Causal Regression Models II: Unconfoundedness and Causal Unbiasedness

By Rolf Steyer, Siegfried Gabler, Alina A. von Davier, and Christof Nachtigall

Abstract

We consider regression models with discrete units and a discrete treatment variable. In this framework, individual and average causal effects as well as causal unbiasedness of conditional expected values $E(Y|X=x)$ and of their differences were defined in a previous paper where it was also noted that a hypothesis of causal unbiasedness is not empirically testable outside the randomized experiment. Therefore, we study a stronger causality criterion which we call “unconfoundedness”. To our knowledge, this is the weakest empirically testable condition implying causal unbiasedness of the conditional expected values $E(Y|X=x)$. Unconfoundedness holds in randomized experiments, but it may hold in nonrandomized experiments, as well. We derive theorems about sufficient and necessary conditions, about sufficient conditions, and about necessary conditions for unconfoundedness. The latter identify the hypotheses to be tested in nonrandomized experiments when it comes to testing the weakest empirically testable sufficient condition for conditional expected values $E(Y|X=x)$ to be causally unbiased.

Keywords: Causality; Confounding; Regression Models; Randomization; Rubin’s Approach to Causality

1 We would like to thank Albrecht Iseler (Free University Berlin) for a discussion in which the idea for some parts of this paper has been developed. We would also like to extend our gratitude to Angelika von der Linde (University of Edinburgh), to Donald B. Rubin (Harvard University), and to Peter Kischka (FSU Jena) who all gave helpful critical comments on previous versions of this paper.
2 Address for correspondence: Prof. Dr. Rolf Steyer, Friedrich-Schiller-Universität Jena, Am Steiger 3 – Haus 1, D-07743 Jena, Germany. Email: rolf.steyer@uni-jena.de
3 Center for Surveys, Methods, and Analyses (ZUMA), Mannheim, Germany
4 Friedrich-Schiller-University Jena, Germany
In a previous paper (Steyer, Gabler, von Davier, Nachtigall, & Buhl, 2000) we reformulated the theory of individual and average causal effects in terms of classical probability theory and illustrated it by some examples. This theory goes back to Neyman (1923/1990; 1935) and has been adopted and enriched by Rubin (1973a, b, 1974, 1977, 1978, 1985, 1986, 1990), Holland and Rubin (1983), Holland (1986, 1988a, b), Rosenbaum and Rubin (1983a, b, 1984, 1985a, b), Rosenbaum (1984a, b, c), and Sobel (1994, 1995), for instance. More specifically, we introduced a probabilistic framework: a potential causal regression model with discrete units and discrete treatment variable, which consists of three components:

- a probability space \((\Omega, \mathcal{A}, P)\), which represents the random experiment, i.e., the empirical phenomenon considered,
- a regression \(E(Y|X)\), the potential causal interpretation of which is focused, where \(X\) represents a discrete (but not necessarily univariate) treatment variable and \(Y\) a real-valued (and not necessarily continuous) response variable,
- a nonnumeric random variable \(U\), the value of which is the observational unit drawn.

The probability space is chosen such that \(X, Y,\) and \(U\) are random variables on \((\Omega, \mathcal{A}, P)\), i.e., \(X, Y,\) and \(U\) have a joint distribution. In this framework and notation, we presented the basic concepts of the theory of individual and average causal effects: the individual conditional expected values \(E(Y|X = x, U = u)\), the individual causal effects \(ICE_u(i, j) := E(Y|X = x_i, U = u) - E(Y|X = x_j, U = u)\) of a treatment \(x_i\) vs. a treatment \(x_j\), the causally unbiased expected value [denoted \(CUE(Y|X = x)\)], and the average causal effect \(ACE(i, j)\) (see Figure 1 of Steyer et al., 2000). We also showed how these concepts are related to the conditional expected values \(E(Y|X = x)\) and the prima facie effects \(PFE(i, j) = E(Y|X = x_i) - E(Y|X = x_j)\) that are usually focused in statistical estimation and hypothesis testing. Specifically, it has been shown that both, (a) stochastic independence of \(U\) and \(X\) (that can be created via random assignment of units to treatment conditions) as well as (b) unit-treatment homogeneity, [i.e., \(E(Y|X, U = E(Y|X)]\) imply the equations \(E(Y|X = x) = CUE(Y|X = x)\), and \(PFE(i, j) = ACE(i, j)\) [i.e., causal unbiasedness of \(E(Y|X = x)\) and of \(PFE(i, j)\)].

We argued that the theory of individual and average causal effects reviewed above has some major limitations. The first one is that a proposition about causal unbiasedness is not empirically falsifiable: Postulating that the prima facie effect \(PFE(i, j)\) is equal to the average causal effect \(ACE(i, j)\) or that the conditional expected values \(E(Y|X = x)\) are equal to the causally unbiased expected values \(CUE(Y|X = x)\) of \(Y\) given \(x\) does not imply anything one could show to be wrong in an empirical appli-

---

5 For a more general formal framework, see Steyer (1992).
cation. The reason is that the computation of the $ACE(i, j)$ and the $CUE(Y | X = x)$ involve all individual conditional expected values $E(Y | X = x, U = u)$ in the population and that it is not possible to estimate all of them. (See the fundamental problem of causal inference in Holland, 1986 or Steyer et al., 2000).

The second limitation is that even if both $PFE(i, j) = ACE(i, j)$ and $E(Y | X = x) = CUE(Y | X = x)$ hold in the total population, the corresponding equations may not hold in the subpopulations, for instance, in the subpopulations of males and females. Hence, causal unbiasedness of the prima facie effects in the subpopulations might not hold although being correct in the total population.

Because of the nonfalsifiability we argued that, although the theory of individual and average causal effects provides indispensable concepts, it is not really complete as a methodological basis for causal modeling outside the randomized experiment in which treatment assignment may not be random. Completing this methodological basis is the central goal of this paper. Specifically, we aim at enriching the theory of individual and average causal effects by introducing the concept of unconfoundedness of a regression $E(Y | X)$. Unconfoundedness will be defined such that:

(a) a proposition that a regression $E(Y | X)$ is unconfounded is empirically falsifiable,

(b) unconfoundedness of $E(Y | X)$ implies that the regressions $E_{W = w}(Y | X)$ in the subpopulations, are causally unbiased as well.\(^6\)

Among all criteria fulfilling these requirements we prefer that criterion which is logically the weakest, i.e., which is implied by the others, but does not imply one of the others.

If requirement (a) were not fulfilled, everybody could make causal propositions, nobody would have a chance to falsify it. Requirement (b) means that causal interpretations such as causal unbiasedness should be transferable from the total population into its subpopulations and in this sense be “stable” or “invariant”.

Hence, the goal of this paper is to introduce a causality criterion that fulfills the requirements (a) and (b) mentioned above, compare it to some other criteria, illustrate it by an example and study its sufficient and necessary conditions. More specifically, we will present and discuss different causality criteria in section 1, including strong ignorability and unconfoundedness. The latter is, according to requirements (a) and (b), the most favorable one. In section 2 we will illustrate the different criteria by an example. Section 3 is devoted to a necessary and sufficient condition of unconfoundedness. Next we will study sufficient conditions of unconfoundedness (section 4), and then sufficient conditions of strong ignorability (section 5). Then we turn to the necessary conditions of

---

\(^6\) Each value $w$ of $W$ represents a subpopulation.
unconfoundedness (section 6). One of these necessary conditions is causal unbiasedness of \(E(Y|X)\). Section 7 focusses two general procedures to falsify unconfoundedness, which will be illustrated by an example in section 8. Section 9 deals with necessary conditions for unconfoundedness under special assumptions such as linearity and \(W\)-conditional linearity of \(E(Y|X, W)\), where \(W\) is a variable each value of which represents a subpopulation. Finally, the merits and limits of the theory presented are discussed in section 10.

1. Some Causality Criteria

The weakest criterion we consider is causal unbiasedness of the treatment regression \(E(Y|X)\) in the total population, i.e.,

\[
E(Y|X = x) = CUE(Y|X = x), \quad \text{for each value } x \text{ of } X, \quad (1)
\]

with

\[
CUE(Y|X = x) := \sum_u (Y|X = x, U = u) P(U = u), \quad (2)
\]

where the summation is across all observational units \(u\) in the (total) population \(\Omega_U\) (i.e., the set of all units considered). This term has been defined to be the causally unbiased expected value of \(Y\) given \(x\) (Steyer et al., 2000). Equation (1) means that each \(E(Y|X = x)\) is equal to the expected value of the individual expected values \(E(Y|X = x, U = u)\) across the distribution of the observational units. Equation (1) is sufficient to imply \(PFE(i, j) = ACE(i, j)\), i.e., causal unbiasedness of the prima facie effects in the total population (see Steyer et al., 2000). The equation \(E(Y|X = x) = CUE(Y|X = x)\), defines causal unbiasedness of a conditional expected value \(E(Y|X = x)\).

From the perspective of the theory of individual and average causal effects, there is no doubt that Equation (1), and with it, the concept of causal unbiasedness, is a desirable and indispensable property of the conditional expected values \(E(Y|X = x)\) in meaningful substantive applications. However, as mentioned before, Equation (1) neither implies anything that could be falsified in an empirical application [see requirement (a)] nor does it imply unbiasedness of the regressions \(E_{W=w}(Y|X)\) in the subpopulations represented by \(W = w\) [see requirement (b)]. Hence, Equation (1) [causal unbiasedness of the regression \(E(Y|X)\)] is too weak to qualify as a satisfactory causality criterion applicable to experimental studies with nonrandom assignment.\(^7\) 

\^ In a similar vein Pearl (1998) argues that unbiasedness might be incidental which necessitates the need for “stable unbiasedness” (see also Example III in Steyer et al., 2000).
gument applies, of course, to the criterion \( PFE(i, j) = ACE(i, j) \), causal unbiasedness of the prima facie effects. Both criteria lack empirical falsifiability.

