NEPS
National Educational Panel Study

NEPS Working Papers

Götz Rohwer

Functional Models and Causal Interpretations

NEPS Working Paper No. 9

Bamberg, May 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

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).
The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS consortium.

The NEPS Working Papers are available at
http://www.uni-bamberg.de/neps/publikationen/neps-working-papers/

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

Functional Models and Causal Interpretations

Götz Rohwer, Ruhr-Universität Bochum

May 2012

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

Bibliographische Angaben:

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

Functional Models and Causal Interpretations

Abstract

Invited by the National Educational Panel Study (NEPS), during the Winter 2011/12, the author gave a series of lectures about 'Statistical methods in sociological research of education'. This text comprises an elaboration of three of these lectures discussing the understanding of statistical models and their relationships with substantive research questions. A basic idea is to distinguish between two goals of statistical methods: Description of statistical facts concerning frequency distributions defined for a sample or population, and finding rules for relationships between variables which can be used for predictions and explanations. Correspondingly, there are two kinds of statistical generalization: Descriptive generalizations (from sample to population), and modal generalizations (from data to rules).

The paper is then mainly concerned with functional models understood as tools for modal generalizations. Such models can be used both for predictions and for explanations. In social research, they are mainly used for explanations. Important questions then concern relationships between explanatory variables which can be of different kinds. In particular, one has to distinguish between interactions and functional relationships. Further questions concern causal interpretations. It is argued that causal relationships cannot be defined in terms of functional models (variables and their functional relations), but require an interpretation in terms of the substantive application. How to understand such interpretations also depends on whether the model is intended to serve explanations or to assess effects of treatments. This distinction is particularly important for social research dealing with processes involving activities of human agents. The paper finally discusses, and contrasts, two quite different understandings of notions of 'potential outcomes'.

Keywords

descriptive generalization, modal generalization, functional model, causal interpretation

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 three of these lectures discussing the understanding of statistical models and their relationships with substantive research questions.

As a starting point, I distinguish two goals of statistical methods: (a) Description of statistical facts concerning frequency distributions defined for a sample or population. (b) Finding rules for relationships between variables which can be used for predictions and explanations. Correspondingly, there are two kinds of statistical generalization: Descriptive generalizations (from sample to population), and modal generalizations (from data to rules). This will be discussed in the first chapter where I introduce a distinction between descriptive models (= tools for descriptive generalizations) and analytical models (= tools for modal generalizations). This chapter also proposes a general definition of functional models.

The second chapter starts from the observation that analytical models most often require the inclusion of several explanatory variables. It is then important to explicitly consider relationships between these variables. I distinguish between interaction and functional relationships, and discuss both kinds at some length.

The third chapter is concerned with using functional models for causal considerations. I argue that causal relationships cannot be defined in terms of such models (variables and their functional relations), but require an interpretation in terms of the substantive application. They nevertheless can provide useful formal notions of (causally interpretable) effects. I then introduce a distinction between explanatory and treatment models which is particularly important for social research dealing with processes involving activities of human agents. Finally, I discuss the 'potential outcomes approach' and distinguish between a descriptive and a rule-based version of this approach. I show that only a rule-based version is compatible with functional models.

1. Descriptive and analytical models

I distinguish two goals of statistical methods: Description of statistical distributions in a specified set of units (a sample or population), and finding rules, most often about relationships between variables, which can be used for predictions and explanations. There correspondingly are two ways of using data:

- descriptive generalization: from descriptive statements about a sample to descriptive statements about a corresponding population, and
- modal generalization: from statistical data to the formulation of more general relationships between variables.

This distinction will be discussed in the first section. I then go on to distinguish two kinds of statistical models. In the second section I very briefly consider descriptive models which aim to describe distributions of statistical variables defined for a sample or population. In the third section I consider analytical models serving to formulate theoretical hypotheses about relationships between variables.

1.1 Two kinds of generalization

1. Descriptive statistical statements. I use the following definition: Descriptive statistical statements are statements about the frequency distribution of properties (or quantities derived thereof) in a specified set of units. As a formal framework, I use statistical variables (Rohwer 2010). The symbolic notation is $X : \Omega \longrightarrow \mathcal{X}$. X is the name of the variable, Ω is the

reference set, a finite set of actually observed or assumed cases, and $\mathcal{X}$ is the property space (domain). If the reference set consists of not actually observed cases, it is nevertheless required, for descriptive statements, that one can reasonably assume that the cases do exist, or have existed in the past. For example, 'all children who attended Kindergarten in Germany in 2010', but not: 'all children who (possibly) attend Kindergarten in Germany in 2020'.

2. Defining descriptive generalization. Statistical variables provide a useful framework for the definition of descriptive generalization. Starting point is a statistical variable, $X : \Omega \longrightarrow \mathcal{X}$, representing the observations. Ω, the set of observed cases, is then considered as subset of another set, say Ω^*, for which one can assume an analogously defined statistical variable: $X^* : \Omega^* \longrightarrow \mathcal{X}$, having the same property space as X.

This framework allows one to define: A *descriptive generalization* consists in using the observed values of X for making descriptive statements about the distribution of X^* in Ω^*. It is noteworthy that the desired generalization has the same linguistic form as the statistical statements derived from the observations; there only is a change in the reference set.

I will not discuss here problems of statistical inference. It is obvious, however, that the justification of a descriptive generalization must be based on the data generating process that has generated the observations. Note that I use the following distinction:

- 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.

3. Limitations of descriptive generalizations. Descriptive generalization intends to enlarge the knowledge about statistical facts, meaning here statistical distributions as they are actually realized in specified populations. This very interest requires a narrow understanding of 'population'.

Limitations become obvious when the justification of descriptive generalization is based on probability sampling. This requires that Ω can be viewed as a probability sample from Ω^*. Consequently, Ω^* can only consist of units having a positive selection probability *when and where* the sample is drawn.

Particular difficulties arise when the interest concerns historical processes. The basic question then is, How to define a population of processes? From a methodological point of view, such populations are best defined as cohorts. However, being interested in descriptive generalizations, this requires to adopt a historical perspective that is confined to mostly completed processes.

A special problem occurs in the NEPS which is based on different samples from different populations: Would it be possible to combine the data in order to get a picture of overarching educational processes? This will not be possible in the form of just one descriptive generalization (since no combined sampling design is available). It might be possible, however, to use the separate samples for modal generalizations (predictive rules) which, taken together, would allow making comprehensive statements about educational processes.

4. Modal generalizations with rules. I now consider a different kind of generalization where the goal is, not a descriptive statement about a set of units, but a predictive rule. I use the term

'rule' in a general sense for statements having the form

If ..., then ...

Different kinds of rules can be distinguished w.r.t. the modalities used in formulating the *then*-part; for example: If ..., then ... is possible, or probable, or necessary, or normatively required.

Empirical research is primarily interested in *predictive rules*. Example: Let ω denote an individual who has finished school in Germany: If at least one of ω's parents has finished school with an Abitur, then it is highly probable that also ω has an Abitur. Note that this is a *generic* rule, meaning that its object is specified only by values of variables.

An important distinction can be made between static and dynamic predictive rules.

- A *static predictive rule* formulates a relationship between properties of a unit. The general form is: If ω has property x, then ω (probably) has property y. This kind of predictive rule is exemplified by the above example.
- A *dynamic predictive rule* relates to a fact-generating process that generates an outcome that is to be predicted.
  Example: If ω (a generically specified individual) regularly participates in the instructions, she will (probably) be successful in the final exam.

5. Formulating predictive rules with variables. In the following, I only consider predictive rules which include a probabilistic qualification of the prediction. When formulating such rules with variables, a first question concerns how to understand the probabilistic qualification. There are two forms:

a) Qualitative: If $X{=}x$, then $Y{=}y$ is probable (in some qualified sense).

b) Quantitative: If $X = x$, then $\Pr(Y = y) = \ldots$ [a specific, actually given or assumed, numerical value].

Empirical research with statistical methods regularly uses quantitative formulations. (Since there is a formal equivalence of frequency and probability functions, researchers often ignore the conceptual distinction and present their observed frequencies in terms of probabilities.)

The presupposition of quantifiable probabilities allows one to use mathematical functions for formulating the relationship between the *if*- and the *then*-part of the rule. As a general form one can use

$$x \longrightarrow \Pr[Y \mid X{=}x]$$

to be read as a function that assigns to each value x in the domain of X a conditional probability distribution of Y. If Y is a discrete variable, one can also use specific functions having the form

$$x \longrightarrow \Pr(Y{=}y \mid X{=}x)$$

for each value y in the domain of Y. Another often used special form is

$$x \longrightarrow \mathrm{E}(Y \mid X{=}x)$$

which formulates the relationship with conditional expectations of Y.

Starting from such general formulations, one can think of more specific parametric forms. However, whatever the finally chosen functional form, these forms must be distinguished from numerically specified functions which actually allow one to calculate values of the function.

6. Predictive rules vs. descriptive statements. Predictive rules must be distinguished from descriptive statistical statements.

- While descriptive statistical statements concern a reference set of particular units, a predictive rule concerns a generic unit which is only specified by values of variables.
- Correspondingly, there is a conceptual difference between frequencies, $\mathrm{P}(Y{=}y \,|\, X{=}x)$, which presuppose a finite reference set, and probabilities, $\Pr(Y{=}y \,|\, X{=}x)$, which concern a generic unit. I therefore use different symbols: P for frequencies, and Pr for probabilities.

A random generator can serve to illustrate the distinction. I use 'throwing a die' as an example. The random generator can be defined by a rule, e.g., 'If the die is thrown, there are six possible outcomes, each can occur with the same probability (1/6).' This rule is to be distinguished from a descriptive statement about frequencies of outcomes in an actually realized set of throws.

Assume the die is thrown 100 times. Results can be represented by a statistical variable $Z : \Omega \longrightarrow \mathcal{Z} := \{1, \ldots, 6\}$. $\mathrm{P}[Z]$, the distribution of Z, must be distinguished (numerically and conceptually) from the probability distribution which is used in the formulation of the rule describing the random generator.

Since predictive rules are different from descriptive statements (and different from analytical truths), they cannot be true or false. They can only be pragmatically justified, that is, with arguments showing that, and how, a rule can help people in their activities.

7. Statistical and modal variables. The conceptual distinction between descriptive statements and predictive rules suggests to make a corresponding distinction between the kinds of variables involved. As already explained, descriptive statistical statements are derived from statistical variables which are known, or assumed, to represent realized properties of existing units. This is also true for conditional frequencies: $\mathrm{P}(Y{=}y \,|\, X{=}x)$ is derived from a statistical variable, (X, Y), which is defined for a particular reference set.

When considering instead conditional probabilities, $\Pr(Y{=}y \,|\, X{=}x)$, one must recognize that there are two different conceptual frameworks:

- One can assume that the conditional probabilities are derived from a random variable (X, Y). This understanding presupposes the existence of joint and marginal probability distributions of the two variables.
- The situation is different when conditional probabilities serve to formulate probabilistic predictive rules. In this framework, X is used to formulate a hypothetical assumption, and so it is neither a random variable (having a probability distribution) nor a statistical variable (having a statistical distribution). Consequently, also Y has no unconditional distribution, but can only be viewed as a random variable for specified values of X.

