NEPS
National Educational Panel Study

NEPS Working Papers

Götz Rohwer

Selection, Choice and Causal Interpretations

NEPS Working Paper No. 18

Bamberg, December 2012

SPONSORED BY THE
Federal Ministry
of Education
and Research

Working Papers of the German National Educational Panel Study (NEPS)
at the University of Bamberg

Contact: German National Educational Panel Study (NEPS) − University of Bamberg − 96045 Bamberg − Germany − contact.neps@uni-bamberg.de

Selection, Choice and Causal Interpretations

Götz Rohwer, Ruhr-Universität Bochum

December 2012

E-Mail-Adresse des Autors:
goetz.rohwer@rub.de

Bibliographische Angaben:

Rohwer, G. (2012). Selection, Choice and Causal Interpretations (NEPS Working Paper No. 18). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Selection, Choice and Causal Interpretations

Abstract

Invited by the National Educational Panel Study (NEPS), during the Winter 2011/12, I gave a series of lectures about ‘Statistical methods in sociological research of education’. This text comprises an elaboration of one of these lectures dealing with different kinds of selection problems. As a starting point, I distinguish between data-generating and fact-generating processes. The term ‘data-generating process’ is used to refer to a process that generates data, that is, information about already existing facts. In contrast, when referring to processes that generate new facts (outcomes), I use the term ‘fact-generating process’. I argue that it is important whether a selection problem concerns a data-generating or a fact-generating process. In the first case, one should be concerned with possible bias. In the second case, one has to consider the selection process as an integral part of the substantive process generating the outcome of interest. In this paper, I am mainly concerned with selections that are part of fact-generating processes. In the first section, I begin with distinguishing different kinds of selection problems, and then consider situations in which a selection process creates a necessary precondition for an outcome. In the second section, I consider selection processes that result from decisions of individual or collective actors. I propose a definition of ‘choice variables’, and distinguish between explanatory and evaluative choice models. In the third section, I discuss how to think of causal effects of choice variables.

Keywords

selection, choice, causal interpretations, functional models, choice models, omitted confounder, educational processes

Invited by the National Educational Panel Study (NEPS), during the Winter 2011/12, I gave a series of lectures about 'Statistical methods in sociological research of education'. This text comprises an elaboration of one of these lectures dealing with different kinds of selection problems.

As a starting point, I distinguish between data-generating and fact-generating processes. The term 'data-generating process' is used to refer to a process that generates data, that is, information about already existing facts. In contrast, when referring to processes that generate new facts (outcomes), I use the term 'fact-generating process'. As an example think of a learning frame in which students can acquire capabilities of a specified kind, and assume that individual learning results can be captured by values of a variable, say Y. One can firstly think of a fact-generating process in which each student eventually acquires a particular capability. Afterwards, a data-generating process can take place, that is, a process in which a researcher represents students' capabilities by particular values of Y.

I argue that it is important whether a selection problem concerns a data-generating or a fact-generating process. In the first case, one should be concerned with possible bias. In the second case, one has to consider the selection process as an integral part of the substantive process generating the outcome of interest.

In this paper, I am mainly concerned with selections that are part of fact-generating processes. As a formal framework, I use functional models (see Rohwer (2010, 2012) for notations and definitions). In the first section, I begin with distinguishing different kinds of selection problems, and then consider situations in which a selection process creates a necessary precondition for an outcome. I argue that one should distinguish between 'selection problems' and problems resulting from omitted confounders. In the second section, I consider selection processes that result from decisions of individual or collective actors. I propose a definition of 'choice variables', and distinguish between explanatory and evaluative choice models. In the third section, I discuss how to think of causal effects of choice variables, and I consider problems resulting from omitted confounders.

1. Different kinds of selection

1. Distinguishing selection problems. I distinguish three situations where Y is a variable of interest, and there is a further variable S which in some sense involves a selection.

a) S is a binary variable, and values of Y can be observed if $S = 1$, and cannot be observed if $S = 0$. This can properly be called a 'sample selection problem' because S only concerns the observability of Y but is not a causally relevant condition for Y. Here it is presupposed that Y has a distribution that exists independently of S.

b) S is a binary variable, and $S = 1$ is a necessary precondition for Y to have a distribution. For example, being employed ($S = 1$) is a necessary precondition for receiving a wage (Y).

c) S can take two or more different values, and it is assumed that the distribution of Y causally depends on the value of S. For example, values of S represent school types, and Y is a measure of educational attainment.

In the first case (a), the selection concerns a data-generating process, and it is obvious that one has to think about possible bias. In cases (b) and (c), however, the selection concerns a fact-generating process, and it is not clear in which sense there might be a selection problem. In the following, I begin with a very brief consideration of (a) and then focus on (b). The discussion is continued in Section 2 where also (c) will be considered.

2. Selection in data-generating processes. Sample selection problems can be conceptualized in two ways. First, Y and S are statistical variables,[1] say

$$(Y, S) : \Omega \longrightarrow \mathcal{Y} \times \{0, 1\}$$

One knows the conditional distribution $\mathrm{P}[Y|S\!=\!1]$, but is interested in the unconditional distribution $\mathrm{P}[Y]$. As an example, one can think that S is a response indicator in a survey: $S\!=\!1$ if a sampled unit provides a value of Y, and $S\!=\!0$ otherwise.

Another framework uses random variables, say $\dot{Y}$ and $\dot{S}$. $\dot{S}$ is again a binary variable and records whether a value of $\dot{Y}$ can be observed. So it is assumed that one knows the conditional distribution $\Pr[\dot{Y}|\dot{S} = 1]$, but is interested in the unconditional distribution $\Pr[\dot{Y}]$. Of course, the interest could also concern distributions of $\dot{Y}$ (or Y in the first framework) which depend on further covariates.