As a second criterion we consider “strong ignorability” described, e.g., by Rosenbaum and Rubin (1983a). They defined potential response variables \( Y_i: \Omega \rightarrow \mathbb{R} \) for each treatment condition \( x_i \). In our notation we could define these variables for each value \( x_i \) by \( Y_i(\omega) = f(u) = E[Y|X = x_i, U = u] \), for all values \( u \) of \( U \). Defined in this way, it is obvious that the variables \( Y_i, i = 1, \ldots, n_\omega \), are functions of \( U \) and therefore have a joint distribution with \( X \), because \( U \) and \( X \) have joint distributions. Strong ignorability implies causal unbiasedness of the conditional expected values and the prima facie effects. However, it is neither empirically testable, and it is unknown whether or not it implies causal unbiasedness in the subpopulations. (See Note 1 in Appendix A).

Let us now discuss several other sufficient conditions for causal unbiasedness that are empirically falsifiable. A third criterion we might consider is stochastic independence of \( X \) and \( U \). This criterion, in fact, fulfills both requirements discussed above. It is empirically falsifiable and it implies causal unbiasedness of the conditional expected values \( E(Y|X = x, W = w) \) in the subpopulations represented by \( W = w \) (see Th. 3 and Th. 6). However, Theorem 2 of Steyer et al. (2000) shows that there is a fourth criterion that is also falsifiable and implies causal unbiasedness of the conditional expected values \( E(Y|X = x, W = w) \): unit-treatment homogeneity, i.e., \( E(Y|X, U) = E(Y|X) \). Hence, independence of \( U \) and \( X \) would be unnecessarily strong.

The fifth criterion to be considered is “\( X \) and \( U \) are independent or \( E(Y|X, U) = E(Y|X) \)”.

This would also fulfill both requirements discussed above. In fact, this criterion is already very close to our favorite one which is somewhat weaker and still fulfills the two requirements (a) and (b).

The sixth criterion will be called unconfoundedness of the regression \( E(Y|X) \). It is defined as follows. Note that throughout the paper we will presume that \( \langle (\Omega, \mathcal{A}, P), \ E(Y|X), U \rangle \) is a potential causal regression model.

---

8. In Rubin’s notation, the causally unbiased expected values would be denoted \( E(Y_i) \), i.e., \( E(Y_i) = CUE(Y_i|X = x) \).

9. For simplicity, we restrict our discussion to the case where there is no concomitant variable or covariate. Rosenbaum and Rubin consider the case with a covariate. Also note that strong ignorability implies \( E(Y_i) = E(Y_i|X = x) \) for each pair \((i, x)\) of indices \( i \in \{1, \ldots, n_\omega \} \) and values \( x \) of \( X \).

10. This criterion is very close to what has been proposed by Pearl (1998) in his definitions 2 and 3.

11. In previous papers (see e.g. Steyer, Gabler, & Rucai, 1996 and Steyer, von Davier, Gabler, & Schuster, 1997), we have defined unconfoundedness in a different but equivalent way (see Th. 2).
Definition 1. \( E(Y \mid X) \) is called unconfounded if for each value \( x \) of \( X \):

\[
P(X = x \mid U = u) = P(X = x) \quad \text{for each value } u \text{ of } U
\]

or

\[
E(Y \mid X = x, U = u) = E(Y \mid X = x) \quad \text{for each value } u \text{ of } U.
\]

This sixth criterion differs from the fifth one by the fact that “Equation (3) or Equation (4)” is postulated within each value \( x \) of \( X \). (For an example, see Table 2.) Hence, within each treatment condition we require equal treatment assignment probabilities for each unit \( u \) [see Eq. (3)] or homogeneity of the units. Only one of these equations has to be true within each treatment condition \( x \). In contrast, the fifth criterion requires Equation (3) to be true for all treatment conditions \( x \) or Equation (4) to be true for all treatment conditions \( x \). Whereas from a substantive point of view this difference seems negligible, it is important from a mathematical point of view: the two criteria are not equivalent to each other. In fact, the fifth criterion implies the sixth criterion, but not vice versa. Table 1 displays a summary of the causality criteria discussed above.

Table 1. Summary of the causality criteria discussed in the paper

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Causal unbiasedness of ( E(Y \mid X) )</td>
<td>( E(Y \mid X = x) := \sum_u E(Y \mid X = x, U = u) P(U = u) ), for each value ( x ) of ( X )</td>
</tr>
<tr>
<td>2. Strong ignorability</td>
<td>( Y_1, \ldots, Y_n \perp X ) and ( 0 &lt; P(X = x_i) &lt; 1 )</td>
</tr>
<tr>
<td>3. Independence of ( X ) and ( U )</td>
<td>( X \perp U )</td>
</tr>
<tr>
<td>4. Unit-treatment homogeneity</td>
<td>( E(Y \mid X, U) = E(Y \mid X) )</td>
</tr>
<tr>
<td>5. Independence of ( X ) and ( U ) or unit-treatment homogeneity</td>
<td>( X \perp U ) or ( E(Y \mid X, U) = E(Y \mid X) )</td>
</tr>
</tbody>
</table>
| 6. Unconfoundedness of \( E(Y \mid X) \) | For each value \( x \) of \( X 
\[
P(X = x \mid U = u) = P(X = x) \quad \text{for each value } u \text{ of } U
\]
or

\[
E(Y \mid X = x, U = u) = E(Y \mid X = x) \quad \text{for each value } u \text{ of } U.
\]

2. An Example

Table 2 gives an example illustrating the criteria discussed above. The example is constructed in such a way that unconfoundedness and causal unbiasedness but none of the other causality criteria hold. We consider three treatment conditions \( x_1 \) to \( x_3 \) and a
population of eight units. We assume that each unit has the probability \( P(U = u) = 1/8 \) to be sampled. Let us first look at treatment \( x_1 \). Column 4 displays the individual assignment probabilities for treatment \( x_1 \). They are all the same, namely \( 1/2 \), for each individual unit. For this specific treatment condition \( x_1 \) we have different individual conditional expected values \( E(Y \mid X = x_1, U = u) \) (see column 5). For treatments \( x_2 \) and \( x_3 \) things are different. Column 6 contains different treatment assignment probabilities \( P(X = x_2 \mid U = u) \) for treatment condition \( x_2 \), and column 7 displays equal individual conditional expected values \( E(Y \mid X = x_2, U = u) \). The same is true for treatment condition \( x_3 \). Again, we have different treatment assignment probabilities \( P(X = x_3 \mid U = u) \) but equal individual conditional expected values \( E(Y \mid X = x_3, U = u) \).

In order to check causal unbiasedness, we may first compute the three conditional expected values \( E(Y \mid X = x_1) = 115 \), \( E(Y \mid X = x_2) = 105 \), and \( E(Y \mid X = x_3) = 110 \) from Equation

\[
E(Y \mid X = x) := \sum_u E(Y \mid X = x, U = u) \cdot P(U = u \mid X = x), \tag{5}
\]

which always holds for a discrete random variable \( U \). Using

\[
P(U = u \mid X = x) = P(X = x \mid U = u) \cdot P(U = u) / P(X = x)
\]

yields \( P(U = u \mid X = x_1) = (1/2) \cdot (1/8) / (1/2) = 1/8 \) for each unit \( u \) and

\[
E(Y \mid X = x_1) = 82 \cdot 1/8 + \ldots + 152 \cdot 1/8 = 115.
\]

Analogously, we can compute the conditional expected value \( E(Y \mid X = x_2) = 105 \cdot (1/20 + 1/20 + 2/20 + 2/20 + 3/20 + 3/20 + 4/20 + 4/20) = 105 \).
Table 2. An example in which the treatment regression $E(Y \mid X)$ is unconfounded and causally unbiased but none of the other causality criteria discussed in this paper hold.

<table>
<thead>
<tr>
<th>Observational-unit variable $U$</th>
<th>$P(U = u)$</th>
<th>$W$ (gender)</th>
<th>$P(X = x_1 \mid U = u)$</th>
<th>$E(Y \mid X = x_1, U = u)$</th>
<th>$P(X = x_2 \mid U = u)$</th>
<th>$E(Y \mid X = x_2, U = u)$</th>
<th>$P(X = x_3 \mid U = u)$</th>
<th>$E(Y \mid X = x_3, U = u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>1/8</td>
<td>$m$</td>
<td>1/2</td>
<td>82</td>
<td>1/10</td>
<td>105</td>
<td>4/10</td>
<td>110</td>
</tr>
<tr>
<td>$u_2$</td>
<td>1/8</td>
<td>$m$</td>
<td>1/2</td>
<td>89</td>
<td>1/10</td>
<td>105</td>
<td>4/10</td>
<td>110</td>
</tr>
<tr>
<td>$u_3$</td>
<td>1/8</td>
<td>$m$</td>
<td>1/2</td>
<td>101</td>
<td>2/10</td>
<td>105</td>
<td>3/10</td>
<td>110</td>
</tr>
<tr>
<td>$u_4$</td>
<td>1/8</td>
<td>$m$</td>
<td>1/2</td>
<td>108</td>
<td>2/10</td>
<td>105</td>
<td>3/10</td>
<td>110</td>
</tr>
<tr>
<td>$u_5$</td>
<td>1/8</td>
<td>$f$</td>
<td>1/2</td>
<td>118</td>
<td>3/10</td>
<td>105</td>
<td>2/10</td>
<td>110</td>
</tr>
<tr>
<td>$u_6$</td>
<td>1/8</td>
<td>$f$</td>
<td>1/2</td>
<td>131</td>
<td>3/10</td>
<td>105</td>
<td>2/10</td>
<td>110</td>
</tr>
<tr>
<td>$u_7$</td>
<td>1/8</td>
<td>$f$</td>
<td>1/2</td>
<td>139</td>
<td>4/10</td>
<td>105</td>
<td>1/10</td>
<td>110</td>
</tr>
<tr>
<td>$u_8$</td>
<td>1/8</td>
<td>$f$</td>
<td>1/2</td>
<td>152</td>
<td>4/10</td>
<td>105</td>
<td>1/10</td>
<td>110</td>
</tr>
</tbody>
</table>

Note: The (unconditional) probabilities for the three treatments are $P(X = x_1) = 1/2$, $P(X = x_2) = P(X = x_3) = 1/4$.

