes that the worker stays at the job forever. All workers are assumed to be ex-ante identical and face the same distribution of offered wages, which we denote by F. The reservation wage w^r is the point at which the individual is indifferent between taking the job and continued search. It is defined implicitly by the formula

$$c + w^r = \frac{\lambda}{r} \int_{w^r}^{\infty} (x - w^r) \mathrm{d}F(x)$$

where c is search cost, r is the interest rate, and λ is the hazard rate of finding a job.

Flinn and Heckman (1982) assume that one observes the time until finding a job (T_i) and the wage a worker receives conditional on finding the job. The only source of heterogeneity in the model comes from the timing of the job offers and the draw from the wage offer distribution. Clearly one can identify the distribution of accepted wage offers which is the distribution of observed wages. The reservation wage is the lowest acceptable wage, so one can identify w^r as the minimum observed wage. Then they can identify

$$\frac{f(x)}{1 - F(w^r)} \quad \text{for } x \geq w^r.$$

They can also identify the hazard rates of job finding which is

$$\lambda(1 - F(w^r)).$$

However, this is all that can be identified. In particular, one cannot separate λ from $(1 - F(w^r))$. Furthermore, the distribution of wage offers below the reservation wage is not identified. This is quite intuitive. Since nobody works at a salary below the reservation wage, we do not have any information from the data on what that distribution might look like.[22] Furthermore, identification of the model above relies on the strong assumption that people are identical. All dispersion in observed wages comes from identical people with identical skills being offered different wages. It also implies a constant hazard rate of finding jobs λ, which is at odds with the data.

By using exclusion restrictions and using some of the ideas from the Roy model with the arguments from the mixed proportional hazard model, most of the components of the model can be identified. In particular let the arrival rate of job offers be

$$\lambda_i = \phi(X_{\lambda i}, X_{0i})\omega_i \tag{6.9}$$

where now $X_{\lambda i}$ is an exclusion restriction that influences the arrival rate, but not any other aspect of the model. We assume that search cost is defined as

$$\log(C_i) = g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}. \tag{6.10}$$

Finally we assume the wage offer that individual i would receive at time t is

$$\log(W_{fit}) = g_f(X_{fi}, X_{0i}) + \varepsilon_{fit}. \tag{6.11}$$

The complicated aspect of this model is that workers may reject the first offer they receive, and then receive a second different offer. Thus we need the time subscript on ε_{fit} to denote that this draw can be different. The second issue is that one would expect the distribution of offered ε_{fit} to not be identical across workers. We assume that the distribution of ε_{fit} is individual specific coming from distribution $F_{i\varepsilon_f}$. That is each time a worker gets a new offer it is a draw from the distribution of $F_{i\varepsilon_f}$. As above X_i is observable and independent of $(\nu_i, \varepsilon_{fit}, \varepsilon_{hi})$.

Using the Lippman and McCall (1976) model, define W_i^* as the solution to the equation

$$C_i + W_i^* = \frac{\lambda_i}{r} \int_{\log(W_i^*) - g_f(X_{fi}, X_{0i})}^{\infty} (\mathrm{e}^{g_f(X_{fi}, X_{0i}) + \varepsilon_{fit}} - W_i^*) \mathrm{d}F_{i\varepsilon_f}(\varepsilon_{fit}). \tag{6.12}$$

[22] Of course this raises an interesting question. What does it mean for a firm to make an offer that it knows no worker would ever take? In most wage posting models, a firm would never post a wage that no worker would take (see e.g. Burdett and Mortensen, 1998). However, if there is a job match component, one can also write down a model in which one could define the counterfactual wage at which a worker would be paid at a job in which he would never take (whether that offer is actually "extended" or not is largely a semantic issue).

The reservation wage is defined as

$$W_i^r \equiv \max\{W_i^*, 0\}. \tag{6.13}$$

If search costs are sufficiently high, W_i^* could be negative. But because the distribution of wages is bounded below at 0, the reservation wage would be 0.

The added assumptions to identify the model are completely analogous to those we used for the Roy model earlier

Assumption 6.1. $(\varepsilon_{fit}, \varepsilon_{hi}, \nu_i)$ is continuously distributed with support $\mathbb{R}^3$, and is independent of X_i.

Assumption 6.2. $\text{supp}(\phi(X_{\lambda i}, X_{0i}), g_f(X_{fi}, x_0), g_h(X_{hi}, x_0)) = \mathbb{R}^+ \times \mathbb{R}^2$ for all $x_0 \in \text{supp}(X_{0i})$.

Assumption 6.3. The marginal distributions of ε_{fit},ε_{hi}, and ν_i have expected values equal to zero. Moreover, the expected value of $e^{\varepsilon_{fit}}$ is finite.

Assumption 6.4. $X_i = (X_{fi}, X_{hi}, X_{\lambda i}, X_{0i})$ can be written as $(X_{fi}^c, X_{fi}^d, X_{hi}^c, X_{hi}^d, X_{\lambda i}^c, X_{\lambda i}^d, X_{0i}^c, X_{0i}^d)$ where the elements of $X^c = (X_{fi}^c, X_{hi}^c, X_{\lambda i}^c, X_{0i}^c)$ are continuously distributed (no point has positive mass), and $X^d = (X_{fi}^d, X_{fi}^d, X_{\lambda i}^d, X_{0i}^d)$ is distributed discretely (all support points have positive mass).

Assumption 6.5. For any $(x_f^d, x_h^d, x_\lambda^d, x_0^d) \in \text{supp}(X_{fi}^d, X_{hi}^d, X_{\lambda i}^d, X_{0i}^d)$, $g_f(x_f^c, x_f^d, x_0^c, x_0^d)$, $g_h(x_h^c, x_h^d, x_0^c, x_0^d)$, and $\phi(x_\lambda^c, x_\lambda^d, x_0^c, x_0^d)$ are almost surely continuous across $(x^c) \in \text{supp}(X_i^c \mid X_i^d = x^d)$.

Theorem 6.2. *Under Assumptions 6.1–6.5 and that ϕ and the distribution of ω_i satisfy the assumptions in Heckman and Honoré (1989), given that we observe T_i and w_{fiT_i} from the model determined by Eqs (6.9)–(6.13), we can identify ϕ and g_f on their support, and g_h up to location on a set $\mathcal{X}^*$ that has measure* 1.

(Proof in Appendix.)

Unlike some of the other models, we have not completely identified the error structure (or the location of g_h). This is probably not surprising given the complexity of $F_{i\varepsilon_f}$ and the relatively modest data conditions.[23]

[23] Some aspects of the distribution of wages can be identified. For example identification of the marginal distribution of ω_i is straightforward. Describing the distribution of $F_{i\varepsilon_f}$ is difficult because it is a distribution of distributions. Given the cost in setting up notation to discuss this, we do not try to characterize this distribution. A typical assumption would be that we could write $\varepsilon_{fit} = \epsilon_{fi} + \zeta_{fit}$, where ϵ_{fi} is an individual specific term that does not vary across wages and ζ_{fit} is iid.

We conclude this section after making three comments. First, it is not clear that one cares about the location of g_h. That is, for many interesting policy counterfactuals, identification of the aspects above should be sufficient. Second, with more structure, more features of the model should be identified.[24] Third, if a researcher observes multiple spells on the same worker, this can add much identifying information. The identification problem arises because if we see one worker making more than another we do not know if it is because the first worker is more productive or if they just happened to get a fortunate draw from offer distribution. With panel data, if we see that the first worker consistently earns more money across many employers, this would suggest that the difference has more to do with ability than with draws from the offer distribution.

