

Exam 5106/9106B – Fall 2025

IMPORTANT: Always motivate your answers. Show knowledge and understanding of the concepts taught in the course. Your answers should be as short as possible and as long as necessary. Total marks of the examination is 150. Each sub-question carries 10 marks. Please submit your responses to these questions alongside your term project. For guidelines on the term project, see the instructions posted on Canvas.

Each question in this exam connects to Joseph Altonji and Zhengren Zhu, "Returns to Specific Graduate Degrees: Estimates Using Texas Administrative Records," NBER Working Paper 33530 (2025), <https://doi.org/10.3386/w33530>.

Using various methods, Altonji and Zhu (2025) claim to estimate causal effects of different graduate degrees on log earnings. Throughout this exam, we will consider a simplified version of their data and models. Suppose we have data from two periods: before graduate school ($t=0$) and after graduate school ($t=1$). We will also consider a simplified version of their estimating equations. Each estimator we consider will be of the form:

$$Y_{it} = \sum_{g>0} \gamma_g G_{git} + \lambda_t + u_{it}. \quad (1)$$

where

- $Y_{it} \equiv \log(y_{it})$ denotes the log earnings of individual i in year t (level earnings having value y_{it})
- $g(i)$ denotes the graduate degree i (eventually) receives.
- Let $g(i) = 0$ if an individual never receives a graduate degree.
- The dummy variable G_{git} is turned on if individual i has obtained graduate degree g in year t (note all such dummies are turned off in $t = 0$)
- λ_t is a calendar year fixed effect.

Importantly, note that Y_{it} is undefined for individuals who have zero earnings (meaning any (i, t) with $y_{it} = 0$ is effectively dropped from our data).

Moreover, let's abstract from undergraduate degrees and consider potential outcomes. Let $Y_{it}(g)$ denote potential log earnings if one were to obtain the graduate degree g . Let $Y_{it}(0)$ denote the potential outcome with no graduate degree.

1. The authors focus on three estimators for returns to degrees:

- **OLS:** views u_{it} as an idiosyncratic residual term.
- **FE_{cg}** (which simplifies to 'FEG' given the above assumptions): substitutes

$$u_{it} = b_{g(i)} + \eta_{it}$$

where $b_{g(i)}$ is a fixed effect for the degree $g(i)$ received by person i in $t = 1$ and η_{it} an idiosyncratic residual term.

- **FE substitutes**

$$u_{it} = e_i + \varepsilon_{it}$$

where e_i is an individual fixed effect and ε_{it} an idiosyncratic residual.

- (a) The paper is not precise about the estimands (parameters of interest) each of the three estimators for γ_g is targeting. Suppose for each estimator, the estimand is merely the given estimator's population value. Without imposing any restrictions on heterogeneity in potential outcomes, how do these estimands compare? Express them in terms of expectations of potential outcomes.

OLS: the limiting value of γ_g is $E[Y_{i1}(g)|g(i) = g, y_{i1} > 0] - E[Y_{i1}(0)|g(i) = 0, y_{i1} > 0]$

FEcg: the limiting value of γ_g is : $(E[Y_{i1}(g)|g(i) = g, y_{i1} > 0] - E[Y_{i0}(0)|g(i) = g, y_{i1} > 0]) - (E[Y_{i1}(0)|g(i) = 0, y_{i1} > 0] - E[Y_{i0}(0)|g(i) = 0, y_{i0} > 0])$

FE: the limiting value of γ_g is: $E[Y_{i1}(g) - Y_{i0}(0)|g(i) = g, y_{i0} > 0, y_{i1} > 0] - E[Y_{i1}(0) - Y_{i0}(0)|g(i) = 0, y_{i0} > 0, y_{i1} > 0]$

The OLS estimand is simply a cross-sectional comparison of log earnings in $t = 1$ between those with degree g versus no degree, among those with positive earnings.

The FEcg estimand is a difference-in-difference estimator, comparing trends in average log earnings among those with graduate degree g to those with no graduate degree.

The FE estimand is also a difference-in-difference estimator, but it differs from FEcg in that it compares within-person trends among individuals who have positive earnings in both periods.

- (b) Suppose we used the FE model to estimate an ATT for two different graduate programs g_1 and g_2 . Call the corresponding coefficient estimates $\hat{\gamma}_{g_1}$ and $\hat{\gamma}_{g_2}$ respectively. Your classmate wants to interpret $\hat{\gamma}_{g_1} - \hat{\gamma}_{g_2}$ as a consistent estimator of the ATT associated with receiving graduate degree g_1 instead of graduate degree g_2 among those actually receiving graduate degree g_1 . That is, our parameter of interest is:

$$E[Y_{it}(g_1) - Y_{it}(g_2)|g(i) = g_1, y_{i1} > 0, y_{i0} > 0]$$

State the common trends assumption that would permit this interpretation. Need it be the case that common trends between those in program g_1 and no graduate degree (or g_2 and no graduate degree) hold?

