Lecture 7 - Modelling and Analysis

Topics in Econometrics - M2 ENS Lyon

Vincent Bagilet

2025-10-14

Introduction

Short feedback form

https://forms.gle/wxwZNFXyPxprLELT7

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

- Main focus in causal inference is often **identification**
- So far, we have: a good question and a convincing quasi-random allocation of the treatment
- How do we make inference for the population?
- We make causal claims based on significance $\Rightarrow$ need **modelling assumptions to hold** and **reliable SEs**
- Aim to give intuition about *some* key points and provide you with resources to learn more about them

Modelling

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

Exercise

- Generate fake data and analyse the impact of violations of some of these assumptions
    1. Non-linearity
    2. Perfect collinearity
    3. Endogeneity
    4. Autocorrelation
    5. Heteroskedasticity
    6. Non-normal errors

Non-linear

Code | Regression table | Graph

```r
n <- 1000
alpha <- 10
beta <- 2
mu_x <- 2
sigma_x <- 1
sigma_u <- 2

data_non_linear <- tibble(
  x = rnorm(n, mu_x, sigma_x),
  u = rnorm(n, 0, sigma_u),
  y = alpha + beta*x^2 + u
)

reg_non_linear <- lm(y ~ x, data = data_non_linear)

# list("Non-linear" = reg_non_linear) |>
#   modelsummary(gof_omit = "IC|Adj|F|RMSE|Log")
```

Collinearity

Code | Perfect collinearity | Almost perfect collinearity

```r
gamma <- 0.2

data_collin <- tibble(
  x = rnorm(n, mu_x, sigma_x),
  w = 0.3*x,
  u = rnorm(n, 0, sigma_u),
  y = alpha + beta*x + gamma*w + u
)

reg_collin <- lm(y ~ x + w, data = data_collin)

# list("Perfect collin." = reg_collin) |>
#   modelsummary(gof_omit = "IC|Adj|F|RMSE|Log")

data_almost_collin <- tibble(
  x = rnorm(n, mu_x, sigma_x),
  w = 0.3*x + rnorm(n, 0, 0.01),
  u = rnorm(n, 0, sigma_u),
  y = alpha + beta*x + gamma*w + u
)

reg_almost_collin <- lm(y ~ x + w, data = data_almost_collin)

# list("Almost perfect collin." = reg_almost_collin) |>
#   modelsummary(gof_omit = "IC|Adj|F|RMSE|Log")
```

Endogeneity

Code | Regression table

```r
data_endog <- tibble(
  x = rnorm(n, mu_x, sigma_x),
  u = 0.5 * x + rnorm(n, 0, sigma_u),
  y = alpha + beta*x + u
)

reg_endog <- lm(y ~ x, data = data_endog)

# list("Endogeneity" = reg_endog) |>
#   modelsummary(gof_omit = "IC|Adj|F|RMSE|Log")
```

Autocorrelation

Code | Regression table

```r
u_auto <- numeric(n)
for (i in 2:n) u_auto[i] <- 0.9*u_auto[i-1] + rnorm(1, 0, sigma_u)

data_auto <- tibble(
  x = rnorm(n, mu_x, sigma_x),
  u = u_auto,
  y = alpha + beta*x + u
)

reg_auto <- lm(y ~ x, data = data_auto)

# reg_auto |>
#   modelsummary(gof_omit = "IC|Adj|F|RMSE|Log", vcov = c("classical", "HAC"))
```

Heteroskedasticity

Code | Regression table | Graph

```r
data_heterosked <- tibble(
  x = rnorm(n, mu_x, sigma_x),
  u = rnorm(n, 0, sigma_u + x^2),
  y = alpha + beta*x + u
)

reg_heterosked <- lm(y ~ x, data = data_heterosked)

# reg_heterosked |>
#   modelsummary(gof_omit = "IC|Adj|F|RMSE|Log", vcov = c("classical", "robust"))
```

Non-normal errors

Code | Regression table

```r
df <- 2

data_non_normal <- tibble(
  x = rnorm(n, mu_x, sigma_x),
  u = rt(n, df), #fatter tails
  y = alpha + beta*x + u
)

reg_non_normal <- lm(y ~ x, data = data_non_normal)

# reg_non_normal |>
#   modelsummary(gof_omit = "IC|Adj|F|RMSE|Log", vcov = c("classical", "bootstrap"))
```

Limited outcome models

- Often, y is limited: **binary**, **categorical**, **censored**, etc
- Linear regression is not appropriate:
  - It does not take the constraints on y into account
  - *ie* wrongly assumes linearity and errors non-normal
- Use **Generalized Linear Models** (GLMs):
  - Idea: uses an invertible link function g to transform a limited y into a continuous variable
  - $g\left(\mathbb{E}\left[y|X\right]\right) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$

