Lecture 7 - Modelling

Topics in Econometrics - M2 ENS Lyon

Vincent Bagilet

Introduction

Steps of an Econometrics Analysis

- **Design**: decisions of data collection and measurement
  - *eg*, decisions related to sample size and ensuring exogeneity of the treatment
- **Modelling**: define statistical models
  - *eg* selecting variables, functional forms, etc
- **Analysis**: estimation and questions of statistical inference
  - *eg* standard errors, hypothesis tests, and estimator properties

Goal of the session

- Explore some modeling
- Main focus in causal inference is often **identification**
- So far, we have: a good question and a convincing quasi-random allocation of the treatment
- Aim to give intuition about *some* key points and provide you with resources to learn more about them

What is a model?

- **Model**: formal representation of a theory $\rightarrow$ necessary simplified
- **Statistical model**: a mathematical model of the DGP generating a random sample of which we observe a realization
- **Regression model**:
  - Conditional distribution of $Y|X$:
  - Need to choose the **functional form** for $\mathcal{F}$
  - In particular, choosing which variables to include
  - Often informed by theory

Type of models

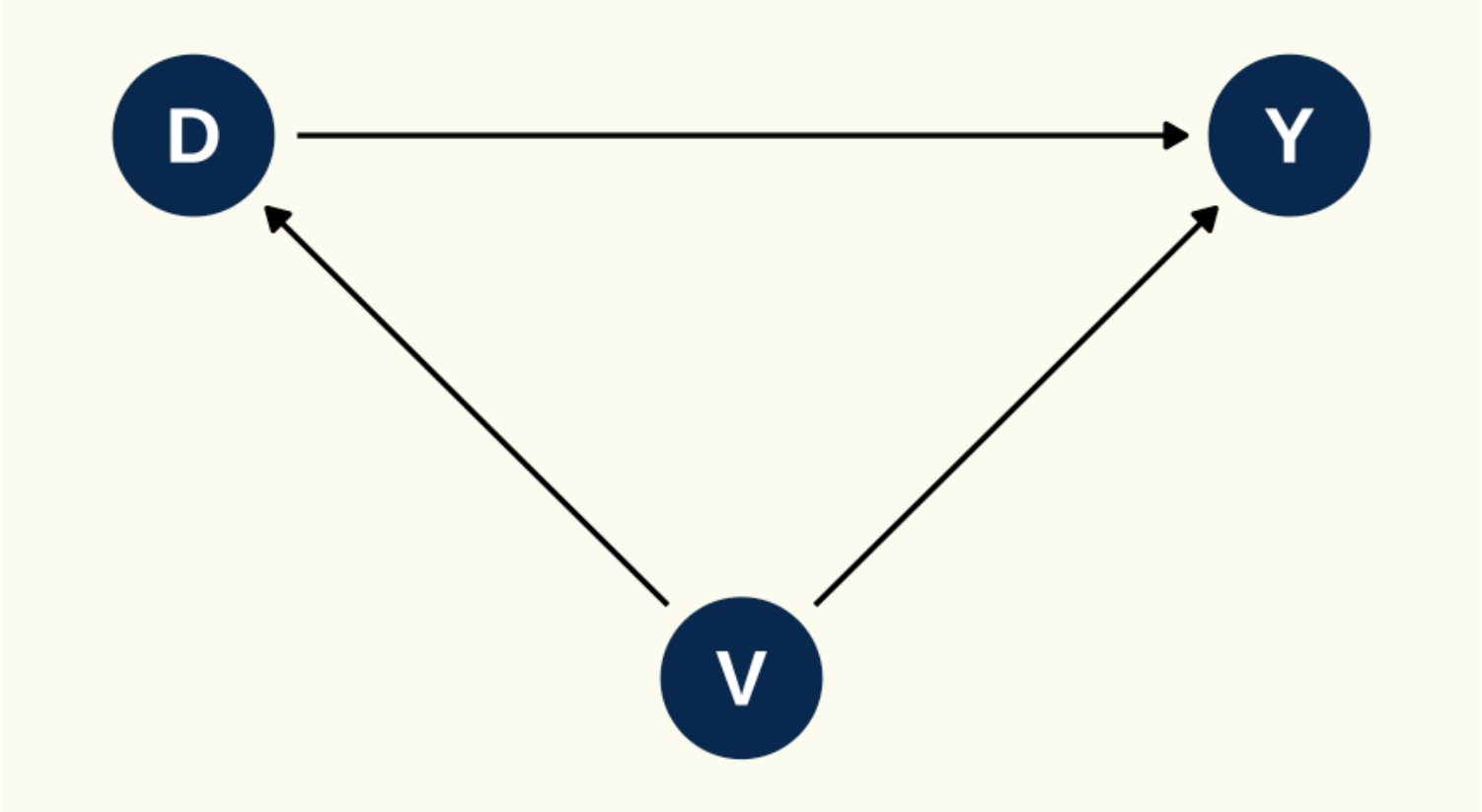
- **Parametric**:
  - $Y|X$ is defined by a finite number of parameters: $Y|X \sim \mathcal{F}(X)$
  - These parameters are **identifiable** when different parameters imply different distributions
- **Non-parametric**: makes no assumptions about a parametric distribution
- **Semi-parametric**: contains both a finite-dimensional set of parameters of interest and an infinite-dimensional nuisance component

Causal Graph Framework

Causal Graphs and DAGs

- Structural causal model framework, represented through **Directed Acyclic Graphs (DAGs)**
- Developed mainly in Pearl (2000)
- Potential outcomes framework and the structural causal model framework are complementary approaches to causal inference
- A DAG is a picture of the assumptions we make. It makes explicit:
  - Which variables we must adjust for
  - Which variables we must **not** adjust for

