Unit 3 Practice Exam — Solutions

There are 42 total points.

1. *[12 points]* A researcher investigating the determinants of crime in the United Kingdom has data for 42 police regions over 22 years. She estimates by OLS the following regression:

$$ln(cmrt)_{it} = \beta_1 unrtm_{it} + \beta_2 proyth_{it} + \beta_3 ln(pp)_{it} + \alpha_i + \phi_t + u_{it}$$
$$i = 1, ..., 42; \quad t = 1, ..., 22$$

where *cmrt* is the crime rate per head of population, *unrtm* is the unemployment rate of males, *proyth* is the proportion of youths, *pp* is the punishment rate, measured as (number of convictions)/(number of crimes reported). α and ϕ are area and year fixed effects, where α_i equals one for area i and is zero otherwise for all i, and ϕ_t is one in year t and zero for all other years for $t = 2, ..., 22$. Note that there is no intercept and that ϕ_1 is not included.

(a) What is the purpose of excluding ϕ_1? What are the terms α and ϕ each likely to pick up? Discuss the advantages of using panel data for this type of investigation. *[4 points]*

(a) Since there is no constant in addition to the entity and time fixed effects, setting ϕ_t to one in year t and zero for all other years for $t = 1, ..., 22$ would result in perfect multicollinearity (the sum of all ϕ_t dummies plus the α_i fixed effects would span the same space as the included set).

(b) α picks up omitted variables that are specific to police regions and do not vary over time — for example, persistent attitudes toward crime or geographic features of a region.

(c) ϕ picks up effects that are common to all police regions in a given year — for example, nationwide macroeconomic shocks or changes in national policing policy that affect all regions equally.

(d) Panel data advantages: (1) We can control for unobserved, time-invariant region-specific factors that would otherwise cause omitted variable bias. (2) We can exploit within-region variation over time, which is more informative than a single cross-section.

(a) The results of the previous estimation are reported below, where the fixed effects are included but the coefficients are not reported. Heteroskedasticity and autocorrelationconsistent standard errors are reported below each coefficient.

$$\widehat{ln(cmrt)}_{it} = \underset{(0.109)}{0.063} unrtm_{it} + \underset{(0.179)}{3.739} proyth_{it} - \underset{(0.024)}{0.588} ln(pp)_{it}$$
$$R^2 = 0.904$$

Discuss the meaning of the three reported coefficients and their statistical significance. What is the effect of a ten-percent increase in the punishment rate? What assumptions, if any, are necessary in order to interpret this as the *causal* impact of the punishment rate on crime rates? *[5 points]*

(a) A higher male unemployment rate and a higher proportion of youths increase the crime rate, while a higher punishment rate decreases it. The coefficients on punishment (-0.588, SE $= 0.024$) and proportion of youths (3.739, SE $= 0.179$) are statistically significant at conventional levels, while the male unemployment rate (0.063, SE $= 0.109$) is not. The $R^2 = 0.904$ indicates the model explains about 90% of variation in log crime rates.

A 10% increase in the punishment rate changes $\ln(pp)$ by $\ln(1.10) \approx 0.0953$, so the effect on $\ln(cmrt)$ is $-0.588 \times 0.0953 \approx -0.056$.

Approximation: Since changes in logs are approximately equal to percent changes for small changes, this corresponds to roughly a **5.6**% decrease in the crime rate.

Exact: $\Delta cmrt \approx (e^{-0.056} - 1) \times 100 \approx -5.45\%$. The approximation overstates the magnitude slightly, but both round to about -5.5% to -5.6%.

For causal interpretation, we need to assume there are no other time-varying, area-specific factors correlated with the punishment rate and the crime rate — i.e., that the fixed effects have controlled for all relevant confounders.

(a) You now want to analyze what happens to the coefficients and their standard errors when the equation is re-estimated without fixed effects. In the resulting regression, $\widehat{\beta_2}$ and $\widehat{\beta_3}$ do not change by much, although their standard errors roughly double. However, the coefficient on *unrtm*, $\widehat{\beta_1}$, is now 1.340 with a standard error of 0.234. Why do you think that is? *[3 points]*

(a) This result suggests that male unemployment rates change slowly within a given police district over time and that this slow-moving variation is largely absorbed by the entity fixed effects. Without fixed effects, the coefficient on *unrtm* is much larger and now significant, which suggests that regions with persistently higher unemployment also have persistently higher crime — a level difference that reflects omitted time-invariant factors (captured by α_i in the fixed-effects model) rather than the causal effect of unemployment on crime.