In order to remind of the second context, I speak of *modal variables* and use a special notation: $\ddot{X}$ instead of X, and $\dot{Y}$ instead of Y. The symbolic notation for a probabilistic predictive rule then becomes $x \longrightarrow \Pr[\dot{Y} \,|\, \ddot{X}{=}x]$.

8. Applications of predictive rules. Why do we need predictive rules? Part of the answer is obvious: they are required for making observations relevant for predicting possible (future) outcomes. However, can predictive rules also be used for explanations?

An influential tradition has sought to use probabilistic rules for explanations of individual outcomes (often called 'inductive-statistical explanations'). This proposal was followed by a long-standing critical discussion. One of the main points of criticism is easily understandable: That A makes B to some degree probable does not show why B occurred.

The discussion mainly concerns differences between 'explanation' and 'prediction' and can therefore be ignored if one is only interested in predictions. On the other hand, being interested in explanations, one should begin with rethinking the questions that should be answered by an

explanation and, in particular, distinguish 'why' and 'how' questions (see, e.g., Cross (1991) and Faye (1999)).

I will not take up this discussion which primarily concerns the explanation of particular outcomes in individual cases. Instead, I briefly consider statistical explanations which are concerned with statistically defined explananda (= statistical distributions or quantities derived from such distributions).

9. Statistical explanations. Statistical explanations, as I use this term here, are concerned with the explanation of statistical distributions. Let $\mathrm{P}[Y]$, the distribution of a statistical variable $Y : \Omega \longrightarrow \mathcal{Y}$, denote the explanandum. In my understanding, a statistical explanation uses two premisses:

- a statistical distribution, $\mathrm{P}[X]$, X being defined for the same reference set Ω, and
- a probabilistic rule: $x \longrightarrow \Pr[\dot{Y}|\ddot{X}{=}x]$, where $\ddot{X}$ and $\dot{Y}$ correspond, respectively, to X and Y.

The formal part of the explanation then consists in using

$$\Pr(\dot{Y}{=}y) := \sum\nolimits_x \Pr(\dot{Y}{=}y \,|\, \ddot{X}{=}x)\, \mathrm{P}(X{=}x) \tag{1}$$

to derive a probability distribution $\Pr[\dot{Y}]$. (This formulation presupposes that all variables are discrete.)

The predictive claim is that $\Pr[\dot{Y}]$ is approximately equal to $\mathrm{P}[Y]$. This claim is trivially valid if the predictive rule is derived from the joint distribution of X and Y. The idea that the predictive rule is a generalization must therefore be taken seriously, and should be explicitly considered.

However, one also must reflect explanatory claims which are entailed neither by the formal framework nor by any particular degree of predictive success. I propose that the following considerations are important.

a) Whether values of explanatory variables can be understood as conditions for processes generating values of the explanandum variable.

b) Whether there are relationships between explanatory variables, and one can distinguish mediating and exogenous explanatory variables.

c) Whether there are potentially important explanatory variables not explicitly considered, and what might follow from their omission.

d) Whether, and to which degree, the rule used in the explanation depends on the particular distribution of the explanatory variables ('distribution-dependent causation').

e) Whether, and to which degree, the rule used in the explanation is historically stable.

1.2 Descriptive models

1. Descriptive models based on statistical Variables. Descriptive models, as understood in this text, are tools for describing distributions of statistical variables. Starting point is a statistical variable, $X : \Omega \longrightarrow \mathcal{X}$, often consisting of several components. A descriptive model aims to describe the distribution of X, denoted by $\mathrm{P}[X]$, or aspects of this distribution, by using a simpler mathematical form.

As an example, one can think of describing the distribution of students' 'ability scores' by a normal distribution (Jackson et al. 2007).

2. Regression models with statistical variables. If X consists of two or more components, one is often interested in descriptions of conditional distributions. This is done with regression functions and regression models. The starting point is given by a two-dimensional statistical variable, say $(X, Y) : \Omega \longrightarrow \mathcal{X} \times \mathcal{Y}$. A general *regression function* is a function

$$x \longrightarrow \mathrm{P}[Y \,|\, X{=}x]$$

which assigns to each value $x \in \mathcal{X}$ the conditional frequency distribution of Y, as given by the statistical variable (data). In order to create a descriptive regression model, one uses a simpler mathematical representation of the conditional distribution, say $g(x;\theta) \approx \mathrm{P}[Y \,|\, X{=}x]$, where θ is a parameter vector. A general *regression model* is then given by the function $x \longrightarrow g(x;\theta)$.

Special regression models are used to represent aspects of $\mathrm{P}[Y \,|\, X{=}x]$. Of widespread use is regression with conditional mean values: $m(x;\theta) \approx \mathrm{M}(Y \,|\, X{=}x)$. As an example, one can think of a linear model,

$$\mathrm{M}(Y \,|\, X{=}x) \approx \alpha + x\beta$$

This model approximates the conditional mean value of Y by a linear function of the values of X that are used as conditions.

3. Descriptive models and descriptive generalizations. Descriptive models are primarily tools for comprehending aspects of complex data sets. As suggested by a famous statistician, R. A. Fisher (1922: 311), this is a primary task of statistical methods:

> Briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.

Descriptive models are also useful tools for descriptive generalizations. Since these models relate to statistical variables, there is no conceptual difference whether the reference set is a sample or a population. Starting from a model intended to describe a population allows one to think in terms of estimating its parameters with the information from a sample.

Notice that the term 'estimation' has a clear meaning in this context: It means that one aims to find values of model parameters which are defined by their hypothetical calculation for the complete population. This entails that already their definition depends not only on the specified model, but also on a particular method to calculate its parameters.

1.3 Analytical models

1. Relationships between variables. Analytical models, as understood in this text, are tools for thinking about relationships between variables. The basic formal tool are functions (mathematically understood) which connect variables. So one can speak of 'functional relationships' between variables, and the models are also called 'functional models' (Rohwer 2010).

Two kinds of such functional relationships must be distinguished. Consider two variables, X with domain $\mathcal{X}$ and Y with domain $\mathcal{Y}$.

- A *deterministic* functional relationship consists of a function

  $$x \longrightarrow y = f(x)$$

  which assigns to each value $x \in \mathcal{X}$ exactly one value $f(x) \in \mathcal{Y}$.

- A *probabilistic* functional relationship consists of a function

  $$x \longrightarrow \Pr[Y|X{=}x]$$

  which assigns to each value $x \in \mathcal{X}$ a conditional probability distribution.

Notice that $\Pr[Y|X=x]$ is itself a function. If Y is a discrete variable, this function can be written as

$$y \longrightarrow \Pr(Y{=}y \,|\, X{=}x) \tag{2}$$

to be interpreted as the probability of $Y{=}y$ given that $X{=}x$.

Since (2) is formally identical with a probabilistic predictive rule as introduced in **1.1.6**, functional models can also be understood as tools for formulating probabilistic predictive rules.

2. A general notion of functional models. A general definition of functional models can be given as follows:

a) The structure of the model is given by a directed acyclic graph.

b) To each node of the graph corresponds a variable. Variables with indegree zero are called exogenous variables and marked by two dots. All other variables are called endogenous variables and marked by a single dot. (The notation corresponds to the convention introduced in **1.1.8**.)

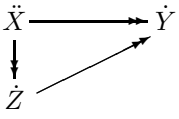
c) For each endogenous variable, there is a deterministic or probabilistic function showing how the variable (its values or probability distribution) depends on values of the immediately preceding variables.

d) Without further assumptions, exogenous variables do not have an associated distribution.

This is a formal framework. The arrows between variables have no specified meaning. In many applications they can be understood as indicating some kind of dependence relation.

3. Illustration with a simple example. To illustrate the notion of functional model, I begin with a simple example which concerns the educational outcome of a generic child.

Model 1.3.1



For simplicity, the variables are assumed to be binary and defined as follows:

$\dot{Y}$ child's educational outcome (1 successful, 0 otherwise)
$\ddot{X}$ parents' educational level (1 high, 0 low)
$\dot{Z}$ school type (1 or 2)

The model contains two probabilistic functions:

$$x \longrightarrow \Pr[\dot{Z} \,|\, \ddot{X}{=}x]$$
$$(x, z) \longrightarrow \Pr[\dot{Y} \,|\, \ddot{X}{=}x, \dot{Z}{=}z]$$

Since $\dot{Z}$ and $\dot{Y}$ are binary variables, it suffices to consider the functions

$$x \longrightarrow \Pr(\dot{Z}=1 \,|\, \ddot{X}=x)$$
$$(x,z) \longrightarrow \Pr(\dot{Y}=1 \,|\, \ddot{X}=x, \dot{Z}=z)$$

The first function is intended to show how the probability of attending a specified school type depends on the parents' educational level. The second function is intended to show how the probability of educational success depends both on the parents' educational level and on the school type.

4. Assuming distributions for exogenous variables? Exogenous variables of a functional model do not have an associated distribution. However, there sometimes are reasons for assuming distributions for exogenous variables.

a) Using the model for a statistical explanation (as defined in **1.1.9**) w.r.t. a particular reference set. Distributions of the model's exogenous variables are then identified with the distributions of the corresponding statistical variables.

b) Using the model for predicting the outcome for an individual that is (only) known to belong to a particular reference set. One then employs a reduced model that is derived from the original model by integrating over the distributions of the unobserved exogenous variables. Assume, for example, that one wants to use the model of the previous paragraph for predicting a child's educational success. Not knowing the educational level of the parents, one cannot use the model. However, substituting $\ddot{X}$ by a variable $\dot{X}$ with a probability distribution $\Pr[\dot{X}|\ddot{Z}=z]$, one can derive a *reduced model*

$$\Pr(\dot{Y}=1 \,|\, \ddot{Z}=z) = \sum\nolimits_x \Pr(\dot{Y}=1 \,|\, \ddot{Z}=z, \dot{X}=x) \Pr(\dot{X}=x|\ddot{Z}=z)$$

which only requires knowledge of the child's school type.

c) Using the model for predicting the value of an exogenous variable based on knowing values of endogenous variables. In order to apply Bayesian inference, one must begin with a prior distribution for the exogenous variables.

5. Defining effects of explanatory variables. Assume that $\dot{Y}$ depends on an exogenous variable $\ddot{X}$. To think of an effect of $\ddot{X}$ means to compare

$$\Pr[\dot{Y}|\ddot{X}=x'] \quad \text{and} \quad \Pr[\dot{Y}|\ddot{X}=x'']$$

for (at least) two values, x' and x'', of $\ddot{X}$. This comparison concerns conditional distributions and cannot, in general, be summarized by a single number.

One therefore often uses a simplified definition which only compares expected values:

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

However, one has to take into account that $\dot{Y}$ also depends on further variables. Then, in general, effects cannot simply be attributed to a change in $\ddot{X}$, but are *context-dependent*. Formally, assume that $\dot{Y}$ also depends on $\ddot{Z}$. The effect of a change in $\ddot{X}$ must then be written as

$$\mathrm{E}(\dot{Y}|\ddot{X}=x'', \ddot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x', \ddot{Z}=z) \tag{4}$$

and, in general, depends on the *covariate context* specified by $\ddot{Z}=z$.