Components of DAGs



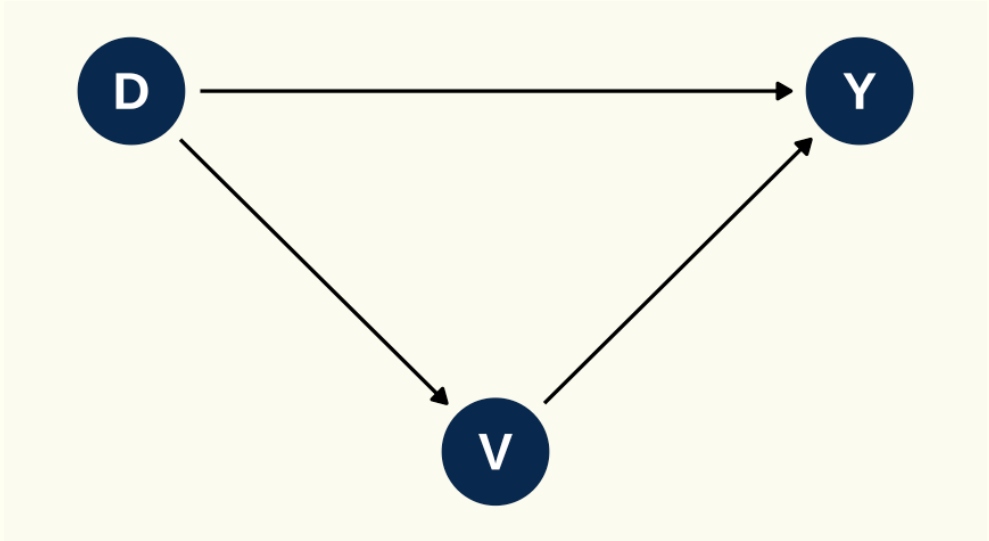
- Nodes are random variables (here, in blue if observed, in beige if unobserved)
- Arrows represent possible direct causal relationships
- A path is any sequence of adjacent edges, *regardless of the direction of the arrows*
  - There are two paths between D and Y: $D \rightarrow Y$ (direct) and $D \leftarrow V \rightarrow Y$

Components of DAGs

Open and closed paths

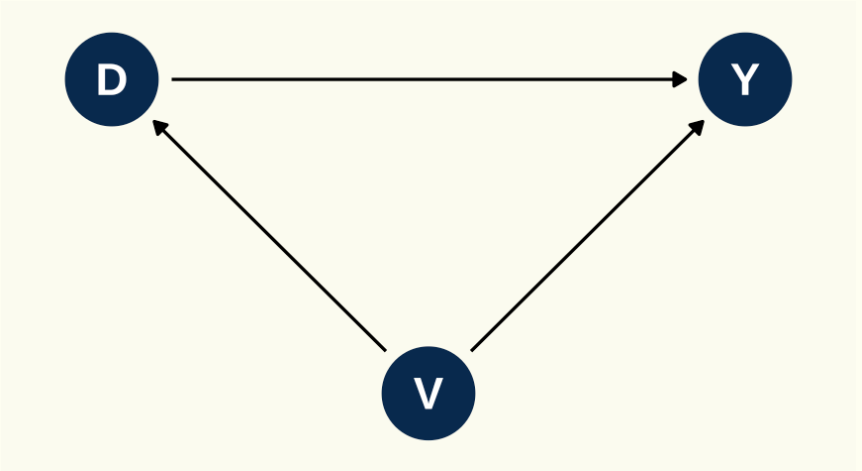
- A path is **open** if it can generate an association, **closed** otherwise
- Conditioning on a variable along a path may open or close it
- **Acyclic**: no directed path leaving a variable comes back to it
  - $\Rightarrow$ no simultaneity, no feedback loop

Mediating path



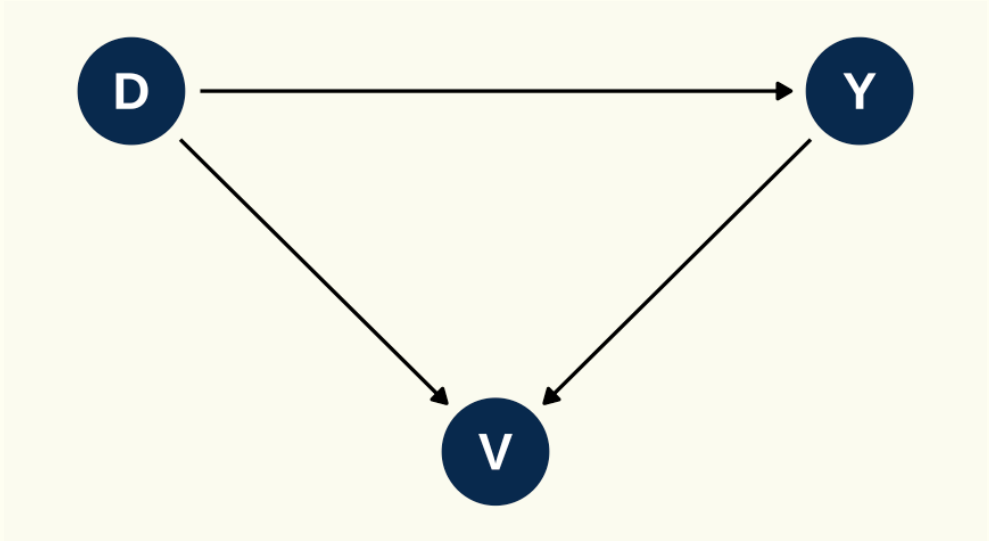
- D causally affects Y through V: the path is part of the effect
- Not conditioning on V identifies the **total** effect
- Conditioning on $V \Rightarrow$ identifies the **direct** effect only

Confounding path



- V causally affects both D and Y: it creates a **non-causal** association between D and Y
- The path is open by default
- Conditioning on V blocks it and identifies the total causal effect

Colliding path



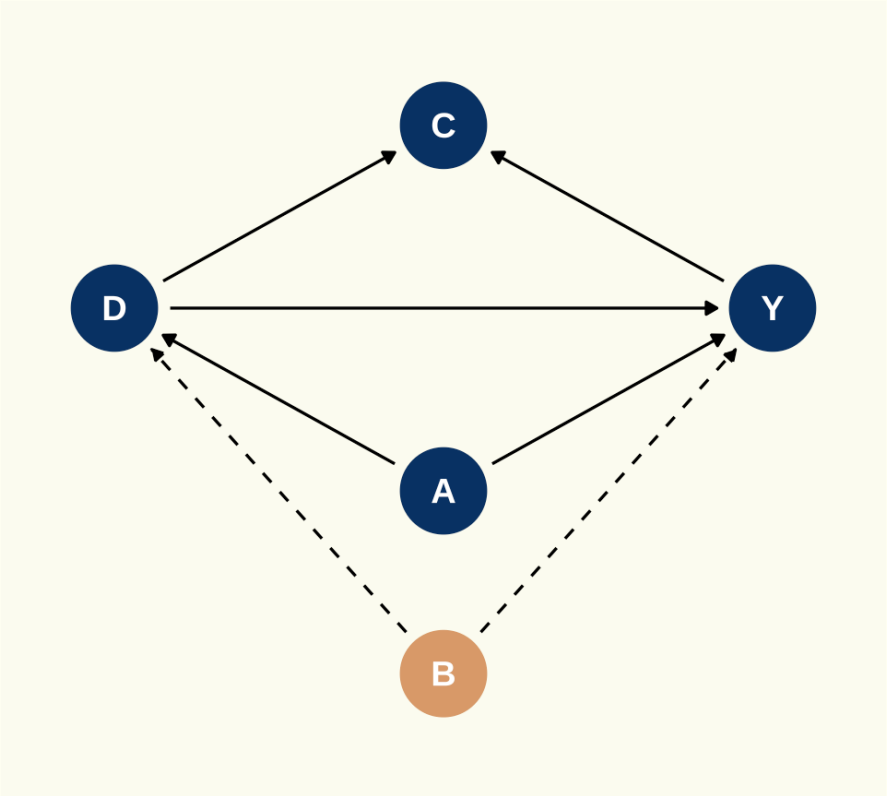
- V is caused by both D and Y
- The path through V is blocked by default
- Conditioning on V, or a descendant, opens it and *creates* an association

