ECON 4261: Final

May 6, 2023

Instructions

- Read every question **carefully**!
- Please ensure that your answers are **neat** and **legible**.
- There are 40 points available. Good luck!

Question 1 - 14 Points

In this section you are going to use some data to diagnose economic inequality. You have a cross-sectional dataset:

$$\{W_n, F_n, B_n, Yf_n, Y_n, G_n\}_{n=1}^N$$

where:

- W_n is log annual earnings in dollars
- $F_n \in \{0, 1\}$ indicates if person n has ever had children.
- B_n is person n's year of birth.
- Yf_n is the year in which person n had their first child (data missing if $F_n = 0$
- Y_n is the calendar year for the observation.
- $G_n = \{M, F\}$ indicates gender.

We want to use this dataset to analyze the relationship between fertility, gender, and wages. Let $X_n = (F_n, B_n, Yf_n, Y_n)$.

(a)[**4 points**] Start with the model:

$$\mathbb{E}[W_n|X_n, G_n = g] = \sum_a \mathbf{1}\{Age_n = a\}\mu_{a,g} + \sum_y \mathbf{1}\{Y_n = y\}\gamma_{y,g} + \alpha_g F_n$$

Describe how you would estimate this model and calculate standard errors for the coefficients. Provide enough details that someone could follow your instructions. *Do not* just say what command you would use in `R`.

- Split the data by gender
- Define the matrix $\mathbf{X}_g = [\mathbf{D}_{g,a}\mathbf{D}_{g,y}\mathbf{F}_g]$ appropriately with dummy variables for calendar year, and age. Drop the dummy for for the first calendary year to satisfy rank restrictions.
- Calculate
$$\hat{\beta}_g = (\mathbf{X}_g'\mathbf{X}_g)^{-1}\mathbf{X}_g'\mathbf{W}_g$$
where $\mathbf{W}_g$ is the column vector of all wages.

- Assuming homoskedasticity of errors, calculate the covariance matrix as:

  $$\hat{V}_g = (\mathbf{X}_g'\mathbf{X}_g)s_g^2$$

  where s_g^2 is the sample variance of the residuals. The standard errors are in the diagonal.

- You should use the heteroskedastic robust formula for SEs if you don't assume homoskedasticity.

(b)[3 points] In this model, which combination of parameters represents the *difference* in the gender gap between individuals with children ($F_n = 1$) and those without ($F_n = 0$):

$$\begin{aligned}\text{Diff in Gender Gap} = &(\mathbb{E}[W_n|X_n, F_n = 1, G_n = M] - \mathbb{E}[W_n|X_n, F_n = 1, G_n = F]) \\ &- (\mathbb{E}[W_n|X_n, F_n = 0, G_n = M] - \mathbb{E}[W_n|X_n, F_n = 0, G_n = F])\end{aligned}$$

This is given by $\alpha_M - \alpha_F$. Full marks as long as the answer is correct, part marks for trying to work with the conditional expectations implied by the model to get the answer.

(c) [3 points] Explain how you would calculate a 95% confidence interval for the difference in the gender wage gap (defined in part (b)) using your estimates and standard errors from part (a). You can assume that the set of male-specific coefficients is uncorrelated (i.e. zero covariance) with the set of female-specific coefficients.

For each set of estimates, the last row and last column of $\hat{V}_g$ contains the variance estimate for $\hat{\alpha}_g$. Call this number $\hat{v}_{\alpha,g}$. Hence, $Var(\hat{\alpha}_M - \hat{\alpha}_F) = v_{\alpha,M} + v_{\alpha,F}$, since the male and female coefficients have zero covariance. Thus, the confidence interval for the estimate is:

$$\hat{\alpha}_M - \hat{\alpha}_F + \pm 1.96\sqrt{v_{\alpha,M} + v_{\alpha,F}}$$

It's fine if the student doesn't know the critical values as long as they define them appropriately.

(d)[2 points] Explain how you would change the model to allow for the difference in the gender wage gap (defined in part(b)) to have a separate value for each calendar year. You only have to write the model, you do not have to describe estimation.

Many ways to write the model but one way is:

$$\mathbb{E}[W_n|X_n, G_n = g] = \sum_a \mathbf{1}\{Age_n = a\}\mu_{a,g} + \sum_y \mathbf{1}\{Y_n = y\}\left(\gamma_{y,g} + \alpha_{y,g}F_n\right)$$

where we would now drop a dummy for age instead of year so as not to complicate interpretation of the year-specific coefficients.