Note that with our simplifying assumptions, part (a) shows $\hat{\gamma}_{g_1} = \hat{E}[Y_{i1} - Y_{i0}|g(i) = g_1, y_{i1} > 0, y_{i0} > 0] - \hat{E}[Y_{i1} - Y_{i0}|g(i) = 0, y_{i1} > 0, y_{i0} > 0]$ and $\hat{\gamma}_{g_2} = \hat{E}[Y_{i1} - Y_{i0}|g(i) = g_2, y_{i1} > 0, y_{i0} > 0] - \hat{E}[Y_{i1} - Y_{i0}|g(i) = 0, y_{i1} > 0, y_{i0} > 0]$. The second expectation is the same in both of these terms, so they cancel out when differencing and we are left with a conventional diff-in-diff:

$$\hat{\gamma}_{g_1} - \hat{\gamma}_{g_2} = \hat{E}[Y_{it} - Y_{i0}|g(i) = g_1, y_{i1} > 0, y_{i0} > 0] - \hat{E}[Y_{it} - Y_{i0}|g(i) = g_2, y_{i1} > 0, y_{i0} > 0]$$

so we need only that individuals with degree g_1 would have had the same trends in outcomes as those with degree g_2 , had they instead taken degree g_2 .

- (c) Suppose OLS, FE, and FEcg each return very different estimates. Is that evidence that one (or more) of these models is misspecified? Discuss one other possible explanation for this.

Because these identify average treatment effects over different subpopulations, treatment effect heterogeneity is one possible explanation for these differences.

2. Now let us consider the FE model. Let's assume that the residual term ε_{it} is normally distributed with mean zero and unknown variance σ_ε^2 . For this question, you can assume y_{it} is zero at random. For brevity, let λ denote the vector of calendar year fixed effects, γ the vector of graduate program effects, and e the vector of individual fixed effects.

- (a) Write down the minimization problem that the FE regression model solves, and its first order conditions. Next, using the assumptions above, write the log likelihood function for the data, $\log(L(\lambda, \gamma, e|Y, G))$ and the first order conditions that MLE uses to identify the model parameters. Compare the first order conditions from these estimators and what they imply about the resulting parameter estimates.

OLS solves

$$\min_{\lambda, \gamma, e} \sum_{i,t} (Y_{it} - \lambda_t - \sum_g \gamma_g G_{git} - e_i)^2$$

while MLE is equivalent to solving

$$\min_{\lambda, \gamma, e} \sum_{i,t} \frac{1}{2} \log(\sigma_\varepsilon^2) + \frac{(Y_{it} - \lambda_t - \sum_g \gamma_g G_{git} - e_i)^2}{2\sigma_\varepsilon^2}$$

and therefore have the same first order conditions, up to an additional condition separately pinning down the variance parameter for MLE. As the other first order conditions are the otherwise the same, FEcg and MLE will return the same point estimates.

- (b) Now suppose we further assume that the individual fixed effects (e_i) are themselves drawn from a normal distribution with an unknown mean μ and un-

known variance σ_e^2 , independently from ε_{it} . Using that assumption to reduce the model's parameter space, write down the log likelihood function for the data, $\log(L(\lambda, \gamma, \mu, \sigma_e^2, \sigma_\varepsilon^2|Y, G))$ and its first order conditions. How do those first order conditions compare to those from FE? Are limiting estimates of γ different from the models in part (a)?

With ε_{it} and e_i both being normally distributed, Y_{it} is the sum of two independent normals. However, e_i is common across all periods for person i . The most straightforward way to proceed is to recognize this implies Y_i —the vector including Y_{it} for all t —follows a *multivariate* normal distribution. We know the mean and variance-covariance matrix for this distribution. The mean is $\lambda + \gamma_{g(i)} + \mu$. The variance-covariance matrix has off-diagonals of: $\text{cov}(Y_{it}, Y_{is}) = \sigma_e^2$ for all $s \neq t$, and a diagonal of: $\text{Var}(Y_{it}) = \sigma_e^2 + \sigma_\varepsilon^2$. We therefore know the exact multivariate density of Y_i , $f(Y_i|G_{g(i)})$. From here getting the total density of the data is straightforward:

$$f(Y|G) = \prod_{i=1}^N f(Y_i|G_{g(i)})$$

and the log likelihood function we derive by simply taking $\log(f(Y|G))$. By inspection, the resulting likelihood function will clearly have different first order conditions from those in part (a), and so in general different limiting estimates of α and γ .

- (c) Suppose you were interested in studying the distribution of individual fixed effects. Would you prefer to plot the distribution of effects estimated by FE, or the normal distribution with mean μ and variance σ_e^2 , from part b? Intuitively describe the trade-off underlying this choice.

