Slides 2021 05 25

Adam Tallman

25/05/2021

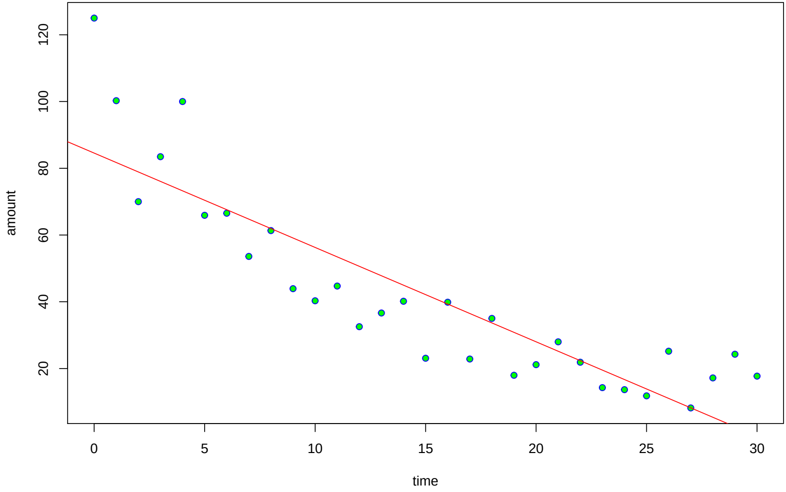
From last lecture

- Anova model
- Linear regression
- Overfitting
- AIC

For today's lecture

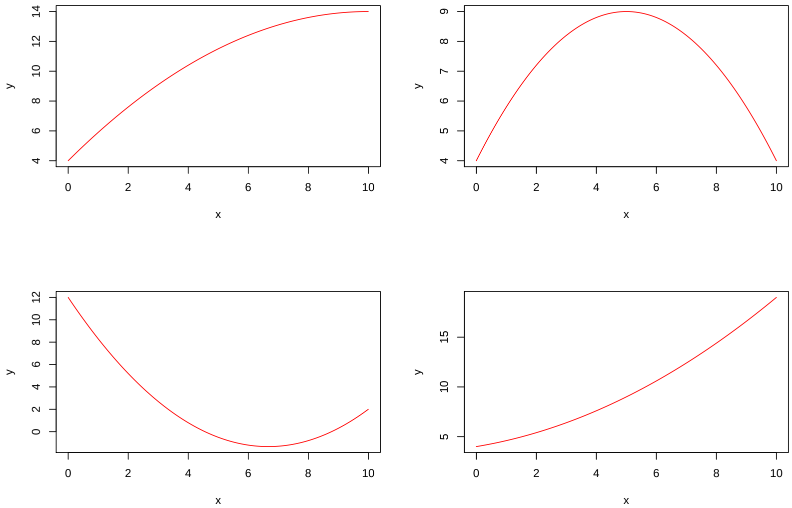
- Multiple regression
- Causal inference
- Model selection procedure
- Interactions

Regression on decay



Polynominal regression

- There are different types of relationship we can construct by adding variables to our regression equation in different ways



Polynomial regression

- model2 <- lm(amount~time)
- model3 <- lm(amount~time+I(time^2))
- summary(model2)
- summary(model3)