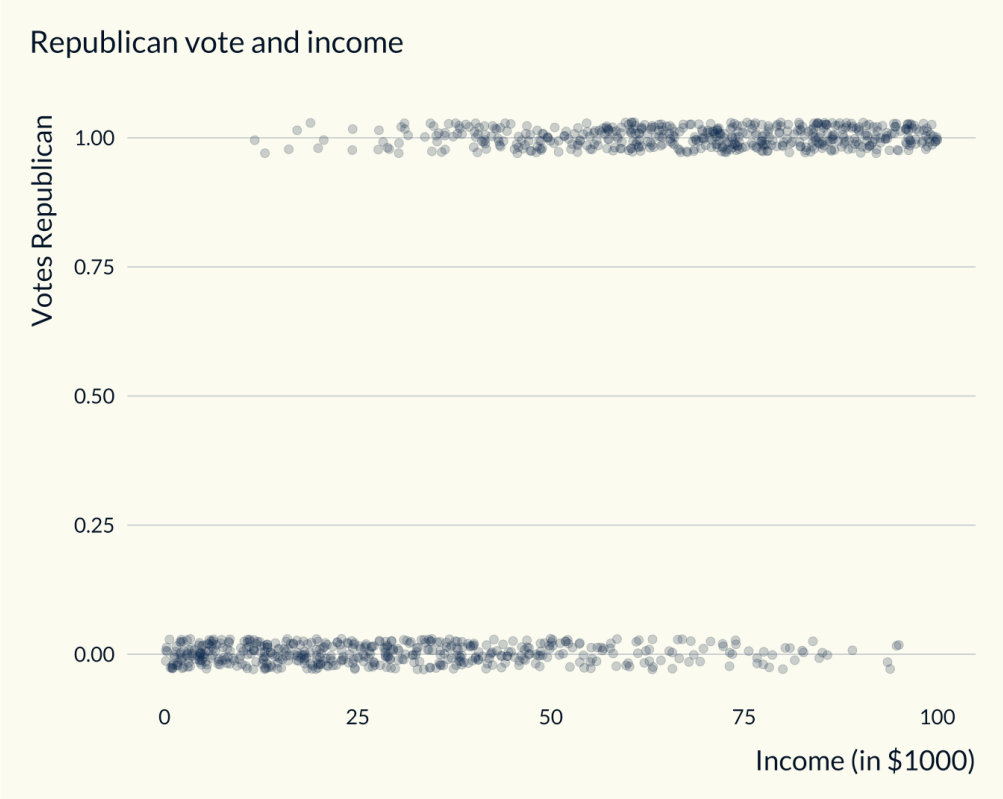
Limited outcome models

Limited y	Example	Regression model
Binary	$y \in \{0, 1\}$	Probit, 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: $y|X \sim Ber(\pi)$
- A regression model expresses the conditional probability $\pi = P[y = 1|X]$ as a function of X and β: $\pi = g^{-1}(X'\beta)$



Binary outcome

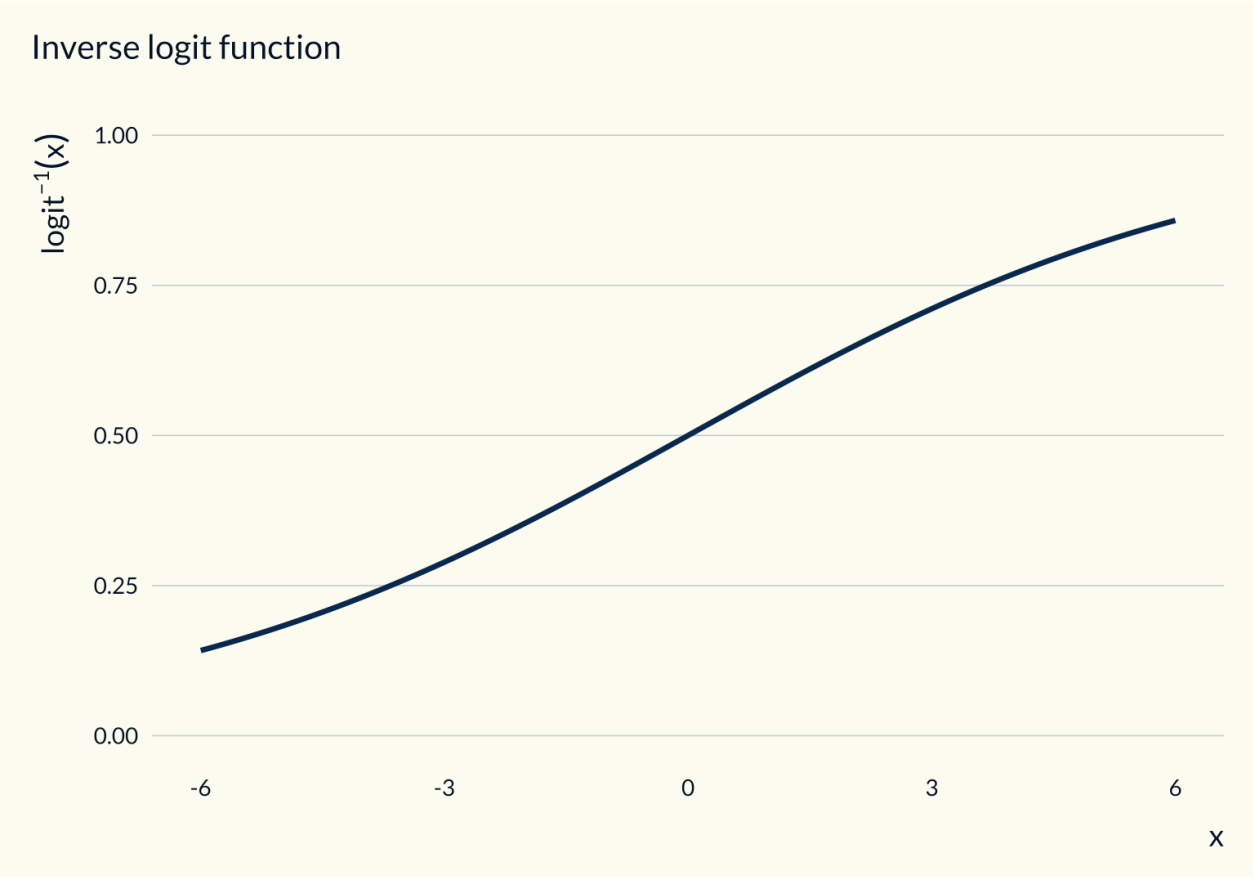
Models

- **Linear Probability Model**
  - $g: x \mapsto x$
  - Almost always yields biased and inconsistent estimates
- **Logistic regression model**
  - $g: x \mapsto log\left(\frac{x}{1-x}\right)$, the logit function
- **Probit regression model**
  - g: probit, *ie* the quantile function of the standard normal distribution (and g^{-1} is the CDF of the normal)
  - Rule of thumb: coefs $\simeq$ logit coefs divided by 1.6
  - Very similar to logit; sometimes easier to implement