The three structures

Structure	Path	Open by default?	Condition on V?
Mediating	$D \to V \to Y$	Yes, it is part of the effect	Only for the *direct* effect
Confounding	$D \leftarrow V \to Y$	Yes	Yes, it blocks the path
Colliding	$D \to V \leftarrow Y$	No	No, it would open the path

- **Back-door path** from D to Y: any path between them whose first edge is an arrow pointing to D
- One should condition on confounders, never on mediators or colliders

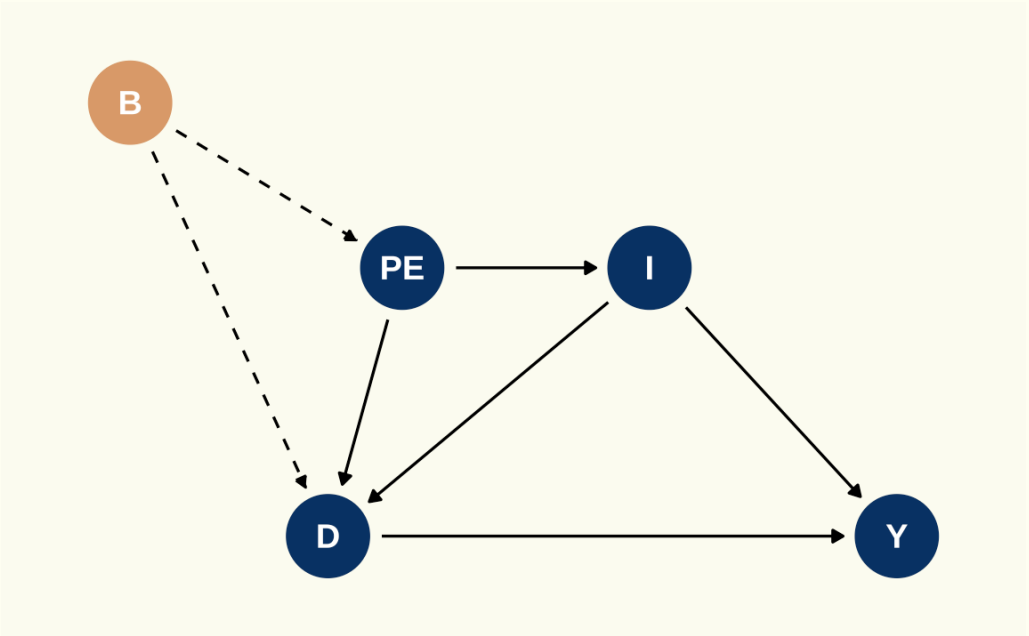
An example summary



- Four paths linking D to Y:
  - $D \to Y$: the direct causal path
  - $D \leftarrow A \to Y$: back-door, blocked if we condition on A
  - $D \leftarrow B \to Y$: back-door, open as B is unobserved
  - $D \to C \leftarrow Y$: colliding, blocked unless we condition on C

⚠ The absence of an arrow is not an absence of assumption: it is a structural **exclusion restriction**

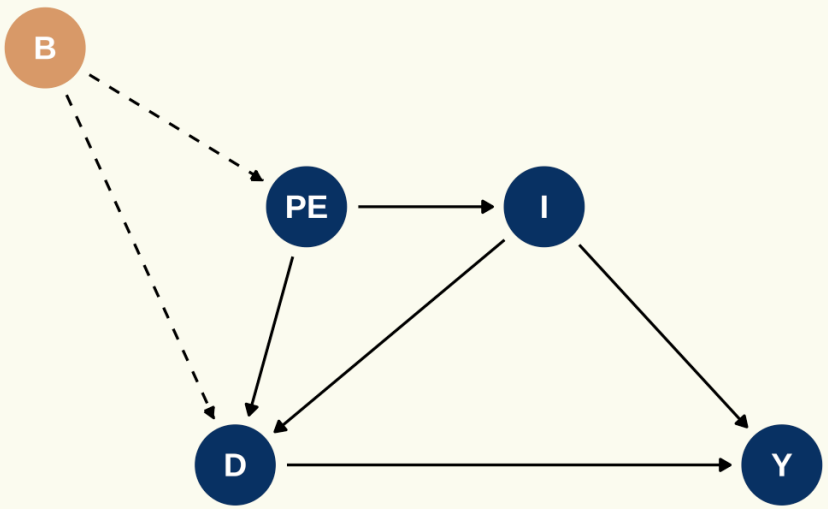
An example: college and earnings



- Effect of college D on earnings Y (Cunningham 2021)
- PE parental education, I family income
- B unobserved background (ability, environment, networks)
- Each arrow is a **direct** causal relationship we are willing to assume
- There is no $B \rightarrow Y$ arrow: background matters only through PE, I and D