In order to check causal unbiasedness, the conditional expected values computed above have to be compared to the causally unbiased conditional expected values, which may be computed by Equation (2). For the treatment condition $x_1$, this equation yields

$$CUE(Y \mid X = x_1) := \sum_u E(Y \mid X = x_1, U = u) \cdot P(U = u)$$

$$= (82 + 89 + 101 + 108 + 118 + 131 + 139 + 152) \cdot 1/8 = 115,$$

for treatment condition $x_2$

$$CUE(Y \mid X = x_2) := \sum_u E(Y \mid X = x_2, U = u) \cdot P(U = u) = (105 + \ldots + 105) \cdot 1/8 = 105,$$

and for treatment condition $x_3$

$$CUE(Y \mid X = x_3) := \sum_u E(Y \mid X = x_3, U = u) \cdot P(U = u) = (110 + \ldots + 110) \cdot 1/8 = 110.$$

Hence, in this example, the conditional expected values are unbiased, i.e., $E(Y \mid X = x) = CUE(Y \mid X = x)$ for each value $x$ of $X$, and the prima facie effects $PFE(i, j) = E(Y \mid X = x_i) - E(Y \mid X = x_j)$ are unbiased as well, i.e., $PFE(i, j) = ACE(i, j)$, for each pair of treatment conditions $x_i$ and $x_j$. 
We now check the other causality criteria discussed in the last section. The second causality criterion, strong ignorability, does not hold in this example. The independence condition \( Y_1, \ldots, y_n \perp X \) implies the conditional expected values \( E(Y_1 \mid X = x) \) and \( E(Y_1 \mid X = x_2) \) to be identical. However, in our example:

\[
E(Y_1 \mid X = x_1) = \sum_u Y_1(u) P(Y_1 = Y_1(u) \mid X = x_1) = \sum_u Y_1(u) P(U = u \mid X = x_1) = \sum_u \left[ Y_1(u) \frac{P(X = x_1 \mid U = u) \cdot P(U = u)}{P(X = x_i)} \right]
\]

\[
= 82 \cdot 1/8 + \ldots + 152 \cdot 1/8 = 115
\]

and

\[
E(Y_1 \mid X = x_2) = \sum_u Y_1(u) P(Y_1 = Y_1(u) \mid X = x_2) = \sum_u Y_1(u) P(U = u \mid X = x_2) = \sum_u \left[ Y_1(u) \frac{P(X = x_2 \mid U = u) \cdot P(U = u)}{P(X = x_2)} \right]
\]

\[
= (82 + 89) \cdot 1/20 + (101 + 108) \cdot 2/20 + (118 + 131) \cdot 3/20 + (139 + 152) \cdot 4/20 = 125.
\]

Hence, we can conclude that strong ignorability does not hold in this example.

The third criterion, stochastic independence of \( X \) and \( U \) does not hold, because, unlike for treatment condition \( x_1 \), the individual assignment probabilities \( P(X = x_2 \mid U = u) \) and \( P(X = x_3 \mid U = u) \) are not the same for each observational unit \( u \) for treatment conditions \( x_2 \) and \( x_3 \).

The fourth causality criterion, unit-treatment homogeneity also does not hold. Although there are equal individual conditional expected values \( E(Y \mid X = x, U = u) \) in treatment conditions \( x_2 \) and \( x_3 \), these values differ in treatment condition \( x_1 \).

The fifth causality criterion, stochastic independence of \( X \) and \( U \) or unit-treatment homogeneity does not hold as well. However, the sixth causality criterion, unconfoundedness does hold. In each treatment condition \( x \) we have either equal assignment probabilities \( P(X = x \mid U = u) \) (for \( x_1 \)) or equal individual conditional expected values \( E(Y \mid X = x, U = u) \) (for \( x_2 \) and \( x_3 \)).

3. Some Equivalent Formulations of Unconfoundedness

We will now study several conditions which are equivalent to unconfoundedness as defined above. These conditions will provide the theoretical basis for empirical falsifiability. All these conditions involve random variables \( W \) for which there exists a map-
ping $f$ such that $W = f(U)$ is the composition of $U$ with $f$. As can easily be seen, such a variable $W$ partitions the population $\Omega_U$ into subpopulations (such as the set of male and the set of female persons). Presuming a potential causal regression model $\langle (\Omega, A, P), E(Y|X), U \rangle$ and using these variables, we may formulate the following theorem.$^{12}$

**Theorem 1.** $E(Y|X)$ is unconfounded if and only if for each value $x$ of $X$ and for each $W = f(U)$

$$P(X = x| W = w) = P(X = x) \quad \text{for each value } w \text{ of } W$$

(7)

or

$$E(Y|X = x, W = w) = E(Y|X = x) \quad \text{for each value } w \text{ of } W.$$  

(8)

According to this theorem, unconfoundedness postulates, within each treatment condition $x$, equal treatment probabilities for each subpopulation $w$ or equal conditional expected values across all subpopulations (represented by the different values $w$ of $W$). Note that, in empirical applications, this theorem already provides a basis for falsification of unconfoundedness, because, given a specific $W = f(U)$, each term in these two equations is empirically estimable. At least one of the two equations above has to be true for each $W = f(U)$ and a given value $x$ of $X$. If neither Equation (7) nor Equation (8) holds for a given $W = f(U)$, we can conclude that $E(Y|X)$ is confounded.

The next theorem will show the surprising fact that unconfoundedness as defined by the sixth criterion (see Def. 1) is equivalent to unconfoundedness as defined by Steyer, Gabler, and Rucai (1996).$^{13}$

**Theorem 2.** $E(Y|X)$ is unconfounded if and only if for each $W = f(U)$

$$E(Y|X = x) = \sum_w E(Y|X = x, W = w) P(W = w), \quad \text{for each value } x \text{ of } X.$$  

(9)

Note that the summation is across all values $w$ of $W$. According to this theorem, postulating Equation (9) for each variable $W = f(U)$ is equivalent to (general) unconfoundedness of the regression $E(Y|X)$. Also note that Equation (9) is not equivalent with Equation (7) or Equation (8). Instead Theorem 2 formulates (via Th. 1) the equivalence between “Equation (9) for each $W = f(U)$” and “Equation (7) or Equation (8) for each $W = f(U)$”. According to Equation (9) unconfoundedness implies that each conditional expected value $E(Y|X = x)$ is the average of the corresponding ex-

---

$^{12}$ For proofs see Appendix B.

$^{13}$ This formulation of unconfoundedness has also been called weak causality by Steyer (1992). There one will also find a formulation which is not restricted to discrete treatments and a discrete unit variable.
expected values $E(Y|X = x, W = w)$ in the subpopulations represented by the values $w$ of $W$.

**Corollary 1.** If the treatment regression $E(Y|X)$ is unconfounded, then for each $W = f(U)$:

$$PFE(i, j) = \sum_w [E(Y|X = x_i, W = w) - E(Y|X = x_j, W = w)] P(W = w). \quad (10)$$

According to this corollary, unconfoundedness implies that each prima facie effect, i.e., each difference $E(Y|X = x_i) - E(Y|X = x_j)$ between two values of the regression $E(Y|X)$, is the average of the differences $E(Y|X = x_i, W = w) - E(Y|X = x_j, W = w)$ across all subpopulations represented by different values $w$ of $W$.

Also note that Theorem 2 holds irrespective of any specific parameterization of the regression $E(Y|X)$, be it linear or not. For instance, if $Y$ is dichotomous with values 0 and 1 and $X$ is numerical, Theorem 2 also holds for a logistic regression $E(Y|X) = P(Y = 1|X) = \exp(\gamma_0 + \gamma_1 X) / [1 + \exp(\gamma_0 + \gamma_1 X)]$.

One might expect that Equation (9) is always true if $W$ is discrete. However, Example Ia of Steyer et al. (2000) (with $W = U$) shows that this expectation is wrong. An equation for the conditional expected values $E(Y|X = x)$ that is always true if $W$ is discrete is:

$$E(Y|X = x) = \sum_w E(Y|X = x, W = w) P(W = w|X = x), \text{ for each value } x \text{ of } X. \quad (11)$$

Note that Equation (9) also provides a basis for empirical falsification of unconfoundedness, because each term in this equation is empirically estimable. Hence, both Theorems 1 and 2 provide possibilities for falsifying unconfoundedness.

It will be useful to supplement the general concept of unconfoundedness by the concept of unconfoundedness with respect to a specific variable $W$. The definition will be such that general unconfoundedness is equivalent to unconfoundedness with respect to each $W = f(U)$.

**Definition 2.** $E(Y|X)$ is called unconfounded with respect to $W = f(U)$ if Equation (9) holds for each value $x$ of $X$.

