ECON3150/4150: Introductory Econometrics – Postponed Exam Spring 2026

Be brief and to the point.

Always motivate your answers.

Use the tables at the end of the exam where necessary.

Within a question, all sub-questions have equal weight.

1. [4 points] For each of the following statements, indicate whether it is true or false and justify your answer.

   (a) In a regression of Y on a single binary variable D (with a constant), the estimated slope coefficient equals the difference in sample means of Y between the two groups.

   ANSWER HINT: *True.* For a binary regressor $D \in \{0, 1\}$, OLS delivers fitted values that match the conditional means in each group. The intercept is $\bar{Y}_{D=0}$ and the slope is $\bar{Y}_{D=1} - \bar{Y}_{D=0}$.

   (b) If the error terms in an OLS regression are heteroskedastic, both the coefficient estimates and the standard errors are biased.

   ANSWER HINT: *False.* Heteroskedasticity does **not** bias the OLS coefficient estimates — they remain unbiased and consistent. What it does invalidate are the *homoskedastic* standard errors, which is why we use heteroskedasticity-robust standard errors. So only the standard errors (not the coefficients) are affected.

   (c) Including additional control variables in a regression always brings the estimated coefficient of interest closer to its true causal effect.

   ANSWER HINT: *False.* While adding "good" controls (confounders) can reduce omitted-variable bias, controlling for a **bad control** — a variable that is itself affected by the regressor of interest (a mediator or post-treatment variable) — can introduce bias by blocking part of the causal effect or conditioning on a collider. More controls is not automatically better.

   (d) A two-sided p-value of 0.04 implies that the probability of the null hypothesis being true is 4 percent.

   ANSWER HINT: *False.* The p-value is the probability of observing a test statistic at least as extreme as the one obtained, **assuming the null is true**. It is NOT the probability that the null is true. In our case, the 4% refers to how often we would see data this extreme if H_0 held — it says nothing about how likely H_0 is to be true given the data.

---

2. [4 points] A government awards a full-tuition merit scholarship to students whose national entrance exam score is at or above 80 out of 100. Students below the cutoff receive no scholarship. Table 1 reports mean earnings ten years after graduation for students with scores close to the cutoff.

Table 1: Earnings ten years after graduation, by entrance exam score

Entrance score range	N	Mean earnings ($)	Std. dev. ($)
Just below cutoff (75–79)	200	35000	10000
Just above cutoff (80–84)	200	44000	10000

(a) Compute the regression discontinuity (RD) estimate of the effect of the scholarship on earnings.

ANSWER HINT: The sharp RD estimate is the discontinuous jump in mean earnings at the cutoff: $\hat{\tau} = 44{,}000 - 35{,}000 = \mathbf{\$9{,}000}$. The scholarship is associated with about $9,000 higher annual earnings ten years after graduation.

(b) State the key identifying assumption needed for this estimate to be causal. What could invalidate it, and how could you test for that?

ANSWER HINT: **Continuity**: in the absence of the scholarship, the conditional expectation of earnings given the entrance score would be a continuous function of the score at the cutoff.

The assumption is invalidated if students can **precisely manipulate** their score (e.g. retakes targeting ≥ 80, or graders adjusting borderline scores upward), since those just above would then differ systematically from those just below. Standard checks: (i) a **density (bunching) test** for a jump in the histogram of scores at 80, and (ii) **balance checks** that pre-treatment covariates evolve continuously through the cutoff.

(c) A colleague proposes instead to compare mean earnings of all students who scored above 80 with all students who scored below 80, using the full sample (not just students near the cutoff). Discuss the pros and cons of this approach relative to the RD estimate from (a).

ANSWER HINT:
Pros: The full-sample comparison uses many more observations, yielding **smaller standard errors and a more precise estimate**.
Cons: The estimate is **biased for the causal effect** of the scholarship. Students with much higher scores (e.g. 95) would typically earn more than students with much lower scores (e.g. 60) even without the scholarship, because high scorers differ in ability, motivation, family background, etc. The full-sample comparison therefore conflates the scholarship effect with these confounders. The RD estimate, by restricting attention to nearly-identical students just on either side of the cutoff, isolates the scholarship effect — at the cost of larger standard errors and a more local interpretation.