Reading the paths

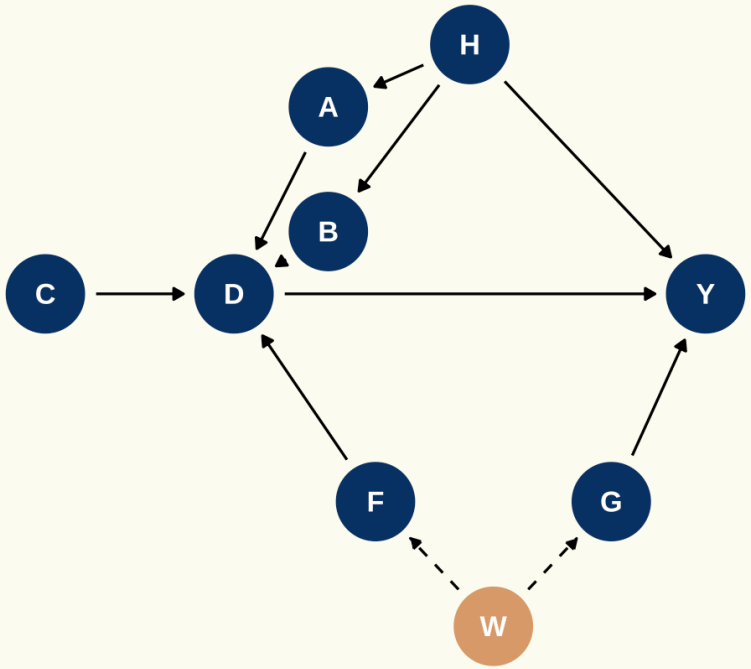
- Four paths between D and Y:
  - $D \rightarrow Y$, the causal effect we want
  - $D \leftarrow I \rightarrow Y$
  - $D \leftarrow PE \rightarrow I \rightarrow Y$
  - $D \leftarrow B \rightarrow PE \rightarrow I \rightarrow Y$
- The last three are back-door paths



- No collider on any of them $\Rightarrow$ all three are open
- All end with $I \rightarrow Y \Rightarrow$ conditioning on income alone closes them

Blocking back-door or instruments

- If all back-door paths are blocked, the observed association identifies the total causal effect
- Here, three back-door paths:
  - $D \leftarrow A \leftarrow H \rightarrow Y$
  - $D \leftarrow B \leftarrow H \rightarrow Y$
  - $D \leftarrow F \leftrightarrow G \rightarrow Y$
- We block all of them by conditioning on:
  - H, and either F or G
  - A and B, and either F or G
- Or use an instrument: C is a candidate here, F too after conditioning on G



Pros and cons

- \+ Make every assumption explicit, arrows *and* missing arrows
- \+ Non-parametric: no functional form assumed
- \+ Give a mechanical rule to pick the adjustment set
- – Say nothing about functional form, monotonicity or continuity (*ie* nothing about LATE or RD)
- – Do not display effect heterogeneity nor compliance types
- – Cannot represent simultaneity or feedback loops

Take-away message

A DAG does not identify anything *for* us, it forces us to state what we assume and tells us what follows from it

Gauss–Markov conditions

Modelling assumptions matter

- OLS valid under a set of assumptions: the Gauss–Markov conditions
- If these assumptions, or any modelling one, do not hold we cannot make reliable inference
- Let's see why!
- **Modelling matters**: specification choices affect the results of our studies and our ability to make reliable inference

Gauss-Markov conditions

Assumption	Idea	If violated
1. Linearity	Model linear in its parameters	Estimates misspecified ⇒ biased/inconsistent
2. No perfect collinearity	$(X'X)^{-1}$ exists	Coefficients not identifiable
3. Exogeneity	$E[u_i \mid X_i] = 0$	Estimator biased
4.a. Independent errors	$Cov(u_i, u_j \mid X) = 0$ for $i \neq j$ (*eg* no autocorrelation)	Invalid inference
4.b. Homoskedasticity	$Var(u_i \mid X_i) = \sigma^2 = \text{cst}$	Inefficient

- 4.a + 4.b = spherical errors

Implications

- Finite sample properties:
  - 1 $\rightarrow$ 3: $\widehat{\beta}_{OLS}$ unbiased
  - 1 $\rightarrow$ 4: $\widehat{\beta}_{OLS}$ efficient (Best Linear Unbiased Estimator)
- Asymptotically:
  - Unbiased
  - Normally distributed
  - Efficient

Normally distributed estimator

- Required for making inference (*eg* computing confidence intervals or p-values)
- Additional assumption: normal errors $\Rightarrow$ $\widehat{\beta}_{OLS}$ normally distributed
- If errors non-normal
  - Alternative way to compute SE (*eg* bootstrap)
  - If n is large enough, Central Limit Theorem (CLT) + Weak Law of Large Numbers (WLLN) $\Rightarrow$ $\widehat{\beta}_{OLS}$ approximately normal

Controls

Types of controls

- Two types of controls:
  1. Those necessary for identification
     - To ensure the exogeneity of the treatment
  2. Those that improve precision
     - To better explain y: $\nearrow$ R^2 $(= 1 - FUV)$ and $\searrow$ thus σ_u^2
- Adjusting for pre-treatment covariates may
  - $\searrow$ variation in y, $\nearrow$ precision
  - $\searrow$ variation in x, $\searrow$ precision

Controls and DAGs

Role of the control	In the DAG	Adjusting for it
Confounder	$X \leftarrow C \rightarrow Y$	Removes bias
Predictor of y	$B \rightarrow Y$ only	Does not change bias, more precise
Mediator	$X \rightarrow M \rightarrow Y$	Removes part of the effect
Collider	$X \rightarrow S \leftarrow U \rightarrow Y$	Creates bias

Saturated models

- A model is **saturated** when it has one parameter per combination of the regressors
  - *eg* with x binary, $y_i = \alpha + \beta x_i + u_i$ already is: two parameters, two cells
- \+ No restriction left on $\mathbb{E}[y|X]$: **nothing to misspecify**
- – Quickly needs a lot of controls (and continuous variables need to be binned first)
- – **Removes a lot of variation** in the treatment

Impact of modelling on design

- Modelling choices (*eg* which controls to include) can affect the **design** of our studies
- They can affect **amount of identifying variation** available:
  - As illustrated through the FWL theorem, controls can remove variation in the treatment variable
- → can affect statistical power and create exaggeration