6. Explained variance and statistical explanation. Authors who estimate regression functions often report some measure of 'explained variance'. It is noteworthy that this notion cannot

immediately be applied to functional models. Assume a simple functional model: $\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{Y}$. Since there is no distribution for $\ddot{X}$, also $\dot{Y}$ has no distribution, and the idea of 'explained variation' cannot be applied.

It would be possible to refer to the variance of $\dot{Y}$ conditional on values of $\ddot{X}$: $\mathrm{V}(\dot{Y}|\ddot{X}{=}x)$. This could be used for quantifying the uncertainty of predictions; but this is a different idea.

However, the notion of 'explained variance' can sensibly be used when functional models are considered as tools for statistical explanations. As proposed in **1.1.9**, this application starts from a statistical variable, say $(X, Y) : \Omega \longrightarrow \mathcal{X} \times \mathcal{Y}$. One then uses $\mathrm{P}[X]$, the distribution of X, and a functional model $\ddot{X} \longrightarrow\!\!\!\!\!\rightarrow \dot{Y}$ to construct a statistical variable, say $\hat{Y}$, whose distribution approximates $\mathrm{P}[Y]$. Two approaches to the construction of $\mathrm{P}[\hat{Y}]$ can be distinguished:

a) Using a prediction rule for individual outcomes to define individual values of $\hat{Y}$, e.g. $\hat{Y}(\omega) := \mathrm{E}(\dot{Y} \,|\, \ddot{X} = X(\omega))$.

b) Directly deriving a distribution of $\hat{Y}$.

The first approach depends on specifying a prediction rule which can be done in several different ways, in particular when the outcome variable is qualitative. I therefore focus on the second approach which does not require prediction rules for individual outcomes. Following this approach, the construction of $\hat{Y}$'s distribution begins with conditional values:

$$\mathrm{P}(\hat{Y}{=}y \,|\, X{=}x) := \Pr(\dot{Y} = y \,|\, \ddot{X}{=}x)$$

One then uses the known distribution of X to derive the corresponding distribution of $\hat{Y}$:

$$\mathrm{P}(\hat{Y}{=}y) = \sum\nolimits_x \mathrm{P}(\hat{Y}{=}y \,|\, X{=}x)\, \mathrm{P}(X{=}x)$$

The construction can finally be assessed with two considerations:

- One can compare $\mathrm{P}[\hat{Y}]$ with $\mathrm{P}[Y]$ ('goodness of distributional fit').
- One can calculate the part of the variation of $\hat{Y}$ which can be attributed to variation of X ('explained variance').[1]

7. Numerical illustration. To illustrate the construction, I use Model 1.3.1 and assume the following data:

X	Z	$Y{=}0$	$Y{=}1$
0	1	300	300
0	2	80	320
1	1	40	160
1	2	80	720

The goodness of distributional fit depends on the parametric model that is used to approximate the functional model. Using a saturated model, the fit would be perfect. In the example:

$$\Pr(\dot{Y}{=}1|\ddot{X}{=}x, \dot{Z}{=}z){=}\mathrm{P}(Y{=}1|X{=}x, Z{=}z) \implies \mathrm{P}[\hat{Y}]{=}\mathrm{P}[Y]$$

The fit would not be perfect if one had used, for example, a logit model without an interaction term. And, of course, there will be no perfect fit when the statistical explanation concerns a set of data different from those that are used to estimate the model.

[1] Of cause, one could use other measures of variation instead of variance.

Finally, one can calculate the explained variance, that is, the part of the variance of $\hat{Y}$ (not of Y) which can be attributed to variation of X. Since the joint distribution of X and $\hat{Y}$ is known, one can apply a standard variance decomposition:

$$\mathrm{V}(\hat{Y}) = \mathrm{V}[\mathrm{M}(\hat{Y}|X)] + \mathrm{M}[\mathrm{V}(\hat{Y}|X)]$$

The first part can be interpreted as *explained variation:*

$$\mathrm{V}[\mathrm{M}(\hat{Y}|X)] = \sum\nolimits_x [\mathrm{M}(\hat{Y}|X{=}x) - \mathrm{M}(\hat{Y})]^2 \, \mathrm{P}(X{=}x)$$

The second part is the *residual variation*:

$$\mathrm{M}[\mathrm{V}(\hat{Y}|X)] = \sum\nolimits_x \mathrm{V}(\hat{Y}|X{=}x) \, \mathrm{P}(X{=}x)$$

In our illustration, assuming a saturated model, one finds $\mathrm{M}(\hat{Y}) = \mathrm{M}(Y) = 0.75$ and $\mathrm{V}(\hat{Y}) = \mathrm{V}(Y) = 0.1875$, and finally:

- explained variation: $\mathrm{V}[\mathrm{M}(\hat{Y}|X, Z)] = 0.0285$,
- residual variation: $\mathrm{M}[\mathrm{V}(\hat{Y}|X, Z)] = 0.1590$,
- proportion of explained variation: $0.0285/0.1875 \approx 15\%$.

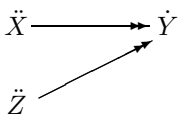
2. Relationships between explanatory variables

Statistical explanations most often consider many explanatory variables. The formulation of explanatory claims must include, then, an explication of relationships between the explanatory variables. This chapter discusses how functional models can be used to consider such relationships. The first section deals with interactions between explanatory variables and points to implications for understanding effects. The second section considers functional relationships between explanatory variables (which are different from interactions). In both sections, the discussion is general without presupposing a specific parametric form. The third section considers relationships between explanatory variables in the context of linear models for expectations.

2.1 Interactions between explanatory variables

1. A general definition of interaction. The leading idea is: $\ddot{X}$ and $\ddot{Z}$ are *interactive conditions* for the distribution of $\dot{Y}$ if the effect of a change in $\ddot{X}$ $[\ddot{Z}]$ depends on values of $\ddot{Z}$ $[\ddot{X}]$. The formulation shows that the presence of interaction also depends on the definition of 'effect'. To illustrate, I use

Model 2.1.1



with $\dot{Y}$ = indicator of a child's educational success; $\ddot{X}$ = educational level of the child's parents (0 low, 1 high); $\ddot{Z}$ = type of school the child is attending (0 or 1). The corresponding probabilistic function is

$$(x, z) \longrightarrow \Pr[\dot{Y}|\ddot{X}{=}x, \ddot{Z}{=}z] \tag{5}$$

The right-hand side denotes the probability distribution of $\dot{Y}$ in a situation where $\ddot{X} = x$ and $\ddot{Z} = x$. Assuming that $\dot{Y}$ is discrete, specific values of this distribution will be denoted by $\Pr(\dot{Y} = y | \ddot{X} = x, \ddot{Z} = z)$.

A simple definition of the effect of a change from $\ddot{X} = x'$ to $\ddot{X} = x''$ is given by

$$\Delta^s(\dot{Y}; \ddot{X}[x', x''], \ddot{Z} = z) := \mathrm{E}(\dot{Y} | \ddot{X} = x'', \ddot{Z} = z) - \mathrm{E}(\dot{Y} | \ddot{X} = x', \ddot{Z} = z) \tag{6}$$

Using this definition, one can easily invent examples with and without interaction between the two explanatory variables:

with interaction			without interaction		
x	z	$\mathrm{E}(\dot{Y} \mid \ddot{X} = x, \ddot{Z} = z)$	x	z	$\mathrm{E}(\dot{Y} \mid \ddot{X} = x, \ddot{Z} = z)$
0	0	0.5	0	0	0.5
0	1	0.7	0	1	0.7
1	0	0.8	1	0	0.7
1	1	0.9	1	1	0.9

(7)

2. Interaction in parametric models. The above definition of interaction between explanatory variables is independent of the parametric form of the function which relates these variables to the distribution of an outcome variable. When using parametric models, it depends on the parametric form whether, and how, interactions can be made visible.

Linear models for mean values require an explicit formulation of interaction terms (most often defined by multiplying variables). In contrast, almost all nonlinear models entail interaction effects already by virtue of their mathematical form. However, even then it depends on the details of the parametric form which interaction effects can be made visible.

The logit model can serve as an example. For Model 2.1.1, the standard formulation would be:

$$\Pr(\dot{Y} = 1 | \ddot{X} = x, \ddot{Z} = z) \approx \frac{\exp(\alpha + x\beta_x + z\beta_z)}{1 + \exp(\alpha + x\beta_x + z\beta_z)}$$

This formulation implies an interaction effect when using the effect definition (6), but not when using odds ratios:

$$\frac{\Pr(\dot{Y} = 1 | \ddot{X} = x'', \ddot{Z} = z) / \Pr(\dot{Y} = 0 | \ddot{X} = x'', \ddot{Z} = z)}{\Pr(\dot{Y} = 1 | \ddot{X} = x', \ddot{Z} = z) / \Pr(\dot{Y} = 0 | \ddot{X} = x', \ddot{Z} = z)} \approx \exp\left((x'' - x')\,\beta_x\right)$$

Of course, it will often be sensible to explicitly add an interaction term:

$$\Pr(\dot{Y} = 1 | \ddot{X} = x, \ddot{Z} = z) \approx \frac{\exp(\alpha + x\beta_x + z\beta_z + xz\beta_{xz})}{1 + \exp(\alpha + x\beta_x + z\beta_z + xz\beta_{xz})}$$

3. Implications for understanding effects. Interaction has an important consequence: If two explanatory variables interact, no one can be attributed a unique effect. Instead, one must speak of *context-dependent effects*. This notion is symmetrical: each of the interacting variables can be considered as providing a context for the other one.

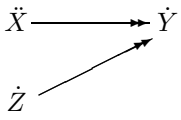
To illustrate, consider Model 2.1.1 as exemplified on the left-hand side of (7). The effect of parents' educational level depends on the school type:

$$\Delta^s(\dot{Y}; \ddot{X}[0, 1], \ddot{Z} = 0) = 0.8 - 0.5 = 0.3$$
$$\Delta^s(\dot{Y}; \ddot{X}[0, 1], \ddot{Z} = 1) = 0.9 - 0.7 = 0.2$$

Conversely, the effect of the school type depends on the parents' educational level.

