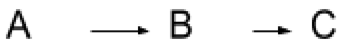
1.Introduction

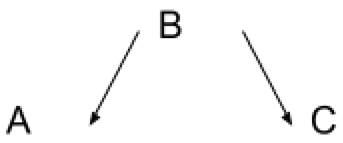
Rubin’s framework can sometimes be is the dominant approach in applied statistics, but the latter approach can sometimes highlight unexpected results that inform the proper analysis of observational data
https://zhuanlan.zhihu.com/p/33860572

1.1 Preliminary

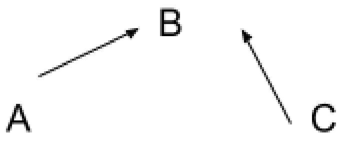
Keep in mind the following three structures. They are very important for you to understand the logic behind the basic idea in the next section.



A is correlated with c. when conditioned on B, however, A is independent of C on this path. **This means, conditioning on B is a way to cut-off the correlation between A and C on this path**



A is correlated with c. when conditioned on B, A is independent of C on this path.**This means, conditioning on B is a way to cut-off the correlation between A and C on this path**

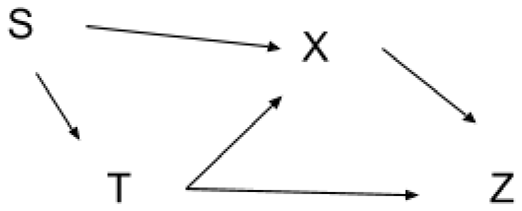


A and C is independent. When conditioned on B, however, A is correlated with C. This is intuitive: when A changes, to make B constant, C has to change. **This means, conditioning on B is a way to admit the correlation between A and C**

2.Basic idea

2.1 Estimating Objective

You first draw an graph showing the ‘causal relation’ of the variables you are interested in according to your knowledge. For example:



,

Now we want to study pr(z|x). How to do this?
First, we want to search for a set V of variables from all the variables other than z and x. Using the V, we estimate to the effect of x on z by conditioning on variables in V:

$$pr(z|x) = \sum_{V} pr(z|x, V)pr(V)$$

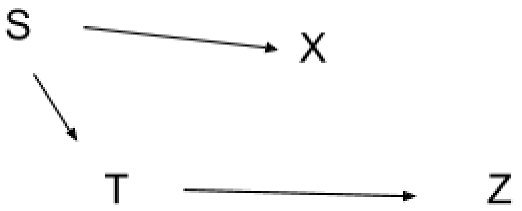
So here the task is to find out a V. Now we have candidates S,T. We want to select some from these three variables by some criterion.

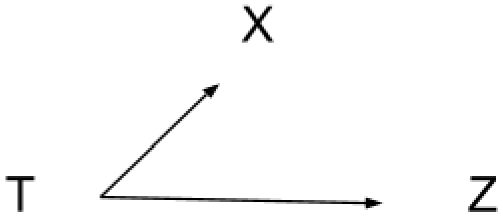
2.2 Backdoor Criterion

This criterion is called ‘Back-door criterion’. It says:

1. V does not include the descendent of x. Both S and T satisfy this.
2. V can block **every path** between **X and Z** that **has arrow pointing to x**.(**such a path is called backdoor path**)

   We first find out all the backdoor paths from X to Z. There are two backdoor paths in the graph





What is the definition of ‘block’? **Nodes set M can block a path** iff:
(1)If the path contains a chain (i.e., i→j→k) and j belongs to M, or
(2)if the path contains a fork (i.e., i←j→k) and j belongs to M, or
(3)if the path contains a inverted fork (i→j←k) and j and j’s descendants do not belong to M.

The definition here means that M can block the path if **ANY** of the three conditions are met.

Intuitively, why? In the left graph, we know that the change of X may be due to the change of S, while the change of T leads to change of T, which lead to the change of Z. This channel will obviously mix into the effect of x on z.

Therefore, we need to condition on T or S. By condition on T or S, we simply ‘cut-off’ the above channel: As is known from our basic structure, by conditioning on T, the S and Z becomes independent (on this path), therefore, Although variation of X is linked to the variation of S, the variation of S now has nothing to do with the variation of Z. On the other hand, by conditioning on T, the X and T becomes independent (on this path). Although variation of T is linked to the variation of Z, the variation of X has nothing to do with the variation of T.

In the right graph, the reasoning is the same. The change of X may be due to the change of T, while the change of T leads to the change of Z. by conditioning on T, the variation of X is independent of the variation of Z (on this path!), thus exclude this channel.