Binary outcome models

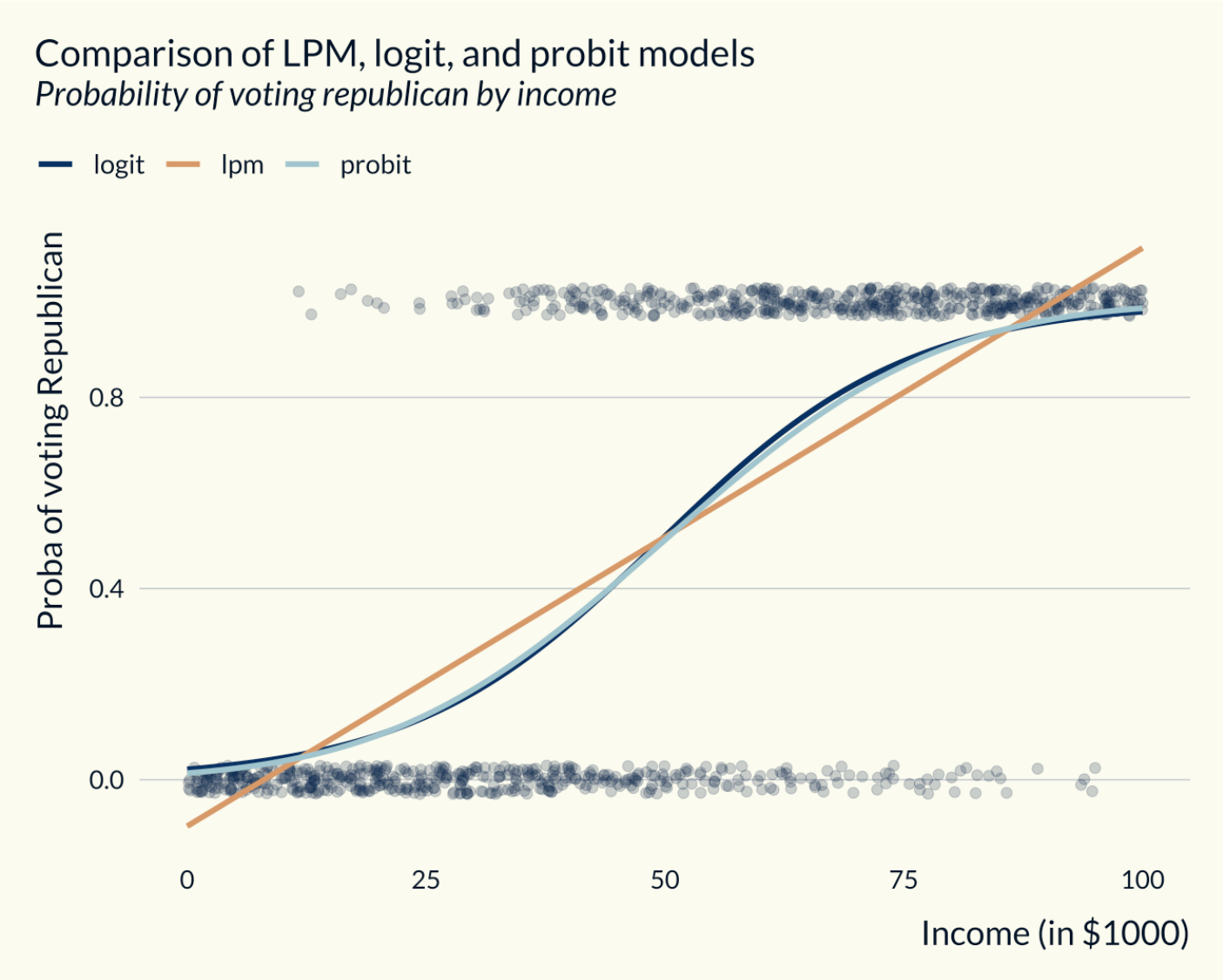
Interpreting coefficients

- β gives the **direction** BUT not the magnitude
- The **marginal effect** depends on X: effects largest in the middle of the distribution



Binary outcome models

Example fits



Count data models

- **Poisson regression model**
  - The Poisson distribution $\text{Pois}(\lambda)$ models the number of events occurring in a fixed interval if when events occur idependently and at a constant mean rate
  - Choice function: ln
  - Imposes $\mathbb{V}[y|X] = \mathbb{E}[y|X]$
- **Negative binomial model**
  - The negative binomial distribution $NB(p, r)$ models the umber of successes in a sequence of iid $Ber(p)$ trials before r failures occur
  - Allows $\mathbb{V}[y|X] \neq \mathbb{E}[y|X]$

A note on controls

- Overall two reasons to include controls:
  1. **To ensure random allocation** of the treatment
     - Necessary for identification
  2. **To improve precision**
     - To better explain y: $\nearrow$ R^2 $(= 1 - FUV)$ and $\searrow$ σ_u^2)
- Adjusting for pre-treatment covariates may
  - Reduce the variation in y, increases precision 👍
  - Reduce the variation in x, decreases precision 👎

> ⚠ **Bad controls**
>
> Do not adjust for post-treatment variables that may be affected by the treatment

Simulating bad controls

Code | Table

```
#reuse the parameters from above
n <- 10000 #to limit sampling variation
mu_b <- 3
sigma_b <- 1
gamma <- 5
kappa <- 4
sigma_a <- 2
delta <- 0

data_bad_control <- tibble(
  x = rnorm(n, mu_x, sigma_x),
  u = rnorm(n, 0, sigma_u),
  b = rnorm(n, mu_b, sigma_b),
  a = kappa*x + rnorm(n, 0, sigma_a),
  y = alpha + beta*x + gamma*b + delta*a + u
)

reg_short <-  lm(data = data_bad_control, y ~ x)
reg_pre <- lm(data = data_bad_control, y ~ x + b)
reg_post <- lm(data = data_bad_control, y ~ x + a)
reg_pre_post <- lm(data = data_bad_control, y ~ x + b + a)
```

Analysis

SEs: what are they?

