

Problem Set 1

Q1: Notions of dependence

In econometrics, we use at least three concepts to describe the relationship between two random variables: independent, mean-independent, and uncorrelatedness. These concepts appear frequently in econometric assumptions, so it is important to understand how they are related and how they differ.

Start with the definitions. For random variables A and B with the relevant finite moments:

- Independent: A and B are independent if their joint distribution factorizes.
- Mean-independence of A from B : $E[A | B] = E[A]$.
- Uncorrelatedness: $Cov(A, B) = 0$ (If $Var(A), Var(B) > 0$, this $\equiv Corr(A, B) = 0$)

It is straightforward to show that the following relationships hold:

$$\text{independence} \Rightarrow \text{mean independence} \Rightarrow \text{uncorrelatedness}.$$

The reverse implications do not generally hold. To show that, it is enough to find counterexamples. Let's start with a counterexample showing that mean independence does not imply independence. Take independent $B \sim U[0, 2]$ and $C \sim U[-1, 1]$, and let $A = (1 + B)C$. Then

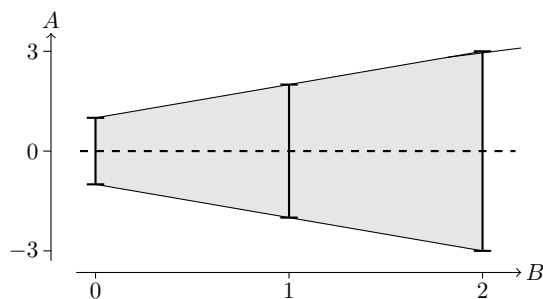
$$E[A | B = b] = (1 + b)E[C] = 0 = E[A],$$

so A is mean-independent of B . But

$$Var(A | B = b) = (1 + b)^2 Var(C) = \frac{(1 + b)^2}{3}$$

varies with b . Independence would require the entire conditional distribution of A given $B = b$ to be the same for all b , not merely its mean. Here the spread changes with B , so A and B are not independent. We can see that in the figure below.

Counterexample 1: $B \sim U[0, 2]$ and $C \sim U[-1, 1]$ are independent. Let $A = (1 + B)C$.



Provide a counterexample showing that uncorrelatedness does not imply mean independence.

Q2: Causal interpretation of regression coefficients

When analyzing randomized controlled trials, we often add covariates X to our regression model to improve precision. For example, if we were analyzing a class size RCT, we might include a control for family background (e.g., parental income). Intuitively, when we do this, we know we are identifying the causal effects of the treatment (in this class size) but not the causal effect of the X 's (in this case parental income). However, you should be able to show this formally, and doing so is a nice application of causal models and linear projections.

Let's start by considering the following common linear causal model

$$Y_i(d, x) = Y_i(0, 0) + \beta_1 d + \beta_2 x.$$

Here β_1 is the causal effect of an additional unit of D holding x fixed, and β_2 is the causal effect of an additional unit of X holding d fixed. Define

$$\beta_0 = E[Y_i(0, 0)], \quad \epsilon_i = Y_i(0, 0) - \beta_0.$$

The observed outcome is $Y_i = Y_i(D_i, X_i)$, so an augmented version of equation (1) is:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \epsilon_i.$$

where β_1 and β_2 are causal parameters.

Next, let's formalize the idea that D but not X is randomly assigned. Specifically, assume:

A1 D_i is randomly assigned and therefore independent of both X_i and ϵ_i .

A2 $Cov(X_i, \epsilon_i) \neq 0$.

A3 Relevant variances are positive and finite.

What is identified by the population regression of Y_i on D_i and X_i ? Which, if either, of the coefficients on D_i and X_i has a causal interpretation? Explain.

Hint: Consider the population linear projection

$$\epsilon_i = p_0 + p_1 D_i + p_2 X_i + v_i.$$

The general projection formulas are

$$p_1 = \frac{Cov(D_i, \epsilon_i)Var(X_i) - Cov(X_i, \epsilon_i)Cov(D_i, X_i)}{Var(D_i)Var(X_i) - Cov(D_i, X_i)^2},$$

$$p_2 = \frac{Cov(X_i, \epsilon_i)Var(D_i) - Cov(D_i, \epsilon_i)Cov(D_i, X_i)}{Var(D_i)Var(X_i) - Cov(D_i, X_i)^2}.$$
¹

Use A1-A3 to simplify the expressions for (p_1, p_2) , then substitute the projection into the causal model.

¹The projection coefficients can be derived from the defining properties of a linear projection: $E(v_i) = 0$, $Cov(D_i, v_i) = 0$, and $Cov(X_i, v_i) = 0$. You are not asked to derive these formulas here.

Q3: Monte Carlo: conventional and robust inference

For this question you will perform a Monte Carlo exercise to examine conventional and heteroskedasticity-robust standard errors. Specifically, you will draw many samples from a DGP with heteroskedasticity and, in each sample, construct confidence intervals and test a true null. You will then measure the proportion of intervals containing the true value (coverage) and the proportion of tests rejecting the true null (the empirical size). Correctly calibrated inference would produce approximately 95% coverage and 5% size. Finally, you will compare conventional and robust inference across sample sizes.

