

Econometrics

Practical Session 13

Midterm Solution



CATÓLICA
LISBON
BUSINESS & ECONOMICS

Ricardo Gouveia-Mendes
rgouveiamendes@ucp.pt

Spring 2025-26

Católica-Lisbon School of Business and Economics

- Variable transformation: $\text{crim}_* = 10 \times \text{crim}$
 - Affects coefficient estimate $\hat{\beta}_{\text{crim}}$ and SE $\hat{\sigma}_{\hat{\beta}_{\text{crim}}}$
 - Sample size does not change...
- Writing abstract models with $-\beta_j$
- Omitted variable bias:

$$\log(\text{medv}) = \beta_0 + \beta_{\text{crim}} \text{crim} + \dots + u$$

$$\hat{\beta}_{\text{crim}} \xrightarrow{p} \beta_{\text{crim}} + \rho_{\text{crim},u} \frac{\sigma_{\text{crim}}}{\sigma_u}$$

$$u = \beta_{\text{nox}} \text{nox} + v, \quad \beta_{\text{nox}} < 0 \quad \Rightarrow \quad \rho_{\text{crim},u} < 0$$

Strict Exogeneity ✓ OLS unbiased

$$\mathbb{E}[u_i \mid X] = 0$$

- About the **mean** of u
- Implies $\text{Cov}(X_j, u) \equiv \mathbb{E}[X_j \cdot u] - \mathbb{E}[X_j] \cdot \mathbb{E}[u] = 0$ for all j
- Violated by: omitted variables, reverse causality
- **Consequence:** $\hat{\beta} \nrightarrow \beta$, OLS is **biased**

Homoskedasticity ✓ OLS efficient

$$\text{Var}(u_i \mid X) = \sigma^2$$

- About the **variance** of u
- Errors have **constant spread** across X
- Violated by: variance growing with X
- **Consequence:** $\hat{\beta}$ still unbiased, but **inefficient** and wrong SEs

Question 1

The following table presents OLS estimates from a regression of log crime rates on county-level characteristics in the US, using data from North Carolina ($N = 630$ county-year observations):

Call: `lm(formula = log(crmrte) ~ unem + I(unem^2) + prbarr + urban)`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.4826	0.2153	-6.89	<0.001	***
unem	0.1200	0.0214	5.61	<0.001	***
I(unem^2)	-0.0060	0.0011	-5.45	<0.001	***
prbarr	-1.4253	0.1882	-7.57	<0.001	***
urban	0.4861	0.0507	9.59	<0.001	***

Residual standard error: 0.518 on 625 degrees of freedom

Multiple R-squared: 0.4126, Adjusted R-squared: 0.4087

F-statistic: 109.7 on 4 and 625 DF, p-value: < 0.001

Question 1

where $crmrte$ is the number of crimes per 1000 residents, $unem$ is the county unemployment rate (in percent), $prbarr$ is the probability of arrest (the proportion of crimes that lead to an arrest, ranging from 0 to 1), and $urban$ is a binary indicator equal to 1 for metropolitan counties.

a) Interpret the coefficient on $unem$ in precise, quantitative terms. A classmate argues: "Since unemployment is positive and highly significant, unemployment causes crime to rise." Do you agree with this causal claim? Briefly justify your answer.

- The coefficient on $unem$ is a semi-elasticity
- A 1p.p. increase in the unemployment rate is associated with $100(e^{0.12} - 1) = 12.75\%$ **increase in the crime rate**

Question 1

- The causal claim is **not warranted**:
 - **Reverse causality**: High crime may itself reduce economic activity, raising unemployment (simultaneity bias).
 - **Omitted variables**: Poverty, racial composition, and policing intensity all affect both crime and unemployment, violating exogeneity $\mathbb{E}[u|x] = 0$.

Question 1

b) Write the expression for the marginal effect of $unem$ on log crime rates. At what unemployment rate is predicted log crime maximised? Does this turning point make economic sense? Briefly discuss.