3. *(15 points)* Consider the following model of education and fertility, using a 1988 survey of women in Botswana.

$$children_i = \beta_0 + \beta_1 edu_i + \beta_2 age_i + \beta_3 age_i^2 + u_i$$

Define *children* as the number of living children that women i has, *edu* is her years of completed education, and *age* is her reported age in years.

```
. regress children educ age agesq

      Source |       SS       df       MS              Number of obs =    4361
-------------+------------------------------           F(  3,  4357) = 1915.20
       Model |  12243.0295     3  4081.00985           Prob > F      =  0.0000
    Residual |  9284.14679  4357  2.13085765           R-squared     =  0.5687
-------------+------------------------------           Adj R-squared =  0.5684
       Total |  21527.1763  4360  4.93742577           Root MSE      =  1.4597

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0905755   .0059207   -15.30   0.000     -.102183   -.0789679
         age |   .3324486   .0165495    20.09   0.000     .3000032     .364894
       agesq |  -.0026308   .0002726    -9.65   0.000    -.0031652   -.0020964
       _cons |  -4.138307   .2405942   -17.20   0.000    -4.609994    -3.66662
------------------------------------------------------------------------------
```

(a) Interpret the coefficient on *edu*, including units. *[2 points]*

(b) Does *age* have a non-linear relationship with the number of live children? How do you know? *[2 points]*

(a) An additional year of education is associated with a reduction of approximately 0.09 children (or 0.9 children per 10 additional years of schooling), holding age constant. This estimate is highly statistically significant ($t = -15.30$, $p < 0.001$).

(b) Yes, age has a non-linear (specifically concave) relationship with the number of children. We know this because the coefficient on age^2 is negative (-0.0026) and statistically significant ($t = -9.65$, $p < 0.001$), allowing us to reject the null hypothesis that the relationship is linear.

(a) Ronalda hypothesizes that we can use the variable $frsthalf$, which is a dummy equal to one if the woman was born during the first six months of the year, as an instrument for education. Her intuition is that in some countries, students can drop out at a certain age, so some students born earlier in the year may complete less schooling. What assumption(s) need to hold in order for her instrument to be valid? Which, if any, can be tested? *[4 points]*

(b) Compare the OLS results (above) and instrumental variable results (below). How do the returns to education differ between the two specifications? Explain why you think that is. *[3 points]*

```
. ivregress 2sls children (educ = frsthalf) age agesq

Instrumental variables (2SLS) regression          Number of obs =     4361
                                                  Wald chi2(3)  =  5300.22
                                                  Prob > chi2   =   0.0000
                                                  R-squared     =   0.5502
                                                  Root MSE      =     1.49

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.1714989   .0531553    -3.23   0.001    -.2756813   -.0673165
         age |   .3236052   .0178514    18.13   0.000     .2886171    .3585934
       agesq |  -.0026723   .0002796    -9.56   0.000    -.0032202   -.0021244
       _cons |  -3.387805   .5478988    -6.18   0.000    -4.461667   -2.313943
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   age agesq frsthalf
```

(c) Name one specific example of a potential threat to the validity of the instrument $frsthalf$. Explain why it is a threat. *[2 points]*

(d) Provide a specific example of reverse causality that might arise in this model. *[2 points]*

(a) For $frsthalf$ to be a valid instrument, three conditions must hold:

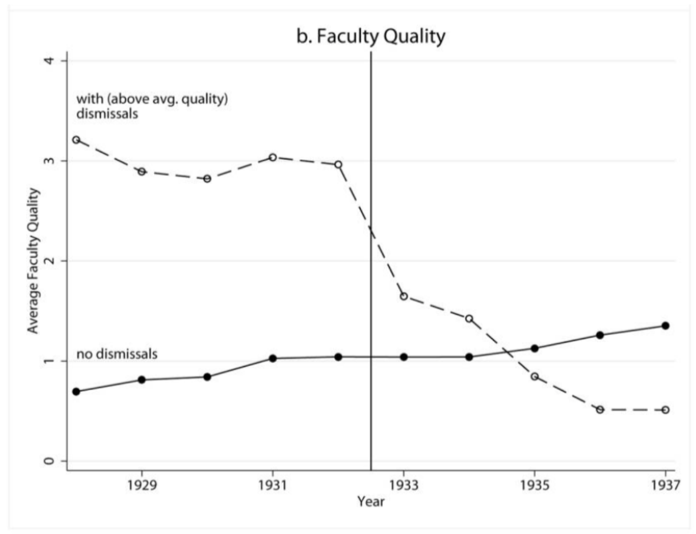
- **Relevance**: Women born in the first half of the year must actually complete less schooling (i.e., $frsthalf$ must be correlated with *edu*). This *can* be tested by running the first-stage regression and checking whether the coefficient on $frsthalf$ is significant and whether the F-statistic exceeds 10.
- **Exogeneity**: $frsthalf$ must be uncorrelated with the error term — i.e., birth month must not be systematically related to other factors that affect fertility (e.g., family wealth or season-of-birth health effects). This *cannot* be directly tested.
- **Exclusion restriction**: $frsthalf$ must affect *children* only through *edu*, not directly. This *cannot* be directly tested.

(b) After using an IV, we find that additional education now results in a stronger negative effect on fertility (-0.171 vs. -0.091 OLS). If the instrument is valid, this suggests that OLS was underestimating the true negative effect of education on fertility, likely due to omitted variable bias (e.g., ability or family background positively correlated with both education and desired family size, attenuating the negative relationship).

(c) A specific threat to validity: Women from poorer households may be more likely to be born in the first half of the year (e.g., due to seasonal variation in conception patterns correlated with economic conditions), and poverty independently reduces both education and the number of children women choose to have. If so, $frsthalf$ would be correlated with the error term, violating exogeneity.