4. Assuming distributions for exogenous variables. In general, whether it is possible to define unique effects also depends on whether the interacting variables are exogenous or endogenous (as specified in a functional model). In Model 2.1.1, both explanatory variables are exogenous. In order to define a unique effect of one of the variables, say $\ddot{X}$, it could be sensible to assume a distribution for the other variable. $\ddot{Z}$ is then substituted by a variable $\dot{Z}$ which is still an exogenous variable with a distribution not depending on $\ddot{X}$. This leads to a modified model:

Model 2.1.2



This modified model would allow one to define a mean effect:

$$\Delta^s(\dot{Y}; \ddot{X}[x', x''], \dot{Z}) := \sum\nolimits_z \big(\mathrm{E}(\dot{Y}|\ddot{X}=x'', \dot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x', \dot{Z}=z)\big)\,\Pr(\dot{Z}=z)$$

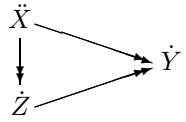
However, given that $\ddot{X}$ and $\dot{Z}$ interact, the effect still depends on the distribution of $\dot{Z}$. This can easily be seen when using the data on the left-hand side of (7):

$$\Delta^s(\dot{Y}; \ddot{X}[0,1], \dot{Z}) = (0.8 - 0.5)\,\Pr(\dot{Z}=0) + (0.9 - 0.7)\,\Pr(\dot{Z}=1)$$

2.2 Functional relations between explanatory variables

1. Functional relationships and interactions. The idea of interaction concerns the probabilistic function which relates the explanatory variables to the outcome variable. A different question concerns functional relationships between explanatory variables. As an example, I consider a modification of Model 2.1.1:

Model 2.2.1



This model, in addition to the functional relationship

$$(x, z) \longrightarrow \Pr[\dot{Y}|\ddot{X}=x, \dot{Z}=z] \tag{8}$$

which corresponds to (5), also assumes a relationship

$$x \longrightarrow \Pr[\dot{Z}|\ddot{X}=x] \tag{9}$$

that shows how the child's school type depends on the parents' educational level. Whether there is interaction between $\ddot{X}$ and $\dot{Z}$ w.r.t. $\dot{Y}$ is completely independent of this function relating $\ddot{X}$ and $\dot{Z}$. For later illustrations, I assume

$$\Pr(\dot{Z}=1 \,|\, \ddot{X}=0) = 0.4 \quad \text{and} \quad \Pr(\dot{Z}=1 \,|\, \ddot{X}=1) = 0.8 \tag{10}$$

2. Mediator and moderator variables. Authors often distinguish between mediator and moderator variables (e.g. Baron and Kenny 1986, MacKinnon 2008). In the context of functional models, one can use the following definitions:

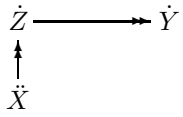
$\dot{Z}$ is a *mediator variable* for $\ddot{X}$ [or $\dot{X}$] w.r.t. another variable, $\dot{Y}$, if $\dot{Z}$ lies on a directed path leading from $\ddot{X}$ [or $\dot{X}$] to $\dot{Y}$.

$\dot{Z}$ is a *moderator variable* w.r.t. a relationship between $\ddot{X}$ [or $\dot{X}$] and $\dot{Y}$ if the effect of $\ddot{X}$ [or $\dot{X}$] on $\dot{Y}$ depends on values of $\dot{Z}$.

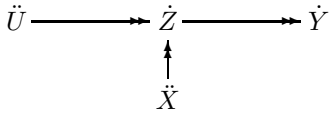
Using these definitions, $\dot{Z}$ is a mediator variable in Model 2.2.1. If $\dot{Z}$ interacts with $\ddot{X}$, it is also a moderator variable. On the other hand, in Model 2.1.1, $\ddot{Z}$ is not a mediator variable, but it is a moderator variable if it interacts with $\ddot{X}$.

One can think, of course, of many different constellations. In the following, I will consider mainly four types:

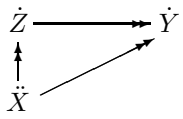
Model 2.2.2a



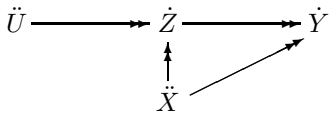
Model 2.2.2c



Model 2.2.2b



Model 2.2.2d



The leading question is, How to define effects of the explanatory variables in these models?

3. Effects of exogenous variables. For models 2.2.2a and 2.2.2b one can derive a *total effect* of $\ddot{X}$, that is, an effect that integrates mediating variables. Starting from

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

the total effect in 2.2.2b is:

$$\begin{aligned}\Delta^s(\dot{Y};\ddot{X}[x',x'']) = \sum\nolimits_z \Big(&\mathrm{E}(\dot{Y}|\ddot{X}=x'', \dot{Z}=z)\,\Pr(\dot{Z}=z|\ddot{X}=x'')- \\ &\mathrm{E}(\dot{Y}|\ddot{X}=x', \dot{Z}=z)\,\Pr(\dot{Z}=z|\ddot{X}=x')\Big)\end{aligned} \tag{11}$$

The total effect in 2.2.2a is:

$$\Delta^s(\dot{Y};\ddot{X}[x',x'']) = \sum\nolimits_z \mathrm{E}(\dot{Y}|\dot{Z}=z)\left(\Pr(\dot{Z}=z|\ddot{X}=x'') - \Pr(\dot{Z}=z|\ddot{X}=x')\right)$$

To illustrate the calculation of a total effect in Model 2.2.2b, I use the data in (7) and (10).

Left-hand side of (7):

$$\mathrm{E}(\dot{Y}|\ddot{X}=1) = 0.8\cdot 0.2 + 0.9\cdot 0.8 = 0.88$$
$$\mathrm{E}(\dot{Y}|\ddot{X}=0) = 0.5\cdot 0.6 + 0.7\cdot 0.4 = 0.58$$
$$\text{Total effect} = 0.88 - 0.58 = 0.3$$

Right-hand side of (7):

$$\mathrm{E}(\dot{Y}|\ddot{X}=1) = 0.7\cdot 0.2 + 0.9\cdot 0.8 = 0.86$$
$$\mathrm{E}(\dot{Y}|\ddot{X}=0) = 0.5\cdot 0.6 + 0.7\cdot 0.4 = 0.58$$
$$\text{Total effect} = 0.86 - 0.58 = 0.28$$

4. Effects of endogenous variables. When thinking of effects of endogenous variables, a first difficulty concerns that such variables have distributions which depend on values of other variables. To circumvent this difficulty, I assume that one can nevertheless hypothetically fix values of endogenous variables. Given this presupposition, one can immediately define effects of $\dot{Z}$ in models 2.2.2a and 2.2.2c:

$$\Delta^s(\dot{Y}; \dot{Z}[z', z'']) = \mathrm{E}(\dot{Y}|\dot{Z}=z'') - \mathrm{E}(\dot{Y}|\dot{Z}=z')$$

In models 2.2.2b and 2.2.2d, effects of $\dot{Z}$ are context-dependent on values of $\ddot{X}$. For example, in 2.2.2b:

$$\Delta^s(\dot{Y}; \dot{Z}[z', z''], \ddot{X}=x) = \mathrm{E}(\dot{Y}|\dot{Z}=z'', \ddot{X}=x) - \mathrm{E}(\dot{Y}|\dot{Z}=z', \ddot{X}=x)$$

5. Direct and indirect effects. Model 2.2.2b (= 2.2.1) leads to the further question of whether one can define, not only a total, but also a *direct effect* of $\ddot{X}$ on the expectation of $\dot{Y}$. A positive answer requires that it makes sense to hypothetically held constant a value of $\dot{Z}$, although its distribution changes with $\ddot{X}$.

Even given this presupposition, one can only define a direct effect if $\ddot{X}$ and $\dot{Z}$ do not interact. If there is no interaction, the effect

$$\Delta^s(\dot{Y}; \ddot{X}[x', x''], \dot{Z}=z)$$

is independent of z and can sensibly be interpreted as a direct effect of $\ddot{X}$. This is illustrated by the data on the right-hand side of (7):

$$\Delta^s(\dot{Y}; \ddot{X}[0,1], \dot{Z}=0) = \Delta^s(\dot{Y}; \ddot{X}[0,1], \dot{Z}=1) = 0.2$$

An indirect effect can then be defined as the difference between the total and the direct effect. Starting from the total effect (11), the indirect effect can be written as

$$\sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}=x', \dot{Z}=z) \left(\Pr(\dot{Z}=z|\ddot{X}=x'') - \Pr(\dot{Z}=z|\ddot{X}=x')\right) \tag{12}$$

Using again the right-hand side of (7), the indirect effect is 0.08, and the total effect is 0.2+0.08 = 0.28.

If, however, $\ddot{X}$ and $\dot{Z}$ do interact, it is not possible to define a unique direct effect even if one assumes that values of $\dot{Z}$ can be fixed. This is illustrated by the data on the left-hand side of (7):

$$\Delta^s(\dot{Y}; \ddot{X}[0,1], \dot{Z}=0) = 0.3 \text{ and } \Delta^s(\dot{Y}; \ddot{X}[0,1], \dot{Z}=1) = 0.2$$

Consequently, there also is no unique indirect effect.

6. Counterfactual effect decompositions. If there is an interaction between $\ddot{X}$ and the mediating variable $\dot{Z}$, direct effects can only be defined for each value of $\dot{Z}$ separately. It is, of course, possible to define versions of mean direct effects. Following this idea, some authors have proposed a 'natural direct effect' (e.g. Pearl 2001, Petersen et al. 2006). In the framework of functional models, this form of a mean direct effect can be defined as

$$\sum\nolimits_z \left(\mathrm{E}(\dot{Y}|\ddot{X}=x'', \dot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x', \dot{Z}=z)\right) \Pr(\dot{Z}=z|\ddot{X}=x') \tag{13}$$

The idea is to use the distribution of $\dot{Z}$ that corresponds to the initial value of $\ddot{X}$, that is x', and to assume ('counterfactually') that this distribution would not change if the value of $\ddot{X}$ changes from x' to x''.

Of course, one also may use other distributions of $\dot{Z}$. In any case, the total effect of $\ddot{X}$ can be divided into a mean direct and a mean indirect effect. Using (13) for the mean direct effect, one gets

$$\begin{aligned} &\mathrm{E}(\dot{Y}|\ddot{X}{=}x'') - \mathrm{E}(\dot{Y}|\ddot{X}{=}x') = \\ &\quad \sum\nolimits_z \big[\mathrm{E}(\dot{Y}|\ddot{X}{=}x'', \dot{Z}{=}z) - \mathrm{E}(\dot{Y}|\ddot{X}{=}x', \dot{Z}{=}z)\big] \Pr(\dot{Z}{=}z|\ddot{X}{=}x') + \\ &\quad \sum\nolimits_z \mathrm{E}(\dot{Y}|\ddot{X}{=}x'', \dot{Z}{=}z) \left[\Pr(\dot{Z}{=}z|\ddot{X}{=}x'') - \Pr(\dot{Z}{=}z|\ddot{X}{=}x')\right] \end{aligned} \tag{14}$$

where the second term on the right-hand side is then interpreted as a mean indirect effect.

To illustrate the decomposition for a situation with interaction, I use the data on the left-hand side of (7) together with (10). The mean direct effect is

$$(0.8 - 0.5)\,0.6 + (0.9 - 0.7)\,0.4 = 0.26$$

the mean indirect effect is

$$0.8\,(0.2 - 0.6) + 0.9\,(0.8 - 0.4) = 0.04$$

and the total effect is $0.26 + 0.04 = 0.3$.

7. Confounders and independent context variables. There is no unique definition of 'confounding variables' (see, e.g., Weinberg 1993). In the context of functional models, I use the following definition:

> A variable $\dot{X}$ [or $\ddot{X}$] is a confounder w.r.t. a functional dependence of $\dot{Y}$ on $\dot{Z}$ if there is a directed path from $\dot{X}$ to $\dot{Y}$ and (a) there is a directed path from $\dot{X}$ [or $\ddot{X}$] to $\dot{Z}$ (direct confounding), or (b) there is a further variable, say $\dot{U}$ [or $\ddot{U}$], and a directed path leads from $\dot{U}$ to $\dot{Z}$ and from $\dot{U}$ to $\dot{X}$ (indirect confounding).