Limited outcome models

Limited outcome models

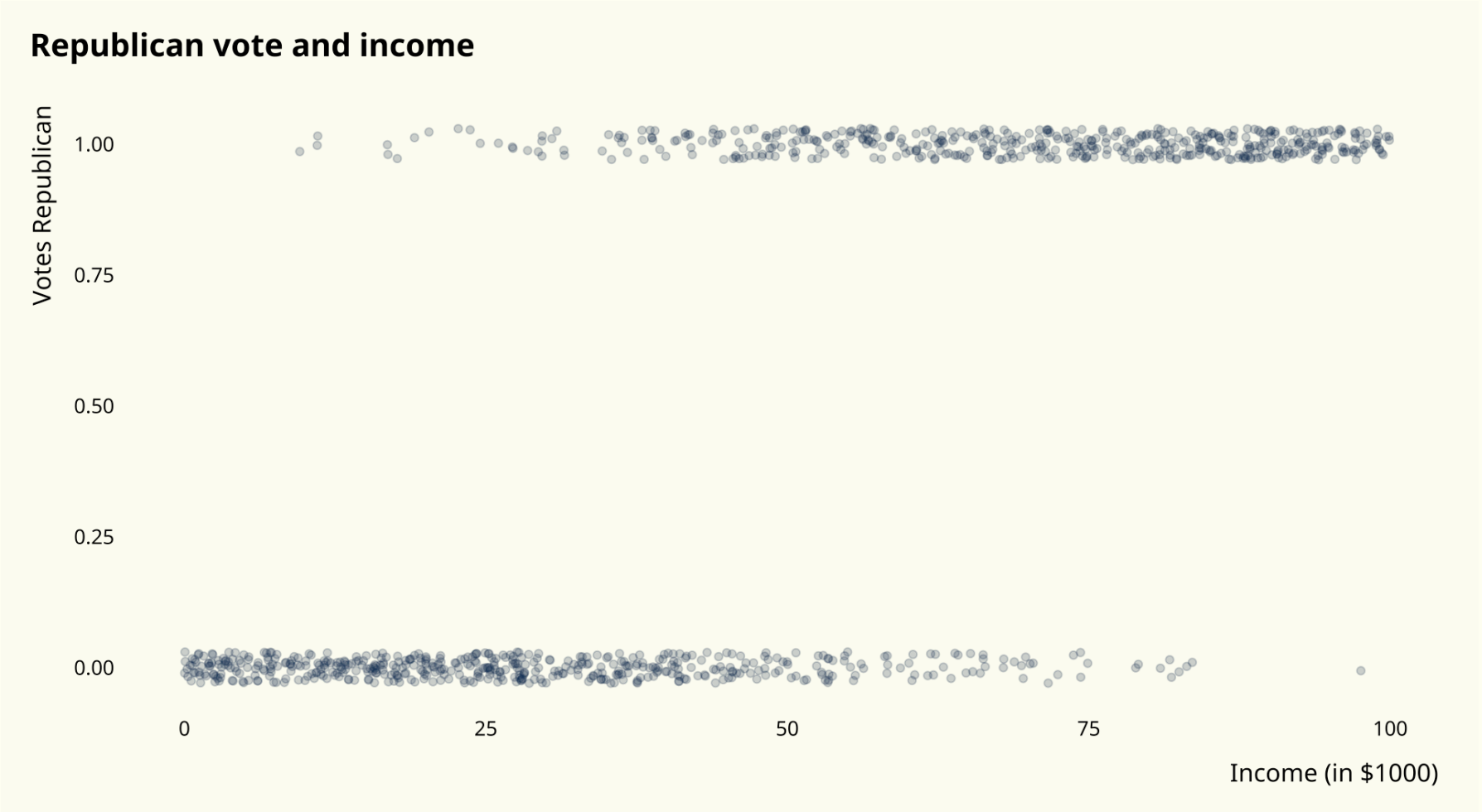
- Often, y is limited: **binary**, **categorical**, **censored**, etc
- Linear regression ignores the constraints on y:
  - Predictions fall outside the range of y
  - *eg* if the outcome is binary, predictions might fall out of $[0, 1]$
- Use **Generalized Linear Models** (GLMs):
  - A **link function** g maps $\mathbb{E}[y|X]$ onto the real line
  - $g(\mathbb{E}[y|X]) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$
  - The differences across models come from the use of different g functions

Types of limited outcomes

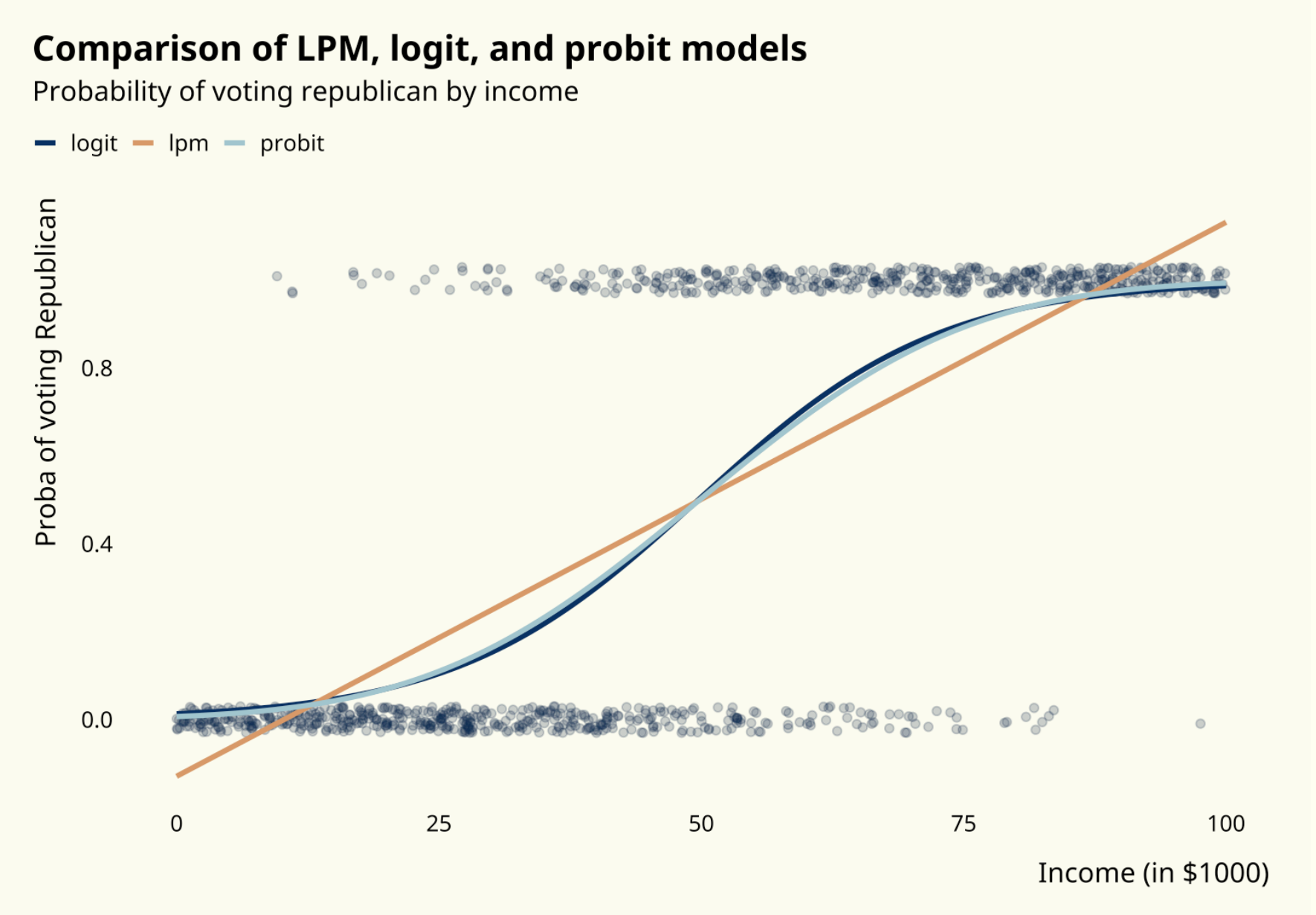
Limited y	**Example**	**Regression model**
Binary	$y \in \{0, 1\}$	Probit, logit, …
Categorical	$y \in \{a, b, c, ...\}$	Multinomial logit, …
Count	$y \in \{0, 1, 2, 3, ...\}$	Poisson, negative binomial, …
Censored	*eg* $y = \max\left(0, y^*\right)$	Censored regression models (*eg* tobit)

