Bill Foote

Causal Cats

Contents

Causal Cats

Why cats? First, I am surrounded by four felinoform beings. So I muse, perhaps far too much and too often, what must they be thinking? I am convinced they do complex computations in a mathematical mechanics I do not understand at all, just to jump 2 meters from a corner perch on a table 1 meter lower than a destination on a 5 centimeter wide window sill.

Yes, I anthropomorphize! They are a prime example of "I have no idea what they are up to, but let me hypothesize anyway. . . " Dogs barking for no apparent reason might be the model of why the car alarm goes off in the middle of night for also no apparent reason, incessantly. This is realm of efficient causality, the fine art of understanding the sequence and pattern of how point A goes to B and C at the same time and B goes to D thereafter. I'm sure as I watch my cat jump that she is thinking of the effectiveness and efficiency of achieving her goal of reaching the sill.

Anyway, it begins with cats. Find your own causal totems to guide you. Oh, do we remember Erwin Schrodinger's cat? There, an example from exactly the domain of our concern, uncertainty . . . and a cat.

We do study causation for any number of reasons. First, we want, desire, to know how and why anything at all works, is related to anything else, and especially how and why it matters at all. Second, we study causation because we need to ask intelligent questions about data, further to ask reflective questions about our understanding of data, and then to ask further questions about the rationality of our understanding, and finally, use these answers to formulate and and aid us when we act on our choices.

In this process we learn of and value our successes and failures, identify blind alleys and dead ends, and even get to ask questions of *what if.* This last question wanders into the dialectical realm of *but for* and *if it weren't for* and *unless*, as we fill in the blanks of the counter-factual intervention analysis set up in a causal analysis.

George Gilder says we should waste transistors (that is chips). Gilder makes the fairly obvious point that we must use transistors (lots of them in an integrated circuit) or go out of business. They are ubiquitous. And arrived everywhere in a very short amount of time to boot. If you do not use them you lose control of your cost structure. Anything you build will be too expensive, too heavy, too big, too slow, too lacking in quality. [1]

Michael Schragge builds on Gilder's ironic hyperbole about transistors and analogizes that we should "waste simulations."[2] If we do not *waste* prototyping, rapid development, simulating potential problems, solutions, we will also be liable to misjudge the trajectory of all of our activity. By misjudging the trajectory we mean, ultimately, to take a line of least resistance where we might miss a dead-end, blind alley, or the opportunity to correct a wrong assumption. This further implies we must simulate until we drop! The alternative is that we will miss the one opportunity to improve on that, as yet unkown, one error that throws us off the arc of our plan. Of course the point he makes is that it iss not a *waste*, rather it is the point that we should never shy away from working the problem, simulating the art of the possible, until time and space constraints enforce themselves.

So what is the value added of a prototype, which is simply a working model? It is about information, and information is a surprise, a deviation from a trend. Schragge believes that testing a hypothesis just gets us to the point of saying we seem, in probability that is, to have a trend going on here. In the world of growth, opportunity, error and ignorance, having a trend is barely the beginning of our journey. It is the deviation from the trend that matters.

Schragge quotes Louis Pasteur: "Chance favors the prepared mind." Here the prepared mind is a product of simulations, the rapidly developed prototypes,

[1] See Gilder's comments here:https://gilderpress.com/2020/03/12/investors-should-ignore-materialistic-superstitions/. He goes on to a further idea: build billions of 1-chip interconnected systems (our mobile phones that are really computers) and waste chips that way instead of manufacturing billion chip data centers. According to Moore's law we will eventually get to a near zero-cost chip. We can checkout any retailer: I just saw an offer for 300 assorted transistors for $9.90 or about $0.033 per transistor. Yes, near-zero cost. Yet some might say it took about $50 million to develop a chip in the 40-64nm range.

[2] Here is a taste of Schrage's points of view. He compiled the "wasting prototyping" paradigm into this book a couple of decades ago. We are not at zero-cost computing. However, in 1989 I purchased a laptop for about$1,100, a huge personal investment in those days. Today a far more powerful computer now costs about $700 at a big-box discount store.

Fleming used agar and discovered penicillin – completely unexpected! Dan Bricklin developed the spreadsheet IBMDOS/Apple IIe program Visicalc.[3] As a complete surprise this product was able to be used by millions of people to rapidly simulate other products and services. [4]

What surprises will we discover with our generative models as we waste causal simulations? A working prototype should be a sandbox where everyone is willing to get in and play. It has at least to be durable enough to get to the first slate of useful comments and suggestions for further improvement. Onward![5]

Types of causal patterns

Efficient causality is a syllogism, a chain of reasoning. There are series (a sequence one after another) and parallel paths (simultaneous moves). In the end, there are only three basic causal patterns.

1. **Chains**, similar to a syllogism, really a series, sequence, chain of reasoning. We also call these **pipes**. *Modus ponens* and *Modus tollens* are simple examples of valid logically conclusive chains.
2. **Forks**, leading to an analysis of confounds of causes and effects, often combined in chains. These are the simultaneous moves of a parallel pattern of reasoning.
3. **Colliders** … invert the forking process. Where they exist we need not contend with correcting confounds as we already have the causal solution. These are also the moves of a parallel pattern of reasoning.

Well, let's look at those separately and askance!

Chained melodies

To chain three objects together is to make them dependent upon one another. What we want are unchained causes and effects. Anything getting in the way

[3] Here is a summary of his work at https://en.wikipedia.org/wiki/Dan_Bricklin His innovation with Visicalc was to transform 20 hours of work into 15 minutes, almost of almost play at the time. Visicalc first ran on the Apple IIe. Dan is working on a web-based WikiCalc these days.

[4] Steve Jobs credited Visicalc with the success of the Apple IIe and Macintosh in a WGBH-PBS interview in 1990 at minute 4:27 in https://www.youtube.com/watch?v=iEi8UxEncic He discussed the two explostions in desktop computing, spreadsheets and desktop publishing. He opined on a third explosion, the development of an electronic organization to have clusters of people on a common task across geography and hierarchical organizations. He called this a collaborative model and extends personal computing into interpersonal computing. Now people had a reason to buy the PC.

[5] If you want to read other non-statistical stories, just skip ahead to the **Cassiopeia Effect**. Mythology ever with its grain of truth provides further hermeneutical grist for our causal cats.

will confound our understanding of how a cause becomes an effect. So how do we remove the links in a chain? We get out the bolt-cutters. In probability and statistics linkages are dependencies and dependencies are conditions. We remove or add them to our list of supposed dependencies.

Another term used for a chain is the pipe. We will continue to use the very visual and evocative term change. In this chain of causality literally one thing leads to another. In logical terms the objects in the chain form a line of reasoning about going from an initiating entity to another entity, onto yet another, ultimately landing at a terminus. Each of the intermediate entities are instrumental in getting us to the terminal entity. There is a beginning, leading to a link in the middle, and resulting in an end.

