Statistics for Linguistics 2021-06-01

Adam Tallman

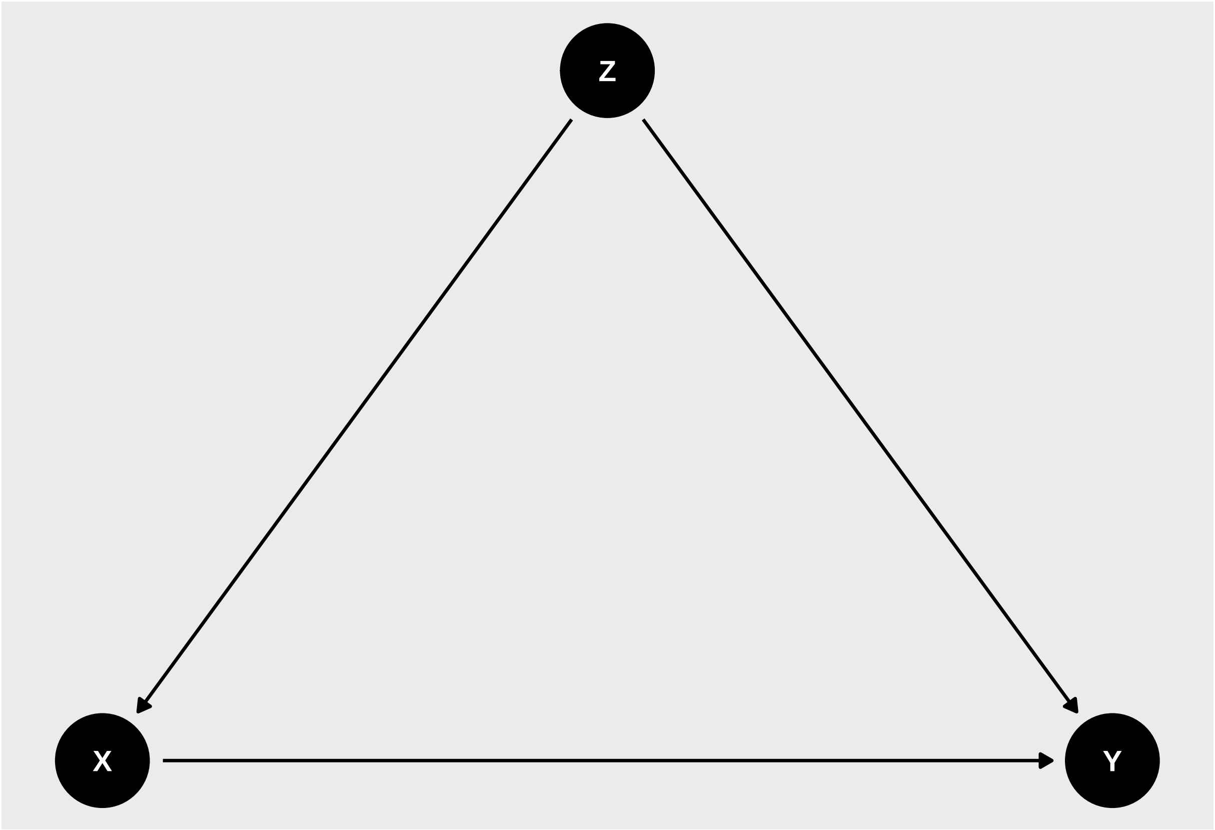
31/05/2021

Last class we talked about multiple regression, causal inference, model selection procedures and interactions. Today we will introduce the notion of a generalized linear model, logistic regression and continue with a few points about causal inference, including Simpson's paradox.