For example, in Model 2.2.2b, $\ddot{X}$ is a directly confounding variable w.r.t. the dependence of $\dot{Y}$ on $\dot{Z}$. In contrast, $\dot{Z}$ is a mediator variable, not a confounder, w.r.t. the relationship between $\ddot{X}$ and $\dot{Y}$. Notice that the definition presupposes a functional model with directional relationships between variables and cannot be formulated in terms of 'correlation'.

The proposed definition distinguishes confounding variables from independent context variables. Without an arrow from $\ddot{X}$ to $\dot{Z}$ in Model 2.2.2b, $\ddot{X}$ would be an *independent context variable*, not a confounder.

The distinction between these two kinds of covariates concerns possibilities to define effects. Different difficulties show up, in particular, when the variables are not observed. Consider the independent context variable, $\ddot{X}$, in Model 2.1.1. It is a context variable for effects of $\ddot{Z}$ on $\dot{Y}$:

$$\Delta^s(\dot{Y}; \ddot{Z}[z', z''], \ddot{X}{=}x) = \mathrm{E}(\dot{Y}|\ddot{Z}{=}z'', \ddot{X}{=}x) - \mathrm{E}(\dot{Y}|\ddot{Z}{=}z', \ddot{X}{=}x)$$

If $\ddot{X}$ is not observed, one can think instead of a variable, $\dot{X}$, having an unknown distribution. Since this distribution does not depend on $\ddot{Z}$, the observed effect can be considered as a mean effect w.r.t. the unknown distribution of $\dot{X}$:

$$\Delta^s(\dot{Y}; \ddot{Z}[z', z''], \dot{X}) = \sum\nolimits_x \Delta^s(\dot{Y}; \ddot{Z}[z', z''], \dot{X}{=}x) \Pr(\dot{X} = x)$$

Moreover, if there is no interaction between $\ddot{Z}$ and $\dot{X}$, one can attribute the effect uniquely to $\ddot{Z}$.

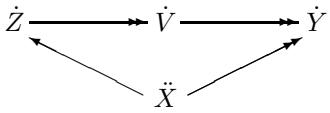
A similar consideration is not possible for unobserved confounding variables. Consider the confounding variable $\ddot{X}$ in Model 2.2.2b (= 2.2.1). If this variable is not observed, one can instead assume a random variable $\dot{X}$ having an unknown distribution. An effect of $\dot{Z}$ can then be expressed as

$$\mathrm{E}(\dot{Y}|\dot{Z}=z'') - \mathrm{E}(\dot{Y}|\dot{Z}=z') = \sum\nolimits_x \Big(\mathrm{E}(\dot{Y}|\dot{Z}=z'',\dot{X}=x)\Pr(\dot{X}=x|\dot{Z}=z'') - \mathrm{E}(\dot{Y}|\dot{Z}=z',\dot{X}=x)\Pr(\dot{X}=x|\dot{Z}=z')\Big)$$

This effect is not only due to different values of $\dot{Z}$, but also to different distributions of $\dot{X}$ (associated with different values of $\dot{Z}$). The effect cannot, therefore, be interpreted as a mean effect w.r.t. an unknown, but common, distribution of the confounding variable.

8. Effects of mediator variables. Interestingly, the mentioned difficulties do not arise w.r.t. mediator variables. Consider the following

Model 2.2.3



which is similar to Model 2.2.2b but in addition contains the mediator variable $\dot{V}$. As before, if the confounder, $\ddot{X}$, is not observed, the observable relationship between $\dot{Y}$ and $\dot{Z}$ is difficult to interpret. The situation is different, however, w.r.t. the mediator variable $\dot{V}$.

In order to see that, substitute $\ddot{X}$ by a random variable $\dot{X}$ having an unknown distribution. The model entails the relationships

$$\Pr[\dot{Y}|\dot{X}=x,\dot{V}=v,\dot{Z}=z] = \Pr[\dot{Y}|\dot{X}=x,\dot{V}=v] \tag{15}$$

$$\Pr[\dot{V}|\dot{Z}=z,\dot{X}=x] = \Pr[\dot{V}|\dot{Z}=z] \tag{16}$$

and therefore[1]

$$\Pr[\dot{X}|\dot{V}=v,\dot{Z}=z] = \Pr[\dot{X}|\dot{Z}=z] \tag{17}$$

Using these relationships, one finds

$$\mathrm{E}(\dot{Y}|\dot{V}=v,\dot{Z}=z) = \sum\nolimits_x \mathrm{E}(\dot{Y}|\dot{V}=v,\dot{Z}=z,\dot{X}=x)\Pr(\dot{X}=x|\dot{V}=v,\dot{Z}=z) = \sum\nolimits_x \mathrm{E}(\dot{Y}|\dot{V}=v,\dot{X}=x)\Pr(\dot{X}=x|\dot{Z}=z)$$

from which one derives

$$\mathrm{E}(\dot{Y}|\dot{V}=v'',\dot{Z}=z) - \mathrm{E}(\dot{Y}|\dot{V}=v',\dot{Z}=z) = \sum\nolimits_x \big[\mathrm{E}(\dot{Y}|\dot{V}=v'',\dot{X}=x) - \mathrm{E}(\dot{Y}|\dot{V}=v',\dot{X}=x)\big]\Pr(\dot{X}=x|\dot{Z}=z)$$

This shows that, conditional on values of $\dot{Z}$, effects of $\dot{V}$ can be interpreted as mean effects w.r.t.

[1] Starting from

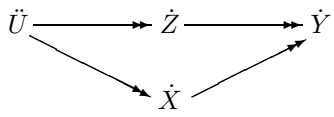
$$\Pr[\dot{X},\dot{V}|\dot{Z}] = \Pr[\dot{X}|\dot{V},\dot{Z}]\Pr[\dot{V}|\dot{Z}] = \Pr[\dot{V}|\dot{X},\dot{Z}]\Pr[\dot{X}|\dot{Z}]$$

and using (16), (17) immediately follows.

an unknown distribution of the confounding variable $\dot{X}$. Moreover, if there is no interaction between $\dot{V}$ and $\dot{X}$, one can attribute these effects uniquely to $\dot{V}$.

9. Indirectly connected explanatory variables. So far the discussion dealt with directly confounding variables. I now consider indirectly confounding variables. The following model illustrates this case:

Model 2.2.4



In this model, $\dot{X}$ is a confounder w.r.t. to the dependence of $\dot{Y}$ on $\dot{Z}$ according to the second part of the above definition. If $\dot{X}$ is observed, effects of $\dot{Z}$ can be calculated conditional on values of $\dot{X}$. (Further considerations depend on whether $\dot{X}$ and $\dot{Z}$ interact, see above.)

If $\dot{X}$ is not observed, one can begin with explicitly conditioning on values of $\ddot{U}$ and use the assumption entailed by the model that $\dot{Z}$ and $\dot{X}$ are independent conditional on values of $\ddot{U}$:

$$\begin{aligned}\mathrm{E}(\dot{Y}|\dot{Z}=z,\ddot{U}=u) = \sum\nolimits_x \mathrm{E}(\dot{Y}|\dot{Z}=z,\dot{X}=x,\ddot{U}=u)\Pr(\dot{X}=x|\dot{Z}=z,\ddot{U}=u) = \\ \sum\nolimits_x \mathrm{E}(\dot{Y}|\dot{Z}=z,\dot{X}=x)\Pr(\dot{X}=x|\ddot{U}=u) \qquad (18)\end{aligned}$$

This shows that, conditional on values of $\ddot{U}$, the observed effect of $\dot{Z}$ can be interpreted as a mean effect w.r.t. an unknown distribution of the confounding variable $\dot{X}$.

Of course, also $\ddot{U}$ is a confounding variable, and no reliable conclusions about effects of $\dot{Z}$ on $\dot{Y}$ can be drawn if also $\ddot{U}$ is not observed.

2.3 Linear regression models

The discussion so far was general and did not rely on any parametric assumptions. I now consider some of the previously discussed relationships in the context of linear models for expectations.

1. Linear models for expectations. Corresponding to Model 2.1.1, a simple linear version of a parametric model for the expectation of $\dot{Y}$ can be written as

$$\mathrm{E}(\dot{Y}|\ddot{X}=x,\ddot{Z}=z) = \beta_0 + x\beta_x + z\beta_z \qquad (19)$$

The model entails that $\ddot{X}$ and $\ddot{Z}$ do not interact. Effects of $\ddot{X}$ are then given by

$$\mathrm{E}(\dot{Y}|\ddot{X}=x'') - \mathrm{E}(\dot{Y}|\ddot{X}=x') = (x''-x')\beta_x \qquad (20)$$

and do not depend on values of $\ddot{Z}$.

What happens if $\ddot{Z}$ is omitted? In order to derive a reduced model, $\ddot{Z}$ must be substituted by a random variable, $\dot{Z}$, having an unknown distribution. If one further assumes that $\dot{Z}$ is independent of $\ddot{X}$, there is a simple result:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = \sum\nolimits_z (\beta_0 + x\beta_x + z\beta_z)\Pr(\dot{Z}=z) = \beta_0 + x\beta_x + \mathrm{E}(\dot{Z})\beta_z$$

showing that effects of $\ddot{X}$ are still given by (20).

2. Linear models with interactions. In its general formulation, Model 2.1.1 does not make assumptions about interactions between $\ddot{X}$ and $\ddot{Z}$. Such assumptions belong to the form of a parametric model. When using linear models, interaction effects must be explicitly included, in our example:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x,\ddot{Z}=z) = \beta_0 + x\beta_x + z\beta_z + xz\beta_{xz} \tag{21}$$

Effects of $\ddot{X}$ now depend on values of $\ddot{Z}$:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x'',\ddot{Z}=z) - \mathrm{E}(\dot{Y}|\ddot{X}=x',\ddot{Z}=z) = (x''-x')(\beta_x + z\beta_{xz})$$

This also changes the consequences of omitting $\ddot{Z}$. If one substitutes $\ddot{Z}$ by $\dot{Z}$ and still assumes that $\dot{Z}$ is independent of $\ddot{X}$, one finds

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = (\beta_0 + \mathrm{E}(\dot{Z})\beta_z) + x\,(\beta_x + \mathrm{E}(\dot{Z})\beta_{xz})$$

Effects of $\ddot{X}$ now also depend on an unknown mean value of $\dot{Z}$.