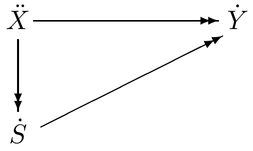
If $\dot{S}$ is independent of $\dot{Y}$, one can use $\Pr[\dot{Y}|\dot{S} = 1]$ to estimate $\Pr[\dot{Y}]$. Problems occur if, and because, $\dot{S}$ and $\dot{Y}$ are correlated. This suggests to find another variable, say $\ddot{V}$ (often consisting of several components), such that

$$\dot{S} \perp\!\!\!\perp \dot{Y} \mid \ddot{V}\!=\!v$$

is approximately true. This would allow one to use $\Pr[\dot{Y}|\ddot{V}\!=\!v, \dot{S}\!=\!1]$ to estimate $\Pr[\dot{Y}|\ddot{V}\!=\!v]$.[2]

3. Selection as a necessary precondition. I now consider the case where the selection variable represents a necessary precondition for values of the variable of interest. As an example, I take $\dot{Y}$ to represent the outcome of a university education, and $\dot{S}\!=\!1$ if a person actually begins with university studies, and $\dot{S}\!=\!0$ otherwise. It is further assumed that $\dot{Y}$ depends on a variable $\ddot{X}$ representing the level of academic performance the person has reached just before the selection variable gets a particular value (the two dots indicate that this is an exogenous variable of the model). A functional model could then be graphically depicted as follows:

Model 1.1



The picture is possibly misleading, however, because it suggests that $\dot{Y}$ has a distribution even if $\dot{S}\!=\!0$. But, obviously, a process generating a value of $\dot{Y}$ can only take place if $\dot{S} = 1$. This entails that there are two rules (functional relationships):

$$x \longrightarrow \Pr(\dot{Y}\!=\!y|\ddot{X}\!=\!x, \dot{S}\!=\!0) = 0 \tag{1}$$

and

$$x \longrightarrow \Pr(\dot{Y}\!=\!y|\ddot{X}\!=\!x, \dot{S}\!=\!1) \tag{2}$$

Nevertheless, both $\ddot{X}$ and $\dot{S}$ are causally relevant for $\dot{Y}$. Referring to expectations of $\dot{Y}$, $\dot{S}$ can be viewed as an event variable having the effect $\mathrm{E}(\dot{Y}|\ddot{X}\!=\!x, \dot{S}\!=\!1)$ (see Rohwer 2012: 22f). Effects of $\ddot{X}$ can be defined, of course, only conditional on $\dot{S}\!=\!1$:

$$\mathrm{E}(\dot{Y}|\ddot{X}\!=\!x'', \dot{S}\!=\!1) - \mathrm{E}(\dot{Y}|\ddot{X}\!=\!x', \dot{S}\!=\!1) \tag{3}$$

[1] For a formal definition of statistical variables see Rohwer (2010: 2, or 2012: 3f).

[2] As introduced in Rohwer (2012), I use a single dot to indicate that a variable has a probability distribution (given values of further variables), and I use two dots to indicate that a variable does not have an associated distribution but simply serves as a placeholder for assuming specific values.

Note that in this model there can be no interaction of $\ddot{X}$ and $\dot{S}$ w.r.t. $\dot{Y}$.

4. Counterfactual and modal questions. Although $\dot{S}=1$ is a necessary precondition for values of $\dot{Y}$, one can ask hypothetical questions. Two forms of such questions must be distinguished:

a) *Counterfactual questions* presuppose that $\dot{S}$ already has taken a particular value. For example: Given $\ddot{X}=x$ and $\dot{S}=0$, what value of $\dot{Y}$ might be expected if $\dot{S}$ had taken the value 1 instead of 0 ? (The complementary question obviously has a trivial answer.)

b) *Modal questions* presuppose a situation where $\dot{S}$ has not already taken a particular value, and a process that might generate a value of $\dot{Y}$ has not yet started.

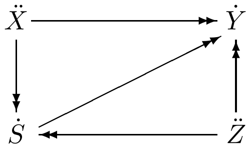
I will not discuss the counterfactual questions. To answer the modal questions, one can use the rules (1) and (2), respectively. The selection variable $\dot{S}$ then only serves to distinguish the two modal questions and does not provide any further information. (This will be further discussed in Section 2.)

5. Is there a sample selection problem? The rule (1) can be established by referring to the institutional framework without reference to sampled data. Numerically specified versions of the rule (2) must be estimated from sampled data. However, as suggested by the rule's formulation, one can simply use the data for those students who actually began a university education ($\dot{S}=1$).

Note that selection on $\dot{S}=1$ has nothing to do with a sample selection problem. The condition $\dot{S}=1$ simply determines the scope of the rule to be estimated. It is possible, of course, that the available data are selective; for example, response rates can depend on $\dot{Y}$. There is then a sample selection problem that leads to biased estimates of $\Pr(\dot{Y}=y|\ddot{X}=x,\dot{S}=1)$. But this bias is not due to conditioning on $\dot{S}=1$, instead, its definition would presuppose this condition.

6. Consideration of omitted confounders. There is, however, another problem that must be considered: the omission of possibly relevant confounding variables. This requires to refer to an enlarged model that explicitly contains a further confounding variable. I use the following

Model 1.2



where $\ddot{Z}$ is a further variable on which $\dot{S}$ depends, e.g., an indicator of a person's educational aspiration. Moreover, it is assumed that also $\dot{Y}$ depends on $\ddot{Z}$ entailing that $\ddot{Z}$ is a confounding variable.