$$y = a + bx$$

$$y = a + bx + cx^2$$

Polynomial regression

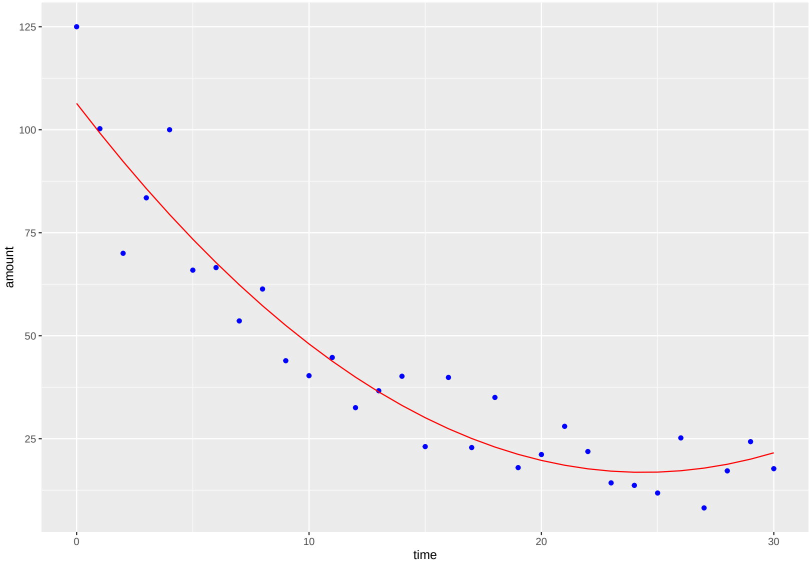
- Here is the linear model

$$y = a + bx$$

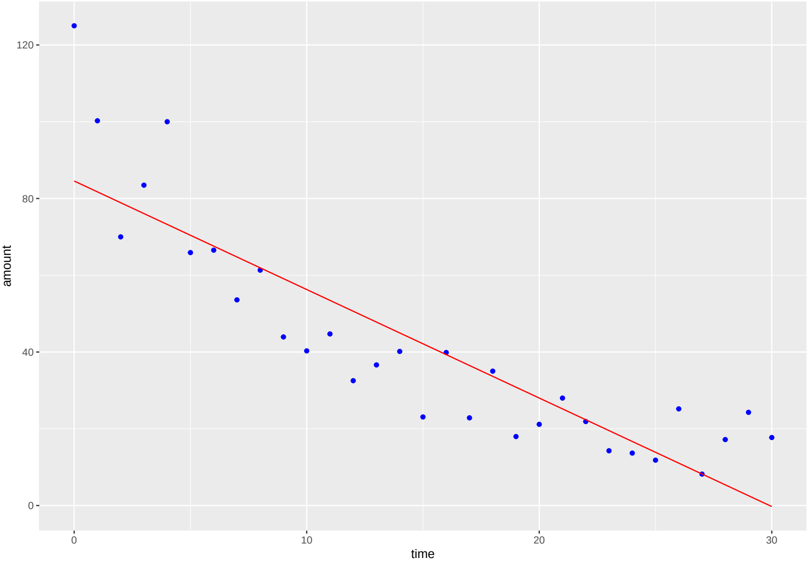
```
##
## Call:
## lm(formula = amount ~ time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.065 -10.029  -2.058   5.107  40.447
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.5534     5.0277   16.82  < 2e-16 ***
## time         -2.8272     0.2879   -9.82 9.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Polynomial regression

- Here is the polynomial equation



Polynomial regression



Polynomial regression

- What about this?

$$y = a + bx + cx^2 + dx^3 + ex^4 + fx^5 ...$$

Akaike Information Criterion

- Complex models have a tendency to not extend beyond the data they are modelling
- Or, to what extent are you modelling noise by adding so many parameters
- Akaike information criterion (and its friends) is a criterion for model selection that
- The statistic model incurs penalties for its complexity
- k = the number of parameters
- $ln(\hat{L})$ = the loglikelihood

$$AIC = 2k - 2ln(\hat{L})$$

Multiple Regression

Model fitting philosophy

- A. The more variance the model accounts for the better
- B. The simpler the model the better (Occam's razor)
- A and B are in conflict
- If you overdo B you're model could simply be **misspecified** and (perhaps) come to **spurious associations** based on not considering enough factors
- If you overdo B, you will **overfit**

Overfitting

- What is so bad about over-fitting?
  - The model will not extend beyond the data you fit it to
  - You are unlikely to have a meaningfully testable hypothesis

Underfitting

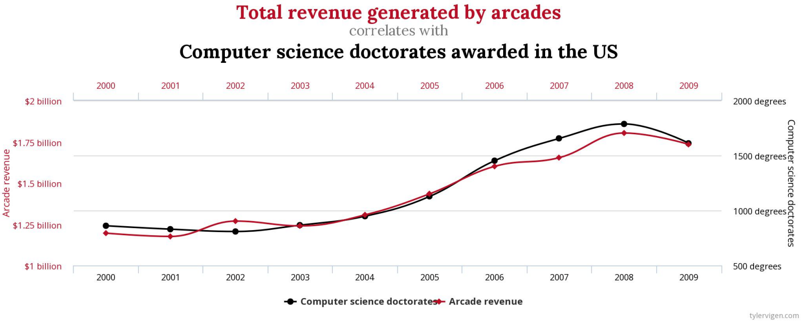
- What is so bad about underfitting?
  - You cannot make reliable **causal inferences** - you just have an association

What is a causal inference