Work with the following DGP:

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \quad \beta_0 = 1, \quad \beta_1 = 0,$$

where

$$D_i \sim U[-\sqrt{3}, \sqrt{3}], \quad \eta_i \sim U[-\sqrt{3}, \sqrt{3}], \quad D_i \perp \eta_i, \quad u_i = D_i \eta_i.$$

Use $n = 20, 50, 200, 1000$, with 5,000 replications at each sample size. In each replication:

- Draw a sample of size n from the DGP.
- Regress Y_i on an intercept and D_i .
- Construct a 95% CI for β_1 using the conventional standard error.
- Construct a 95% CI using HC1 heteroskedasticity-robust standard errors. In Stata, use “vce(robust)”. In R, use “sandwich::vcovHC(..., type=“HC1”)”.¹
- Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at the 5% level using each of those standard errors.
- Use $z_{0.975} \simeq 1.96$ for both procedures: the CI is $\hat{\beta}_1 \pm z_{0.975} se(\hat{\beta}_1)$, and reject if $|\hat{\beta}_1 / se(\hat{\beta}_1)| > z_{0.975}$.

- (a) Show that $E[u_i | D_i] = 0$ but $Var(u_i | D_i) = D_i^2$. What do these results imply for estimation of β_1 and for the choice between conventional and robust standard errors?

¹Other heteroskedasticity-robust variance estimators are available. HC0 is the original Huber–White estimator; HC1 applies a degrees-of-freedom correction; HC2 and HC3 additionally adjust for leverage. These estimators are asymptotically equivalent but can differ when n is small. In this question, use HC1 throughout so that everyone implements the same procedure.

- (b) Carry out the Monte Carlo exercise and complete the table below, reporting all rates as percentages. Since each test and CI use the same SE and critical value, rejection occurs exactly when the CI excludes the true value. Thus rejection rate equals one minus coverage in this exercise. Use this as a check on your code.

n	Conventional coverage	Robust coverage	Conventional rejection	Robust rejection
20				
50				
200				
1000				

- (c) Compare conventional and robust inference as n increases. Which procedure, if either, appears to deliver valid large-sample inference under this DGP? Explain using the table.
- (d) Use the robust-SE results to discuss the idea of the large-sample approximation. How well does the large-sample approximation appear to work at $n = 20, 50, 200, 1000$?

Q4: Clustering in practice

Section 3.6 gave the special-case approximation

$$\text{Var}(\hat{\beta}_1) \simeq \frac{\sigma_u^2}{n\sigma_D^2} [1 + (m-1)\rho_D\rho_u],$$

where m is group size and ρ_D and ρ_u are within-group treatment and error correlations respectively. This special case suggests that we should cluster when both correlations are present, although this is not a general theorem. Use this intuition to explain each inference choice below. You do not need to read the papers to answer this question, but can if you want to!

- (a) **Baby’s First Years.** Noble et al. (2025) study the “Baby’s First Years” experiment. Mothers were individually randomized within four study sites to receive either \$333 or \$20 per month. The authors analyzed child outcomes at age four using individual-level regressions and do not cluster their standard errors. Why is the decision not to cluster natural given the intuition from Section 3.6?
- (b) **Oregon Health Insurance Experiment.** Finkelstein et al. (2012) study a lottery allocating the opportunity to apply for Medicaid. Names were drawn individually, but selection also made other household members on the lottery list selected and eligible to apply. Outcomes are individual-level; SEs are clustered by household. Why is household clustering natural in light of Section 3.6?
- (c) **Gifted Education.** Card and Giuliano (2016) study a U.S. district in which elementary schools place fourth-grade students into gifted/high-achiever (GHA) and regular classrooms. Among non-gifted students, priority for open GHA seats depends on prior test-score rank within the school/cohort. Crossing a school/cohort-specific cutoff sharply increases the probability of GHA placement, which the authors exploit in a regression-discontinuity design. Observations are individual students, so gifted participation varies within schools. The data contain multiple cohorts from the same schools. The authors cluster SEs by school. Why might school clustering be sensible even though gifted participation varies within a school? Does this contradict the special-case intuition? Explain.

Hint: Our special case in which individual randomization can make clustering unnecessary relies on treatment being independently randomized across individuals. Is gifted treatment generated by independent individual random assignment within schools?

References

- Card, D. and Giuliano, L. (2016). Can tracking raise the test scores of high-ability minority students? *American Economic Review*, 106(10):2783–2816.
- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Group, O. H. S. (2012). The Oregon health insurance experiment: evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106.
- Noble, K., Magnuson, K., Duncan, G., Gennetian, L. A., Yoshikawa, H., Fox, N. A., Halpern-Meehin, S., Troller-Renfree, S., Han, S., Egan-Dailey, S., et al. (2025). The effect of a monthly unconditional cash transfer on children’s development at four years of age: a randomized controlled trial in the us. Technical report, National Bureau of Economic Research.