There is no problem if data are available for both $\ddot{X}$ and $\ddot{Z}$. But now assume that data on $\ddot{Z}$ are not available so that one can only estimate a reduced version of the model; in terms of expectations:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{S}=1)=\sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{Z}=z,\dot{S}=1)\Pr(\dot{Z}=z|\ddot{X}=x,\dot{S}=1) \tag{4}$$

where $\dot{Z}$ is used instead of $\ddot{Z}$ in order to allow thinking of conditional distributions. This shows that effects of $\ddot{X}$, as defined in (3), do not have a balanced formulation in the reduced model. Of course, this is the general problem resulting from the omission of confounding variables.

Here it is important, however, that the problem does not result from conditioning on $\dot{S}$. It is true that $\dot{S}$ is a collider, entailing that conditioning on $\dot{S}$ changes the correlation between

Table 1.1 Fictitious data for the example.

x	z	s	$y=0$	$y=1$	cases
0	0	0	–	–	800
0	0	1	100	100	200
0	1	0	–	–	600
0	1	1	120	280	400
1	0	0	–	–	400
1	0	1	120	480	600
1	1	0	–	–	200
1	1	1	80	720	800

values of $\ddot{X}$ and $\ddot{Z}$ in a reference set (sample). But this is not relevant here because effect definitions are conditional on $\dot{S}=1$; in the example, they only concern students who actually begin with university studies. Problems resulting from the omission of confounding variables should therefore be clearly distinguished from 'selection problems'.

7. Illustration with a numerical example. To illustrate the argument, I use the fictitious data in Table 1.1. Based on these data, one can estimate the rule

$$(x,z) \longrightarrow \mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{Z}=z,\dot{S}=1) \tag{5}$$

and finds the values

x	z	$\mathrm{E}(\dot{Y}\mid\ddot{X}=x,\dot{Z}=z,\dot{S}=1)$
0	0	0.5
0	1	0.7
1	0	0.8
1	1	0.9

(6)

Conditioning on $\dot{S}=1$ changes the relationship between $\ddot{X}$ and $\dot{Z}$:

x	$\Pr(\dot{Z}=1\mid\ddot{X}=x,\dot{S}=0)$	$\Pr(\dot{Z}=1\mid\ddot{X}=x,\dot{S}=1)$
0	0.429	0.667
1	0.333	0.571

(7)

This is not relevant, however, because the rule (5) only applies to situations where $\dot{S}=1$.

Now assume that values of $\dot{Z}$ are not available. The data then lead to the following estimate of an effect of $\ddot{X}$:

$$\mathrm{E}(\dot{Y}|\ddot{X}=1,\dot{S}=1) - \mathrm{E}(\dot{Y}|\ddot{X}=0,\dot{S}=1) = 0.857 - 0.633 = 0.224 \tag{8}$$

This effect is not balanced. As can be seen from (7), the distribution of the omitted variable $\dot{Z}$ is different for $\ddot{X}=0$ and $\ddot{X}=1$. This means that the effect cannot be attributed solely to a difference in $\ddot{X}$. Of course, without observations on $\dot{Z}$ one cannot assess the contribution of this variable.

8. Selection and omitted confounders. In order to stress that 'selection problems' and problems resulting from unobserved confounders should be distinguished, I briefly consider a further example: How does the risk of divorce depends on the women's level of education? Researchers often found that education has a positive impact on the risk of divorce, but there also are other findings. In a recent study, Bernardi and Martinez-Pastor (2011) discuss the hypothesis that the observed relationships between education and risk of divorce might result, at least in part, from ‘selection effects’. However, it is not immediately clear how to understand this hypothesis.

To see this, one can use Model 1.2 with the following interpretation. $\dot{S}=1$ if a woman is married, and $\dot{S}=0$ otherwise. $\dot{Y}$ is an indicator variable for becoming divorced;[3] $\ddot{X}$ records the level of education, and $\ddot{Z}$ is an omitted confounder. Obviously, without being married there can be no risk of divorce. Consequently, $\dot{S}=1$ is also a necessary precondition for the variables’, $\ddot{X}$ and $\ddot{Z}$, having an impact on the risk of divorce.

Of course, one can assume that distributions of values of $\ddot{X}$ and $\ddot{Z}$ exist for married and unmarried women in some specified population.[4] So one can think that marriage changes these distributions and their correlation, and consider this as a ‘selection effect’. But this selection effect cannot change relationships between these variables and the risk of divorce; simply because these relationships are only defined conditional on $\dot{S}=1$.

However, an important part of the research question concerns the observation of historically changing relationships between women’s education and the risk of divorce. One has then to consider at least two periods ($t = 1, 2$), say

$$(\ddot{X}_1, \dot{Z}_1, \dot{S}_1, \dot{Y}_1) \;\longrightarrow\; (\ddot{X}_2, \dot{Z}_2, \dot{S}_2, \dot{Y}_2) \tag{9}$$

$\dot{Z}_t$ represents the distribution of the omitted confounder in period t. This allows one to think of the observed relationships in the following way:

$$\mathrm{E}(\dot{Y}_t|\ddot{X}_t=x, \dot{S}_t=1) = \sum\nolimits_z \mathrm{E}(\dot{Y}_t|\ddot{X}_t=x, \dot{Z}_t=z, \dot{S}_t=1)\,\Pr(\dot{Z}_t=z|\ddot{X}_t=x, \dot{S}_t=1) \tag{10}$$

So it is quite possible that differences between the observed relationships can be due to both

a) changes in the conditional expectation $\mathrm{E}(\dot{Y}_t|\ddot{X}_t=x, \dot{Z}_t=z, \dot{S}_t=1)$ which, presumably, has a causal interpretation, and
b) changes in the conditional distributions of the omitted confounder, $\Pr[\dot{Z}_t|\ddot{X}_t=x, \dot{S}_t=1]$.