- A causal inference is what you get when you assume that your statistical model is close enough to reality for you to make a claim about causation.
- In most complex systems you'll need to think about whether the relationship is causal

Can you think of a non-causal association?

Can you think of a non-causal association?



How do we distinguish between causation and association?

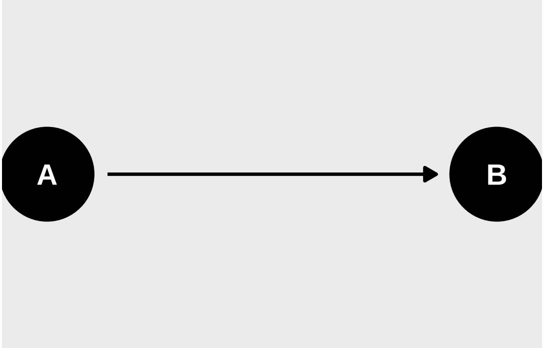
- There's a fairly new field of **causal inference** that is concerned with distinguishing between association and causation
- We won't learn all the details of causal inference in this course, **but** interpreting multiple regression properly does implicitly involves some causal thinking.
- So if you work your way into more advanced courses it worth it to read more about causal inference

Modeling causal assumptions

- Structural causal model (SCM) describes how nature assigns values to the variables of interest
- "Variable X is a *direct cause* of a variable Y if X appears in the function that assigns Y's value. X is a *cause* of Y it it is a direct cause of Y, or of any cause of Y." (Pearl, Glymour, Jewell 2015)
- You can associate every SCM with a graphical causal model

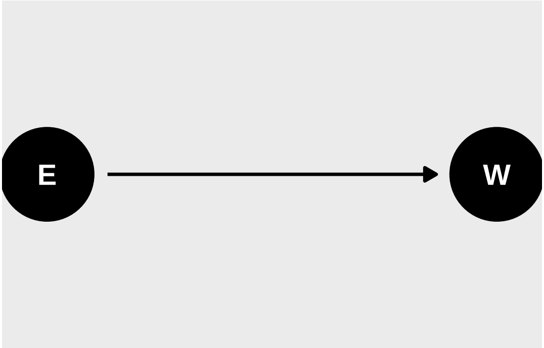
Graphical causal model

- A node with an arrow from A to B represents a causal function from the variable A to B



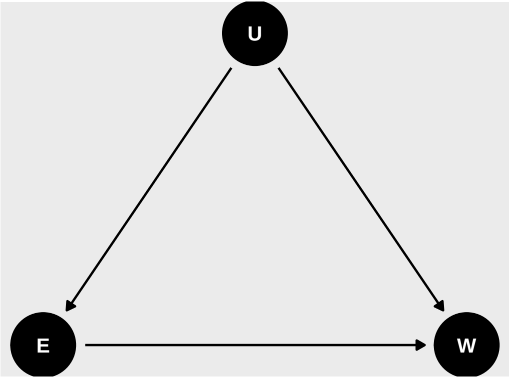
Graphical causal model

- Imagine we want to figure out whether education (E) causes higher wages (W)



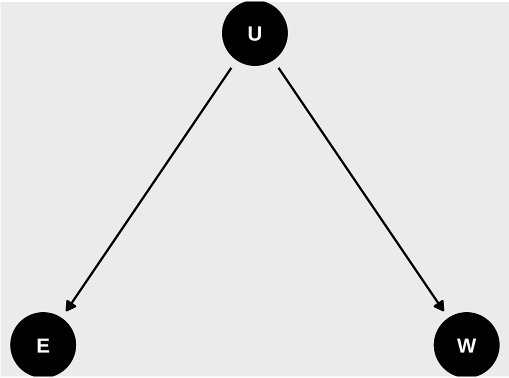
Graphical causal model

- But there is a third variable that is causes a higher education and higher wages.
- By virtue of this unknown variable E and W are correlated with one another, (but, in fact, there may be no reliable causal relationship)



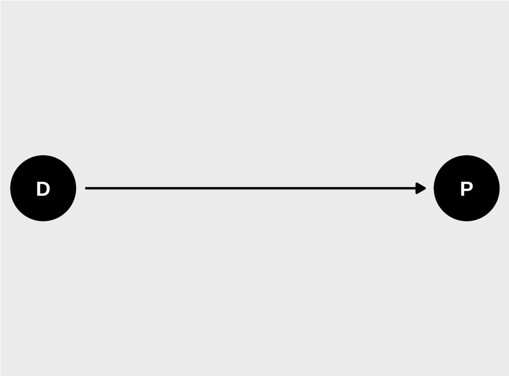
Graphical causal model

- U would be what we call a “confound"
- You get an association between W and E because of a **non-causal path** between the two, but not because one causes the other.