Causal inference overview

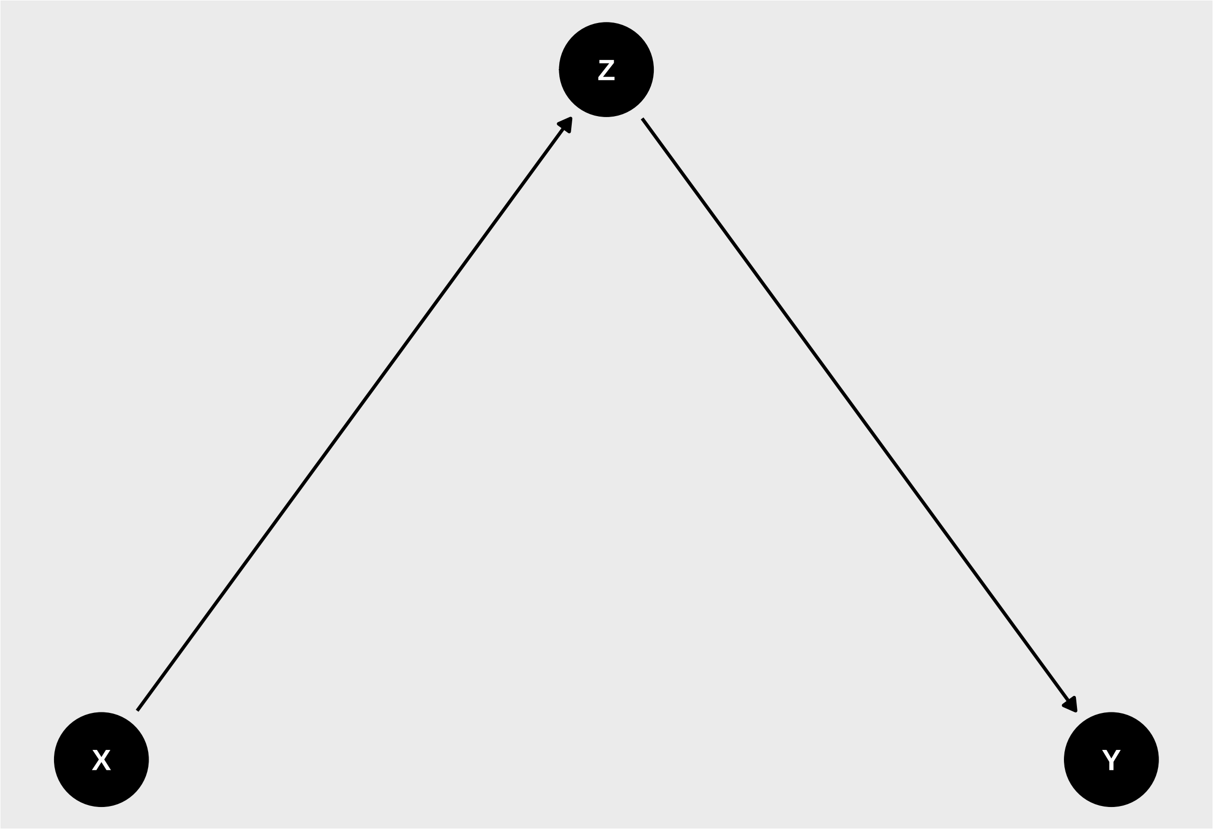
One of the important functions of multiple regression is to help us infer whether certain correlations between variables are spurious or not. Causal inference consists partially involves a set of graphical techniques for visualizing confounds. Last lecture we considered the The Fork and The Pipe. The Thor Fork is when you want to figure out the influence of X on Y, but a variable Z causes both of these variables.

```
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.5

##
## Attaching package: 'ggdag'

## The following object is masked from 'package:stats':
##
##     filter
```