3. Mediator variables and total effects. So far I have assumed that $\dot{Z}$ is independent of $\ddot{X}$. I now assume that it is a mediator variable for $\ddot{X}$ w.r.t. $\dot{Y}$, corresponding to Model 2.2.1. Then starting from (19) without interaction, one gets

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = \beta_0 + x\beta_x + \mathrm{E}(\dot{Z}|\ddot{X}=x)\beta_z \tag{22}$$

If one assumes a linear model for the expectation of $\dot{Z}$,

$$\mathrm{E}(\dot{Z}|\ddot{X}=x) = \gamma_0 + x\gamma_x \tag{23}$$

one also gets a linear formulation for the reduced model:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = (\beta_0 + \gamma_0\beta_z) + x\,(\beta_x + \gamma_x\beta_z)$$

Since $\dot{Z}$ is a mediator variable, $\beta_x + \gamma_x\beta_z$ can be interpreted as representing the total effect of $\ddot{X}$ on $\dot{Y}$:

$$\mathrm{E}(\dot{Y}|\ddot{X}=x'') - \mathrm{E}(\dot{Y}|\ddot{X}=x') = (x''-x')(\beta_x + \gamma_x\beta_z)$$

One gets a more involved formulation when starting from (21) which includes an interaction effect. Again assuming (23), one gets

$$\mathrm{E}(\dot{Y}|\ddot{X}=x) = (\beta_0 + \gamma_0\beta_z) + x\,(\beta_x + \gamma_x\beta_z + \gamma_0\beta_{xz}) + x^2\gamma_x\beta_{xz}$$

$\dot{Y}$ now depends in a nonlinear way on $\ddot{X}$, but $\mathrm{E}(\dot{Y}|\ddot{X}=x'') - \mathrm{E}(\dot{Y}|\ddot{X}=x')$ can still be interpreted as a total effect of $\ddot{X}$ on $\dot{Y}$.

4. Direct and indirect effects. In both cases, the total effect can be decomposed into a direct and an indirect effect. Starting from (19) without interaction, one can follow the consideration in **2.2.5**. The direct effect is given by $(x''-x')\beta_x$, and the indirect effect can be calculated according to (12):

$$(\mathrm{E}(\dot{Z}|\ddot{X}=x'') - \mathrm{E}(\dot{Z}|\ddot{X}=x'))\,\beta_z = (x''-x')\,\gamma_x\beta_z$$

When starting from model (21), which includes an interaction effect, one can follow the approach that was described in **2.2.6**. Based on the decomposition (14), one finds the counterfactual direct effect

$$(x''-x')(\beta_x + \gamma_0\beta_{xz} + x'\gamma_x\beta_{xz})$$

and the counterfactual indirect effect

$$(x'' - x')(\gamma_x \beta_z + x'' \gamma_x \beta_{xz})$$

5. Omitting a confounding variable. Consider again Model 2.2.1, but now assume that one has omitted the variable $\ddot{X}$. This is a confounding variable w.r.t. the effect of $\dot{Z}$ on $\dot{Y}$. As was argued in **2.2.7**, the effect of $\dot{Z}$ in the reduced model is not easily interpretable. Linear regression models offer no advantage. Starting from (19), the reduced model omitting $\dot{X}$ (which is the substitute for $\ddot{X}$) becomes:

$$\mathrm{E}(\dot{Y}|\dot{Z}=z) = \beta_0 + z\beta_z + \mathrm{E}(\dot{X}|\dot{Z}=z)\beta_x$$

While formally similar to (22), there is no similar interpretation. In (22), one can interpret $\mathrm{E}(\dot{Z}|\ddot{X}=x)\beta_z$ as part of the total effect of $\ddot{X}$ on $\dot{Y}$. In contrast, $\mathrm{E}(\dot{X}|\dot{Z}=z)$ cannot be interpreted as a substantive dependency relation. Starting from the dependency relation assumed in Model 2.2.1, which goes from $\ddot{X}$ to $\dot{Z}$, one can derive

$$\mathrm{E}(\dot{X}|\dot{Z}=z) = \frac{\sum_x x \Pr(\dot{Z}=z|\dot{X}=x)\Pr(\dot{X}=x)}{\sum_x \Pr(\dot{Z}=z|\dot{X}=x)\Pr(\dot{X}=x)}$$

So it depends essentially on the distribution of $\dot{X}$ which is determined outside of the model.

3. Causal interpretations

The functional relationships which are posited when specifying a functional model do not automatically have a causal meaning. In general, their presupposition only entails the claim that variables used as arguments in a function can contribute to predicting conditional distributions of the dependent variable. For example, one can use a person's educational level to predict the educational level of her parents; but there obviously is no corresponding causal relationship. So the question arises in which sense one sometimes can claim that relationships in a functional model also have a causal meaning.

This question concerns the understanding of causal relationships and must be distinguished from the further question of how one can use statistical data for estimating quantitative (numerically specified) forms of such relationships. In the present chapter, I am mainly concerned with the conceptual question. The first section proposes a rudimentary understanding of the idea that variables can be causally related. In the second section, I introduce a distinction between explanatory and treatment models, and suggest that there correspondingly are some differences in the understanding of causal effects. The third section takes up the often used notion of 'potential outcomes' and shows that there are two quite different understandings.

3.1 Causal relations between variables

1. Causally relevant variables. Consider a functional relationship $x \longrightarrow \Pr[\dot{Y}|\ddot{X}=x]$. What is entailed by the idea that this can be viewed as a causal relationship? I propose that there are basically two claims:

(1) that one can refer to a fact-generating process generating values of $\dot{Y}$,[1] and

(2) that this process depends on values of $\ddot{X}$.

[1] See the distinction between data-generating and fact-generating processes introduced in **1.1.2**.

I then call $\ddot{X}$ a variable which is *causally relevant* for $\dot{Y}$.

These two claims are at the core of viewing causation as a generative process (see also Cox 1992, Goldthorpe 2001, Blossfeld 2009). Of course, they also provide a starting point for further questions: How do processes generating values of $\dot{Y}$ depend on $\ddot{X}$, and which role is played by further variables on which $\dot{Y}$ possibly depends?

2. Thinking of causes as events. In order to think about how processes generating values of $\dot{Y}$ depend on $\ddot{X}$ one must be more specific about the meaning of the values of $\ddot{X}$. I begin with assuming that values of $\ddot{X}$ represent events. In the most simple case, $\ddot{X}$ is a binary variable, and $\ddot{X}=1$ means that an event of a specified kind has occurred, and $\ddot{X}=0$ means that such an event has not (yet) occurred.

One can then distinguish two kinds of relationships between $\ddot{X}$ and a process generating values of $\dot{Y}$:

a) The event initiates a process that generates a value of $\dot{Y}$. As an example, think of $\dot{Y}$ as the outcome of a student's participating in a learning frame. The student's *beginning* to participate in the learning frame can be considered as an event that initiates a process that eventually generates a value of $\dot{Y}$.

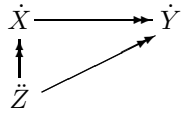
b) The event occurs while a process that eventually generates a value of $\dot{Y}$ already takes place. So one can think of two such processes: one during which the event did occur, and another one in which the event did not occur. The impact of the event, if it occurs, must then be understood as modifying an ongoing process. As an example, one can think that the student becomes severely ill while participating in the learning frame.

Most often, already the definition of a process requires to refer to an event that initiates the process. The causal relevance of that event is then easily stated: Its occurrence is a necessary condition for the process to take place.

Variables representing the occurrence of events will be called *event variables* (Rohwer 2010, ch. 7). These need not be binary variables which refer to just one event type. In general, if $\dot{X}$ is an event variable, its domain will be denoted by $\mathcal{X} = \{0, 1, \ldots, m\}$, with values having the following meaning: If $j > 0$, $\dot{X}=j$ means that an event of the type j has occurred; $\dot{X}=0$ means that no event of the specified kinds has yet occurred. The definition shows that using event variables at least implicitly requires a temporal view.

3. Causally relevant conditions. Causally relevant variables need not be event variables. As another kind one can think of variables representing conditions on which a process depends. I then speak of *context variables*. To illustrate, I use our standard example in which a student's educational success, $\dot{Y}$, depends on the school type, $\dot{X}$, and on the parents' educational level, $\ddot{Z}$. The functional model looks as follows:

Model 3.1.1



In this model, one can think of $\dot{X}$ as a variable representing the occurrence of an event:

$$\dot{X} = \begin{cases} 1 & \text{if the student starts participating in learning frame } \sigma_1 \\ 2 & \text{if the student starts participating in learning frame } \sigma_2 \end{cases}$$

In contrast, $\ddot{Z}$, recording the parents' educational level, is a context variable. Its values can sensibly be understood as characterizing the context in which the process that generates a value of $\dot{X}$ takes place. Similarly, one can understand the causal relevance of $\ddot{Z}$ for processes generating values of $\dot{Y}$. The leading idea is that parents' activities through which they support a child's education depend on their own educational level.

How to think of the causal relevance of $\dot{X}$? There are two considerations. First, thinking of the event $\dot{X} = j$, it can be understood as initiating a process that eventually generates a value of $\dot{Y}$. In this view, $\dot{X}$ is causally relevant because without $\dot{X}$'s taking a positive value a process generating a value of $\dot{Y}$ cannot take place. Second, as soon as one of the possible events did occur, it can be viewed as having generated a specific context. If $\dot{X} = j$, it is the context σ_j, a particular learning frame, in which the process generating a value of $\dot{Y}$ takes place.

This can be generally stated: As soon as an event variable has a positive value, the variable can be considered as a context variable for a process that begins at the point in time when the event occurred.

4. Comparative and dynamic effects. To speak of a causally relevant variable, say $\dot{X}$, presupposes a functional model in which the variable has a particular place and can be functionally related to other variables representing possible effects. However, it is not the model, understood as a system of mathematical functions, that provides the causal meaning. To give a variable a causal meaning requires considerations which cannot be expressed in terms of mathematical functions.

The functional model is silent about the meaning of its functional relationships. But given a causal interpretation, it can be used to formally define causal effects. One specifies a variable, say $\dot{Y}$, representing the outcomes of interest, and considers all variables on which $\dot{Y}$ functionally depends. Assume that these are the variables $\dot{X}$ and $\dot{Z}$. Both can then be used to define effects. For example, an effect of $\dot{X}$ can be defined by

$$\Delta^s(\dot{Y}; \dot{X}[x', x''], \dot{Z} = z) = \mathrm{E}(\dot{Y}|\dot{X} = x'', \dot{Z} = z) - \mathrm{E}(\dot{Y}|\dot{X} = x', \dot{Z} = z) \tag{24}$$

This definition compares the expectation of $\dot{Y}$ in two situations: one in which $\dot{X} = x'$ and another one in which $\dot{X} = x''$, and further presupposes that $\dot{Z} = z$ in both situations. In this sense, the definition formulates a *comparative effect*, and can be used for all kinds of causally relevant variables. If $\dot{X}$ is an event variable, also another effect definition becomes possible:

$$\Delta^d(\dot{Y}; \dot{X}[j], \dot{Z} = z) = \mathrm{E}(\dot{Y}|\dot{X} = j, \dot{Z} = z) - \mathrm{E}(\dot{Y}|\dot{X} = 0, \dot{Z} = z) \tag{25}$$

This definition formulates a *dynamic effect*; it compares a situation where the event $\dot{X} = j$ occurred with a situation where no event (of the kinds specified by $\dot{X}$) occurred.[2]

To illustrate, consider the example introduced in **3.1.3**. A comparative effect compares the educational outcomes in the two learning frames, σ_1 and σ_2. In contrast, a dynamic effect compares the outcome of $\dot{X} = j$ with $\dot{X} = 0$, and since a positive value of $\dot{X}$ is a necessary condition for a value of $\dot{Y}$, the dynamic effect is given by $\mathrm{E}(\dot{Y}|\dot{X} = j, \dot{Z} = z)$.