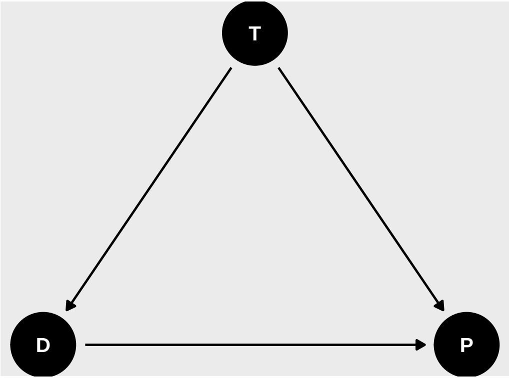
Graphical causal model

- Let's look at the relationship between the pitch at the end of a vowel and vowel duration -Let's say we think that an increase in vowel duration is associated with a pitch increase.



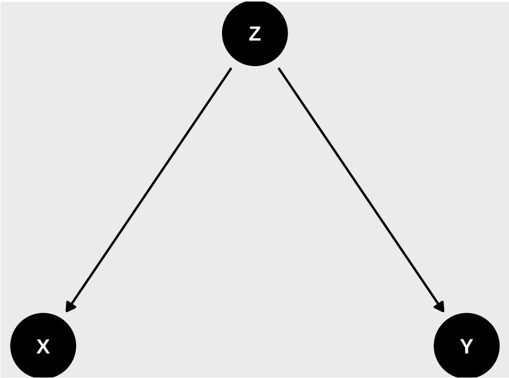
Graphical causal model

- The problem is that the pitch and the duration might be correlated with each other because they are both correlated with difference in tones.



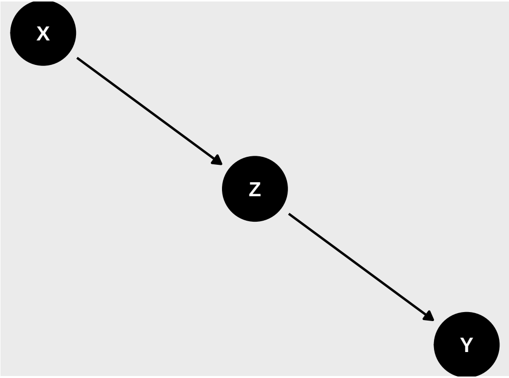
Confounds, the Fork

- The association between duration and end pitch we just saw is considered a backdoor association (it doesn't do through causal paths)
- We have to condition on one of the variables to close the backdoor
- This specific relationship is called a "fork"
- We say that Z forks the relationship between X and Y



Confound 2, The pipe

- Another type of confound is called **The Pipe**
- We block the path between X and and Y by conditioning on Z, but in this case it is about direct vs. indirect causation.



Causal inference

- There are two more types of confounds, which we'll introduce next lecture
- **The Collider** and **The Descendant**

Basic model fitting / simplification

Basic model fitting

- Once you understand the causal relationships you are interested, you have to have some methodology for weighing simplicity against accuracy/fit
- The simplest way of doing this is by starting with a maximal model and moving to a minimal adequate model

Some basic types of models

- From Crawley (2016: 195)
- **Saturated model**: There is one parameter for every data point (perfect fit)
- **Maximal model**: Contains all factors, interactions and covariates that might be of interest
- **Minimal adequate model**: A simplified model which has removed superfluous variables
- **Null model**: Just one paramter, the overall mean

