**International School of Economics at TSU**

**Econometrics 2**

**Lasha Chochua**

## Problem Set 8

**Instructions:** You are encouraged to solve the problems before the recitation. Additionally, you are encouraged to work in groups. It is **not mandatory** to submit solutions unless stated otherwise. However, if you would like to share your solution, I would be happy to review it.

**Problem 1:** Consider a model for the health of an individual:

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + u_1 \quad (1)$$

where *health* is some quantitative measure of the person's *health*; *age*, *weight*, *height*, and *male* are self-explanatory; *work* is weekly hours worked; and *exercise* is the hours of exercise per week.

**a.** Why might you be concerned about *exercise* being correlated with the error term $u_1$?

**b.** Suppose you can collect data on two additional variables, *disthome* and *distwork*, the distances from home and from work to the nearest health club or gym. Discuss whether these are likely to be uncorrelated with $u_1$.

**c.** Now assume that *disthome* and *distwork* are in fact uncorrelated with $u_1$, as are all variables in equation (1) with the exception of *exercise*. Write down the reduced form for *exercise*, and state the conditions under which the parameters of equation (1) are identified.

**Problem 2:** Consider the following model to estimate the effects of several variables, including cigarette smoking, on the weight of newborns:

$$\log(bwght) = \beta_0 + \beta_1 male + \beta_2 parity + \beta_3 \log(faminc) + \beta_4 packs + u, \quad (2)$$

where *male* is a binary indicator equal to one if the child is male, *parity* is the birth order of this child, *faminc* is family income, and *packs* is the average number of packs of cigarettes smoked per day during pregnancy.

**a.** Why might you expect *packs* to be correlated with $u$?

**b.** Suppose that you have data on average cigarette price in each woman's state of residence. Discuss whether this information is likely to satisfy the properties of a good instrumental variable for *packs*.

**c.** Use the data in BWGHT.RAW to estimate equation (2). First, use OLS. Then, use 2SLS, where *cigprice* is an instrument for *packs*. Discuss any important differences in the OLS and 2SLS estimates.

**d.** Estimate the reduced form for *packs*. What do you conclude about identification of equation (2) using *cigprice* as an instrument for *packs*? What bearing does this conclusion have on your answer from part c?

**Problem 3:** A researcher is interested in estimating the variance of the error term in Equation (1) in the slides.

**a.** Suppose she uses the estimator from the second-stage regression of TSLS:

$$\hat{\sigma}_a^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS}\hat{X}_i\right)^2,$$

where $\hat{X}_i$ is the fitted value from the first-stage regression. Is this estimator consistent? (For the purposes of this question, suppose that the sample is very large and the TSLS estimators are essentially identical to $\beta_0$ and $\beta_1$.)

**b.** Is

$$\hat{\sigma}_b^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS}X_i\right)^2$$

consistent?

**Problem 4 (Replication): Card (1995) – Returns to Schooling**

This problem asks you to replicate the main findings of **Card, D. (1995)**: *"Using Geographic Variation in College Proximity to Estimate the Return to Schooling."* The paper estimates the causal effect of education on earnings using proximity to a four-year college as an instrument for years of schooling. The data come from the NLS Young Men Cohort and are available at:

`https://github.com/scunning1975/mixtape/raw/master/card.dta`

The structural equation of interest is:

$$\log(wage_i) = \alpha + \delta \cdot educ_i + \gamma' X_i + \varepsilon_i$$

where $educ_i$ is years of schooling, $X_i$ includes experience, race, region, marital status, and urban indicators, and $\varepsilon_i$ captures unobserved ability. The instrument is $nearc4_i$, a dummy equal to one if the individual grew up in a county with a four-year college.

**a.** Estimate the structural equation by OLS. Report the coefficient on *educ* and its standard error. Why might this estimate be biased?

**b.** Estimate the first-stage regression of *educ* on *nearc*4 and all controls. Report the coefficient on *nearc*4, its standard error, and the first-stage $F$-statistic. Is the instrument strong by the Stock–Yogo criterion? What does the first-stage coefficient tell us economically?

**c.** Estimate the structural equation by 2SLS using *nearc*4 as the instrument for *educ*. Report the 2SLS coefficient on *educ* and compare it to the OLS estimate. Is the 2SLS estimate larger or smaller? What does this imply about the direction of omitted variable bias in OLS?

**d.** Implement 2SLS manually by running the two stages by hand. Verify that the point estimate from the manual second stage matches the `ivreg()` output. Do the standard errors match? Explain why or why not, connecting your answer to Problem 3 above.

**e.** Interpret the 2SLS estimate as a Local Average Treatment Effect (LATE). Who are the compliers in this setting? Why might the LATE differ from the Average Treatment Effect (ATE)?

**f.** Compute the Anderson–Rubin test of $H_0 : \delta = 0$. Does it agree with the 2SLS $t$-test? Comment on whether the two tests could in principle disagree, and what that would imply.

### Problem 5 (Replication): Graddy (1995) – Fulton Fish Market

This problem asks you to replicate the main findings of **Graddy, K. (1995)**: *"Testing for Imperfect Competition at the Fulton Fish Market."* The paper estimates the price elasticity of demand for whiting using weather conditions at sea as instruments for price. The dataset contains daily observations from the Fulton Fish Market in New York.

The structural demand equation is:

$$\log(qty_i) = \alpha + \delta \cdot \log(price_i) + \gamma' X_i + \varepsilon_i$$

where $qty_i$ is the quantity of whiting sold, $price_i$ is the average daily price, and $X_i$ includes day-of-week dummies and a time trend. The instrument is $stormy_i$, a dummy equal to one if weather conditions at sea were stormy in the days prior to the observation.

**Data:** The dataset `fulton_fish_markets_data.csv` is available on **Moodle**.

**a.** Explain the identification strategy. Why does stormy weather provide a valid instrument for price? Discuss both the relevance condition and the exclusion restriction. Is the exclusion restriction likely to hold exactly?

**b.** Estimate the demand equation by OLS without any controls. Report the coefficient on $\log(price)$ and interpret it as a price elasticity. Why is this estimate likely to be biased?

**c.** Add day-of-week dummies and a rainy weather indicator as controls and re-estimate by OLS. Does the price elasticity change? Add month fixed effects and repeat. Discuss what each set of controls is doing.

**d.** Estimate the demand equation by 2SLS using *stormy* as the instrument for $\log(price)$, without controls. Report the 2SLS elasticity estimate and compare it to the OLS estimate. Is the 2SLS estimate larger or smaller in absolute value? What does this imply about the nature of simultaneity bias in this market?

**e.** Re-estimate by 2SLS including day-of-week dummies, the rainy indicator, and month fixed effects. Report the first-stage $F$-statistic. Is the instrument strong? Is the 2SLS elasticity estimate robust to the inclusion of controls?

**f.** Based on your results, does the Fulton Fish Market appear to be competitive? What would a perfectly competitive market imply about the relationship between stormy weather and quantity sold?