Statistics for Linguistics

Adam J.R. Tallman

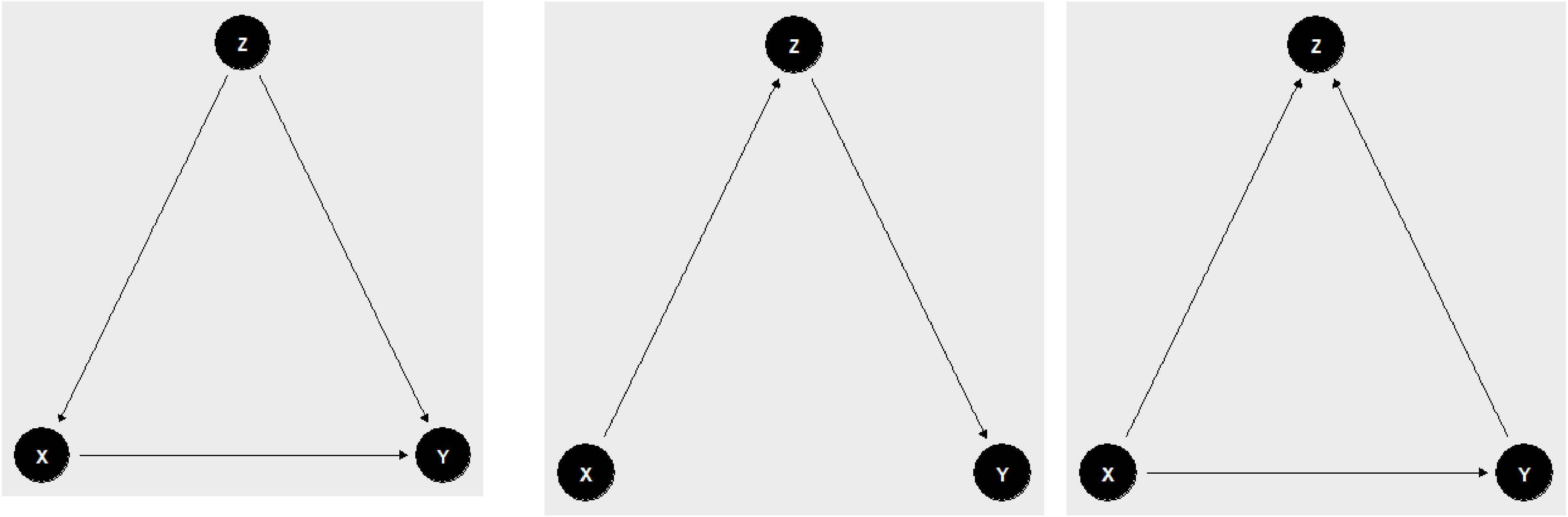
2021-06-01

From last class

- Causal inference
- Confounds (forks, pipes)
- Model selection
- interactions

Today’s class

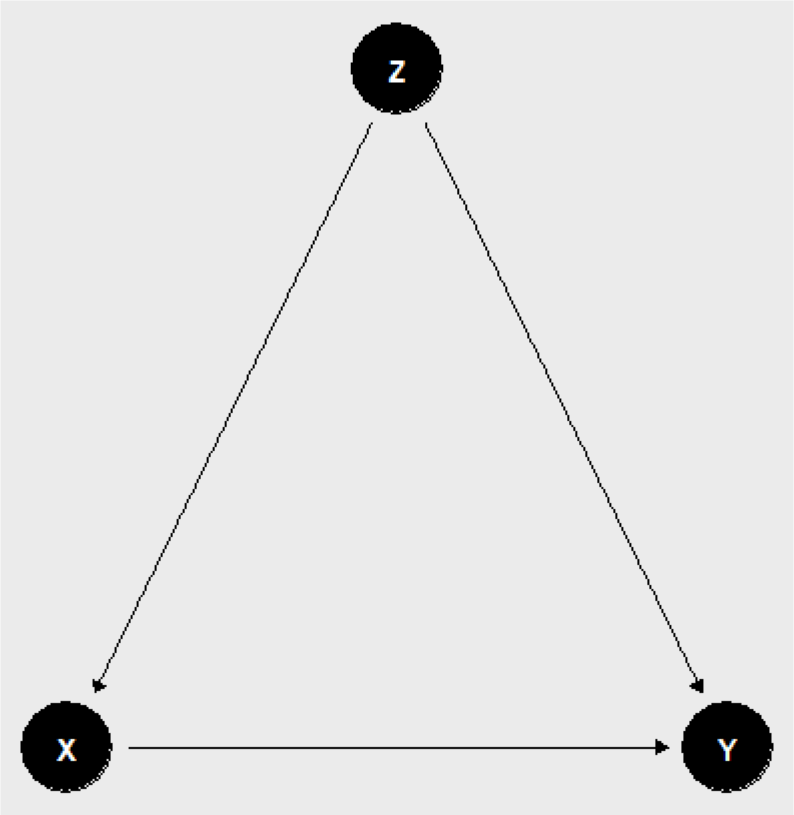
- Confound (pipes, colliders)
- Probability vs. Likelihood
- Generalized linear models
- Logistic regression
- **Final project draft is due July 13th**