(d) There might be reverse causality if having children causes women to drop out of school, reducing their years of completed education. In that case, the causal arrow also runs from *children* to *edu*, and OLS estimates of β_1 would be biased.

3. *(15 points)* Waldinger (2010) looks at the impact of faculty quality on the outcomes of PhD students. He uses a natural experiment caused by the dismissal of scientists in Nazi Germany. Upon gaining power in 1933, the new Nazi government fired all Jewish and "politically unreliable" scholars from German universities. Depending on the composition of their faculty, some universities had to dismiss more than half of their faculty, while others were not affected at all. The following figure summarizes the impact of this policy on faculty quality between universities with high levels of dismissals (dashed line) and those with no dismissals (solid line).



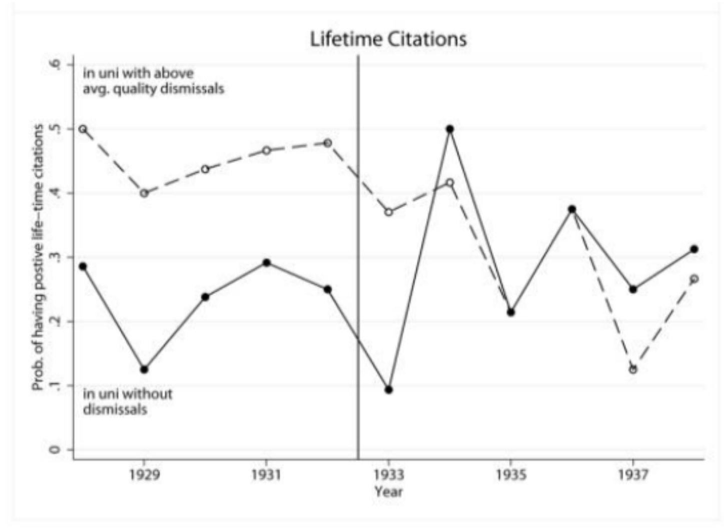
(a) Suppose you have panel data on graduate productivity and faculty quality at the university-year level. Aside from potential measurement error, provide one specific reason why the following equation would be unlikely to measure the causal impact of faculty quality on PhD outcomes. *[3 points]*

$$PhDCitations_{uni,y} = \beta_0 + \beta_1 FQual_{uni,y} + u_{uni,y}$$

$PhDCitations_{uni,y}$ is the number of lifetime citations the graduating y cohort from university uni receives. $FQual_{uni,y}$ is the number of citation-weighted publications per faculty member at university uni in year y.

(b) Based on the description above, propose a difference-in-differences model that you could use to estimate the causal impact of faculty quality on graduate student productivity. Define any new variables you use. Indicate which coefficient(s) will show the causal impact. *[5 points]*

(c) Name what, if any, assumptions you need for the model you write in part (b) to reflect a causal relationship. *[2 points]*



(d) The above figure shows the average level of *PhDCitations* by year separately for universities with a high number of dismissals and those without dismissals. Based on that figure, does faculty quality affect graduate student productivity? Explain how you know. *[3 points]*

(e) What, if any, concerns do you have about the *external validity* of your model? Explain. *[2 points]*

(a) Omitted variable bias (OVB): well-funded institutions may attract both higher-quality faculty and higher-ability graduate students, so $FQual$ would be correlated with the error term. Alternatively, reverse causality: productive graduate students may attract higher-quality faculty to a university, making the direction of causation ambiguous.

(b) Define two binary variables: $Post$, equal to one for years 1933 onward and zero otherwise, and $HighDismiss$, equal to one for universities that experienced a high level of dismissals.

$$PHDCitations_{uni,y} = \beta_0 + \beta_1 HighDismiss_{uni} + \beta_2 Post_y + \delta_1 (HighDismiss_{uni} \times Post_y) + u_{uni,y}$$

The coefficient of interest is δ_1, which captures the differential change in PhD citations at high-dismissal universities after 1933, relative to the trend at low-dismissal universities.

(c) The key assumption is **parallel trends**: in the absence of the Nazi dismissals, PhD student productivity would have followed the same trend at high-dismissal and low-dismissal universities. This is plausible if, prior to 1933, the two groups of universities were on similar trajectories in graduate student outcomes (which the pre-1933 period in the figure can be used to assess).

(d) Yes, faculty quality appears to affect graduate student productivity. Before 1933, universities with above-average-quality dismissals had higher PhD citations than those with no dismissals. After the dismissals in 1933, the gap closes dramatically — the high-dismissal universities' PhD citation rates fall toward those of the low-dismissal universities. This convergence is consistent with the reduction in faculty quality causing the decline in graduate student productivity.

(e) This is a very particular natural experiment: the context involves a massive, politically-forced reduction in faculty — far larger and more abrupt than normal faculty turnover. The results may not generalize to more typical settings (e.g., one or two faculty departures at a university). Additionally, there may have been broader disruptions to university functioning during this period (e.g., student emigration, research funding cuts) that confound the faculty quality channel.