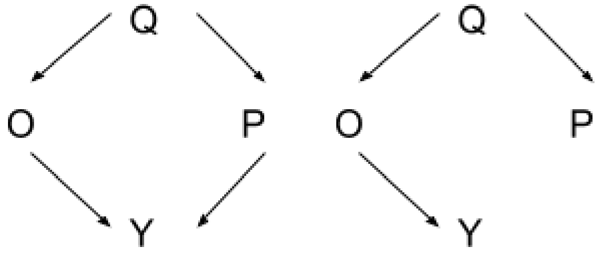
According to this definition, What is the possible set that can block the first back door path? {S} {T} {S,T} all can be such a set. What is the possible set that can block the second back door path? Only {T} can be such a set. **Combine both, we find out that {T}{S,T} can be V.**For example we set V to be {T}.Then we can estimate the effect of x on z using :

$$pr(z|x) = \sum_{T} pr(z|x,T)pr(T)$$

That is, we identify a set of nodes in {V} to condition on such that:
We block all spurious paths from X to Z. (that is, block the backdoor paths)
We leave all directed paths from X to Z unperturbed.
We do not inadvertantly create new spurious paths via conditioning on colliders or their descendants.’(the (3) of block definition)

3.Examples

Example 1:



We want to see $pr(p|y)$. First, we find out the candidate variables, which are O and Q. we find out all the back-door paths from p to y, i.e., path that contains arrow to p (the right hand side of above figure)

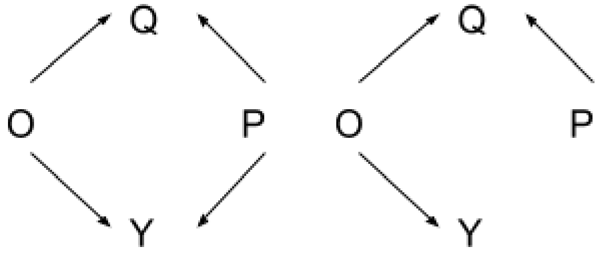
From the definition of block, both O and Q can block this backdoor path. And since O and Q is not descendent of X, either O or Q or both can be V.

The reasoning of why we should conditio on O or Q is exactly the same as the first example.

$$pr(z|x) = \sum_{O} pr(z|x, O)pr(O)$$

$$pr(z|x) = \sum_{Q} pr(z|x, Q)pr(Q)$$

Example 2:



We want to see $pr(p|y)$

The candidates variables for V is still O and Q. according to backdoor criterion, Q cannot be in V since it is a descendant of P. So V={} (no element) can actually block the path since the (3) of the block definition is satisfied. Of course, since O can block the backdoor path (right hand side of above figure) (satisfies the (2) of the block definition), it is also OK to have V= {O}.

The reasoning is that, if we condition on Q, notice that since both O and P can cause the change of Q, then when P changes, in order to make Q constant, O must also changes, therefore in this path the variation of P is correlated to the variation of O, which is correlated to the variation of Y. Therefore we should not condition on Q. according the basic structure, when we to not condition on Q, the variation of P is independent on the variation of O (in this path), thus this channel is excluded.

$pr(z|x) = \sum_{O} pr(z|x, O)pr(O)$

Example 3: (with median variable)

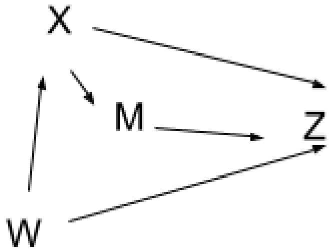
x affects y directly, and indirectly through M is a bit complicated, but still easy to understand.

In summary, if we want to see the **overall effect of x on y**, then still use our rule.

If we want to see the direct effect of x on y, then things are a bit complicated.

Sub-example 1

Suppose that you are dealing with the following example:



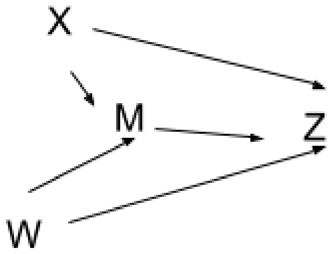
Suppose that you want to estimate the effect of X on y. Here we can see that the M is a median variable of X to Z.

According to our rule, we should condition on w. Should we condition on M?

- If we want to see the overall effect of x on y, then we follow our rule, which states that M, which is a descendant of X, should not be included. Finally the V={w}
- If we want to see the DIRECT effect of x on y, then we can condition on M too. Finally the V={W,M}

This is not difficult. But what about the next example?

Sub-example 2(important)



We can see that there is a variable W that both affects media variable M and the result variable Z.

- If we want to see the total effect of X on z:
- Then We still follow our rules. M, which is a descendant of x, should not be conditioned (in fact, in this setting, another bigger reason is that if we condition on M, according to the basic structure, x and w will be correlated, thus leading to spurious channel.). Since x and w are independent, it implies we do not need any condition variables. V={}
- If we want to see the direct effect of X on z:
- Then here are some problems. Since we want to see the direct effect, we need to condition on M. However, when we condition on the M, x and w becomes correlated, thus leading to spurious channel. How can we deal with this ?
  We can fix the value of x and m at the same time(not condition on!)

$$pr(z|x, M) = \sum_{W} pr(z|x, M, w)pr(w)$$

From the above we know the expression of $pr(z|x, M)$ we can then get **the direct effect of x on z when the M is fixed at a given level m**

