Economic History

Week 10: Essential Tools in Cliometrics – Linear Regression

Ricardo Gouveia-Mendes
rgouveiamendes@ucp.pt

1st Semester 2023-24

Católica-Lisboon School of Business and Economics

Basic Concepts

The Simplest Relationship Between Two Variables

- Suppose you want to study the impact **of one extra euro spent in TV advertising** tv **on sales** s. What do you need to do?
- To collect data! For each month $i = 1, \ldots, n$ you would record the expenditure in TV ads tv_i and the sales level s_i
- But how do these variables relate to each other?
- The simplest quantitative relationship we can establish is **linear**:

$$s_i = \beta_0 + \beta_1 tv_i + \varepsilon_i$$

The Simplest Relationship Between Two Variables

- Suppose you want to study the impact **of one extra euro spent in TV advertising** tv **on sales** s. What do you need to do?
- To collect data! For each month $i = 1, \ldots, n$ you would record the expenditure in TV ads tv_i and the sales level s_i
- But how do these variables relate to each other?
- The simplest quantitative relationship we can establish is **linear**:

$$s_i = \beta_0 + \beta_1 tv_i + \varepsilon_i$$

The Simplest Relationship Between Two Variables

- Suppose you want to study the impact **of one extra euro spent in TV advertising** tv **on sales** s. What do you need to do?
- To collect data! For each month $i = 1, \ldots, n$ you would record the expenditure in TV ads tv_i and the sales level s_i
- But how do these variables relate to each other?
- The simplest quantitative relationship we can establish is **linear**:

$$s_i = \beta_0 + \beta_1 tv_i + \varepsilon_i$$

The Simplest Relationship Between Two Variables

- Suppose you want to study the impact **of one extra euro spent in TV advertising tv on sales s.** What do you need to do?
- To collect data! For each month $i = 1, \ldots, n$ you would record the expenditure in TV ads tv_i and the sales level s_i
- But how do these variables relate to each other?
- The simplest quantitative relationship we can establish is **linear**:

$$s_i = \beta_0 + \beta_1 tv_i + \varepsilon_i$$

The Simplest Relationship Between Two Variables

- Suppose you want to study the impact **of one extra euro spent in TV advertising** tv **on sales** s. What do you need to do?
- To collect data! For each month $i = 1, \ldots, n$ you would record the expenditure in TV ads tv_i and the sales level s_i
- But how do these variables relate to each other?
- The simplest quantitative relationship we can establish is **linear**:

$$s_i = \beta_0 + \beta_1 tv_i + \varepsilon_i$$

The Ordinary Least Squares Estimators

$$s_i = \beta_0 + \beta_1 tv_i + \varepsilon_i$$

- **How to estimate the β's?**
- Solving the following problem:

$$\min_{\{\hat\beta_0, \hat\beta_1\}} \sum_{i=1}^{n} \hat\varepsilon_i^2 = \sum_{i=1}^{n} (s_i - \hat s_i)^2 = \sum_{i=1}^{n} \left(s_i - (\hat\beta_0 + \hat\beta_1 tv_i)\right)^2$$

- The solutions are the **OLS estimators**:

$$\hat\beta_0 = \overline{s} - \hat\beta_1 \overline{tv} \qquad \text{and} \qquad \hat\beta_1 = \frac{\sum_{i=1}^{n} \left(tv_i - \overline{tv}\right)\left(s_i - \overline{s}\right)}{\sum_{i=1}^{n} \left(tv_i - \overline{tv}\right)^2} = \rho_{s,tv} \frac{\sigma_s}{\sigma_{tv}}$$

The Ordinary Least Squares Estimators

$$s_i = \beta_0 + \beta_1 tv_i + \varepsilon_i$$

- **How to estimate the β's?**
- Solving the following problem:

$$\min_{\{\hat{\beta}_0, \hat{\beta}_1\}} \quad \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (s_i - \hat{s}_i)^2 = \sum_{i=1}^{n} \left(s_i - (\hat{\beta}_0 + \hat{\beta}_1 tv_i)\right)^2$$

- The solutions are the **OLS estimators**:

$$\hat{\beta}_0 = \overline{s} - \hat{\beta}_1 \overline{tv} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(tv_i - \overline{tv}\right)\left(s_i - \overline{s}\right)}{\sum_{i=1}^{n} \left(tv_i - \overline{tv}\right)^2} = \rho_{s,tv} \frac{\sigma_s}{\sigma_{tv}}$$

The Ordinary Least Squares Estimators

$$s_i = \beta_0 + \beta_1 tv_i + \varepsilon_i$$

- **How to estimate the β's?**
- Solving the following problem:

$$\min_{\{\hat{\beta}_0,\hat{\beta}_1\}} \quad \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (s_i - \hat{s}_i)^2 = \sum_{i=1}^{n} \Big(s_i - (\hat{\beta}_0 + \hat{\beta}_1 tv_i)\Big)^2$$

- The solutions are the **OLS estimators**:

$$\hat{\beta}_0 = \overline{s} - \hat{\beta}_1 \overline{tv} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(tv_i - \overline{tv}\right)\left(s_i - \overline{s}\right)}{\sum_{i=1}^{n} \left(tv_i - \overline{tv}\right)^2} = \rho_{s,tv} \frac{\sigma_s}{\sigma_{tv}}$$

Convenient Properties ... Under Some Underlying Assumptions

The OLS estimators are...