Most probably, they are due to both kinds of changes so that one would like to learn about the quantitative relevance of omitted confounders. But, of course, this would require to observe the confounder.

The question remains whether one can think of (b) as a ‘selection effect’. Obviously not in the static sense referred to above. One would need to consider the function

$$(x, z) \longrightarrow \Pr(\dot{S}_t=1|\ddot{X}_t=x, \dot{Z}_t=z) \tag{11}$$

A change of this function would describe a historically changing selection into marriage. However, the changes referred to in (b) do not only result from a change in the function (11). As seen from

$$\Pr(\dot{Z}_t=z|\ddot{X}_t=x, \dot{S}_t=1) \;=\; \frac{\Pr(\dot{S}_t=1|\ddot{X}_t=x, \dot{Z}_t=z)\,\Pr(\dot{Z}_t=z|\ddot{X}_t=x)}{\Pr(\dot{S}_t=1|\ddot{X}_t=x)} \tag{12}$$

they can also result from a change in the conditional distributions of the omitted confounder.

Consequently, without observing the confounder (whose supposed existence motivates the discussion) one cannot draw any clear conclusions. On the other hand, if one could observe $\dot{Z}_t$, one could immediately consider the relationship

$$(x, z) \longrightarrow \mathrm{E}(\dot{Y}_t|\ddot{X}_t=x, \dot{Z}_t=z, \dot{S}_t=1) \tag{13}$$

[3] Bernardi and Martinez-Pastor use a duration model, but this is not important for the present conceptual discussion.

[4] To make this precise, one would need statistical variables instead of the modal variables, $\ddot{X}$ and $\ddot{Z}$.

and recognize that the risk of divorce not only depends on the women's level of education, but also on another identifiable variable, $\dot{Z}$.

2. Choice-based selection

I now consider selection variables which get their values from decisions of individual or collective actors. Such variables will be called 'choice variables'.

1. A notion of choice variables. As I will use the term, a *choice variable*, say C, has the following features:

a) The domain of C is a set of $m \geq 2$ alternatives, numerically represented by $\mathcal{C} = \{1, \ldots, m\}$.

b) Referring to a choice variable entails that there is an individual or collective agent who can choose, or already has chosen, a particular value of the variable. The agent associated with a choice variable C will be denoted by $A[C]$.

c) It is presupposed that the agent has the power to select one of the alternatives. In other words, $\mathcal{C}$ must only contain states which can be realized by the agent.

d) The agent is assumed to have considered the alternatives before one of them is actually chosen. I do not assume that the agent is 'rational' in any particular sense.

Following this definition, choice is a specific kind of selection, namely a selection which is reflexively generated by an actor. This is meant by the expression 'choice-based selection'.[5]

2. Two contexts for using choice variables. Choice variables can be used in two different contexts. In one context, one conceives of a choice variable as a kind of event variable. $C = c$ then means that the agent, $A[C]$, has chosen the alternative $c \in \mathcal{C}$. In addition, $C = 0$ means that such an event has not yet occurred.

In this understanding, choice variables can be used in explanatory models, both as explanatory variables and as dependent variables. A model that attempts to explain choices, considered as events, will be called an *explanatory choice model.* Such a model can be depicted as

$$\ddot{X} \twoheadrightarrow \dot{C}$$

where $\dot{C}$ is the choice variable, and $\ddot{X}$ denotes the explanatory variable (possibly consisting of several components). For example, one can think that the model is intended to represent the choice of a child's school type; $A[\dot{C}]$ refers to the child's parents, and $\ddot{X}$ denotes the parents' educational level. Note that this understanding entails that the parents have the power to choose a school type for their child.

In a quite different context, choice variables serve to consider possible consequences of hypothetically chosen alternatives. To hypothetically give a choice variable a value is obviously different from an actual choice, it simply means to consider that alternative. Thus, in this understanding, a choice variable is not an event variable; and in a functional model it can only be used as an exogenous variable. As an example, think again of the parents' choice of a school type for their child. One can consider the following model:

[5] So it has nothing to do with 'choice-based sampling', that is, sampling designs which use a stratification w.r.t. realized choices; see Scott and Wild (1989).

Model 2.1

$$\begin{tikzcd}
\ddot{X} \arrow[r, twoheadrightarrow] & \dot{Y}^{*} \\
\ddot{C} \arrow[ur, twoheadrightarrow] &
\end{tikzcd}$$

The model relates to a generic $A[\ddot{C}]$, the parents who are assumed to consider possible values of $\ddot{C}$ (school types). As before, $\ddot{X}$ denotes the parents' educational level. $\dot{Y}^{*}$ is used to assess possible outcomes: the child's educational success that can be expected if, given $\ddot{X}$, the parents would choose $\ddot{C}=c$. The corresponding function is

$$(x, c) \longrightarrow \mathrm{E}(\dot{Y}^{*} | \ddot{X}=x, \ddot{C}=c) \tag{14}$$

A model of this kind will be called an *evaluative choice model*. Its aim is not to predict realized choices. Instead, it is intended to serve thinking about modal questions (in the sense introduced in § 1.4).[6] Consequently, also $\dot{Y}^{*}$ cannot be understood as representing realized outcomes, and must be distinguished from an outcome variable in an explanatory model.

There obviously are similarities between an evaluative choice model and a treatment model (as this term was introduced in Rohwer 2012: 24). But also note the different tasks. A treatment model serves to formulate a generic rule about causal effects of treatments (= events which can be deliberately generated). An evaluative choice model serves an agent to consider the consequences of available alternatives. While randomization might be used for a treatment model, randomization w.r.t. the choice variable would contradict the idea of a choice.