5. Functional and causal mechanisms. Without presupposing possible effects one cannot think of 'causes'. In a statistical approach to causality, possible effects are conceptualized by an outcome variable, $\dot{Y}$. In social research, being interested in processes generating values of $\dot{Y}$, it is seldom reasonable to consider only a single causal condition, say $\ddot{X}$. In most applications one has to take into account further causally relevant conditions. The question 'What is the causal effect of $\ddot{X}$ on $\dot{Y}$?', without further qualification, is then not appropriate.

[2] If the occurrence of an event specified by $\dot{X}$ is a necessary condition for $\dot{Y}$'s getting a value, I use the convention that $\mathrm{E}(\dot{Y}|\dot{X} = 0, \dot{Z} = z) = 0$; see Rohwer (2010, ch. 7).

An alternative is to think of 'mechanisms'.[3] Here I use the term in this sense: A *mechanism* is an explicitly defined framework for thinking of processes generating values of an outcome variable. More specifically, I use the term *functional mechanism* to denote a functional model that shows how an outcome variable depends on other variables; and it will be called a *causal mechanism* if at least some of the functional relationships can be given a causal interpretation.

Given these definitions, a mechanism is a (formal) framework and must be distinguished from the processes which, possibly, take place according to the rules of the mechanism. This entails that a mechanism is not by itself a dynamic entity. While one can sensibly think that a process can generate an outcome, this cannot be said of a mechanism. But note that only the mechanism has an explicit representation (in terms of variables and functional relations). To think of processes that actually generate outcomes requires a causal interpretation of the mechanism.

3.2 References to human actors

1. Primary and secondary actors. Models in social research most often concern processes which depend on the behavior of human actors. In the following, I call these the *primary actors*. In contrast, I speak of *secondary actors* when referring to those who construct and use models.

For example, think of the models dealing with students' educational outcomes in different learning frames. Primary actors are the students, their parents, the teachers; in general, all actors to which one refers when interpreting the models and reflecting about causal relationships. In contrast, the secondary actors are those who construct, discuss, and possibly use these models for one reason or another. Note that the distinction presupposes the reference to a model. Only w.r.t. a model can one distinguish primary and secondary actors in the proposed sense.

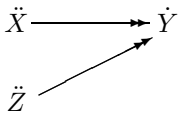
2. Explanatory and treatment models. There are many reasons why one could be interested in models. Here I want to mention just one distinction that also suggests to distinguish two kinds of models.

One interest concerns the primary actors, the conditions and outcomes of their behavior. Models are then constructed as tools for understanding and explaining conditions and outcomes of the behavior of the primary actors. I then speak of *explanatory models*.

In contrast, secondary actors could be interested in the possibility of interventions supporting their particular goals. Models are then constructed as tools for assessing the possible effects of interventions (often called 'treatments'). I then speak of *intervention* or *treatment models*.

3. Functional models of experiments. Discussions of causally interpretable models often presuppose an interest in effects of treatments. Grounded in a long tradition, treatment models are preferably related to an experimental context. I briefly mention some ways in which functional models can be used as a formal framework. A first possibility is illustrated by the following model.

Model 3.2.1



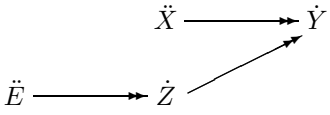
Values of $\ddot{X}$ represent the treatments whose causal effects are of primary interest. $\ddot{Z}$ records

[3] For a discussion of many of the meanings which are given to this term in the sociological literature, see Mahoney (2001).

further conditions which, presumably, are causally relevant for $\dot{Y}$. Both are exogenous variables because their values are deliberately fixed by the experimenter.

In another kind of experiment, the experimenter randomly selects some of the conditions for the experiment but still deliberately generates values of the treatment variable. This can be represented by the following model.

Model 3.2.2

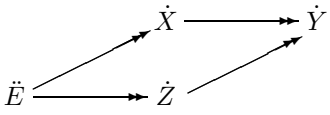


$\ddot{X}$ is still an exogenous variable, but $\dot{Z}$ is now an endogenous variable which depends in a specified way on $\ddot{E}$, an event variable initiating the experiment.

In both situations, assuming that the experiment concerns processes depending on the behavior of primary actors, the experimenter is a secondary actor. There also is then a potential conflict. Being interested in predictable effects of treatments, the experimenter has reason to control and regulate the behavior of the primary actors as far as possible.

4. Randomly assigned causal conditions. There is a further kind of experiment in which also values of the treatment variable, $\ddot{X}$, are randomly generated. The functional model then looks as follows:

Model 3.2.3



The set-up entails that, conditional on $\ddot{E} = 1$ (initiation of the experiment), $\dot{X}$ and $\dot{Z}$ are independent. In an often used notation, this can be written as

$$\dot{X} \perp\!\!\!\perp \dot{Z} \,|\, \ddot{E} = 1$$

This allows one to express effects of $\dot{X}$ as follows:

$$\Delta^s(\dot{Y}; \dot{X}[x', x''], \ddot{E}=1) = \sum\nolimits_z \Delta^s(\dot{Y}; \dot{X}[x', x''], \dot{Z}=z, \ddot{E}=1) \Pr(\dot{Z}=z|\ddot{E}=1) \tag{26}$$

This shows that a randomized experiment allows one to compare effects of different treatments (values of $\dot{X}$) in a balanced way: the distribution of further possibly relevant variables, $\dot{Z}$, does not depend on the values of $\dot{X}$, the treatments, that one intends to compare. But note that this does not entail that effects of $\dot{X}$ are independent of $\dot{Z}$. If $\dot{X}$ and $\dot{Z}$ interact in the generation of values of $\dot{Y}$, the effect defined in (26) still depends on the distribution of $\dot{Z}$.

Notice that a reference to the variable $\ddot{E}$ is required in order to think of an experimental context which includes an experimenter who at least initiates the experiment. Without the variable $\ddot{E}$, Model 3.2.3 only contains random variables getting their values from processes not represented in the model. Also note that this is an intervention model in a very specific sense: randomization is used as a fact-generating process, that is, a process generating events (values of $\dot{X}$).

5. Two contexts for randomization. Is randomization useful in social research? In thinking about this question one should distinguish between two different contexts for randomization.

First, one can think of data-generating processes. The most relevant application concerns the selection of units (for further observation). There are good arguments that samples should be generated randomly, that is, according to in some way fixed and known selection probabilities.

An essentially different context is experimentation. Performing an experiment requires, first of all, a fact-generating process. In this context, randomization not only, if at all, concerns the selection of a sample of units for the experiment; but also is a method of creating facts ('treatments') whose possible effects one intends to study.

While randomization in data generation is certainly useful in social research, the value of randomized experiments seems questionable. The argument is not that such experiments are seldom possible. The relevant point is that randomization would change, in an essential way, the processes to be studied.

The example depicted in Model 3.1.1 can show this. In this example, one can be interested in effects of the learning frames ($\dot{X}$). The model realistically assumes that the selection of learning frames depends on the student's family background (represented by $\ddot{Z}$). This is a fact-generating process. Randomization would substitute this by another fact-generating process that randomly assigns students to learning frames.

3.3 Rule-based and descriptive approaches

In the statistical literature, discussions of causal effects often use notions of 'potential outcomes' (e.g. Rubin 2005; Morgan and Winship 2007; Angrist and Pischke 2009; Gangl 2010). Different understandings are possible. In this section I distinguish, and contrast, rule-based and descriptive understandings of potential outcomes.

1. Rule-based understanding of potential outcomes. The approach to understanding causal relationships that was sketched in Section 3.1 uses functional models as a formal framework. This is appropriate when one is interested in causally interpretable rules. Such rules – I briefly speak of *causal rules* – concern potential outcomes which can be linked to different values of causally relevant variables. Continuing with the notation introduced in Section 3.1, such a rule has the form

$$(x, z) \longrightarrow \mathrm{E}(\dot{Y}|\dot{X}=x, \dot{Z}=z) \tag{27}$$

It is a causal rule if $\dot{X}$ and $\dot{Z}$ can be interpreted as variables which are causally relevant for the generation of values of $\dot{Y}$. The dependent variable, $\dot{Y}$, represents *potential outcomes* which can be expected in a generic situation where $\dot{X}$ and $\dot{Z}$ have specified values. Given the rule, one can compare potential outcomes for different values of $\dot{X}$ and $\dot{Z}$, and, as defined in (24) and (25), interpret their differences as causal effects.

2. Descriptive notion of potential outcomes. In the statistical literature, one also finds a descriptive notion of potential outcomes which is not based on rules. This notion relates to a specified set of particular units, say Ω, and presupposes three (or more) statistical variables.[4] A variable, say $X : \Omega \longrightarrow \mathcal{X}$, represents causally relevant factors (conditions or events). As is often done in the literature, in order to simplify notations, I assume that X is a binary variable:

$$X = \begin{cases} 1 & \text{if a specified causal factor is present} \\ 0 & \text{otherwise} \end{cases}$$

The interest concerns outcomes in situations where $X = 1$ or $X = 0$. It is assumed that these

[4] See, for example, Holland (1986). Similarly, Rubin (2005: 323) refers to an array of given, partly observed values.

outcomes can be represented by statistical variables

$$Y_j : \Omega \longrightarrow \mathcal{Y} \tag{28}$$

($j = 0$ or 1) having the following meaning: If $X(\omega) = j$, the outcome of interest has the value $Y_j(\omega)$. One can then formally define, for each unit $\omega \in \Omega$, a causal effect $Y_1(\omega) - Y_0(\omega)$. Of course, these individual causal effects cannot be observed. One therefore aims to estimate an *average causal effect* which can be defined for Ω by

$$\mathrm{M}(Y_1) - \mathrm{M}(Y_0) \tag{29}$$

However, observations only allow estimation of conditional mean values, $\mathrm{M}(Y_j|X = j)$. So the question arises under which conditions one can think of these conditional mean values as unbiased estimates of $\mathrm{M}(Y_j)$. A sufficient condition would be that X and Y_j are approximately independent;[5] formally: $Y_j \perp\!\!\!\perp X$ (for $j = 0, 1$).

This independence condition suggests that a critical question concerns the generation of values of X. If possible, such values should be randomly assigned to the members of Ω. This would justify to consider $\Omega_j := \{\omega \in \Omega | X(\omega) = j\}$ as a simple random sample from Ω, and therefore to assume that Y_j has approximately the same distribution in Ω_j and Ω.

3. Including further causally relevant variables. As described in the previous paragraph, the descriptive approach aims to define a causal effect that can be attributed to a single variable, X. A somewhat extended formulation is required if effects also depend on further variables. Assume that outcomes depend not only on X, but also on values of a variable Z (possibly consisting of several components). Instead of (29), one has to consider the effect definition

$$\mathrm{M}(Y_1|Z = z) - \mathrm{M}(Y_0|Z = z) \tag{30}$$

As before, values of Y_j can only be observed if $X = j$; the observable conditional mean values are $\mathrm{M}(Y_j|X = j, Z = z)$. They provide unbiased estimates of $\mathrm{M}(Y_j|Z = z)$ if

$$Y_j \perp\!\!\!\perp X \,|\, Z = z \quad (\text{for } j = 0, 1) \tag{31}$$

This shows that it would suffice to perform the randomization (the random assignment of values of X to the members of Ω) separately for each value of Z.

