Microeconometrics

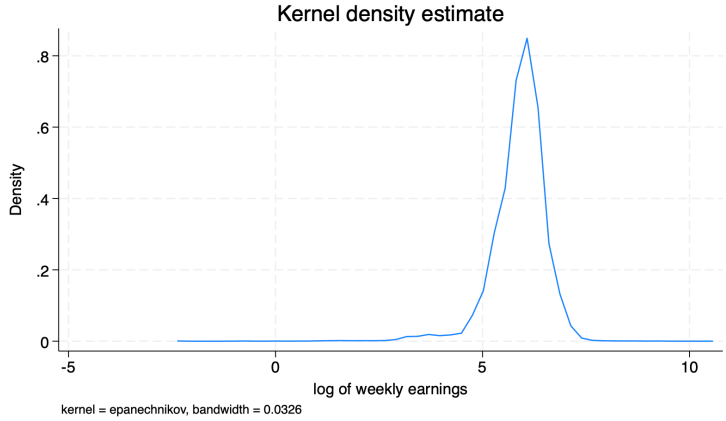
Problem set 1 - Solutions

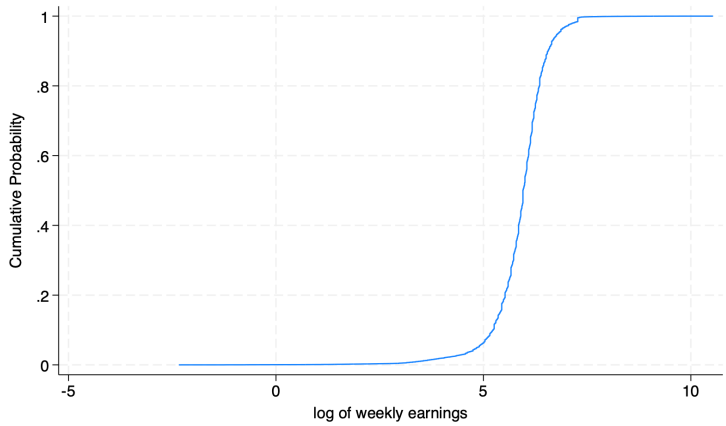
1. (100%) **Instrumental variables.** Consider the dataset **QOB** that is comprised of a subset of the data used in the seminal paper by Angrist and Krueger (1991) "*Does compulsory school attendance affect schooling and earnings?*". In the paper the quarter of birth of individuals is used as an instrument for education in order to estimate the impact of compulsory school on earnings. The authors use samples from Census data for men born in 1920s, 1930s, and 1940s. IMPORTANT: focus on the sub-sample of men born in 1930-1939. Year of birth is the variable YOB and includes only the last 2 digits.

   (a) Compute the non-parametric cdf and pdf of the (log-)wage and interpret the results.

   SOLUTION: We begin by dropping observations that are outside the sample that needs to be used (men not born in 1930-1939). The non-parametric estimates of the cdf and pdf of the log-wage indicates that the variable is distributed similarly to a normal distribution and is relatively symmetric around the mean.

   Stata command: keep if YOB>= 30 & YOB<= 39
   kdensity LWAGE
   cdfplot LWAGE

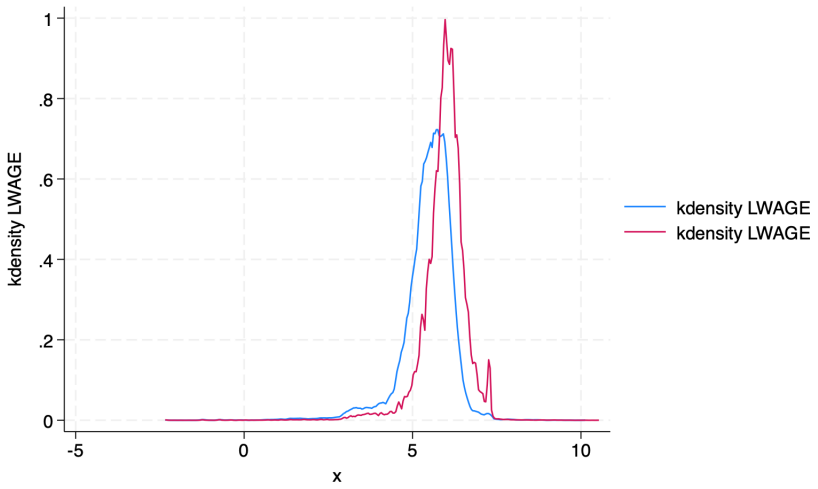




(b) Compute the non-parametric pdf of the (log-)wage for individuals with less or equal than 8 years of education and for individuals with more than 8 years of education and compare the two.

SOLUTION: The distribution of log-wage for individuals with more than 8 years of education is shifted to the rights and is slightly less sparse (density is higher in the central part of the distribution), indicating lower variance. Overall, this is indicative of returns to education (higher education increases wages). Additional note is that, as the number of observations goes down, the estimated pdf is less smooth with the pre-selected bandwidth.

Stata command:
tw (kdensity LWAGE if EDUC <= 8)(kdensity LWAGE if EDUC > 8 & !mi(EDUC))



(c) Estimate the returns to education by OLS using age and squared age as control variables.[1] Interpret the results and explain why the estimated returns to education might not have a causal interpretation.

SOLUTION: The marginal impact on wages of getting older by one year is given by $(0.0711 - 2 \times 0.0034 AGE) \times 100\%$.

One additional year of schooling increases predicted wages by approximately 7.1%, on average, ceteris paribus. This impact may not have a causal interpretation because there are concerns about omitted variable bias in this model. Perhaps, innate ability is clearly correlated with both years of schooling and wages.

