# Midterm Summary · Sessions 1 to 6

## MATH60033A · Quantitative Methods in International Business

Thierry Warin, PhD · HEC Montréal · A2026

**How to use these pages.** A map, not a substitute for the sessions: for each one, the ideas to *state* and *explain*. The formulas are here because they make the ideas precise; the midterm asks you to understand them, not to compute with them. The last page lists what loses marks. Then sit the sample paper (`assessments/midterm/EXAM.md`) under exam conditions, and only afterwards open `SOLUTIONS.md`.

**The paper.** 3 hours, on paper, in pen. A simple calculator; no notes, no computer, no internet. It tests **understanding**, with very little arithmetic. Six parts: **I** *Europe 2031* and what AI is, in words · **II** one of the readings: its question, its model, two critiques, another model · **A** 20 multiple-choice questions · **B** short answers (state a concept precisely) · **C** reading output (interpret a table in words) · **D** judgement (say what a result does and does not license). The midterm examines **Sessions 1 to 6**, all covered here. Session 7, the replication workshop, reviews them on the five papers' own data.

## Session 1 · What the course asks of you

The course is about **claims**: what a number lets you say, and what it does not. Three habits run through every answer. **Every number has units**: "one more point of X is associated with 0.31 more units of Y, holding Z fixed", not "a slope of 0.31". **Association is not causation**: unless the design isolates a cause, write *is associated with*, never *drives* or *causes*. **Say what you would check**: name the diagnostic that could prove you wrong.

**Europe 2031 and what AI is.** Know the scenario's argument and its premises, the seam between its fact-based prologue and its speculative narrative, and why a scenario is not a forecast (its numbers are parameters of the story). Be ready to explain, in words, what AI is and is not — the ideas of our Session 1 conversation — and to argue a position about Europe's options.

## Session 2 · Describing a variable before you model it

### Centre

- **Mean** $\bar{x} = \frac{1}{n}\sum_i x_i$: the balance point. Pulled by extreme values.
- **Median**: the middle observation. Robust to outliers; answers "what is typical?".
- **Trimmed mean**: the mean after dropping a fixed share (e.g. 10%) from each tail; a compromise.
- Mean > median suggests a long **right** tail; mean < median a long **left** tail. For a skewed variable, report the median (or both) and say why.

### Spread

- **Sample variance** $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$. Divide by $n-1$ because $\bar{x}$ was estimated from the same data: the deviations sum to zero, so only $n-1$ are free. Dividing by $n$ would underestimate the variance.
- **Standard deviation** $s = \sqrt{s^2}$, in the units of $x$. **IQR** $= Q_3 - Q_1$, the spread of the middle half, robust to outliers.

### Shape

$$g_1 = \frac{1}{n}\frac{\sum_i (x_i - \bar{x})^3}{s^3} \qquad g_2 = \frac{1}{n}\frac{\sum_i (x_i - \bar{x})^4}{s^4} - 3$$

- $g_1 > 0$: right-skewed; $g_1 < 0$: left-skewed. $g_2 > 0$: heavier tails than a normal (**excess** kurtosis; a normal has $g_2 = 0$).
- **Tests.** Substantially skewed if $|g_1| > 2\sqrt{6/n}$; substantially kurtic if $|g_2| > 4\sqrt{6/n}$. Always report the statistic **with** its threshold.
- **Empirical rule** (68–95–99.7% within 1, 2, 3 SDs) holds only for a roughly **normal** variable. Symmetric is not the same as normal, and the rule does not apply to a variable you have shown to be skewed.

**Example, in words.** Firm size in a sample of exporters: the mean is 250 employees, the median 18. The mean sits far above the median because a few very large firms pull it up — a long **right** tail, which the skewness statistic confirms when it exceeds its threshold. So the typical exporter is small, not medium-sized: report the **median** and the IQR, say the distribution is right-skewed, and do not apply the 68–95–99.7 rule to it.

## Session 3 · Can the regression be trusted?

### The model and the fit

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$. **Least squares** picks the $\hat{\beta}$ that minimise $\sum_i (y_i - \hat{y}_i)^2$; in matrix form $\hat{\beta} = (X^\top X)^{-1} X^\top y$. The residual is $e_i = y_i - \hat{y}_i$.