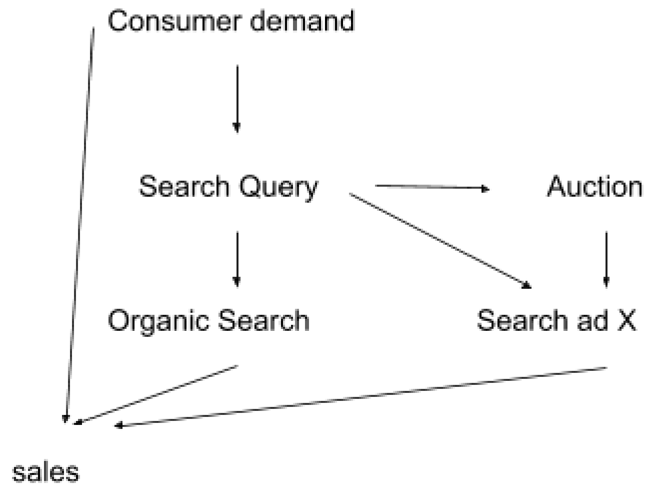
$$pr(z|x', m) - pr(z|x, m)$$

What is the conditions under which we can identify the direct effect of x on y with the presence of m in this way? Intuitively,

There exists a set V1 of variables that blocks all backdoor paths from M to Z.

There exists a set V2 of variables that blocks all backdoor paths from X to Z after deleting all arrows entering M. (The second of these is met automatically given the lack of parents for X)

Example 4: A real situation example

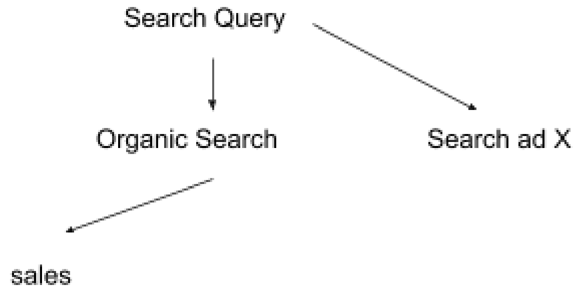


Suppose that we want to see the effect of search ad (X) on sales (Y).

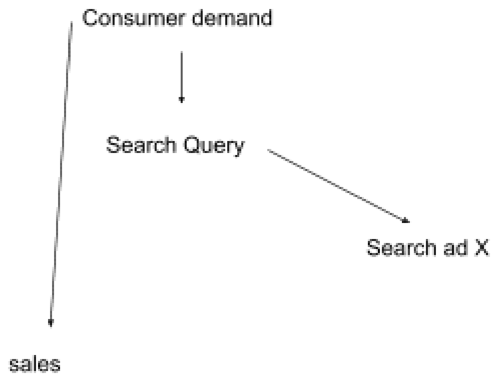
$$Y = \beta X + \varepsilon_0 + \varepsilon_1$$

In which ε_0 is the effect of consumer demand on the sales, and ε_1 is the effect of organic search on sales. It is easy to see from the figure that both error terms are correlated to X, leading to an inconsistent estimation. Therefore we need to find out some covariables to the cut off such correlation.

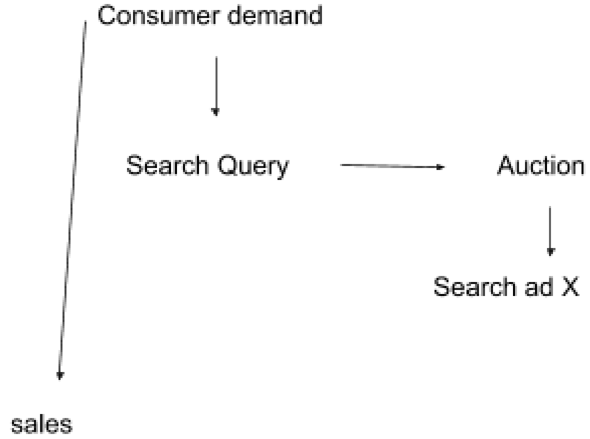
To do this, we first get the backdoor paths.



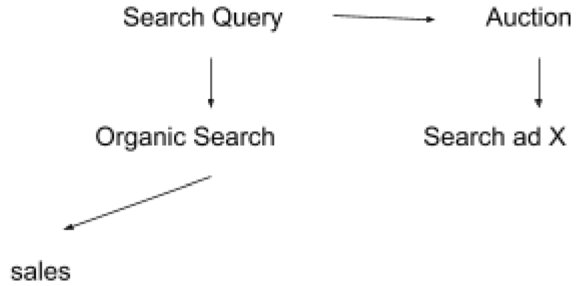
We can use Search Query or Organic Search to block this path.i.e., to make $cov(X, \varepsilon_1 | V) = 0$



We can use Search Query or Consumer Demand to block this path.i.e., to make $cov(X, \varepsilon_0|V) = 0$



We can use the Consumer demand, Search query, or auction to block This back door path, $cov(X, \varepsilon_0|V) = 0$



We can use the organic search, search query, auction to block the backdoor. $cov(X, \varepsilon_1|V) = 0$

The covariates variable set V must be able to block every blackdoor path (or, to block every ε). From the above analysis we find out that for both we **can set V to include only Search Query.**

$E(y|x, V) = \beta X + E(\varepsilon_0| x, V) + E(\varepsilon_1| x, V) = \beta X + E(\varepsilon_0| V) + E(\varepsilon_1| V)$