We have barely scratched the surface of identification of search models. Many papers being estimated today are based on equilibrium models such as Mortensen and Pissarides (1994), Burdett and Mortensen (1998), or Postel-Vinay and Robin (2002). We think there is much work to be done on identification in these models.[25]

7. FORWARD LOOKING DYNAMIC MODELS

In this section we discuss an extension of the generalized Roy model into a dynamic framework with uncertainty and forward looking behavior. We show that the basic identification ideas presented above can be generalized to dynamic models. The identification results for the simple models on which we focus can be extended to more complicated environments. We begin with a two period model in which there are three choices made over two periods. We then discuss some general issues with identifying the components of the Bellman Equation. Finally we present a dynamic Generalized Roy model that one can use for dynamic treatment effect evaluation. Once again, we do not provide a full review of the literature, but focus on expanding the generalized Roy model into a forward looking dynamic model. Abbring (2010) includes a more complete discussion.[26]

7.1. Two period discrete choice dynamic model

We begin with the framework of Taber (2000) who considers a simple version of a dynamic model. To think of this model as an extension of the basic Roy model we go from two occupational choices to three. While we could modify the fishing/hunting example to a dynamic context, it is easiest to think about this in terms of an education

[24] Proving identification in nonlinear models such as this one is often quite difficult. This might not be problematic in practice as researchers can search for multiple solutions in the data. If there are multiple solutions, all can be reported. If only one solution exists, this should give a consistent estimate of the truth.

[25] Canals-Cerda (2010) provides a recent example which adds measurement error in wages to the Flinn and Heckman (1982) framework. Barlevy (2008) shows how to non-parametrically identify the wage offer distribution in the presence of measurement error in wages and unobserved heterogeneity in skills.

[26] Recent papers that cover aspects of identification not discussed here include Kashara and Shimotsu (2009) and Hu and Shum (2009).

model as Taber does. In particular, a student first decides whether to graduate from high school or not. After graduating from high school, she decides whether to attend college or enter the labor market directly. Extending beyond 3 choices is straightforward, but as in Taber we stick to the 3 choice model for expositional purposes. We focus on identification of the choice model and ignore data on earnings until Section 7.3.

First consider the case in which there was no uncertainty or dynamics. We specify the model using the three value functions

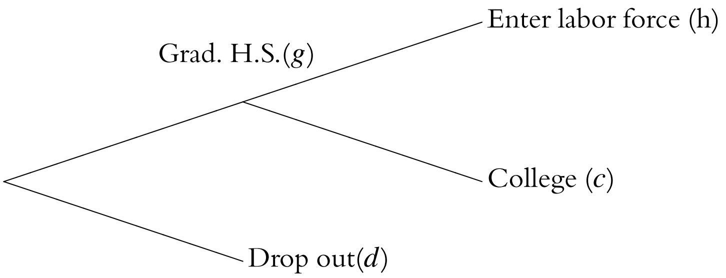
$$\begin{aligned} V_{ci} &= g_c(X_{ci}, X_{0i}) + \varepsilon_{ci} \\ V_{di} &= g_d(X_{di}, X_{0i}) + \varepsilon_{di} \\ V_{hi} &= 0 \end{aligned}$$

where V_{ci} is the value function for a college student, V_{hi} the value function for an individual with exactly a high school degree, and V_{di} the value function for high school dropout. Individuals choose the option with the highest value function. That is

$$J_i = \operatorname{argmax} \{V_{di}, V_{hi}, V_{ci}\}.$$

If there were no uncertainty in this model it would be a simple polychotomous choice model. Matzkin (1993) considers identification a general class of polychotomous choice modes under a number of different assumptions. One result is that since choices are only identified up to monotonic transformations, $V_{hi} = 0$ is a location normalization that we impose at this point. Adding dynamics and uncertainty does not change this result.

Our goal now is to add dynamics and uncertainty to the model. The timing can be seen in the following figure



In the first period the agent chooses whether to graduate from high school. If she graduates in the first period, she then chooses whether to go to college in the second.

The key aspect of the model is that information will be revealed between the first and second period. The agent's preferences are summarized by lifetime reward function V_{ji} at each terminal state $j \in \{c, h, d\}$. Taber defines V_{di} so that it is known at the time the high school graduation choice is made. Then in period two, V_{ci} and V_{hi} are known when the choice between c and h is made. That is, in period one the agent does not know X_{ci} or ε_{ci}. The first period information is assumed to be contained in $(X_{0i}, X_{1i}, \varepsilon_{1i})$, where X_{1i} is observable in period one and will be informative about X_{ci} while ε_{1i} is unobservable and informative about ε_{ci}. We assume that decisions are made in order to maximize expected lifetime reward. Thus the reward function at node g in the first period takes the value

$$V_g(x_1, x_d, x_0, \epsilon_1) \equiv E[\max\{V_{ci}, V_{hi}\} \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0), \varepsilon_{1i} = \epsilon_1].$$

The agent chooses node d if $V_{di} > V_g(X_{1i}, X_{di}, X_{0i}, \varepsilon_{1i})$ and chooses node g otherwise. If she chooses g in the first period she chooses node c in the second if $V_{ci} > V_{hi}$ and node h otherwise.

We let $G(X_{ci} \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0))$ denote the distribution of X_{ci} conditional on $(X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0)$. We can summarize the information structure as follows

Known to the Agent at time one	**Learned by the Agent at time two**	**Observed by the Econometrician**
$\varepsilon_{1i}, \varepsilon_{di}$	ε_{ci}	X_{0i}, X_{1i}, X_{di}
X_{0i}, X_{1i}, X_{di}	X_{ci}	X_{ci}
$G(X_{ci} \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0))$		J_i

We first consider identification of g_c and g_d up to monotonic transformations. We follow Taber (2000) closely except that we use our notation and use stronger assumptions than he does to avoid adding more notation.[27]

Assumption 7.1. For any $(x_c, x_0) \in \text{supp}\,\{X_{ci}, X_{0i}\}$,

$$\text{supp}\{\varepsilon_{di}\} = \mathbb{R} = \text{supp}\{g_d(X_{di}, x_0) \mid (X_{ci}, X_{0i}) = (x_c, x_0)\}$$
$$\text{supp}\{\varepsilon_{ci}\} = \mathbb{R}.$$

This assumption is analogous to what we have been assuming all along. In order to estimate the full model, we need full support of g_d conditional on (X_{ci}, X_{0i}).

[27] Taber (2000) allows for the possibility that the support of the error term could be bounded, which allows for weaker support condition on the observables.

Assumption 7.2. For any $(x_d, x_0) \in \text{supp}\{X_{di}, X_{0i}\}$, $y \in \mathbb{R}$, and $a \in (0, 1)$, there exists a set $\mathcal{X}_1(x_f, x_0, y, a)$ with positive measure such that for $x_1 \in \mathcal{X}_1(x_f, x_0, y, a)$,

(a) $\Pr\left(g_c(X_{ci}, x_0) < y \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0)\right) > a$.

(b) The distribution of $g_c(X_{ci}, x_0)$ conditional on $(X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0)$ is stochastically dominated by the unconditional distribution of $g_c(X_{ci}, x_0)$.

This is a stochastic analogue of a support condition. In the case in which X_{ci} were known at time one so that $X_{1i} = X_{ci}$, this would be implied to be a standard support condition. However, it is general enough to allow for the distribution of X_{ci} to not be known at time one, but we still need a time one variable X_{1i} that is useful in forecasting X_{ci}. For example X_{ci} could be a variable like family income while the child is in college while X_{1i} is a variable like family income while the child is in high school. This assumption states that we can condition on the value of this variable so that the conditional probability that the agent chooses option c in the second period can become arbitrarily small. In the family income example this means we could condition on families whose income while the child is in high school are sufficiently low that college seems like a very unlikely outcome for the child.

Assumption 7.3. $(\varepsilon_{1i}, \varepsilon_{di}, \varepsilon_{ci})$ is independent of $(X_{1i}, X_{di}, X_{ci}, X_{0i})$, for any $\epsilon_1 \in \text{supp}(\varepsilon_{1i})$,

$$E(|\varepsilon_{ci}| \mid \varepsilon_{1i} = \epsilon_1) < \infty$$

and for any $(x_1, x_d, x_0) \in \text{supp}(X_{1i}, X_{di}, X_{0i})$,

$$E\left(|g_c\left(X_{ci}, x_0\right)| \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0)\right) < \infty.$$

Assumption 7.3 is the separable independent assumption that we have been making throughout this chapter. We also need to assume that the stochastic components have finite expectations so that V_g is finite.

