# Financial Econometrics
# Shanghai Jiao Tong University
# Summer School – 2025
# Empirical Exercise – 2B –
# **With Solutions**

Fatih Kansoy

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

*Data:* `Earnings_and_Height.csv` *from* `http://www.pearsonglobaleditions.com`

**Empirical Exercise 4.2.** On the text website, you will find the data file `Earnings_and_Height`, which contains data on earnings, height, and other characteristics of a random sample of U.S. workers. A detailed description is given in `Earnings_and_Height_Description`, also available on the website. In this exercise, you will investigate the relationship between earnings and height.

**(a)** What is the median value of height in the sample?

(i) Estimate average earnings for workers whose height is at most 67 inches.

(ii) Estimate average earnings for workers whose height is greater than 67 inches.

(iii) On average, do taller workers earn more than shorter workers? How much more? What is a 95% confidence interval for the difference in average earnings?

**(b)** Construct a scatterplot of annual earnings (*Earnings*) on height (*Height*). Notice that the points on the plot fall along horizontal lines. (There are only 23 distinct values of *Earnings*). Why? (*Hint:* Carefully read the detailed data description.)

**(c)** Run a regression of *Earnings* on *Height*.

(i) What is the estimated slope?

(ii) Use the estimated regression to predict earnings for a worker who is 67 inches tall, for a worker who is 70 inches tall, and for a worker who is 65 inches tall.

**(d)** Suppose height were measured in centimeters instead of inches. Answer the following questions about the *Earnings* on *Height* (in cm) regression.

(i) What is the estimated slope of the regression?

(ii) What is the estimated intercept?

(iii) What is the $R^2$?

(iv) What is the standard error of the regression?

**(e)** Run a regression of *Earnings* on *Height*, using data for female workers only.

(i) What is the estimated slope?

(ii) A randomly selected woman is 1 inch taller than the average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

**(f)** Repeat (f) for male workers.

**(g)** Do you think that height is uncorrelated with other factors that cause earnings? That is, do you think that the regression error term, $u_i$, has a conditional mean of 0 given *Height* ($X_i$)? (You will investigate this more in the *Earnings and Height* exercises in later chapters.)

**Solution:**

## Data Overview

The dataset contains 17,870 observations on U.S. workers:

| **Variable** | **Description** | **Mean** | **Std. Dev.** |
|---|---|---|---|
| `earnings` | Annual earnings ($) | 46,875 | 26,923 |
| `height` | Height (inches) | 66.96 | 3.97 |
| `sex` | Gender (0 = Female, 1 = Male) | 0.44 | — |

**Important Data Feature:** Earnings are reported in 23 discrete brackets, not as continuous values. Each bracket is assigned a single average value, which is why the scatterplot shows horizontal bands rather than a continuous scatter.

| **Sample** | **Observations** | **Percentage** |
|---|---|---|
| Female (`sex` = 0) | 9,974 | 55.8% |
| Male (`sex` = 1) | 7,896 | 44.2% |
| Total | 17,870 | 100% |

## Key Notation and Formulas

The simple linear regression model:

$$\text{Earnings}_i = \beta_0 + \beta_1 \times \text{Height}_i + u_i$$

OLS estimates minimize the sum of squared residuals:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

The coefficient of determination:

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}$$

## (a) Median Height

| | Mean | Std. Dev. | Min | Median | Max |
|---|---|---|---|---|---|
| Height (inches) | 66.96 | 3.97 | 48 | **67** | 84 |

**Answer:** The median height in the sample is **67 inches** (5 feet 7 inches).

This median will be used in part (b) to divide workers into "shorter" ($\leq 67$ inches) and "taller" ($> 67$ inches) groups.

## (b) Average Earnings by Height Group

### (i) Workers with Height $\leq$ 67 inches

| | Mean | Std. Error | 95% Confidence Interval |
|---|---|---|---|
| Earnings ($) | 44,488 | 265 | (43,968, 45,009) |

$n = 10,114$

### (ii) Workers with Height $>$ 67 inches

| | Mean | Std. Error | 95% Confidence Interval |
|---|---|---|---|
| Earnings ($) | 49,988 | 305 | (49,389, 50,587) |

$n = 7,756$

### (iii) Difference: Tall − Short

| | Difference | Std. Error | 95% Confidence Interval |
|---|---|---|---|
| Tall − Short | $5,499 | 405 | ($4,706, $6,293) |

$t = 13.59, \quad p < 0.0001$

**Summary Table**

| Height Group | Mean Earnings | Std. Dev. | n |
|---|---|---|---|
| Short ($\leq$ 67 in) | $44,488 | $26,700 | 10,114 |
| Tall ($>$ 67 in) | $49,988 | $26,897 | 7,756 |
| **Difference** | **$5,499** | — | — |

**Answer:** Yes, taller workers earn significantly more. The difference of **$5,499 per year** represents more than 10% of average earnings. The 95% confidence interval ($4,706, $6,293) excludes zero, confirming this difference is statistically significant ($p < 0.0001$).

