NYU STERN
NEW YORK UNIVERSITY LEONARD N. STERN SCHOOL OF BUSINESS

Department of Economics

Econometric Analysis of Panel Data
Assignment 1

This exercise will be based on a data set from the World Health Organization that relates to aggregate health care outcomes in world economies. (See Slide 31 in Notes 1.) The original data set is an unbalanced panel of 191 countries observed yearly from 1993 to 1997. We will use the balanced panel data for the 140 countries for which all five years of data are available. The data are stored in two forms, .csv (comma separated values) which is portable to any software and .lpj (limdep project) which is the native "save" format for limdep or nlogit. There are also two files of each type:
http://people.stern.nyu.edu/wgreene/Econometrics/WHO-balanced-panel.csv and .lpj
contains the full panel, all 140 sets of 5 observations. The smaller files
http://people.stern.nyu.edu/wgreene/Econometrics/WHO1997.csv and .lpj
contains the 140 observations from the same countries for 1997 only. This is a cross section.

The variables in the data sets aer

YEAR	1993, …, 1997
COMP, LOGCOMP	composite measure of health care attainment
DALE, LOGDALE	disability adjusted life expectancy
HLTHEXP,LOGHEXP	health expenditure
EDUC, LOGEDUC	average education
LOGHEXP2	square of LOGHEXP
LOGEDUC2	square of LOGEDUC
LOGED_EX	LOGHEXP * LOGEDUC
GINI	Gini coefficient of income distribution
TROPICS	dummy variable for tropical location
POPDEN, LOGPOPDEN	population density
PUBTHE	proportion of health expenditure that is public
GDPC,LOGGDPC	per capita income
T93,…,T97	year dummy variables
GEFF	World Bank measure of government effectiveness
VOICE	World Bank measure of extend of Democracy
OECD	dummy variable, member of UN OECD in 1997
MEANLCMP	country (5 year mean) of LOGCMP
MEANLHC,MEAHLHC2	country mean and its square of LOGEDUC
MEANLEXP	country mean of LOGHEXP

Tips: Do the computations for this exercise using any software you wish: Stata, R, NLOGIT, MatLab, Gauss, SAS, etc. You can import either csv file directly into R, Stata, NLOGIT, or most other programs. (Use Project:Import->Variables in nlogit.) You can import the lpj files into NLOGIT by using File:Open Project…) To use the cross section data, it will be inconvenient to use a subset of the panel. In nlogit, it is only necessary to precede the analysis with a command **REJECT;year#1997$**, or use the cross section data set.

Part I. Linear Regression Analysis Based on the 1997 Data Only

Data for this exercise are on the course website – please use the "Cornwell and Rupert Returns to Schooling Data." The computations can be done with Stata, NLOGIT, SAS, R, MatLab, Gauss or any other software you wish to use. We begin with the linear regression model (using the variable names in the data set),

(A) $\text{LOGCOMP}_i = \beta_1 + \beta_2\text{LOGEDUC}_i + \beta_3\text{LOGHEXP}_i + \beta_4\text{LOGED\_EX}_i + \varepsilon_i$

1. Compute the linear least squares regression results and report the coefficients, standard errors, 't-ratios,' R^2, adjusted R^2, residual standard deviation, and F statistic for testing the joint significance of all the variables in the equation.

2. Test the hypothesis that the log of education is not a significant determinant of the expected log COMP. Use an F (Wald), likelihood ratio (assuming normality of ε), and a Lagrange multiplier (also assuming normality) test. In each case, document in minute detail exactly how you are computing your results and what conclusion you reach. Note that education appears in two terms, so you are testing the joint hypothesis that β_2 and β_4 are both zero. Hints:You can find how to carry out the LM test in Notes 2 in your class notes. For the linear regression model with homoscedastic disturbances, the likelihood ratio statistic can be computed from the F statistic using

$$\text{LR} = -2(\log \text{L}_0 - \log \text{L}_1) = \text{n} \log\left[1 + \frac{\text{JF}}{\text{n} - \text{K}}\right]$$

where J = the number of restrictions and K = the number of variables in the larger model. (J and n-K are the degrees of freedom in the F statistic.) You are invited to prove this result. The inverse transformation is F = [(n-K)/J] × [exp(LR/n) – 1]. A direct manipulation of the LR statistic reveals

$$\text{LR} = \text{n} \ln(\mathbf{e}^{*\prime}\mathbf{e}^* / \mathbf{e}'\mathbf{e})$$

where $\mathbf{e}^*$ is the residuals from the restricted model. In this case, $\mathbf{e}^{*\prime}\mathbf{e}^* = \Sigma_{it}(y_{it} - \overline{y})^2$

3. The data contain a dummy variable for OECD membership. Add OECD to the regression model and reestimate. I.e., what is the economic meaning of the value you computed for this coefficient? Test the hypothesis that this coefficient equals zero.