Binary outcome: example

- The outcome follows a Bernoulli distribution: $y|X \sim Ber\left(p\right)$
- We want to express $p = P\left[y = 1|X\right]$ as a function of X and β: $p = g^{-1}\left(X'\beta\right)$



Example fits



Linear Probability Model

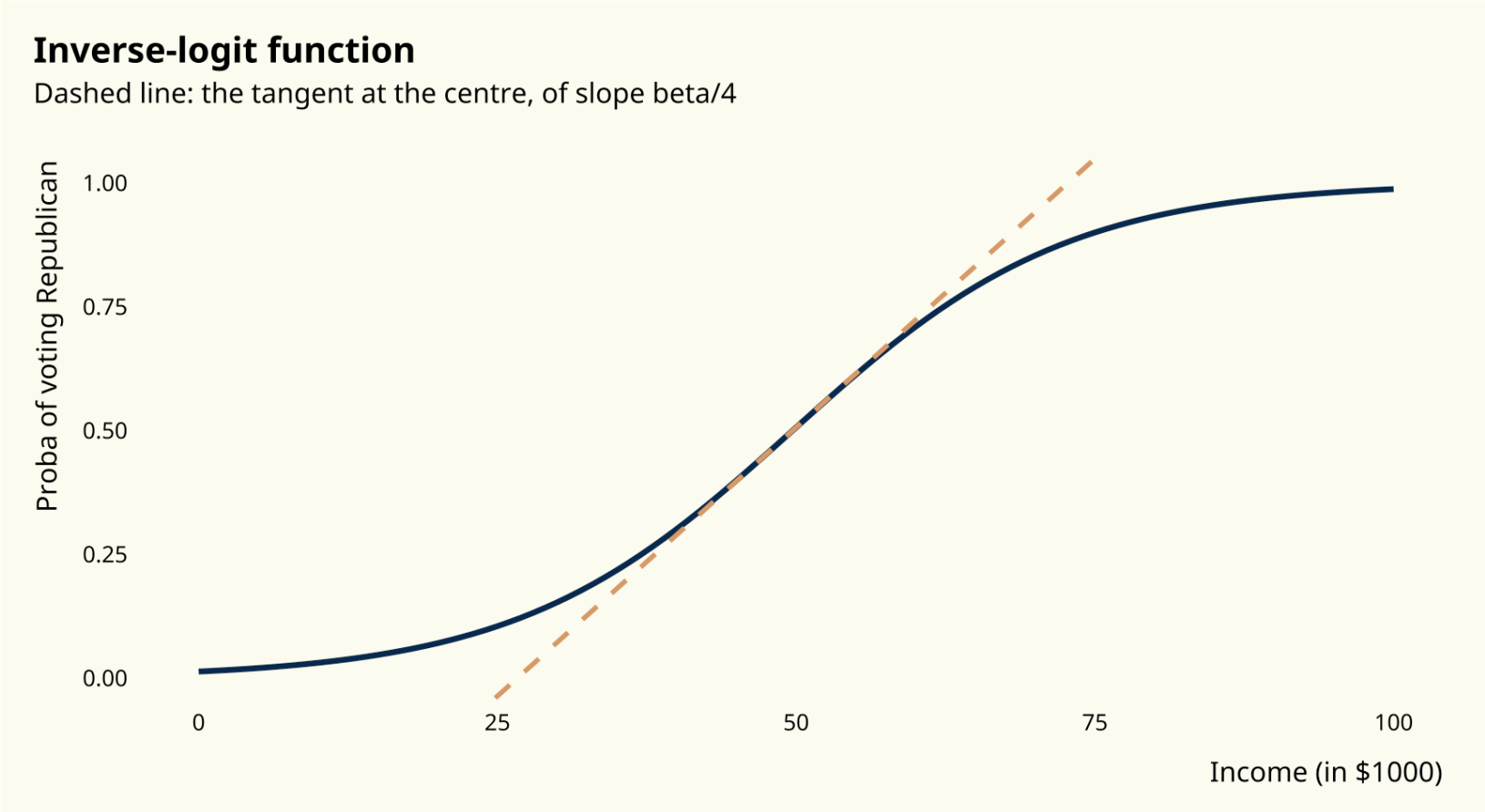
- Link function g: $x \mapsto x$, the identity
- This is an OLS on a binary outcome
- Why not stick with this?
  - Predicted probabilities are *not* constrained to $[0, 1]$
  - $P[y = 1|X]$ is highly non-linear in X: the effect of an increase of 1 in X is not the same if $P[y = 1|X] = 0.5$ or 0.99
  - LPM is thus **biased** and **inconsistent**
- LPM fine for an **ATE**: with a saturated model it *is* the difference in means

