Introduction to Stata, Data manipulation and representation, Logit, Post-estimation

Part 1: Getting Started

1. "`Indiv.csv`" contains individual-specific time-invariant information and `Employment.csv`" contains employment-specific information. Read the two datasets. Merge them using the common variable **numenq**. Save the resulting dataset in Stata format.

2. Produce some descriptive statistics of variables *wage* and *sch_lev*. Produce two-way tables of frequency counts of *man* and *sch_lev*. Are men often more educated than women in the sample?

3. Produce easy-to-read histograms of age and wage.

4. How does wage evolve with age in the sample? Produce a scatterplot of wage on age, and fit a line through it.

Note: Useful commands:

`import delimited, browse, describe, summarize, bysort:, tabulate, histogram, twoway, scatter, lfit, duplicates, label, replace, keep, local, global, forvalues, foreach`

Part 2: Data Cleaning

1. Identify and drop outliers in wage and age. Remove missing values until you obtain a dataset where no characteristic is missing - except employment variables for non-employed individual, and unemployment variables for employed individuals.
   Note: Some variables take the value '99' when missing. Stata only understands '.' as a missing value. Replace those '99' by '.' when necessary.

Part 3: Explaining the Employment Probability

We would like to see what explains the fact that an individual is employed. Use gender, mother's and father's origin, school entry age and the schooling level as explanatory variables.

1. Create the necessary variables to conduct the estimation (dummy variables, interaction terms).
   Hint: Some variables from Part 1 and 2 can be reused.
   Note: Never include all dummy values of a covariate in an regression with a constant. Why?
   In a regression, the constant β_0 is in fact the coefficient corresponding to a variable taking the value 1 all the time: $\beta_0 = \beta_0 \times 1$.
   Let's take the variable "man". If you create dummies $\mathbb{1}\{man = 0\}$ and $\mathbb{1}\{man = 1\}$, the sum of those will be 1 for every observation. This means it is perfectly collinear with the variable to which the constant corresponds (the never-changing 1). Therefore, we need to drop one dummy, so we will only include $\mathbb{1}\{man = 1\}$ in our regressions.

2. Run a pooled linear regression. Compute heterosckedasticity-robust standard errors.

   ### Linear Probability Model (LPM)

   $$\begin{aligned} Employed_i = \beta_0 + \beta_1 Gender_i + \beta_2 FiF_i + \beta_3 MiF_i + \beta_4 FMaF_i + \beta_5 SchEntryAge_i \\ + \sum_{k \in Levels \setminus \{\text{No qualification}\}} \gamma_k \mathbb{1}(SchLevel_i = k) + \epsilon_i \end{aligned}$$

   with $Levels = \{\text{No qualification}, \text{Vocational high school}, \text{High school}, ...\}$ and $E[\epsilon_i | X_i] = 0$.

3. Estimate a logit model.

   The Logit model assumes errors follow a logistic distribution:

   $$\begin{aligned} P\left(Emp_i = 1 \mid X_i\right) = F(\beta X) = F\Big(\beta_0 + \beta_1 Gender_i + \beta_2 FiF_i + \beta_3 MiF_i + \beta_4 FMaF_i \\ + \beta_5 SchEntryAge_i + \sum_k \gamma_k \mathbb{1}(SchLevel_i = k)\Big) \end{aligned}$$

   where F(.) is the cdf of the logistic distribution.

4. Interpret the parameter estimates and calculate the marginal effects. How do we interpret marginal effects?

   We estimate the model by maximum likelihood (MLE). The estimation table gives us coefficient estimates. We can comment on their sign and significance, but not on their magnitude. Indeed, in order to say something like "having a father born in France increases the probability of employment by x", we need $\frac{\partial \widehat{Employed_i}}{\partial X_i}$. In the linear probability model, this is exactly equal to β_2, but not in the Logit model. The *margins* command computes this derivative for us and gives us interpretable results.

5. What are the differences between a linear model and a logit model for binary variables?

   Disadvantages of the LPM:

   - The errors are always heteroskedastic. Let Y_i be a binary variable. The LPM writes:
   $$Y_i = \alpha + \beta X_i + \varepsilon_i$$
   $$\Rightarrow \varepsilon_i = \begin{cases} 1 - \alpha - \beta X_i & \text{if } Y_i = 1 \\ -\alpha - \beta X_i & \text{if } Y_i = 0 \end{cases}$$
   So ε always depends on X, which is the definition of heteroskedasticity.
   - Fitted values resulting from LPM are not necessarily between 0 and 1. Let $\hat{Y}_i$ be the fitted value of our binary dependent variable.
   $$\hat{Y}_i = \widehat{\mathbb{P}[Y_i = 1]} = \hat{\alpha} + \hat{\beta} X_i$$
   Although this is a probability, it is not constrained to stay between 0 and 1. Therefore, we cannot do prediction with the LPM. However, this does not affect the interpretation of coefficients.

   Advantages of the LPM:

   - Easy to interpret (no need for the *margins* command, which is appreciated when there are many variables).
   - Faster to run on a large dataset.
   - Much easier to control for endogeneity with an instrumental variable (you can do 2SLS, whereas a maximum-likelihood estimation requires much more complicated tools).

6. Compute the predicted probabilities, the predicted residuals and plot one against each other. Do you observe a pathological structure of the residuals?