This is such an abstract way to characterize this otherwise simplistic causal chain. Here is a simple chain. Any chain requires at least three entities, e.g., gender, education, wages. We will use X, Y, and Z.

```
library( gganimate )
library( animation )
tidy_ggdag <- dagify(
  z ~ x,
  y ~ z
) %>%
  tidy_dagitty()
p <- ggdag( tidy_ggdag )
```

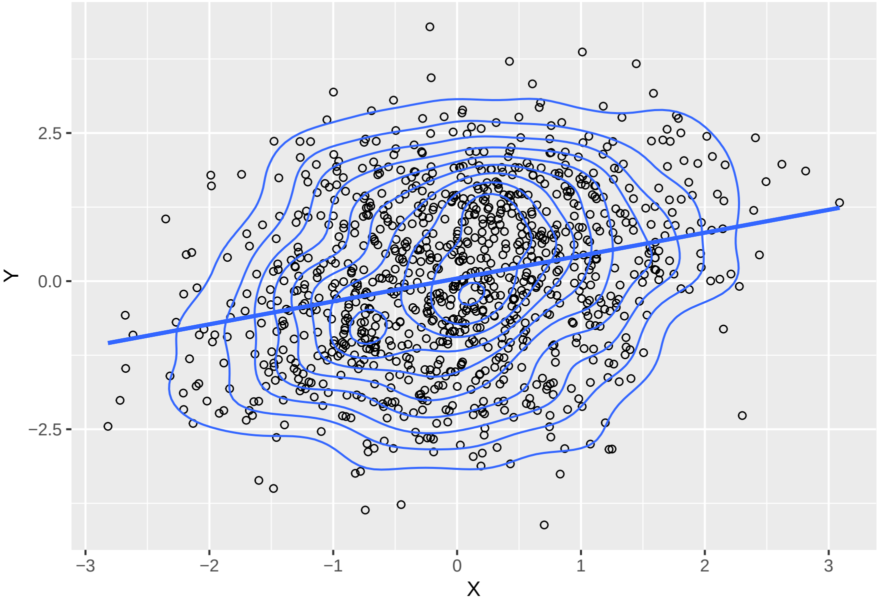
A causal model of the chain is the coupling of a trio of entities linked with two relationships, each with a distribution to back them up to help us understand how the linkages might interact under various conditions. We simulate to stimulate further understanding.

$$
\begin{aligned}
X &\sim \text{Normal}(\mu, \sigma) \qquad (1)\\
Z &\sim \text{Bernoulli}(InverseLogit(Z)) \qquad (2)\\
Y &\sim \text{Normal}(2Z - 1, \sigma) \qquad (3)
\end{aligned}
$$

The Z variable effectively splits the relation between X and Y into two, possibly overlapping, sets. In the R language, the `rnorm()` will draw numbers from the Gaussian, Pearson's Normal, distribution. The `rbern()` is the Bernoulli distribution, a kernel of the binomial distribution. The $InverseLogit$ function maps a continuous value into the $(0, 1)$ interval, mimicking a probability that feeds the *Bernoulli* distribution. We pull the chained entities' outcomes into a tibble and plot the results.

```
N <- 1000
X <- rnorm(N)
```

```
Z <- rbern(N,inv_logit(X))
Y <- rnorm(N,(2*Z-1))
dat <- tibble(
  X = X,
  Z = Z,
  Y = Y
)
dat %>%
  ggplot( aes( X, Y )) +
  geom_point( shape = 1 ) +
  geom_density2d() +
  geom_smooth( method = lm, se = FALSE ) +
  geom_smooth( aes( X, Y ), method = lm, se = FALSE )
```

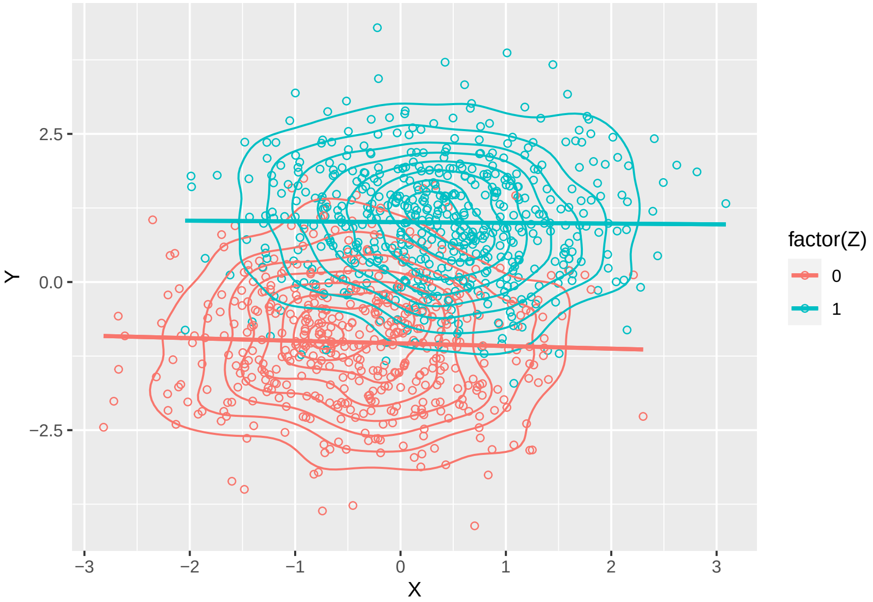


The two-dimensional density plot demonstrates the chain's positive impact of X on Y. There is a concentration of impacts around point $(0, 0)$, very multi-normal indeed, with a sprinkling of outlying, less frequently occuring, variates in the plane surrounding this statistical hill, viewed from above.

This is a classic identification problem. In economic impact studies we often observe the upward sloping *statistical* relationship between price and quantity. We even might be nudged into thinking that if prices rise, people might buy more. This seems to go against common sense, if common sense dictates buy low and sell high, then our statistical estimation is, to put it mildly, off kilter. The relationship instead might be simply confounded.

In the next experiment we plot two ways in which X might statistically influence Y. We have already encoded, through the chain DAG, and in the generative model, the intermediate factor Z as a Bernoulli process resulting in zeros and ones. This time we instruct the plotter to notice the possible differences between $(X, Y) \mid Z = 0$, pairings when $Z = 0$, and $(X, Y) \mid Z = 1$, pairings when $Z = 1$. Color coding the factor Z combinations aids our comprehension here. Regression lines through the tops of these two statistical hills tells the story. In effect we condition the joint determination of X and Y on Z.

```
dat %>%
  ggplot( aes( X, Y, color = factor(Z) )) +
  geom_point( shape = 1 ) +
  geom_density2d() +
  geom_smooth( method = lm, se = FALSE ) +
  geom_smooth( aes( X, Y ), method = lm, se = FALSE )
```



```
# Add linear regression lines
```

No longer is Y positively impacted by X. In fact, Y is negatively impacted by X, *when we account for the differentiator* Z. In the economic impact case of mis-identification, the factor Z, which might represent two different markets, delineates two different and negatively sloped, statistically that is, demand curves. *We have unconfounded the chain by conditioning the relationship between* Y *and* X *with the factor* Z*.*

We have backed into Pearl (2016)'s *Rule 1*:

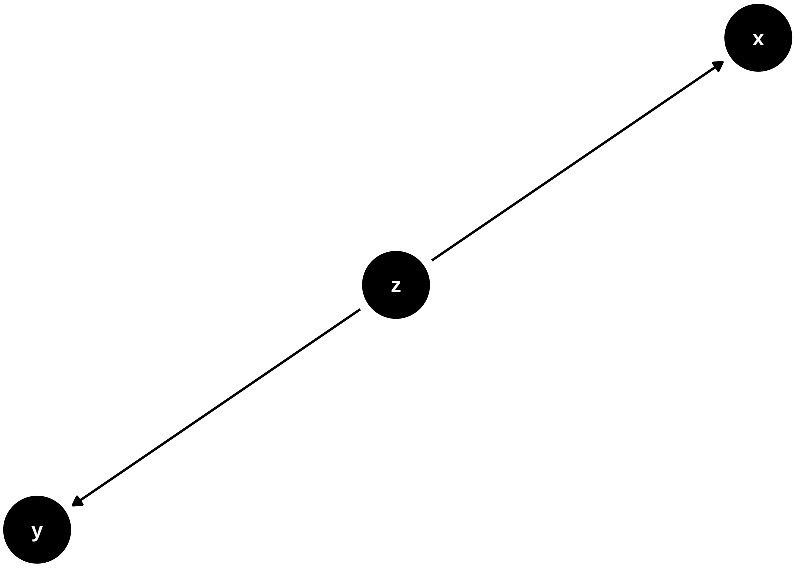
Theorem 1 (Rule 1: Conditional Independence in Chains)**.** *Let Z be a set of variables along the unidirectional path that connects variables X with Y. Then X is independent of Y if X and Y are conditional on Z.*

Forked causes

The fork starts with a fact which can *simultaneously* impact the supposed cause X and effect Y. The adverb *simultaneously* can be thought of in two ways. In one way, in the time domain, the factor acts in the same nano-second, really at exactly the same *time*, on both cause and effect. But another way of thinking about *simultaneity* would have the factor influence the effect sometimes, and at all other times the cause. We simply do not know that an influence occurs, but we do know, or at least opine we know what we know here, that it is plausible that the factor will affect either the cause or the effect. The accumulation of factor's influence on the cause might be less then, equal to, or greater than the factor's influence on the effect. In any case, would everyone agree that the factor truly can confound any thought of a relationship between cause X and effect Y? Let's suppose so.

Here is a graphical representation of the way one variable, Z, would simultaneously share information with two other variables, X and Y. Here there is no back sharing of information from X or Y to Z. If there were we would have quite a different discussion!

```r
tidy_ggdag <- dagify(
  x ~ z,
  y ~ z
) %>%
tidy_dagitty()
ggdag( tidy_ggdag ) +
  theme_dag()
```



Okay, it still looks like a chain. We would just stretch Z to the northwest corner of this drawing to notice the *forkedness* of the diagram, or we could just pretend someone stepped on the fork and flattened it. The point of this extended discussion is that what we see is not what's at stake. What is at stake if that Z influences *both Y and X*, somehow. But our hypothesis is still that X is the cause and Y is the effect.

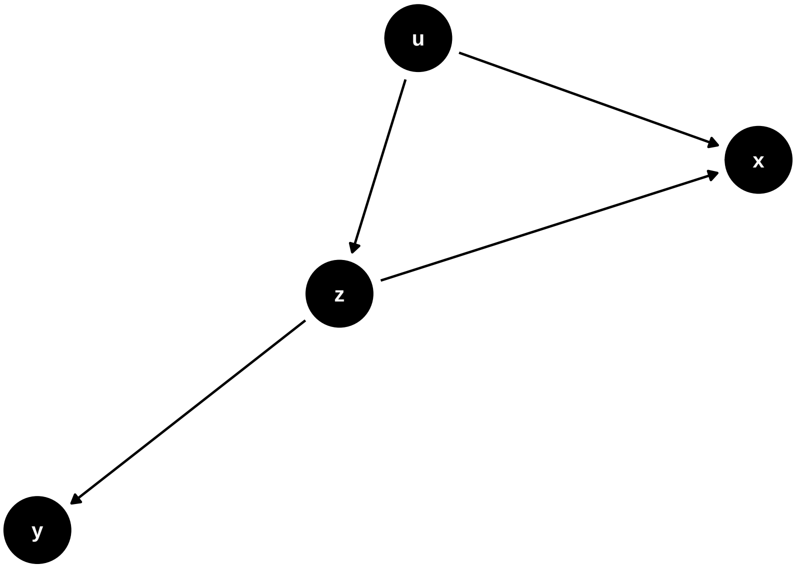
A statistically minded causal model of the fork is this trio of relationships, each with a distribution to back them up.

$$
\begin{aligned}
Z &\sim \text{Bernoulli}(p) && (4)\\
X &\sim \text{Normal}(\text{2Z-1}, \sigma) && (5)\\
Y &\sim \text{Normal}(2Z-1, \sigma) && (6)
\end{aligned}
$$

Fashioned in this way Z again can take on one of two discrete values which might, in our story, represent two different treatments, genders, countries, or status. The status palpably affects the means of the normal distributions both of X and Y. (Merton 1967 the *Matthew Effect.*)

We can complicate this a bit to show how another fork could influence two of the three variables. We might call this interloper U, an unobserved variable.

```
tidy_ggdag <- dagify(
  x ~ z,
  x ~ u,
  z ~ u,
  y ~ z
) %>%
  tidy_dagitty()
ggdag( tidy_ggdag ) +
  theme_dag()
```



In a linear sort of way, we can model U in this slightly amended causal arrangement.

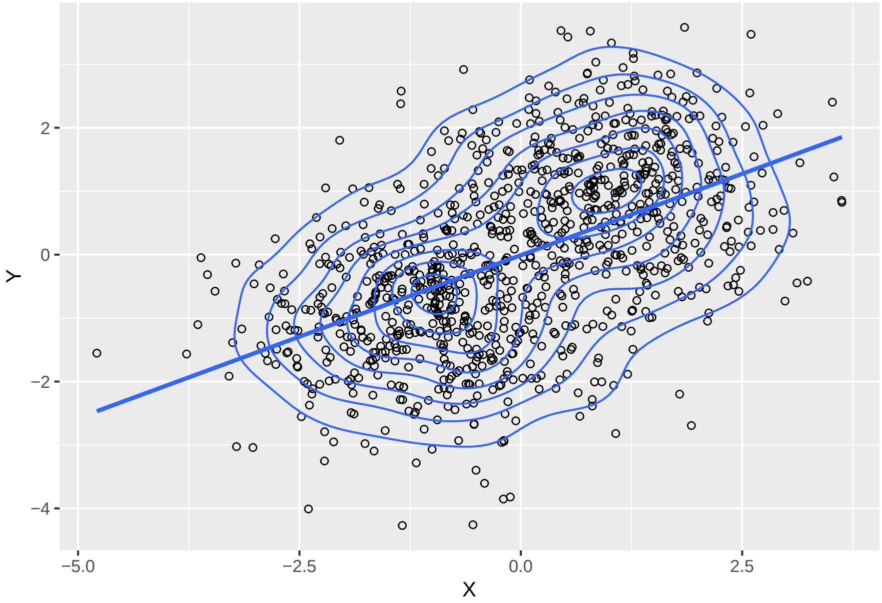
$$\begin{aligned}
U &\sim \text{Normal}(0,1) && (7)\\
Z &\sim \text{Bernoulli}(\text{InverseLogit}(U)) && (8)\\
X &\sim \text{Normal}(2Z-1+U,\sigma) && (9)\\
Y &\sim \text{Normal}(2Z-1,\sigma) && (10)
\end{aligned}$$

Thus we might tell a potential saga of missing variables in a study, or the omission of data in an investigation, or the ignorance of a factor, or a cause, in a chain of reasoning.