(d) Test the null hypothesis that the scholarship has no effect on earnings at the

5 percent significance level. (Assume the two groups are independent samples and use the normal approximation.)

ANSWER HINT: The standard error of the difference in means is

$$\text{SE} = \sqrt{\frac{\sigma^2_{\text{below}}}{n_{\text{below}}} + \frac{\sigma^2_{\text{above}}}{n_{\text{above}}}} = \sqrt{\frac{10{,}000^2}{200} + \frac{10{,}000^2}{200}} = \sqrt{1{,}000{,}000} = 1{,}000.$$

The t-statistic is $t = 9{,}000/1{,}000 = 9$. Since $|t| = 9 > 1.96$, we **reject** the null at the 5 percent level: the scholarship has a statistically significant positive effect on earnings.

---

3. [6 points] You have data on 500 workers. The relevant variables are:

```
# wage:       hourly wage ($)
# educ:       years of education
# experience: years of work experience
# female:     =1 if female
# union:      =1 if union member
```

```
##                mean        SD min   max   N
## wage       24.26898 8.8622037 5.9 61.74 500
## educ       13.97000 2.4814139 8.0 20.00 500
## experience 14.85600 6.0501631 0.0 31.00 500
## female      0.52800 0.4997153 0.0  1.00 500
## union       0.16000 0.3669732 0.0  1.00 500
```

You estimated the following models:

```
reg1 = feols(log(wage) ~ educ, wages_df)
reg2 = feols(log(wage) ~ educ + experience + female + union, wages_df)
reg3 = feols(wage ~ educ + experience + female + union, wages_df)
```

```
##                           reg1           reg2          reg3
## Dependent Var.:     log(wage)      log(wage)          wage
##
## educ            0.079 (0.005)  0.078 (0.005)  1.88 (0.120)
## experience                     0.015 (0.002) 0.375 (0.049)
## female                        -0.206 (0.024) -4.91 (0.597)
## union                          0.113 (0.033)  2.90 (0.813)
## Constant         2.02 (0.077)   1.90 (0.077)  -5.49 (1.92)
## ---------------- ------------- -------------- -------------
## S.E. type                 IID            IID           IID
## SSR                      45.7           35.2      21,949.0
## Observations              500            500           500
```

a. Given that the standard deviation of log(wage) in the sample is 0.36, compute the R-squared of `reg1`.

ANSWER HINT: $R^2 = 1 - \dfrac{\text{SSR}}{\text{TSS}}$, where $\text{TSS} = (n-1)\,\widehat{\text{Var}}(\log \text{wage}) = 499 \times 0.36^2 \approx 64.7$ and $\text{SSR}_1 = 45.7$ from the table:

$$R^2 = 1 - \frac{45.7}{64.7} \approx$$

0.29. About 29 percent of the variation in log(wage) is explained by education alone.

b. Construct and interpret the 95 percent confidence interval for the coefficient on `educ` in `reg1`.

ANSWER HINT: From the normal table, $\Phi(1.96) = 0.975$, so the critical value is $z_{0.025} = 1.96$. The 95% CI is $0.079 \pm 1.96 \times 0.005 = [0.069,\ 0.089]$. We are 95% confident that each additional year of education is associated with between a 6.9% and an 8.9% higher hourly wage.

c. Using `reg2`, compute the predicted hourly wage for a non-unionized female worker with 16 years of education and 10 years of experience. Report your answer in dollars.

ANSWER HINT: Plug into `reg2`:

$$\widehat{\log(\text{wage})} = 1.90 + (0.078)(16) + (0.015)(10) + (-0.206)(1) + (0.113)(0) \approx 3.09.$$

Predicted wage $= \exp(3.09) \approx$ **$22.0 per hour**.

d. Interpret the coefficient on `female` in `reg2`. Roughly how much less per hour does a female worker earn than an otherwise identical male worker, evaluated at the sample mean wage?

