11.3.4 Data vs. knowledge in covariate selection

What then can be done in the absence of a causal graph? One way is to postulate a plausible graph, based on one’s understanding of the domain, and check if the data refutes any of the statistical claims implied by that graph. In our case, the graph of Figure 11.8 (b) advertises several such claims, cast as conditional independence constraints, each associated with a missing arrow in the graph:

\[
\begin{align*}
V \perp X | Z_1, W_1 & \quad V \perp W_2 & \quad Z_2 \perp W_1 | W_2 \\
V \perp Y | X, Z_2, W_2 & \quad Z_1 \perp Z_2 | V, W_1, W_2 & \quad Z_1 \perp W_2 | W_1 \\
V \perp W_1 & \quad X \perp Z_2 | Z, W_1, W_2 & \quad X \perp \{V, Z_2\} | Z_1, W_1, W_2
\end{align*}
\]

Satisfying these constraints does not establish of course the validity of the causal model postulated because, as we have seen in Chapter 2, alternative models may exist which satisfy the same independence constraints yet embody markedly different causal structures, hence, markedly different admissible sets and effect estimands. A trivial example would be a complete graph, with arbitrary orientation of arrows which, with a clever choice of parameters, can emulate any other graph. A less trivial example, one that is not sensitive to choice of parameters lies in the class of equivalent structures, in which all conditional independencies emanate from graph separations. The search techniques developed in Chapter 2 provide systematic ways of representing all equivalent models compatible with a given set of conditional independence relations.

For example, the model depicted in Figure 11.9 is observationally equivalent to that of Figure 11.8 (b). In particular, it satisfies all the conditional independencies implied by the latter, and none others. However, in contrast to Figure 11.8 (b), the sets \(\{Z_1, W_1 W_2\}, \{V, W_1, W_2\}\), and \(\{Z_2, W_1, W_2\}\) are admissible. Adjusting for the latters would remove bias if the correct model is Figure 11.9 and would produce bias if the correct model is 11.8 (b).

Figure 11.9: A model that is statistically indistinguishable from that of Figure 11.8 (b), in which the irreducible sets \(\{Z_1, W_1 W_2\}, \{V, W_1, W_2\}\), and \(\{W_1, W_2, Z_2\}\) are admissible.

Is there a way of telling the two models apart? Although the notion of “observational equivalence” precludes discrimination by statistical means, substantive causal knowledge may provide discriminating information. For example, the model of Figure 11.9 can be
ruled out if we have good reasons to believe that variable $W_2$ cannot have any influence on $X$ (e.g., it may occur later than $X$) or that $W_1$ could not possibly have direct effect on $Y$.

The power of graphs lies in offering investigators a transparent language to reason about, and discuss the plausibility of such assumptions and, when consensus is not reached, to isolate differences of opinion and identify what additional observations would be needed to resolve differences. This facility is lacking in the potential outcome approach where, for most investigators, “strong ignorability” remains a mystical black box.

In addition to serving as carriers of substantive judgments, graphs also offer one the ability to reject large classes of models without testing each member of the class. For example, all models in which $V$ and $W_1$ are the sole parents of $X$, thus rendering $\{V, W_1\}$ (as well as $C$) admissible, could be rejected at once if the condition $X \perp Z_1 | V, W_1$ does not hold in the data.

A covariate set $S$ that contains all parents of $X$ exhibits two useful properties: (1) admissibility and (2) monotonicity, i.e. $S$ remains admissible under any expansion with new elements. Such a set is rightly called complete. Testing for completeness cannot be done by statistical means, of course, because one can never rule out the possibility that the entire set of measured covariates are $X$’s parents (unless some are known to occur later than $X$.) We can however reject the hypothesis that some proper subset of $S$ is $X$’s parents. This can be established statistically if we find that every subset $S_1$ of $S$ that is $\phi$-equivalent to $S$ violates condition $C_1$ (with $S_2$ any other subset of $S$.) Equivalently, if we find that the only minimal subset of $S$ satisfying $C_1$ is $S$ itself.

In Figure 11.8 (b), for example, we cannot reject the hypothesis that $Z_1, W_1$ and $W_2$ are $X$’s parents because this set is $\phi$-equivalent to $C$ and satisfies $C_1$ (by $d$-separating $X$ from $\{V, Z_2\}$.) Indeed, in the model of Figure 11.9, $C$ is shown to properly contain all parents of $X$, namely, $Z_1, W_1$ and $W_2$. 