More confounds

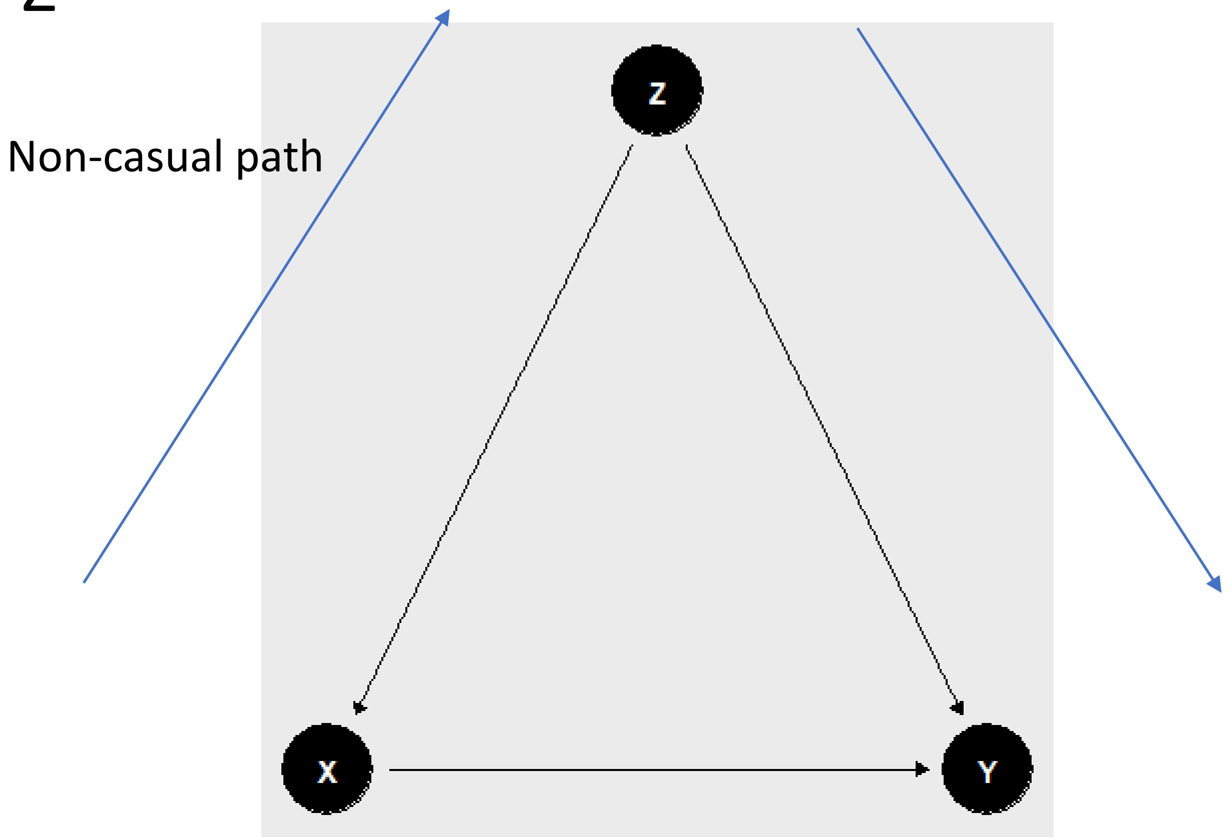
The fork

- You want to estimate the effect of X on Y, but the relationship is forked by another variable Z
- What do you do?



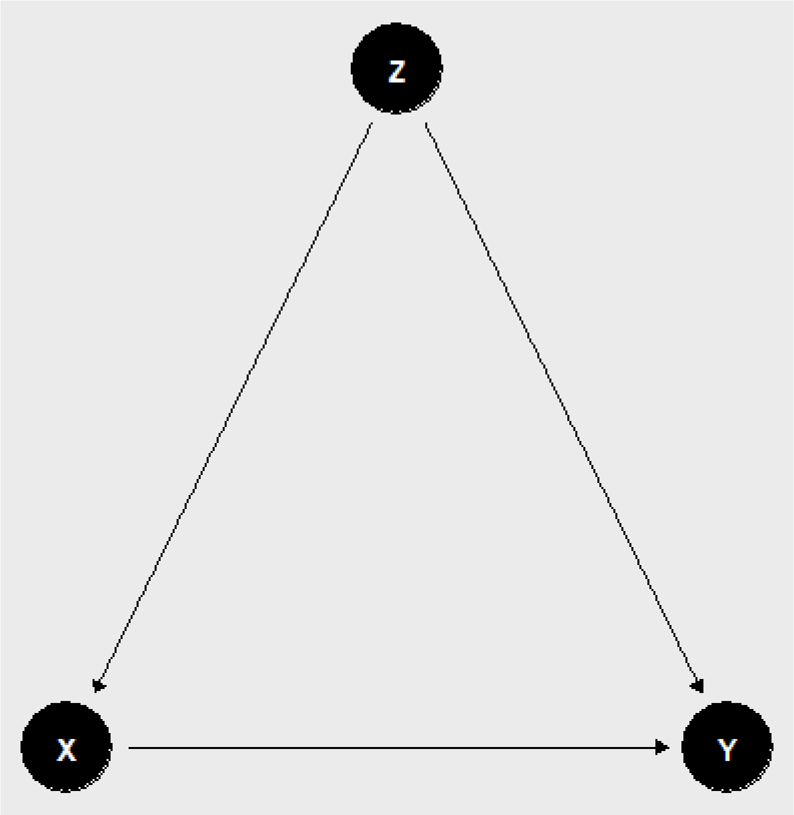
The fork

- You want to estimate the effect of X on Y, but the relationship is forked by another variable Z
- What do you do?