In this causal scenario, U simply confounds the generation of X and Z by sometimes sharing more or less information with X and othertimes more or less with Z. U can represent missing observations, model misspecification and selection bias, sometimes all at once. Shall we say it muddies the water a bit more than the simple fork?

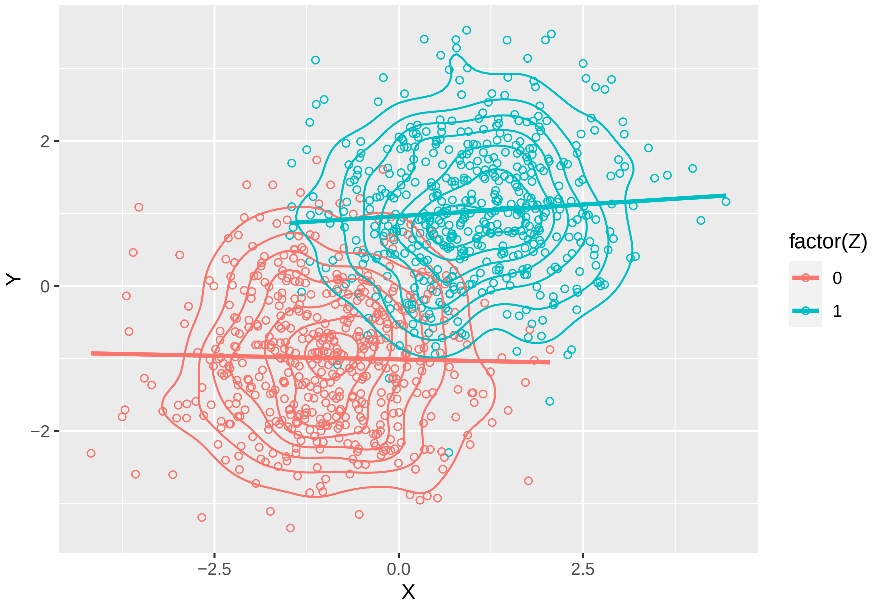
First, let's simulate without the U confounder to create a baseline case of the simple fork's simulated causes and effects. Initially, we will plot only the effect and the cause, as we did with the chain.

```
N <- 1000
Z <- rbern(N)
X <- rnorm(N,2*Z-1)
Y <- rnorm(N,(2*Z-1))
dat <- tibble(
  X = X,
  Z = Z,
  Y = Y
)
dat %>%
  ggplot( aes( X, Y )) +
  geom_point( shape = 1 ) +
  geom_density2d() +
  geom_smooth( method = lm, se = FALSE ) +
  geom_smooth( aes( X, Y ), method = lm, se = FALSE )
```



Clearly, X statistically effects changes in Y. Now let's disaggregate the effects as we did with the chain. Here we plot the impact of the Z factor's contribution to the effect and to the cause.

```
N <- 1000
Z <- rbern(N)
X <- rnorm(N,2*Z-1)
Y <- rnorm(N,(2*Z-1))
dat <- tibble(
  X = X,
  Z = Z,
  Y = Y
)
dat %>%
  ggplot( aes( X, Y, color = factor(Z) )) +
  geom_point( shape = 1 ) +
  geom_density2d() +
  geom_smooth( method = lm, se = FALSE ) +
  geom_smooth( aes( X, Y ), method = lm, se = FALSE )
```

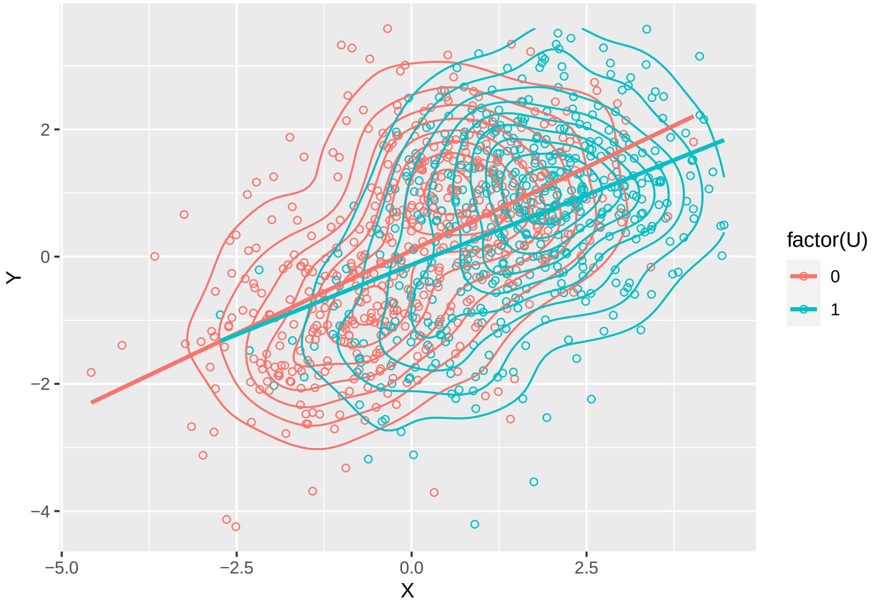


Remarkably we see that both X and Y are not at all related to one another for either value of Z, even when we use the chain technique to differentiate the two cases encoded in Z. This is the confounding effect itself! Factor Z donates information *pari passu* and lickety-split to X and Y.

When we add another confounder U and plot the interaction of X and Y conditional on the status of the U confounder we get this story.

```
N <- 1000
U <- rbern(N)
```

```
Z <- rbern(N, inv_logit(U))
X <- rnorm(N,2*Z-1 + U )
Y <- rnorm(N,(2*Z-1))
dat <- tibble(
  X = X,
  Z = Z,
  Y = Y,
  U = U
)
dat %>%
  ggplot( aes( X, Y, color = factor(U) )) +
  geom_point( shape = 1 ) +
  geom_density2d() +
  geom_smooth( method = lm, se = FALSE ) +
  geom_smooth( aes( X, Y ), method = lm, se = FALSE )
```

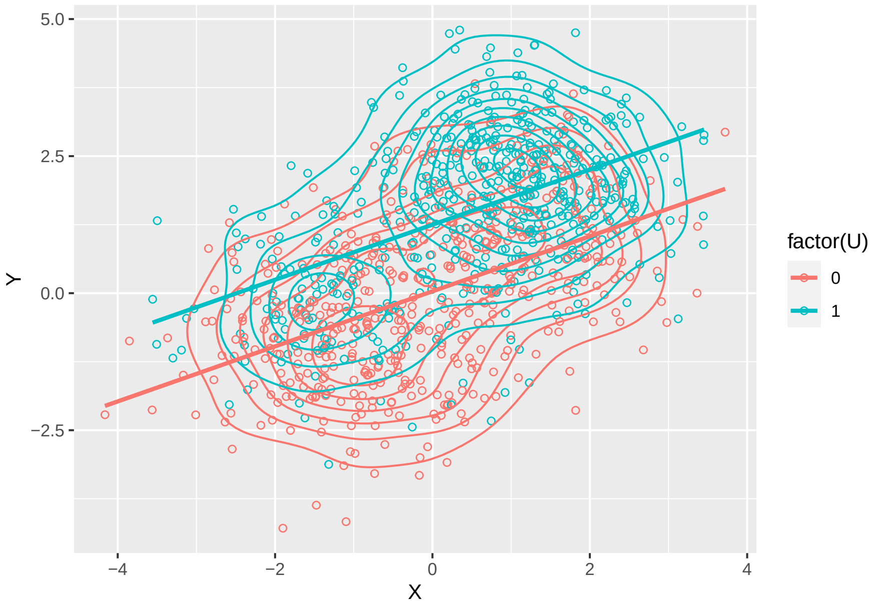


In this panel we see that both X and Z depend on U. Amazingly (but is it?) we get back the same relationship we basically (except for a slight shift) saw with the pure X to Y plot.