- Standard errors = estimate of **sampling variability**
- Tell us how precise estimates are, how much $\widehat{\beta}$ would vary across repeated samples
- They determine confidence intervals and p-values
- An example will better illustrate

Illustration of sampling variability

Relationship between education and wage in fake data

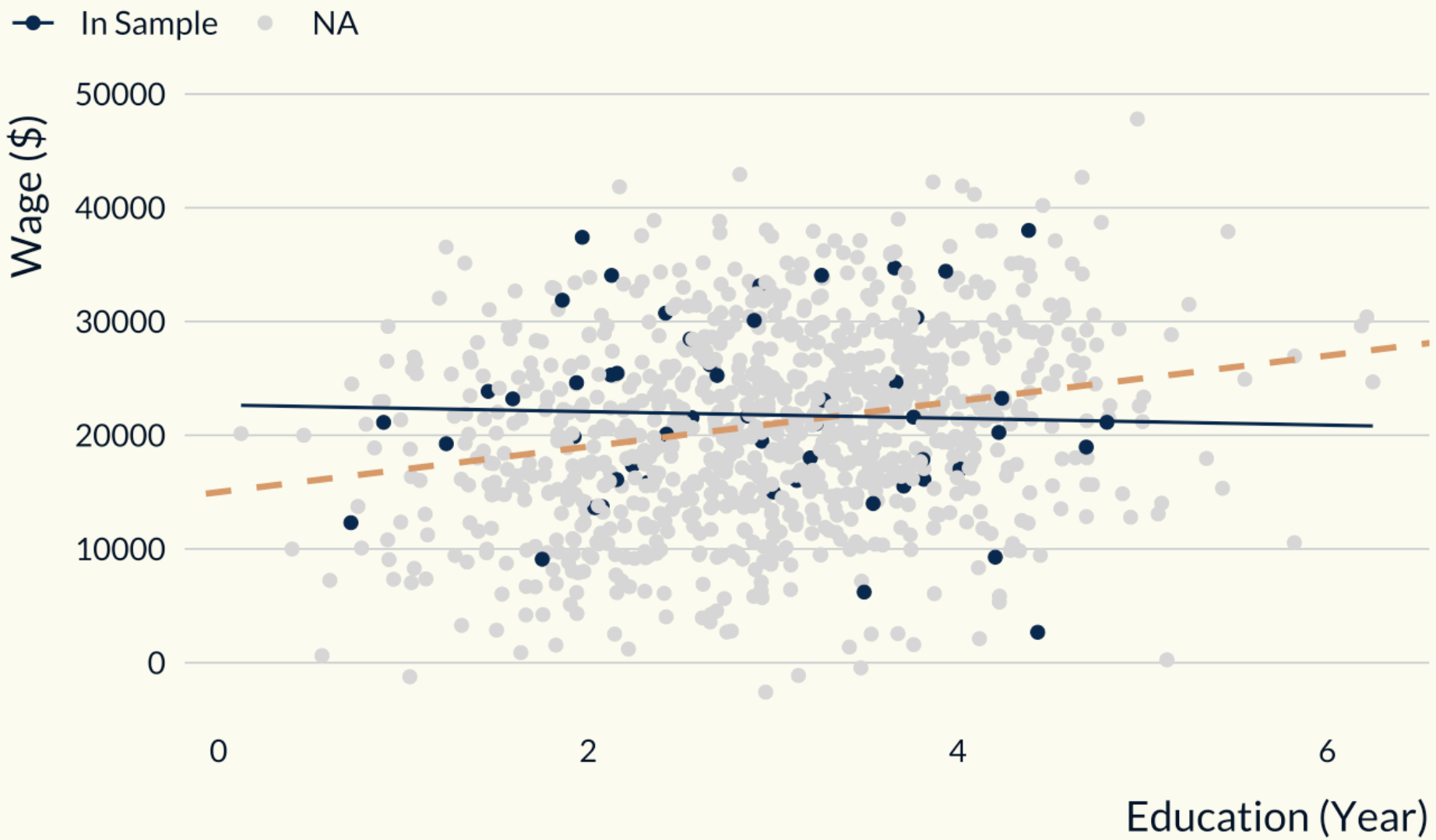