Logit and probit

- In a **logit** model, $g: x \mapsto log\left(\frac{x}{1-x}\right)$
- In a **probit**, g = quantile function of the standard normal (the inverse of the CDF of the normal distribution)
- These models are **non-linear** regression models
- Probit and logit are very similar, except in the tails, *ie* for very rare events
- Rule of thumb: probit coefs $\simeq$ logit coefs divided by 1.6
- A logit/probit point estimate gives the **direction**, not the magnitude
- The marginal effect depends on X: effects largest in the middle of the distribution

Divide-by-4 rule

- The inverse-logit curve is steepest at $p = 0.5$, where its **slope is** $\beta/4$
- $\Rightarrow$ $\beta/4$ is an upper bound of the relation, in percentage points
- Ok approximation when $p \simeq 0.5$



Estimation results

	Outcome: votes Republican		
	LPM	**Logit**	**Probit**
Income (in 1000 USD)	0.01261	0.08684	0.04939
	(0.00037)	(0.00508)	(0.00253)
Observations	1000	1000	1000

Standard errors in parentheses. Intercept estimated but not reported.

- Only the LPM is already in percentage points
- The others are not directly interpretable, they need to be “converted”

Interpretating logit results

- Three ways to turn $\widehat{\beta}_{logit} = 0.087$ into a probability:
  - **Divide by 4**: maximum relationship between income and republican share (0.022 pp)
  - **Average Marginal Effect**: 0.0096
- The Average Marginal Effect averages the flat parts of the curve
- The LPM slope is approximately equal to the AME $\Rightarrow$ the LPM is often fine

Odds

- The odds of an event compare how likely it is to happen to how likely it is not: $\frac{p}{1-p}$
  - *eg* $p = 0.75 \Rightarrow$ odds of 3, *ie* it is 3 times more likely to happen than not
- Odds map to $[0, +\infty)$ and log odds to $\mathbb{R}$ while probabilities only take values in $[\,0, 1\,]$
- A logit assumes that the log odds are linear in x: $\log \frac{p}{1-p} = \alpha + \beta x$
- *Why use this?* A probability cannot be linear in x without leaving $[\,0, 1\,]$, the log odds can

Odds ratio

- A one unit larger x is associated with β larger log odds, *ie* odds are multiplied by e^{β}
- e^{β} is the **odds ratio**
  - *ie* here a 1000 USD larger income is associated with 9% larger odds of voting Republican, wherever we start from
- *Why use this?* Summary of $\widehat{\beta}$ that does not depend on where we start from
- The odds ratio is the same everywhere, the change in probability is not
- Yet, odds ratios are **not straightforward to interpret**

Exercise

- Switzerland, 1981: how does a married woman's labor force participation relate to the household's other income?

```
data("SwissLabor", package = "AER")
```

1. Regress `participation` on `income` using `glm`, with a logit then a probit link
2. How would you interpret the coefficient?
3. Run the LPM and compare the results

Categorical outcome

- Sometimes y is **categorical** but unordered: mode of transport, occupation, brand bought
- Use a **multinomial logit**: essentially a logit for each category, modeled against a reference category
- $\log \frac{P[y = j|X]}{P[y = \text{ref}|X]} = X'\beta_j \Rightarrow$ one set of coefficients per category
- Why use this? The probabilities of all categories must sum to one, separate logits would not impose it
- Relies on **Independence of Irrelevant Alternatives** (IIA): the relative odds of two categories do not depend on the others available
- `nnet::multinom()`, or `MASS::polr()` when the categories are ordered