- Differentiating with respect to $unem$:

$$\frac{\partial \log(crmrte)}{\partial unem} = \hat{\beta}_{unem} + 2\hat{\beta}_{unem^2} \cdot unem = 0.1200 - 0.0120 \cdot unem$$

- Setting equal to zero:

$$unem^* = \frac{0.1200}{0.0120} = \mathbf{10.0\%}$$

- **Economic sense:** at very high unemployment, most *marginal* crimes have already been committed. The turning point at 10% is broadly plausible.

Question 1

c) Southbrook is a metropolitan county with an unemployment rate of 8% and a probability of arrest of 0.25. Hilldale is a rural county with an unemployment rate of 4% and a probability of arrest of 0.30. Compute the predicted log crime rate for each county. Express the difference as a percentage.

- **Southbrook:**

$$\hat{y}_S = -1.4826 + 0.1200(8) - 0.0060(64) - 1.4253(0.25) + 0.4861 = -\mathbf{0.7768}$$

- **Hilldale:**

$$\hat{y}_H = -1.4826 + 0.1200(4) - 0.0060(16) - 1.4253(0.30) + 0 = -\mathbf{1.5262}$$

Question 1

- Difference: $\hat{y}_S - \hat{y}_H \approx 0.75$
- $100(e^{0.75} - 1) \approx 112\%$: Southbrook's predicted crime rate is **$\approx 112\%$ higher** than Hilledale's

Question 1

d) Construct a 95% confidence interval for the coefficient on urban. Interpret the interval — both statistically and economically.

$$CI_{\beta_{\text{urban}}}^{95\%} = \{\hat{\beta}_{\text{urban}} \pm Z_{97.5\%} \times s_{\hat{\beta}_{\text{urban}}}\} = \{0.4861 \pm 1.96 \times 0.0507\} = [\mathbf{0.387}, \mathbf{0.585}]$$

- **Statistical:** 0 outside the interval → reject $H_0 : \beta_{\text{urban}} = 0$ at 5%
- **Economic:** $100(e^{0.387} - 1) \approx 47.2\%$ and $100(e^{0.585} - 1) \approx 79.5\%$ → metropolitan counties have between **47.2% and 79.5% higher** crime rates than rural counties, holding everything else constant

Question 1

e) You wish to test $H_0 : 3\beta_{\text{unem}} = 2\beta_{\text{urban}}$ using only a t -test (no F -test). Show explicitly how to reparametrize the model so that this hypothesis can be tested as a single coefficient equals zero. State clearly the new regression you would run and the new null hypothesis.

- Define $\theta \equiv 3\beta_{\text{unem}} - 2\beta_{\text{urban}}$
- Null hypothesis: $H_0 : \theta = 0$
- Solve: $\beta_{\text{unem}} = (\theta + 2\beta_{\text{urban}})/3$ and substitute into the model:

$$\begin{aligned} \log(\text{crmrte}) = & \beta_0 + \theta \cdot \frac{\text{unem}}{3} + \beta_{\text{urban}} \left(\frac{2}{3} \text{unem} + \text{urban} \right) + \\ & + \beta_{\text{unem}^2} \text{unem}^2 + \beta_{\text{prbarr}} \text{prbarr} + u \end{aligned}$$

Question 1

f) Suppose the true model also includes $povrate$ — the county poverty rate — but you omit it from the regression.

- Explain why omitting $povrate$ is likely to bias $\hat{\beta}_{unem}$.*
- In which direction is the bias? Justify your answer.*

- OVB formula:

$$\hat{\beta}_{unem} \xrightarrow{p} \beta_{unem} + \rho_{unem,u} \frac{\sigma_{unem}}{\sigma_u}$$

$$u = \beta_{prate} prate + v, \quad \beta_{prate} > 0 \quad \Rightarrow \quad \rho_{unem,u} > 0$$

- **Bias is positive:** $\hat{\beta}_{unem}$ **overstates** the causal effect of unemployment on crime

Question 1

g) The F -statistic reported above equals 109.7. Write the null hypothesis it tests. What can you conclude from the fact that it is highly significant? Is rejecting the joint null sufficient to conclude that each individual regressor matters?