Part II. Structural Change Based on the Panel Data Set

1. The implication of the specification of OECD in the model in Part II is that the extent of the difference between OECD and non-OECD countries is captured in a parallel shift of the regression function (based on a change in the intercept alone). Consider, instead, the hypothesis that different regression functions apply to the two groups. Fit the model separately for OECD=0 and OECD=1, then use a Chow test to test the null hypothesis that the same equation applies to both groups. (Note, for purposes of this exercise, your model will not contain the OECD variable.) The model is

(B) $\text{LOGCOMP}_{i,\text{OECD}} = \beta_1 + \beta_2\text{LOGEDUC}_{i,\text{OECD}} + \beta_3\text{LOGHEXP}_{i,\text{OECD}} + \beta_4\text{LOGED\_EX}_{i,\text{OECD}} + \varepsilon_i$

Completely document your analysis. Include in your results a table that shows the results of the three regressions, male, female and pooled, so that the reader can easily see the comparison of the estimated coefficients. What is the result of the test? (Note: This exercise is based on the cross section data.)

2. Looking ahead to our work in panel data modeling, repeat this analysis for the 5 years of data in the sample. That is, compute the regression in (B) using the full pooled data set, then again for each of the 5

years. (There are 140 observations for each of the 5 years.) Using a Chow (F) test, test the null hypothesis that the same model applies to all 5 years. (This will require you to use the panel data set.)

3. To investigate whether a structural change might be explained by a simple shift of the function, fit the model

(C) $$LOGCOMP_{it} = \beta_1 + \beta_2 LOGEDUC_{it} + \beta_3 LOGHEXP_{it} + \beta_4 LOGED\_EX_{it} + \delta_{1994} T94 + \delta_{1995} T95 + \delta_{1996} T96 + \delta_{1997} T97 + \varepsilon_{it}$$

where T94,...T97 are 4 dummy variables for the 4 years, omitting the first. Test the null hypothesis that the 4 dummy variable coefficients all equal zero and report all results. Intepret your findings. Test the hypothesis that the set of time coefficients are jointly zero. Note the pattern of the coefficients on the dummy variables. Interpret the results.

Part III. A Nonlinear Regression

1. The model (A) above contains a nonlinearity, LOGED_EX = LOGEDUC*LOGHEXP. Therefore,

(D) $$\gamma = \partial E[LOGCOMP|\mathbf{x}]/\partial LOGEDUC = \beta_2 + \beta_4 LOGHEXP$$

Compute the value of γ at the mean value of LOGHEXP. Compute an asymptotic standard error for this estimator of γ then test the "hypothesis" that θ equals zero.

2. In the original study published by the WHO, the researchers began with a "translog" specification that included LOGHEXP, LOGEDUC, LOGHEXP2, LOGEDUC2 and LOGED_EX. In the published version of the model, they dropped LOGHEXP2 and LOGED_EX. The reason for dropping the variables was not based on a test of whether the associated coefficients were zero; it was based on the shape of the quadratic function. Carry out the test based on the panel data set (which is the one they used) and report your finding.

Part IV. Instrumental Variables

Does the economy achieve greater health because it is more educated, or does a healthy economy increase education? The ambiguity in this statement suggests that in (A), it is possible that LOGEDUC (and LOGED_EX) are endogenous. If so, 2SLS would be a preferable estimator. The cross section data set contains a number of variables that could be used as suitable instrumental variables. Reestimate (A) using 2SLS. Report your results and comment on whether the result in (D) above has changed.

Part V. Partial Effects

In the regression model, E[y|x] = m(x,β), the interesting quantities are usually the partial effects, $\partial E[y|x]/\partial x$. An issue in the methodology of econometric modeling is the difference between the "partial effects at the means," and the "average partial effects." Show that these are exactly the same if the regression is linear (m(x,β) = x$'\beta$), but not the same if the function is nonlinear (such as m(x,β) = exp(x$'\beta$)). To see how the difference depends on the data, consider the APE for one variable, $APE(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(x_i, \beta)}{\partial x_i}$. Now, approximate APE(x) by using a second order Taylor series, with the expansion point $x_i^0 = \overline{x}$. What do you find?