Theorem 7.1. *Under Assumptions 7.1–7.3, from data on $(X_{1i}, X_{di}, X_{ci}, X_{0i}, J_i)$, g_d and g_c are identified up to monotonic transformation.*

(Proof in Taber (2000).)

The basic strategy used in this proof is a stochastic extension of "identification at infinity." This should not be surprising as this looks very much like the type of selection problem we have discussed throughout this chapter: we can not observe the choice between c and h unless individuals have already rejected d.

We identify g_c in almost exactly the same way as we identified g_f as presented for the Roy Model. With an exclusion restriction we can condition on g_d arbitrarily low so that the probability of selecting node d is close to zero. This leaves us with a simple binary choice model in which the agents choose between h and c. The type of exclusion restriction used here is a variable that enters g_d, but does not influence g_c directly. One can see this in the following expression

$$\begin{aligned}
&\lim_{g_d(x_d,x_0)\to-\infty} \Pr(J_i = c \mid X_i = x) \\
&\quad = \lim_{g_d(x_d,x_0)\to-\infty} \Pr[g_d(x_d,x_0) + \varepsilon_{di} \le V_g(x_1,x_d,x_0,\varepsilon_{1i}), g_c(x_c,x_0) + \varepsilon_{ci} > 0] \\
&\quad = \Pr[g_c(x_c,x_0) + \varepsilon_{ci} > 0].
\end{aligned}$$

Using standard identification strategies for the binary choice model described in the first step of identification of the Roy model, g_c is identified.

Identification of g_d is somewhat trickier, but one can use essentially the same idea. In a static model one could use an identification at infinity argument by eliminating c as an option and could compare the binary choice of d versus h. In this stochastic case this is can not be done because the value of X_{ci} is not known at time 1. Thus we need a somewhat different type of exclusion restriction, a variable known at time one that does not enter g_d directly, but does have predictive power for the distribution of g_c above and beyond X_{di}. To see how this works, suppose we have a variable X_{1i} that satisfies these conditions and that as x_1 gets small the conditional distribution of g_c shifts to the left. In this case

$$\lim_{x_1\to-\infty} E\left[\max\left(g_c(X_{ci},x_0)+\varepsilon_{ci},0\right) \mid (X_{1i},X_{di},X_{0i}) = (x_1,x_d,x_0), \varepsilon_{1i} = \epsilon_1\right] = 0,$$

so that

$$\begin{aligned}
&\lim_{x_1\to-\infty} \Pr(J_i = d \mid X_i = x) \\
&\quad = \lim_{x_1\to-\infty} \Pr\left[g_d(x_d,x_0) + \varepsilon_{di} > E\left[\max\left(V_{ci},0\right) \mid (X_{1i},X_{di},X_{0i})\right.\right. \\
&\qquad \left.\left. = (x_1,x_d,x_0), \varepsilon_{1i} = \epsilon_1\right]\right] \\
&\quad = \Pr[g_d(x_d,x_0) + \varepsilon_{di} > 0].
\end{aligned}$$

From this piece we can identify g_d up to a monotonic transformation. This type of variable will satisfy Assumption 7.2. Note that the type of exclusion restriction we need here is something that is known at time 1, is useful in forecasting X_{ci}, but does not affect V_{di}.

Taber (2000) goes on to consider identification of the distribution of the error terms. The most general version of the full model above can not be identified without further assumptions, so he instead studies a few interesting cases. Identification of the error terms requires a different kind of exclusion restriction. His key assumption requires variation in $g_c(x_c)$ holding x_1 fixed. Thus we need some uncertainty from the point of view of the agents. The full model is not identified if agent's have perfect information about future values of X_{ci}. A natural way to satisfy this exclusion restriction is with time varying observables. The details can be found in Taber (2000).

7.2. Identification of the components of the Bellman equation

While the model above is dynamic, we have not used Bellman's equation. A natural way to parameterize the model would be to define period specific utility functions $u_h\left(X_{hi}, X_{0i}, \varepsilon_{hi}\right)$, $u_c\left(X_{ci}, X_{0i}, \varepsilon_{ci}\right)$, and $u_g(X_{1i}, X_{0i}, \varepsilon_{1i})$ in each of the three nodes above other than the dropout node. If we think of the model as a two period model we can define $u_d(t, X_{di}, X_{0i}, \varepsilon_{di})$ to be the period specific utility of individual i if she drops out at time t. Conditional on graduating, she enters college if

$$u_c(X_{ci}, X_{0i}, \varepsilon_{ci}) > u_h(X_{hi}, X_{0i}, \varepsilon_{hi}).$$

The Bellman equation for the high school graduate is

$$\begin{aligned} V_g(x_1, x_d, x_0, \epsilon_1) \equiv u_g(x_1, x_0, \epsilon_1) + \beta E[\max\{u_c(X_{ci}, X_{0i}, \varepsilon_{ci}), \\ u_h(X_{hi}, X_{0i}, \varepsilon_{hi})\} \mid (X_{1i}, X_{di}, X_{0i}) = (x_1, x_d, x_0), \varepsilon_{1i} = \epsilon_1]. \end{aligned}$$

Mapping back to the notation in the subsection above, the rest of the value functions are defined as

$$\begin{aligned} V_{di} &= u_d(1, X_{di}, X_{0i}, \varepsilon_{di}) + \beta u_d(2, X_{di}, X_{0i}, \varepsilon_{di}) \\ V_{hi} &= u_g(X_{1i}, X_{0i}, \varepsilon_{1i}) + \beta u_h(X_{hi}, X_{0i}, \varepsilon_{hi}) \\ V_{ci} &= u_g(X_{1i}, X_{0i}, \varepsilon_{1i}) + \beta u_c(X_{ci}, X_{0i}, \varepsilon_{ci}). \end{aligned}$$

An obvious question arises as to whether one can separately identify the components of the value functions β, u_h, u_c, and u_d. Unfortunately, in general one can not do this. Consider a full certainty version of the model. In this case the decision of which occupation to enter would depend on V_{di}, V_{hi}, and V_{ci} only. One can choose any $\beta > 0$ and any u_g, but then always find a value of u_c and u_h to leave V_{ci} and V_{hi} unchanged. For a simple model such as the one Taber (2000) presents, parameterizing the model in terms of the terminal value functions (i.e. V_{di}, V_{hi}, and V_{ci}) avoids this problem as one does not need to decompose them into their components.

However, Taber's parameterization is clearly not feasible for an infinitely lived model. Furthermore, it is not convenient in an finite time model with many periods and state variables. It does not take advantage of the dimension reducing advantages of the Bellman formulation: the functions would depend on the whole history of state variables rather than just the current set.

Next we consider Rust's (1994) model. Note that we use his notation exactly even though it is inconsistent with our previous notation. Let S_i represent the current state and D_i represent the discrete choice. In general S_i will contain elements that are both observed and unobserved by the econometrician. He writes the Bellman equation as

$$v(s,d) = u(s,d) + \beta \int \max_{D_i' \in D(S_i')} [v(S_i', D_i')] p(dS_i' \mid S_i = s, D_i = d)$$

where v is the value function, u is the period specific utility function, β is the discount rate, $D(s)$ is the choice set in state of the world s, and p is the transitional probability distribution of the state variables. Rust (1994) shows that one can not separately identify the model above from an alternative with the same β and p, but with

$$\bar{u}(s,d) = u(s,d) + f(s) - \beta \int f(S_i') p(ds' \mid S_i = s, D_i = d).$$

Intuitively this is close to the discussion above in the simple model in which you can change the timing at which the innovation to utility takes place, without changing the value function.