The fork

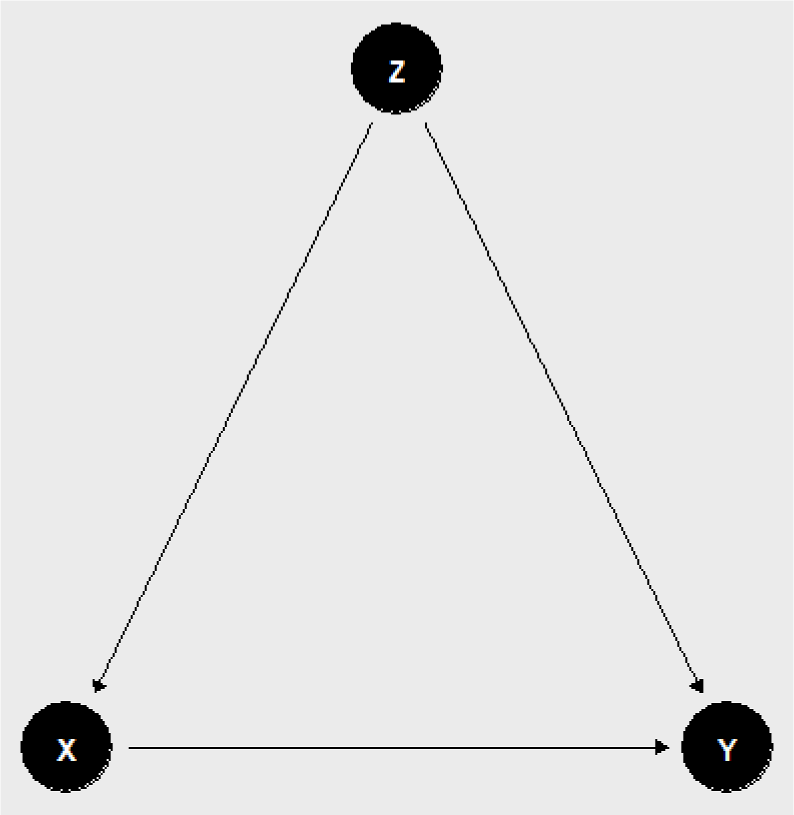
- You want to estimate the effect of X on Y, but the relationship is forked by another variable Z
- What do you do?
- You condition on Z



The fork

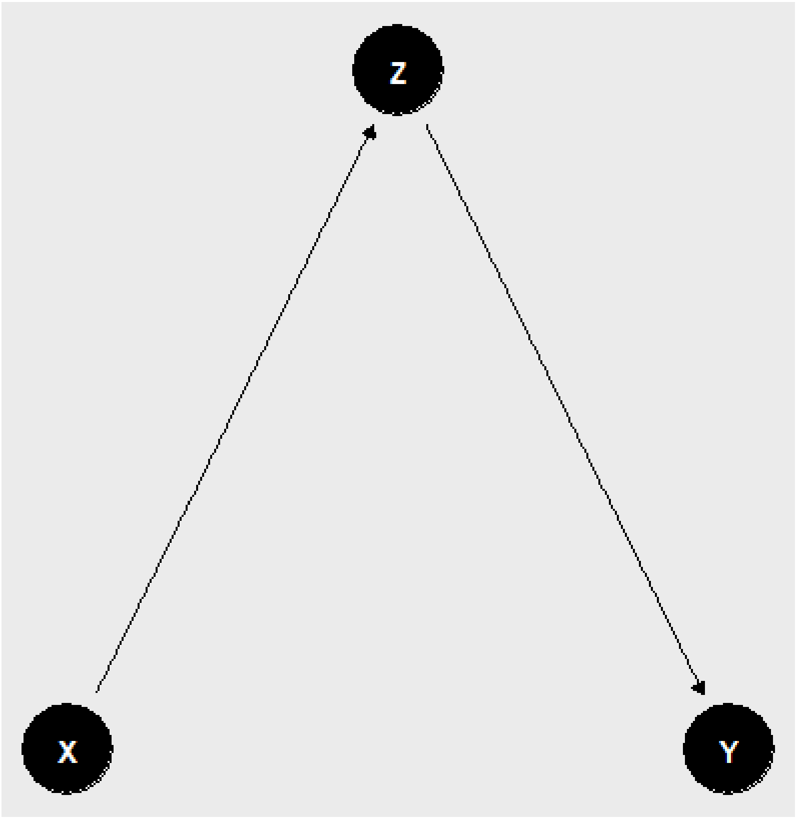
- You want to estimate the effect of X on Y, but the relationship is forked by another variable Z
- What do you do?
- You condition on Z
- That means putting Z in as a