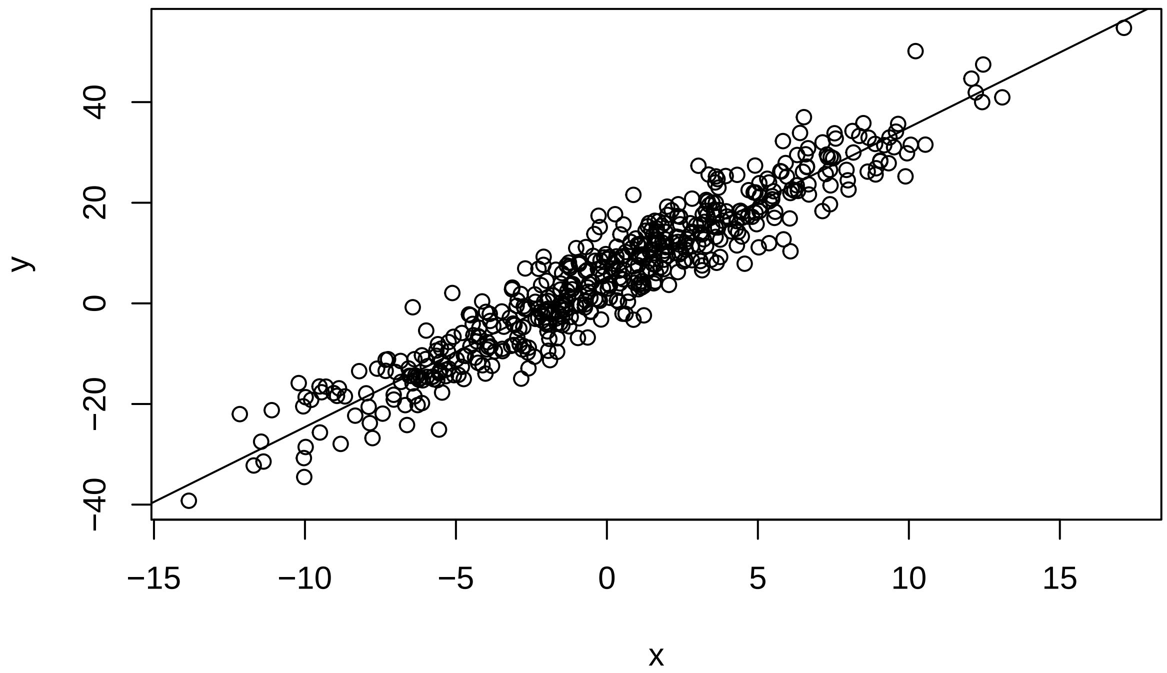


Z opens up a non-causal path between X and Y. You remove Z by conditioning on it, which in terms of a multiple regression translates into including Z in the model.

Last lecture we also considered a Pipe. A piped relationship is where you are interested in inferring the influence of X on Y, and there is another variable Z which is causally influenced by X and causally influences Y.



If this is your causal model, then you should not condition on Z, because this will block the causal relationship between the two variables. The relationship between X and Y mediated by Z is a causal relationship, in contrast to what we see with the fork about. To illustrate simulate the following data based on the causal relationship stated in the graph above.



The effect of x on y is strong as can be seen from the results of the following model.