3. Primary and secondary actors of choice models. In Rohwer (2012: 24) I introduced a distinction between primary actors (= agents of the social processes that are the topic of a model) and secondary actors (= those who construct and use a model). The distinction can easily be applied to explanatory choice models. These models are concerned with choices made by primary actors. The secondary actors, in contrast, are those who construct and use these models for one reason or another.

Now consider an evaluative choice model. Since the model serves an agent to think about available alternatives and possible consequences of choosing one of them, the agent is a secondary actor w.r.t. the model. On the other hand, the model is concerned with the agent's choice, and so the agent can also be considered as a primary actor. However, an evaluative model has not the task to predict the agent's choice. 'To choose' and 'to predict the outcome of a choice' are obviously different activities.

4. Consideration of modal questions. Let me stress the connection between modal questions (as introduced in § 1.4) and evaluative choice models. An explanatory choice model cannot be used to consider modal questions (or must be reinterpreted in some way). As an illustration, I refer to Model 2.1. This model is intended to show how expectations of an outcome variable, $\dot{Y}^{*}$, depend on exogenously given values of $\ddot{X}$ and hypothetically chosen values of $\ddot{C}$.

Of course, in order to become useful, one needs a quantification of the function (14). Data might come from a sample that relates to the following model where it is assumed that $\dot{Y}$ has the same meaning as $\dot{Y}^{*}$.

[6] Of course, an evaluative choice model only allows an agent to consider conditional expectations and cannot be used to formally derive a particular decision.

Model 2.2

$$\begin{tikzcd}
\ddot{X} \arrow[r, two heads] \arrow[d] & \dot{Y} \\
\dot{C} \arrow[ur, two heads] &
\end{tikzcd}$$

In this model, $\dot{C}$ is an event variable, representing alternatives chosen by the primary agents to which the model relates. Since $\dot{C}$ depends on $\ddot{X}$, part of the model consists in an explanatory choice model.

Only the evaluative model 2.1, not the explanatory model 2.2, can be used for modal questions. In the explanatory model, values of $\dot{C}$ come into being according to a function, $\ddot{X} \twoheadrightarrow \dot{C}$, that predicts the choices actually made. This model can therefore not be used for a situation where a choice variable can hypothetically be given different values because a choice event has not yet occurred.

Nevertheless, data corresponding to the explanatory model 2.2 can also be used to estimate the function (14). Such data would allow one to estimate $\mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{C}=c)$, and then use the rule

$$\text{Estimate } \mathrm{E}(\dot{Y}^*|\ddot{X}=x,\ddot{C}=c) \text{ by } \mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{C}=c) \qquad (15)$$

As assumed by Model 2.2, the data result from ‘self selection’ in the sense of choices, made by primary actors, which depend on a variable $\ddot{X}$; $\dot{S}$ stochastically depends on $\ddot{X}$.[7] But this does not entail a sample selection problem that might create a bias when using the rule (15). As mentioned in § 1.6, it is quite possible that Model 2.2 misses a relevant confounder and could be replaced by a better model (if the necessary data would be available). But this has nothing to do with the fact that the model contains an event variable that gets its values by choices of primary actors.

5. Modal questions w.r.t. necessary preconditions. I now consider a situation where the choice concerns a necessary precondition for a possible outcome. To illustrate, I continue with the example that was introduced in § 1.3: beginning with university studies and consideration of possible outcomes. Details depend on the set-up of the choice situation. I consider two situations.

(1) I begin with a situation where an agent has the power to decide whether a person (the agent herself or someone else) will begin with university studies. There is then a choice variable, $\ddot{C}$, having the domain $\mathcal{C} = \{1, 2\}$; and $\ddot{C}=1$ means ‘beginning’ and $\ddot{C}=2$ means ‘not beginning’ with university studies. The outcome assumed to be relevant for the choice is the success of the university education; it will be denoted by $\dot{Y}^*$. I further assume that expectations about $\dot{Y}^*$ depend on an exogenous variable, $\ddot{X}$, representing properties of the person who possibly begins university studies that are known in the choice situation.

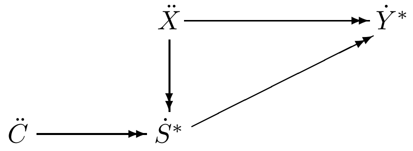
With this notation, one can again use Model 2.1 as an evaluative model and Model 2.2 as a corresponding explanatory model. The evaluative model is to be used for the two modal questions. First, what would be the outcome if $\ddot{C}=2$? Obviously, without beginning university studies there could be no successful outcome ($\dot{Y}^*$ is then either undefined or has the value zero)

Now consider the other question, What is the expected value of $\dot{Y}^*$ if $\ddot{C}=1$, given $\ddot{X}=x$? To answer this question, one needs an estimate of $\mathrm{E}(\dot{Y}^*|\ddot{X}=x,\ddot{C}=1)$. This can be derived from the explanatory model 2.2 by using the rule (15). One only needs $\mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{C}=1)$. Whether also $\mathrm{E}(\dot{Y}|\ddot{X}=x,\dot{C}=2)$ can be given a sensible interpretation is obviously irrelevant.

[7] Note that this is not the case in the evaluative model 2.1. In fact, in that model it is not even possible to imagine a dependence of $\ddot{C}$ on $\ddot{X}$.

(2) The argument in (1) presupposes that the choice variables in the two models, $\ddot{C}$ and $\dot{C}$, have the same meaning. I now consider a situation where an agent cannot immediately decide whether to begin, or not begin, with university studies, but only whether to apply for admission. The choice variable, $\ddot{C}$, has again the domain $\mathcal{C} = \{1, 2\}$, but now $\ddot{C} = 1$ means 'to apply' and $\ddot{C} = 2$ means 'not to apply'. An evaluative choice model can now be depicted as follows.

