Statistics II

Week 3: **Revisiting Regression Estimators of Causal Effects**

Content for Today

Today we will have a look at **regression** and how it can be used under certain conditions to gather **causal estimates**. Additionally, we will learn how putting our qualitative assumptions in **causal graphs** can help us in a very intuitive way to adjust and rid our estimates of bias.

Content for Today

1. OLS and regression from a causal perspective
2. Causal graphs and the backdoor criterion
3. Thinking about bias
4. A reminder of regression in R

Lecture Review

Ordinary Least Squares (OLS)

Addresses a simple mechanical problem. How to minimize the sum of the square deviations from a line.

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

← Minimize this

It creates a linear fit.

* think about it as what is our best guess for y given a particular x

Ordinary Least Squares (OLS)

We can have a **bivariate regression** where the slope of the line will be calculated as:

$$\hat{\beta}_1 = \frac{cov(x,y)}{var(x)} = \frac{\sum(x_i - \hat{x}_i)(y_i - \hat{y}_i)}{\sum(x_i - \hat{x}_i)^2}$$

The intercept can be derived as:

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

Multiple regression

As we have seen during the last two weeks, more often than not bivariate relationships are subject of noise coming from other variables. In this cases **multiple regression** can aid us in partially accounting for the noise.

We can think about conditional independence/ignorability assumption from last week's lecture*

*(Session 2 - Slide 21)

Regression from a causal perspective

As we have discussed regression addresses a simple mechanical problem, namely, what is our best guess of y given an observed x.

- Regression can be utilized without thinking about causes as a *predictive* or *summarizing* tool.
- It would not be appropriate to give causal interpretations to any β , unless we establish the fulfilment of certain assumptions.

Regression from a causal perspective

If we put this structural model,

$$y_i = \beta_0 + \beta_1 D + e_i$$

in POF notation:

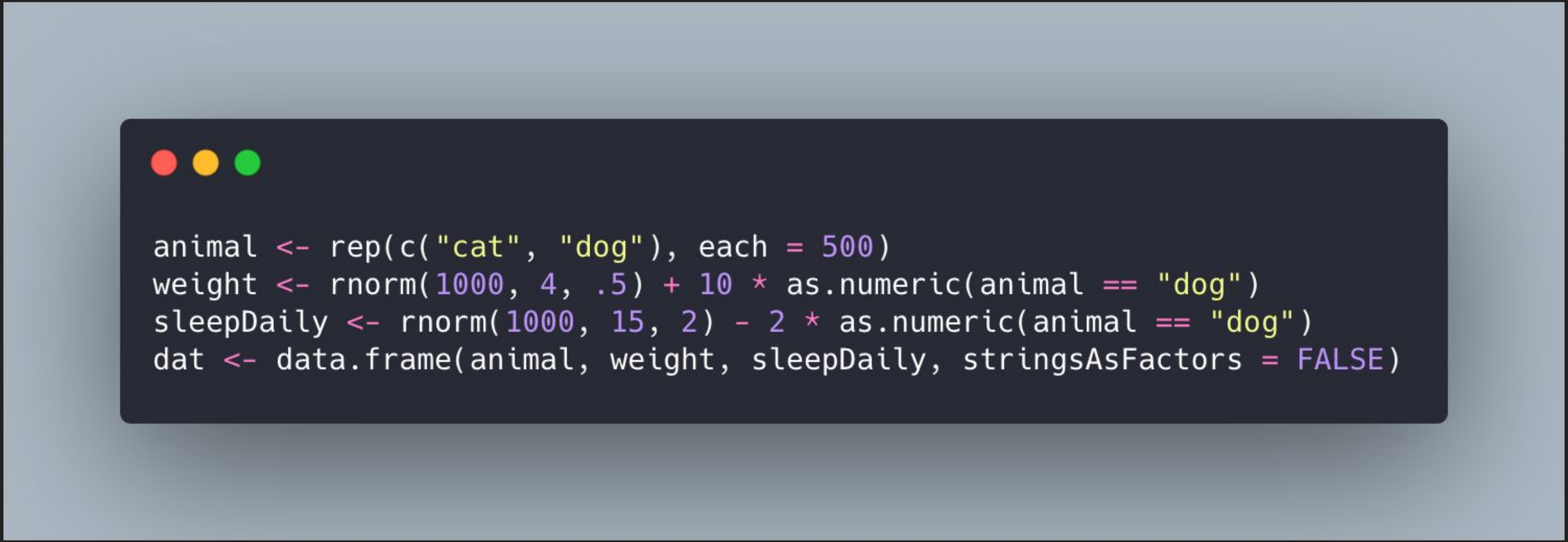
$$E(Y^0 | D = 0) = \beta_0$$

$$E(Y^1 | D = 1) = \beta_0 + \beta_1$$

$$\beta_1 = NATE$$

Let’s think about our cats and dogs simulation from last week!

Regression from a causal perspective



```r
animal <- rep(c("cat", "dog"), each = 500)
weight <- rnorm(1000, 4, .5) + 10 * as.numeric(animal == "dog")
sleepDaily <- rnorm(1000, 15, 2) - 2 * as.numeric(animal == "dog")
dat <- data.frame(animal, weight, sleepDaily, stringsAsFactors = FALSE)
```

$$weight = 4 + 10(dog) + r_i$$

Regression from a causal perspective

If **D** and the **error term** are independent the our β_1 could be the **ATE.**

In order to achieve this we need to have the true model or else our bias will be relegated to the error term.

This is where putting our qualitative assumptions in **causal graphs** can help us lay out our models in a very intuitive way.

Causal graphs

Directed Acyclic Graphs (DAGs) can be utilized to identify if we can meet the **adjustment criterion** with a set of observables.

- Back-door paths -> non-causal paths that start with an arrow into **D**

Paths are **open** at mediators and confounders

Paths are **closed** at colliders

Let's look at the lecture slides for week 3!

Causal graphs and regression

1. Lay out your assumptions in a DAG based on empirical a theoretical knowledge
2. Identify the causal and non-causal paths from **D** to **Y**
3. Identify the adjustments that would close the non-causal paths
4. Include the identified variables in the model specifications
5. Withstand the temptation to give any other coefficient than that for D a causal interpretation – the status of covariates is path-specific!

Let’s look at the lecture slides for week 3!

Thinking about bias

1. A path is **open** or **unblocked** at non-colliders (confounders or mediators)
2. A path is **(naturally) blocked at colliders**
3. An **open path induces statistical association between two variables**
4. Absence of an open path implies statistical independence
5. Two variables are **d-connected** if there is an open path between them
6. Two variables are **d-separated** if the path between them is blocked

Let’s look at the lecture slides for week 3!

Thinking about bias

Conditioning on a **mediator** leads to **overcontrol** or **post-treatment bias**

Thinking about bias

Conditioning on a **collider** (or a descendant) leads to **collider bias** or **endogenous bias**

Thinking about bias

Failing to condition on a **confounder** leads to **omitted variable bias**

Let’s move to R!