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7217  -3.2330  -0.2265   3.2867  14.4964
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.19757    0.21740   23.91   <2e-16 ***
## x            2.97782    0.04433   67.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.859 on 498 degrees of freedom
## Multiple R-squared:  0.9006, Adjusted R-squared:  0.9004
## F-statistic:  4512 on 1 and 498 DF,  p-value: < 2.2e-16
```

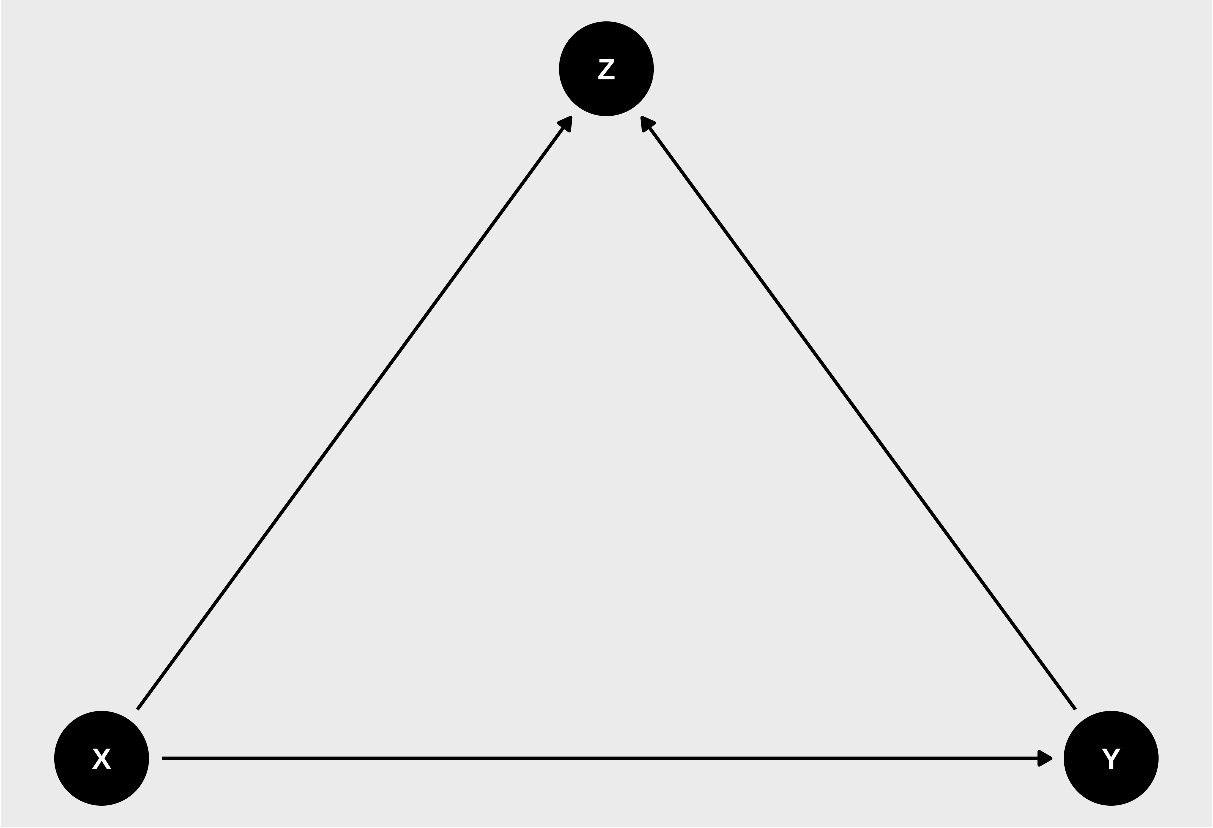
If you add the Z variable, the relationship between X and Y dissappears.

