Financial Econometrics
Shangai Jiao Tong University
Summer School – 2025-2026
Problem Set - Lecture 2
With Solutions

Fatih Kansoy

Reference: Stock, J.H. and Watson, M.W., Introduction to Econometrics (4th Edition, Global Edition)

Chapter 4: Linear Regression with One Regressor

Question 1. A regression of average monthly expenditure (AME, measured in euros) on average monthly income (AMI, measured in euros) using a random sample of college-educated full-time workers earning €100 to €1.5 million yields the following:

$$\widehat{AME} = 710.7 + 8.8 \times AMI, \quad R^2 = 0.030, \quad SER = 540.30$$

(a) Explain what the coefficient values 710.7 and 8.8 mean.

(b) The standard error of the regression (SER) is 540.30. What are the units of measurement for the SER? (Euros? Or is it unit free?)

(c) The regression R^2 is 0.030. What are the units of measurement for the R^2? (Euros? Or is R^2 unit free?)

(d) What does the regression predict will be the expenditure of a person with an income of €100? With an income of €200?

(e) Will the regression give reliable predictions for a person with an income of €2 million? Why or why not?

(f) Given what you know about the distribution of earnings, do you think it is plausible that the distribution of errors in the regression is normal?

Solution:

(a) **Interpretation of coefficients:**

- $\hat{\beta}_0 = 710.7$ (**Intercept**): This is the predicted average monthly expenditure when income is €0. It represents the baseline or autonomous consumption—expenditure that occurs regardless of income level.

  Economic interpretation: This could represent essential spending on necessities.

- $\hat{\beta}_1 = 8.8$ (**Slope**): For each additional €1 increase in average monthly income, average monthly expenditure is predicted to increase by €8.8.

  Wait—this seems wrong! A slope of 8.8 means spending increases by €8.8 for every €1 of income, which implies spending more than 100% of additional income. This is likely because AMI is measured in hundreds or thousands of euros, or there's a scaling issue in the data.

  More sensible interpretation: If AMI is in thousands of euros, then for each €1,000 increase in income, expenditure increases by €8.80.

(b) **Units of SER:**

The SER (Standard Error of the Regression) measures the typical size of the residuals. Since residuals have the same units as the dependent variable:

$$SER = 540.30 \textbf{ euros}$$

Interpretation: The typical prediction error is about €540.

(c) **Units of R^2:**

$R^2 = 1 - \frac{SSR}{TSS} = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}$

Since R^2 is a ratio of variances (both in euros^2), the units cancel:

$$R^2 \text{ is } \textbf{unit-free} \text{ (dimensionless)}$$

Interpretation: Only 3% of the variation in expenditure is explained by income—a very weak relationship.

(d) **Predictions:**

- At AMI = €100: $\widehat{AME} = 710.7 + 8.8 \times 100 = 710.7 + 880 = \boxed{1,590.70}$
- At AMI = €200: $\widehat{AME} = 710.7 + 8.8 \times 200 = 710.7 + 1760 = \boxed{2,470.70}$

(e) **Reliability for €2 million income:**

No, the regression will not give reliable predictions for €2 million.

Reasons:

- **Extrapolation:** The sample only includes incomes up to €1.5 million. Predicting at €2 million is *extrapolation* beyond the observed data range.
- **Linearity assumption:** The linear relationship may not hold at extreme income levels (diminishing marginal propensity to consume).
- **Prediction:** $\widehat{AME} = 710.7 + 8.8 \times 2,000,000 = 17,600,710$—spending €17.6 million monthly on €2 million income is impossible.

(f) **Normality of errors:**

No, the errors are unlikely to be normally distributed.

Reasons:

- The distribution of earnings is **positively skewed** (right-tailed), with many low earners and few very high earners.
- The regression errors inherit this skewness.
- Normal distributions are symmetric and have no minimum value, but earnings have a natural lower bound (€0).
- The kurtosis of earnings is typically higher than the normal distribution (heavy tails).

Implication: While OLS estimates remain consistent, inference based on normality (small-sample t-tests) may be unreliable. Large-sample inference using the CLT is more appropriate.