Stata command:
reg LWAGE EDUC AGE AGESQ

(d) Explain why the quarter of birth might be a good instrument for education when estimating returns to education.

SOLUTION: "If the fraction of students who desire to leave school before they reach the legal dropout age is constant across birthdays, a student's birthday should be expected to influence his or her ultimate educational attainment. This relationship is expected because, in the absence of rolling admissions to school, students born in different months of

[1] Returns to education are estimated using a linear regression of wages (generally in logs) on years of schooling.

the year start school at different ages. This fact, in conjunction with compulsory schooling laws, which require students to attend school until they reach a specified birthday, produces a correlation between date of birth and years of schooling.
Students who are born early in the calendar year are typically older when they enter school than children born late in the year." (Angrist and Krueger, 1991)

(e) Construct a dummy variable, *first_qob*, which equals one for men born in the first quarter of the year and zero otherwise. Compute the IV estimate $IV = (Z'X)^{-1}Z'Y$ of returns to education considering *first_qob* as instrumental variable. Compare the estimate with the OLS returns to education estimated in the regression of *lwage* on a constant and years of education.

SOLUTION: In the IV setting, one additional year of schooling is associated with a 10.2% increase in wages, on average, ceteris paribus. OLS regression suggests that one additional year of schooling is associated with a 7.1% increase on wages, on average, ceteris paribus.
Stata commands:
gen first_qob= (QOB==1)
ivregress 2sls LWAGE (EDUC = first_qob)

(f) Suppose in the following that we have three instrumental variables, $Z1$, $Z2$, and $Z3$ representing dummy variables for first-, second-, and third-quarter births, respectively. Generate these three instrumental variables.

SOLUTION: Stata commands:
gen Z1 = first_qob
gen Z2= (QOB==2)
gen Z3= (QOB==3)

i. Describe and estimate the first-stage equation with multiple instruments. Include the following explanatory variables as additional control variables: *age*, *agesq*, *race*, *married*, and *smsa*.

SOLUTION: In the first-stage equation, the dependent variable is the endogenous variable (EDUC) and the explanatory variables are the set of instruments and the exogenous variables. Notice however that including all instruments leads to collinearity between the set of instruments and *agesq*. You can therefore drop *agesq* and run the following regression:

Stata command: reg EDUC Z1 Z2 Z3 AGE RACE MARRIED SMSA

ii. Compute an F-test under the null hypothesis that the quarter of birth dummy variables have no effect on the total years of education. Are these instruments valid?

SOLUTION: Let δ_1, δ_2, and δ_3 denote the coefficients on $Z1$, $Z2$, and $Z3$, respectively. The null and alternative hypotheses in a test of joint significance of the three quarter of birth dummy variables are:

$$H_0: \quad \delta_1 = \delta_2 = \delta_3 = 0$$
$$H_a: \quad \text{Not } H_0$$

The joint significance test yields an F-statistic equal to 12.37 and the associated p-value is 0.000. Therefore, we reject the null hypothesis in favor of the alternative hypothesis and the instruments are jointly statistically significant (a rule of thumb

for validity of the instruments in the first stage is an F-statistic greater than 10).

Stata: test Z1 Z2 Z3

iii. Estimate the returns to education by 2SLS and compare the results to standard OLS estimates [consider the same set of control variables as in part (i)].

SOLUTION: In this regression model, an additional year of schooling is associated with a 12.5% increase in wages, on average, ceteris paribus. OLS regression yields an estimate of the returns to schooling equal to 0.0642, i.e., an additional year of schooling is associated with a 6.42% increase in wages, on average, ceteris paribus.

Stata: ivregress 2sls LWAGE (EDUC= Z1 Z2 Z3) AGE RACE MARRIED SMSA

Notice that including *agesq* leads to collinearity. Therefore estimating the following regression will make the software drop one of the instruments.

Stata: ivregress 2sls LWAGE (EDUC= Z1 Z2 Z3) AGE AGESQ RACE MARRIED SMSA

iv. Are the instruments exogenous?

SOLUTION: If quarter of birth is uncorrelated with unobserved determinants of wages (once you control for age, race, marital status,..), then dummy variables for each value of the quarter of birth variable should also be uncorrelated.
If covered in class before the problem set deadline: we can perform a test of instrument exogeneity. The overidentification test can be described as follows: obtain the residuals from 2SLS regression such that $uhat\_IV = y - x\widehat{\beta}_{IV}$; then run a regression where the dependent variable are the residuals and the independent variables are the instruments and the exogenous variables ($uhat\_IV = z\gamma + \epsilon$). Test the joint significance of the instruments: if the instruments are exogenous then we should fail to reject the null hypothesis that all the coefficients are equal to zero ($H_0 : \gamma = 0$). The test statistic is equal to mF and follows a χ^2_{m-k}.
More specifically, the null hypothesis $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$ in the following equation stands for exogeneity of the instruments, whereas the alternative hypothesis H_a : Not H_0 says that at least one instrument is not exogenous.

$$uhat\_IV = \gamma_0 + \gamma_1 Z1 + \gamma_2 Z2 + \gamma_3 Z_3 + v$$

The F-statistic of global significance of this regression equals 1.78 (with associated p-value equal to 0.1480) and therefore we fail to reject the null hypothesis that the instruments are exogenous at 10% significance level.

Stata commands:
ivregress 2sls LWAGE (EDUC= Z1 Z2 Z3) AGE RACE MARRIED SMSA
predict uhat_IV, res
reg uhat_IV Z1 Z2 Z3
Alternatively: ivregress 2sls LWAGE (EDUC= Z1 Z2 Z3) AGE RACE MARRIED SMSA, robust
estat overid