- **Slope in units:** a one-unit rise in $x_j$ is associated with a $\hat{\beta}_j$-unit change in $y$, **holding the other regressors fixed**.
- **What "controlling for" means (Frisch–Waugh–Lovell).** $\hat{\beta}_j$ equals the slope from regressing the part of $y$ not explained by the other regressors on the part of $x_j$ not explained by them. It is a **partial** relationship, not a marginal one.
- **Omitted variable bias.** If the true model includes $z$ with coefficient $\beta_z$, and $z = \delta_0 + \delta_1 x + v$, then the short regression of $y$ on $x$ alone gives $\mathbb{E}[\hat{\beta}_1^{short}] = \beta_1 + \beta_z \delta_1$. The bias has the sign of $\beta_z \times \delta_1$.
- **Unbiasedness** needs linearity, $\mathbb{E}[\varepsilon \mid X] = 0$ (exogeneity) and full column rank. Constant variance is **not** needed for unbiasedness.

### Adequacy: how much does the line explain?

$R^2 = 1 - \frac{SSR}{SST}$, the share of the variance of $y$ explained. **Adjusted** $R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$ penalises extra regressors. **RSE** $= \sqrt{SSR/(n-k-1)}$, the typical size of a residual, in units of $y$. A high $R^2$ says nothing about causation or about whether the assumptions hold.

### Significance

A coefficient is significant at 5% when its $p$-value is below 0.05: the data can tell it apart from zero. **Not significant is not "no effect"**: it means the data cannot distinguish the effect from zero — it may be small, or real but imprecisely measured.

### Validity: do the assumptions hold?

The **residuals-versus-fitted** plot is the instrument. Look for:

- **Linearity** — a curve in the residuals means the functional form is wrong (add a quadratic, transform).
- **Nearly normal residuals** — a Q–Q plot; matters most for intervals in small samples.
- **Constant variance** — a **fan** (residuals spread out as fitted values grow; clearer on a **scale–location** plot) is heteroskedasticity. The coefficients stay unbiased, but the precision the model reports (its $p$-values) is unreliable.
- **Multicollinearity** — a variance inflation factor $VIF_j = 1/(1 - R_j^2)$ above about 10 means $x_j$ is nearly a combination of the others: imprecise coefficients, unstable signs.

### Robustness: which observations move the answer?

- **Outlier**: a large residual. **Leverage point**: an unusual $x$. **Influential point**: one whose removal changes the fit — typically high leverage *and* a large residual.
- **Leverage** $h_{ii}$ is the $i$-th diagonal of the hat matrix $H = X(X^\top X)^{-1}X^\top$; $0 \le h_{ii} \le 1$ and $\sum_i h_{ii} = p$, the number of parameters. Average leverage is $p/n$.
- **Cook's distance** combines leverage and residual into one number per observation; inspect the spikes (a common flag is $D_i > 4/n$). Investigate an influential point; never delete it silently.

### Parsimony: comparing candidate models

$AIC = -2\ell + 2p$ and $BIC = -2\ell + p \ln n$, where $\ell$ is the maximised log-likelihood. **Lower is better**, and only on the **same rows**. BIC penalises extra parameters more heavily, so the two can disagree. Comparing a few candidates you chose is honest; letting a search pick the model is not model selection (Session 5 shows why).

**Example, in words.** GDP per capita on productivity, across European country-years. *The sentence:* comparing two country-years one index point apart, the more productive one has on average about €1,100 more GDP per capita. *The intercept* is the prediction at productivity zero, far outside the data — do not interpret it. $R^2 = 0.63$*:* productivity accounts for about two-thirds of the variation, which says nothing about cause.
*Omitted variable:* if investment raises income *and* goes with higher productivity, leaving it out lets productivity take credit for investment: the slope is biased **upward**. Adding investment changes the question the slope answers — and its size.