What has been achieved so far? We have three different but equivalent ways of formulating unconfoundedness (see Table 3). This gives us a deeper understanding of this concept. Specifically, we have shown that, in contrast to causal unbiasedness and strong ignorability, unconfoundedness is empirically falsifiable (see versions 2 and 3 in Table 3).
### Table 3. Three equivalent formulations of unconfoundedness

1. For each value $x$ of $X$:
   
   \[ P(X = x | U = u) = P(X = x) \quad \text{for each value } u \text{ of } U \]
   
   or
   
   \[ E(Y | X = x, U = u) = E(Y | X = x) \quad \text{for each value } u \text{ of } U \]

2. For each variable $W = f(U)$ and for each value $x$ of $X$:
   
   \[ P(X = x | W = w) = P(X = x) \quad \text{for each value } w \text{ of } W \]
   
   or
   
   \[ E(Y | X = x, W = w) = E(Y | X = x) \quad \text{for each value } w \text{ of } W \]

3. For each variable $W = f(U)$ and for each value $x$ of $X$:
   
   \[ E(Y | X = x) := \sum_w E(Y | X = x, W = w) P(W = w), \text{ for each value } x \text{ of } X \]

### 4. Sufficient Conditions for Unconfoundedness

Contrasting unconfoundedness to other criteria, we will now show more formally that unconfoundedness is a comparatively weak sufficient condition for causal unbiasedness. This will be achieved studying some sufficient conditions for a treatment regression $E(Y | X)$ to be unconfounded. The sufficient conditions will also help to understand the concept and how it is related to the experiment with random assignment of units to experimental conditions.

**Theorem 3.** Each of the following conditions is sufficient for unconfoundedness of the regression $E(Y | X)$:

- (i) $U$ and $X$ are stochastically independent;
- (ii) (Unit-treatment homogeneity) $E(Y | X, U) = E(Y | X)$;
- (iii) (Strong causality) For each $W = f(U)$ there exists a function $h$ such that

\[
E(Y | X, W) = E(Y | X) + h(W) .
\]

Note that each of these conditions is sufficient for unconfoundedness of $E(Y | X)$ but not necessary, i.e., each of them implies unconfoundedness of $E(Y | X)$, but unconfoundedness of $E(Y | X)$ neither implies (i), nor (ii), nor (iii). Condition (i), independence of $X$ and $U$, is important for understanding the role of randomization, i.e., random assignment of units to treatment conditions. Randomization is the only way for the experimenter to deliberately create one of the sufficient conditions of unconfoundedness.
Equation (12) postulates the additive decomposability of the regression $E(Y|X, W)$ into a function $g(X)$ and a function $h(W)$, i.e.,

$$E(Y|X, W) = g(X) + h(W)$$ (13)

with the additional requirement $g(X) = E(Y|X)$, which, if Equation (13) holds, is equivalent to $E[h(W)|X] = E[h(W)]$.

For $W = U$, Equation (13) may be called unit-treatment additivity. It postulates the additive decomposability of the regression $E(Y|X, U)$ into a function $g(X)$ and a function $h(U)$, with the additional requirement $g(X) = E(Y|X)$. Equation (13) does not allow for interactions between $X$ and $W$ in the sense of analysis of variance (or moderator effects in terms of moderator regression models.) Strong causality as defined by condition (iii) requires invariance of the individual causal effects across all observational units, because, for $W = U$, Equation (13) implies

$$E(Y|X = x_1, U = u) - E(Y|X = x_2, U = u) = g(x_1) + h(u) - [g(x_2) + h(u)] = g(x_1) - g(x_2).$$ (14)

It should be noted that there is no sufficient condition for such an invariance of individual causal effects that could deliberately be created by the experimenter. Specifically, Equation (13) is not necessarily (and in fact seldom) true in the randomized experiment, in contrast, to unconfoundedness that always holds if there random assignment of units to treatment conditions. Obviously, unit-treatment homogeneity is a special case of unit-treatment additivity with $h(W) = 0$.

5. Sufficient Conditions for Strong Ignorability

In this section we state some sufficient conditions for strong ignorability. According to our next theorem, the first two sufficient conditions of unconfoundedness mentioned in Theorem 3 are also sufficient for strong ignorability.

**Theorem 4.** Each of the following conditions is sufficient for strong ignorability:

(i) $U$ and $X$ are stochastically independent;

(ii) (Unit-treatment homogeneity) $E(Y|X, U) = E(Y|X)$;

Note that each of these conditions is sufficient for strong ignorability but not necessary, i.e., each of them implies strong ignorability but strong ignorability neither implies (i) nor (ii).
6. Necessary Conditions of Unconfoundedness

The first necessary condition formulated in this section relates unconfoundedness to causal unbiasedness (Th. 5). Next we show that unconfoundedness and, therefore, causal unbiasedness, can be transferred to each subpopulation. Other necessary conditions treated are important for causal modeling in nonrandomized experiments, because they are the logical basis for falsifying a hypothesis of unconfoundedness.

**Theorem 5.** If the treatment regression \( E(Y|X) \) is unconfounded, then each of the following propositions hold:

(i) \( E(Y|X) \) is causally unbiased, i.e., \( E(Y|X = x) = CUE(Y|X = x) \) for each value \( x \) of \( X \);

(ii) the prima facie effects are causally unbiased, i.e., \( PFE(i, j) = ACE(i, j) \) for all pairs \((x_i, x_j)\) of values of \( X \).

We now turn to a very powerful theorem according to which unconfoundedness of a regression \( E(Y|X) \) allows to conclude that unconfoundedness and, therefore, causal unbiasedness (see Th. 5), can be transferred to each subpopulation. This is not only important for the interpretation of expected values and their differences in subpopulations but also allows to conclude that all properties of an unconfounded regression also hold for the corresponding regressions within subpopulations.

**Theorem 6.** If \( E(Y|X) \) is unconfounded and \( W = f(U) \), then, for each value \( w \) of \( W \), the regression \( E_{W=w}(Y|X) \) of \( Y \) on \( X \) in the subpopulation represented by \( W = w \) is unconfounded as well.

As mentioned before, this theorem implies that all properties of an unconfounded regression also apply for the regressions \( E_{W=w}(Y|X) \) within the subpopulations. Instead of rewriting all propositions for the subpopulations we just consider the property of causal unbiasedness of the regressions \( E_{W=w}(Y|X) \) of \( Y \) on \( X \) in the subpopulation represented by \( W = w \).

**Corollary 2.** If \( E(Y|X) \) is unconfounded and \( W = f(U) \), then the regressions \( E_{W=w}(Y|X) \) of \( Y \) on \( X \) in the subpopulations represented by \( W = w \) are causally unbiased, i.e.,

\[
E(Y|X = x, W = w) = \sum_u E(Y|X = x, U = u, W = w) \ P(U = u | W = w), \quad (15)
\]

for each pair \((x, w)\) of values of \( X \) and \( W \).\(^{14}\)

---

\(^{14}\) Note that \( E(Y|X = x, U = u) = E(Y|X = x, U = u, W = w) \), because we presuppose that \( W \) is a measurable function of \( U \) (see the definitions of \( U \) in Steyer et al., 2000 and of \( W \) above).
Theorem 6 and Corollary 2 show that *unconfoundedness* is a stable property in the sense that it carries over from the total population to all subpopulations. In contrast, *causal unbiasedness* may be incidentally true in the total population but fail in any subpopulation (see Example III of Steyer et al., 2000).

Table 4. Necessary conditions of unconfoundedness

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>The regressions $E_{W \rightarrow w}(Y \mid X)$ in each subpopulation $W = w$ are unconfounded and, therefore, causally unbiased</td>
</tr>
<tr>
<td>2.</td>
<td>The prima facie effects $PFE(i, j)$ in the total population as well as the prima facie effects $PFE_{W \rightarrow w}(i, j)$ in each subpopulation are causally unbiased</td>
</tr>
<tr>
<td>3.</td>
<td>For each variable $W = f(U)$ and for each value $x$ of $X$: $P(X = x \mid W = w) = P(X = x)$ for each value $w$ of $W$, or $E(Y \mid X = x, W = w) = E(Y \mid X = x)$ for each value $w$ of $W$</td>
</tr>
<tr>
<td>4.</td>
<td>For each variable $W = f(U)$ and for each value $x$ of $X$: $E(Y \mid X = x) := \sum_w E(Y \mid X = x, W = w) P(W = w)$, for each value $x$ of $X$</td>
</tr>
</tbody>
</table>

Table 4 gives a summary of the most important necessary conditions of unconfoundedness. The first two necessary conditions are important for substantive interpretations, because they add meaning to the concept of unconfoundedness which is otherwise not easily seen. The second two necessary conditions are also sufficient. Conditions 1, 3, and 4 may be used when it comes to falsification trials of the hypothesis of unconfoundedness. It should be noticed that, *for a given variable $W$ “Equation (7) or (8)” is stronger than “Equation (9)”, i.e., “Equation (7) or (8)” implies “Equation (9)” but not vice versa.*

7. General Procedures to Falsify Unconfoundedness

An important goal of this paper is to present a theory that may serve as a methodological foundation of causal modeling also outside the randomized experiment. Since a hypothesis that the conditional expected values $E(Y \mid X = x)$ are causally unbiased is not falsifiable, we introduced unconfoundedness of $E(Y \mid X)$ as the weakest empirically falsifiable sufficient condition for causal unbiasedness. How to proceed if we want to try to falsify the hypothesis of unconfoundedness? A first general principle is to choose a variable $W = f(U)$ and compare the estimates on the right-hand side of Equation (9)
to the estimates on the right-hand side of Equation (11), i.e., to compare, for each value \( x \) of \( X \),

\[
E_{adj \ for \ W}(Y \mid X = x) := \sum_w E(Y \mid X = x, W = w) \ P(W = w)
\]

to

\[
E(Y \mid X = x) = \sum_w E(Y \mid X = x, W = w) \ P(W = w \mid X = x).
\]

Note that the last equation is always true if \( W \) is discrete. If, in a specific empirical application, the estimate of \( E_{adj \ for \ W}(Y \mid X = x) \) cannot be assumed to be an estimate of \( E(Y \mid X = x) \), then unconfoundedness, too, cannot be assumed to hold.