   Residuals are computed by doing $\hat{\epsilon}_i = Employed_i - \mathbb{P}[\widehat{Employed_i} = 1]$. Given $Employed_i \in \{0, 1\}$, residuals will be split in 2 groups, the positive ones (when $Employed_i = 0$) and the negative ones (when $Employed_i = 1$). This makes the plot of residuals against fitted values hard to read.
   The command *predict residPearson, residuals* creates a variable called "residPearson" which corresponds to the residuals of the model for each observations, weighted by a term which smoothes the value of residuals within observations sharing a covariate pattern (= with the same values of *man*, *FiF*, *MiF* etc.). This *residualPearson* will no longer be divided in 2 groups, and we can plot it against fitted values.

   When making such a residuals-vs-fitted plot, you want to look at the following:

- Are residuals well centered around 0? If you see that residuals value is correlated with the fitted value, it means that the model is misspecified. That can mean for instance that a key variable is missing. You can plot the residuals against some variables you did not include in your regression. If there seems to be a correlation, it can be a good idea to try adding this variable to your regression.
- Do you see observations with very high and/or very low levels of the residuals? Those could be outliers.
- Does it seem that the variance of the residuals is higher around some values of $\mathbb{P}[\widehat{Employed_i} = 1]$? That could indicate heteroskedasticity.

Appendix

Appendix 1: "Génération 98": Description of the Database

"Enquête Génération 98", provides detailed information on the socio-demographic background and labor market characteristics of young individuals who left school in one specific year (1998). Interviews are conducted three years after the school-exit date. The aim of the Enquête Génération is to document many aspects of early labor market transitions (spells of employment, unemployment, and training) experienced between school completion (labor market entrance) and the date of the survey. The database is an extract from the original data where labor market information has been extracted at 3 periods: In April 1999 (month=16), in April 2000 (month=28) and in April 2001 (month=40). The data are organized in the usual convention with panel data: 1 row corresponds to 1 period observation per individual. Also, the data are split in two datasets: "`Indiv.csv`" contains individual (time unvarying) information, while "`Eployment.csv`" contains the employment-specific information.

Variables included in the dataset are the following:

`Indiv.csv`: Individual (time unvarying) characteristics:

- numenq: individual identifier
- man: indicator for male
- age98: age in 1998 country
- country_birth: country of birth origin (France = 0)
- origin_father: country of birth of the father origin (France = 0)
- origin_mother: country of birth of the mother (France = 0)
- region_school: region of the school in 1998
- location: location at the end of school

  1 = Urban;
  2 = Rural;
  3 = Overseas French territories;
  4 = Abroad;
  99 = Unknown

- urban_area: urban area at the end of school
- sch_lev: highest schooling level at the end of school (in 1998)

  1 = No qualification;
  2 = Vocational sigh school degree (CAP, BEP);
  3 = High school degree (baccalaureat);
  4 = Some higher education (without graduating);
  5 = 2 years higher education degree (baccalaureat and 2 years);
  6 = Intermediate university degree (baccalaureat and 3 or 4 years);
  7 = Advanced university degree (DEA, DESS, doctorat), elite business or engineering school degree

- age_jun_hs: age at entry into junior high school (6th grade)

  9 = 9 years old (2 years in advance on the normal age);
  10 = 10 years old (1 year in advance on the normal age);
  11 = 11 years old (“normal" age, without grade repetition);
  12 = 12 years old (1 year after the normal age);
  13 = 13 years old (2 years after the normal age);
  14 = 14 years old and more (3 years and more after the normal age)

- occup_father/occup_mother: Father’s /Mother’s last occupation in 1998

  1 = Farmer;
  2 = Craftsman, tradesman, company director;
  3 = Senior executive, engineer, teacher;
  4 = Technical, middle manager;
  5 = White collar;
  6 = Blue collar;
  7 = Inactive;
  99 = Unknown

`Employment.csv`: Labor market characteristics

- numenq: individual identifier
- month: month of the interview

  16 = April 1999;
  28 = April 2000;
  40 = April 2001.

- stat: labor market status

  1 = Employment;
  2 = Unemployment;
  3 = Out of the labor force;
  4 = Training;
  5 = Schooling;
  6 = National service

Employment characteristics for employed individuals (i.e. who have stat=1)

- tenure: on the job tenure (in months)
- firm_loc: Location of the firm

  1 = Urban;
  2 = Rural;
  3 = Overseas French territories;
  4 = Abroad;
  99 = Unknown

- firm_status: Status of the firm

  1 = Public sector;
  2 = Private sector;
  99 = Unknown;

- firm_sector: Economic sector of the firm

  1 = Car industry;
  2 = Non-market services;
  3 = Market services;
  4 = Construction;
  5 = Other industry;
  6 = Hotel, restaurant;
  7 = Other;
  99 = Unknown

- firm_size: number of employees in the firm

  1 = 0 employee;
  2 = 1 to 2 employees;
  3 = 3 to 9 employees;
  4 = 10 to 49 employees;
  5 = 50 to 199 employees;
  6 = 200 to 499 employees;
  7 = 500 employees and more;
  99 = Unknown

- contract: Type of employment contract

  1 = Permanent contract (CDI, fonctionnaire);
  2 = Fixed Term Contract (CDD, temporary work);
  3 = Publicly subsidized contract;
  4 = Other

- occup: Job occupation

  1 = Farmer;
  2 = Craftsman, tradesman, company director;
  3 = Senior executive, engineer, teacher;
  4 = Technical, middle manager;
  5 = White collar;
  6 = Blue collar;
  99 = Unknown

- work_time: Working time

  1 = Full-time employment;
  2 = Part-time employment;
  3 = Unknown

- wage: Monthly wage (including bonuses), in euros

Characteristics for non-employed individuals

- tenure_unemp: tenure in the unemployment spell (in months)