## (c) Scatterplot and Discrete Earnings Values

*[See Figure: Scatterplot of Earnings vs Height]*

**Observation:** The points in the scatterplot fall along **23 horizontal lines** rather than forming a continuous scatter.

**Explanation:** According to the data documentation, individual earnings were reported in **23 income brackets** (e.g., "\$20,000–\$25,000"), and a single average value is assigned to all workers within each bracket. Consequently, the dataset contains only 23 distinct values of earnings.

| Earnings Value | Frequency | Percentage |
|---|---|---|
| \$84,055 (top bracket) | 5,566 | 31.1% |
| \$33,713 | 1,725 | 9.7% |
| \$23,363 | 1,689 | 9.5% |
| ... | ... | ... |
| \$4,726 (bottom bracket) | 52 | 0.3% |
| **Total distinct values** | **23** | |

**Implication:** The discreteness of the earnings variable does not invalidate OLS regression, but it does affect the visual appearance of the scatterplot and may slightly affect standard error estimates.

## (d) Regression: Earnings on Height (All Workers)

### Regression Output

| | Coefficient | Std. Error | t-statistic | p-value |
|---|---|---|---|---|
| Height | 707.67 | 50.49 | 14.02 | 0.000 |
| Constant | −512.73 | 3,386.86 | −0.15 | 0.880 |
| $n = 17,870$, $R^2 = 0.0109$, Root MSE = \$26,777 | | | | |

### Estimated Regression Equation

$$\boxed{\widehat{\text{Earnings}} = -512.73 + 707.67 \times \text{Height}}$$

### (i) Estimated Slope

**Answer:** The estimated slope is **\$707.67 per inch**.

**Interpretation:** Each additional inch of height is associated with \$707.67 higher annual earnings, on average.

### (ii) Predictions

Using $\widehat{\text{Earnings}} = -512.73 + 707.67 \times \text{Height}$:

| Height | Calculation | Predicted Earnings |
|---|---|---|
| 65 inches | $-512.73 + 707.67 \times 65$ | **$45,486** |
| 67 inches | $-512.73 + 707.67 \times 67$ | **$46,901** |
| 70 inches | $-512.73 + 707.67 \times 70$ | **$49,024** |

**Note:** The difference between predictions for 70 and 65 inches is $\$49,024 - \$45,486 = \$3,538 = 5 \times \$707.67$, confirming the slope interpretation.

## (e) Regression with Height in Centimeters

**Unit Conversion:** 1 inch = 2.54 cm, so Height (cm) = Height (in) $\times$ 2.54

### Regression Output

| | **Coefficient** | **Std. Error** | **t-statistic** | **p-value** |
|---|---|---|---|---|
| Height (cm) | 278.61 | 19.88 | 14.02 | 0.000 |
| Constant | −512.73 | 3,386.86 | −0.15 | 0.880 |
| $n = 17,870$, | $R^2 = 0.0109$, | Root MSE = \$26,777 | | |

### Answers

| **Quantity** | **Value** | **Units** |
|---|---|---|
| (i) Slope | **278.61** | \$/cm |
| (ii) Intercept | −512.73 | \$ |
| (iii) $R^2$ | **0.0109** | (unitless) |
| (iv) Standard Error (SER) | **\$26,777** | \$ |

### Understanding the Unit Conversion

The slope changes when units change, but the relationship between coefficients is predictable:

$$\hat{\beta}_1^{(\text{cm})} = \frac{\hat{\beta}_1^{(\text{in})}}{2.54} = \frac{707.67}{2.54} = 278.61 \text{ \$/cm}$$

**Verification:** $278.61 \times 2.54 = 707.67$ ✓

**Key Insight:** The $R^2$ and SER are **unchanged** by the unit conversion. These measures depend on the fit of the model, not on the scale of measurement. The slope changes proportionally with the unit conversion factor.

## (f) Regression: Female Workers Only

### Regression Output

| | **Coefficient** | **Std. Error** | **t-statistic** | **p-value** |
|---|---|---|---|---|
| Height | 511.22 | 98.90 | 5.17 | 0.000 |
| Constant | 12,650.86 | 6,383.74 | 1.98 | 0.048 |
| $n = 9,974$, $R^2 = 0.0027$, Root MSE = \$26,801 | | | | |

### Estimated Regression Equation (Females)

$$\boxed{\widehat{\text{Earnings}}_{\text{female}} = 12,650.86 + 511.22 \times \text{Height}}$$

### Answers

**(i) Estimated Slope: \$511.22 per inch**

**(ii) Prediction for a woman 1 inch taller than average:**

- Average female height: 64.5 inches
- A woman 1 inch above average has height = 65.5 inches
- Her predicted earnings are \$511.22 *higher* than average female earnings

**Answer:** A woman who is 1 inch taller than average is predicted to earn **\$511.22 more per year** than the average woman in the sample.