- F -statistic tests the **joint null**:

$$H_0 : \beta_{\text{unem}} = \beta_{\text{unem}^2} = \beta_{\text{prbarr}} = \beta_{\text{urban}} = 0$$

- $F = 109.7$ ($p < 0.001$): strongly reject → **at least one** regressor is significantly related to crime
- **Not sufficient** to conclude each regressor individually matters → individual t -tests

Question 2

A researcher estimates a hedonic pricing model for residential housing in the Lisbon metropolitan area (AML — Área Metropolitana de Lisboa, $N = 508$):

```
Call: lm(formula = log(medv) ~ rooms + log(dist) + crim + stratio)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.4961	0.4892	11.23	<0.001	***
rooms	0.3067	0.0164	18.70	<0.001	***
log(dist)	-0.1913	0.0467	-4.10	<0.001	***
crim	-0.0120	0.0013	-9.21	<0.001	***
stratio	-0.0519	0.0091	-5.70	<0.001	***

Residual standard error: 0.2652 on 503 degrees of freedom

Multiple R-squared: 0.6486, Adjusted R-squared: 0.6458

Question 2

where *medv* is median home value (in €1000s), *rooms* is the average number of rooms per dwelling, *dist* is the weighted distance to Lisbon employment centers (in km), *crim* is the per-capita crime rate (per 100 residents), and *stratio* is the pupil-to-teacher ratio.

a) Interpret the coefficient on $\log(\text{dist})$ precisely. What functional form relationship does including *dist* in logs — rather than levels — imply between distance and house prices?

- Log-log specification: coefficient is a **price elasticity w.r.t. distance**
- “A 1% increase in distance from Lisbon employment centres is associated with a **0.19% decrease** in median house prices, all else equal”

Question 2

- Contrast with level specification:
 - Constant euro effect per km
 - log-log is more realistic: the first km from employment matters more than the n -th km

Question 2

b) Looking at the results above, a policy analyst concludes: "Rooms is a much more important determinant of house prices than crime, since its t -statistic is roughly twice as large." Explain the distinction between statistical significance and economic significance, and explain precisely why the analyst's comparison is misleading. Under what conditions could a coefficient be statistically significant yet economically negligible?

- **Statistical significance** (t -stat, p -value): is the effect distinguishable from zero?
- **Economic significance**: is the **magnitude** large enough to matter?

Question 2

- Both are **statistically significant**
- What matters: **one extra room** $\rightarrow 100(e^{0.3067} - 1) \approx 35.9\%$ in house value; **one extra crime per 100 residents** $\rightarrow -1.2\%$.

Question 2

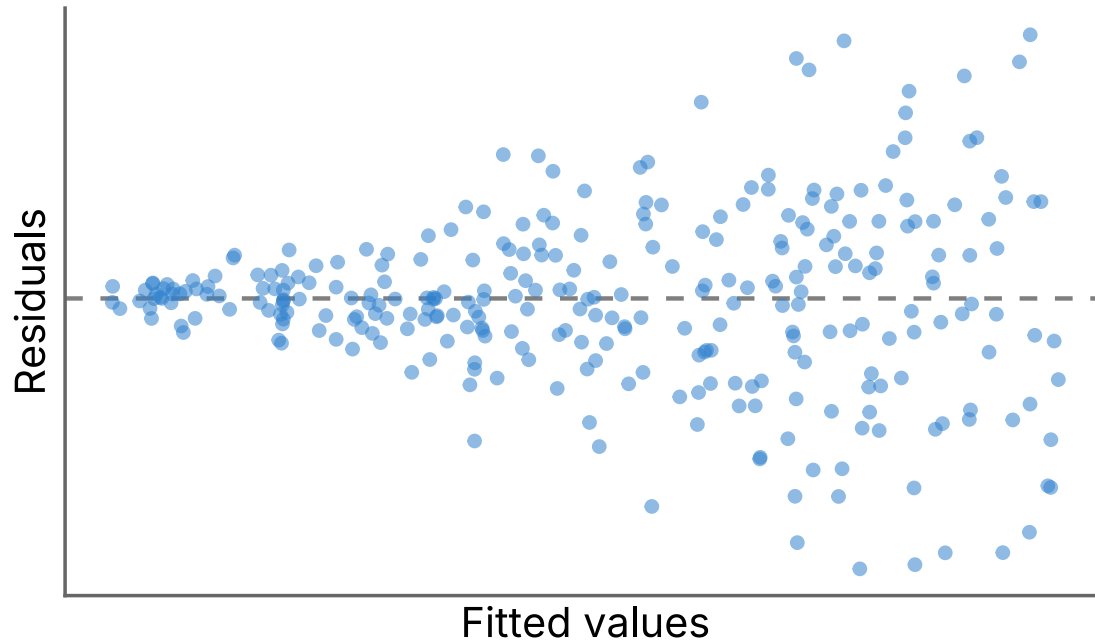
c) A colleague recodes *crim* from “per 100 residents” to “per 1000 residents”. Without re-running the regression, state the new coefficient on *crim*. What happens to its standard error, *t*-statistic, *p*-value, and the model’s R^2 ? Explain your reasoning.