According to the considerations above, falsifying in a specific application the hypothesis of unconfoundedness of \( E(Y \mid X) \) can be achieved by falsifying the null hypothesis of unconfoundedness of \( E(Y \mid X) \) with respect to a specific \( W = f(U) \), i.e., by falsifying the hypothesis:

\[
E_{adj \ for \ W}(Y \mid X = x) - E(Y \mid X = x) = 0, \quad \text{for each value } x \text{ of } X.
\]

This hypothesis is equivalent to Equation (9). It may also equivalently be written:

\[
\sum_w E(Y \mid X = x, W = w) \ [P(W = w) - P(W = w \mid X = x)] = 0, \quad \text{for each } x \text{ of } X.
\]

If the unconditional probabilities \( P(W = w) \) and the conditional probabilities \( P(W = w \mid X = x) \) are known, Equation (19) is a linear hypothesis about the parameters \( E(Y \mid X = x, W = w) \).

A second general procedure can be based on condition 4 in Table 4. Again, choose a variable \( W = f(U) \) and check if for each given value \( x \) of \( X \) condition

\[
P(X = x \mid W = w) = P(X = x) \quad \text{for each value } w \text{ of } W
\]

holds or

\[
E(Y \mid X = x, W = w) = E(Y \mid X = x) \quad \text{for each value } w \text{ of } W.
\]

If there is a value \( x \) of \( X \) for which none of these conditions hold, then unconfoundedness does not hold as well.

8. Another Example

The example displayed in Table 5 serves to illustrate the general procedures to falsify unconfoundedness outlined in the last section. Of course, a simple look at treatment
condition $x_2$ reveals that the requirements formulated in Definition 1 do not hold in this example. For $x_2$ we neither have constant assignment probabilities across units nor constant individual conditional expected values $E(Y|X = x_2, U = u)$ across units. As a consequence, the conditional expected value $E(Y|X = x_2)$ is causally biased. Whereas the causally unbiased expected value is $CUE(Y|X = x_2) = 105$, the conditional expected value $E(Y|X = x_2)$ is equal to 115. Hence, the prima facie effect $PFE(1, 2) = E(Y|X = x_1) - E(Y|X = x_2) = 115 - 115 = 0$ is considerably different from the average causal effect $ACE(1, 2) = 115 - 105 = 10$. Whereas the average causal effect is positive, the prima facie effect is zero.

In the last paragraph we used Definition 1 to check unconfoundedness. However, in empirical applications, the data on the individual unit level are neither available nor estimable. Hence, Definition 1 cannot be used in empirical falsification trials. What is estimable, however, are the data within subpopulations such as gender (see column 3 of Table 5). The conditional expected values $E(Y|X = x, W = w)$ may be computed by the formula

$$E(Y|X = x, W = w) = \sum_u E(Y|X = x, U = u, W = w) P(U = u|X = x, W = w),$$

where $E(Y|X = x, W = w, U = u) = E(Y|X = x, U = u)$, because $W = f(U)$. Hence, for this purpose we have to compute the conditional probabilities $P(U = u|X = x, W = w)$. Using the data of Table 5 we receive for $X = x_2$ and $W = m$:

$$E(Y|X = x_2, W = m) = 68 \cdot 1/6 + 81 \cdot 1/6 + 89 \cdot 2/6 + 102 \cdot 2/6 = 88.5,$$

and for $X = x_2$ and $W = f$:

$$E(Y|X = x_2, W = f) = 112 \cdot 3/14 + 119 \cdot 3/14 + 131 \cdot 4/14 + 138 \cdot 4/14 = 126.36.$$

The unconditional probabilities $P(W = w)$ are

$$P(W = f) = P(W = m) = 1/2,$$

whereas the conditional probabilities $P(W = w|X = x_2)$ are

$$P(W = m|X = x_2) = 3/10 \quad \text{and} \quad P(W = f|X = x_2) = 7/10.$$

Inserting these results into

$$\sum_w E(Y|X = x_2, W = w) [P(W = w) - P(W = w|X = x_2)],$$

[see Eq. (19)] yields

$$88.5 \cdot (1/2 - 3/10) + 126.36 \cdot (1/2 - 7/10) = 17.7 - 25.272 = -7.572.$$
Hence, this is a contradiction to the hypothesis that the regression \( E(Y|X) \) is unconfounded in the example of Table 5 [see Eq. (19)].

The second general procedure outlined in section 7 is easier to follow in this example. It easily seen from Table 5 that neither Equation (20) nor Equation (21) holds for \( X = x_2 \) and \( W = m \), for instance.

It should be emphasized again that all parameters necessary for the general procedures to falsify unconfoundedness illustrated in this example are estimable in “real” empirical applications. In this sense unconfoundedness is empirically falsifiable.

Table 5. An example in which the treatment regression \( E(Y|X) \) is confounded

| Observational-unit variable | \( W \) (gender) | \( P(X = x_1 | U = u) \) | \( E(Y|X = x_1, U = u) \) | \( P(X = x_2 | U = u) \) | \( E(Y|X = x_2, U = u) \) | \( P(X = x_3 | U = u) \) | \( E(Y|X = x_3, U = u) \) |
|-----------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| \( u_1 \)                  | 1/8 \( m \)      | 1/2              | 82               | 1/10             | 68               | 4/10             | 110              |
| \( u_2 \)                  | 1/8 \( m \)      | 1/2              | 89               | 1/10             | 81               | 4/10             | 110              |
| \( u_3 \)                  | 1/8 \( m \)      | 1/2              | 101              | 2/10             | 89               | 3/10             | 110              |
| \( u_4 \)                  | 1/8 \( m \)      | 1/2              | 108              | 2/10             | 102              | 3/10             | 110              |
| \( u_5 \)                  | 1/8 \( f \)      | 1/2              | 118              | 3/10             | 112              | 2/10             | 110              |
| \( u_6 \)                  | 1/8 \( f \)      | 1/2              | 131              | 3/10             | 119              | 2/10             | 110              |
| \( u_7 \)                  | 1/8 \( f \)      | 1/2              | 139              | 4/10             | 131              | 1/10             | 110              |
| \( u_8 \)                  | 1/8 \( f \)      | 1/2              | 152              | 4/10             | 138              | 1/10             | 110              |

Note: The (unconditional) probabilities for the three treatments are \( P(X = x_1) = 1/2 \), \( P(X = x_2) = P(X = x_3) = 1/4 \).

9. Special Procedures to Falsify Unconfoundedness

In the last two sections, we did not presuppose anything about the functional form of the regression \( E(Y|X, W) \). In empirical applications, the procedure outlined in these sections can be used if the sample is big enough to estimate each conditional expected value \( E(Y|X = x, W = w) \) sufficiently well and if the probabilities \( P(X = x|W = w) \) are known or can also be estimated sufficiently well. If, however,
compared to the sample size, the number of different combinations \((x, w)\) of values of \(X\) and \(W\) is too large, then there might even be pairs \((x, w)\) for which there are no observations at all. If, however, we are able to make an assumption about the functional form of the regression \(E(Y|X, W)\) sparseness of data are less of a problem. We will consider two such assumptions: linearity [Eq. (22)] and conditional linearity [Eq. (32)]. The case of linearity has already been considered by Pratt and Schlaifer (1988), for instance.

In the next theorem, \(X = (X_1, ..., X_p)'\) denotes a \(p\)-dimensional regressor and \(W = (W_1, ..., W_q)'\) a \(q\)-dimensional function of \(U\). Both, \(X\) and \(W\) are column vectors.\(^{15}\)

**Theorem 7.** If (a) \(W = f(U)\), (b) the treatment regression \(E(Y|X)\) is unconfounded (or at least unconfounded with respect to \(W\)), (c) the covariance matrix \(\Sigma\) of the \((p + q)\)-dimensional row vector \((X', W')\) is regular, and (d) \(E(Y|X, W) = \beta_0 + X' \beta_X + W' \beta_W, \quad \beta_0 \in \mathbb{IR}, \beta_X \in \mathbb{IR}^p, \beta_W \in \mathbb{IR}^q, \quad (22)\)

then

\[
E(Y|X) = \alpha_0 + X' \alpha_X, \quad \alpha_0 \in \mathbb{IR}, \alpha_X \in \mathbb{IR}^p
\]

\[
\alpha_X = \beta_X, \quad (24)
\]

\[
E(W'|X) \beta_W = E(W') \beta_W, \quad (25)
\]

\[
\Sigma_{XW} \beta_W = 0. \quad (26)
\]

In Equation (26) \(\Sigma_{XW}\) denotes the covariance matrix of \(X\) and \(W\).

Hence, under the linearity assumption [Eq. (22)], unconfoundedness of \(E(Y|X)\) can be falsified in a specific empirical application, if it is shown that \(\alpha_X = \beta_X\) does not hold. Alternatively, one might try to show that \(\Sigma_{XW} \beta_W = 0\), does not hold. Hence, tests of \(\alpha_X = \beta_X\) and tests of \(\Sigma_{XW} \beta_W = 0\) are tests of unconfoundedness, provided that \(W = f(U)\) and Equation (22) is true. Usually, a test of linearity or goodness of fit of Equation (22) should precede a test of unconfoundedness. Also note that, given the presuppositions of Theorem 7 [including Eq. (22)], Equations (24) and (26) are equivalent.