Magnac and Thesmar (2002) discuss this issue in much greater detail. They not only show that the model is not identified, but document the extent of underidentification. They additionally assume that one can write

$$u(S_i, d) = u_d(X_i) + \varepsilon_{di}$$

where X_i is the observable part of the state space and the unobservable ε_{di} is mean independent of x and independent across periods (conditional on x and d). That is S_i represents the state space, so if one knows S_i, they also know X_i and ε_{di}. They show that given knowledge of β and the joint distribution of the ε_{di}, one can identify

$$\begin{aligned} U_d(x) \equiv u_d(x) + \beta \int \max_{D_i' \in D(D_i')} [v(S_i', D_i')] p(dS_i' \mid X_i = x, D_i = d) - u_k(x) \\ + \beta \int \max_{D_i' \in D(S_i')} [v(S_i', D_i')] p(dS_i' \mid X_i = x, D_i = k) \end{aligned}$$

where k is one of the elements of $D(s)$. They further explore the model with additional identifying information and correlated random effects.

How problematic is it that the model is not fully identified? The answer to this question depends on the purpose of the model. That is, even if the model is not fully identified, one may still be able to identify policy counterfactuals of interest. Ichimura and Taber (2002) provide one example of a case in which the policy counterfactual can be identified. They start with the model of Keane and Wolpin (2001) and show how one can estimate a semiparametric reduced form version of this model and use it to evaluate the effect of a tuition subsidy on college enrollment. They key is having enough structure on the model to map variation in the data to the counterfactual tuition subsidy.

Aguirregabiria (2010) presents a different and somewhat more general example of policy evaluation in a finite time dynamic discrete choice model. We do not get into the details as it is different from the types of labor models we study here, but he shows that, despite the fact that his full model is not identified, the welfare effect function resulting from the policy change can be identified. Thus one can do welfare analysis even though the full model is not identified.

7.3. Dynamic generalized Roy model

Heckman and Navarro (2007) provide another example showing that one can identify interesting counterfactuals even when the full model is not identified. Their study complements the discussion in this chapter as it extends the work on identification in dynamic discrete choice models into the treatment effects literature discussed in Section 5 above. They consider a finite time optimal stopping problem. Using the notation used above in Section 7.2, D_i is either zero or one, and once it is one it remains one forever. Their main example is a schooling model in which students decide at which time to leave school (assuming that after leaving they cannot come back). The model is essentially a dynamic generalized Roy model. Let T_{ia} and L_{ia} respectively denote the level of schooling and a dummy for whether individual i is out of school at age a. Using a somewhat modified version of their notation we can write time a earnings as

$$Y_{i,a,t,\ell} = \mu(a, t, \ell, X_i) + \varepsilon_{i,a,t,\ell}$$

where t and ℓ represent potential outcomes of $T_{i,a}$ and $L_{i,a}$. Heckman and Navarro (2007) also assume that the cost of schooling can be written as

$$C_{i,t} = \Phi(t, X_i, Z_i) + \omega_{i,t}.$$

In order to keep our notation complete and consistent across sections we will assume that random variable $\Theta_{i,a}$ summarizes all information (both observables and unobservables) that individual i has at age a. This means that if we know $\Theta_{i,a}$ we also know $(X_i, Z_i, T_{i,a}, L_{i,a}, \varepsilon_{i,a,t,l}, \omega_{i,t})$, so when we condition on $\Theta_{i,a} = \theta$, we are conditioning on $(X_i, Z_i, T_{i,a}, L_{i,a}, \varepsilon_{i,a,t,l}, \omega_{i,t}) = (x, z, t, \ell, \epsilon_{a,t,\ell}, \omega_t)$. We will make use of this notation below.

Once a student leaves school they make no further decisions, so if a student leaves school at age a with t years of schooling, lifetime utility discounted to the time one leaves school is written as

$$R(a,t,\theta)=E\left(\sum_{j=0}^{\bar{T}}\left(\frac{1}{1+r}\right)^{j}Y_{i,a+j,t,1}\mid\Theta_{i,a}=\theta\right).$$

The only decision that agents make is whether they will drop out of school or not. For a student at age a with t years of schooling the value function when they make this decision is written as

$$\begin{aligned}V(a,t,\theta)=\max\Big\{&R(a,t,\theta),\mu(a,t,0,x)+\epsilon_{a,t,0}-\Phi(t,x,z)-\omega_t\\&+\left(\frac{1}{1+r}\right)E\left[V(a+1,t+1,\Theta_{i,a+1})\mid\Theta_{i,a}=\theta\right]\Big\}.\end{aligned}$$

This is basically a dynamic version of the generalized Roy model. Identification follows by essentially combining the arguments used by Taber (2000) for the dynamic aspects of the model with the arguments for identification of the generalized Roy model. Heckman and Navarro (2007) use higher level assumptions to avoid the use of exclusion restrictions.[28] They also use a factor structure on the distribution of the error term to reduce dimension. We refer readers interested in these generalizations and in the details of their proof to their paper. Here we attempt to give an intuitive feel for identification of this model and show how it is related to identification of the generalized Roy model presented in Section 3.3.

Identification of reduced form choice model

In this case they do not derive an explicit reduced form, but note that

$$\Pr(T_{i,a}=t\mid X_i=x,Z_i=z)$$

can be identified directly from the data.

Identification of the earnings equation μ

With exclusion restrictions this can be done in exactly the same way as in the static model. Assuming that $\varepsilon_{i,a,t,\ell}$ has a zero mean,

$$\lim_{\Pr(T_{i,a}=t\mid(X_i,Z_i)=(x,z))\to 1}E\left[Y_{i,a+j,t,1}\mid(X_i,Z_i)=(x,z)\right]=\mu(a+j,t,1,x).$$

[28] This relates back to our discussion of identification and exclusion restrictions in the sample selection model at the very end of Section 3. Exclusion restrictions prevent one from setting $\widetilde{g}_f(x)=g_f(x)+h(g(x))$ but shape restrictions on g and g_f can do this as well. Their "higher level assumptions" are essentially assuming that we make restrictions on g_f so that we can not add $h(g(x))$ to it and remain in the permissible class of g_f functions.

$$\lim_{\Pr(T_{i,a}>t\mid(X_i,Z_i)=(x,z))\rightarrow 1} E\left[Y_{i,a,a,0} \mid (X_i, Z_i) = (x, z)\right] = \mu(a, a, 0, x).$$

Thus this is a version of an "identification at infinity argument." Heckman and Navarro (2007) do not use this explicit argument because they avoid exclusion restrictions with a higher order assumption. However, they do use identification at infinity.

Identification of Φ

Next consider the identification of the cost of schooling function Φ. The best way to think about identification in these types of models is to start with the final period and work backward.

Since the maximum length of schooling is $\bar{T}$, the final decision is made when the individual has $\bar{T} - 1$ years of schooling. At that point the student decides whether to attend the final year of school or not. Heckman and Navarro (2007) use an "identification at infinity" argument so that $\Pr(T_i > \bar{T} - 2 \mid X_i = x, Z_i = z) \approx 1$. Then the problem becomes analogous to a static problem.[29] That is

$$\begin{aligned}
&\lim_{\Pr(T_i>\bar{T}-2\mid X_i=x,Z_i=z)\rightarrow 1} \Pr(T_{i\bar{T}} = \bar{T} \mid X_i = x, Z_i = z) \\
&= \Pr\Bigg(R(\bar{T}-1, \bar{T}-1, \Theta_{i,\bar{T}-1}) < \mu(\bar{T}-1, \bar{T}-1, 0, x) + \varepsilon_{i,\bar{T}-1,\bar{T}-1,0} \\
&\quad - \Phi(\bar{T}-1, x, z) - \omega_{i,\bar{T}-1} + \left(\frac{1}{1+r}\right) \\
&\quad \times\ E\left[R(\bar{T}, \bar{T}, \Theta_{i\bar{T}}) \mid \Theta_{i,\bar{T}-1}\right] \mid X_i = x, Z_i = z\Bigg).
\end{aligned}$$

This is analogous to identification of the g_h function in the Roy model.[30]

Now one can just iterate backward given knowledge of all variables at $\bar{T}$ and $\bar{T} - 1$. That is, the distribution of $(\frac{1}{1+r})E\left[V(\bar{T}-1, \bar{T}-1, \Theta_{i,\bar{T}-1}) \mid \Theta_{i,\bar{T}-2}\right]$ has been identified so once again we can use the identification approach of the static problem and can use the same basic style of proof. That is we can condition on a set of variables so that $\Pr(t > \bar{T} - 2 \mid X_i = x, Z_i = z) \approx 1$ so that identification is analogous to the static problem. Consider the decision with $\bar{T} - 2$ years of schooling.