- New variable: $\text{crim}^* = 10 \times \text{crim}$. For fitted values to be unchanged:

$$\hat{\beta}_{\text{crim}}^* = \frac{\hat{\beta}_{\text{crim}}}{10} = \frac{-0.0120}{10} = -\mathbf{0.0012}$$

- SE scales by the same factor: $\hat{\sigma}^* = 0.0013/10 = 0.00013$
- ***t*-statistic and *p*-value** unchanged
- R^2 unchanged

Question 2



d) After estimation, you obtain the following residual plot. Name the phenomenon visible in this plot. Is any of Stock and Watson's four Least Squares Assumptions violated? What are the consequences for (i) the OLS coefficient estimates, and (ii) the standard errors and inference? What remedy does the Stock and Watson framework recommend?

Question 2

- **Heteroskedasticity:** $\text{Var}(u_i | X_i)$ is not constant → here it grows
- **S&W assumptions:** None of the four assumptions violated
- **Consequences:**
 - (i) **Coefficient estimates:** OLS remains **unbiased and consistent**
 - (ii) **Standard errors:** wrong → inference compromised
- **Remedy:** use **heteroskedasticity-robust (HC) standard errors**

Question 2

e) Define Adjusted R^2 and explain why it differs from R^2 . After adding 15 extra control variables, a researcher reports that R^2 rose from 0.6486 to 0.6520 while Adjusted R^2 fell from 0.6458 to 0.6442. What does this tell you about the additional variables?

- R^2 rose mechanically (it always does when variables are added)

Question 2

$$\bar{R}^2 = 1 - \frac{SSR / (n - k - 1)}{SST / (n - 1)} = 1 - (1 - R^2) \cdot \frac{n - 1}{n - k - 1}$$

- Adds a **degrees-of-freedom penalty**: $\bar{R}^2$ rises only if the new variable reduces SSR **more** than the penalty requires
- $\bar{R}^2$ fell → the 15 extra variables **collectively add no explanatory power** → do not include them

Question 2

f) The researcher realizes that the variable nox (nitrogen oxide concentration, a measure of air pollution, in parts per 10 million) was accidentally left out of the model. Economic reasoning suggests:

- Higher pollution lowers house prices: $\beta_{nox} < 0$.*
- Pollution tends to be higher in high-crime areas: $Cov(crim, nox) > 0$.*

Does this make the estimated effect of $crim$ on house prices appear more negative or less negative than the true effect? What does this imply for the reliability of the regression?

Question 2

- OVB formula:

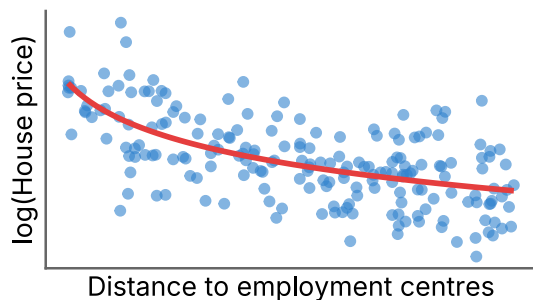
$$\hat{\beta}_{\text{crim}} \xrightarrow{p} \beta_{\text{crim}} + \rho_{\text{crim},u} \frac{\sigma_{\text{crim}}}{\sigma_u}$$

$$u = \beta_{\text{nox}} \text{nox} + v, \quad \beta_{\text{nox}} < 0 \quad \Rightarrow \quad \rho_{\text{crim},u} < 0$$

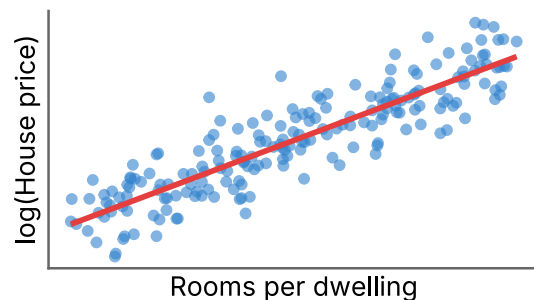
