Financial Econometrics
Shangai Jiao Tong University
Summer School – 2025-2026
Problem Set - Lecture 5
With Solutions

Fatih Kansoy

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

Chapter 7: Hypothesis Tests and Confidence Intervals in Multiple Regression

Questions 7.1–7.6 refer to the following regression table using data on employees in a developing country ($n = 10,973$). The dependent variable is log average weekly earnings (AWE).

Regressor	(1)	(2)	(3)
High school graduate (X_1)	0.352	0.373	0.371
	(0.021)	(0.021)	(0.021)
Male (X_2)	0.458	0.457	0.451
	(0.021)	(0.020)	(0.020)
Age (X_3)		0.011	0.011
		(0.001)	(0.001)
North (X_4)			0.175
			(0.037)
South (X_5)			0.103
			(0.033)
East (X_6)			-0.102
			(0.043)
Intercept	12.84	12.471	12.390
	(0.018)	(0.049)	(0.057)
F-statistic (regional effects = 0)			21.87
SER	1.026	1.023	1.020
R^2	0.0710	0.0761	0.0814

Question 1. For each of the three regressions, add * (5% level) and ** (1% level) to the table to indicate the statistical significance of the coefficients.

Solution:

Compute t-statistics for each coefficient: $t = \hat{\beta}/SE(\hat{\beta})$.

Critical values: 5% level: $|t| > 1.96$; 1% level: $|t| > 2.576$

Column (1):

- High school: $t = 0.352/0.021 = 16.76^{**}$
- Male: $t = 0.458/0.021 = 21.81^{**}$
- Intercept: $t = 12.84/0.018 = 713.3^{**}$

Column (2):

- High school: $t = 0.373/0.021 = 17.76^{**}$
- Male: $t = 0.457/0.020 = 22.85^{**}$
- Age: $t = 0.011/0.001 = 11.0^{**}$
- Intercept: $t = 12.471/0.049 = 254.5^{**}$

Column (3):

- High school: $t = 0.371/0.021 = 17.67^{**}$
- Male: $t = 0.451/0.020 = 22.55^{**}$
- Age: $t = 0.011/0.001 = 11.0^{**}$
- North: $t = 0.175/0.037 = 4.73^{**}$
- South: $t = 0.103/0.033 = 3.12^{**}$
- East: $t = -0.102/0.043 = -2.37^{*}$
- Intercept: $t = 12.390/0.057 = 217.4^{**}$

All coefficients are significant at the 1% level ($**$), except East in Column (3) which is significant only at the 5% level ($*$).

Question 2. Using the regression results in column (1):

(a) Is the high school earnings difference estimated from this regression statistically significant at the 5% level? Construct a 95% confidence interval of the difference.

(b) Is the male–female earnings difference estimated from this regression statistically significant at the 5% level? Construct a 95% confidence interval for the difference.

Solution:

(a) **High school premium:**

$t = 0.352/0.021 = 16.76 > 1.96 \Rightarrow$ **Yes, significant at 5% level**

95% CI: $0.352 \pm 1.96 \times 0.021 = 0.352 \pm 0.041 = \boxed{[0.311, 0.393]}$

Interpretation: High school graduates earn 31–39% more than non-graduates (since the dependent variable is log earnings).

(b) **Gender premium:**

$t = 0.458/0.021 = 21.81 > 1.96 \Rightarrow$ **Yes, significant at 5% level**

95% CI: $0.458 \pm 1.96 \times 0.021 = 0.458 \pm 0.041 = \boxed{[0.417, 0.499]}$

Interpretation: Men earn 42–50% more than women on average.

Question 3. Using the regression results in column (2):

(a) Is age an important determinant of earnings? Use an appropriate statistical test and/or confidence interval to explain your answer.

(b) Suppose Alvo is a 30-year-old male college graduate, and Kal is a 40-year-old male college graduate. Construct a 95% confidence interval for the expected difference between their earnings.

Solution:

(a) **Statistical significance of age:**

$t = 0.011/0.001 = 11.0$

This is far greater than 2.576 (1% critical value), so age is **highly significant**.

95% CI for age coefficient: $0.011 \pm 1.96 \times 0.001 = \boxed{[0.009, 0.013]}$

Interpretation: Each additional year of age is associated with a 0.9–1.3% increase in weekly earnings.

(b) **Earnings difference between Alvo (30) and Kal (40):**

The difference in predicted log earnings:

$$\Delta \log\widehat{(AWE)} = 0.011 \times (40 - 30) = 0.011 \times 10 = 0.11$$

Standard error of the difference:

$$SE(\Delta) = |40 - 30| \times SE(\hat{\beta}_{\text{Age}}) = 10 \times 0.001 = 0.01$$

95% CI:

$$0.11 \pm 1.96 \times 0.01 = 0.11 \pm 0.0196 = \boxed{[0.090, 0.130]}$$

Interpretation: Kal (40 years old) is expected to earn 9.0–13.0% more than Alvo (30 years old).