[29] Once again, Heckman and Navarro (2007) use higher order assumptions that do not require exclusion restrictions. For example they allow for either an exclusion restriction or a cost variable to identify the scale (such as tuition described in Section 4 above).

[30] Note that we have violated one convention in this chapter which is to make conditioning explicit such as $E(\cdot \mid X_i = x)$. When we condition on $\Theta_{i,\bar{T}-1}$ we cannot do this explicitly because while the expectation inside the expression conditions on its outcome, the probability expression (immediately after the = sign) treats $\Theta_{i,\bar{T}-1}$ as a random variable.

$$
\begin{aligned}
&\lim_{\Pr(T_i > \bar{T}-3 \mid X_i = x, Z_i = z) \to 1} \Pr(T_{i,\bar{T}-1} = \bar{T} - 1 \mid X_i = x, Z_i = z) \\
&\quad = \Pr\Bigg(R(\bar{T}-2, \bar{T}-2, \Theta_{i,\bar{T}-2}) \\
&\quad < \mu(\bar{T}-2, \bar{T}-2, 0, x) + \varepsilon_{i,\bar{T}-2,\bar{T}-2,0} - \Phi(\bar{T}-2, x, z) \\
&\quad - \omega_{i,\bar{T}-2} + \left(\frac{1}{1+r}\right) E[V(\bar{T}-1, \bar{T}-1, \Theta_{i,\bar{T}-1}) \mid \Theta_{i,\bar{T}-2}] \mid X_i = x, Z_i = z \Bigg).
\end{aligned}
$$

One can keep iterating on this procedure so that Φ is identified in all periods.

Identification of the distribution of the error terms

Heckman and Navarro (2007) impose a factor structure so that

$$
\begin{aligned}
\varepsilon_{i,a,t,\ell} &= \alpha'_{a,t,\ell}\tau_i + \varepsilon_{i,a,t,\ell} \\
\omega_{i,t} &= \lambda'_t \tau_i + \xi_{i,t}
\end{aligned}
$$

where τ_i is a vector random variable, the $\varepsilon' s$ and $\xi' s$ are all independently distributed, and the α and λ terms are factor loadings. Given this structure and that the other components of the model have been identified, identification of the distribution of the error terms and factor loadings can be done by varying the indices in much the same way as in the static model. We do not show this explicitly.

8. CONCLUSIONS

In this chapter we have presented identification results for models of the labor market. The main issue in all of these models is the issue of sample selection bias. We start with the classic Roy model and devote much space to explaining how this model can be identified. We then show how these results can be extended to more complicated cases, the generalized Roy model, treatment effect models, duration data, search models, and forward looking dynamic models. We show the importance of both exclusion restrictions and support conditions for all of these models.

TECHNICAL APPENDIX

Proof of Theorem 2.1. Let $\mathcal{X}^*$ be the set of points (x^c, x^d) at which g is continuous in x^c. For any $(x^c, x^d) \in \mathcal{X}^*$ and $\delta > 0$, $E(Y_i \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d)$ is identified directly from the data.

Since g is continuous at (x^c, x^d),

$$
\lim_{\delta \downarrow 0} E(Y_i \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d) = g(x^c, x^d),
$$

so $g(x^c, x^d)$ is identified on $\mathcal{X}^*$. By Assumption 2.2, $\mathcal{X}^*$ has measure one. □

Proof of Theorem 3.1. Let $\mathcal{X}^*$ be the set of points $(x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d)$ at which g_h and g_f are continuous in x^c.

First notice that for any $x = (x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d) \in \mathcal{X}^*$,

$$\begin{aligned}\lim_{\delta \downarrow 0} \Pr(J_i = f \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d) &\equiv \Pr(J_i = f \mid X_i = x) \\ &= g(x)\end{aligned}$$

is identified.

Thus we have thus established that we can write the model as $J_i = f$ if and only if $g(X_i) > \varepsilon_i$, where ε_i is uniform $[0, 1]$ and that g is identified.

Next consider identification of g_f at the point (x_f, x_0). This is basically the standard selection problem. As long as g is continuous on the continuous covariates at this point, we can identify

$$\begin{aligned}&\lim_{\delta \downarrow 0} \mathrm{Med}(Y_i \mid \|X_{fi}^c - x_f^c\| < \delta, X_{fi}^d = x_f^d, \|X_{0i}^c - x_0^c\| < \delta, \\ &\quad X_{0i}^d = x_0^d, |1 - g(X_i)| < \delta, J_i = f) \\ &\quad = g_f(x_f, x_0) + \lim_{\delta \downarrow 0} \mathrm{Med}(\varepsilon_{fi} \mid \|X_{fi}^c - x_f^c\| < \delta, X_{fi}^d = x_f^d, \|X_{0i}^c - x_0^c\| < \delta, \\ &\qquad\quad X_{0i}^d = x_0^d, |1 - g(X_i)| < \delta, J_i = f) \\ &\quad = g_f(x_f, x_0).\end{aligned}$$

Thus g_f is identified. Note that having an exclusion restriction with strong support conditions is necessary to guarantee that the measure of the set of X_i satisfying $|1 - g(X_i)| < \delta$ is not zero.

Next we show how to identify g_h. Note that for any (x_h, x_0) where g is continuous in the continuous covariates and $\delta > 0$ we can identify the set

$$\begin{aligned}\mathcal{X}(x_h, x_0, \delta) \equiv \{\widetilde{x} \in \mathcal{X}^* : \|\tilde{x}_h^c - x_h^c\| < \delta, \\ \tilde{x}_h^d = x_f^d, \|\tilde{x}_0^c - x_0^c\| < \delta, \tilde{x}_{0i}^d = x_0^d, |0.5 - g(\widetilde{x})| < \delta\}\end{aligned}$$

where $\tilde{x} = (\tilde{x}_f, \tilde{x}_h, \tilde{x}_0)$. Under our assumptions it has positive measure.

The median zero assumption guarantees that

$$\begin{aligned}\lim_{\delta \downarrow 0} \mathcal{X}(x_h, x_0, \delta) &= \left\{\tilde{x} \in \mathcal{X}^* : \tilde{x}_h = x_h, \tilde{x}_0 = x_0, 0.5 = \Pr(J_i = F \mid X_i = \tilde{x})\right\} \\ &= \left\{\tilde{x} \in \mathcal{X}^* : \tilde{x}_h = x_h, \tilde{x}_0 = x_0, 0.5 = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(\tilde{x}_f, x_0) - g_h(x_h, x_0))\right\} \\ &= \left\{\tilde{x} \in \mathcal{X}^* : \tilde{x}_h = x_h, \tilde{x}_0 = x_0, g(\tilde{x}_f, x_0) = g_h(x_h, x_0)\right\}\end{aligned}$$

is identified. Since $g(\tilde{x}_f, x_0)$ is identified, g_h is identified.