Model 2.3



$\ddot{X}$ and $\dot{Y}^*$ (and correspondingly $\dot{Y}$) have the same meaning as before. In addition, there is now the variable $\dot{S}^*$ representing the decision of the admission committee: $\dot{S}^* = 1$ if the applicant is admitted, and $\dot{S}^* = 0$ otherwise. Both $\dot{Y}^*$ and $\dot{S}^*$ are starred in order to distinguish these variables from corresponding variables in an explanatory model.

The model can be used for thinking about modal questions in two steps. In a first step, the agent, $A[\ddot{C}]$, can consider $\mathrm{E}(\dot{S}^* | \ddot{X} = x, \ddot{C} = 1)$. Given information from a corresponding explanatory model (see Model 1.1), one can use $\mathrm{E}(\dot{S} | \ddot{X} = x, \dot{C} = 1)$ and a suitably modified version of rule (15). In a second step, one can think about the final outcome, $\dot{Y}^*$, conditional on $\dot{S}^* = 1$. In this step, $\mathrm{E}(\dot{Y} | \ddot{X} = x, \dot{S} = 1)$ can be used as an estimate of $\mathrm{E}(\dot{Y}^* | \ddot{X} = x, \dot{S}^* = 1)$. Finally, both steps can be combined:

$$\text{Estimate } \mathrm{E}(\dot{Y}^* | \ddot{X} = x, \ddot{C} = 1) \text{ by } \mathrm{E}(\dot{Y} | \ddot{X} = x, \dot{S} = 1)\, \mathrm{E}(\dot{S} | \ddot{X} = x) \tag{16}$$

Notice that the evaluative model 2.3 relates to a choice situation where the committee has not yet decided about $A[\ddot{C}]$'s application (simply because $A[\ddot{C}]$ has not yet decided whether to apply). Possibly useful information can therefore only result from estimates of $\mathrm{E}(\dot{S} | \ddot{X} = x)$.

3. Causal effects of choice variables

1. Three different choice situations. How to think of causal effects of choice variables depends on the kind of choice situation. I propose to distinguish three situations where $\ddot{C}$ always is a binary choice variable with domain $\mathcal{C} = \{0, 1\}$.[8]

a) The choice concerns a necessary precondition for a process generating values of an outcome variable, say $\dot{Y}$, to take place. So the causal effect of $\ddot{C}$ can simply be stated: $\ddot{C} = 1$ is a necessary precondition for $\dot{Y}$. In order to quantify the effect, one can consider $\ddot{C}$ as an event variable and use

$$\mathrm{E}(\dot{Y} | \ddot{X} = x, \ddot{C} = 1) \tag{17}$$

where $\ddot{X}$ represents conditions on which the process generating values of $\dot{Y}$ depends.

b) Associated with the two alternatives are two qualitatively different outcome variables, say $\dot{Y}_0$ and $\dot{Y}_1$. This implies that one has to consider two qualitatively different effects, one effect of $\ddot{C} = 0$ and another one of $\ddot{C} = 1$. Both should be considered separately as indicated in (a).

c) The choice concerns two different ways to generate values of a single outcome variable, $\dot{Y}$. So one can compare the alternatives w.r.t. expectations of $\dot{Y}$, and use the effect definition

$$\Delta^s(\dot{Y}; \ddot{C}[0,1], \ddot{X} = x) := \mathrm{E}(\dot{Y} | \ddot{X} = x, \ddot{C} = 1) - \mathrm{E}(\dot{Y} | \ddot{X} = x, \ddot{C} = 0) \tag{18}$$

[8] This differs from the convention introduced in § 2.1, but eases the notation.

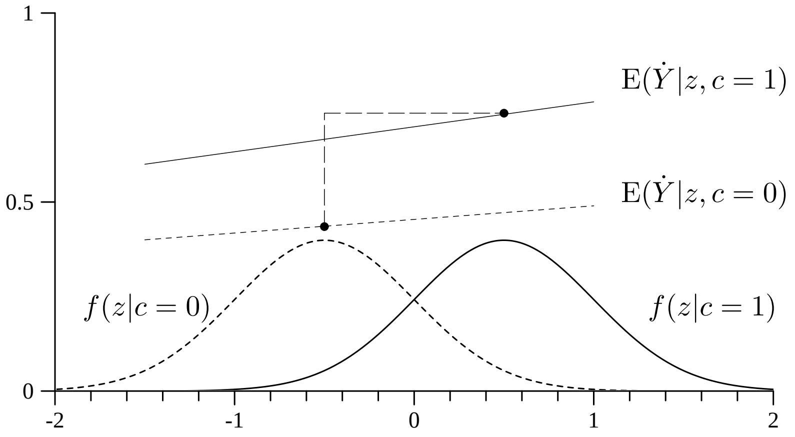
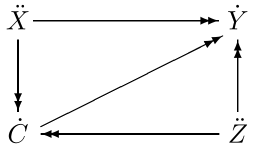


Figure 3.1 Illustration of an unbalanced effect (values of $\dot{Z}$ on the abscissa).

In the following, I only discuss the situation (c), in particular, how to think about omitted confounders. Note that this problem is of no particular relevance in the situations (a) and (b). Of course, the conditional expectation referred to in (17) might well depend on further variables. But the effect of $\ddot{C}$ can always be interpreted as an average effect w.r.t. distributions of these variables. So this is different from the situation discussed in Section 1 where one is interested in effects of $\ddot{X}$ (e.g. education of married women).

2. The problem: omitted confounders. To focus the discussion, I consider

Model 3.1



The structure is the same as in Model 1.2, but it is now assumed that $\dot{Y}$ has a distribution for both values of $\dot{C}$. As an example, one can think that the choice is between two learning frames, σ_0 and σ_1, where a person can acquire competencies represented by $\dot{Y}$, and it is assumed that the processes generating values of $\dot{Y}$ also depend on two further variables, $\ddot{X}$ and $\ddot{Z}$, having values already fixed in the choice situation.