(Given V, X and $\varepsilon_1(\varepsilon_0)$ are independent)

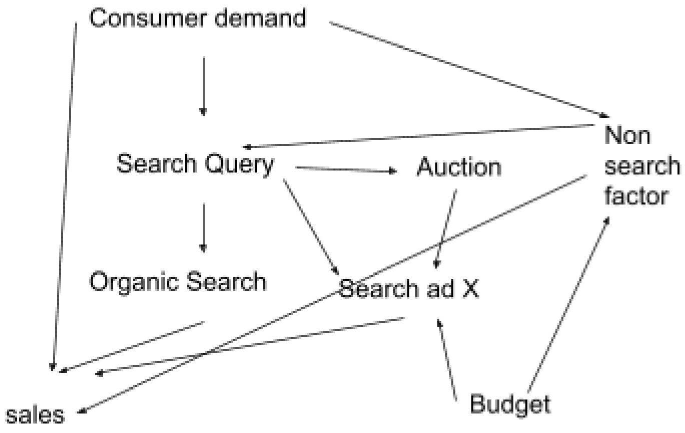
Define $E(\varepsilon_0|\,V\,) + E(\varepsilon_1|\,V\,) \equiv f(V)$, we can then write

$$E(y|x, V) = \beta X \;+\; f(V)$$

This is a typical basic-type semi-parametric model, and we can apply robinson' s method to easily estimate it. After working out $E(y|x, V)$, we can easily compute

$$E(y|x) = E_v[E(y|x, V)]$$

Example 5: An even complicated situation.

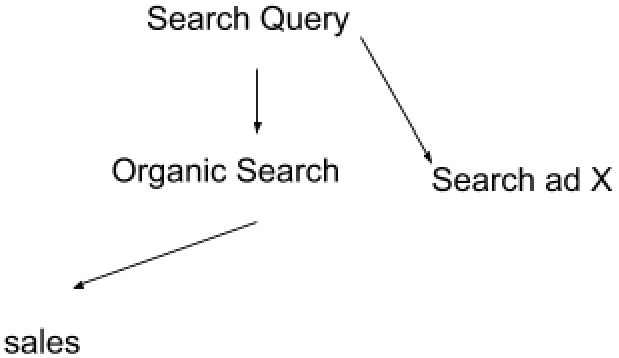


We still want to see the effect of search ad (x) on sales (Y).

$$Y \;=\; \beta X \;+\; \varepsilon_0 + \varepsilon_1 + \varepsilon_2$$

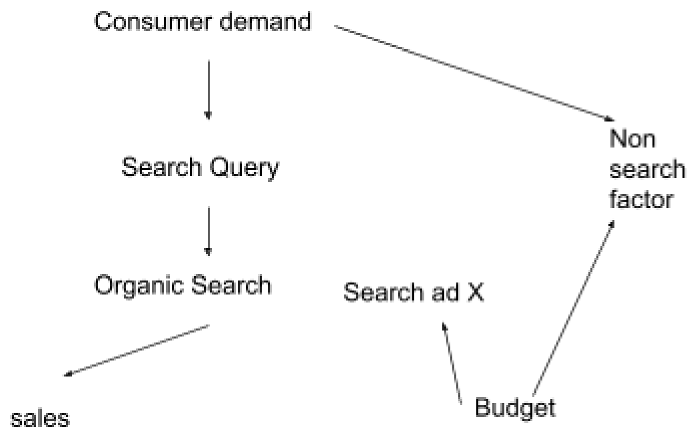
In which ε_0 is the effect of consumer demand on the sales, and ε_1 is the effect of organic search on sales, and ε_2 is the effect of non search factors on sales. It is easy to see from the figure that all error terms are correlated to X, leading to an inconsistent estimation. Therefore we need to find out some covariables to the cut off such correlation.

(1)



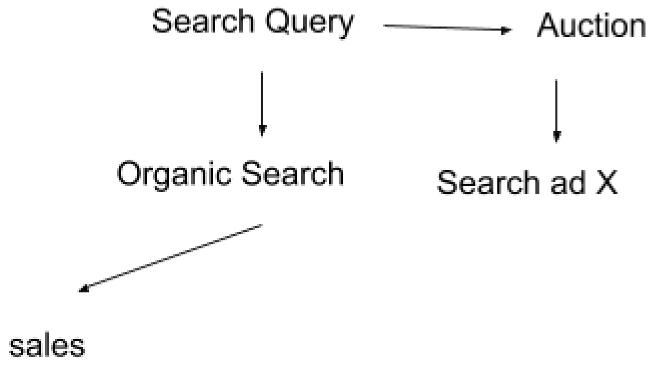
Use Search Query or Organic Search to block ε_1

(2)



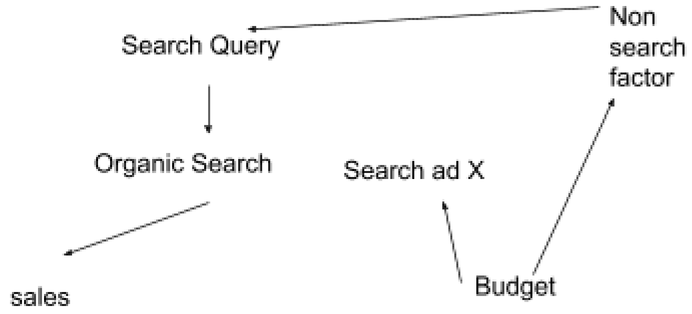
Use Search Query or Organic Search or consumer demand or budget to block ε_1 (Do not use non-search factor ALONE since it is a collider here)