Finally consider identification of G given g_f and g_h. Note that from the data one can identify

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = f, \log(Y_{fi}) < s \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d) \\
&\quad = \lim_{\delta \downarrow 0} \Pr(g_h(X_{hi}, X_{0i}) + \varepsilon_{hi} \leq g_f(X_{fi}, X_{0i}) + \varepsilon_{fi}, g_f(X_{fi}, X_{0i}) + \varepsilon_{fi} \\
&\qquad \leq s \mid \left\|X_i^c - x^c\right\| < \delta, X_i^d = x^d) \\
&\quad = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(x_f, x_0) - g_h(x_h, x_0), \varepsilon_{fi} \leq s - g_f(x_f, x_0))
\end{aligned}$$

which is the cumulative distribution function of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$ evaluated at the point $(g_f(x_f, x_0) - g_h(x_h, x_0), s - g_f(x_f, x_0))$. By varying the point of evaluation one can identify the joint distribution of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$ from which one can derive the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$. □

Proof of Theorem 4.1. As in the proof of Theorem 3.1, let $\mathcal{X}^*$ be the set of points $(z^c, z^d, x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d)$ at which g_h, g_f, φ_h and φ_f are continuous in $(z^c, z^d, x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d)$.

First notice that for any $(z, x) = (z^c, z^d, x_f^c, x_f^d, x_h^c, x_h^d, x_0^c, x_0^d) \in \mathcal{X}^*$,

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = f \mid \|X_i^c - x^c\| < \delta, \|Z_i^c - z^c\| < \delta, (Z_i^d, X_i^d) = (z^d, x^d)) \\
&\quad = \Pr(\nu_i \leq \varphi(z, x)) \\
&\quad = \varphi(z, x).
\end{aligned}$$

Thus φ is identified on the relevant set. Next consider g_f and the joint distribution of $(\nu_i, \varepsilon_{fi})$. Note that for all $(z, x_f, x_h, x_0) \in \mathcal{X}^*$ and any $y \in \mathbb{R}$, we can identify

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = f, Y_{fi} \leq y \mid \|X_i^c - x^c\| < \delta, \|Z_i^c - z^c\| < \delta, (Z_i^d, X_i^d) = (z^d, x^d)) \\
&\quad = \Pr(\nu_i \leq \varphi(z, x), g_f(x_f, x_0) + \varepsilon_{fi} \leq y)
\end{aligned}$$

which is the joint distribution of $(\nu_i, g_f(x_f, x_0) + \varepsilon_{fi})$ evaluated at $(\varphi(z, x), y)$. Holding (x_f, x_0) constant and varying $(\varphi(z, x), y)$ we can estimate this joint distribution. Since the median of ε_{fi} is zero, g_f is identified and given g_f the joint distribution of $(\nu_i, \varepsilon_{fi})$ is identified. Since the model is symmetric in h and f, g_h and the joint distribution of $(\nu_i, \varepsilon_{hi})$ are identified using the analogous argument. □

Proof of Theorem 4.2. The first part is analogous to step three of identification of the Roy model presented in the text. Note that for any (z, x_0) and δ we can identify the set

$$\begin{aligned}
\mathcal{X}(z, x_0, \delta) \equiv \{(\tilde{z}, \tilde{x}) \in \mathcal{X}^* : \|\tilde{z}^c - z^c\| < \delta, \tilde{z}^d = z^d, \|\tilde{x}_0^c - x_0^c\| < \delta, \\
\tilde{x}_0^d = x_0^d, |0.5 - \varphi(\tilde{z}, \tilde{x})| < \delta\}
\end{aligned}$$

and it has positive measure where the elements of $(\tilde{z}, \tilde{x})$ are defined in the obvious way. The median zero assumption guarantees that

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \mathcal{X}(z, x_0, \delta) \\
&\quad = \{(\tilde{z}, \tilde{x}) \in \mathcal{X}^* : \tilde{z} = z, \tilde{x}_0 = x_0, 0.5 = \Pr(J_i = F \mid (Z_i, X_i) = (\tilde{z}, \tilde{x}))\} \\
&\quad = \{(\tilde{z}, \tilde{x}) \in \mathcal{X}^* : \tilde{z} = z, \tilde{x}_0 = x_0, 0.5 = \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(\tilde{x}_f, x_0) \\
&\qquad + \varphi(z, x_0) - g_h(\tilde{x}_h, x_0)) - \varphi(z, x_0)\} \\
&\quad = \{(\tilde{z}, \tilde{x}) \in \mathcal{X}^* : \tilde{z} = z, \tilde{x}_0 = x_0, \varphi_f(z, x_0) - \varphi_h(z, x_0) \\
&\qquad = g_h(\tilde{x}_h, x_0) - g_f(\tilde{x}_f, x_0)\}.
\end{aligned}$$

Since g_h and g_f are identified by Theorem 4.1, $\varphi_f(z, x_0) - \varphi_h(z, x_0)$ is also identified. Given this we can identify the distribution of $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{fi})$ and $(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}, \varepsilon_{hi})$ since in general

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = f, Y_{fi} \leq y \mid \|Z_i^c - z^c\| < \delta, Z_i^d = z^d, \|X_i^c - x^c\| < \delta, X_i^d = x_0^d) \\
&\quad = \Pr(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi} \leq g_f(x_f, x_0) + \varphi_f(z, x_0) - g_h(x_h, x_0) \\
&\qquad - \varphi_h(z, x_0), \varepsilon_{fi} \leq y - g_f(x_f, x_0)),
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{\delta \downarrow 0} \Pr(J_i = r, Y_{hi} \leq y \mid \|Z_i^c - z^c\| < \delta, Z_i^d = z^d, \|X_i^c - x^c\| < \delta, X_i^d = x_0^d) \\
&\quad = \Pr(-\left(\varepsilon_{hi} + \nu_{hi} - \varepsilon_{fi} - \nu_{fi}\right) \leq g_h(x_h, x_0) + \varphi_h(z, x_0) - g_f(x_f, x_0) \\
&\qquad - \varphi_f(z, x_0), \varepsilon_{hi} \leq y - g_h(x_h, x_0)). \quad \square
\end{aligned}$$

Proof of Theorem 5.1. Theorem 4.1 shows that the marginal distributions of ε_{fi} and ε_{hi} are identified. Since their expectations are finite, $E(\varepsilon_{fi})$ and $E(\varepsilon_{hi})$ are identified. We also showed that g_f and g_h are identified over a set of measure 1. Note that $E(\pi_i) = E(Y_{fi}) - E(Y_{hi}) = E(g_f(X_{fi}, X_{0i}) + \varepsilon_{fi}) - E(g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}) = g_f(X_{fi}, X_{0i}) - g_h(X_{hi}, X_{0i}) + E(\varepsilon_{fi}) - E(\varepsilon_{hi})$. Because all the components of $E(\pi_i)$ are identified, $E(\pi_i)$ is identified as well. □

Proof of Theorem 5.2. The marginal distribution of X_i, the joint distribution of (X_i, Y_{fi}) conditional on $J_i = f$ and the joint distribution of (X_i, Y_{hi}) conditional on $J_i = h$ are identified directly from the data. Assumption 5.2 guarantees that for both fishing and hunting ($j \in \{f, h\}$), the conditional distribution of Y_{ji} conditional on X_i and $J_i = j$ is the same as the conditional distribution of Y_{ji} conditional on X_i alone. From each of these conditional distributions and the marginal distribution of X_i, one can identify $E(Y_{ji})$, and thus the average treatment effect is identified by taking the difference between the two. □

Proof of Theorem 6.2. Let $\mathcal{X}^*$ be the set of points (x^c, x^d) at which the functions are all continuous in x^c.

First note that in this model the hazard rate of finding for any individual can be written as

$$\phi(X_{\lambda i}, X_{0i})\nu_i[1 - F_{i\varepsilon_f}(\log(W_i^r) - g_f(X_{fi}, X_{0i}))].$$

Our first goal is for any $(x_f, x_{\lambda,} x_0) \in \mathcal{X}^*$, to identify the values of x_h that send $g_h(x_h, x_0)$ arbitrarily large so that all offers are accepted. Since the reservation wage is strictly decreasing in g_h, the hazard rate is strictly increasing in g_h, we can do this by fixing (X_{fi}, X_{0i}) within some neighborhood of (x_f, x_0) and finding the value of x_h that minimizes the job finding rate.