(e)[2 points] Explain how you would change the model to allow for the difference in the gender wage gap (define in part(b)) to grow linearly with the number of years since birth of the first child $(Y_n - Yf_n)$. You only have to write the model, you do not have to describe estimation.

Now the model is:

$$\mathbb{E}[W_n|X_n, G_n = g] = \sum_a \mathbf{1}\{Age_n = a\}\mu_{a,g} + \sum_y \mathbf{1}\{Y_n = y\}\gamma_{y,g} + \alpha_{g,1}F_n + \alpha_{g,2}F_n(Y_n - Yf_n)$$

Question 2 - 14 Points

Consider a staggered rollout of a federal subsidy for community college across counties in the US. Suppose you have the following data:

$$\{E_{c,t}, Y_{c,t}, P_{c,t}\}_{c=1,t=1}^{C,T}$$

where

- $P_{c,t}$ is a binary variable that indicates whether the policy has been introduced in county c at time t
- $E_{c,t}$ is the community college enrollment rate in county c at time t.
- $Y_{c,t}$ is average earnings in county c at time t.

(a) [**3 points**] Propose a difference-in-differences strategy for estimating the effect of the tuition subsidy policy on community college enrollment. Write a model where the effect of the policy is α, which is the same across counties and is immediate after the policy is introduced.

The diff in diff model is:

$$Y_{c,t} = \mu_c + \gamma_t + \alpha P_{c,t} + \epsilon_{c,t}$$

where μ_c are county fixed effects and γ_t are time fixed effects. A similar model works for enrollment.

(b) [**3 points**] Write an event-study model of the effect of the policy on enrollment and earnings where α_s is the effect of the policy s periods after introduction, and the effect of the policy is constant after 10 periods.

The event study model is:

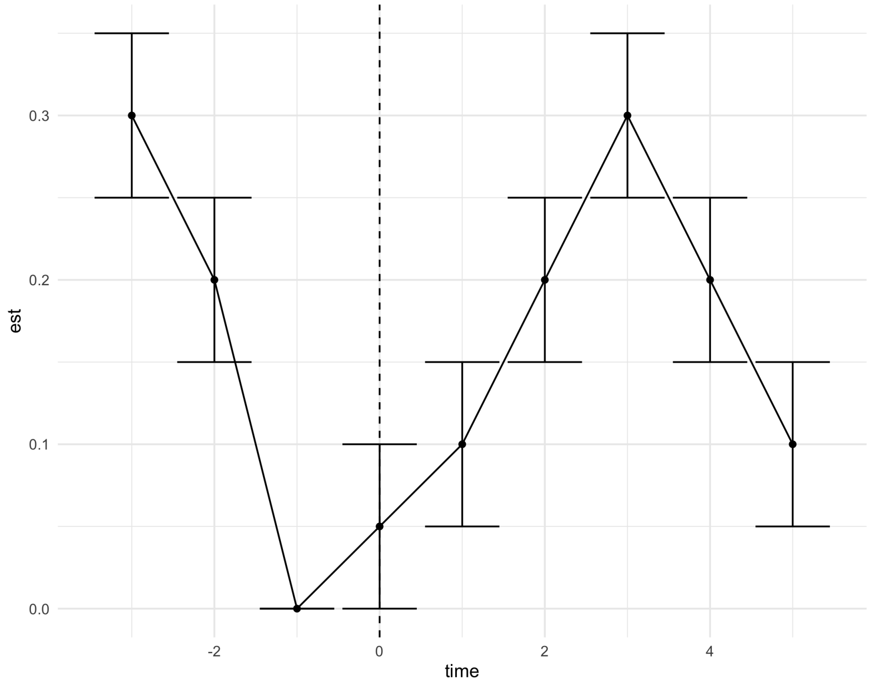
$$Y_{c,t} = \mu_c + \gamma_t + \sum_{\tau=-3}^{9} \alpha_\tau \mathbf{1}\{t_c^* - t = \tau\} + \alpha_{10}\mathbf{1}\{t_c^* - t \geq 10\} + \epsilon_{c,t}$$

where t_c^* is the time period in which the policy is introduced in county c. This model starts by estimating effects up to three periods before the policy but any number less than or equal to zero is valid.

(c) [4 points] Describe how you would estimate this event-study model with standard errors. You can be brief by provide enough details that someone could follow the steps. *Do not* just write what command you would use in `R`.

- Let $\mathbf{D}_t$ and $\mathbf{D}_c$ be a matrix of time and county dummies for the full dataset. Let $\mathbf{D}_\tau$ be a matrix of dummies for "event times" between -3 and 9, with a dummy in the final column if the event time is greater than or equal to 10.
- Drop the first time dummy and the event-time dummy corresponding to $\tau = -1$ (equivalent to normalizing $\alpha_{-1} = 0$).
- With this make the full matrix of covariates $\mathbf{X}$ and the steps become the same as Question (1)(a).
- Any normalizations/ommitted dummies are valid.