Illustration of sampling variability

Relationship between education and wage in fake data

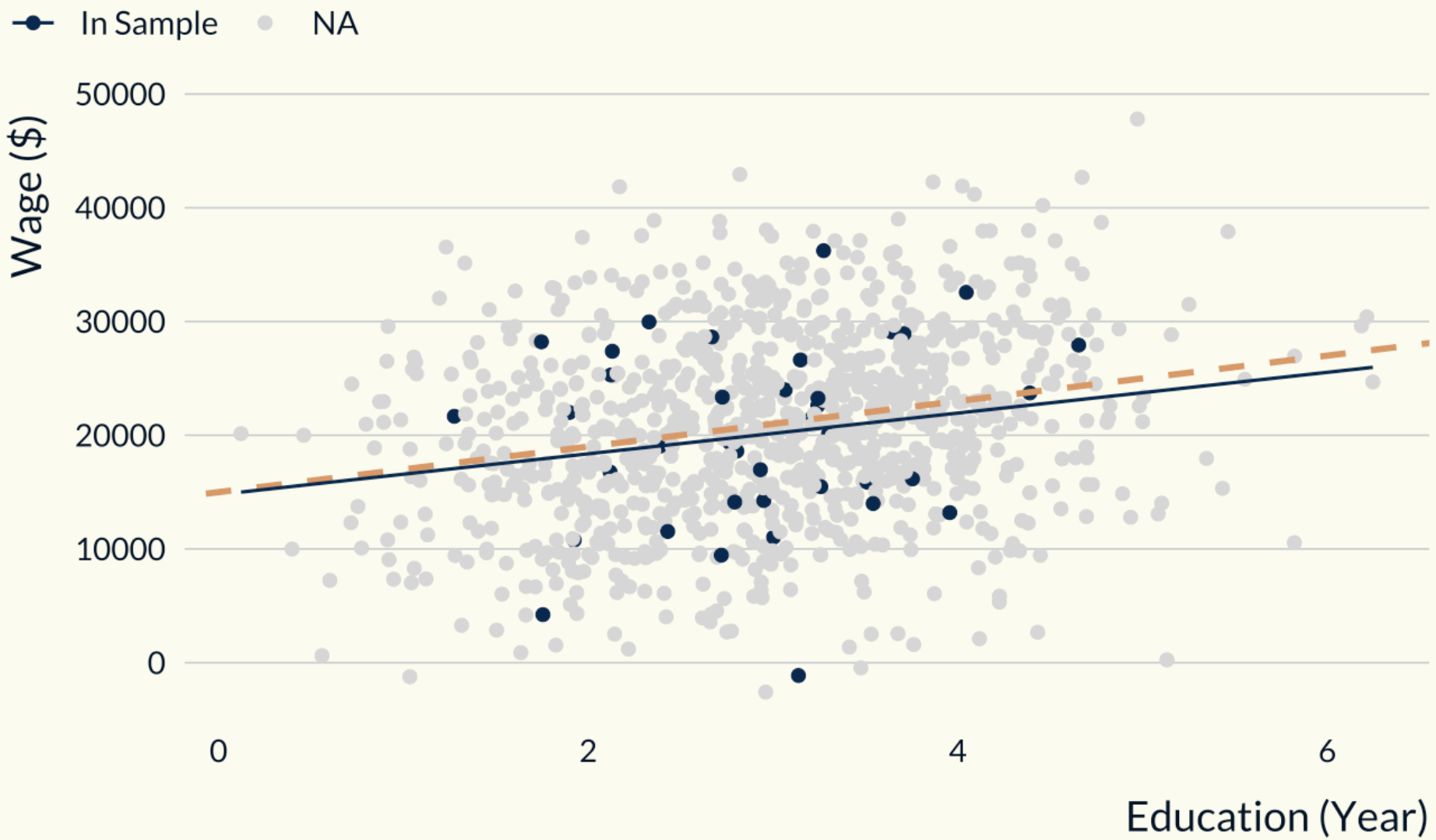


Illustration of sampling variability