More formally for any (x_f, x_λ, x_0) and δ,define

$$x_h(\delta) \equiv \text{argmin} E(T_i \mid \|X_i^c - (x_f^c, x_h^c(\delta), x_\lambda^c, x_0^c)\| < \delta, X_i^d = (x_f^d, x_h^d(\delta), x_\lambda^d, x_0^d)).$$

Note that this minimum will be such that as $\delta \to 0$, $W_i^r \to 0$ so that

$$\lim_{\delta \downarrow 0} \Pr(\log(T_i) < t, \log(W_{fit}) < w \mid \|X_i^c - (x_f^c, x_h^c(\delta), x_\lambda^c, x_0^c)\| < \delta, X_i^d$$
$$= (x_f^d, x_h^d(\delta), x_\lambda^d, x_0^d)) = G_{\omega^*,\varepsilon}(t + \log(\phi(x_\lambda, x_0)), w - g_f(x_f, x_0))$$

where G is the joint distribution between a convolution of ω_{it} and an extreme value and of ε_{fit}. Given G, applying the identification arguments for the mixed proportional hazard model one can identify ϕ. Furthermore, g_f can be identified through the standard argument for identification of the regression model.

Finally, recovering g_h can be done in an analogous way as for the Roy model. Notice that the reservation wage is scalable so that if we increase both C_i and W_{it} by 10%, then the reservation wage increases by 10% and the probability of job acceptance does not change. That is for any $\delta > 0$ if w_i^* solves

$$e^{g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}} + w_i^* = \frac{\lambda_i}{r} \int_{\log(w_i^*) - g_f(X_{fi}, X_{0i})}^{\infty} (e^{g_f(X_{fi}, X_{0i}) + \varepsilon_{fit}} - w_i^*) dF_{i\varepsilon_f}(\varepsilon_{fit})$$

then $w_i^* e^\delta$ solves

$$e^{g_h(X_{hi}, X_{0i}) + \delta + \varepsilon_{hi}} + w_i^* e^\delta$$
$$= \frac{\lambda_i}{r} \int_{\log(w_i^*) - g_f(X_{fi}, X_{0i})}^{\infty} (e^{g_f(X_{fi}, X_{0i}) + \delta + \varepsilon_{fit}} - w_i^* e^\delta) dF_{i\varepsilon_f}(\varepsilon_{fit}),$$

but the probability of accepting a job and thus the expected duration remains the same.

Thus as in the identification of the slope that we discuss in Step 2 of the identification of the Roy model, for any (x_h, x_0) and $(\tilde{x}_h, \tilde{x}_0)$ suppose we want to identify

$g_h\left(x_h, x_0\right) - g_h\left(\tilde{x}_h, \tilde{x}_0\right)$. Fix x_λ and $\tilde{x}_\lambda$ so that $\phi(x_\lambda, x_0) = \phi(\tilde{x}_\lambda, \tilde{x}_0)$. Then the key here is finding values x_f and $\tilde{x}_f$ so that

$$\begin{aligned}&\lim_{\delta\downarrow 0} E(\log(Z(T_i)) \mid \|X_i^c - x^c\| < \delta, X_i^d = x^d)\\&\quad= \lim_{\delta\downarrow 0} E(\log(Z(T_i)) \mid \|X_i^c - \tilde{x}^c\| < \delta, X_i^d = \tilde{x}^d).\end{aligned}$$

But if this is the case it must be that

$$g_f(x_f, x_0) - g_h(x_h, x_0) = g_f(\tilde{x}_f, \tilde{x}_0) - g_h(\tilde{x}_h, \tilde{x}_0)$$

but then

$$g_h(x_h, x_0) - g_h(\tilde{x}_h, \tilde{x}_0) = g_f(x_f, x_0) - g_f(\tilde{x}_f, \tilde{x}_0)$$

where the right hand side has already been identified. Thus g_h is identified up to location on the set $\mathcal{X}^*$. □

REFERENCES

Abbring, J., 2010. Identification of dynamic discrete choice models. Annual Review of Economics 2, 367–394.

Abbring, J., Ridder, G., 2009. A note on the non-parametric identification of generalized accelerated failure-time models. Unpublished manuscript, Tilburg University.

Abbring, J., Heckman, J., 2007. Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation. In: Heckman, Leamer (Eds.), Handbook of Econometrics. North Holland, Amsterdam, pp. 5145–5303.

Abbring, J., van den Berg, G., 2003. The identifiability of the mixed proportional hazards competing risks model. Journal of the Royal Statistical Society, Series B (Statistical Methodology) 3, 701–710.

Aguirregabiria, V., 2010. Another look at the identification of dynamic discrete decision processes: an application to retirement behavior. Journal of Business and Economic Statistics 28, 201–218.

Altonji, J., Elder, T., Taber, C., 2005a. Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. Journal of Political Economy 113.

Altonji, J., Elder, T., Taber, C., 2005b. An evaluation of instrumental variable strategies for estimating the effects of Catholic schooling. Journal of Human Resources.

Angrist, J., Imbens, G., 1999. Comment on James J. Heckman, "Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations". The Journal of Human Resources 34 (4), 823–827.

Angrist, J., Imbens, G., Rubin, D., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91 (June).

Angrist, J., Pischke, S., 2010, The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. NBER working paper 15794.

Barlevy, G., 2008. Identification of search models using record statistics. Review of Economic Studies 75 (1), 29–64.

Bloom, H., Orr, L., Bell, S., Cave, G., Doolittle, F., Lin, W., Bos, J.s, 1997. The benefits and costs of JTPA title II-A programs: key findings from the national job training partnership act study. Journal of Human Resources 32 (3), 549–576.

Blundell, R., Gosling, A., Ichimura, H., Meghir, C., 2007. Changes in the distribution of male and female wages accounting for employment composition using bounds. Econometrica 75, 323–363.

Buera, F., 2006. Non-parametric identification and testable implications of the Roy model. Unpublished manuscript, UCLA.

Burdett, K., Mortensen, D.T., 1998. Wage differentials, employer size and labor market equilibrium. International Economic Review 39, 257–273.

Canals-Cerda, J., 2010. Identification in empirical search models when ages are measured with errors. unpublished manuscript. Federal Reserve Bank of Philadelphia.

Carneiro, P., Heckman, J., Vytlacil, E. 2010. Estimating marginal returns to education. unpublished manuscript, University College London.

Carneiro, P., Lee, S., 2009. Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. Journal of Econometrics 149, 191–208.

Chamberlain, G., 1986. Asymptotic efficiency in semiparametric models with censoring. Journal of Econometrics 32, 189–218.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. In: Handbook of Econometrics. North-Holland (Chapter 76).

Das, M., Newey, W., Vella, F., 2003. Nonparametric estimation of sample selection models. The Review of Economic Studies 70 (1), 33–58.

Davidson, J., 1994. Stochastic Limit Theory. Oxford University Press, Oxford.

Deaton, A.. 2009. Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development. NBER working paper 14690.

DiNardo, J., Lee, D., 2011. Program evaluation and research designs. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. 4a. Elsevier Science, pp. 463–536.

Doyle, J., 2007. Child protection and child outcomes: measuring the effects of foster care. The American Economic Review 97 (5), 1583–1610.

Eckstein, Z., van den Berg, G, 2007. Empirical labor search: a survey. Journal of Econometrics 136, 531–564.

Elbers, C., Ridder, G., 1982. True and spurious duration dependence: the identifiability of the proportional hazard model. Review of Economic Studies 64, 403–409.

Evans, W., Schwab, R., 1995. Finishing high school and starting college: do Catholic schools make a difference?. Quarterly Journal of Economics 110, 947–974.