## Session 4 · Binary, ordinal and multinomial outcomes

### Why not OLS on a 0/1 outcome?

The **linear probability model** $P(Y=1)=\beta_0+\beta_1 x$ can predict probabilities below 0 or above 1, worst at extreme $x$. The **logistic** function keeps them in $(0,1)$:

$$p(x)=P(Y=1\mid x)=\frac{1}{1+e^{-x^\top\beta}},\qquad \ln\frac{p}{1-p}=x^\top\beta.$$

- $\beta_j$ is a change in the **log-odds**. $e^{\beta_j}$ is an **odds ratio**: the odds are multiplied by $e^{\beta_j}$ per unit of $x_j$. An OR of 1.30 means odds 30% higher, **not** a probability 30% higher.
- **Marginal effect** on the probability: $\partial p/\partial x_j = p(1-p)\,\beta_j$. It depends on where you evaluate it, and is largest at $p=0.5$.
- Estimated by **maximum likelihood**. **Wald test**: one coefficient at a time (its $p$-value). **Likelihood-ratio test**: $LR=2(\ell_{full}-\ell_{restricted})\sim\chi^2_q$, for $q$ coefficients jointly (e.g. all the dummies of one categorical variable).
- **McFadden's pseudo-$R^2$** $=1-\ell_{model}/\ell_{null}$. Not a share of variance; values of 0.2–0.4 already indicate a good fit.

### Ordinal logit (ordered categories)

$\ln\frac{P(Y\le j)}{P(Y>j)}=c_j-x^\top\beta$, with $J-1$ **cutpoints** $c_1<c_2<\ldots$ and **one** $\beta$ per predictor. Positive $\beta$ pushes up the scale. Category probabilities are **differences** of cumulative ones: $P(Y=j)=F(c_j-\eta)-F(c_{j-1}-\eta)$. **Proportional odds**: the same odds ratio applies at every cut — check it by fitting a binary logit at each cut.

### Multinomial logit (unordered categories)

Pick a **baseline** category. Each other category $k$ gets its own equation $\ln\frac{P(k)}{P(\text{base})}=x^\top\beta_k$. $e^{\beta}$ is a **relative risk ratio** against the baseline. Probabilities: $P(k)=e^{\eta_k}/\sum_m e^{\eta_m}$, with $\eta_{base}=0$. A negative coefficient does **not** mean that category's probability falls: every equation moves at once.

### From a probability to a decision

- A threshold turns $p$ into a yes/no. If a missed positive costs $c_{FN}$ and a false alarm $c_{FP}$, the cost-minimising threshold is $\tau^\star=\frac{c_{FP}}{c_{FP}+c_{FN}}$ — not 0.5.
- **Accuracy** is meaningless with a rare outcome: with a 3% base rate, "never" is 97% accurate. Report the base rate, **precision** $=TP/(TP+FP)$ and **recall** $=TP/(TP+FN)$.
- **Discrimination** (AUC: does the model rank positives above negatives?) is not **calibration** (does $p=0.2$ happen 20% of the time?). A high AUC does not make the probabilities usable.

**Example, in words.** A bank models default. *The coefficient* on being a first-time borrower is positive: first-time borrowers are more likely to default — the sign is readable directly. *Its odds ratio,* 1.65, means their **odds** of default are 65% higher than those of comparable repeat borrowers — not that their probability is 65% higher. *The probability* depends on the whole profile: the model may give 0.50 for one applicant and 0.10 for another, and the same odds ratio moves the first far more than the second. *The decision:* if a missed default costs much more than a false alarm, the bank should flag applicants well below 50%: the threshold follows the costs, not the default of 0.5.

## Session 5 · Regularisation

### Why a biased estimator can win

Test error = bias$^2$ + variance + irreducible noise. OLS is unbiased, but with few rows and many correlated predictors its **variance** is enormous. Accepting a little bias for a large cut in variance lowers the total. The penalty $\lambda$ is the dial: $\lambda = 0$ is OLS; $\lambda \to \infty$ sets every coefficient to 0 (predict the mean).

**Stepwise selection** makes one in-or-out decision at a time; resample the rows and it picks a different model, noise included. "The data chose these variables" is not a finding.