predictor in your model



The pipe

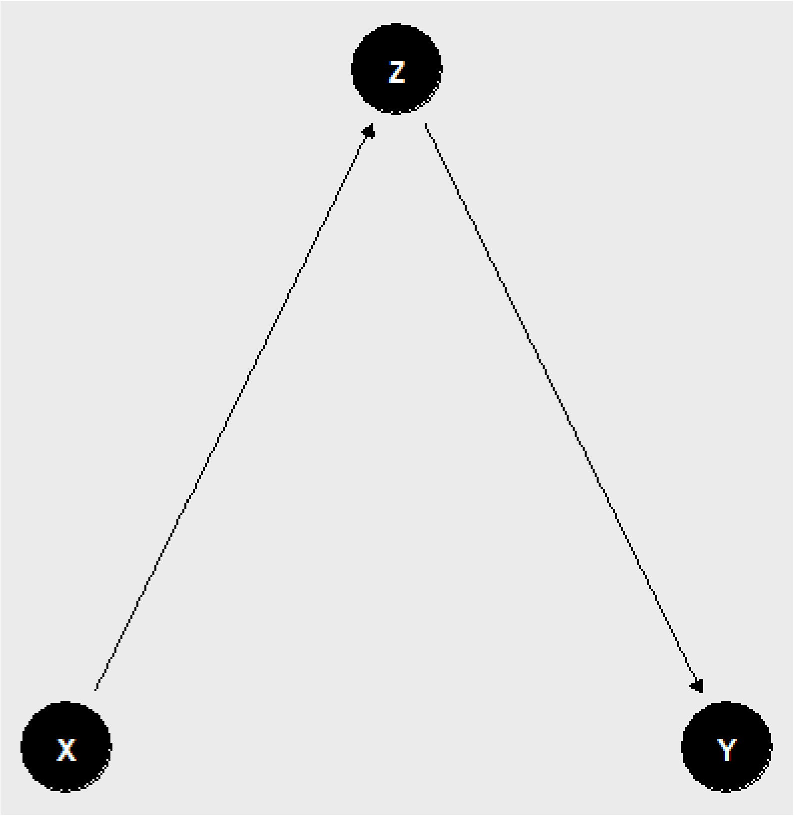
- You want to estimate the relationship of X on Y, but the relationship is piped by Z.
- What do you do?



The pipe

- You want to estimate the relationship of X on Y, but the relationship is piped by Z.
- What do you do?

If your question is how a change in X will influence a change in Y, then notihing
You do not want to add variables that block off causal pathways

