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**.

Content for Today

1. OLS and regression from a causal perspective
2. A reminder of regression in R

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:

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

Regression from a POF 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 POF 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 ethnic prejudice example from last week!

Regression from a POF perspective

Student (i)	Prejudice			Contact
	y_{0i}	y_{1i}	δ_i	
1	6			0
2		2		1
3	4			0
4	6			0
5		1		1
6		2		1
7	8			0
8	4			0

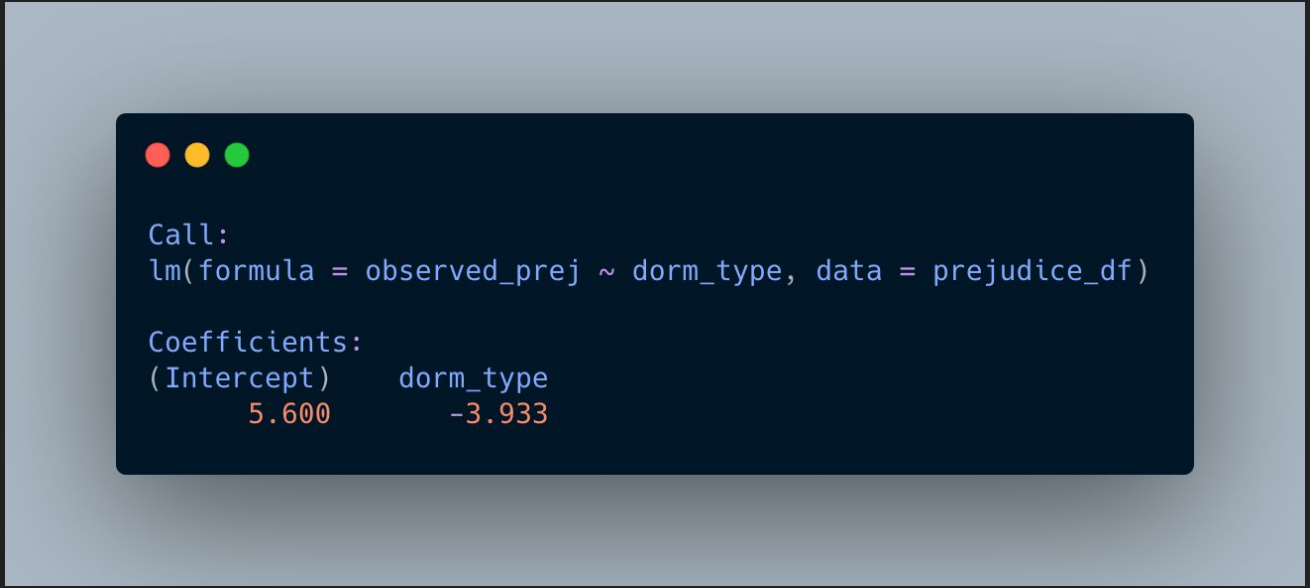
Information we *do* have

Regression from a POF perspective

$$
\begin{aligned}
NATE &= E[y_{1i}|d_i = 1] - E[y_{0i}|d_i = 0] \\
&= \frac{2+1+2}{3} - \frac{6+4+6+8+4}{5} \\
&= 1.666 - 5.6 \\
&= -3.933
\end{aligned}
$$

*In this case, we know that the NATE is not equal to the ATE

Regression from a POF perspective



```
Call:
lm(formula = observed_prej ~ dorm_type, data = prejudice_df)

Coefficients:
(Intercept)    dorm_type
      5.600       -3.933
```

$$ObsPrejudice = 5.6 - 3.93(dorm\_type) + e_i$$

Regression from a POF perspective

If **D** and the **error term** are independent then our β_1 , or **NATE**, could be the **ATE.**

In order to achieve this, we need to have the true model or else we will have biased estimates (as is the case for the prejudice example)

This is where putting our qualitative assumptions in **causal graphs** can help us lay out our models in a very intuitive way. We will learn more about this next week.

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 these cases **multiple regression** can aid us in partially accounting for the noise.

We will learn about the backdoor criterion and conditional independence/ignorability assumption next week*

Multiple regression

Assumptions granted, we can utilize multiple regression to **condition on observables** to render unbiased estimates.

We will learn about the backdoor criterion and conditional independence/ignorability assumption next week*

Let’s move to R!