ANSWER HINT: $\hat{\beta}_{\text{female}} \approx -0.206$. Holding education, experience, and union status constant, women earn approximately **20.6 percent less** per hour than men (linear approximation; the exact semi-elasticity is $\exp(-0.206) - 1 \approx -18.6\%$ — either answer is fine). At the sample mean wage of \$24.27, this translates to roughly $24.27 \times 0.206 \approx$ **$5.00 less per hour** (or $24.27 \times 0.186 \approx \$4.50$ using the exact figure).

e. Compute the F-statistic (assuming homoskedasticity) that tests the null hypothesis that the coefficients on `experience`, `female`, and `union` in `reg2` are jointly zero. What is the critical value at the 5% level?

ANSWER HINT: Under the null, the restricted model is `reg1` ($\text{SSR}_R = 45.7$); the unrestricted model is `reg2` ($\text{SSR}_U = 35.25$). With $q = 3$ restrictions and $n - k - 1 = 500 - 4 - 1 = 495$ denominator d.f.,

$$F = \frac{(\text{SSR}_R - \text{SSR}_U)/q}{\text{SSR}_U/(n-k-1)} = \frac{(45.7 - 35.25)/3}{35.25/495} = \frac{3.483}{0.0712} \approx 48.9.$$

The critical value from the $F_{3,\infty}$ table at the 5% level is **2.60**. Since $48.9 \gg 2.60$ we **reject** the null: `experience`, `female`, and `union` are jointly significant.

f. Using `reg3`, compute the predicted hourly wage for a non-unionized male worker with no years of education and no work experience. What does this result suggest about the linear specification?

ANSWER HINT: Plug zeros into `reg3`:

$$\widehat{\text{wage}} = -5.49 + 1.88(0) + 0.375(0) + (-4.91)(0) + 2.90(0) = -\$5.49 \text{ per hour.}$$

This is **negative** and therefore economically meaningless: a wage cannot be below zero. The result reflects two limitations of the linear specification: (i) we are **extrapolating** far outside the data range (the minimum education in the sample is 8 years), and (ii) the linear model imposes constant marginal effects throughout, so nothing prevents predictions from crossing zero.

4. [8 points] A government runs a randomized evaluation of a publicly-funded job training program. Unemployed workers are randomly assigned to receive an offer for free training (assignment $Z = 1$) or no offer ($Z = 0$). Among those offered, some workers complete the training and some do not. Let $D = 1$ if the worker actually completes the training (and $D = 0$ otherwise), and let Y denote monthly earnings (in USD) one year after assignment. Table 2 reports the slope coefficient (with standard error in parentheses) from two OLS regressions involving Y, D, and Z. The first column reports the average Y (and its standard deviation in parentheses) for the control group ($Z = 0$).

Table 2: Job training experiment ($N = 2{,}000$)

Outcome	Control mean ($Z = 0$)	Y on Z	D on Z
Monthly earnings ($)	1500 (700)	200 (50)	0.50 (0.02)

a. Interpret the reduced-form (ITT) estimate. What does it tell you about the program?

ANSWER HINT: The reduced form is the regression of the outcome Y on the assignment Z, given in the column "Y on Z": estimate = **$200**. Being *offered* training raises monthly earnings by approximately $200 on average — the policy-relevant impact of the offer itself, regardless of whether the worker actually completed the training.

b. Interpret the first-stage estimate. Is the instrument strong?

ANSWER HINT: The first stage is the regression of the treatment received D on the assignment Z, given in the column "D on Z": estimate = **0.50**. Being offered training raises the probability of actually completing training by 0.50 (50 percentage points). The t-statistic is $0.50/0.02 = 25$ (equivalently $F = 625$), far above the rule-of-thumb threshold of 10. The instrument is **strongly relevant**.

c. Compute the instrumental variables (Wald) estimate of the effect of completing the training on monthly earnings. Briefly explain why this estimate differs from the reduced-form (ITT) estimate in (a).