- Bias is **negative** $\rightarrow \hat{\beta}_{\text{crim}}$ is **more negative** than the true β_{crim}

Question 2

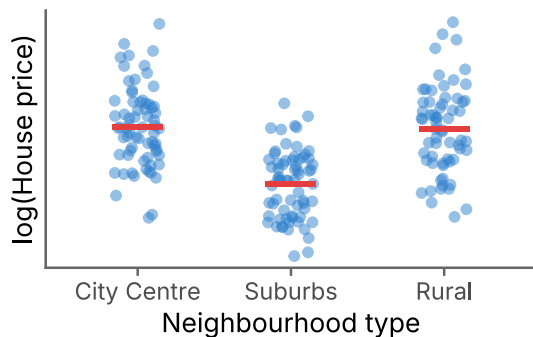
Plot A



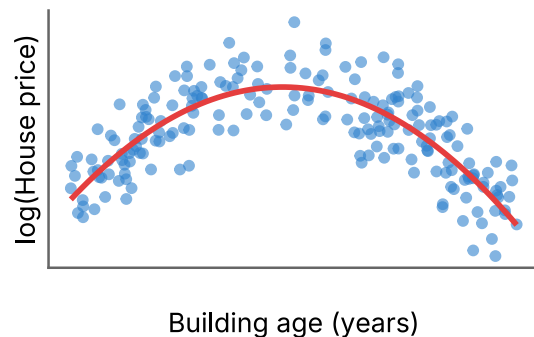
Plot B



Plot C



Plot D



g) Before specifying the regression model, a researcher creates four scatter plots of $\log(\text{medv})$ against different variables. The observed patterns are shown on the left. For each plot, write down the model specification that is most appropriate. Provide a brief econometric and economic justification for each choice.

Question 2

- **Plot A** → concave, decelerating decline → log:

$$\log(\text{medv}) = \beta_0 + \dots + \beta_1 \log(\text{dist}) + u$$

- **Plot B** → linear pattern, no curvature → levels:

$$\log(\text{medv}) = \beta_0 + \dots + \beta_1 \text{rooms} + u$$

- **Plot C** → categorical → dummies (one as reference):

$$\log(\text{medv}) = \beta_0 + \dots + \beta_1 D_1 + \beta_2 D_2 + u$$

- **Plot D** → inverted-U → quadratic:

$$\log(\text{medv}) = \beta_0 + \dots + \beta_1 \text{age} + \beta_2 \text{age}^2 + u$$

Question 3

Card (1990) exploits the unexpected arrival of approximately 125,000 Cuban immigrants in Miami between May and September 1980 — the so-called Mariel Boatlift — as a natural experiment to study the effect of immigration on local labor markets. This sudden influx increased Miami's labor force by roughly 7% almost overnight. Card compares wages of low-skilled non-Cuban workers in Miami against a group of comparable US cities (Atlanta, Houston, Los Angeles, and Tampa) that experienced no comparable shock. The table below reports average log hourly wages before and after the boatlift:

Group	Before (1979)	After (1981)
Miami	1.85	1.81
Comparison cities	1.80	1.74

Model: $\text{wage}_i = \beta_0 + \beta_1 \text{Miami}_i + \beta_2 \text{after}_i + \beta_3 (\text{Miami}_i \times \text{after}_i) + \epsilon_i$

a) Explain the economic interpretation of each coefficient $\beta_0, \beta_1, \beta_2, \beta_3$. Which one captures the causal effect of the boatlift on wages? Why are the other three necessary in the model?

$$\text{wage}_i = \beta_0 + \beta_1 \text{Miami}_i + \beta_2 \text{after}_i + \beta_3 (\text{Miami}_i \times \text{after}_i) + \epsilon_i$$

Question 3

	Before	After	Diff
Comparison	β_0	$\beta_0 + \beta_2$	β_2
Miami	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Diff	β_1	$\beta_1 + \beta_3$	β_3

Question 3

- β_0 : baseline wage in comparison cities **before** the boatlift
- β_1 : **pre-existing** Miami wage premium (absorbs permanent differences)
- β_2 : **common time trend** (macro shock affecting both groups)
- β_3 : **DiD estimator** — causal effect of the boatlift, under parallel trends

Question 3