1. **Best**: they have the minimum variance σ^2_{β} among all the unbiased linear estimators (a.k.a. *efficient*)
2. **Linear**: they are linear functions of the observed values
3. **Unbiased**: $\mathbb{E}\left[\hat{\beta}|tv\right] = \beta$

... if and only if:

1. Linearity
2. Random Sampling
3. No perfect colinearity
4. $\mathbb{E}\left[\varepsilon|tv\right] = 0$
5. Homoskedasticity: $\mathrm{Var}\left[\varepsilon|tv\right] = \sigma^2$

Convenient Properties ... Under Some Underlying Assumptions

The OLS estimators are...

1. **Best**: they have the minimum variance σ_β^2 among all the unbiased linear estimators (a.k.a. *efficient*)
2. **Linear**: they are linear functions of the observed values
3. **Unbiased**: $\mathbb{E}\left[\hat{\beta}|tv\right] = \beta$

... if and only if:

1. Linearity
2. Random Sampling
3. No perfect colinearity
4. $\mathbb{E}\left[\varepsilon|tv\right] = 0$
5. Homoskedasticity: $\mathrm{Var}\left[\varepsilon|tv\right] = \sigma^2$

Convenient Properties ... Under Some Underlying Assumptions

The OLS estimators are...

1. **Best**: they have the minimum variance σ^2_{β} among all the unbiased linear estimators (a.k.a. *efficient*)
2. **Linear**: they are linear functions of the observed values
3. **Unbiased**: $\mathbb{E}\left[\hat{\beta}|tv\right] = \beta$

... if and only if:

1. Linearity
2. Random Sampling
3. No perfect colinearity
4. $\mathbb{E}\left[\varepsilon|tv\right] = 0$
5. Homoskedasticity: $\mathrm{Var}\left[\varepsilon|tv\right] = \sigma^2$

Convenient Properties ... Under Some Underlying Assumptions

The OLS estimators are...

1. **Best**: they have the minimum variance σ^2_β among all the unbiased linear estimators (a.k.a. *efficient*)
2. **Linear**: they are linear functions of the observed values
3. **Unbiased**: $\mathbb{E}\left[\hat{\beta}|tv\right] = \beta$

... if and only if:

1. Linearity
2. Random Sampling
3. No perfect colinearity
4. $\mathbb{E}\left[\varepsilon|tv\right] = 0$
5. Homoskedasticity: $\mathrm{Var}\left[\varepsilon|tv\right] = \sigma^2$

Convenient Properties ... Under Some Underlying Assumptions

The OLS estimators are...

1. **Best**: they have the minimum variance σ^2_β among all the unbiased linear estimators (a.k.a. *efficient*)
2. **Linear**: they are linear functions of the observed values
3. **Unbiased**: $\mathbb{E}\left[\hat{\beta}|tv\right] = \beta$

... if and only if:

1. Linearity
2. Random Sampling
3. No perfect colinearity
4. $\mathbb{E}\left[\varepsilon|tv\right] = 0$
5. Homoskedasticity: $\mathrm{Var}\left[\varepsilon|tv\right] = \sigma^2$

Convenient Properties ... Under Some Underlying Assumptions

The OLS estimators are...

1. **Best**: they have the minimum variance σ^2_{β} among all the unbiased linear estimators (a.k.a. *efficient*)
2. **Linear**: they are linear functions of the observed values
3. **Unbiased**: $\mathbb{E}\left[\hat{\beta}|tv\right] = \beta$

... if and only if:

1. Linearity
2. Random Sampling
3. No perfect colinearity
4. $\mathbb{E}\left[\varepsilon|tv\right] = 0$
5. Homoskedasticity: $\mathrm{Var}\left[\varepsilon|tv\right] = \sigma^2$

Convenient Properties ... Under Some Underlying Assumptions

The OLS estimators are...

1. **Best**: they have the minimum variance σ_{β}^2 among all the unbiased linear estimators (a.k.a. *efficient*)
2. **Linear**: they are linear functions of the observed values
3. **Unbiased**: $\mathbb{E}\left[\hat{\beta}|tv\right] = \beta$

... if and only if:

1. Linearity
2. Random Sampling
3. No perfect colinearity
4. $\mathbb{E}\left[\varepsilon|tv\right] = 0$
5. Homoskedasticity: $\mathrm{Var}\left[\varepsilon|tv\right] = \sigma^2$

Convenient Properties ... Under Some Underlying Assumptions

The OLS estimators are...

