# Bias

Econ 140 Spring 2025, Section 3

---

Jonathan Old

Syllabus/OH bCourses Website Feedback form (**Always open**) RStudio

# Roadmap

1. Time for your questions
2. Recap
3. Selection bias in action: Omitted Variable Bias
4. Time for your questions
5. Inference and hypothesis testing
6. Exam practice

# Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

# Recap

# Recap: Regression

See on blackboard

# Recap: Selection bias

We saw that whenever we do a difference-in-means comparison (or a regression), we get:

**Estimating the effect of iPads on grades**

Let us start with a difference-in-means comparison:

$$\Delta = E[\text{Grade}_i|\text{iPad}_i = 1] - E[\text{Grade}_i|\text{iPad}_i = 0]$$

Add and subtract $E[\text{Grade}_i(0)|\text{iPad}_i = 1]$ :

$$= E[\text{Grade}_i(1)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 1] + E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

Use properties of expectations:

$$= E[\text{Grade}_i(1) - \text{Grade}_i(0)|\text{iPad}_i = 1] + E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

$$= \text{ATT} + \text{Selection bias}$$

Selection bias: Students with and without iPad have different potential grades: **even if they both *had* iPads, they would be different**.

6

Causal Effect

+

Selection Bias

- We cannot observe them - so we can never be sure!
- Econometrics is all about uncertainty. You can always state that there are different possibilities and you cannot know for sure

# Recap: How to think about Selection Bias

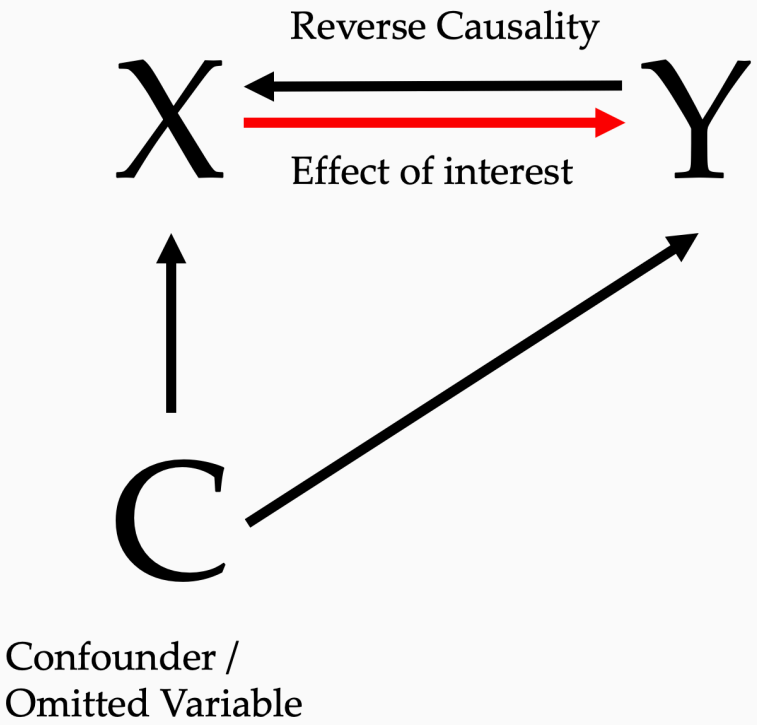



**Figure 1:** Selection bias

# Selection bias in action: Omitted Variable Bias

# The most important slide of this course (until the midterm)

Let $Y_i$ be the outcome variable, $X_i$ our regressor of interest, $W$ a series of control variables, and $Z_i$ the "omitted" variable.

$$\text{[Long regression]} \quad Y_i = \alpha_L + \beta_L X_i + \lambda Z_i + W\beta_{WL} + e_i^S$$

$$\text{[Short regression]} \quad Y_i = \alpha_S + \beta_S X_i + W\beta_{WS} + e_i^L$$

$$\text{[Auxiliary regression]} \quad Z_i = \pi_0 + \pi_1 X_i + W\pi_{WP} + v_i$$

Then, the **Omitted variable bias formula** states that:

$$\underbrace{\beta_S}_{\text{Short}} = \underbrace{\beta_L}_{\text{Long}} + \underbrace{\lambda}_{\text{Omitted}} \cdot \underbrace{\pi_1}_{\text{Included}}$$

The OVB formula describes what happens to our coefficient of interest, $\beta$, as we include one additional variable $Z$ in the regression. We call $\lambda\pi_1$ the **omitted variable bias**. Direction of bias: multiply our guesses for the signs of $\lambda$ and $\pi_1$.

# Control variables

Control variables are additional variables (or covariates) **included in a regression**. We do this for **various reasons** (in decreasing order of importance):

- To remove selection bias / omitted variable bias
- To increase precision of our estimates
- To know about the (conditional/partial) correlation of other variables
- To better predict the outcome

Let's see *graphically* how control variables work!

# Your questions

Any questions?