Question 2. A researcher runs an experiment to measure the impact of a short nap on memory. There are 200 participants and they can take a short nap of either 60 minutes or 75 minutes. After waking up, each participant takes a short test for short-term recall. Each participant is randomly assigned one of the nap times, based on the flip of a coin. Let Y_i denote the number of points scored on the test by the ith participant ($0 \leq Y_i \leq 100$), let X_i denote the amount of time for which the participant slept prior to taking the test ($X_i = 60$ or 75), and consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

(a) Explain what the term u_i represents. Why will different participants have different values of u_i?

(b) What is $E(u_i|X_i)$? Are the estimated coefficients unbiased?

(c) What concerns might the researcher have about ensuring compliance among participants?

(d) The estimated regression is $\hat{Y}_i = 55 + 0.17X_i$.

 (i) Compute the estimated regression's prediction for the average score of participants who slept for 60 minutes before taking the test. Repeat for 75 minutes and 90 minutes.

 (ii) Compute the estimated gain in score for a participant who is given an additional 5 minutes to nap.

Solution:

(a) **Interpretation of u_i:**

The error term u_i represents **all factors other than nap duration that affect test performance**, including:

- **Individual characteristics:** Inherent cognitive ability, natural memory capacity, age, health status
- **Pre-experiment factors:** Prior sleep quality, caffeine consumption, stress levels
- **During-experiment factors:** Sleep quality during the nap, environmental disturbances
- **Post-nap factors:** Alertness upon waking, motivation during the test
- **Measurement error:** Random variation in test performance

Why u_i differs across participants:

Each participant has unique cognitive abilities, physiological responses to sleep, and experiences random factors during the experiment. These individual-specific unobserved factors cause variation in u_i.

(b) **Conditional mean of u_i and unbiasedness:**

Because of **random assignment** (coin flip):

$$E(u_i|X_i) = E(u_i) = 0$$

Reasoning:

- Random assignment ensures X_i is **independent** of all participant characteristics (observed and unobserved).
- Since u_i captures unobserved characteristics, $X_i \perp u_i$.
- Independence implies $E(u_i|X_i) = E(u_i)$, and we normalize $E(u_i) = 0$.

Unbiasedness:

Since the first least squares assumption $E(u_i|X_i) = 0$ is satisfied (thanks to randomization), the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased**:

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad E(\hat{\beta}_1) = \beta_1$$

(c) **Compliance concerns:**

The researcher might worry about:

- **Non-compliance with assigned nap time:** Participants assigned 75 minutes might wake early; those assigned 60 minutes might oversleep.
- **Inability to fall asleep:** Some participants might not actually sleep for the full assigned duration.
- **Differential compliance:** If compliance varies systematically (e.g., anxious participants sleep less), this could introduce bias.

Impact on estimates:

If compliance is **equally imperfect** across both groups (non-differential measurement error in X), the estimates remain unbiased but lose precision (larger standard errors, attenuation bias toward zero).

If compliance differs systematically between groups, **bias** could be introduced.

(d) **Predictions from $\hat{Y}_i = 55 + 0.17X_i$:**

(i) **Predicted scores:**

$$\begin{aligned}
X = 60: &\quad \hat{Y} = 55 + 0.17 \times 60 = 55 + 10.2 = \boxed{65.2} \\
X = 75: &\quad \hat{Y} = 55 + 0.17 \times 75 = 55 + 12.75 = \boxed{67.75} \\
X = 90: &\quad \hat{Y} = 55 + 0.17 \times 90 = 55 + 15.3 = \boxed{70.3}
\end{aligned}$$

Caution: The prediction for 90 minutes is an **extrapolation** beyond the experimental design (only 60 and 75 minutes were tested). This prediction assumes the linear relationship continues to hold.

(ii) **Gain from additional 5 minutes:**

$$\Delta\hat{Y} = \hat{\beta}_1 \times \Delta X = 0.17 \times 5 = \boxed{0.85 \text{ points}}$$

Interpretation: Each additional minute of nap time is associated with an increase of 0.17 points on the memory test. Five extra minutes yields about 0.85 additional points.