Classic model selection process (simplified)

- From Crawley (2016:195-196)
  - **Fit the maximal model**: Fit all the factors, interactions and covariates of interest. Note the AIC.
  - **Begin model simplification**: Inspect the parameter estimates using summary(). Remove the least significant terms first, using update(), starting with highest order interactions.
  - **What does the deletion do to the AIC?**
    - If it increases the AIC -> Keep the interaction term and go back to step one looking at another term
    - If it decreases the AIC -> Leave the parameter deleted and continue simplifying the model
    - **Check assumptions**: Use plot() to check model assumptions making sure there is no heteroscedacity (unequal scatter of residuals)

What's wrong with heteroscedacity?

- We want our model to make good predictions within all range of values, not just some of them.
- Heteroscedacity indicates that you cannot trust the model within certain ranges of the values you are interested in.
- Roughly this means under different circumstances your predictions about the world could be complete shit (and this might be dangerous for policy decisions)

Interactions

- Interactions are sometimes hard to interpret
- They arise when the value of the effect of x on y depends on some third variable z
- It's normal to spend some time trying to wrap your head around the meaning of an interaction
- The only reason Crawley seems to recommend removing third way interactions, is because they are conceptually difficult to understand.

Interactions

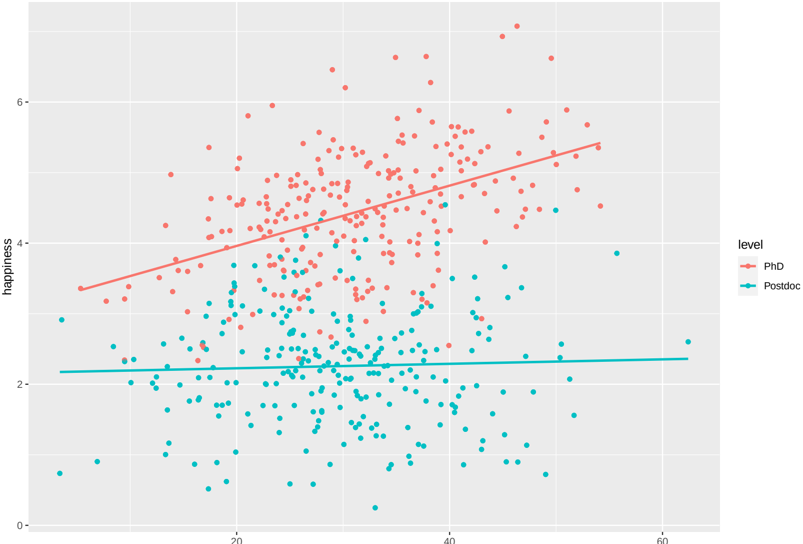
-Simulate a regression without an interaction

```
## `geom_smooth()` using formula 'y ~ x'
```

Interactions

- simulate a regression with interactions

```
## `geom_smooth()` using formula 'y ~ x'
```



Reading and Homework

- Homework 4
- Reading, Levshina (2016) Chapter 12, 13 & 14
- Reading, Baayen (2008) Chapter 5 & 6