b) Using only the information in the table above, derive numerical estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$. Show all steps clearly. What does the sign and magnitude of $\hat{\beta}_3$ suggest about the impact of the Mariel Boatlift on the wages of incumbent low-skilled workers?

$$\hat{\beta}_0 = 1.80$$

$$\hat{\beta}_1 = 1.85 - 1.80 = +0.05$$

$$\hat{\beta}_2 = 1.74 - 1.80 = -0.06$$

$$\begin{aligned}\hat{\beta}_3 &= \underbrace{(1.81 - 1.85)}_{\Delta \text{ Miami} = -0.04} - \underbrace{(1.74 - 1.80)}_{\Delta \text{ Comp} = -0.06} = \\ &= +\mathbf{0.02}\end{aligned}$$

- **Check:** $1.80 + 0.05 - 0.06 + 0.02 = 1.81 \checkmark$
- Miami wages fell **less** than comparison cities after the boatlift
- $\hat{\beta}_3 = +0.02$: **no adverse wage effect**

Question 3

c) State, in formal notation, the null hypothesis corresponding to "the Mariel Boatlift had no effect on wages in Miami."

$$H_0 : \beta_3 = 0$$

Under H_0 : Miami wages changed by exactly the same amount as comparison cities → boatlift had zero causal impact on incumbent low-skilled workers' wages

Question 3

*d) A fellow student proposes adding to the regression a variable treated_post_i , defined as equal to 1 when $\text{Miami}_i = 1$ **and** $\text{after}_i = 1$, and 0 otherwise — alongside Miami_i , after_i , and $\text{Miami}_i \times \text{after}_i$. What problem arises? Explain precisely.*

- By construction: $\text{treated_post}_i \equiv \text{Miami}_i \times \text{after}_i$ for every observation
- This is **perfect multicollinearity**: one column of X is an exact linear combination of others
- $X'X$ is singular $\rightarrow (X'X)^{-1}$ does not exist $\rightarrow$ OLS is undefined
- Software detects this automatically, drops one column, reports its coefficient as NA

Question 3

e) Card uses the Mariel Boatlift as a natural experiment rather than simply regressing wages on the local immigrant share across US cities. Explain why a cross-sectional OLS regression of wages on immigrant share would likely yield a biased estimate of the effect of immigration on wages. In which direction would the bias run, and why? How does the boatlift address this problem?

- Immigrants **self-select** into cities with stronger labor demand and higher wages
- Short regression:

$$\text{wage} = \beta_0 + \beta_{\text{imm_share}} \text{imm_share} + u$$

Question 3

- OVB formula:

$$\hat{\beta}_{\text{imm_share}} \xrightarrow{p} \beta_{\text{imm_share}} + \rho_{\text{imm_share},u} \frac{\sigma_{\text{imm_share}}}{\sigma_u}$$

$$u = \beta_{\text{labor_demand}} \text{labor_demand} + v, \quad \beta_{\text{labor_demand}} > 0 \quad \Rightarrow \quad \rho_{\text{labor_demand},u} > 0$$

- **Positive bias:** OLS makes immigration appear to **raise** wages
- **Boatlift fix:** the shock's timing and destination were determined by the Cuban government (entirely exogenous to Miami's economic conditions) → no self-selection → DiD recovers the true causal effect