What did U do to the confounding influence of Z on both X, cause, and Y, effect? Basically, U shifts the focus of the information flow away from Y to X. It literally loads X with a differentiator which does not impact Y, the effect. It effectually and partially closes the door on the impact of Z on Y.

Just a thought now: what if we partially closed the door on X instead and loaded U into the effect Y?

```
N <- 1000
U <- rbern(N)
Z <- rbern(N, inv_logit(U))
X <- rnorm(N,2*Z-1 )
Y <- rnorm(N,2*Z-1 + U)
dat <- tibble(
  X = X,
  Z = Z,
  Y = Y,
  U = U
)
dat %>%
  ggplot( aes( X, Y, color = factor(U) )) +
  geom_point( shape = 1 ) +
  geom_density2d() +
  geom_smooth( method = lm, se = FALSE ) +
  geom_smooth( aes( X, Y ), method = lm, se = FALSE )
```



This time we magnify the differences in how U influences Y and X. Statistically we can not discern whether or not Y or X are cause or effect at all! All we can do is rise above the statistical statements that one variate is related to another plausibly, and here, positively, as in the generating model.

We are at another rule, the rule to correct, adjust, the relationship between cause and effect due to a fork, a confounder.

Theorem 2 (Rule 2: Conditional Independence in Forks)**.** *Let Z be a common cause of variables X and Y. Let there only be one path between X and Y. Then X and Y are independent conditional on Z.*

Colliding effects

In this third, and final, case two variables independently influence a third, that is, they collide into the third variable. The two causitive variables are independent of one another only in respect of their being immediate causes of the terminal effect.

```
tidy_ggdag <- dagify(
  z ~ x,
  z ~ y
) %>%
  tidy_dagitty()
ggdag( tidy_ggdag ) +
  theme_dag()
```

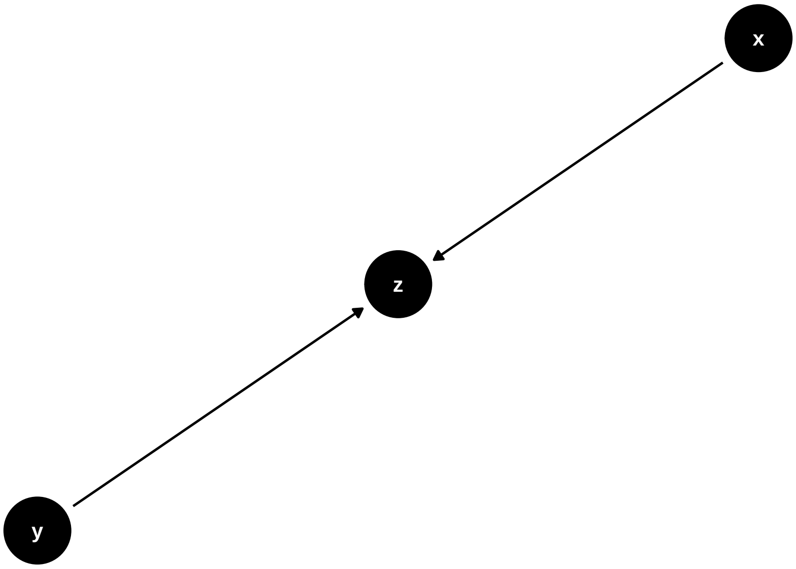


FIGURE 1 Causal model of a collider.

Z is the result of cause X and cause Y, whatever they are. And they are independent of one another in this segment of the story of these three variables.

First here is the depiction of this story of $X \to Y$, that is X causes Y, at least as statistically told. At least we can use this as a baseline.

```
dat %>%
  ggplot( aes( X, Y )) +
  geom_point( shape = 1 ) +
  geom_density2d() +
  geom_smooth( method = lm, se = FALSE )
```

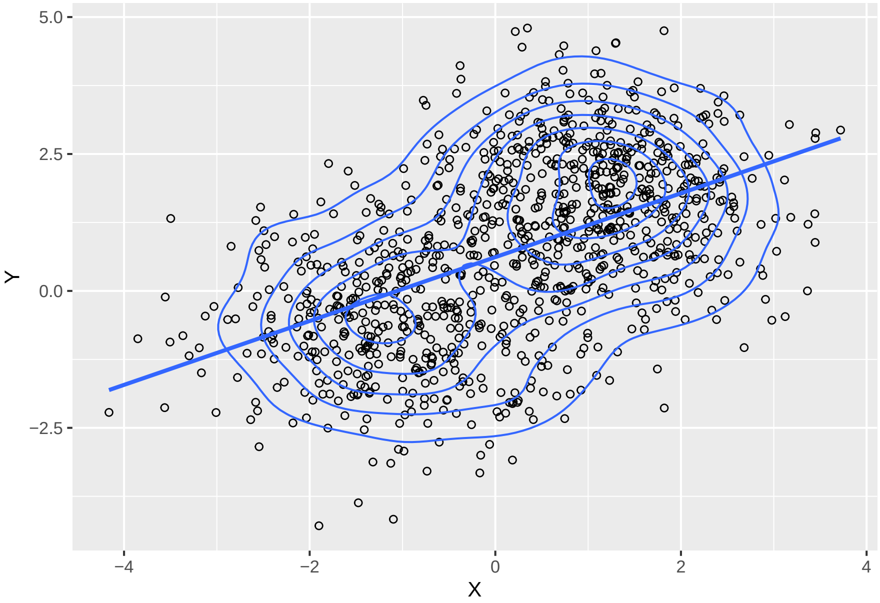


FIGURE 2 Two causes colliding into one effect.