ANSWER HINT: $\widehat{\beta}_{\text{IV}} = \dfrac{\text{ITT}}{\text{first stage}} = \dfrac{\text{coef on } Z \text{ in } Y \text{ on } Z}{\text{coef on } Z \text{ in } D \text{ on } Z} = \dfrac{200}{0.50} =$ **$400 per month**. The IV estimate is **twice** the ITT because only **50 percent** of those offered training actually complete it. The ITT averages the effect of the *offer* over all assigned workers (including non-takers, on whom the offer has no effect), whereas the IV rescales by 1/first stage to recover the effect of actually completing the training.

d. Suppose a researcher ignored the random assignment Z and estimated the effect of training by simply comparing the average earnings of workers who completed the training ($D = 1$) with those who did not ($D = 0$). Why is this comparison likely to be biased, and in which direction?

ANSWER HINT: With imperfect compliance, D is not randomly assigned: workers **self-select** into training based on unobserved characteristics (motivation, ability,

job-search intensity) that also affect earnings. The naive comparison therefore conflates the causal effect of training with these confounders. The bias is most likely **upward** — workers who voluntarily complete training tend to be those who expect to benefit most or who have higher earnings potential anyway, so the comparison overstates the true causal effect. The IV estimate avoids this bias by using only the variation in D induced by the **random** offer Z.

Critical Values for the $F_{m,\infty}$ Distribution

```
## Rows denote degrees of freedom (m), and columns significance level (%)
##          10%     5%     1%
##
##  1 :  2.7055 3.8415 6.6349
##  2 :  2.3026 2.9957 4.6052
##  3 :  2.0838 2.6049 3.7816
##  4 :  1.9449 2.3719 3.3192
##  5 :  1.8473 2.2141 3.0173
##  6 :  1.7741 2.0986 2.8020
##  7 :  1.7167 2.0096 2.6393
##  8 :  1.6702 1.9384 2.5113
##  9 :  1.6315 1.8799 2.4073
##  10 :  1.5987 1.8307 2.3209
##  11 :  1.5705 1.7886 2.2477
##  12 :  1.5458 1.7522 2.1847
##  13 :  1.5240 1.7202 2.1299
##  14 :  1.5046 1.6918 2.0815
##  15 :  1.4871 1.6664 2.0385
##  16 :  1.4714 1.6435 2.0000
##  17 :  1.4570 1.6228 1.9652
##  18 :  1.4439 1.6038 1.9336
##  19 :  1.4318 1.5865 1.9048
##  20 :  1.4206 1.5705 1.8783
##  21 :  1.4102 1.5557 1.8539
##  22 :  1.4006 1.5420 1.8313
##  23 :  1.3916 1.5292 1.8104
##  24 :  1.3832 1.5173 1.7908
##  25 :  1.3753 1.5061 1.7726
##  26 :  1.3678 1.4956 1.7554
##  27 :  1.3608 1.4857 1.7394
##  28 :  1.3541 1.4763 1.7242
##  29 :  1.3478 1.4675 1.7099
##  30 :  1.3419 1.4591 1.6964
```

The Cumulative Standard Normal Distribution Function, $\Phi(z) \equiv \Pr(Z \leq z)$

```
## Rows denote 1st decimal value of z, and columns 2nd decimal value of z

## So for example, P(Z <= 1.08) = 0.860
```

	0	1	2	3	4	5	6	7	8	9
0.0 :	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1 :	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2 :	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3 :	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4 :	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5 :	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6 :	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7 :	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8 :	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9 :	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0 :	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1 :	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2 :	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3 :	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4 :	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5 :	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6 :	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7 :	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8 :	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9 :	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2.0 :	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1 :	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2 :	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3 :	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4 :	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5 :	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6 :	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7 :	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8 :	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9 :	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999