```
##
## Call:
## lm(formula = y ~ x + z)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.3977 -2.6223  0.0958  2.4927 10.7845
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.017368   0.253812   7.948 1.28e-14 ***
## x           -0.005042   0.177985  -0.028    0.977
## z            1.488625   0.087069  17.097  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.86 on 497 degrees of freedom
## Multiple R-squared:  0.9374, Adjusted R-squared:  0.9372
## F-statistic:  3722 on 2 and 497 DF,  p-value: < 2.2e-16
```

But it is wrong to conclude that X and Y do not bear a causal relationship with each other - it is just that the causal relationship is not a direct one. If you are interested in the effect that changing X will have on a change in Y, then you shouldn't condition on Z. The point is that how you interpret a statistical model, depends on the causal model you have in mind.

Another type of confound is referred to as a collider effect. In certain respects this is the mirror image of the fork. You are interested in the causal effect of X on Y, but both of these cause Z.

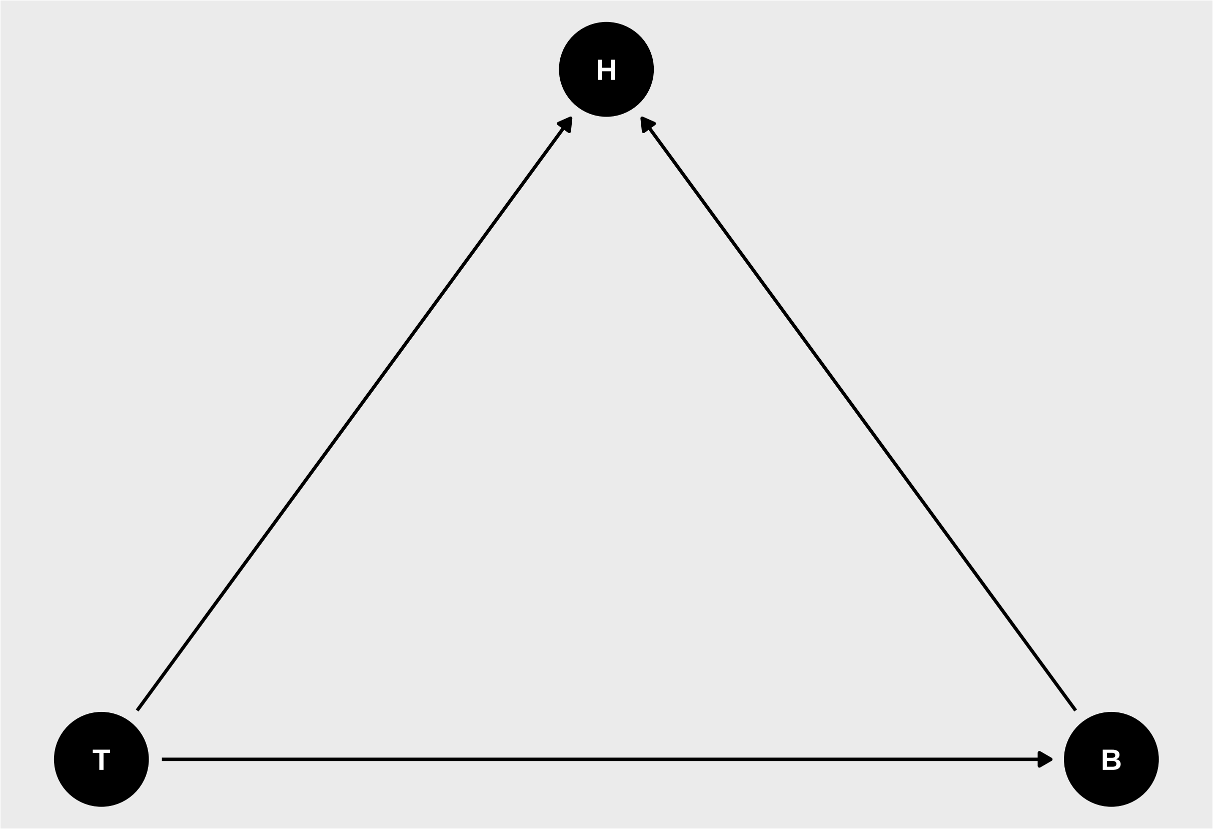


Let's say you are interested in the causal relationship between two covid symptoms. Let's say you think that if you lose your taste, you are more likely (or less likely to have backpain).

Now consider the effect that having a loss in taste sensation and/or backpain could have on whether you end up in the hospital.

```
##
## Call:
## lm(formula = backpain ~ notaste + h)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67693 -0.52495 -0.02099  0.56405  1.77430
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.19566    0.06531  -2.996  0.00309 **
## notaste     -0.37491    0.06761  -5.545 9.38e-08 ***
## hTRUE        1.60768    0.16296   9.865  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8245 on 197 degrees of freedom
## Multiple R-squared:  0.3343, Adjusted R-squared:  0.3275
## F-statistic: 49.47 on 2 and 197 DF,  p-value: < 2.2e-16
```

The model suggests that there is a negative correlation between back pain and having no taste as if having no taste relieves back pain. But this is a spurious correlation based on **adding** a factor to our model. On conditioning on a collider.



Generalized linear model

The function that we used lm() using a method called ordinary least squares regression. A generalized linear model is an extension of OLS. In generalized linear models, parameters are not chosen based on minimizing the sum of squares errors. Generalized linear models use maximum likelihood estimation to find the parameters of the model. MLE is an iterated fitting algorithm.

To get an idea of what MLE is doing, the first starting point is understanding the distinction between probability and likelihood. Probability refers to

$$pr(data|distribution)$$

$$L(distribution|data)$$

Logistic regression

Logistic regression is used when the relationship between

$$logit(y) = b_0 + b_1x_1 + b_2x_2...$$

logit() stands for the log-odds. Odds are calculated as follows

$$logit(p) = log\frac{p}{1-p}$$

$$2^x = 16$$

$$log_2(16) = x$$

log() in R gives you the natural log where the base is e (2.71828. . . .)

$$P\{y = 1|x\} = \frac{1}{1 + exp(-x\beta)}$$

$$Prob\{y = 1|x\} = \frac{1}{1 + exp(-x\beta)}$$