1. **Best**: they have the minimum variance σ^2_{β} among all the unbiased linear estimators (a.k.a. *efficient*)
2. **Linear**: they are linear functions of the observed values
3. **Unbiased**: $\mathbb{E}\left[\hat{\beta}|tv\right] = \beta$

... if and only if:

1. Linearity
2. Random Sampling
3. No perfect colinearity
4. $\mathbb{E}\left[\varepsilon|tv\right] = 0$
5. Homoskedasticity: $\mathrm{Var}\left[\varepsilon|tv\right] = \sigma^2$

Assessing a Linear Regression

The Statistical Significance of the Estimates

- Our **estimators are random numbers** — different samples lead to different estimates
- Can we **rule out the hypothesis that our variable selection is useless**? Can we reject: $H_0 : \beta_1 = 0$?
- If $\beta_1 = 0$, how likely would it be to get an estimate as the one we've got?
  - We need to choose a **level of significance** α (usually $\alpha > 5$)
  - Check the p**-value** (the cumulative probability of getting a lower value to β_1 lower than $\hat{\beta}_1$): reject if p-value $< \alpha$

The Statistical Significance of the Estimates

- Our **estimators are random numbers** — different samples lead to different estimates
- Can we **rule out the hypothesis that our variable selection is useless**? Can we reject: $H_0 : \beta_1 = 0$?
- If $\beta_1 = 0$, how likely would it be to get an estimate as the one we've got?
  - We need to choose a **level of significance** α (usually $\alpha > 5$)
  - Check the p**-value** (the cumulative probability of getting a lower value to β_1 lower than $\hat{\beta}_1$): reject if p-value $< \alpha$

The Statistical Significance of the Estimates

- Our **estimators are random numbers** – different samples lead to different estimates
- Can we **rule out the hypothesis that our variable selection is useless**? Can we reject: $H_0 : \beta_1 = 0$?
- If $\beta_1 = 0$, how likely would it be to get an estimate as the one we've got?
  - We need to choose a **level of significance** α (usually $\alpha > 5$)
  - Check the p**-value** (the cumulative probability of getting a lower value to β_1 lower than $\hat{\beta}_1$): reject if p-value $< \alpha$

The Statistical Significance of the Estimates

- Our **estimators are random numbers** — different samples lead to different estimates
- Can we **rule out the hypothesis that our variable selection is useless**? Can we reject: $H_0 : \beta_1 = 0$?
- If $\beta_1 = 0$, how likely would it be to get an estimate as the one we've got?
  - We need to choose a **level of significance** α (usually $\alpha > 5$)
  - Check the p**-value** (the cumulative probability of getting a lower value to β_1 lower than $\hat{\beta}_1$): reject if p-value $< \alpha$

The Statistical Significance of the Estimates

- Our **estimators are random numbers** – different samples lead to different estimates
- Can we **rule out the hypothesis that our variable selection is useless**? Can we reject: $H_0 : \beta_1 = 0$?
- If $\beta_1 = 0$, how likely would it be to get an estimate as the one we've got?
  - We need to choose a **level of significance** α (usually $\alpha > 5$)
  - Check the p**-value** (the cumulative probability of getting a lower value to β_1 lower than $\hat{\beta}_1$): reject if p-value $< \alpha$

The Quality of the Overall Fit

How good a linear regression is depends on **how well the model describes the data** (how close the points are to the line).

The measure of the overall fit is given by the **Coefficient of Determination**:

$$R^2 = \frac{ESS}{TSS}$$

Total Sum of Squares:

$$TSS = \sum_{i=1}^{n} (s_i - \overline{s})^2$$

Explained Sum of Squares:

$$ESS = \sum_{i=1}^{n} (\hat{s}_i - \overline{s})^2$$

The Quality of the Overall Fit

How good a linear regression is depends on **how well the model describes the data** (how close the points are to the line).

The measure of the overall fit is given by the **Coefficient of Determination**:

$$R^2 = \frac{ESS}{TSS}$$

Total Sum of Squares:

$$TSS = \sum_{i=1}^{n} (s_i - \overline{s})^2$$

Explained Sum of Squares:

$$ESS = \sum_{i=1}^{n} (\hat{s}_i - \overline{s})^2$$

The Quality of the Overall Fit

How good a linear regression is depends on **how well the model describes the data** (how close the points are to the line).
The measure of the overall fit is given by the **Coefficient of Determination**:

$$R^2 = \frac{ESS}{TSS}$$

Total Sum of Squares:

$$TSS = \sum_{i=1}^{n} (s_i - \overline{s})^2$$

Explained Sum of Squares:

$$ESS = \sum_{i=1}^{n} (\hat{s}_i - \overline{s})^2$$

Generalizing OLS

The Multiple Linear Regression

- We may want to consider **more than one explanatory variables**:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \varepsilon_i$$

- The matricial representation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \dots & x_{k,n} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The Multiple Linear Regression

- We may want to consider **more than one explanatory variables**:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \varepsilon_i$$

- The matricial representation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \dots & x_{k,n} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The Multiple Linear Regression

- We may want to consider **more than one explanatory variables**:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \varepsilon_i$$

- The matricial representation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \dots & x_{k,n} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The Multiple Linear Regression

- The OLS estimators are, then, the solution of the problem:

$$\min_{\hat{\boldsymbol{\beta}}} \quad \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$$

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

- The solution:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$