Now, here is one possible rendition of the simple story of X and Y implying Z told through a 1000 iterations. We remember, and reconstruct in our mind, that X would have caused Y in this mind of ours.

```
N <- 1000
X <- rnorm(N)
Y <- rnorm(N)
Z <- rbern(N,inv_logit(2*X+2*Y-2))
dat <- tibble(
  X = X,
  Z = Z,
  Y = Y
)
dat %>%
  ggplot( aes( X, Y, color = factor(Z) )) +
  geom_point( shape = 1 ) +
  geom_density2d() +
  geom_smooth( method = lm, se = FALSE ) +
  geom_smooth( aes( X, Y ), method = lm, se = FALSE )
```

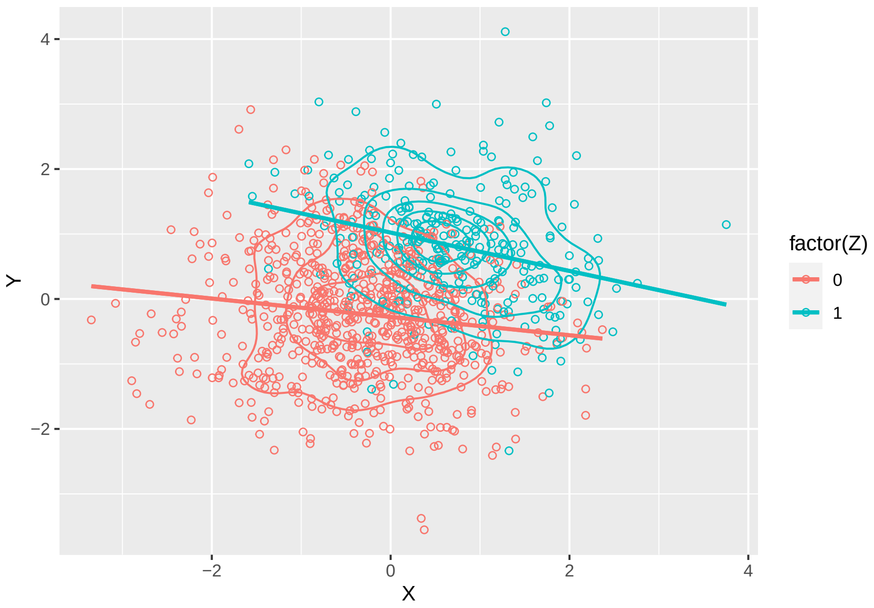


FIGURE 3 Colliding: inverting the fork.

No Z yields the plot of $X \to Y$! Do we not have a third rule to follow from this demonstration?

Theorem 3 (Rule 2: Conditional Independence in Colliders)**.** *Let X and Y separately collide into Z. Then X and Y are unconditional and independent, but are dependent and conditional on Z and descendants of Z.*

Stargazing

The purportedly vain and beautiful Cassiopeia reigned as queen of Aethiopia with her husband King Cepheus. She allegedly committed the sin of hubris by saying that she and her daughter Andromeda were more beautiful than the daughters of the sea god Nereus, known as the Nereids. An enraged Poseidon sent the sea monster Cetus to plague the coasts of Aethiopia. An oracle bade Cepheus and Cassiopeia to sacrifice Andromeda in order to appease Poseidon. They chained her daughter to a rock on the shore as a sacrifice to Cetus. Perseus, her lover, saved her and killed the beast, and married Andromeda.

Poseidon would not be outdone by the mortal Perseus. He still punished Cassiopeia. He tied her to a chair in the heavens forever. She would revolve upside down half of the time as the earth moved through time and the cosmos. The inferential morale of this ancient story is that we should never assume that appeasing our model with multiple sources of causal influences will result in a happy ending of results.

The **Cassiopeia Effect** is the result of our hubris in our models. Also, in the story, no one expected Perseus to arrive on the scene. We might consider as the **Perseus Effect** the amazingly surprising result of an unexpected intervention, also known as a counterfactual condition.

Cassiopeia labors under the stars

Let's investigate labor market gender discrimination to build perhaps a more realistic cause and effect model. Here D is gender discrimination, something not measured, and thus acts as an unobserved variable. We do observe gender with F as female, and some sort of measure of ability, such as years of schooling or scores on a test, with A as ability. We observe labor market outcomes as W for wages.

The theory posits that wages depend on discrimination. The model assumes that gender (F is reporting as female) influences both occupational decisions by labor suppliers and wage determination through discriminatory sorting of labor sources. we observe occupational quality, gender, ability, and wages. We do not directly observe discrimination in the data. Both wages and occupational choices depend on ability. This all makes a lot of theoretical sense. But do any of these forces confuse and confound the path of discrimination to wages?

Let's map our story so far.

```r
tidy_ggdag <- dagify(
  D ~ F,
  O ~ A,
  O ~ F,
  O ~ D,
  W ~ D,
  W ~ O,
  W ~ A
) %>%
  tidy_dagitty()
ggdag( tidy_ggdag ) +
  theme_dag()
```

The shape of this causal map mimics the constellation of Cassiopeia. We might be encouraged by conjecture, and the so-called hubris of Cassiopeia, to regress wages on all of the other influences which collide into wages. However we have two forked confounders, $O \leftarrow F \rightarrow O$ and $O \leftarrow A \rightarrow W$, and one

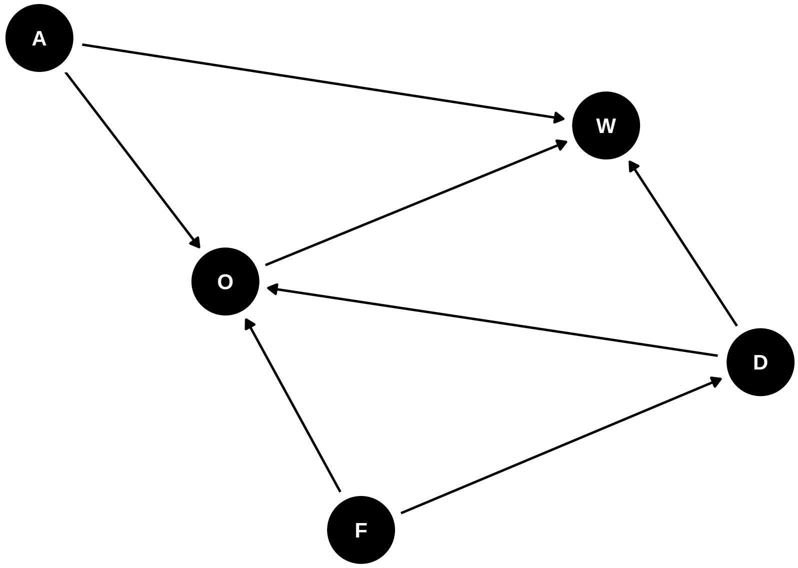


FIGURE 4 Causal impact on wages: occupation and ability impact wage sensitivity to gender.

collider, $F \rightarrow O \leftarrow D$, on the way to wages. These amount to three very sneaky backdoor passes to confusion about how discrimination can cause wages.