If data for both variables are available, one can use $\ddot{Z}=z$ as an additional condition in the effect definition (18). But if data on $\ddot{Z}$ are not available, this variable is an omitted confounder, and the effect definition (18) relates to a reduced model. Substituting $\ddot{Z}$ by a variable $\dot{Z}$ having a distribution, the observable effect is then given by

$$\begin{aligned}\Delta^s(\dot{Y};\dot{C}[0,1],\ddot{X}=x) = \textstyle\sum_z \big[& \mathrm{E}(\dot{Y}|\dot{C}=1,\ddot{X}=x,\dot{Z}=z)\,\Pr(\dot{Z}=z|\dot{C}=1,\ddot{X}=x)- \\ & \mathrm{E}(\dot{Y}|\dot{C}=0,\ddot{X}=x,\dot{Z}=z)\,\Pr(\dot{Z}=z|\dot{C}=0,\ddot{X}=x)\big]\end{aligned} \tag{19}$$

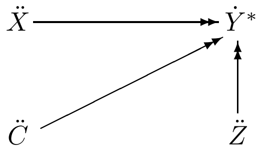
This effect is no longer balanced because

$$\Pr[\dot{Z}|\dot{C}=1,\ddot{X}=x] \neq \Pr[\dot{Z}|\dot{C}=0,\ddot{X}=x]$$

and it is therefore unclear how to think of a causal effect of $\dot{C}$; see the illustration in Figure 3.1. This difficulty motivates an interest in balanced effect formulations.

3. The perspective of evaluative choice models. A further motive for an interest in balanced effect formulations comes from evaluative choice models where an agent is interested in potential effects of the available alternatives. Consider the evaluative choice model that corresponds to Model 3.1:

Model 3.2



The agent, $A[\ddot{C}]$, knows the value of $\ddot{X}$, say x^*. Assuming that only the rule (19) is available, expectations are evaluated by comparing

$$\mathrm{E}(\dot{Y}^*|\ddot{C}=0,\ddot{X}=x^*) \quad \text{and} \quad \mathrm{E}(\dot{Y}^*|\ddot{C}=1,\ddot{X}=x^*)$$

As shown by (19), both are averages w.r.t. different distributions of $\dot{Z}$. However, the agent can assume that there is a particular value of this variable, say z^*, that is identical for both alternatives in the given choice situation. Even if this value is not known, it would probably preferable to use a balanced effect formulation.[9]

4. Constructions of balanced effects. So one should ask, Can balanced effect formulations be constructed? There are three approaches.

a) Randomization w.r.t. all possible confounders. As I have argued in Rohwer (2012), even if possible, this approach is most often problematic in social research because it would change the processes that one aims to investigate.

b) One uses a fixed distribution of the confounding variables. For example, instead of (19) one can consider

$$\sum\nolimits_z \big[\mathrm{E}(\dot{Y}|\dot{C}=1,\dot{X}=x,\dot{Z}=z)\,\Pr(\dot{Z}=z) \; - \mathrm{E}(\dot{Y}|\dot{C}=0,\dot{X}=x,\dot{Z}=z)\,\Pr(\dot{Z}=z) \qquad (20)$$

where $\Pr[\dot{Z}]$ is an arbitrarily defined distribution (e.g., the distribution of a corresponding statistical variable in a sample). Of course, this method can only be used with observed confounders; one does not get effects that are balanced w.r.t. omitted confounders.

c) One uses specific kinds of parametric models which, given sufficient assumptions, allow one to construct balanced effects. This will be further discussed in the next paragraph.

5. Parametric assumptions about confounders. I use the essential ideas of Heckman's probit selection model (Heckman 1979) to illustrate the parametric approach to the construction of balanced effects.[10] The basic idea is to assume that omitted confounders can be implicitly taken into account by a joint parametric distribution for the variables $\dot{Y}$ and $\dot{C}$.

To explain the argument, I start form Model 3.1. If $\ddot{Z}$ is observed, one can begin with a linear model

$$\dot{Y} = g_x(x) + g_z(z) + \dot{C}\gamma + \epsilon' \qquad (21)$$

[9] In this respect, evaluative choice models are similar to treatment models which also motivate an interest in balanced effect formulations.

[10] Here I consider this model as a proposal for coping with omitted confounders ('endogeneity bias'). As proposed by Heckman, the model is primarily used for 'sample selection problems' in the sense defined in § 1.1, that is, in situations where observations are only available if $\dot{C}=1$.

where g_x and g_z are deterministic functions of values of $\ddot{X}$ and $\ddot{Z}$, respectively, and ϵ' is a residual random variable. Assuming that the distribution of ϵ' is independent of $\dot{C}$, one can interpret γ as a balanced effect. If $\ddot{Z}$ is not observed, one can consider the reduced model that is based on assuming, instead of $\ddot{Z}$, a variable $\dot{Z}$ with some unknown but exogenously given distribution. Instead of (21), one gets the reduced model

$$\dot{Y} = g_x(x) + \dot{C}\gamma + \epsilon \tag{22}$$

where $\epsilon := g_z(\dot{Z}) + \epsilon'$. Since $\dot{C}$ depends on $\dot{Z}$, the distribution of ϵ depends on $\dot{C}$, and the effect of $\dot{C}$ is not balanced w.r.t. ϵ. But from (22) one can derive

$$\mathrm{E}(\dot{Y}|x,c) = g_x(x) + c\gamma + \mathrm{E}(\epsilon|x,c)$$

This shows that, in order to estimate γ, one would like to know values of $\mathrm{E}(\epsilon|x,c)$. If observed, they could be used as values of a further variable in a regression model for $\dot{Y}$. Of course, these values cannot be observed; but given enough assumptions, they can be constructed.