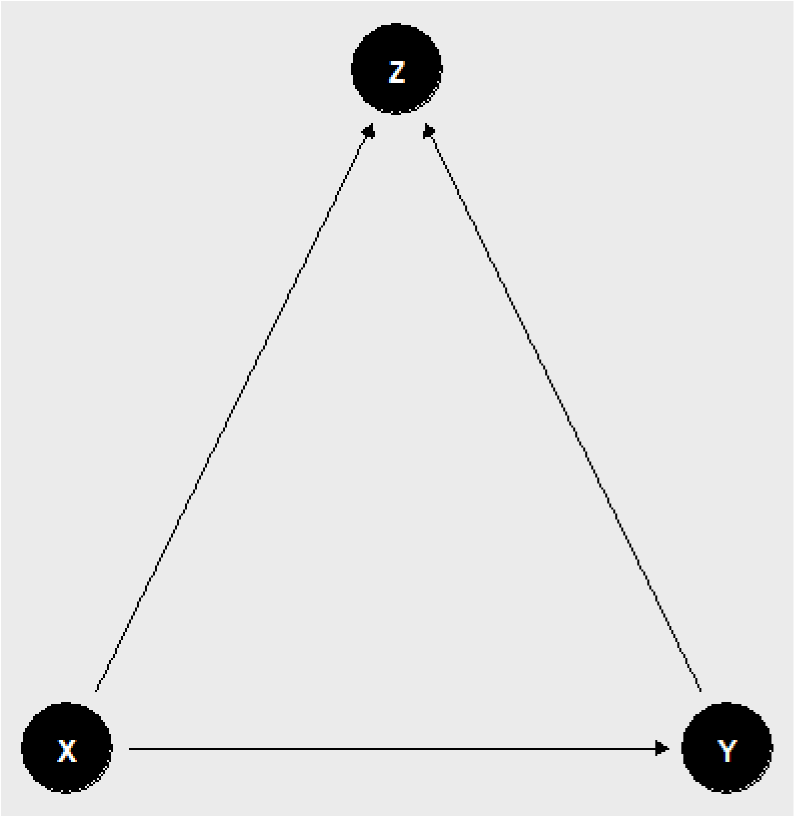


The pipe

- Let’s try a simulation

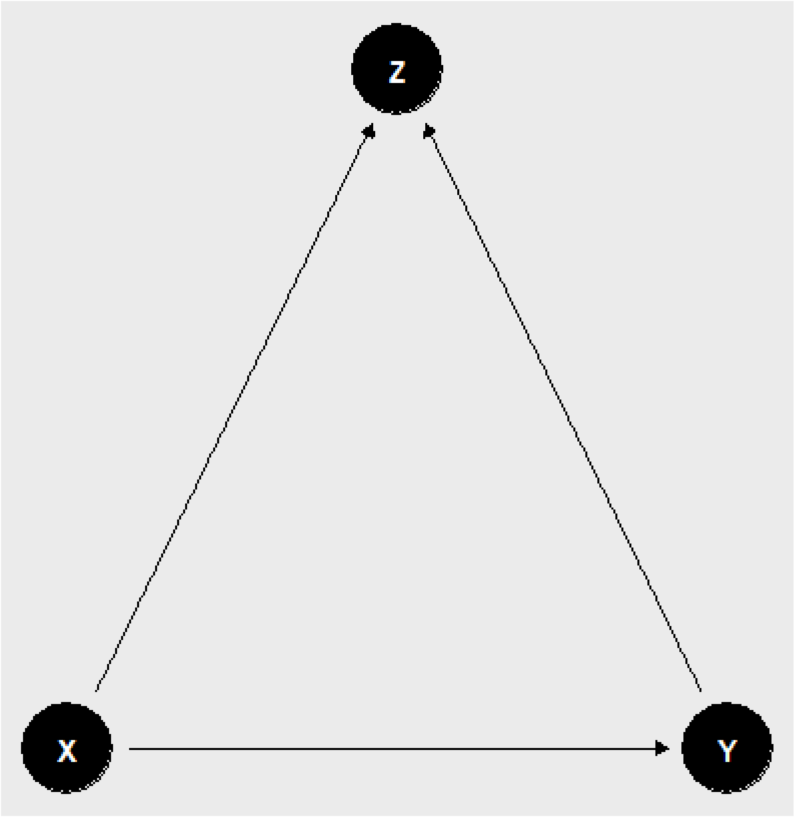
The collider

- We want to assess the relationship of X on Y, *and* both X and Y cause Z.
- Should we condition on Z?



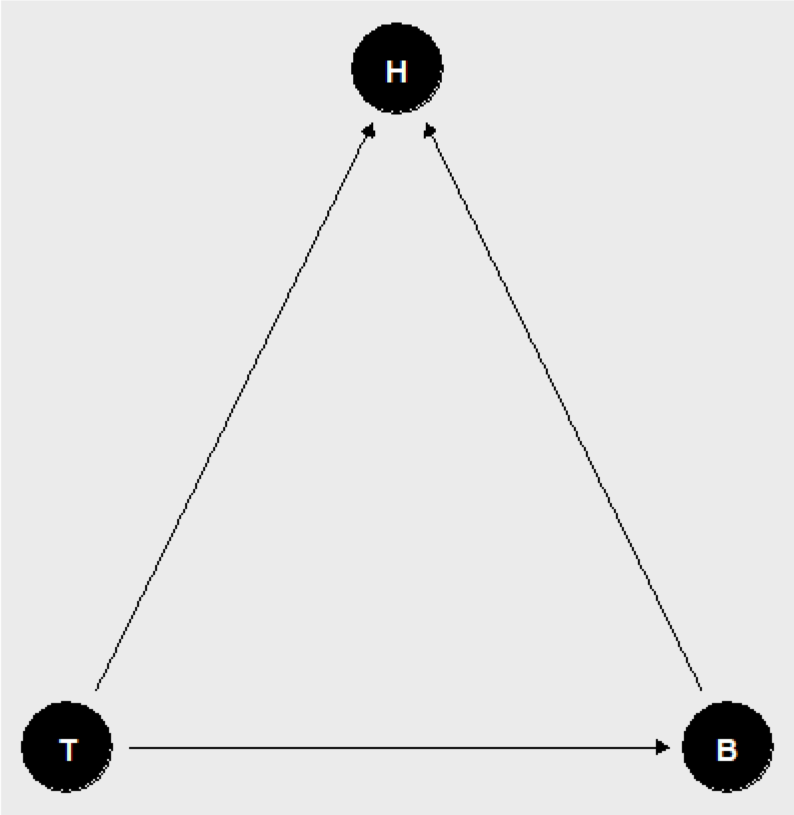
The collider

- We want to assess the relationship of X on Y, *and* both X and Y cause Z.
- Should we condition on Z?
- This is the most counterintuitive confound
- So let's illustrate.



