Continuous Assessment in Applied Econometrics – 2H

Anderson Vil
CY Cergy Paris Université
2026 – 2027
Master 1 – Mock exam 3 – Solutions

Answer the questions in order. All answers must be justified. Calculators are allowed. Throughout, log denotes the natural logarithm. Unless stated otherwise, the assumptions of the classical linear model, including normally distributed errors, are taken to hold. Critical values are read from the tables at the end of the paper. When the number of degrees of freedom is not printed, use the closest lower row, and use the row ∞ above 120. The marking scheme is indicative.

Exercise 1: An intern's report on umbrella sales *(4 points)*

A company runs umbrella kiosks in Paris train stations. An intern studied $N = 20$ autumn days: x_i is the rainfall of the day (mm) and y_i the number of umbrellas sold. He kept the following sums:

$$\sum_i x_i = 60, \qquad \sum_i y_i = 400, \qquad \sum_i x_i^2 = 380, \qquad \sum_i y_i^2 = 8{,}900, \qquad \sum_i x_i y_i = 1{,}560.$$

His report on the OLS regression $y_i = \beta_0 + \beta_1 x_i + u_i$ contains the following statements:

(R1) $\hat{\beta}_0 = 14.6$, $\hat{\beta}_1 = 1.8$ and $R^2 = 0.72$.

(R2) "The residuals of my regression sum to 4.2."

(R3) "When sales are counted in dozens of umbrellas, the R^2 falls to 0.06: the model then explains almost nothing."

(R4) "Without rain, nobody should buy an umbrella, so I also estimated the model without constant. Its SSR is 200, lower than the SSR of the model with constant (252): the model without constant fits better."

1. Check statement (R1) from the sums, showing your calculations. *[1]*

> **Solution.** $\bar{x} = 3$, $\bar{y} = 20$; $S_{xx} = 380 - 180 = 200$, $S_{xy} = 1560 - 1200 = 360$, $S_{yy} = 8900 - 8000 = 900$. $\hat{\beta}_1 = 360/200 = 1.8$, $\hat{\beta}_0 = 20 - 1.8 \times 3 = 14.6$, $R^2 = S_{xy}^2/(S_{xx}S_{yy}) = 129600/180000 = 0.72$. (R1) is correct; SSR $= 900 - 1.8 \times 360 = 252$.

2. For each of the statements (R2), (R3) and (R4), say whether the reported result is possible. Justify your answer with a property of OLS, and give the correct value when the reported one is wrong. *[2.25]*

> **Solution.** (R2) Impossible. With a constant, the first normal equation (first row of $X'\hat{u} = 0$, which follows from $MX = 0$) gives $\sum_i \hat{u}_i = 0$ exactly. *[0.5]*
> (R3) Impossible. Dividing y by 12 divides the residuals and the deviations from the mean by 12, so SSR and TSS are both divided by 144 and $R^2 = 0.72$ is unchanged. The intern divided the R^2 by 12. Only the coefficients change: $\hat{\beta}_0 = 1.217$, $\hat{\beta}_1 = 0.15$ dozen per mm. *[0.75]*
> (R4) Impossible. The model without constant minimises the SSR over a subset of lines (those through the origin), so its SSR cannot be lower than 252. Correct values: $\hat{\gamma}_1 = 1560/380 = 4.105$ and $\text{SSR}_0 = 8900 - 4.105 \times 1560 = 2495.8$, ten times larger. The premise is also wrong: $\hat{\beta}_0 = 14.6$ umbrellas are sold on dry days (forecasts, forgotten umbrellas). *[1]*

3. The intern concludes: "each additional millimetre of rain causes 1.8 extra umbrella sales". In two or three lines, is a causal reading more or less credible here than in most regressions estimated on observational data? *[0.75]*

Solution. More credible: rainfall is determined by the weather, not by the kiosks or the buyers, so it is plausibly unrelated to the other determinants of sales and $E(u \mid x) = 0$ is reasonable. Remaining caveats: a correlation with the day of the week or with temperature (season) could still bias the slope.

Typical mistakes. recomputing (R2) from the data instead of invoking $\sum \hat{u}_i = 0$; "(R3) is possible because units change"; comparing SSRs of nested models in the wrong direction; arguing for causality from the high R^2.