The individual fixed effects are very noisy, as they are based on only a single observed outcome for each individual (the value in $t=0$). On the other hand, the random effects model of part (b) relies on distributional assumptions. One would prefer the random effects model when the individual fixed effects are in truth normally distributed. Otherwise, we face a bias-variance trade-off between imposing structure on the distribution of fixed effects (what the random effects model does) versus estimating that distribution non-parametrically (what the random effects model does).

3. Consider now a design which controls for application/admission portfolios, like the one the authors introduced in Section 5.7.2. However, for this question we will simplify the design. Let us restrict to individuals who enroll in graduate program g_J or no program at all, and restrict to only outcomes in $t = 1$. Letting P_{ip} denote an indicator for whether person i has application/admission portfolio p , our model can be written as:

$$Y_{it} = \gamma_{g_J} G_{g_J} + \sum_{p \in \mathcal{P}} \eta_p P_{ip} + u_{it} \text{ for } t=1 \text{ and } i \text{ such that } g(i) \in \{g_J, 0\} \quad (2)$$

where $\mathcal{P}$ denotes the set of possible application/admission portfolios. For the pur-

pose of this question, let us use the “OLS” notion of u_{it} (treating u_{it} as an idiosyncratic residual term).

- (a) Describe the quasi-experimental design approximated by estimating Equation (2) on its given dataset. Who are the treated units and control units that we are trying to compare? What is held fixed in this quasi-experiment, and what is left to drive variation in treatment (G_{git})?

We are holding fixed the places an individual applied and the options for institutions they can attend. We are then comparing individuals who chose to go to program J (and successfully graduate) to those who chose to attend no graduate program (or choose program J and fail to graduate). The variation used to pin down γ_J is entirely coming from individual choices (and performance in law school).

- (b) Consider a design in which we instead randomly vary whether a student receives an *admission offer* from program J, fixing all other aspects of the application/admission portfolio. Recall the data at hand includes information on all the programs to which individual i applied, the programs into which i was admitted, and the program that i ultimately enrolled. Describe how you would implement that quasi-experiment with the available data.

Controlling for the schools an individual applied to and other observable characteristics, we can use admissions as an instrument for graduation in an IV design: cross-sectionally comparing those who receive an offer to similar individuals who do not. The key assumption would be regarding the conditional exogeneity of program admission offers. Dividing the reduced-form by the first stage then gives a LATE estimate when we further impose LATE assumptions.

- (c) Consider the assumptions under which the quasi-experiments of (a) versus (b) would return an average treatment effect over some subpopulation of interest. Which do you find to be more plausible?

The design of (a) requires that INDIVIDUAL choices conditional on one's enrollment options and other characteristics are as good as random across individuals. The design of (b) requires that UNIVERSITY enrollment choices conditional on individual application profiles and other characteristics are as good as random. The latter seems more likely to hold when it comes to individual outcomes. The additional IV assumptions we need for part (b) seem like they would plausibly hold.

- (d) Now consider taking Equation (2) to a dataset which also includes all available individuals (not just those in program g_J or no program at all). In doing so, we adapt the model as follows:

$$Y_{it} = \gamma_{g_j}^* G_{git} + \sum_{p \in \mathcal{P}} \eta_p P_{ip} + u_{it} \text{ for } t = 1 \text{ and for all } i \quad (3)$$

Where we write $\gamma_{g_j}^*$ to distinguish this parameter in Equation (3) from γ_{g_j} in Equation (2) (note these estimating equations differ only in the data to which

they are applied). Express the population values of the OLS estimators for γ_{g_J} and $\gamma_{g_J}^*$ in terms of expectations over potential outcomes. How do they differ?

$$\begin{aligned}\gamma_{g_J} &= E[Y_{i1}(g_J)|g(i) = g_J, y_{i1} > 0] - E[Y_{i1}(0)|g(i) = 0, y_{i1} > 0] \\ \gamma_{g_J}^* &= E[Y_{i1}(g_J)|g(i) = g_J, y_{i1} > 0] - E[Y_{i1}(g(i))|g(i) \neq g_J, y_{i1} > 0]\end{aligned}$$

- (e) Consider the conditions under which the OLS estimator of $\gamma_{g_J}^*$ is consistent for the ATT associated with taking program g_J over one's next alternative, $g^*(i)$, among those who actually took program g_J . Next, consider the conditions under which the OLS estimator of γ_{g_J} is consistent for the ATT associated with taking program g_J instead of no graduate program at all, among individuals who actually enrolled in program g_J . Is one set of conditions stronger or weaker than the other?