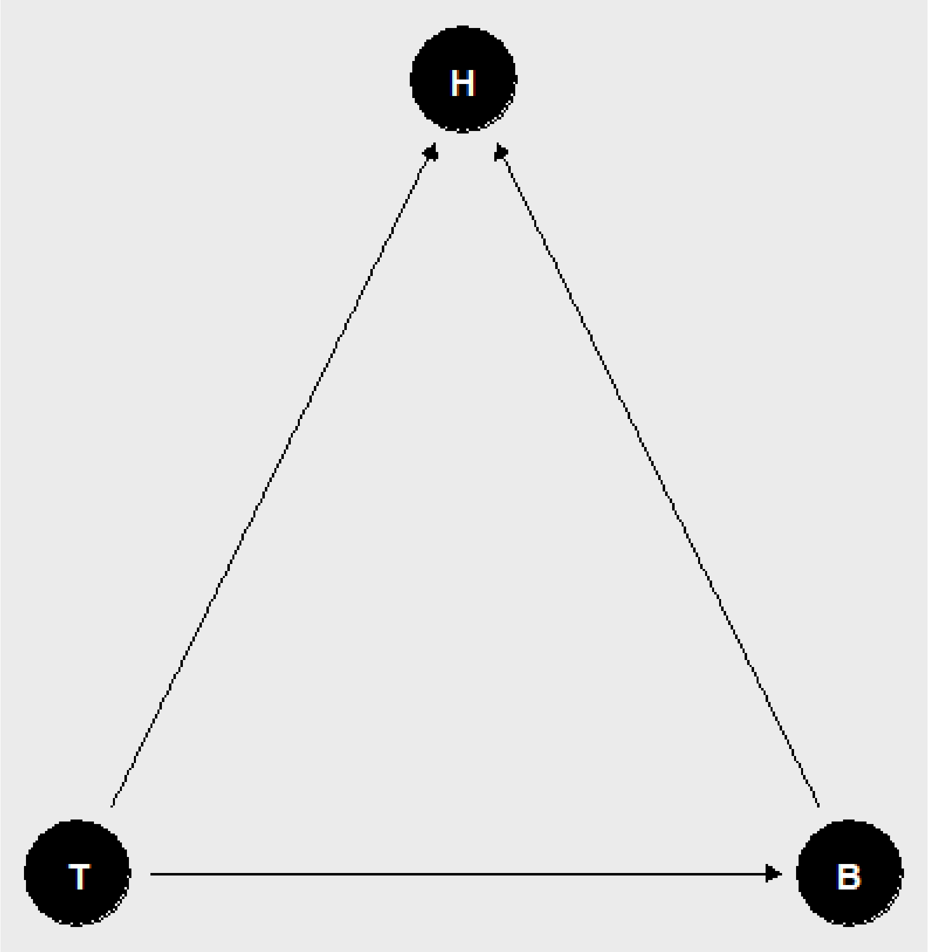
The collider

- Let’s say you are interested in the relationship between a loss of taste and back pain.
- You think that a loss of taste might be a proxy for covid for instance, and you want to see, the effect of covid on back pain.
- But you can only observe patients in a hospital.



The collider

- T = Loss of taste sensation
- B = backpain
- H = hospitalization
- We know that more people go to the hospital if they have a loss of taste
- We know that more people go to the hospital if they have backpain



The collider

- Let’s simulate this
- But imagine you can only observe people in a hospital

```
set.seed(115)
n <- 1000
notaste <- rnorm(n)
backpain <- rnorm(n)

p <- 0.2 #proportion of people who go to hospital
z <- notaste + backpain + rnorm(n) #measure of overall discomfort
q <- quantile(z, 1-p)
h <- ifelse(z>=q, TRUE, FALSE)
model.hospital <- lm(backpain[h]~notaste[h])
summary(model.hospital)
```

The collider

- Should you conclude that covid cures back pain?

```
Call:
lm(formula = backpain[h] ~ notaste[h])
Residuals:
     Min       1Q   Median       3Q      Max
-1.72670 -0.61173 -0.06442  0.52977  2.57790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.00764    0.08018  12.567  < 2e-16 ***
notaste[h]  -0.37545    0.06579  -5.707 4.16e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.8355 on 198 degrees of freedom
Multiple R-squared:  0.1412,  Adjusted R-squared:  0.1369
F-statistic: 32.57 on 1 and 198 DF,  p-value: 4.157e-08
```

The collider

- Let’s say you have access to data from people outside the hospital as well.
- Should you include the hospital factor in your model?

```
model.collider <- lm(backpain~notaste+h)
summary(model.collider)
```

