

Notes on Lab Session 1

Giulia Tani

<https://tanigiulia.github.io/>

TSE - MSc course in Program Evaluation

January 2026

Linear Regression Model

- We want to understand the relationship between a dependent variable y (e.g. employment) and regressors/covariates x (e.g. worker characteristics).
- We can **always** write:

$$y = E(y \mid x) + \epsilon, \quad E(\epsilon \mid x) = 0,$$

where the **conditional mean** $E(y \mid x)$ is the best predictor of y given x among all possible functions of x .

- A regression **model** imposes **restrictions** on $E(y \mid x)$. The **linear regression model** assumes

$$E(y \mid x) = x' \beta,$$

so that

$$y = x' \beta + \epsilon, \quad E(\epsilon \mid x) = 0.$$

- In a sample, we estimate β by OLS:

$$\hat{\beta} = \arg \min_b \sum_{i=1}^n (y_i - x_i' b)^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right).$$

Linear Probability Model

- The **linear probability model** (LPM) is the linear multiple regression model applied to a **binary dependent variable** $y \in \{0, 1\}$:

$$E(y \mid x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

- Because y is binary, $E(y \mid x_1, \dots, x_k) = \text{Pr}(y = 1 \mid x_1, \dots, x_k)$. Then, for the LPM:

$$\text{Pr}(y = 1 \mid x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

- The coefficient β_1 gives the **marginal effect** of x_1 on y , that is, the difference in the probability that $y = 1$ associated with a unit difference in x_1 , holding constant the other regressors:

$$\frac{\partial \text{Pr}(y = 1 \mid x_1, \dots, x_k)}{\partial x_1} = \beta_1.$$

- Regression coefficients can be estimated with OLS.

Disadvantages of the LPM

1) Heteroskedasticity

The conditional variance of the errors is:

$$\text{Var}(\epsilon | x) = \text{Var}(y - x'\beta | x) = \text{Var}(y | x).$$

Since y is binary,

$$\text{Var}(y | x) = \text{Pr}(y = 1 | x) (1 - \text{Pr}(y = 1 | x)).$$

Hence, under the LPM:

$$\text{Var}(\epsilon | x) = x'\beta (1 - x'\beta).$$

The variance of ϵ changes depending on x .

2) Predicted probabilities outside [0,1] interval

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

The effect on \hat{y} of a unit change in x_1 is constant ($\hat{\beta}_1$). For low x_1 , predicted probability \hat{y}_i can drop below 0; for high x_1 , it can be greater than 1.

Logit Model

- The logit model is a **nonlinear regression model** specifically designed for a binary y .
- As before, because y is binary, $E(y \mid x_1, \dots, x_k) = Pr(y = 1 \mid x_1, \dots, x_k)$. But now we impose:

$$Pr(y = 1 \mid x_1, \dots, x_k) = F(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k),$$

where $F(\cdot)$ is the cdf of the logistic distribution:

$$Pr(y = 1 \mid x_1, \dots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}.$$

Hence **predicted probabilities are within [0,1] interval**.

- Coefficients β are estimated by maximum likelihood (MLE):

$$\hat{\beta} = \arg \max_{\beta} \log(L(\beta)), \quad \text{where } L(\beta) = \prod_{i=1}^n [Pr(y_i = 1 \mid x_i)^{y_i} (1 - Pr(y_i = 1 \mid x_i))^{(1-y_i)}]$$

- To get the **marginal effect** of x_1 on y :
 - 1) compute \hat{y} at initial value of x_1 using estimated $\hat{\beta}$
 - 2) compute \hat{y}' at changed value $x_1 + 1$
 - 3) compute the difference: $\hat{y}' - \hat{y}$.

Disadvantages of the Logit Model

- Coefficients are harder to interpret.
- Estimation uses iterative MLE (no closed-form like OLS), so it can be slower on large datasets.
- Standard linear IV/2SLS does not apply; handling endogeneity requires nonlinear IV methods.

- **To download Stata:** <https://intranet.ut-capitole.fr/outils-numeriques/outils-pedagogiques/stata-licence-etudiant-enseignants-personnels>.
- **To set a directory:**
 - 1) Download all files from Moodle and save them in the same folder (e.g. "Lab1").
 - 2) Get the folder path (e.g. "my_path/Lab1":
 - **Windows:** press Shift + right-click the folder and select "Copy as path".
 - **Mac:** drag and drop the folder into Terminal, then copy the path.
 - 3) In the Stata do-file, type cd and paste the path in quotation marks:

```
cd "my_path/Lab1"
```

- **To get information on Stata commands,** use help followed by the command name:

```
help generate
```