4. Balanced effects and kinds of models. Neither the rule-based nor the descriptive approach require that the explicitly represented variables, X and Z, are independent. Relationships between these variables become important, however, if one aims to define causal effects of just one variable, say X. Whether this is possible depends first of all on whether X and Z interact in the generation of outcomes.[6] If they interact, effects cannot be attributed solely to X. It is nevertheless possible to define average effects. Following the rule-based approach, one can use the definition

$$\sum\nolimits_z \left[\mathrm{E}(\dot{Y}|\dot{X} = x'', \dot{Z} = z) - \mathrm{E}(\dot{Y}|\dot{X} = x', \dot{Z} = z)\right] \Pr(\dot{Z} = z) \tag{32}$$

where $\Pr(\dot{Z} = z)$ refers to an (arbitrarily) specified distribution of $\dot{Z}$. This is a *balanced effect*, meaning that the distribution of $\dot{Z}$ is identical for x' and x''. However, if $\dot{X}$ and $\dot{Z}$ interact, the effect still depends on the assumed distribution of $\dot{Z}$.

[5] Note that X and Y_j are statistical variables, defined for a finite reference set Ω. One can therefore think of statistical independence only in an approximate sense.

[6] As was discussed in Chapter 2, this must be distinguished from dependency relations which concern the joint distribution of X and Z (or $\dot{X}$ and $\dot{Z}$).

How to proceed when following the descriptive approach depends on the given data to which the causal statements relate. If the data result from a process which entailed a randomization of X w.r.t. Z, the distribution of Z already is approximately independent of X, and one can interpret (29) as a balanced average effect of X. Again, if X and Z interact, this effect also depends on the distribution of Z in the reference set of units.

If X and Z are not independent, one can construct a balanced effect. This is analogous to the procedure in the rule-based approach. In the descriptive approach, one starts from (30) and (arbitrarily) specifies a distribution of Z. Formally analogous to (32), an average effect can then be defined by

$$\sum\nolimits_z \big[\mathrm{M}(Y_1|Z=z) - \mathrm{M}(Y_0|Z=z)\big]\,\mathrm{P}(Z=z) \tag{33}$$

One might ask whether balanced effects are particularly useful. This depends on the kind of model. With treatment models, one is normally interested in finding an effect that can be attributed solely to the treatment, given that all other possibly relevant conditions are in some sense fixed. This interest suggests to construct balanced effects.

The situation is different with explanatory models. In social research, explanatory models most often relate to situations where at least some of the causally relevant conditions are generated by actions of primary agents. Effects of single variables are then never balanced w.r.t. all causally relevant conditions. It would be possible, of course, to construct balanced effects w.r.t. observed variables; but I believe that a primary interest concerns how the real effects, which are unbalanced, come into being.

5. Contrasting the two approaches. The rule-based and the descriptive approach to the definition of causal effects are in several respects different.

(1) A first difference concerns the notion of potential outcomes. As mentioned in **3.3.1**, the rule-based approach conforms to the understanding that potential outcomes are outcomes which, under specified conditions, possibly will come into existence. Correspondingly, potential outcomes are defined by a rule (a linguistic if-then construction).

The descriptive approach, in contrast, presupposes that potential outcomes (= values of Y_0 and Y_1) already exist before values of X, and all other causally relevant variables, are fixed. To speak of 'potential outcomes' is therefore somewhat misleading. Actually, what is potentially realized is an observation of a hypothetically presupposed fact (value of Y_j).[7] So it would be less confusing to speak of 'potential observations'.

(2) It might be helpful to remember the distinction between fact-generating and data-generating processes. The rule-based approach aims to formulate causal rules for fact-generating processes. The descriptive approach, as it is theoretically formulated, is concerned with data-generating processes which provide partial information about hypothetically presupposed facts (values of Y_0 and Y_1). This approach therefore seems to allow one to think of 'causal inference' in parallel to a missing observation problem.[8]

(3) It is important to understand that the variables Y_0 and Y_1 can only be defined by referring to a set of existing units. For each particular unit, say $\omega \in \Omega$, one can posit values, $Y_0(\omega)$ and $Y_1(\omega)$, representing the outcomes corresponding to $X(\omega)=0$ and $X(\omega)=1$, respectively. This is possible because, and insofar, one can assume that all further conditions on which outcomes depend are implicitly fixed by the reference to ω, a particular unit existing in particular circumstances.[9] The

[7] This is seldom explicitly mentioned; but see Greenland (2004: 4).

[8] See, e.g., Rubin et al. (2004: 105), Winship and Morgan (1999: 664).

[9] Positing values of Y_0 and Y_1 can be done in a deterministic or in a probabilistic way. This entails different

descriptive approach is therefore essentially static and not well suited for causal interpretations of temporally extended processes.

The rule-based approach, in contrast, is not based on a reference to a set of already existing units, but relates to generic units which are only defined by values of variables. It is therefore not possible to define variables corresponding to Y_0 and Y_1. Instead, there is a single outcome variable, $\dot{Y}$, having possible values which only become realized when, and after, $\dot{X}$, and any further variables which define the generic unit, have taken specific values. There are no restrictions for thinking of a temporally extended process connecting $\dot{X}$ and the final outcome, $\dot{Y}$.

(4) As a consequence, the independence requirement (31) cannot be formulated in the conceptual framework of the rule-based approach. Of course, based on the mentioned understanding of $\dot{Y}$, one can define variables $\dot{Y}_j$ having distributions defined by $\Pr[\dot{Y}_j|\dot{Z}=z] = \Pr[\dot{Y}|\dot{X}=j,\dot{Z}=z]$. An independence condition paralleling (31) is then trivially true:

$$\dot{Y}_j \perp\!\!\!\perp \dot{X} \,|\, \dot{X}=j, \dot{Z}=z \quad (\text{for } j=0,1) \tag{34}$$

But this condition has not the same interpretation. (31) can be interpreted as the requirement that X, conditional on values of Z, is approximately independent of all further circumstances which are fixed by the implicit reference to particular units. This is not entailed by (34).

(5) As mentioned, in order to satisfy the independence condition (31), there ideally should be a randomized assignment of values of X to the members of Ω (conditional on values of Z). (34) does not require any randomization procedure. However, the important point is that formulating a causal rule like (27) does not entail the claim that there are no further variables on which the outcome variable depends. Such variables are simply not taken into account. Consequently, also definitions of causal effects which are derived from (27) do not entail anything about further variables on which the outcome variable depends. Consider the causal effect defined in (24). This definition compares two generic units, one with $\dot{X}=x''$ and the other one with $\dot{X}=x'$. Both units have identical values of $\dot{Z}$; but they can differ in all other respects.

6. Omitted causally relevant conditions. If a causal rule does not relate to an artificial random generator, one can almost always think that the rule misses one or more causally relevant conditions. Note that this is true even if the data used to estimate the rule result from a randomized experiment. The point simply is that there probably are causally relevant conditions not explicitly referred to in the rule's formulation. It is therefore not reasonable to require that a causal rule entails the claim that one has taken into account all causally relevant conditions.

Moreover, except when dealing with artificial random generators, already the assumption that one can 'theoretically' refer to a complete set of variables which are causally relevant for an outcome variable seems obscure. The descriptive approach to potential outcomes avoids this assumption and instead requires the conditional independence (31). This independence is viewed as a precondition for thinking of a causal effect of X. However, as already mentioned, (31) cannot be formulated in a rule-based approach. In a rule-based approach one would need to refer to explicitly defined variables which, in addition to Z, are causally relevant for Y. If one could refer to a list of such variables, say $(U_1, U_2, \ldots)$, one could use the formulation

$$X \perp\!\!\!\perp (U_1, U_2, \ldots) \,|\, Z=z \tag{35}$$

However, the formulation is not useful because one cannot define, not even clearly think of, such a list of variables.

understandings of individual causal effects, but the essential features of the descriptive approach are independent of this distinction.

Of course, it is often quite possible to think of a particular variable, say U, which is left out in the formulation of a causal rule, but should be taken into account in order to get a better understanding of the causal mechanism. The original model, that was used to derive the causal rule, must then be enlarged by incorporating U; and this also demands to specify U's relationship with the other variables in the model. How to do this depends on the intended use of the model. If the model is intended to represent a randomized experiment, one can assume in the formulation of the enlarged model that $\dot{X} \perp\!\!\!\perp \dot{U} \mid \dot{Z} = z$.

However, as I have argued in **3.2.5**, in social research an explanatory model can almost never be formulated as a model representing a randomized experiment. It then depends on the details of the model how to think of $\dot{U}$'s role in the mechanism generating values of the outcome variable. In any case, one would need observations of $\dot{U}$ in order to quantify its causal role.

References

Angrist, J. D., Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton: Princeton University Press.

Baron, R. M., Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology* 51, 1173–82.

Blossfeld, H.-P. (2009). Causation as a Generative Process. In: H. Engelhardt, H.-P. Kohler, A. Prskawetz (eds.), *Causal Analysis in Population Studies*, 83–109. Berlin: Springer.

Cox, D. R. (1992). Causality: Some Statistical Aspects. *Journal of the Royal Statistical Society* A 155, 291–301.

Cross, C. B. (1991). Explanation and the Theory of Questions. *Erkenntnis* 34, 237–60.

Faye, J. (1999). Explanation Explained. *Synthese* 120, 61–75.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London.* Series A, Vol. 222, 309–368.

Gangl, M. (2010). Causal Inference in Sociological Research. *Annual Review of Sociology* 36, 21–47.

Goldthorpe, J. H. (2001). Causation, Statistics, and Sociology. *European Sociological Review* 17, 1–20.

Greenland, S. (2004). An Overview of Methods for Causal Inference from Observational Studies. In: A. Gelman, X.-L. Meng (eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 3–13. New York: Wiley.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* 81, 945–968.

Jackson, M., Erikson, R., Goldthorpe, J. H., Yaish, M. (2007). Primary and Secondary Effects in Class Differentials in Educational Attainment. *Acta Sociologica* 50, 211–229.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. New York: Lawrence Erlbaum.

Mahoney, J. (2001). Beyond Correlational Analysis: Recent Innovations in Theory and Method. *Sociological Forum* 16, 575–593.

Morgan, S. L., Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles*. Cambridge: Cambridge University Press.

Pearl, J. (2001): Direct and Indirect Effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420. San Francisco, CA.

Petersen, M. L., Sinisi, S. E., Van der Laan, M. J. (2006). Estimation of Direct Causal Effects. *Epidemiology* 17, 276–284.

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

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association* 100, 322–331.

Rubin, D. B., Stuart, E. A., Zanutto, E. L. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics* 29, 103–116.

Weinberg, C. R. (1993). Toward a Clearer Definition of Confounding. *American Journal of Epidemiology* 137, 1–8.

Winship, C., Morgan, S. L. (1999). The Estimation of Causal Effects from Observational Data. *Annual Review of Sociology* 25, 659–706.