Econometrics II. Assignment 2: Quantile Regression, Static linear panel data model

Due date: Sunday, January 15, 11.59 pm. Hand in your solutions as a **single .pdf file** including your code via Canvas. Assignments can be made in teams of two students. Both teammates have to submit solutions via Canvas.

Question 1: Quantile regression

In this exercise you will use a dataset on medical expenditures `assignment2a_2023.csv(\.dta)`. The dataset includes the following variables:

Variable	Description
lntotexp	log of total medical expenditure
age	age
female	1 if female
white	1 if white
totchr	number of chronic problems
suppins	1 if has a supplementary private insurance

Consider the following quantile regression model for q^{th} quantile:

$$Y_q(LnTotExp_i|X_i) = \beta_{0q} + \beta_{1q}TotChr_i + \beta_{2q}SuppIns_i + \beta_{3q}Age_i + \beta_{4q}Female_i + \beta_{5q}White_i + u_i$$

(i) Create a quantile plot for the log of total medical expenditure. Draw lines to indicate the median, the 10^{th} percentile, and the 90^{th} percentile. Describe your plot.

(ii) Estimate the model for the quantiles $q = 0.1$, 0.25, 0.5, 0.75 and 0.9. Briefly explain your result. Compare quantile regression results to OLS estimates.

(iii) Graph the estimated coefficients from the quantile regressions for q from 0.05 to 0.95 in increments of 0.05, together with their 95% confidence interval and the corresponding estimates from a linear regression (and their 95% confidence interval). Discuss your findings.

Question 2: Fixed and random effects

(i) Why does the process of taking each observation relative to its individual-level mean have the effect of "controlling for individual effects"?

(ii) Two-way fixed effects with terms for both individual and time are often referred to as "controlling for individual and time effects". Why might a researcher want to do this rather than just taking individual fixed effects and adding a linear/polynomial/etc. term for time?

(iii) Why random effects is likely to do a better job of estimating the individual effects than fixed effects, if its assumptions hold?

Question 3: Discrimination and returns to schooling

The dataset `assignment2b_2023.csv(\.dta)` is a panel dataset on male workers, working at least 30 hours a week. It has data for the years 1980-1994, 1996, 1998 and 2000. It contains the following variables:

Variable	Description
id	personal identifier
time	year − 1980
earnings	hourly wage
age	age
agesq	age squared
school	years of schooling
ethblack	1 if black ethnicity
urban	1 if living in urban area
regne	1 if region north-east
regnc	1 if region north-central
regw	1 if region west
regs	1 if region south
asvabc	index test score

The variable *asvabc* is a composite test score for skills like arithmetic reasoning, word knowledge, paragraph comprehension. A higher value of *asvabc* goes together with a higher test score. It is generally considered to be a good indicator for ability. It is measured only once, upon entering the panel, and therefore it is constant over time. A researcher is interested in estimating the effect of schooling on the hourly wage. A basic equation is

$$\ln Earnings_{it} = \beta_0 + \beta_1 School_{it} + \beta_2 Age_{it} + \beta_3 AgeSq_{it} + \delta' X_{it} + u_{it}$$

where $\mathbf{X}_{it}$ includes remaining variables.

(i) First use pooled OLS to check the impact of including and excluding *asvabc* on the estimate of β_1. Present and explain the result.

The researcher is asked to analyze whether returns to schooling, as measured by the parameter β_1, is different for workers with a black ethnicity. A difference in returns to schooling by ethnicity is sometimes interpreted as an indication of discrimination on the labour market. Two basic ways

to model heterogeneity in the schooling parameter by ethnicity are (1) including a cross effect of schooling and ethnicity and (2) estimating separate equations by ethnicity.

(ii) Perform a pooled OLS analysis to obtain insight in the heterogeneity of returns to schooling by ethnicity. Present the results and comment on the outcomes. What are the conclusions based on this?

So far, the panel nature of the data has hardly been exploited. Random effects estimation can improve efficiency of the estimates compared to pooled OLS.

(iii) Perform the analysis for heterogenous schooling effects using the random effects model. Present the results and compare the outcomes with the pooled OLS results obtained before. Interpret the outcomes.

Alternatively, panel data can be exploited to perform fixed effects (or within group) estimation.

(iv) A priori, would you plead for using fixed effects estimation or random effects estimation? Explain your answer.

(v) Apply the fixed effects estimator to analyze the heterogenous schooling effects. Interpret the outcomes.

(vi) Fixed effects estimation may not be as efficient as random effects estimation, but is robust to correlation between regressors and the random effect. Can we perform a Hausman test in this context? Perform the test you propose.

(vii) Perform Mundlak estimation of the model. Present the results of estimation and test for the joint significance of the within-group means.

(viii) What are your overall conclusions from the analysis of heterogeneity in returns to schooling by ethnicity?

(ix) To gain insights in the impact of nonresponse and attrition, the researcher applies a variant of the Verbeek and Nijman-test. He defines the dummy variable d_i which is 1 if the individual is in the panel for more than 5 waves, and is zero otherwise. Apply the Verbeek and Nijman test with this definition of d_i (otherwise equal to the definition in the lecture slides). Draw conclusions and address practical problems you possibly met in implementing the test.