Problem Set 1: Answer Key (answers in **bold)**

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: $E[A \mid B] = E[A]$.
- 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

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

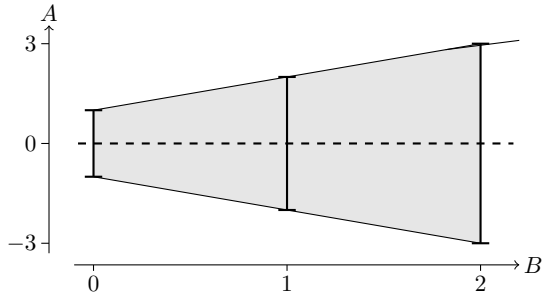
so A is mean-independent of B. But

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

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.

Take $B \sim U[-1,1]$ and $A = B^2$. Symmetry gives $E[B] = 0$ and $E[B^3] = 0$, so

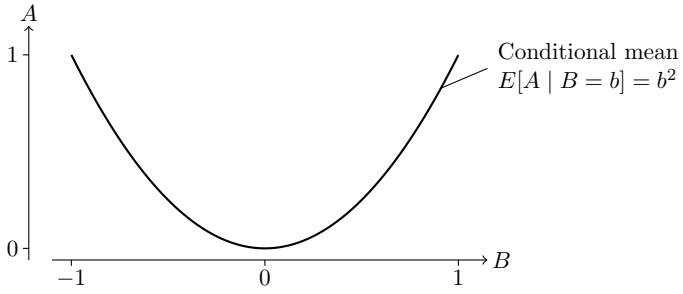
$$\begin{aligned} Cov(A,B) &= E[AB] - E[A]E[B] \\ &= E[B^3] - E[B^2]E[B] = 0. \end{aligned}$$

Hence A and B are uncorrelated. But

$$E[A \mid B = b] = b^2, \qquad E[A] = E[B^2] = \frac{1}{3}.$$

Since b^2 is not equal to $1/3$ for all b, the conditional mean is not constant at $E[A]$. Thus A is not mean-independent of B.

Counterexample 2: $B \sim U[-1,1]$ and $A = B^2$.



Mean independence is stronger than uncorrelatedness: uncorrelatedness rules out only linear association.

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

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

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.

Random assignment A1 implies

$$Cov(D_i, \epsilon_i) = 0, \qquad Cov(D_i, X_i) = 0.$$

Substituting into the supplied formulas,

$$p_1 = \frac{0 \cdot Var(X_i) - Cov(X_i, \epsilon_i) \cdot 0}{Var(D_i)Var(X_i) - 0^2} = 0,$$
$$p_2 = \frac{Cov(X_i, \epsilon_i)Var(D_i) - 0 \cdot 0}{Var(D_i)Var(X_i) - 0^2} = \frac{Cov(X_i, \epsilon_i)}{Var(X_i)} \neq 0.$$

By A3 the denominator is positive, and by A2 $Cov(X_i, \epsilon_i) \neq 0$**. Substituting the projection into the causal model gives**

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_i + \beta_2 X_i + p_0 + p_1 D_i + p_2 X_i + v_i \\ &= \beta_0' + \beta_1 D_i + \left[\beta_2 + \frac{Cov(X_i, \epsilon_i)}{Var(X_i)}\right] X_i + v_i, \end{aligned}$$

where $\beta_0' = \beta_0 + p_0$**. By construction,** $E(v_i) = 0$**,** $Cov(D_i, v_i) = 0$**, and** $Cov(X_i, v_i) = 0$**. Thus this is the population linear projection of** Y_i **on an intercept,** D_i**, and** X_i**. The population regression therefore identifies** β_1 **and** $\beta_2 + \frac{Cov(X_i, \epsilon_i)}{Var(X_i)}$**. The second is the causal effect plus OVB, and hence only the coefficient on** D_i **has a causal interpretation.**

[1] The 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:

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

where

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

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

[1] Other 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.

Independence of D_i and η_i gives

$$E[u_i \mid D_i] = E[D_i\eta_i \mid D_i] = D_iE[\eta_i \mid D_i]$$
$$= D_iE[\eta_i] = 0.$$

Also, $\eta_i \sim U[-\sqrt{3}, \sqrt{3}]$ has variance $(2\sqrt{3})^2/12 = 1$, so

$$Var(u_i \mid D_i) = Var(D_i\eta_i \mid D_i) = D_i^2Var(\eta_i \mid D_i)$$
$$= D_i^2Var(\eta_i) = D_i^2.$$

The zero conditional mean implies $E[u_i] = 0$ and $E[D_iu_i] = 0$, so the OLS identifies β_1. The conditional variance depends on D_i, so the errors are heteroskedastic and conventional inference is not generally valid.

(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				

See `stata/ps1_mc_heterosk.do` for the Stata code. The table below imports its automatically generated results; all rates are percentages.

n	**Conventional coverage**	**Robust coverage**	**Conventional rejection**	**Robust rejection**
20	**82.1**	**89.3**	**17.9**	**10.7**
50	**83.8**	**92.5**	**16.2**	**7.5**
200	**85.6**	**94.7**	**14.4**	**5.3**
1000	**85.3**	**94.7**	**14.7**	**5.3**

(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.

Robust inference appears to deliver valid inference as n becomes large; conventional inference does not. With conventional standard errors, coverage remains well below 95% and the rejection rate remains well above 5% even at the largest sample sizes. With robust standard errors, coverage moves toward 95% and the rejection rate moves toward 5% as n increases. This is what the theory predicts. The conventional variance formula is invalid under the heteroskedasticity in this DGP, so increasing n does not fix it. The robust variance estimator is consistent under heteroskedasticity, so the associated confidence intervals and tests become correctly calibrated as $n \to \infty$. The results need not improve monotonically with n because the reported rates themselves have Monte Carlo sampling variation.

(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 $n = 20, 50, 200, 1000$?

The asymptotic result tells us that robust inference becomes correctly calibrated as $n \to \infty$, but in a finite sample we use that limiting result as an approximation. At $n = 20$, the approximation is not very good: robust coverage is only 89.3% and the rejection rate is 10.7%, rather than 95% and 5%. At $n = 50$, performance improves but there is still noticeable undercoverage and overrejection. By $n = 200$, robust inference is close to its large-sample benchmark, and the $n = 1000$ results are also close. Thus the Monte Carlo shows directly that the quality of the large-sample approximation can depend on sample size.

Q4: Clustering in practice

Section 3.6 gave the special-case approximation

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

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?

Treatment is randomized to individual mothers within each site, rather than assigned to entire sites. In the special case model, individual-level random assignment implied no positive within-group treatment correlation, so group-level outcome shocks alone did not generate the clustering variance inflation. The Baby's First Years design therefore gives no reason to cluster at the site level merely because mothers were recruited in four sites.

(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?

The shared lottery opportunity makes treatment correlated within households. Shared income, health environment, access to care, and shocks make error correlation plausible too. Both ingredients are present: $\rho_D > 0$ and $\rho_u > 0$. Household clustering allows for this dependence.

(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?

GHA participation varies within schools, but treatment is not independently assigned across students. Eligibility depends on relative rank and a school/cohort-specific cutoff, so there is no reason to impose the $\rho_D = 0$ simplification from the individual-randomization case. At the same time, errors may be correlated within schools because students share school-level environments and shocks, including across cohorts. Clustering by school allows for this within-school dependence. This does not contradict the special-case intuition: that result was derived from a special case, and the no-clustering conclusion relied on independently assigned individual treatment.

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-Meekin, 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.