Observe that the vectors of regression coefficients \(\beta_X\) and \(\beta_W\) can be computed by

\[
\beta_X = (\Sigma_{XX}^{-1} + \Gamma E^{-1} \Gamma') \Sigma_{XY} - \Gamma E^{-1} \Sigma_{WY} \quad (27)
\]

\(^{15}\)Note that all concepts treated in this paper are not restricted to a univariate treatment variable \(X\). If \(X\) is \(p\)-variate, a value \(x\) of \(X\) consists of \(p\) values \(x_1, ..., x_p\). Since we are now formulating the models in terms of matrix algebra, we will denote the treatment variable by the boldface letter \(\textbf{X}\). The same holds true, of course, for \(\textbf{W}\).
and

$$\beta_w = E^{-1}(\Sigma_{wy} - \Gamma^T \Sigma_{xy}), \quad (28)$$

where $E := \Sigma_{ww} - \Sigma_{wx} \Sigma_{xx}^{-1} \Sigma_{wy}$ and $\Gamma := \Sigma_{xx}^{-1} \Sigma_{xy}$ (see, e.g., Seber, 1984). Also note that, for $p = q = 1$, Equations (27) and (28) simplify to the well-known formulas for partial regression coefficients:

$$\beta_X = \left(\sigma_w^2 \sigma_{xy} - \sigma_{wx} \sigma_{wy}\right) / \left(\sigma_X^2 \sigma_w^2 - \sigma_{wx}^2\right) \quad (29)$$

and

$$\beta_W = \left(\sigma_X^2 \sigma_{wy} - \sigma_{wx} \sigma_{xy}\right) / \left(\sigma_X^2 \sigma_w^2 - \sigma_{wx}^2\right). \quad (30)$$

In Theorem 7 we presuppose that the regression $E(Y|X, W)$ is additive. Especially, Equation (22) does not allow for multiplicative terms such as $X \cdot W$, which occur, e.g., in the following equation:

$$E(Y|X, W) = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 X \cdot W \quad (31)$$

$$= (\beta_0 + \beta_2 W) + (\beta_1 + \beta_3 W) X$$

$$= g_0(W) + g_1(W) X.$$

Note that $E(Y|X, W)$ can always be parameterized in the form of Equation (31), if both $X$ and $W$ are dichotomous. In other cases, however, Equation (31) may not hold for the regression $E(Y|X, W)$. The values of the slope coefficient function $g_i(W)$ are the values of the slope coefficients of the $(W = w)$-conditional linear regressions $E_{W=w}(Y|X) = g_0(w) + g_1(w) X$. If, e.g., $W$ is “gender”, $E_{W=m}(Y|X)$ may represent the regression of $Y$ on $X$ in the subpopulation of males ($m$).

In the next theorem, we consider a special case which allows for a $p$-dimensional regressor $X = (X_1, ..., X_p)$ '. The basic idea behind Equation (32) is that for each given value $w$ of $W$, the regression of $Y$ on $X$ is an additive linear multiple regression with the $p$-variate numerical regressor $X = (X_1, ..., X_p)$ '.

**Theorem 8.** If (a) the treatment regression $E(Y|X)$ is unconfounded, (b) $X: \Omega \rightarrow \mathbb{R}^p$, (c) $W = f(U)$, and (d) there are functions $g_0, ..., g_p$ such that

$$E(Y|X, W) = g_0(W) + g_1(W) X_1 + ... + g_p(W) X_p, \quad (32)$$

then

$$E(Y|X) = \alpha_0 + \alpha_1 X_1 + ... + \alpha_p X_p \quad (33)$$
with

\[\alpha_0 = E[g_0(W)], \quad \alpha_1 = E[g_1(W)], \quad \ldots, \quad \alpha_p = E[g_p(W)].\]  

(34)

According to this theorem, the partial regression coefficients of the regressors \(X_1, \ldots, X_p\) are the expected values of the corresponding \((W = w)\)-conditional partial regression coefficients if unconfoundedness holds. Furthermore, this result may also be used for falsification trials of the hypothesis that \(E(Y \mid X)\) is unconfounded. If the Equations (33) and (34) do not hold, although Equation (32) does, then we can reject the hypothesis that \(E(Y \mid X)\) is unconfounded. For \(p = 1\), this theorem shows how the slope coefficient \(\alpha_1\) of a simple linear regression \(E(Y \mid X) = \alpha_0 + \alpha_1 X\) is related to the slope coefficient function \(g_1(W)\), if unconfoundedness holds. A special case of Equation (32) results if \(X_1 := X^1, \ldots, X_p := X^p\), i.e., if the regression of \(Y\) on \(X\) is a polynomial function of the one-dimensional numeric regressor \(X\).

10. Discussion

In this paper we introduced the concept of unconfoundedness of a treatment regression \(E(Y \mid X)\) and showed its relationship to Neyman’s and Rubin’s concepts of individual and average causal effects. We discussed several criteria that are relevant in causal modeling. The weakest is causal unbiasedness of the conditional expected values \(E(Y \mid X = x)\) [see Eq. (1)] which has the disadvantage of being nonfalsifiable. Other conditions were independence of \(U\) and \(X\), as well as unit-treatment homogeneity: \(E(Y \mid X, U) = E(Y \mid X)\) and the and/or-combination of the two. All these conditions are sufficient conditions for causal unbiasedness but they are unnecessarily strong. We finally introduced unconfoundedness as the weakest empirically falsifiable sufficient condition for causal unbiasedness. We defined a regression \(E(Y \mid X)\) to be unconfounded if, for each value \(x\) of \(X\): (a) \(P(X = x \mid U = u) = P(X = x)\) for each value \(u\) of \(U\), or (b) \(E(Y \mid X = x, U = u) = E(Y \mid X = x)\) for each value \(u\) of \(U\). This concept goes beyond the observational level and still has many implications, some of which are empirically testable.

Unconfoundedness implies that the conditional expected values \(E(Y \mid X = x)\) and their differences \(E(Y \mid X = x_i) - E(Y \mid X = x_j)\), the prima facie effects, are causally unbiased, i.e. \(E(Y \mid X = x_i) - E(Y \mid X = x_j) = ACE(i, j)\). This provides the link to the concept of the average causal effect. Furthermore, we showed that unconfoundedness of \(E(Y \mid X)\) is sufficient to imply that the conditional expected values \(E(Y \mid X = x, W = w)\) and their differences \(E(Y \mid X = x, W = w) - E(Y \mid X = x, W = w)\) are causally unbiased in every subpopulation \(W = w\) (see Corollary 2).
The sufficient conditions for unconfoundedness of $E(Y|X)$ may guide us in the design of experiments. Most important, however, they make clear that random assignment of units to experimental conditions serves to secure stochastic independence of $X$ and $U$, which implies unconfoundedness and causal unbiasedness. Such an experiment with random assignment of observational units to experimental conditions is in fact the only way to deliberately create the circumstances in which average causal effects can be estimated. Although unit-treatment homogeneity, [i.e., $E(Y|X, U) = E(Y|X)$] is also sufficient for unconfoundedness (see Th. 3), this condition can not deliberately be created through techniques of experimental design. Nevertheless, both sufficient conditions for unconfoundedness, independence of $X$ and $U$, and unit-treatment homogeneity, show that causal modeling in randomized and nonrandomized experiments aims at the same goal: to secure causal unbiasedness. Hence, if it were possible to generalize the theory presented such that it could also cover causal modeling in observational (i.e., nonexperimental) studies, in which the $X$-variables do not represent experimental conditions, the two traditions of theorizing about causality in statistics identified by Cox (1992), experimental and observational, could be integrated into a single theory.

The necessary conditions for unconfoundedness of $E(Y|X)$ may guide us in nonrandomized experimental causal modeling, because they provide empirically testable consequences. The theorems presented show which statistical hypotheses have to be tested and rejected in order to falsify the hypothesis of unconfoundedness of $E(Y|X)$ in a specific application.

What happens if a $W$ has been identified with respect to which there is confounding? First, we may look at the regressive dependencies of $Y$ on $X$ within each subpopulation defined by $W = w$, i.e., we may look at the conditional effects. In fact, this provides more detailed information about the conditional regressive dependence of $Y$ on $X$ given $W = w$. Note that this procedure is in full accordance with the framework presented in this paper. One would just replace the total population $\Omega_U$ by the subpopulation defined by $W = w$. This procedure is, in principle, equivalent with statistically controlling for the covariate vector $W$ provided that the equation for the regression $E(Y|X, W)$ is adequately specified. Second, if we are willing to assume that the conditional regressions $E_{W=w}(Y|X)$ are unbiased, we may also compute the average causal effects via Equation (16), because, under this assumption

$$ACE(i, j) = E_{adj for W}(Y \mid X = x_i) - E_{adj for W}(Y \mid X = x_j)$$

(for details, see Wüthrich-Martone, Steyer, Nachtigall & Suhl, 1999).

A limitation requiring a more general formulation of our approach follows from assuming $P(X = x, U = u) > 0$. Whereas this assumption has been useful for a simple presentation, it restricts generality because it is not compatible with the assumption
that $X$ has a normal distribution, for instance, which is often made in nonexperimental causal modeling (see, e.g., Bollen, 1989). Since normality is not really necessary in structural equation models (see, e.g., Browne & Arminger, 1995), this restriction is not too serious. Nevertheless, we should look for a more general theory which is not restricted to discrete treatments and units.