(d) [2 points] Suppose you estimate the event-study model and you plot your coefficients, $\hat{\alpha}_s$, with 95% confidence intervals with the below result:



Suppose that the policy introduction is completely unanticipated. Does this picture propose any challenges to your model assumptions? Which assumption does it appear to violate?

If the policy is unanticipated, there should be no effect of the policy before it is introduced, while the estimates here imply that there is. In other words, the model fails basic placebo test using a placebo time period. This suggests a violation of the parallel trends assumption.

(e) [2 points] Suppose that the policy is introduced *simultaneously* in every county in period 3 ($t = 3$). Will this cause problems for your difference-in-differences study? Which condition is violated?

Any version of the following is valid:

- With no variation in the timing of the policy, the effect of the policy is not identified.
- The rank condition is violated.
- Since the policy is introduced simultaneously everywhere, the time fixed effects cannot be separately identified from the effect of the policy.

Question 3 - 12 Points

Consider again the tuition subsidy from the previous question You are now going to estimate a TSLS model:

$$\begin{aligned} Y_{n,t} &= \mu_c + \kappa_b + \gamma_t + \alpha D_{n,t} \\ D_{n,t} &= \tilde{\mu}_c + \tilde{\kappa}_b + \tilde{\gamma}_t + \delta P_n \end{aligned}$$

where

- $Y_{n,t}$ is the earnings of person n at time t
- $D_{n,t}$ is a dummy that indicates whether person n at time t has been to community college.
- μ_c and $\tilde{\mu}_c$ are county fixed effects.
- γ_t and $\tilde{\gamma}_t$ are calendar time fixed effects.
- κ_b and $\tilde{\kappa}_b$ are year of birth fixed effects.
- P_n is an indicator for whether the tuition policy was in place in person n's county in the year that person n turned 18.

(a) [**4 points**] Describe how you would estimate the parameter α along with standard errors. You can be brief but provide enough details that someone could copy your steps. *Do not* just say what command you would use in `R`.

- Let $\mathbf{X}$ be the full matrix of dummies for county, birth year, and time period. Drop one birth year and one time period dummy from this matrix. In the last column of $\mathbf{X}$ place the vector of the college-attendance dummies for each observation.
- Let $\mathbf{Z}$ be the same matrix of county, birth year, and time period dummies, with the vector of policy variable indicators,P_n, in the final column.
- Calculate

  $$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

  Then get:

  $$\hat{\beta} = (\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}'\mathbf{Y}$$

  where we will find an estimate for α in the last entry of this vector of coefficient estimates.

- Next calculate the sample variance of the residuals given by $\mathbf{Y} - \mathbf{X}\hat{\beta}$, call it s^2.

- Estimate the variance of $\hat{\beta}$ as:

$$V_\beta = s^2(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}$$

  which is valid assuming $\mathbb{V}[\epsilon|Z] = \sigma^2$. The standard error for $\hat{\alpha}$ is the square root of the last digonal element of this matrix.

(b) [3 points] In the model, what is the logic for including the county and birth fixed effects? What might go wrong if we don't include these parameters?

Any one of the following arguments works:

- We want to only use variation in the presence of the policy *within* a county over time as an instrument for college attendance.

- Since outcomes might vary systematically across counties and cohorts, we want to control for potential permanent differences across counties and changes in outcomes over time.

- Without adding county and time fixed effects, variation in the policy is confounded with permanent differences across counties and aggregate cohort trends.

(c) [2 points] Suppose that the effect of community college on earnings is in fact heterogeneous across individuals. How can you interpret the TSLS estimand, α, in this case?

In this case we can interpret α as the local average treatment effect: the average effect of community college on earnings for those who are induced to go to community college by the tuition subsidy policy.

(d) [2 points] Suppose now that

1. You run a diff-in-diff analysis as in Question (2) and you find that the subsidy policy caused the price of community college tuition to increase; and

2. The subsidy is only offered to individuals in the county who are below a certain income threshold.

What condition is violated that was necessary for your interpretion in part (c)?

For part (c) we needed to assume that the instrument was *monotonic*, i.e. that all individuals are either pushed toward or away from the treatment by the subsidy. If the subsidy leads to an increase in the price of community college, then those who are not eligible for the policy might now be less likely to attend, while those who are eligible are more likely to attend. This would violate the monotonicity condition.