Exercise 2: A new substrate for tomato plants *(5 points)*

An agronomy laboratory tests a new growing substrate on $N = 6$ tomato plants. Plants 1 to 3 are grown in the standard substrate and plants 4 to 6 in the new one. The yields (kg per plant) are

$$y = (2,\ 3,\ 4,\ 5,\ 5,\ 8)'.$$

Let $d_i = 1$ if plant i is grown in the new substrate and 0 otherwise, and consider the model $y_i = \beta_0 + \beta_1 d_i + u_i$, with $X = [\iota,\ d]$ where ι is a vector of ones.

1. Compute $X'X$, $(X'X)^{-1}$ and $\hat{\beta}$. Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$. *[1]*

Solution. $X'X = \begin{pmatrix} 6 & 3 \\ 3 & 3 \end{pmatrix}$, $(X'X)^{-1} = \frac{1}{9}\begin{pmatrix} 3 & -3 \\ -3 & 6 \end{pmatrix}$, $X'y = (27, 18)'$, $\hat{\beta} = \frac{1}{9}(81 - 54,\ -81 + 108)' = (3, 3)'$. $\hat{\beta}_0 = 3$: mean yield with the standard substrate; $\hat{\beta}_1 = 3$: difference between the mean yields (6 vs 3 kg).

2. Show that the projection matrix $P = X(X'X)^{-1}X'$ is block diagonal, with two 3×3 blocks whose elements all equal $1/3$. Deduce $\hat{y} = Py$ without using $\hat{\beta}$, and check that $P^2 = P$. *[1.25]*

Solution. $P_{ij} = x_i'(X'X)^{-1}x_j$ with $x_i = (1, d_i)'$. Two standard plants: $(1,0)(X'X)^{-1}(1,0)' = 3/9 = 1/3$; two new plants: $(1,1)(X'X)^{-1}(1,1)' = (3 - 3 - 3 + 6)/9 = 1/3$; one of each: $(1,0)(X'X)^{-1}(1,1)' = (3 - 3)/9 = 0$. Hence $\hat{y} = Py$ replaces each yield by the mean of its group: $\hat{y} = (3, 3, 3, 6, 6, 6)'$. Each block $B = \frac{1}{3}J_3$ (J_3: matrix of ones) satisfies $B^2 = \frac{1}{9}J_3^2 = \frac{1}{9} \times 3J_3 = B$, so $P^2 = P$.

3. Compute $\hat{u} = My$, check that $X'\hat{u} = 0$ and interpret its two equations. Compute $\hat{u}'\hat{u}$. *[1]*

Solution. $\hat{u} = y - \hat{y} = (-1, 0, 1, -1, -1, 2)'$. $\iota'\hat{u} = 0$: residuals sum to zero; $d'\hat{u} = -1 - 1 + 2 = 0$: residuals sum to zero within the new-substrate group (hence also within the standard group). $\hat{u}'\hat{u} = 1 + 0 + 1 + 1 + 1 + 4 = 8$.

4. Compute the trace of P and deduce the trace of M. Use it to compute $\hat{\sigma}^2$, then the standard error of $\hat{\beta}_1$. Test at the 5% level the hypothesis that the new substrate increases the yield. Would a two-sided test lead to the same conclusion? *[1.25]*

Solution. $\mathrm{tr}(P) = 6 \times \frac{1}{3} = 2 = K$, so $\mathrm{tr}(M) = 6 - 2 = 4 = N - K$: the number of degrees of freedom. $\hat{\sigma}^2 = 8/4 = 2$; $\widehat{\mathrm{Var}}(\hat{\beta}_1) = 2 \times \frac{6}{9} = \frac{4}{3}$, se $= 1.155$. $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 > 0$: $t = 3/1.155 = 2.60 > t_{0.95}(4) = 2.132$: rejected. Two-sided: $2.60 < t_{0.975}(4) = 2.776$: not rejected. With only six plants, the conclusion depends on whether the direction of the effect is specified in advance.

5. A technician wants to add the column $s = \iota - d$ (equal to 1 for the standard substrate) to X. In two or three lines, explain why this is impossible. *[0.5]*

Solution. $\iota - d - s = 0$: with $c = (1, -1, -1)'$, $Xc = 0$, so $X'X$ is singular (not positive definite) and $(X'X)^{-1}$ does not exist: perfect collinearity (dummy variable trap).