Another open problem is how the theory presented relates to statistical sampling models. Especially, statistical tests and procedures have to be developed to help us in deciding whether or not unconfoundedness holds in a specific application. For some cases, these tests and procedures have already been presented (see, e.g., Allison, 1995; Clogg et al., 1992; Clogg et al., 1995; Steyer, Gabler, Rucai & Schuster, 1997; von Davier, 2000), for others they still need to be developed.

A significance test of unconfoundedness in very large samples may not really be meaningful in many empirical applications, because the exact null hypothesis will rarely hold in nonrandomized experiments. Instead, one might rather be interested in estimating a coefficient reflecting the strength of confounding of a regression $E(Y|X)$ with respect to a specific variable $W$. Such a coefficient has been proposed by Steyer, Gabler, and Rucai (1996). More research and experience will be necessary to work out principles dealing with the practical relevance of confoundings in concrete empirical studies.

References


**Appendix A: Extended Notes**

**Note 1.** Strong ignorability implies causal unbiasedness.

\[
E(Y \mid X = x_i) = \sum_u E(Y \mid X = x_i, U = u) P(U = u \mid X = x_i) \quad \text{[always true if } U \text{ is discrete]}
\]

\[
= \sum_u Y_i(u) P(U = u \mid X = x_i) \quad \text{[definition of } Y_i]\]

\[
= \sum_u Y_i(u) P(Y_i = Y_i(u) \mid X = x_i) \quad \text{[see comment below]}
\]

\[
= E(Y_i \mid X = x_i) \quad \text{[definition of conditional expected value]}
\]

\[
= E(Y_i) \quad \text{[because of strong ignorability]}
\]
\[
= \sum_u Y_i(u) \ P[Y_i = Y_i(u)] \quad \text{[definition of expected value]}
\]
\[
= \sum_u Y_i(u) \ P(U = u) \quad \text{[see comment below]}
\]
\[
= \sum_u E(Y|X = x_i, U = u) \ P(U = u) \quad \text{[definition of } Y]\]
\[
= \text{CUE}(Y|X = x_i). \quad \text{[definition of the causally unbiased expected value]}
\]

**Comment.** Note that the independence condition \(Y_1, \ldots, Y_n \perp X\) in strong ignorability implies regressive independence \(E(Y_i|X = x) = E(Y_i)\) for each pair \((i, x)\) of indices \(i \in \{1, \ldots, n\}\) and values \(x\) of \(X\) and that the events \(\{Y_i = Y_i(u)\} := \{\omega \in \Omega: Y_i(\omega) \in \{Y_i(u)\}\}\) and \(\{U = u\} := \{\omega \in \Omega: U(\omega) \in \{u\}\}\) are identical subsets of \(\Omega\) and therefore have identical conditional and unconditional probabilities. This does not presume that all the variables \(Y_i\) or one them are one-to-one functions of \(U\). It is still possible that two different units \(u_1\) and \(u_2\) have identical values \(Y_i(u_1) = Y_i(u_2)\).

### Appendix B: Proofs

**Proof of Theorem 1.** In this proof we will use the so-called *tower property* of regressions (see, e.g., Williams, 1991, p. 88 or Steyer & Eid, 2001, p. 357):

\[
E[Y|f(X)] = E[E(Y|X)|f(X)],
\]

(35)

and its special case

\[
E(Y) = E[E(Y|X)]
\]

(36)

for \(f(X) = \text{constant}\). Equation (36) is always true and Equation (35) is always true for every (measurable) mapping \(f(X)\) of \(X\). If unconfoundedness as defined in Definition 1 holds, then, for each value \(x\) of \(X\), Equation (3) or Equation (4) will hold. Let us assume that Equation (4) holds. This equation may also be written

\[
E_{X=x}(Y|U) = E_{X=x}(Y),
\]

(37)

where the \(E_{X=x}\) means that we refer to the expectation (or conditional expection) with respect to the conditional probability measure \(P_{X=x}\). Since \(W\) is a measurable mapping of \(U\) we may apply Equation (35) which yields:

\[
E_{X=x}(Y|W) = E_{X=x}[E_{X=x}(Y|U)|W]
\]

[see Eq. (35)]

\[
= E_{X=x}[E_{X=x}(Y)|W] = E_{X=x}(Y). \quad \text{[see Eqs. (37) and (36)]}
\]

The corresponding argument may be applied to the case in which Equation (3) holds. As \(P(X = x|U = u) = E[I_{X=x}|U = u]\), where \(I_{X=x}\) denotes the indicator of the event \(X = x\), Equation (3) may also be written
\[ E(I_{X=x} | U) = E(I_{X=x}). \]  
(38)

Since \( W \) is a measurable mapping of \( U \) we may again apply Equation (35) which, together with Equation (38) yields:

\[
E(I_{X=x} | W) = E[E( I_{X=x} | U ) | W]  
\]
[see Eq. (35)]

\[
= E[E(I_{X=x}) | W] = E(I_{X=x}).  
\]
[see Eqs. (38) and (36)]

However, this is just another way to write Equation (7).

The other direction of implication is more simple to prove: The observational unit variable \( U \) itself may be written \( f(U) \), because \( U = id(U) \), where \( id \) is the identity mapping.

The following lemma will be needed in the proof of Theorem 2.

**Lemma 1.** If Equation (9) holds for each \( W = f(U) \), then, for each value \( x \) of \( X \) and each value \( w \) of \( W \),

\[
P(W = w | X = x) = P(W = w) \quad \text{or} \quad E(Y|X = x, W = w) = E(Y|X = x). \]  
(39)

**Proof of Lemma 1.** Let \( I_{W=w} \) denote the indicator variable for the event

\[ W = w := \{ \omega \in \Omega: W(\omega) = w \} \] and note that \( I_{W=w} = f(U) \). Hence, Equation (9) implies:

\[
E(Y|X = x) = E(Y|X = x, I_{W=w} = 1) P(W = w) + E(Y|X = x, I_{W=w} = 0) P(W \neq w). \]  
(40)

Another equation for \( E(Y|X = x) \) which is always true [we still assume \( P(X = x, U = u > 0 \) implying \( P(X = x, W = w > 0) \)] is:

\[
E(Y|X = x) = E(Y|X = x, I_{W=w} = 1) P(W = w|X = x) + E(Y|X = x, I_{W=w} = 0) P(W \neq w|X = x). \]  
(41)

If we subtract Equation (41) from (40), we receive:

\[
E(Y|X = x, I_{W=w} = 1) [P(W = w) - P(W = w|X = x)] \]

\[ = E(Y|X = x, I_{W=w} = 0) [P(W = w) - P(W = w|X = x)]. \]

This implies \( P(W = w) = P(W = w|X = x) \) or \( E(Y|X = x, I_{W=w} = 1) = E(Y|X = x, I_{W=w} = 0) \), and \( P(X = x) = P(X = x|W = w) \) or \( E(Y|X = x, W = w) = E(Y|X = x) \), which was to be shown.

**Proof of Theorem 2.** If unconfoundedness as defined in Definition 1 holds, then Equation (7) or (8) will be true (see Th. 1). We first show that Equation (7) as well as Equation (8) imply Equation (9).

If Equation (7) holds, then

\[
P(W = w|X = x) = P(W = w) \quad \text{for each value} \ w \ \text{of} \ W \]
(42)
will hold as well. [We still presume \( P(X = x, U = u) > 0 \), which implies \( P(X = x, W = w) > 0 \), \( P(X = x) > 0 \), and \( P(W = w) > 0 \).]

Hence, using Equation (11) – which is always true – and (42), Equation (9) follows. If Equation (8) holds, then

\[
\sum_w E(Y|X=x, W=w) P(W = w) = \sum_w E(Y|X=x) P(W = w) = E(Y|X=x) \sum_w P(W = w) = E(Y|X=x),
\]

for each value \( x \) of \( X \), which again is Equation (9).

The other direction of the implication is more difficult to prove. We will conduct an indirect proof: Assuming neither Equation (7) nor Equation (8) were true will lead to a contradiction which together with Theorem 1 completes the proof.

Assume, for each \( W = f(U) \), Equation (9) holds and there were a \( W = f(U) \) with two values \( w_1, w_2 \) such that

\[
P(W = w_1|X = x) \neq P(W = w_1) \quad \text{and} \quad E(Y|X = x, W = w_2) \neq E(Y|X = x).
\]

(43)

for the same value \( x \) of \( X \). For \( w_1 = w_2 \), this would contradict Lemma 1. Hence, according to Lemma 1:

\[
P(W = w_1|X = x) \neq P(W = w_1) \quad \text{and} \quad E(Y|X = x, W = w_1) = E(Y|X = x),
\]

(44)

\[
E(Y|X = x, W = w_2) \neq E(Y|X = x) \quad \text{and} \quad P(W = w_1|X = x) = P(W = w_1),
\]

(45)

as well as \( w_1 \neq w_2 \). This is tantamount to assuming that neither Equation (7) nor Equation (8) were true. We now show that this leads to a contradiction. Let \( I_{12} \) be an indicator variable such that \( I_{12} = 1 \) if \( W = w_1 \) or \( W = w_2 \) and \( I_{12} = 0 \) otherwise. Then

\[
P(I_{12} = 1|X = x) = P(W = w_1|X = x) + P(W = w_2|X = x) = P(W = w_1|X = x) + P(W = w_2).
\]