Flinn, C., Heckman, J., 1982. New methods for analyzing structural models of labor force dynamics. Journal of Econometrics 18, 115–168.

French, E., Song, J. 2010. The effect of disability insurance receipt on labor supply. unpublished manuscript. Federal Reserve Bank of Chicago.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–162.

Heckman, J., 1990. Varieties of selection bias. American Economic Review 80.

Heckman, J., 1997. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. The Journal of Human Resources 32 (3), 441–462.

Heckman, J., 1999. Instrumental variables: Response to Angrist and Imbens. The Journal of Human Resources 34 (4), 828–837.

Heckman, J., 2000. Causal parameters and policy analysis in economics: a twentieth century retrospective. Quarterly Journal of Economics 115, 45–97.

Heckman, J., Honoré, B., 1990. The empirical content of the Roy model. Econometrica 58, 1121–1149.

Heckman, J., Honoré, B., 1989. The identifiability of the competing risks model. Biometrika 76, 325–330.

Heckman, J., LaLonde, R., Smith, J., 1999. The economics and econometrics of active labor market programs. In: Ashenfelter, Card (Eds.), Handbook of Labor Economics, vol. 3A. North-Holland, New York, pp. 1865–2097.

Heckman, J., Lochner, L., Taber, C., 1998. Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. Review of Economic Dynamics.

Heckman, J., Navarro, S., 2007. Dynamic discrete choice and dynamic treatment effects. Journal of Econometrics 136, 341–396.

Heckman, J., Robb, R., 1986. Alternative methods for evaluating the impact of interventions. In: Heckman, Singer (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, New York, pp. 156–245.

Heckman, J., Singer, B., 1984a. The identifiability of the proportional hazard model. Review of Economic Studies 51, 231–241.

Heckman, J., Singer, B., 1984b. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52, 271–320.

Heckman, J., Taber, C., 1994. Econometric mixture models and more general models for unobservables in duration analysis. Statistical Methods in Medical Research 3 (3), 279–299.

Heckman, J., Taber, C., 2008. Roy model. In: Durlauf, Blume (Eds.), The New Palgrave Dictionary of Economics Second Edition. Palgrave Macmillan.

Heckman, J., Urzúa, S., 2010. Comparing IV with structural models: what simple IV can and cannot identify. Journal of Econometrics 156 (1), 27–37.

Heckman, J., Vytlacil, E., 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences 96, 4730–4734.

Heckman, J., Vytlacil, E., 2001. Local instrumental variables. In: Hsiao, C., Morimune, K., Powell, J. (Eds.), Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya. Cambridge University Press, Cambridge, p. 145.

Heckman, J., Vytlacil, E., 2005. Structural equations, treatment effects and econometric policy evaluation. Econometrica 73, 669–738.

Heckman, J., Vytlacil, E., 2007a. Econometric evaluation of social programs, Part I: causal models, structural models and econometric policy evaluation. In: Heckman, Leamer (Eds.), Handbook of Econometrics. North Holland, Amsterdam, pp. 4779–4874.

Heckman, J., Vytlacil, E., 2007b. Econometric evaluation of social programs, Part II: using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In: Heckman, Leamer (Eds.), Handbook of Econometrics. North Holland, Amsterdam, pp. 4875–5143.

Honoré, B., 1993. Identification results for duration models with multiple spells. Review of Economic Studies 60, 241–246.

Hu, Y., Shum, M., 2009. Nonparametric identification of dynamic models with unobserved state variables. Working Paper 543, Department of Economics, Johns Hopkins University.

Ichimura, H., Taber, C., 2002. Semiparametric reduced form estimation of tuition subsidies. American Economic Review 92 (2), 286–292.

Imbens, G., 2009. Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). unpublished manuscript, Harvard University.

Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62.

Imbens, G., Wooldridge, J., 2009. Recent developments in the econometrics of program evaluation. Journal of Economic Literature 47 (1), 5–86.

Kashara, H., Shimotsu, K., 2009. Nonparametric identification of finite mixture models of dynamic discrete choice. Econometrica 77 (1), 135–175.

Keane, M., Wolpin, K., 2001. The effect of parental transfers and borrowing constraints on educational attainment. International Economic Review 42, 1051–1103.

Keane, M., Todd, P., Wolpin, K., 2011. The structural estimation of behavioral models: discrete choice dynamic programming methods and applications. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, vol. 4a. Elsevier Science, pp. 331–461.

Lalonde, R., 1986. Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76, 604–620.

Leamer, E., 1983. Let's take the con out of econometrics. American Economic Review 73 (1), 31–43.

Lee, L-F, 1978. Unionism and wage rates: a simultaneous equations model with qualitative and limited dependent variables. International Economic Review 19 (2), 415–433.

Lippman, S., McCall, J., 1976. The economics of job search: a survey, Part I. Economic Inquiry 14, 155–189.

Magnac, T., Thesmar, D., 2002. Identifying dynamic discrete decision processes. Econometrica 70 (2), 801–816.

Manski, C., 1989. Anatomy of the selection problem. The Journal of Human Resources 24 (3), 343–360.

Manski, C., 1990. Nonparametric bounds on treatment effects. American Economic Review 80 (2), 319–323.

Manski, C., 1995. Identification problems in the social sciences. Harvard University Press, Cambridge Mass..

Manski, C., 1997. Monotone treatment response. Econometrica 65 (6), 1311–1334.

Manski, C., Pepper, J., 2000. Monotone instrumental variables with an application to the returns to schooling. Econometrica 68 (4), 997–1010.

Manski, C., Pepper, J., 2009. More on monotone instrumental variables. The Econometric Journal 12 (s1), s200-s216.

Matzkin, R., 1992. Nonparametric and distribution-free estimation of the threshold crossing and binary choice model. Econometrica 60, 239–270.

Matzkin, R., 1993. Nonparametric identification and estimation of polychotomous choice models. Journal of Econometrics 58, 137–168.

Matzkin, R., 2007. Nonparametric identification. In: Heckman, Leamer (Eds.), Handbook of Econometrics. North-Holland, Amsterdam, pp. 5145–5368.

Mortensen, D., Pissarides, C., 1994. Job creation and job destruction in the theory of unemployment. Review of Economic Studies 61, 397–415.

Neal, D., 1997. The effects of catholic secondary schooling on educational attainment. Journal of Labor Economics 15, 98–123.

Neal, D., Grogger, J., 2000. Further evidence on the effects of Catholic secondary schooling. Brookings-Wharton Papers on Urban Affairs 151–193.

Postel-Vinay, F., Robin, J.-M., 2002. Wage dispersion with worker and employer heterogeneity. Econometrica 70 (6), 2295-350.

Ridder, G., 1990. The non-parametric identification of generalized accelerated failure-time models. Review of Economic Studies 57, 167–182.

Rosenzweig, M., Wolpin, K., 2000. Natural 'natural experiments' in economics. Journal of Economic Literature 38 (4), 827–874.

Roy, A.D., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers (New Series) 3, 135–146.

Rust, J., 1994. Structural estimation of Markov decision processes. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. North Holland, Amsterdam, pp. 3082–3139.

Shaikh, A., 2010. Identification in Economics, Lecture Notes for Topics in Econometrics, http://home.uchicago.edu/˜amshaikh/classes/topics_winter09.html, University of Chicago.

Sims, C., 2010. Comment on Angrist and Pischke. unpublished manuscript, Princeton University.

Taber, C., 2000. Semiparametric identification and heterogeneity in dynamic programming discrete choice models. Journal of Econometrics.

Van den Berg, G., 2001. Duration models: Specification, identification and multiple durations. In: Heckman, Leamer (Eds.), Handbook of Econometrics vol. 5. Elsevier.

Vytlacil, E., 2002. Independence, monotonicity, and latent index models: an equivalence result. Econometrica 70, 331–341.

Willis, R., Rosen, S., 1979. Education and self- selection. Journal of Political Economy 87.