(3)



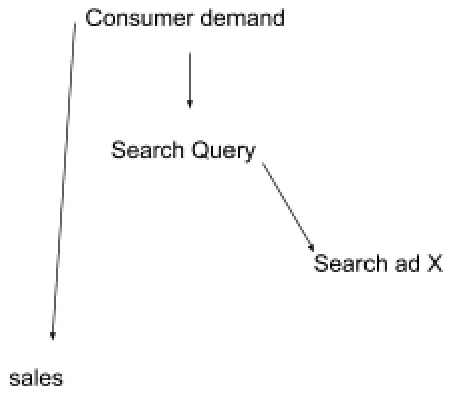
Use Search Query or Organic Search or auction to block ε_1

(4)



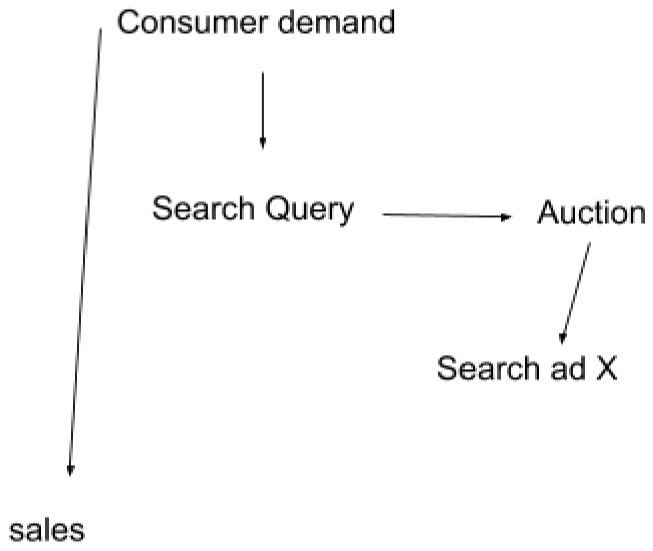
Use organic search or organic search or non search factor or budget to block ε_1

(5)



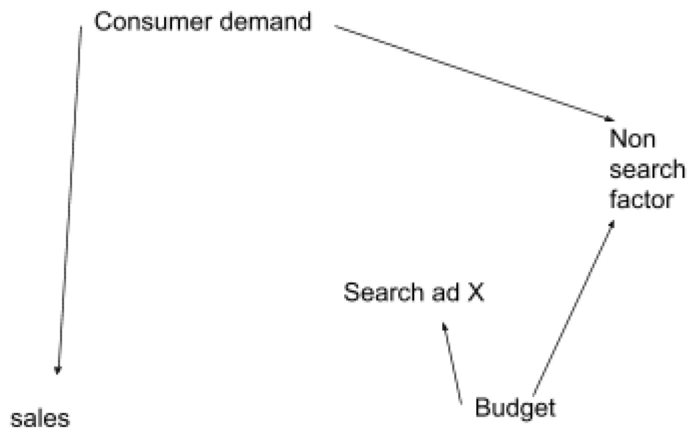
Use Search Query or Consumer demand to block ε_0

(6)

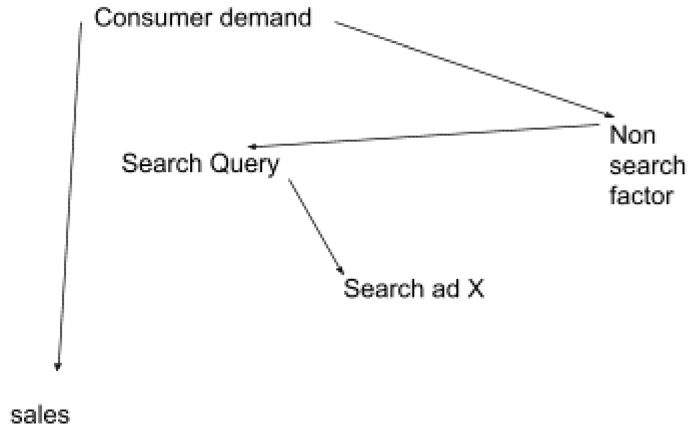


Use Search Query or Consumer demand or auction to block ε_0

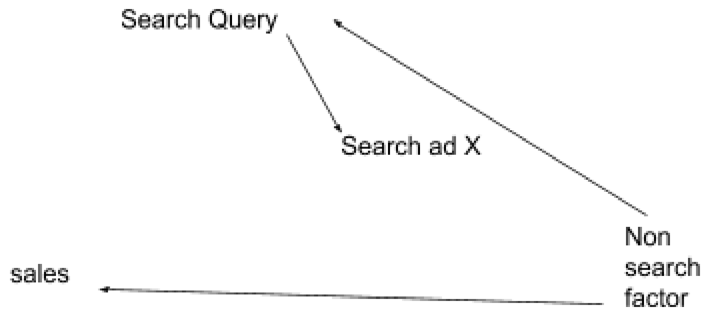
(7)