A generative model will prod us

We will form and examine generative models to challenge our causal assumption. Here is a generative model of our discrimination analysis. Occupational decisions are plausibly abridged by discrimination, thus a negative relationship. Discrimination also negatively impacts wages through channels of suboptimized labor choices. All other relationships are positive. The influence of being female on occupational choice has been eliminated here since this is one backdoor confounder on the path to wages.[6]

[6] We use what I now consider to be a canonical model of causality introduced by Richard McElreath in a series of lectures all posted here: https://github.com/rmcelreath/causal_salad_2021. Any omissions, reversals, and other blemishes are of my own construction, not Richard's.

$$F \sim \text{Bernoulli}(p) \tag{11}$$
$$D = F \tag{12}$$
$$A \sim \text{Normal}(0,\,1) \tag{13}$$
$$O \sim \text{Normal}(1 + 2A + 0F - 2D, \sigma) \tag{14}$$
$$W \sim \text{Normal}(1 - 1D + 1O + 2A, \sigma) \tag{15}$$

The 1-coefficients indicate a single arrow and, thus influence, the 2-coefficients indicate two arrows on paths into W. For example there are 2 paths to W from A, one directly through O, and one directly not through O. There is only 1 path from D to W, but 2 paths to W if we also include O. There are 4 paths in total each into W and O. Finally, since F is the indicator of D, we might think we can detect D using just F. This should be true, however, is $F \to D \to W$ going to be magnified or diminished by O and A, otherwise known as bias?

```
library(tidyverse)
library(rethinking)
library(stargazer)
#
N <- 100
data <- tibble(
  F = rbern( N ),
  D = F,
  A = rnorm( N ),
  O = 1 + 2*A + 0*F - 2*D + rnorm( N ),
  W = 1 - 1*D + 1*O + 2*A + rnorm( N )
)
#
lm_1 <- lm(W ~ F, data)
lm_2 <- lm(W ~ F + O, data)
lm_3 <- lm(W ~ F + O + A, data)
#
stargazer(lm_1,lm_2,lm_3, type = "text", column.labels = c("Biased Unconditional",
```

```
##
## =======================================================================================
##                                              Dependent variable:
##                     -------------------------------------------------------------------
##                                                      W
##                     Biased Unconditional        Biased           Unbiased Conditional
##                             (1)                    (2)                     (3)
## ---------------------------------------------------------------------------------------
## F                         -2.225**               0.568*                 -0.952***
##                           (0.876)                (0.290)                 (0.284)
##
## O                                                1.890***                1.045***
##                                                   (0.063)                 (0.110)
##
## A                                                                        1.992***
##                                                                           (0.235)
##
## Constant                  1.496**                 0.302                  0.942***
##                           (0.619)                (0.199)                 (0.169)
##
## -----------------------------------------------------------------------------------------
## Observations                100                    100                     100
## R2                         0.062                  0.908                   0.948
## Adjusted R2                0.052                  0.907                   0.946
## Residual Std. Error   4.380 (df = 98)        1.375 (df = 97)         1.045 (df = 96)
## F Statistic         6.452** (df = 1; 98) 481.333*** (df = 2; 97) 579.775*** (df = 3; 96
## =========================================================================================
## Note:                                                        *p<0.1; **p<0.05; ***p<0.0
```

The first model would have us believe that F influences W more than 3 times the impact than what is is in the data. This model is definitely biased. It is also the right sign so we may have appeased our theory, but over-emphasized the role of discrimination in the presence of other influences such as occupational quality and ability.

The second model conditions both F and O so that W seems to increase with F and O. The sign of O is correct, higher-quality occupations determine higher wages. The sign of F is not the direction implied by the data, females have a negative effect on wages in the data. This bias comes from the netting of negative discrimination and positive ability through occupational quality on to wages.

A third model has the correct sign and size of influence of $F \rightarrow D \rightarrow W$ discrimination on wages. This model also correctly represents, within sampling error, the size and sign of O and A on W.

That was the conventional OLS approach. Here are the fully probabilistic models. The coefficients are carbon copies of the OLS models. These estimations will allow us to assess the relative plausibilities of each model as well as their informativeness.

```
m_1 <- quap( alist(
  W ~ dnorm( muW, sigma ),
  muW <- aW + bWF*F,
  aW ~ dnorm( 0, 1),
  c( bWF ) ~ dnorm( 0, 1),
  sigma ~ dexp( 1 )
), data = data )
precis( m_1 )
```

```
##              mean        sd        5.5%      94.5%
## aW      0.7937566 0.4763371  0.03247786  1.5550353
## bWF    -1.1134694 0.6282024 -2.11745823 -0.1094806
## sigma   4.2835664 0.2961956  3.81018859  4.7569441
```

Now for the second model.

```
m_2 <- quap( alist(
  W ~ dnorm( muW, sigma ),
  muW <- aW + bWF*F + bWO*O,
  aW ~ dnorm( 0, 1),
  c( bWF, bWO) ~ dnorm( 0, 1),
  sigma ~ dexp( 1 )
), data = data )
precis( m_2 )
```

```
##             mean         sd       5.5%     94.5%
## aW     0.3156304 0.18708417 0.01663374 0.6146270
## bWF    0.5276326 0.27095555 0.09459328 0.9606719
## bWO    1.8801505 0.06134658 1.78210682 1.9781942
## sigma  1.3456439 0.09423574 1.19503694 1.4962508
```

```
#
```

Here is the third model.

```
m_3 <- quap(
  alist(
    W ~ dnorm( muW, sigma ),
    muW <- aW + bWF*F + bWO*O + bWA*A,
    aW ~ dnorm( 0, 1),
    c( bWF, bWO, bWA ) ~ dnorm( 0, 1),
    sigma ~ dexp( 1 )
), data = data )
precis( m_3 )
```

```
##              mean         sd       5.5%      94.5%
## aW      0.8672653 0.15840856  0.6140979  1.1204328
## bWF    -0.8061240 0.26305484 -1.2265365 -0.3857116
## bWO     1.0977591 0.10337475  0.9325463  1.2629719
## bWA     1.8714922 0.21955869  1.5205950  2.2223894
## sigma   1.0204385 0.07183052  0.9056394  1.1352375
```

With each of these models we then pull samples of each of the impact parameters (all unobserved data) to compare and contrast their locations and shapes.

```
library(ggridges)

without_O <- extract.samples( m_1 )$bWF #[ ,1 ]
with_O <- extract.samples( m_2 )$bWF #[ ,2 ]
with_O_A <- extract.samples( m_3 )$bWF #[ ,2 ]

plot_extract <- tibble(
  sim =1:10000,
  without_O,
  with_O,
  with_O_A) %>%
  pivot_longer(
    -sim,
    names_to = "model",
    values_to = "bWF")

plot_extract %>%
  ggplot( aes( bWF, model) ) +
    geom_density_ridges() +
    scale_fill_viridis_d() +
    theme(legend.position = "bottom") +
    labs(title = "Predicted variation of slope sensitivity",
         subtitle = "Occupation and Ability impact Wage sensitivity to gender.")
```

Now for the variability estimated for each model. This parameter is analogous to the standard error of the OLS regression.

```
library(ggridges)
without_O <- extract.samples( m_1 )$sigma #[ ,1 ]
with_O <- extract.samples( m_2 )$sigma #[ ,2 ]
with_O_A <- extract.samples( m_3 )$sigma #[ ,2 ]

plot_extract <- tibble(
  sim =1:10000,
  without_O,
  with_O,
  with_O_A) %>%
  pivot_longer(
    -sim,
    names_to = "model",
    values_to = "sigma")

plot_extract %>%
```