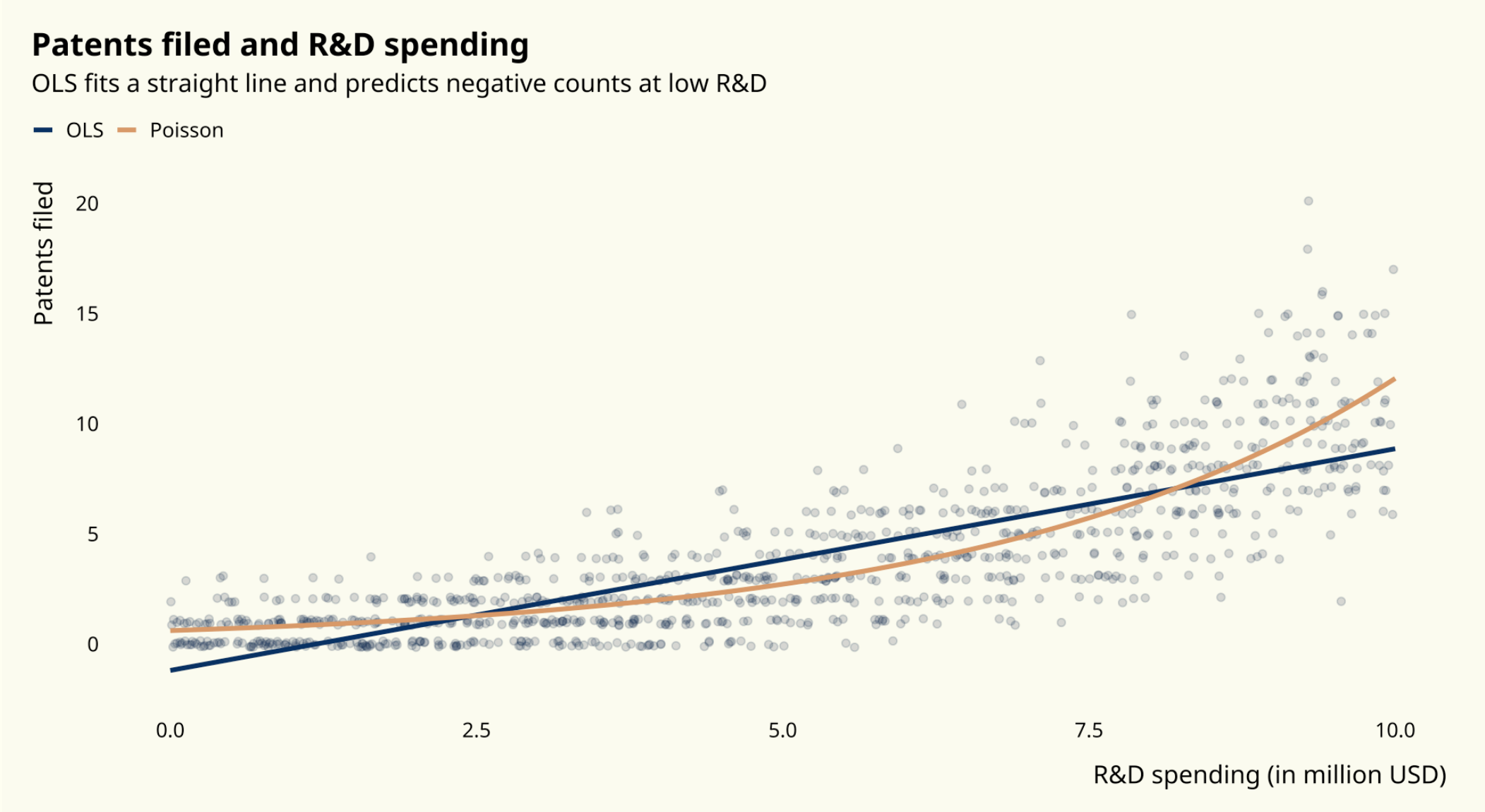
Exercise

- US bank, 1970s: how does education relate to the kind of job someone holds?

```
data("BankWages", package = "AER")
```

1. Regress `job` on `education` using `nnet::multinom()`
2. Which category are the two sets of coefficients compared to?
3. Exponentiate them. How do the relative odds move with a year of schooling?
4. `job` is ordered. Run `MASS::polr()` too: what do you get instead?

Count outcome: example



Poisson model

- Many outcomes are **counts**: patents, deaths, hospital admissions, crimes, trips
- We may want to model them
- A **Poisson regression** assumes $y|X \sim \text{Pois}(\lambda)$
- It assumes that the link function g is the logarithm $\ln$, *ie* $\ln\left(\mathbb{E}\left[y|X\right]\right) = \beta_0 + \beta_1 X$
- The expected count is therefore multiplicative: $\mathbb{E}\left[y|X\right] = e^{\beta_0} \times e^{\beta_1 x_1} \times \ldots$

Why not a simple OLS?

- OLS fits a straight line to a multiplicative relationship: **predictions go below zero**
- It reports a single slope, when the change in y depends on the level of y
- **Counts are heteroskedastic** by construction
  - In a Poisson model $\mathbb{V}\left[y|X\right] = \mathbb{E}\left[y|X\right] \Rightarrow$ default SEs are wrong
- $\log\left(y\right)$ is not defined at $y = 0$

Interpreting a Poisson coefficient

	Outcome: patents filed	
	OLS	**Poisson**
R&D (in million USD)	1.006	0.298
	(0.024)	(0.007)
Observations	1000	1000

Standard errors in parentheses. Intercept estimated but not reported.

- $\widehat{\beta} = 0.3 \Rightarrow$ one million more of R&D multiplies the expected count by $e^{\widehat{\beta}} = 1.35$
- *ie* 35% more patents, whether the firm expects 1 or 10 of them
- Rule of thumb: for a *small* β, $e^{\beta} - 1 \simeq \beta$
- In levels, OLS reports the same 1.01 patents per million, wherever we stand
  - The Poisson adds 0.29 patents from 1 to 2 million of R&D, and 2.31 from 8 to 9

Negative binomial

- A Poisson model has a **single parameter** and imposes $\mathbb{V}\left[y|X\right] = \mathbb{E}\left[y|X\right]$
- Counts are usually more spread out than that: this is overdispersion
- Thus, better to use a **negative binomial** model
- It adds a parameter and allows $\mathbb{V}\left[y|X\right] > \mathbb{E}\left[y|X\right]$
- Interpreting: "a 1-unit difference in X is associated with a β-unit difference in the log of the expected count"
    - *ie* exactly as in a Poisson: the extra parameter changes the variance, not the mean

Estimation

- These models are **non-linear** (except LPM) $\rightarrow$ they cannot be estimated by OLS
- Estimated by **maximum likelihood**: we assume a distribution for $y|X$, then pick the β that makes our data the most likely
- A logit assumes $y|X \sim Ber\left(p\right)$, a Poisson $y|X \sim \text{Pois}\left(\lambda\right)$
- The distribution choice affects and actually defines the standard errors

Pseudo-Poisson Maximum Likelihood

- **PPML**: fitting $\mathbb{E}\left[y|X\right] = e^{X'\beta}$ by maximising a Poisson likelihood, without assuming y is Poisson
- Ok because maximising it = solving $\sum_i X_i\left(y_i - e^{X_i'\beta}\right) = 0$
- Only depends on the mean $\Rightarrow$ $\widehat{\beta}$ is consistent as soon as the mean is right
- *Why use it rather than a log*? $\log\left(y\right)$ drops the zeros, $\log\left(1+y\right)$ depends on the units of y
  - PPML **needs no transformation**: y only has to be non-negative
- $\widehat{\beta}$ then reads directly as a percentage difference in $\mathbb{E}\left[y|X\right]$
- Since s.e. computed from an incorrect likelihood, we need to report robust s.e.

Exercise

- UK, 1969-1984: how do monthly driver deaths relate to the compulsory seat belt law?

```
seatbelts <- as_tibble(Seatbelts)
```

1. Regress `DriversKilled` on `law` and `PetrolPrice` using `glm`, with a log link
2. How would you interpret the coefficient on `law`?
3. Compare the mean and variance of `DriversKilled`. Do Poisson assumptions make sense?
4. Run a negative binomial with `MASS::glm.nb()` and compare the results

Summary

Summary

- Many choices when building regression models
- Choices not straightforward and can affect the **credibility** of our estimates
- The **Causal Graph Framework** can be particularly helpful to think about the relations between variables and causality
- We often deal with **limited outcomes**. That needs to (and can) be accounted for using several types of models.

References

Aihounton, Ghislain B D, and Arne Henningsen. 2021. “Units of Measurement and the Inverse Hyperbolic Sine Transformation.” *The Econometrics Journal* 24 (2): 334–51. https://doi.org/10.1093/ectj/utaa032.

Chen, Jiafeng, and Jonathan Roth. 2024. “Logs with Zeros? Some Problems and Solutions.” *The Quarterly Journal of Economics* 139 (2): 891–936. https://doi.org/10.1093/qje/qjad054.

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press. https://doi.org/10.2307/j.ctv1c29t27.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Santos Silva, J. M. C., and Silvana Tenreyro. 2006. “The Log of Gravity.” *The Review of Economics and Statistics* 88 (4): 641–58. https://doi.org/10.1162/rest.88.4.641.