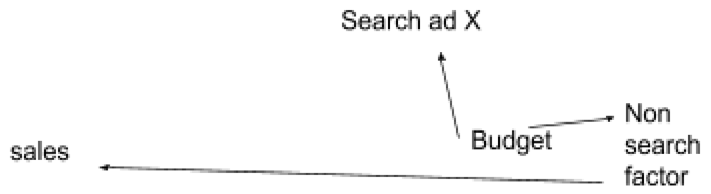
Use budget or Consumer demand or auction or budget to block ε_0. (**Do not use non-search factor ALONE, since it is a collider here)**
(8)



Use search query, consumer demand, or non-search factor to block ε_0.
(9)

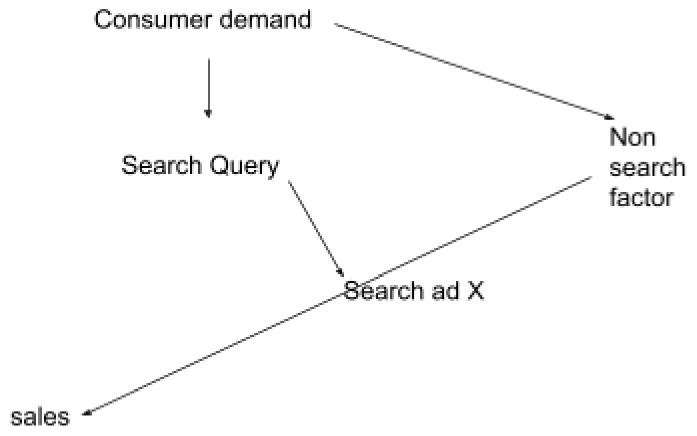


Use Search Query, non search factor to block ε_2
(10)



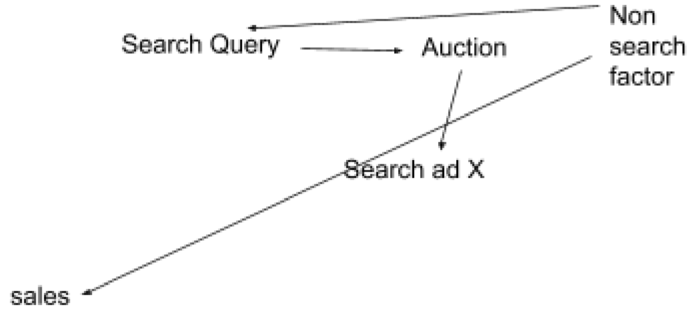
Use budget or non search factor to block ε_2

(11)



Use consumer demand, search query or non search factor to block ε_2

(12)



Use auction, search query or non search factor to block ε_2

Therefore,for ε_1 , we can use use search query to block it .for ε_2 , we can use Non search factor to block all backdoor paths that have it. for $\varepsilon_0,$ we can use search query to block it in (5)(6)(8); in (7) however, since we have a collider (non-search factor), we should not include non-search factor alone. Therefore, using search query alone, we can block the ε_0 in (7). Adding non-search factor, however, will be problematic. We need to add additional variable that can block the path in (7). Either budget or consumer demand will be good. Let's use budget to block

Denote budget as B, search query as V, and other non-search contributors as N.Finally we can write down the following equation:

$$E(y|x, V, N, B) = \beta X + E(\varepsilon_0| x, V, N, B) + E(\varepsilon_1| x, V, N, B) + E(\varepsilon_2| x, V, N, B)$$

$= \beta X + E(\varepsilon_0 | x, V, N, B\ ) + E(\varepsilon_1 | x, V) + E(\varepsilon_2 | x, N)$

The reason is clear: as we argued,for ε_0 , we only need to condition on V and B. For the $E(\varepsilon_1 | x, V, N, B)$, from graph (1)-(4) we know that by conditioning on V, ε_1 is independent of N or B (see (2) and (4), in which V is **between** N(B) and ε_1 . Therefore:

$$E(\varepsilon_1 | x, V, N, B) = E(\varepsilon_1 | x, V)$$

Then same logic stands for ε_2 (in (9)-(12), N lies between B(V) and ε_2). for ε_0 However, things are bit different. Even when you condition on V,B, the N is may not be independent of ε_2 (since N(V) and ε_2 lies on the same side of B in (7)(8)) .