Relationship between education and wage in fake data

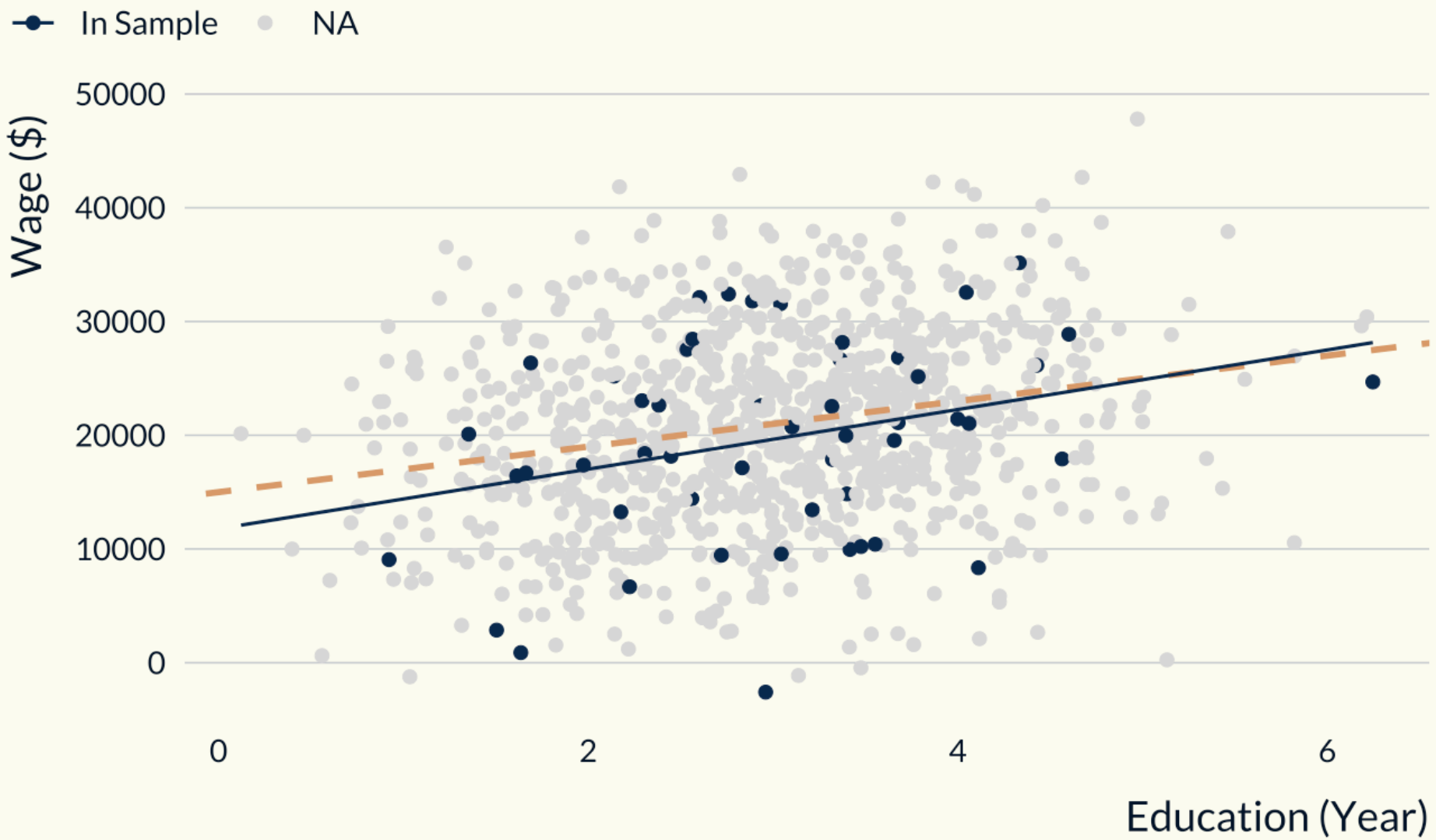
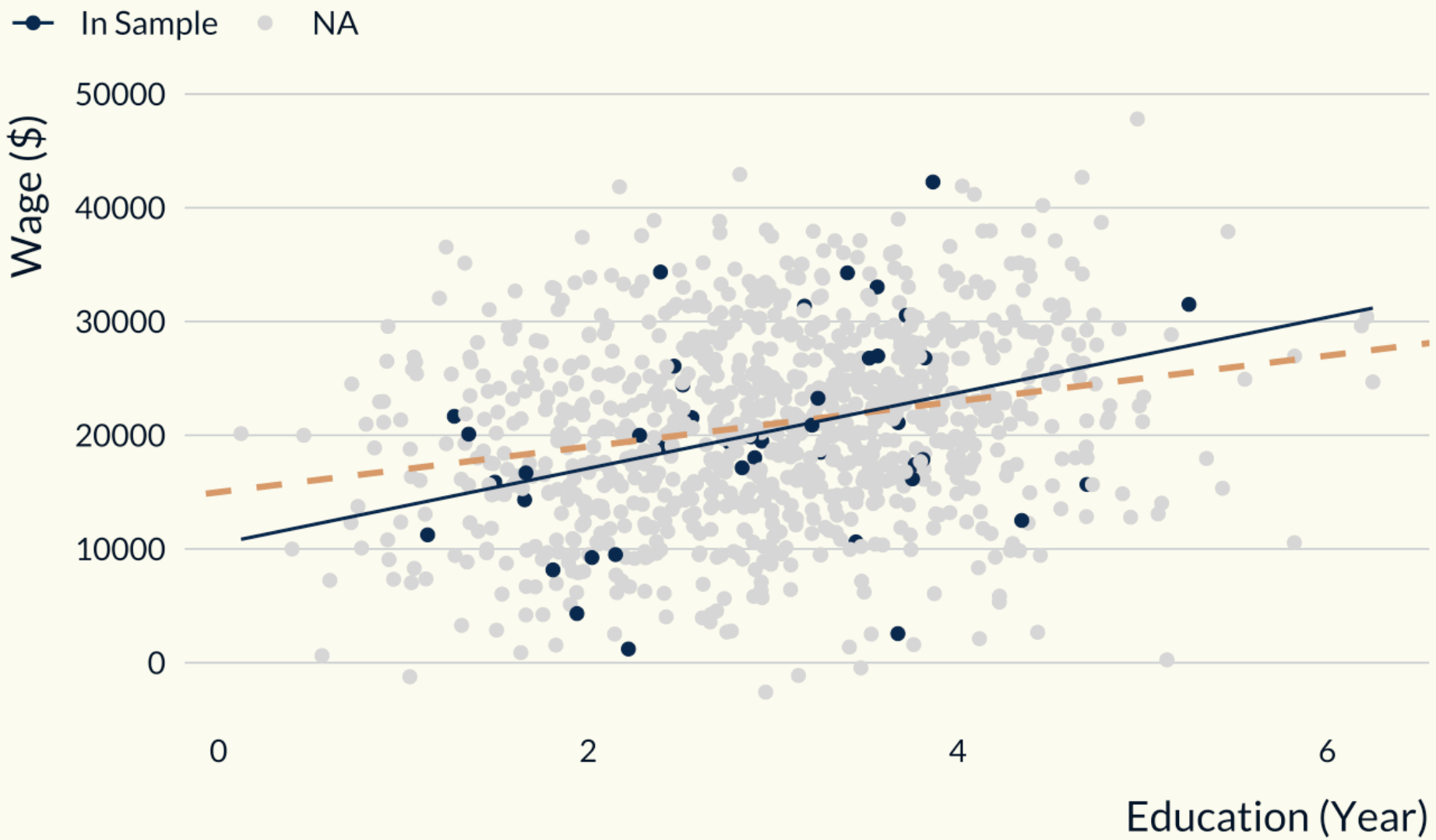
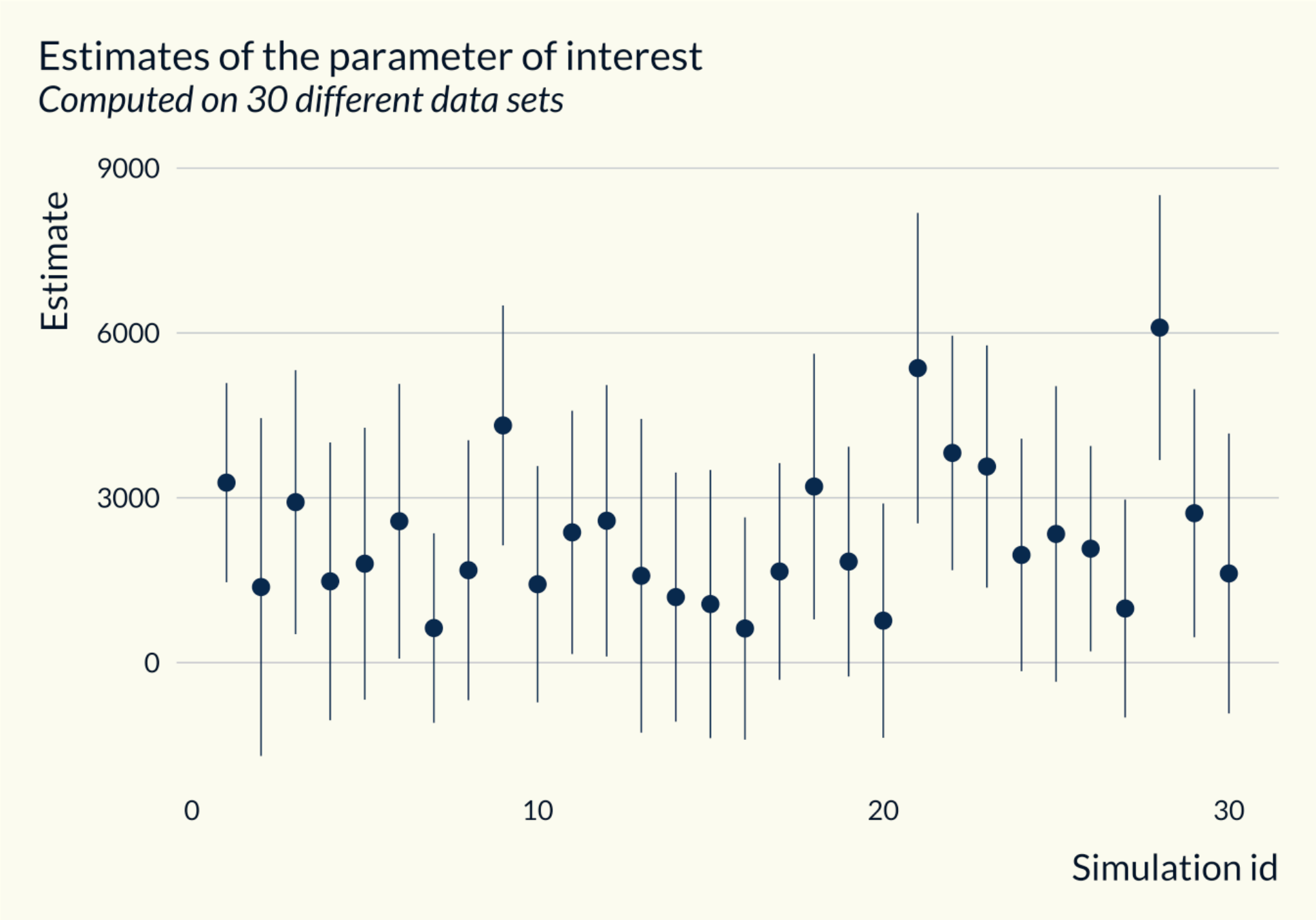
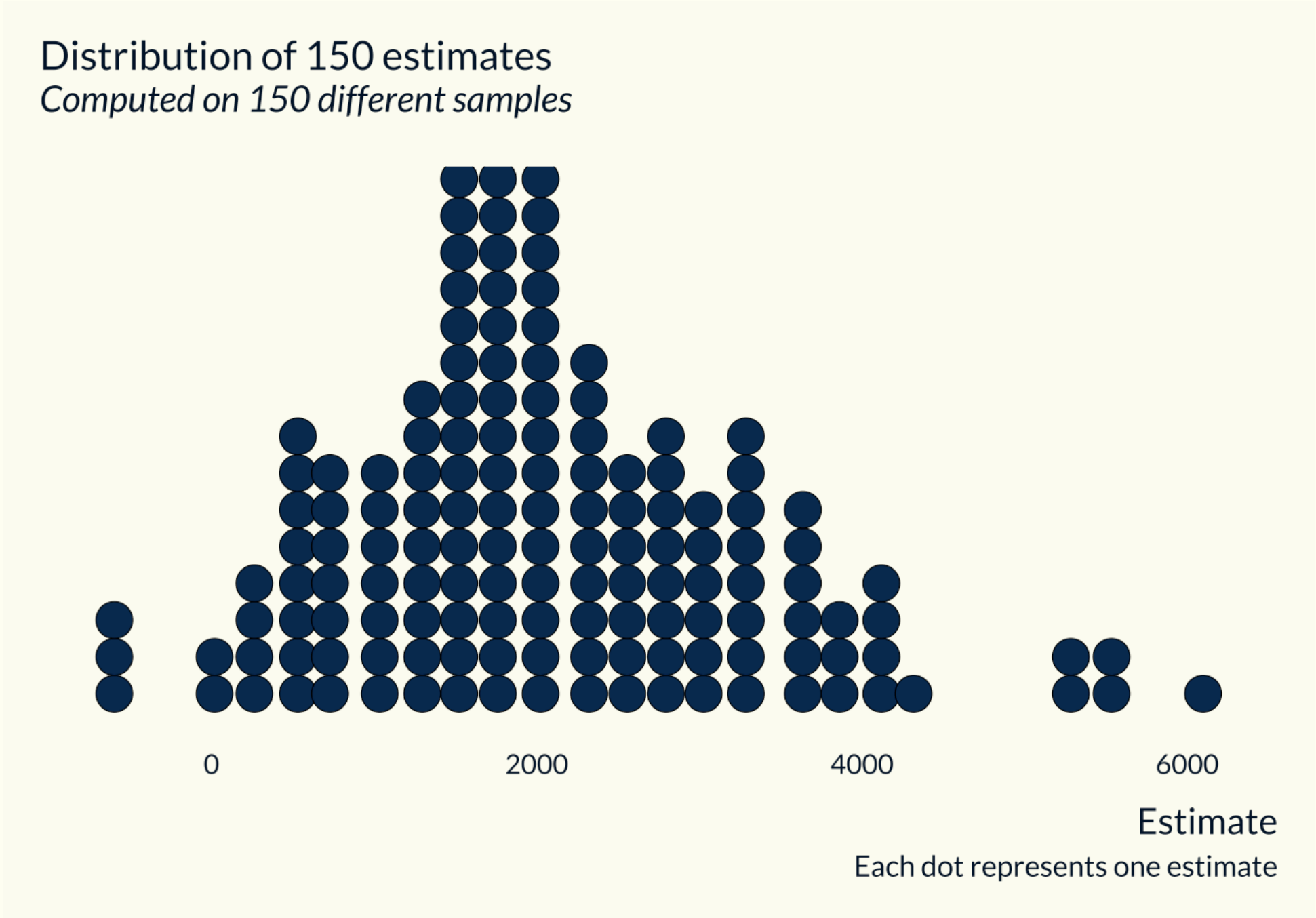