Typical mistakes. $X'X = \begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix}$; computing P as $X'X(X'X)^{-1}$; using $N - 1$ degrees of freedom; a two-sided critical value for a one-sided alternative (or the reverse).

Exercise 3: Taxi fares, in euros and in dollars ***(5 points)***

A city's taxi regulator studies $N = 400$ rides. It estimates by OLS (standard errors in parentheses):

$$\widehat{fare} = \underset{(0.35)}{3.80} + \underset{(0.05)}{1.45}\, dist + \underset{(0.08)}{0.40}\, wait, \qquad R^2 = 0.88,$$

where *fare* is the fare paid (euros), *dist* the distance (km) and *wait* the number of minutes spent stopped in traffic. In the sample, the mean distance is 8 km and the mean waiting time is 5 minutes.

1. Interpret the coefficients on *dist* and *wait*. The official tariff is 1.50 euros per km: test, at the 5% level and against a two-sided alternative, whether the estimated rate per km is compatible with it. *[1]*

Solution. One more km raises the fare by 1.45 euros at given waiting time; one more minute stopped raises it by 0.40 euro at given distance. $H_0 : \beta_1 = 1.50$: $t = (1.45 - 1.50)/0.05 = -1$, $|t| < t_{0.975}(397) \approx 1.960$ (row ∞): not rejected.

2. A travel website re-estimates the equation with the fare in dollars (1 euro = 1.10 dollar) and the distance in miles (1 mile = 1.609 km), the waiting time being unchanged. Give the new coefficients and standard errors. What happens to the t statistics and to the R^2? Justify. *[1.25]*

Solution. $fare_{\$} = 1.1\, fare$ and $dist = 1.609\, miles$, so $\widehat{fare_{\$}} = 4.18 + 2.566\, miles + 0.44\, wait$, with standard errors 0.385, 0.0885 and 0.088 (each se is multiplied by the same factor as its coefficient). The t statistics are unchanged (ratio of two numbers multiplied by the same factor) and the R^2 is unchanged (SSR and TSS both multiplied by 1.1^2).

3. The regulator also estimates $\widehat{\log(fare)} = 1.10 + \underset{(0.02)}{0.78} \log(dist)$, $R^2 = 0.79$. Without re-estimating, give the coefficients and the standard error of the slope in the regression of $\log(fare/dist)$ on a constant and $\log(dist)$. Interpret this slope. *[1]*

Solution. Intercept 1.10, slope $0.78 - 1 = -0.22$, se 0.02 (identical residuals and regressor). A 1% longer ride has a fare per km 0.22% lower.

4. Test at the 5% level the hypothesis that fares are proportional to distance. *[1]*

Solution. H_0: elasticity equal to 1, i.e. slope 0 in the $\log(fare/dist)$ regression. $t = -0.22/0.02 = -11$, $|t| > t_{0.975}(398) \approx 1.960$ (row ∞): rejected.

5. Using the level equation, compute the elasticity of the fare with respect to distance at the mean point of the sample. In two or three lines, explain why both this elasticity and the slope of question 3 are below one. *[0.75]*

Solution. Mean fare: $3.80 + 1.45 \times 8 + 0.40 \times 5 = 17.40$; elasticity $= 1.45 \times 8/17.40 = 0.67$. Both are below one because part of the fare does not grow in proportion to distance: the pick-up charge (3.80), and the waiting charge if waiting time grows less than proportionally with distance, are spread over more km on long rides, so the fare per km falls with distance.

Typical mistakes. dividing (instead of multiplying) by 1.609; leaving the standard errors unchanged; slope -0.78 or $0.78/1$ in question 3; testing proportionality with $H_0 : \alpha_2 = 0$; elasticity computed at $dist = 1$.

Exercise 4: The diffusion of electric vehicles across départements ***(6 points)***

A think tank studies the 96 départements of metropolitan France in 2025. The dependent variable is the share of electric vehicles in new private car registrations (in %). The regressors are *chargers*, the number of public charging points per 10,000 inhabitants, the log of the median disposable income (euros), *urban*, the share of the population living in urban areas (%), and *mount*, a dummy equal to 1 for mountain départements. Column (4) reports the regression of $\log(inc)$ on a constant and *chargers*.