If \( P(I_{12} = 1|X = x) = P(I_{12} = 1) \), then \( P(I_{12} = 1|X = x) = P(W = w_1) + P(W = w_2) \). But this implies \( P(W = w_1|X = x) = P(W = w_1) \), which would contradict (44). Hence,

\[
P(I_{12} = 1|X = x) \neq P(I_{12} = 1) \quad \text{and} \quad E(Y|X = x) = E(Y|X = x, I_{12} = 1),
\]

(46)

again using Lemma 1. We have

\[
E(Y|X = x, I_{12} = 1) = E(Y|X = x, W = w_1) P(W = w_1|X = x) / P(I_{12} = 1|X = x)
\]

\[
= E(Y|X = x, W = w_2) P(W = w_2|X = x) / P(I_{12} = 1|X = x).
\]

(47)

Inserting (46) as well as (44) into (47) yields

\[
E(Y|X = x) - E(Y|X = x) P(W = w_1|X = x) / P(I_{12} = 1|X = x)
\]

\[
= E(Y|X = x, W = w_2) P(W = w_2|X = x) / P(I_{12} = 1|X = x).
\]
Now
\[ P(W = w_1 | X = x) / P(I_{i2} = 1 | X = x) = 1 - P(W = w_2 | X = x) / P(I_{i2} = 1 | X = x) \]
implies
\[ [E(Y | X = x) - E(Y | X = x, W = w_2)] P(W = w_2 | X = x) / P(I_{i2} = 1 | X = x) = 0. \]
It follows that \( E(Y | X = x) = E(Y | X = x, W = w_2) \) which is a contradiction to (45). Hence, we have shown that if Equation (9) holds for each \( W = f(U) \), then Equation (7) or Equation (8) must hold for each \( W = f(U) \). Theorem 1 completes the proof.

**Proof of Corollary 1.** This proposition immediately follows from Theorem 2 and
\[
\text{PFE}(i, j) = \sum_w [E(Y | X = x, W = w) - E(Y | X = x, W = w)] P(W = w) =
\]
\[
\sum_w E(Y | X = x, W = w) P(W = w) - \sum_w E(Y | X = x, W = w) P(W = w). \]

**Proof of Theorem 3.** Propositions (i) and (ii) are immediate consequences of Definition 1. Proposition (iii) can be derived as follows: According to Theorem 2 we have to show that Equation (9) holds for each \( W = f(U) \).
\[
E(Y | X = x) = E(Y | X = x) \cdot 1 = E(Y | X = x) \cdot \sum_w P(W = w) =
\]
\[
\sum_w E(Y | X = x) \cdot P(W = w) =
\]
\[
\sum_w [E(Y | X = x, W = w) - h(w)] \cdot P(W = w) \quad [\text{see Eq. (12)}]
\]
\[
= \sum_w E(Y | X = x, W = w) \cdot P(W = w) - \sum_w h(w) \cdot P(W = w)
\]
\[
= \sum_w E(Y | X = x, W = w) \cdot P(W = w). \]

The last equation follows from \( \sum_w h(w) \cdot P(W = w) = E[h(W)] = 0 \), which is true because: \( E(Y) = E[\sum_{i=1}^{n} Y_i W_i] = E[E(Y | X, W)] = E[Y] + E[h(W)] \).

**Proof of Theorem 4.** (i) The variables \( Y_i \) are defined in such a way that they are measurable functions of \( U \). Hence, independence of \( U \) implies independence of \( Y_1, \ldots, Y_n \) and \( X \).

(ii) Unit-treatment homogeneity implies that each \( Y_i \) is a constant, and constants are always stochastic independent of any random variable.

**Proof of Theorem 5.** Since \( U \) and \( X \) are assumed to be discrete, the following equation is always true for each value \( x \) of \( X \):
\[
E(Y | X = x) = \sum_u E(Y | X = x, U = u) P(U = u | X = x). \quad (48)
\]
If Equation (3) holds for a given value $x$ of $X$, then $P(U = u \mid X = x) = P(U = u)$, for each value $u$ of $U$. Inserting $P(U = u)$ in the last equation yields unbiasedness of each conditional expected value $E(Y \mid X = x)$. If Equation (4) holds, for this value $x$ of $X$, then

$$E(Y \mid X = x) = E(Y \mid X = x) \sum_u P(U = u)$$

$$= \sum_u E(Y \mid X = x) P(U = u)$$

$$= \sum_u E(Y \mid X = x, U = u) P(U = u).$$

This proves proposition (i).

Proposition (ii) now follows from (i) and

$$ACE(i, j) = \sum_u \left[ E(Y \mid X = x_i, U = u) - E(Y \mid X = x_j, U = u) \right] P(U = u).$$

$$= \sum_u E(Y \mid X = x_i, U = u) P(U = u) - \sum_u E(Y \mid X = x_j, U = u) P(U = u)$$

$$= CUE(Y \mid X = x_i) - CUE(Y \mid X = x_j).$$

Proof of Theorem 6. Let $\Omega_U := \{ u \in \Omega : W(\omega) = w \}$, denote the subpopulation represented by $W = w$ and suppose Equation (3) holds. Because of

$$P_W(X = x \mid U = u) = \begin{cases} P(X = x \mid U = u), & \text{if } u \in \Omega_U, \\ \text{undefined, otherwise,} \end{cases}$$

Equation (3) implies

$$P_W(X = x \mid U = u) = \begin{cases} P(X = x) = P_{W \mid U}(X = x), & \text{if } u \in \Omega_U, \\ \text{undefined, otherwise.} \end{cases}$$

Now suppose Equation (4) holds. Then, because of

$$E_{W \mid U}(Y \mid X = x, U = u) = \begin{cases} E(Y \mid X = x, U = u), & \text{if } u \in \Omega_U, \\ \text{undefined, otherwise,} \end{cases}$$

Equation (4) implies

$$E_{W \mid U}(Y \mid X = x, U = u) = \begin{cases} E(Y \mid X = x) = E_{W \mid U}(Y \mid X = x), & \text{if } u \in \Omega_U, \\ \text{undefined, otherwise.} \end{cases}$$

This completes the proof, because we have shown that for each value $x$ of $X$

$$P_W(X = x \mid U = u) = P_{W \mid U}(X = x) \quad \text{for each value } u \text{ of } \Omega_U. \quad (49)$$

or

$$E_{W \mid U}(Y \mid X = x, U = u) = E_{W \mid U}(Y \mid X = x) \quad \text{for each value } u \text{ of } \Omega_U. \quad (50)$$

Proof of Corollary 2. This is an immediate consequence of Theorem 6 and Theorem 5 (i).

Proof of Theorem 7. For each value $x$ of $X$: 
\[ E(Y|X = x) = \sum_{w} E(Y|X = x, W = w) \cdot P(W = w) \quad \text{[see Eq. (9)]} \]
\[ = \sum_{w} (\beta_0 + x' \beta_x + w' \beta_w) \cdot P(W = w) \quad \text{[see Eq. (22)]} \]
\[ = \beta_0 + x' \beta_x + E(W') \beta_w \]
\[ = [\beta_0 + E(W') \beta_w] + x' \beta_x. \]

Hence, the constant \( \alpha_0 \) of the regression \( E(Y|X) \) is identical with \( \beta_0 + E(W') \beta_w \) and the vector of regression coefficients \( \alpha_x \) is equal to \( \beta_x \), which proves equations (23) and (24).

Equation (25) may be derived in the same way, only using
\[ E(Y|X = x) = \sum_{w} E(Y|X = x, W = w) \cdot P(W = w|X = x) \quad \text{[see Eq. (11)]} \]
to start with. This yields
\[ E(Y|X = x) = [\beta_0 + E(W'|X = x) \beta_w] + x' \beta_x. \]
The last equation shows that \( E(Y|X) \) will be of the form \( \alpha + X' \alpha_x \) (see Eq. (23)) with \( \alpha_x = \beta_x \) (see Eq. (24)), if \( E(W'|X = x) \beta_w = E(W') \beta_w \). This proves Equation (25).

Let \( \Sigma_{XW} = \text{Cov}(X, W) \) denote the \( p \times q \) covariance matrix of \( X \) and \( W \). Since \( \text{Cov}(X, W) = \text{Cov}(X, E(W|X)) \),
\[ \Sigma_{XW} \beta_w = \text{Cov}(X, E(W|X)) \beta_w \]
\[ = \text{Cov}(X, E(W'|X) \beta_w] \quad \text{(bilinearity of the covariance)} \]
\[ = \text{Cov}(X, E(W') \beta_w] \quad \text{[see Eq. (25)]} \]
\[ = 0, \]
because \( E(W') \beta_w \) is a constant. This proves Equation (26).

Using \( \text{Cov}(X, Y) = \text{Cov}(X, E(Y|X, W)), \text{Cov}(W, Y) = \text{Cov}(W, E(Y|X, W)) \), and Equation (22) yields:
\[ \Sigma_{XY} = \Sigma_{XX} \beta_x + \Sigma_{XW} \beta_w \]
\[ \Sigma_{YW} = \Sigma_{WX} \beta_x + \Sigma_{WW} \beta_w. \]
Writing these two equations in a single matrix equation, the well-known results for the inverse of a partitioned matrix (e.g., Seber, 1984, p. 519) yield equations (27) and (28).

**Proof of Theorem 8.** For each value \( x \) of \( X \):
\begin{align*}
E(Y | X = x) & = \sum_w E(Y | X = x, W = w) \cdot P(W = w) \quad \text{[see Eq. (9)]} \\
& = \sum_w [g_0(w) + g_1(w)x_1 + \ldots + g_p(w)x_p] \cdot P(W = w) \quad \text{[see Eq. (32)]} \\
& = E[g_0(W)] + E[g_1(W)]x_1 + \ldots + E[g_p(W)]x_p.
\end{align*}