There are two steps. In the first step one assumes a model for $\dot{C}$. This is done by employing a latent variable, η', as follows:

$$\dot{C} = I[\eta' > -h_x(x) - h_z(z)] \tag{23}$$

where h_x and h_z are deterministic functions of values of $\ddot{X}$ and $\ddot{Z}$, respectively ($I[\ldots]$ denotes the indicator function). Again, one can define $\eta := h_z(z)+\eta'$, and rewrite (23) as $\dot{C} = I[\eta > -h_x(x)]$.

In the second step, the distributional assumptions come into play. These concern ϵ', η', $\dot{Z}$, and (ϵ, η):

a) $\epsilon' \sim \mathcal{N}(0, \sigma^2_{\epsilon'})$

b) $\eta' \sim \mathcal{N}(0, 1)$

c) $g_z(\dot{Z}) \sim \mathcal{N}(\mu_{g_z}, \sigma^2_{g_z})$

d) $h_z(\dot{Z}) \sim \mathcal{N}(\mu_{h_z}, \sigma^2_{h_z})$

e) ϵ' and $g_z(\dot{Z})$ are independent.

f) η' and $h_z(\dot{Z})$ are independent.

g) The joint distribution of ϵ and η is bivariate normal with a correlation ρ.

These assumptions entail: $\epsilon \sim \mathcal{N}(\mu_{g_z}, \sigma^2_\epsilon)$ with $\sigma^2_\epsilon = \sigma^2_{g_z} + \sigma^2_{\epsilon'}$, and $\eta \sim \mathcal{N}(\mu_{h_z}, \sigma^2_\eta)$ with $\sigma^2_\eta = \sigma^2_{h_z} + 1$.

From these assumptions and their implications, one can finally derive expressions for the conditional expectations $\mathrm{E}(\epsilon|x,c)$:

$$\mathrm{E}(\epsilon|x, c{=}1) = \mathrm{E}(\epsilon|x, \eta > -h_x(x)) = \mu_{g_z} + \rho\,\sigma_\epsilon\,\frac{\phi(h^*_x(x))}{\Phi(h^*_x(x))}$$

and correspondingly

$$\mathrm{E}(\epsilon|x, c{=}0) = \mathrm{E}(\epsilon|x, \eta \leq -h_x(x)) = \mu_{g_z} - \rho\,\sigma_\epsilon\,\frac{\phi(h^*_x(x))}{\Phi(-h^*_x(x))}$$

where ϕ and Φ denote, respectively, the density and distribution function of the standard normal distribution, and $h^*_x(x) := (h_x(x) + \mu_{h_z})/\sigma_\eta$. Both can be combined as

$$\mathrm{E}(\epsilon|x,c) = \mu_{g_z} + \rho\,\sigma_\epsilon\,\lambda(x,c) \tag{24}$$

with

$$\lambda(x,c) := c\,\frac{\phi(h_x^*(x))}{\Phi(h_x^*(x))} - (1-c)\,\frac{\phi(h_x^*(x))}{\Phi(-h_x^*(x))}$$

It can now be seen how this is an approach to the construction of balanced effects. By defining a new residual variable,

$$\epsilon^* := \epsilon - \rho\,\sigma_\epsilon\,\lambda(x,c) - \mu_{g_z}$$

one can rewrite (22) as

$$\dot{Y} = g(x) + \dot{C}\gamma + \rho\,\sigma_\epsilon\,\lambda(x,c) + \mu_{g_z} + \epsilon^* \tag{25}$$

containing a further regressor, $\lambda(x,c)$, and the new residual, ϵ^*. Now one can derive

$$\mathrm{E}(\epsilon^*|x,c) = 0 \quad \text{and} \quad \mathrm{Cov}(\epsilon^*,\dot{C}|x) = 0$$

showing that the effect of $\dot{C}$ is balanced w.r.t. ϵ^*.

However, it is obvious that this result relies on very particular assumptions about the distributions of omitted variables which are difficult to justify in applications.[11] The most important assumption concerns the distribution of the omitted confounder, $\dot{Z}$. If it is not normally distributed, e.g. if it is a binary variable, the argument will not work. Moreover, in order to actually construct the conditional expectations (24), one would need assumptions about the mean and the variance of the unobserved confounder.

References

Bernardi, F., Martinez-Pastor, J.-I. (2011). Female Education and Marriage Dissolution: Is it a Selection Effect? *European Sociological Review* 27, 693–707.

Briggs, D. C. (2004). Causal Inference and the Heckman Model. *Journal of Educational and Behavioral Statistics* 29, 397–420.

Bushway, S., Johnson, B. D., Slocum, L. A. (2007). Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology. *Journal of Quantitative Criminology* 23, 151–178.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica* 47, 153–161.

Little, R. J. A. (1985). A Note About Models for Selectivity Bias. *Econometrica* 53, 1469–1474.

Rohwer, G. (2010). *Models in Statistical Social Research.* London: Routledge.

Rohwer, G. (2012). Functional Models and Causal Interpretations. *NEPS Working Paper No. 9*. Bamberg: Otto-Friedrich-Universität. Nationales Bildungspanel.

Scott, A. J., Wild, C. J. (1989). Selection Based on the Response Variable in Logistic Regression. In: C. J. Skinner, D. Holt, T. M. F. Smith (eds.), *Analysis of Complex Surveys*, 191–205. New York: Wiley.

[11] For critical discussion see Little (1985), Briggs (2004), Bushway, Johnson and Slocum (2007).