The collider

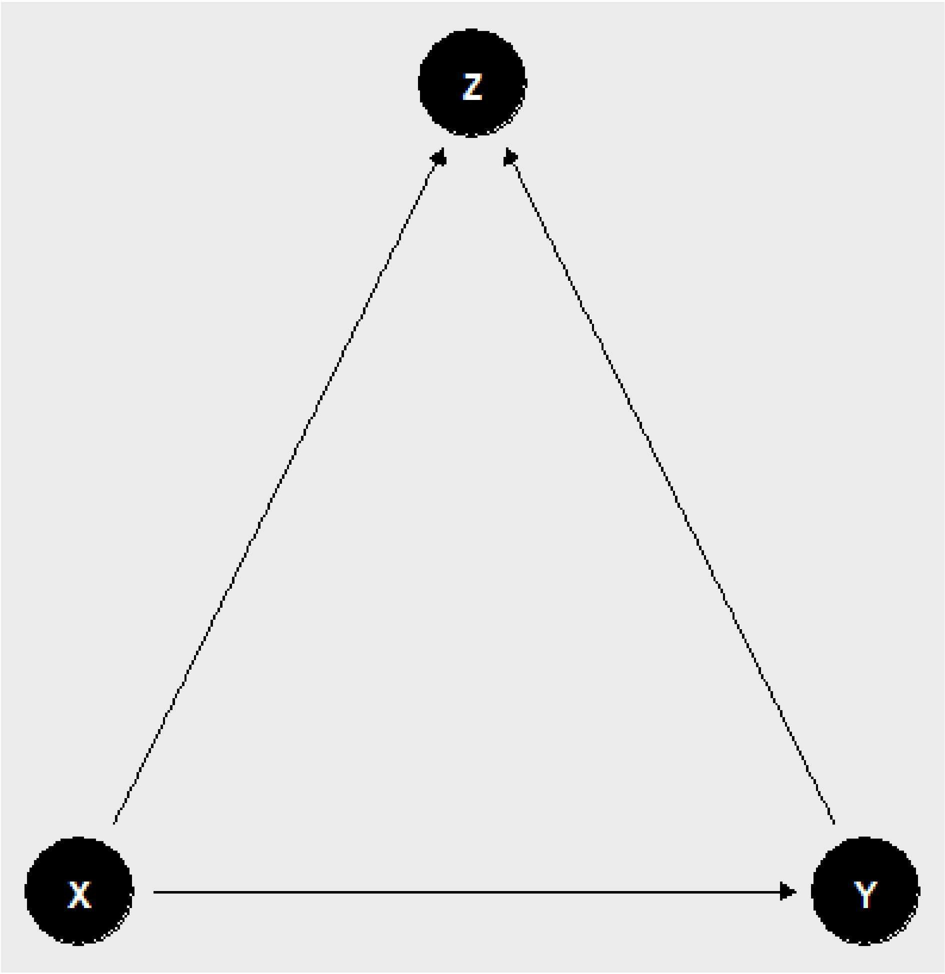
```
Call:
lm(formula = backpain ~ notaste + h)
Residuals:
     Min       1Q   Median       3Q      Max
-2.95109 -0.59268  0.00274  0.55379  2.97088

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.24049    0.03157  -7.617 6.00e-14 ***
notaste     -0.23603    0.02998  -7.874 8.94e-15 ***
hTRUE        1.13326    0.07574  14.962  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8832 on 997 degrees of freedom
Multiple R-squared:  0.1867,  Adjusted R-squared:  0.1851
F-statistic: 114.5 on 2 and 997 DF,  p-value: < 2.2e-16
```

The collider

- You *shouldn't* condition on a collider because it introduces a confound.
- You should avoid designing research questions in such a way that you create collider biases as well.



In the real world

- When scientists advocate more tests for people who show no symptoms or for people who are not hospitalized, they are generally worried about collider effects associated with hospitalization.

Causal inference

- Take-away:

- In later stages of research, don't just throw lots of variables into a model – think about what you are doing in terms of causal structure.

- Statistical inference depends on a causal model.

- Next week we'll look at the last confound (the descendent)

Generalized linear models

Generalized linear models

- Up until now we have been using OLS (ordinary least squares) regression.
- A more powerful tool is a **generalized linear model**
- Generalized linear models use Maximum Likelihood Estimation (MLE) in order to get their parameters.

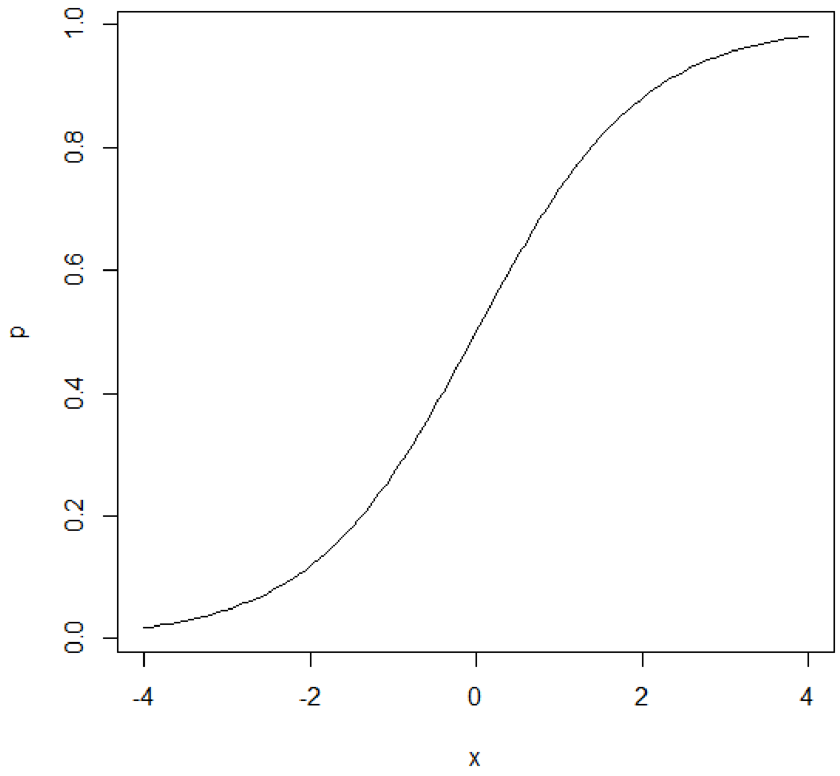
Generalized linear models

- Likelihood and maximum likelihood simulation exercise

$$pr(data|distribution)$$

$$L(distribution|data)$$

Logistic regression



Logistic regression

- Logistic regression is a method for modelling binary data.
- The basic ideas can be extended to non-binary data as long as they are organized into levels.
- It is typically used when the dependent variable is binary and there is an interest in knowing how a change in x effects the probability that something is y.