Question 4. Using the regression results in column (3):

(a) Are there any important regional differences? Use an appropriate hypothesis test to explain your answer.

(b) Juan is a 32-year-old male high school graduate from the North. Mel is a 32-year-old male college graduate from the West. Ari is a 32-year-old male college graduate from the East.

(i) Construct a 95% confidence interval for the difference in expected earnings between Juan and Mel.

(ii) Explain how you would construct a 95% confidence interval for the difference in expected earnings between Juan and Ari.

Solution:

(a) **Joint test for regional effects:**

$H_0 : \beta_{\text{North}} = \beta_{\text{South}} = \beta_{\text{East}} = 0$

The F-statistic is 21.87.

Critical values: $F_{3,\infty}$ at 1% ≈ 3.78, at 5% ≈ 2.60

Since $21.87 \gg 3.78$, we **strongly reject** H_0.

Conclusion: There are significant regional differences in earnings.

(b) **Earnings comparisons:**

(i) **Juan (North, HS grad) vs. Mel (West, college):**

Note: West is the omitted (reference) category.

Difference in predicted log earnings:

$$\begin{aligned}\Delta &= (\text{Juan}) - (\text{Mel}) \\ &= [\beta_{\text{HS}} + \beta_{\text{North}}] - [0] \\ &= 0.371 + 0.175 - 0 = 0.546\end{aligned}$$

Wait—we need to be careful. Juan is HS graduate, Mel is college graduate. But "High school graduate" means HS grad vs. non-HS grad, not vs. college. Let me re-read...

Actually, the variable is "High school graduate (X_1)" which equals 1 if highest degree is HS, 0 otherwise. So if Mel is a "college graduate," then $X_1 = 0$ for Mel.

Juan: $\hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(1) + \hat{\beta}_3(32) + \hat{\beta}_4(1)$

Mel: $\hat{\beta}_0 + \hat{\beta}_1(0) + \hat{\beta}_2(1) + \hat{\beta}_3(32) + 0$ (West is reference)

Difference: $\hat{\beta}_1 + \hat{\beta}_4 = 0.371 + 0.175 = 0.546$

Hmm, but this means Juan (HS grad from North) earns *more* than Mel (college grad from West)? That seems backwards. Let me reconsider...

I think "High school graduate" means having at least a HS diploma (=1) vs. not (=0), and all observations are either HS or college educated. So college grads also have $X_1 = 1$.

If that's the case: Both Juan and Mel have $X_1 = 1$.

Difference: $\hat{\beta}_4 = 0.175$ (Juan in North, Mel in West)

Since we don't have the covariance structure, and North is directly in the regression, the 95% CI is:

$$0.175 \pm 1.96 \times 0.037 = 0.175 \pm 0.073 = \boxed{[-0.248, -0.102]}$$

Wait, that's wrong sign. Juan - Mel = North effect = 0.175, so Juan earns MORE.

$$\boxed{[0.102, 0.248]}$$

(ii) **Juan vs. Ari (East):**

Difference: $\hat{\beta}_4 - \hat{\beta}_6 = 0.175 - (-0.102) = 0.277$

To construct a CI, we need $SE(\hat{\beta}_4 - \hat{\beta}_6)$, which requires the covariance.

Method: Re-run the regression with West omitted *and* East omitted (i.e., include North and South, with East as reference). Then the coefficient on North directly measures the North-East difference with its standard error.

Question 5. The regression shown in column (2) was estimated again, this time using data from 1993 (5000 observations selected at random and converted into 2007 units using the Consumer Price Index). The results are:

$$\log AWE = 9.32 + 0.301 \cdot \text{High school} + 0.562 \cdot \text{Male} + 0.011 \cdot \text{Age}$$

$$\text{(SE:)} \quad (0.20) \quad (0.019) \quad (0.047) \quad (0.002)$$

$$SER = 1.25, \quad R^2 = 0.08$$

Comparing this regression to the regression for 2007 shown in column (2), was there a statistically significant change in the coefficient on High school?

Solution:

Test for change in high school premium:

Let $\hat{\beta}_{1,2007} = 0.373$ with $SE = 0.021$ (from column 2)

Let $\hat{\beta}_{1,1993} = 0.301$ with $SE = 0.019$

Difference: $\hat{\beta}_{1,2007} - \hat{\beta}_{1,1993} = 0.373 - 0.301 = 0.072$

Since the two samples are **independent**, the variance of the difference is the sum of variances:

$$SE(\hat{\beta}_{1,2007} - \hat{\beta}_{1,1993}) = \sqrt{(0.021)^2 + (0.019)^2} = \sqrt{0.000441 + 0.000361} = \sqrt{0.000802} = 0.0283$$

Test statistic:

$$t = \frac{0.072}{0.0283} = 2.54$$

Since $|2.54| > 1.96$, the difference is **significant at the 5% level**.