## (g) Regression: Male Workers Only

### Regression Output

| | **Coefficient** | **Std. Error** | **t-statistic** | **p-value** |
|---|---|---|---|---|
| Height | 1,306.86 | 100.77 | 12.97 | 0.000 |
| Constant | $-43,130.34$ | 7,068.48 | $-6.10$ | 0.000 |
| $n = 7,896$, $R^2 = 0.0209$, Root MSE = \$26,671 | | | | |

### Estimated Regression Equation (Males)

$$\boxed{\widehat{\text{Earnings}}_{\text{male}} = -43,130.34 + 1,306.86 \times \text{Height}}$$

**Answers**

**(i) Estimated Slope: \$1,306.86 per inch**

**(ii) Prediction for a man 1 inch taller than average:**

- Average male height: 70.1 inches
- A man 1 inch above average has height = 71.1 inches
- His predicted earnings are \$1,306.86 *higher* than average male earnings

**Answer:** A man who is 1 inch taller than average is predicted to earn **\$1,306.86 more per year** than the average man in the sample.

## Comparison: All Workers vs. Female vs. Male

| | **All Workers** | **Females Only** | **Males Only** |
|---|---|---|---|
| Sample Size | 17,870 | 9,974 | 7,896 |
| Intercept ($\hat{\beta}_0$) | $-512.73$ | 12,650.86 | $-43,130.34$ |
| Slope ($\hat{\beta}_1$) | 707.67 | 511.22 | **1,306.86** |
| $R^2$ | 0.0109 | 0.0027 | 0.0209 |
| Root MSE | \$26,777 | \$26,801 | \$26,671 |
| Average Height | 66.96 in | 64.5 in | 70.1 in |

**Key Finding:** The height-earnings relationship is **much stronger for men** than for women. The male slope (\$1,307/inch) is more than **2.5 times** the female slope (\$511/inch).

## (h) Is Height Uncorrelated with Other Factors?

**Question:** Does the conditional mean of the error term equal zero given height? That is, does $E(u_i|\text{Height}_i) = 0$?

**Answer: Probably not.**

Height is likely correlated with other factors that affect earnings. This violates the first OLS assumption and implies that $\hat{\beta}_1$ may suffer from **omitted variable bias**.

**Potential Confounding Factors**

| Factor | Mechanism |
|---|---|
| **Gender** | Men are taller on average (70.1 vs. 64.5 inches) AND earn more. Part of the height-earnings relationship may simply reflect the gender wage gap. |
| **Nutrition/Health** | Better childhood nutrition leads to both greater adult height and better health/cognitive outcomes, which may increase earnings. |
| **Socioeconomic Background** | Higher-income families may provide better nutrition ($\rightarrow$ taller children) and better education ($\rightarrow$ higher earnings). |
| **Confidence/Perception** | Taller individuals may be perceived as more authoritative or confident, leading to better labor market outcomes independent of productivity. |
| **Physical Strength** | In some occupations, greater height correlates with physical strength, which may increase productivity and earnings. |

**Implications**

If height is positively correlated with omitted variables that positively affect earnings (e.g., gender, nutrition, family background), then:

$$\hat{\beta}_1 \text{ (estimated effect of height)} > \beta_1 \text{ (true causal effect)}$$

The **true causal effect of height on earnings is likely smaller** than our estimates suggest. The regression captures not only the direct effect of height (if any) but also the effects of correlated factors.

**Conclusion:** Height is almost certainly correlated with other earnings-determining factors. The condition $E(u_i|X_i) = 0$ is likely violated, and our estimates should be interpreted as **associations**, not causal effects. Later chapters will introduce methods (multiple regression, instrumental variables) to address this issue.

## Summary: Key Takeaways

| Question | Finding |
|---|---|
| Median height | 67 inches |
| Tall vs. Short earnings gap | \$5,499 per year (significant) |
| Slope (all workers) | \$707.67 per inch |
| Slope (females) | \$511.22 per inch |
| Slope (males) | \$1,306.86 per inch |
| $R^2$ (all workers) | 1.09% |
| Is height exogenous? | Probably not—omitted variable bias likely |

## The Big Picture

This exercise illustrates several important concepts:

1. **Economic vs. Statistical Significance:** The height-earnings relationship is statistically significant ($t = 14.02$, $p < 0.0001$) but explains only 1.09% of the variation in earnings ($R^2 = 0.011$). Many other factors matter for earnings.

2. **Heterogeneous Effects:** The relationship between height and earnings differs substantially by gender. The male slope is more than double the female slope, suggesting different labor market dynamics.

3. **Unit Conversion:** Changing the units of measurement (inches $\rightarrow$ centimeters) changes the slope coefficient proportionally but leaves $R^2$ and SER unchanged. The underlying relationship is the same regardless of units.

4. **Omitted Variable Bias:** Simple regression captures correlation, not causation. Because height is likely correlated with other earnings determinants (gender, family background, health), our estimates overstate the causal effect of height itself.

5. **Data Structure Matters:** The discrete nature of the earnings variable (23 brackets) affects the scatterplot's appearance but does not invalidate regression analysis.

*Calculations performed using Stata. See accompanying `EE_2B_solution.do` file.*