The OLS estimator of $\gamma_{g_J}^*$ will represent an ATT when $E[Y_{i1}(g(i))|g(i) \neq g_J, y_{i1} > 0] = E[Y_{i1}(g^*(i))|g(i) = g_J, y_{i1} > 0]$, where $g^*(i)$ is the program person i would expect among those in other programs are representative of expectations for individuals in program g_J had they not enrolled in program g_J .

The OLS estimator of γ_{g_J} will most plausibly represent an ATT when $E[Y_{i1}(0)|g(i) = 0, y_{i1} > 0] = E[Y_{i1}(0)|g(i) = g_J, y_{i1} > 0]$, meaning expectations among those who do not go to graduate school are representative of counterfactual expectations for those in program g_J , were they to not go to graduate school at all instead.

While one assumption is not strictly weaker or stronger than the other, the conditions for $\gamma_{g_J}^*$ imposes assumptions only on how individuals select into J . That is, if J is randomly assigned, then we can identify $\gamma_{g_J}^*$. This would not necessarily be enough to identify γ_{g_J} , as we also need to worry about selection into other graduate programs (instead of no graduate program) among those not in J , and how that relates to expected outcomes.

4. Finally, let us consider selection and attrition in this setting.

- (a) Suppose *enrollment* into graduate programs were randomly assigned. However, students know this assignment several years in advance. As a result, many students assigned to law programs would take on very low-paying jobs as a legal interns (law studies are hard and this experience is useful for passing classes). Is this a potential source of bias for the OLS, FEcg, and FE implementations of Equation (1)?

Taking the estimand to be the ATE associated with obtaining a law degree, there are at least two problems. First, as we have written them, the FEEg and FE estimators compares TRENDS over time between those with a given degree and other individuals. Under this channel, the common trends assumption is violated: students studying in JD programs would earn more in the future even if they don't graduate, simply because they'll stop working in their low-paid intern positions.

On the other hand, the OLS estimator only cross-sectionally compares earnings of those with a given degree to those with other degrees, in period $t = 1$. With programs being randomly assigned, this comparison of level differences in post-graduation earnings for individuals coming from different degree programs will give us a consistent estimate of average treatment effects, so long as graduation is exogenous conditional on program assignment.

Second, enrolling in law school is distinct from graduating with a law degree. This advance information is an opportunity for students to non-randomly select into internship programs, in a manner which is related to one's potential outcomes and which affects the probability of graduation conditional on enrollment. Thus, it is also a channel for selection bias for all three estimates.

- (b) You ask ChatGPT how to think about potential bias from attrition in this setting. It responds: "compare graduates of program g to otherwise-similar individuals who obtain no graduate program. If earnings trends in the years before enrollment are comparable across the two groups, then you can take this as evidence that evolving attrition over time is not biasing effect estimates." Suppose we had multiple years of pre-graduate-school data, (e.g., Y_{it} for negative values of t) so you could actually check those trends. Do you agree with this reasoning? Explain. Is there anything else you would want to know, to gauge if this check was potentially useful?

Common pre-trends is not conclusive evidence that common trends would (continue) to hold in the post-period. In this case, we would also want to know if attrition actually changed at all over the pre-period (by the nature of studying in a program, I would think attrition is flat in the pre-period). If individuals are not attriting, then attrition could not be influencing pre-enrollment average earnings.

The authors only observe earnings for individuals who stay in Texas after they graduate. For the next two parts, suppose that the decision to stay in Texas is taken according to the following rule:

$$T_i = \mathbb{1} \{ HighSchool_i \times \alpha + X_i' \beta + v_i > 0 \}$$

where v_i is an unobserved idiosyncratic taste toward living in Texas, and $HighSchool_i$ is an indicator that individual i went to high school in Texas.

- (c) Now suppose that attending *high school* in Texas is a shifter in one's propensity to stay in Texas *after college*. Suppose this characteristic is strictly exogenous and does not otherwise affect outcomes. Under those assumptions, discuss what we learn from the check described on page 5 of Altonji and Zhu (2025): "... restricting our analysis to individuals who attended high school in Texas does not significantly impact our returns estimates, suggesting that the underestimation [from attrition] may not be severe."

Supposing that this shifter is actually relevant, then this suggests that individuals selecting out of Texas do not differ systematically in their earnings from Texas residents (at least among those on the margin of moving out of Texas, who are actually influenced by this shifter). This would suggest that selection bias is not a concern when estimating average earnings by degree.

- (d) Suppose that the exercise of the previous part *did* find different effect estimates for those who went to high school in Texas. Suppose further that v_i and u_{it} are jointly normal. Describe how you could use this assumption to correct OLS effect estimates for bias due to attrition.

This is a straightforward application of a Heckman correction model. We can add a control function to the regression equation, controlling for $\lambda(HighSchool_i \times \alpha + X_i' \beta)$, where λ is the inverse mills ratio. This should eliminate selection bias from the conditional expectations underlying our diff-in-diff design.