...Remember – this is a safe space! Every question is useful!

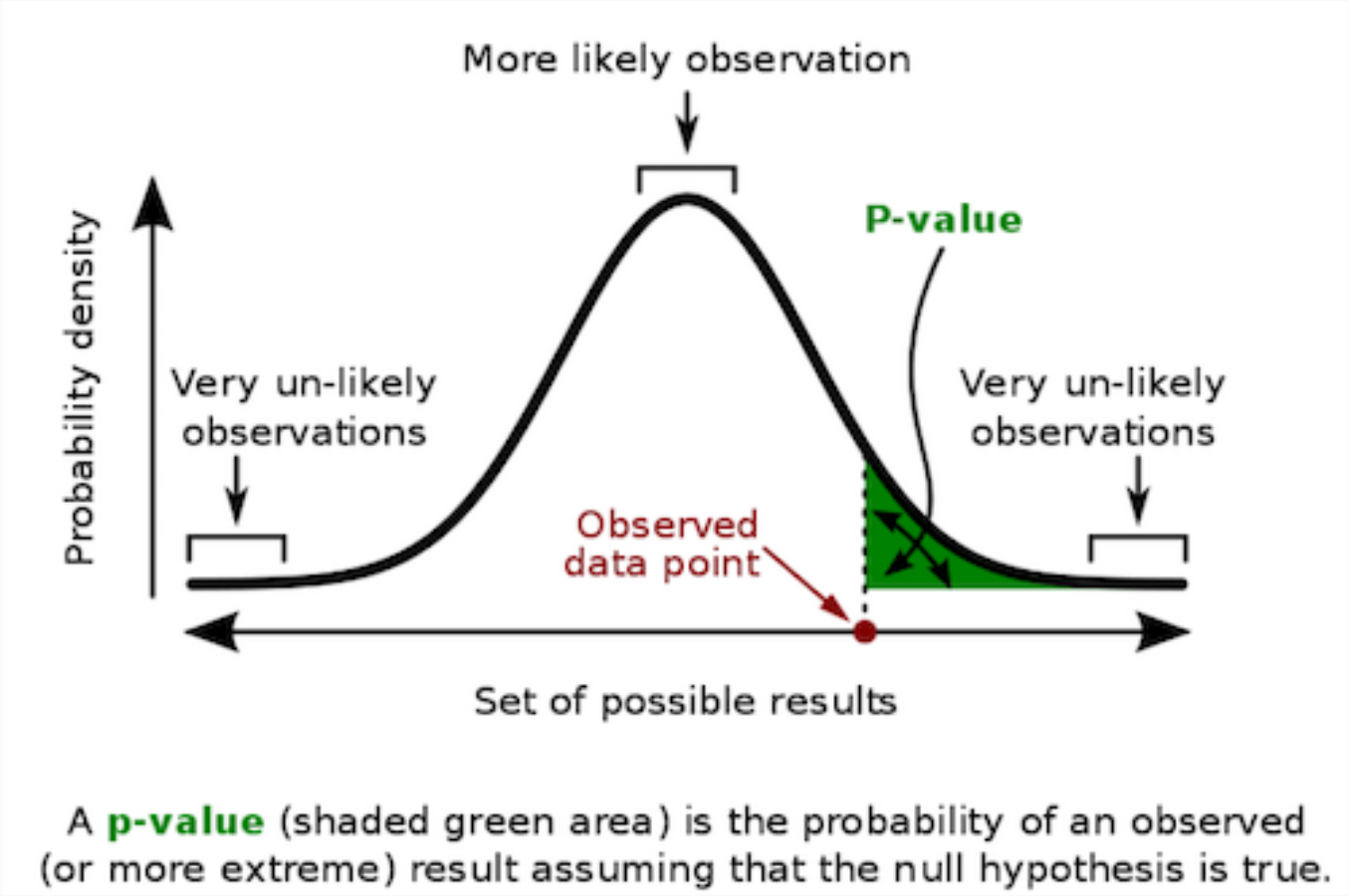
# Inference and hypothesis testing

# Hypothesis testing

We need a few ingredients:

- Random variables: Our estimator is a random variable (randomly drawn from population)
- Standard error: Random variables have a standard deviation, estimators have standard errors. This quantifies their uncertainty
- Statistics: The two keywords are the law of large numbers and the central limit theorem: The sum/mean of many random variables will follow normal distribution
- For hypothesis test: null and alternative hypothesis.
- We assume that the null hypothesis is true and then see how plausible results are, given the null is true.
- If they are implausible – we reject the null hypothesis! Otherwise: Fail to reject.

# It’s all connected




A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

# Hypothesis testing mini-cheatsheet

$$\left| \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} \right| \geq 1.96$$

$$\Leftrightarrow \left| \text{t-stat} \right| \geq 1.96$$

$$\Leftrightarrow \text{p-value} \leq 0.05$$

$$\Leftrightarrow 0 \notin \text{CI}$$

If you are testing the null hypothesis $H_0$: $\beta = 0$ (against the alternative hypothesis $H_0$: $\beta \neq 0$), then all of these are equivalent, and you can use any of these.