Conclusion: Yes, there was a statistically significant increase in the high school premium from 1993 to 2007. The premium increased from about 30% to 37%.

Question 6. In all of the regressions in the previous Exercises, the coefficient of High school is positive, large, and statistically significant. Do you believe this provides strong statistical evidence of the high returns to schooling in the labor market?

Solution:

No, statistical significance alone does not establish **causality**.

Concerns:

1. **Omitted variable bias:** The estimated coefficients may be biased due to omitted variables correlated with both education and earnings:
   - **Ability:** More able individuals may be more likely to complete high school AND earn higher wages.
   - **Family background:** Wealthier families invest more in education AND provide better networks/opportunities.
   - **Motivation:** More motivated individuals are more likely to graduate AND work harder.
2. **Selection bias:** Those who graduate high school may differ systematically from those who don't in ways that affect earnings independently of education.
3. **Reverse causality:** While unlikely for education, in some contexts the direction of causation is unclear.

What we can say:

The regressions show a strong **correlation** between high school completion and earnings. This is **consistent with** the hypothesis that education increases earnings, but does not **prove** it.

To establish causality:

- Natural experiments (e.g., compulsory schooling laws)
- Instrumental variables (e.g., distance to school)
- Randomized controlled trials (rare for education)

Question 7. Question 6.5 reported the following regression (where standard errors have been added):

$$\begin{aligned}
\widehat{Price} = {} & \underset{(22.1)}{109.7} + \underset{(1.23)}{0.567} \cdot BDR + \underset{(9.76)}{26.9} \cdot Bath + \underset{(0.021)}{0.239} \cdot Hsize \\
& + \underset{(0.00072)}{0.005} \cdot Lsize + \underset{(0.23)}{0.1} \cdot Age - \underset{(12.23)}{56.9} \cdot Poor
\end{aligned}$$

with $R^2 = 0.85$ and $SER = 45.8$.

(a) Is the coefficient on BDR statistically significantly different from zero?

(b) Typically, four-bedroom houses sell for more than three-bedroom houses. Is this consistent with your answer to (a) and with the regression more generally?

(c) A homeowner purchases 2500 square feet from an adjacent lot. Construct a 95% confidence interval for the change in the value of her house.

(d) Lot size is measured in square feet. Do you think that another scale might be more appropriate? Why or why not?

(e) The F-statistic for omitting BDR and Age from the regression is $F = 2.38$. Are the coefficients on BDR and Age statistically different from zero at the 10% level?

Solution:

(a) **Significance of BDR:**

$t = \frac{0.567}{1.23} = 0.46$

Since $|0.46| < 1.96$, the coefficient is **not significantly different from zero**.

(b) **Economic interpretation:**

This **is consistent** with the regression. Here's why:

The coefficient on BDR measures the effect of bedrooms **holding house size constant**. But in reality, a 4-bedroom house is typically larger than a 3-bedroom house.

The total effect of an extra bedroom includes:

- Direct effect: +\$567 (from BDR coefficient)
- Indirect effect via size: If an extra bedroom adds 150 sq ft, that's $0.239 \times 150 = \$35,850$

So 4-bedroom houses sell for more primarily because they're **bigger**, not because of the bedroom count per se.

(c) **95% CI for lot size increase:**

Change in predicted price: $\Delta \hat{Price} = 0.005 \times 2500 = 12.5$ (thousands)

Standard error: $SE = 2500 \times 0.00072 = 1.8$

95% CI: $12.5 \pm 1.96 \times 1.8 = 12.5 \pm 3.53 = \boxed{[\$8,970, \$16,030]}$

(d) **Scale of lot size:**

Yes, a different scale would improve **interpretability**:

- Current: 0.005 means $5 per additional square foot
- If measured in 1000 sq ft: coefficient would be 5.0 (easier to interpret)
- If measured in acres: coefficient would be ≈ 218 ($218K per acre)

The choice doesn't affect statistical significance, only readability.

(e) **Joint F-test for BDR and Age:**

$H_0 : \beta_{BDR} = \beta_{Age} = 0$

$F = 2.38$ with $q = 2$ restrictions and $n - k - 1 = 200 - 6 - 1 = 193$ degrees of freedom.

Critical value: $F_{2,\infty}$ at 10% ≈ 2.30

Since $2.38 > 2.30$, we **reject H_0 at the 10% level**.

Conclusion: While BDR and Age are individually insignificant, they are **jointly significant** at 10%. This can happen when regressors are correlated.