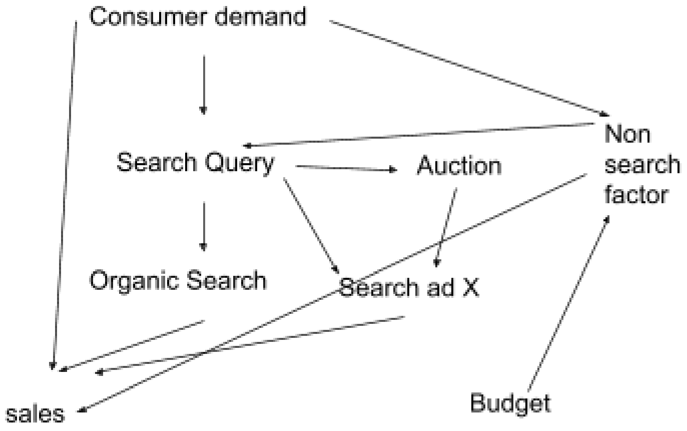
This implies that we can, in fact, do the following estimation process.

$$E(y|x, V, N, B) = \beta X + f(V, N, B) + g(V) + h(N)$$

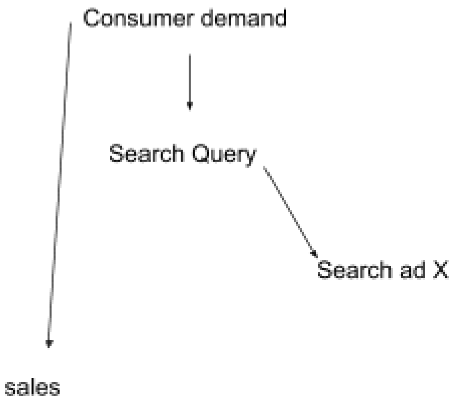
The latter three unknown functions, f,g,h, are non-parametric parts. This is a typical additive model, and we can estimate the beta using the back-fitting algorithm.

Example 6: simpler than E5

We finally practice on a situation which is a bit simpler than E5.

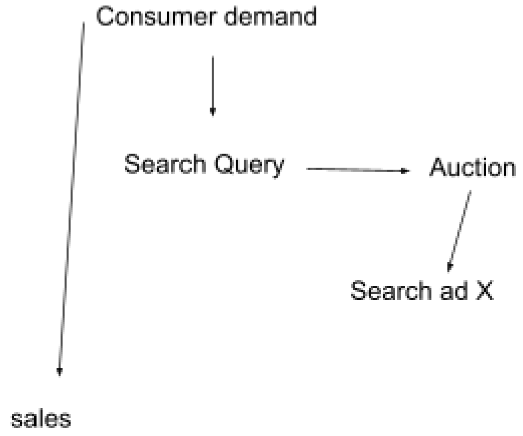


Let’s redo the procedures.
(1)



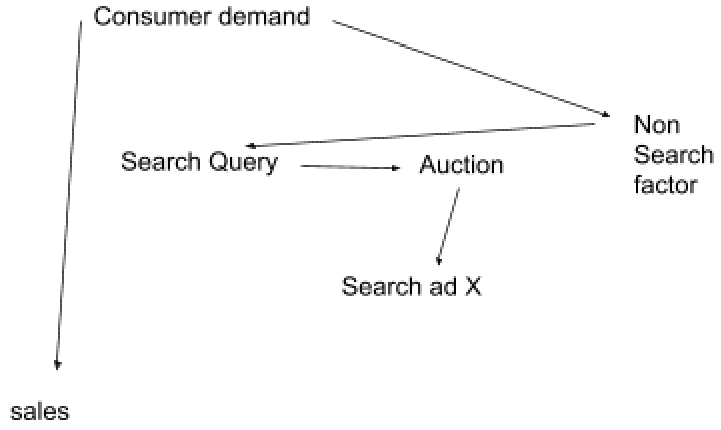
Search query, consumer demand blocks ε_0

(2)



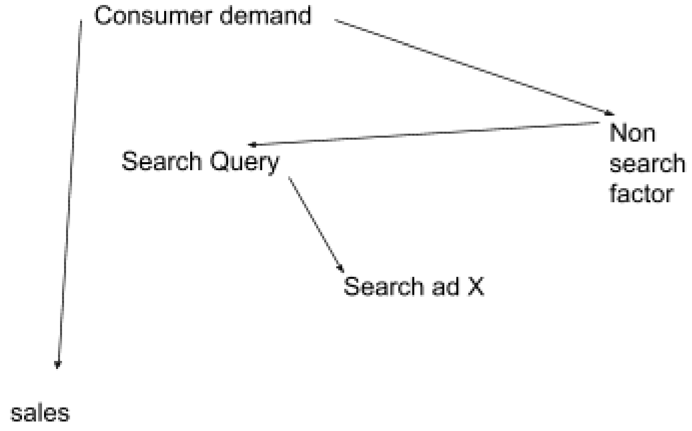
Search query, consumer demand, and auction block ε_0

(3)