# Exam practice

# Practice exam question: 1a)

The Ministry of Truth is interested in a rumour that **air pollution could impact mental health**. One of the most harmful pollutants is fine particulate matter PM2.5, which comes from operations that involve the burning of fuels such as wood, oil, coal, etc. A research team is sent to investigate the rumour. The team **randomly** selects and surveys 19,920 people across 71 districts of the country. The key variable, Exposure $E_i$, is a **dummy variable** equal to 1 if the individual $i$ is exposed to a large amount of PM2.5 in the last 24 hours, and 0 otherwise. The team also conducts a standardised questionnaire to record **depressive symptoms** in the last month, called the Kessler Psychological Distress scale (K6). The questionnaire results in a score, $\text{Depression}_i$, that ranges from 0 to 24; and the higher the score, the more severe the depressive symptoms for individual $i$. The variable has a sample average of 2.96. **Running regressions with Depression $D_i$ as the dependent variable, you obtain the following results**:

# Practice exam question: 1a)

Dependent variable: Depression $_i$

| Regressor | (1) | (2) | (3) |
|---|---|---|---|
| Exposure $_i$ | 0.834 | 0.614 | 0.554 |
| | (0.032) | (0.045) | (0.042) |
| Exposure $_i \times$ Female $_i$ | | 0.065 | |
| | | (0.024) | |
| Female $_i$ | | −0.739 | −0.825 |
| | | (0.036) | (0.066) |
| $\text{Age}_i$ | | | 0.452 |
| | | | (0.132) |
| $\text{Age}_i^2$ | | | 0.524 |
| | | | (0.121) |

Notes: All estimations contain a constant term. Robust standard errors are in the parentheses. $\text{Age}_i$ is the age (years old) of individual $i$, and $\text{Age}_i^2$ is the square of $\text{Age}_i$.

# Practice Exam question: 1a)

**a)** Interpreting the coefficient in Column (1), a journalist, Katherine, claims: ”Since participants are randomly selected, we can infer that exposure to a large amount of PM2.5 does cause depression.”

**i.** Explain carefully why Katherine is wrong. Come up with two confounding variables, specifying the direction of bias(es) if there are any. Which assumption(s) would she need to impose for the causality claim to hold?

**ii.** What is the correct interpretation from Column (1) that Katherine should have made?

# (Detailed) Suggested Answer: 1a)

i. **Random selection is not the same thing as random assignment to treatment!** Survey respondents may be systematically different from each other in ways that are correlated with depression and pollution exposure. Therefore, the results from a regression can not be interpreted causally (and are biased). A priori, it is unclear in which direction the bias would go, but we could imagine that (**Only one explanation needed for exam**):
On rainy days, pollution is lower (-) and people may be reporting more depression symptoms (+), leading to downward bias. More wealthy people choose to live in less polluted areas (-) and they may have less depression (e.g., better access to mental health resources) (-), leading to upward bias.
For the regression causality claim to hold, we need to assume that: People exposed to pollution and those not exposed to pollution would have, on average, the same depression level, had they been exposed to the same level of pollution. In other words: **Both groups would have to have the same potential depression outcomes**.

ii. On average, people that were exposed to pollution had a 0.8 points higher score on the depression scale. The difference between the two groups is significant at the 5% level.

# Pratice Exam question: 1b)

**b)** Interpret column (2) of the regression table

**i.** A colleague notes the the coefficient on $\text{Female}_i$ is significant, and states: ”The effect of being female on depression is significantly different from zero”. Do you agree with the statement? Why or why not?

**ii.** How is pollution exposure related to depression, for men? And how for women?

# (Detailed) Suggested Answer: 1b)

**i. It is difficult to make such interpretations when interaction terms are involved.** Taking partial derivatives, the "effect" of being female is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Female}_i} = -0.739 + 0.065 \cdot \text{Exposure}_i$$

We can do inference (and test significance) at $\text{Exposure}_i = 0$ (just looking at the coefficient for female, $-0.739$ is significant). We can also do it for any other level of exposure, but for that we also need to take the other coefficient into account and cannot just use the table.

**Comment 1: Less relevant for the exam, but important to be aware of!**
**Comment 2: We can always do inference on the interaction term, which is significant here!**

**ii.** Taking partial derivatives, the "effect" of pollution exposure is:

$$\frac{\partial \text{Depression}_i}{\partial \text{Exposure}_i} = 0.614 + 0.065 \cdot \text{Female}_i$$

Hence, the "effect" for Males is 0.614 and the "effect" for Females is larger ($0.614 + 0.065 = 0.779$). The effect of pollution is also **signficantly** larger than for men, because the coefficient on the interaction term is significantly different from zero.

If time permits: Practice exam question, midterm 2022.