Illustration of sampling variability

Relationship between education and wage in fake data




Estimates of the parameter of interest
Computed on 30 different data sets
Estimate
9000
6000
3000
0
0
10
20
30
Simulation id


Distribution of 150 estimates
Computed on 150 different samples
0
2000
4000
6000
Estimate
Each dot represents one estimate

Why care about them?

- Violations of classical assumptions:
  - **Heteroskedasticity**: variance depends on X
  - **Non-independence**: errors correlated within groups
- Implications:
  - t-tests and CIs misleading
  - In general, SEs too small $\Rightarrow$ inference too optimistic

(Non-)spherical errors

- Under assumptions 1 $\rightarrow$ 3 we saw earlier, the asymptotic distribution of $\widehat{\beta}_{OLS}$ is

$$\widehat{\beta}_{OLS} \overset{a}{\sim} \mathcal{N}\left(\beta_0, (X'X)^{-1}X'\Sigma X(X'X)^{-1}\right)$$

- If spherical errors (*ie* $\Sigma = \sigma^2 I$), use unbiased sample variance
- If non-spherical errors, need a covariance matrix estimator that is consistent under this misspecification
- $\Rightarrow$ use **sandwich estimators** of the variance ($\underbrace{(X'X)^{-1}}_{bread}\underbrace{X'\widehat{\Sigma}X}_{meat}\underbrace{(X'X)^{-1}}_{bread}$)
- Heteroskedasticity: compute **White SEs**
- Autocorrelation: compute **Newey-West/Conley SEs** if correlated in time/space

Why clustering SEs

- When errors are correlated within groups (e.g. individuals, firms, regions), clustering adjust for this

$$\mathbb{E}\left[u_i u_j | X\right] \neq 0 \text{ for } i, j \text{ in the same cluster}$$

- **Examples**
  - Panel data: repeated measures of same unit
  - Group-level treatment, *eg* policy at regional level with many individuals per region
- **Intuition**
  - We do not have N independent observations but G clusters
  - *eg* 10 classrooms of 30 students does not correspond to 300 independent observations
  - The real sample size is closer to the number of clusters, not observations

What clustering does

- Do not affect point estimates
- Allows for intra-cluster correlation and adjusts the effective number of independent observations
- Increases SEs to reflect within-group dependence $\Rightarrow$ wider CIs

$$\widehat{\mathrm{Var}_{CR}}\left(\widehat{\beta}\right) = (X'X)^{-1}\left(\frac{1}{n-p}\sum_{g \in G} X_g{}' r_g r_g{}' X_g\right)(X'X)^{-1}$$

where X_g and r_g are data and residuals for cluster g

- Intuition: each cluster provides one independent piece of information about how residuals co-evolve with regressors

Which level to cluster at?

- Clustering accounts for correlation in residuals
- $\Rightarrow$ the level of clustering depends on **where correlation comes from**
- Rule of thumb:
  - Cluster at the level of the shock or treatment variation
  - If unsure, cluster higher rather than lower
- If level of clustering too low, SEs too small, overconfidence in results
- If level of clustering too large, SEs too large, under reject

When is clustering difficult to implement?

- When **few clusters** (*eg* < 30) $\Rightarrow$ asymptotic results unreliable
- When complex dependence or unclear correlation structure
- When model residuals violate independence in unknown ways
- Solution
  - **Resampling** methods (*eg* bootstrap)
  - Hierarchical clustering

Bootstrap

- **Intuition**: simulate the sampling variability by resampling from our data
- **Steps**:
    1. Draw samples (**with replacement**)
    2. Estimate $\widehat{\beta}$ on each sample
    3. Compute the SD of $\widehat{\beta}$ across replications
    4. That SD = bootstrap SE
- When resampling, respect the correlation structure: **sample clusters**

Summaries

Summary of today

- Regression models and OLS estimation rest on a set of **assumptions**
- Ensure estimates are unbiased, efficient, and statistically valid
- When fail, estimates are biased and standard errors misleading
- We reviewed these modelling assumptions and discuss **what happens when they fail**
- Limited outcome models or clustering help **overcome** the issues

Goal of the whole course

Give *us* a deeper understanding of:

- How regression works "**under the hood**": **intuition**
- Causal identification strategies and their assumptions
- How **design**, **modeling**, and **analysis** choices shape empirical results
- Common pitfalls and challenges in empirical work
- How to use **simulations** to explore estimator behavior and diagnose potential problems specific to your own cases
- Existing **references** and where to find additional information on a specific topic

To mention when pitching your analysis

1. **Research question**
    - What causal effect of interest are you trying to estimate?
2. **Ideal experiment**
    - What ideal experiment would capture the causal effect?
3. **Identification strategy**
    - How are the observational data used to make comparisons that approximate such an experiment?
4. **Estimation method** (including assumptions made when constructing standard errors)
5. **Falsification tests** that support the identifying assumptions

To also mention in your pitch

- Motivation
  - Why is your research question important?
- Contributions to the literature
- Methodological contributions
- Internal validity
  - Are the identifying assumptions plausible?
  - Are there unexplained results?
- External validity
  - Gap between policy questions and the analyses performed?
  - Generalization to other populations and settings?

Summary of the entire course

Take away messages

- Your research question and underlying theory are crucial
- Think about what is the identifying variation in your model:
  - What are you estimating exactly with your model?
  - Which observations contribute to identification?
- Always wonder what you are comparing
- Use simulations to explore and understand points

References