Logistic regression (typical uses)

- Psycholinguistic experiments where subjects have to give yes or no answers.
- Various uses in natural language processing
- Predict the risk of developing a specific disease.
- Predict probability that someone will vote for a particular political party

Logistic regression

- A logistic regression or logit model can be represented with the following equation.

$$logit(y) = b_0 + b_1 x_1 + b_2 x_2 ...$$

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

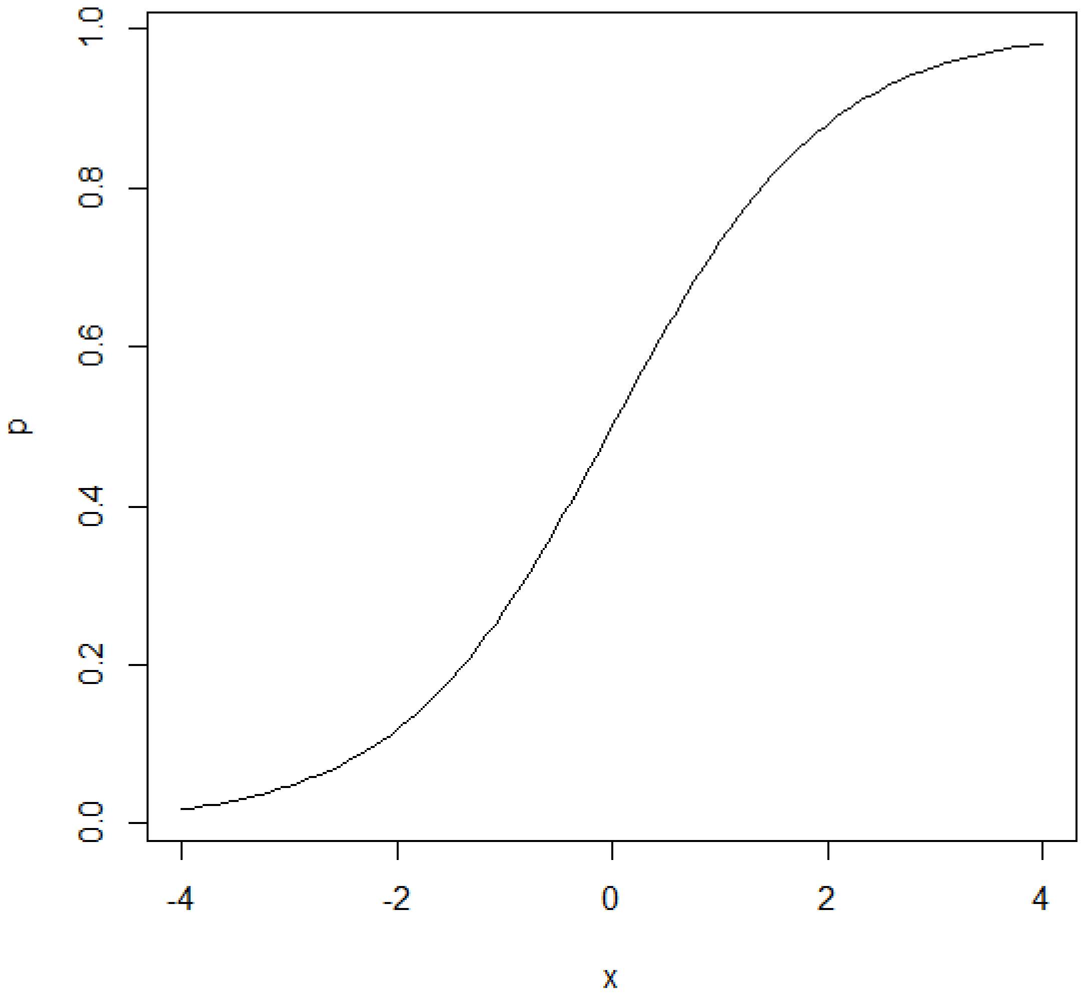
$$Prob\{y = 1 | x\} = \frac{1}{1 + exp(-x\beta)}$$

Odds,log odds, oddrs ratios and log odds ratios

- **Odds**: simple ration of the probability of one event to the probability of another evento (frequency of a / frequency b) are the odds of a over b.
- **Log odds**: Logarithmically transformed odds
- **Odds ratio**: ratio of two odds
- **Log odds ratio**: Logarithmically transformed log odds

Logistic regression

- Y Is bounded to 0 or 1
- The relationship between x and y has a ceiling effect (like logarithms)
- Let’s run a simulation model to get the feel for it.



Two causative constructions in Dutch

Doen relates to direct causation

Laten relates to indirect causation

(1)	*Hij*	***deed***	*me*	*denken*		*aan*	*mijn*	*vader*
	He	**did**	me	think		at	my	father
	‘He reminded me of my father.’							
(2)	*Ik*	***liet***	*hem*	*mijn*	*huis*	*schilderen*		
	I	**let**	him	my	house	Paint		
	‘I had him paint my house.’							

Homework and reading

- Chapter 11 and 12 of Levshina (2016)
- Chapter 7 Baayen (2008)
- Please finish homework!

- Next lecture:
  - Interpreting logistic regression
  - The descendent confound
  - Hierarchical models