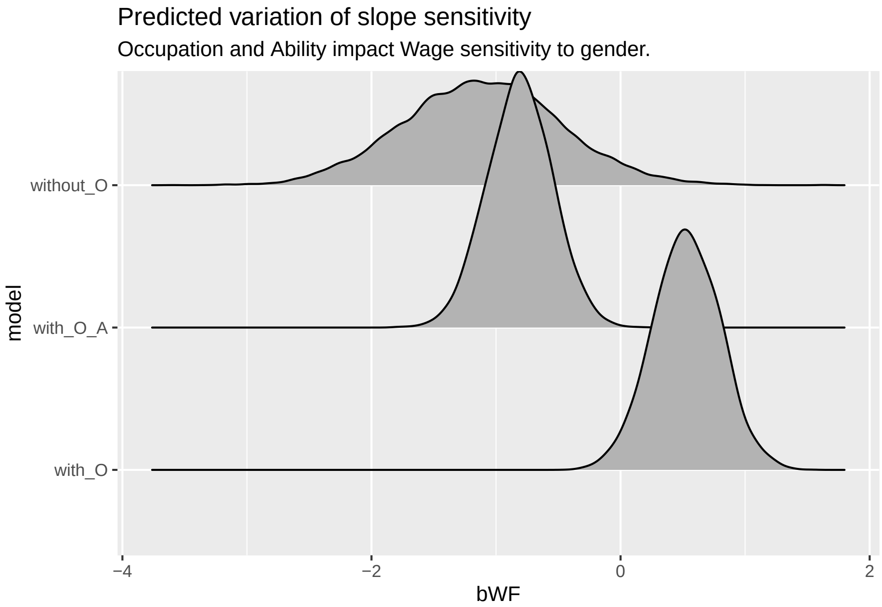


FIGURE 5 Predicted impact on wages: occupation and ability impact Wage sensitivity to gender.

```
ggplot( aes( sigma, model) ) +
  geom_density_ridges() +
  scale_fill_viridis_d() +
  theme(legend.position = "bottom")
```

These probabilistic models point to differences missed by the OLS. The σ credibility intervals not only get successively narrower in range, but also the size of σ gets appreciably smaller as we add ability, A, to the mix of variables. We might be tempted to use the kitchen sink of all variables to improve the overall explanatory power of the model.

Some regression diagnostics are in order, even with our generative data. One more plot for another comparison: Pareto Smoothed Importance Sampling (PSIS) calculated pointwise will help us identify outlier differences among models. PSIS exploits the distribution of potentially outlying and influential observations using the Generalized Pareto distribution to model and measure the data point-wise with the shape parameter $k = \xi$. Any point with $k > 0.5$ will have infinite variance and thus contribute to a concentration of points in an ever-thickening tail of the distribution of uncertainty about the relationships we are modeling.

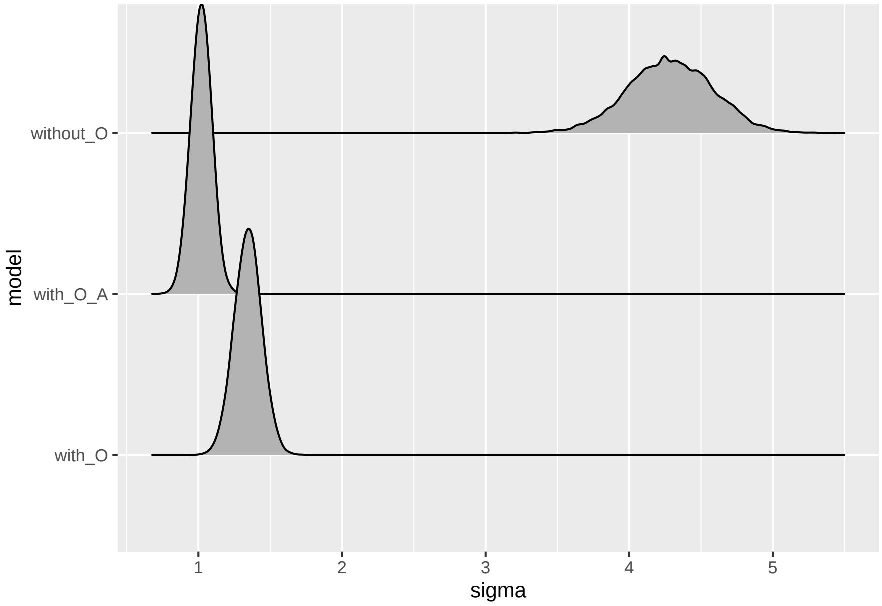


FIGURE 6 Predicted variation of wages: accupation and ability impact wage sensitivity to gender.

```
## R code 7.34 McElreath2020
#library( plotly )
options( digits=2, scipen=999999)
d <- data
set.seed(4284)
PSIS_m3 <- PSIS( m_3, pointwise=TRUE )
PSIS_m3 <- cbind( PSIS_m3, wage=d$W, occupation=d$O, ability=d$A, gender=d$F)
set.seed(4284)
#WAIC_m2.2 <- WAIC(m2.2,pointwise=TRUE)
p1 <- PSIS_m3 |>
  ggplot( aes( x=penalty, y=k, group=wage, color=gender )) +
  geom_point( shape=21 ) +
  xlab("PSIS Pareto k") +
  ylab("PSIS penalty") +
  geom_vline( xintercept = 0.03, linetype = "dashed")
p1
```

The cautionary tale here is that we also need to consider the three potential biases: sign and size of the influence of a proposed variable on outcomes as well as the role and influence of outliers.

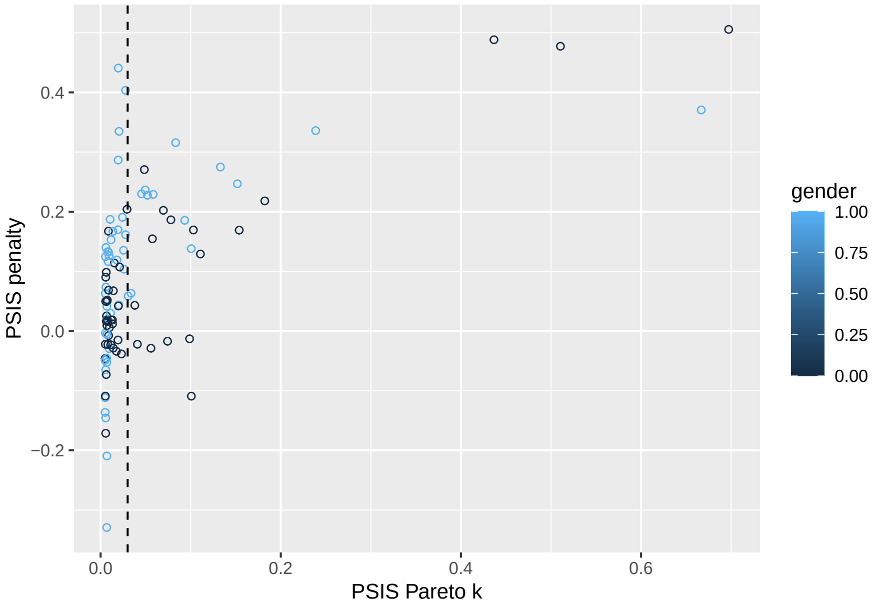


FIGURE 7 Highly influential observations and out-of-sample prediction. Male observations inhabit the NE quadrant with high penalty and Pareto k values. These observations are highly unpredictable.

References

Merton, R. K. 1967. *On Theoretical Sociology*. A Free Press Paperback. Sociology, pt. 1. Free Press. https://books.google.com/books?id=woFqAAAAMAAJ.

Pearl, Judea. 2016. *Causal Inference: A Primer*. Wiley.