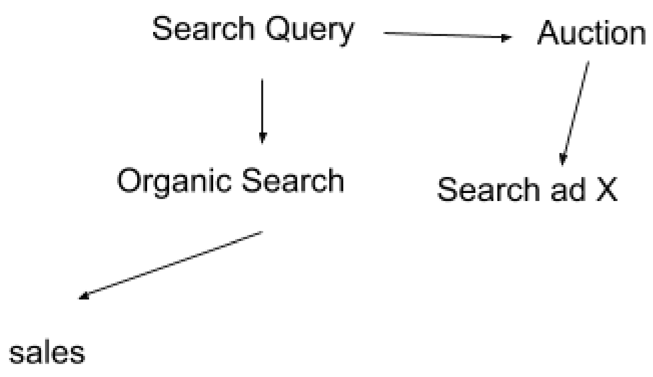


Search query, consumer demand, auction, non-search factors block ε_0

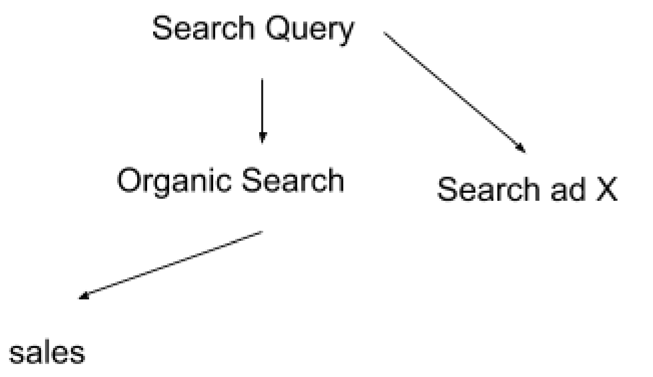
(4)



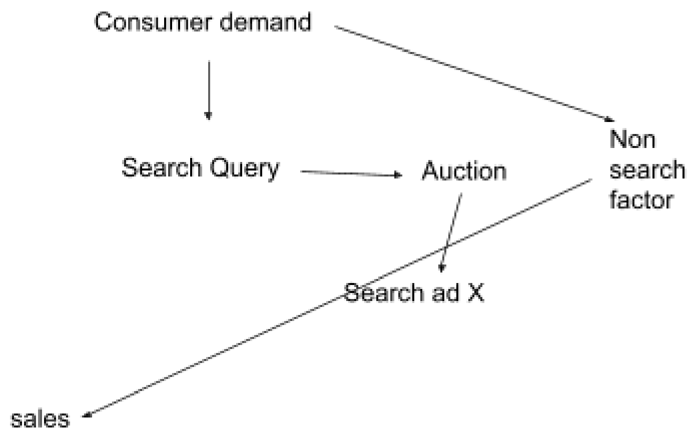
Search query, consumer demand, non-search factor blocks ε_0
(5)



Search query, organic search, and auction block ε_1
(6)

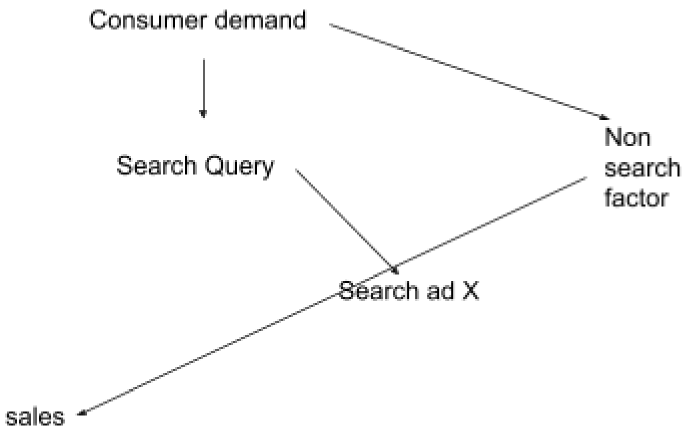


Search query, and organic search block ε_1
(7)



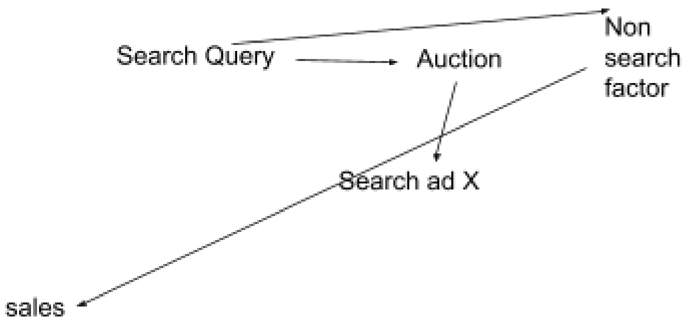
Search query, consumer demand, non-search factor, and auction block ε_2

(8)



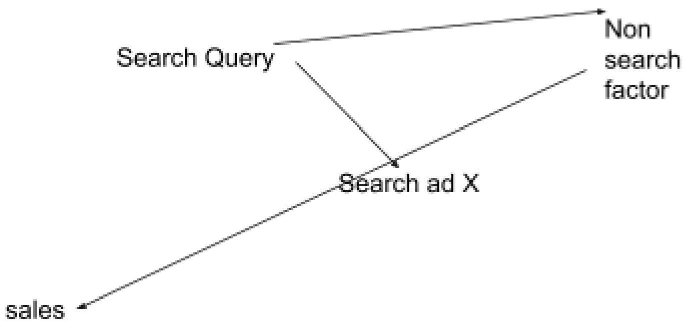
Search query,consumer demand, and non search factor block ε_2

(9)



Search query,auction, and non search factor block ε_2

(10)



Search query,and non search factor block ε_2

It seems that search query alone can block every backdoor path. Therefore we can specify the model exactly the same as in example 4.

Reference:

https://research.google/pubs/pub46861/