Table 1 – Share of electric vehicles in registrations, OLS estimates

Dependent variable	(1) *EVshare*	(2) *EVshare*	(3) *EVshare*	(4) $\log(inc)$
chargers	0.981	0.486	0.503	0.0662
	(0.075)	(0.112)	(0.111)	(0.0050)
$\log(inc)$		7.465	6.995	
		(1.366)	(1.405)	
urban			0.005	
			(0.009)	
mount			−0.517	
			(0.237)	
Constant	0.668	−72.381	−67.936	9.786
	(0.258)	(13.369)	(13.619)	(0.017)
N	96	96	96	96
R^2	0.644	0.731	0.746	0.654

Notes: standard errors in parentheses. *EVshare* is measured in % (e.g. 3.5 for 3.5%).

We denote by $\tilde{\beta}_1$ the coefficient on *chargers* in column (1), by $\hat{\beta}_1$ and $\hat{\beta}_2$ the coefficients on *chargers* and $\log(inc)$ in column (2), and by $\tilde{\delta}_1$ the coefficient on *chargers* in column (4).

1. Interpret the coefficients on *chargers* and on $\log(inc)$ in column (2); for the latter, compute the effect of a 10% higher median income. *[1]*

Solution. One more charging point per 10,000 inhabitants is associated with an EV share higher by 0.49 percentage point, for a given income. A 1% higher income is associated with a share higher by $7.465/100 = 0.075$ point; for $+10\%$: $7.465 \log(1.1) = 0.71$ point (approximation 0.75).

2. Test, in column (2) and at the 5% level, the hypothesis that charging points raise the EV share. State H_0 and H_1. *[0.75]*

Solution. $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 > 0$. $t = 0.486/0.112 = 4.34 > t_{0.95}(93) \approx 1.662$ (row 90): rejected.

3. Using only the R^2 of column (3), test at the 5% level the joint significance of its regressors. *[0.75]*

Solution. $F = \dfrac{0.746/4}{0.254/91} = 66.8 > F_{0.95}(4, 91) \approx 2.47$ (row 90): rejected.

4. (a) Test whether *urban* and *mount* can be jointly excluded from column (3), at the 5% and at the 10% level.

(b) Show that *mount* is individually significant at the 5% level (two-sided). In two lines, explain how this can be reconciled with your answer to (a).

[1.25]

Solution. (a) $F = \dfrac{(0.746 - 0.731)/2}{0.254/91} = \dfrac{0.0075}{0.00279} = 2.69$: $< F_{0.95}(2, 91) \approx 3.10$ (row 90), not rejected at 5%; $> F_{0.90}(2, 91) \approx 2.36$, rejected at 10%.
(b) $t = -0.517/0.237 = -2.18$, $|t| > t_{0.975}(91) \approx 1.987$ (row 90). No contradiction: the F test asks whether the two variables together add explanatory power; the clearly irrelevant *urban* ($t = 0.6$) dilutes the contribution of *mount*. Joint and individual tests answer different questions.

5. (a) Interpret $\tilde{\delta}_1$.

(b) Verify numerically the relation $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$.

(c) Is the effect of charging points overestimated or underestimated in column (1)? By how much? Explain in one sentence.

[1.25]

Solution. (a) One more charging point per 10,000 inhabitants is associated with a median income higher by about 6.6%.
(b) $0.486 + 7.465 \times 0.0662 = 0.486 + 0.494 = 0.980 \approx 0.981$ (rounding).
(c) The bias $+0.49$ is positive: column (1) overestimates the effect, by about one half, because richer départements have both more charging points and more EV buyers.

6. In two or three lines each:

(a) Rural départements (low *urban*) register few new cars, so their EV share is measured from few registrations and varies much more, for given values of the other regressors. What does this imply for your answers to questions 2 to 4?

(b) Operators install charging points where electric vehicles are already numerous. What does this imply for the interpretation of the coefficient on *chargers* in column (3)?

[1]

Solution. (a) Heteroskedasticity: coefficients remain unbiased, but the usual standard errors, hence the t and F tests of questions 2–4, are invalid (robust standard errors needed).
(b) Reverse causality: a high EV share (high u) attracts chargers, so *chargers* is correlated with u and its coefficient is biased upward; even 0.50 cannot be read as the causal effect of installing chargers.

Typical mistakes. interpreting 0.486 as a 48.6% effect; reading 7.465 as an elasticity; F with $N - 3$ degrees of freedom in question 4; concluding that *mount* "is not significant" because the joint test does not reject at 5%; claiming that heteroskedasticity biases the coefficients.