### Ridge

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_i (y_i - x_i^\top \beta)^2 + \lambda \sum_j \beta_j^2 = (X^\top X + \lambda I)^{-1} X^\top y$$

Shrinks every coefficient toward 0, never exactly to 0; shrinks most where predictors are nearly collinear. $X^\top X + \lambda I$ is invertible for any $\lambda > 0$, **even when** $p > n$. **Standardise first**: the penalty is scale-dependent.

### Lasso and soft-thresholding

The lasso penalises $\lambda \sum_j |\beta_j|$. Its diamond-shaped budget has **corners on the axes**, so some coefficients are **exactly 0**: it selects. Coordinate update (standardised columns, $\rho_j$ = covariance of $x_j$ with the partial residual):

$$\hat{\beta}_j = S_\lambda(\rho_j) = \text{sign}(\rho_j)\,(|\rho_j| - \lambda)_+ \qquad \text{(zero if } |\rho_j| \le \lambda\text{; else shrink by } \lambda)$$

Limits to state: with $p > n$ it keeps at most $n$ variables; among correlated predictors it keeps one, almost arbitrarily; the selected set is not "the true model".

### Elastic net

Mixes both penalties with weight $\alpha$ ($\alpha = 1$ lasso, $\alpha = 0$ ridge). Update: $\hat{\beta}_j = S_{\lambda\alpha}(\rho_j) \,/\, (1 + \lambda(1-\alpha))$. The **grouping effect**: correlated predictors enter together rather than one standing in for all.

### Choosing $\lambda$, and what not to claim

- **$K$-fold cross-validation**: fit on $K-1$ folds, score on the held-out fold, repeat, average; keep the best $\lambda$. Larger $K \to$ less bias (each fit uses more data) but more variance and more computation; $K = 5$ or 10 is the compromise. $\lambda_{1se}$ = the largest $\lambda$ whose cross-validation error cannot be told apart from the best: a smaller model at no visible cost.
- **Leakage**: imputation, standardisation and feature screening must happen **inside** each fold; for time series, split by time, never shuffle. Otherwise the CV score is optimistic.
- **p-values after selection are not valid**: the same data chose the variables and then tested them.

**Example, in words.** Two predictors, firms' cloud use and their e-commerce sales, move almost together, and both matter. *OLS* cannot tell them apart and may give one a large positive coefficient and the other a negative one. *Ridge* shrinks both and lets them share the effect: stable, but it keeps every variable, useless ones included. *The lasso* sets the useless predictors to exactly zero — and usually drops one of the two, almost at random. *The elastic net* sets the useless ones to zero and keeps the pair together. In all three, a zero means "not needed for prediction here", never "no effect".

## Session 6 · Transformations: when the effect is not a straight line

Linear regression is linear in the **coefficients**, not necessarily in $x$: you may add $x^2$, $\log x$ and other transformations, and it is still a linear regression.

- **Quadratic** $y = \beta_0 + \beta_1 x + \beta_2 x^2$: the effect of one more unit of $x$ is $\beta_1 + 2\beta_2 x$, so it **depends on** $x$. $\beta_2 < 0$: the effect weakens and can turn negative (an inverted U); $\beta_2 > 0$: it accelerates. Never interpret $\beta_1$ on its own.
- Read the curve **inside the range of your data**; never extrapolate a quadratic beyond it.
- A curve in the residual plot (Session 3) is the usual reason to add a squared term.

**Example, in words.** Productivity rises with a firm's age, but more and more slowly, and eventually falls: a squared term with a negative sign. The effect of one more year is therefore not one number — large for young firms, smaller for mature ones, negative for the oldest. The coefficient on age alone is the effect at age zero; report the effect at a few meaningful ages instead, inside the range of the data.

## What loses marks, session by session

**2** · a summary pasted without interpretation · a skewness or kurtosis without its threshold · calling a variable normal because it is symmetric · the empirical rule on a skewed variable.
**3** · a slope without units or without "holding the others fixed" · reading "not significant" as "no effect" · deleting an influential point silently · comparing AIC on different rows.
**4** · log-odds reported as probabilities · an odds ratio read as a relative risk · accuracy on a rare outcome with no base rate · a 0.5 threshold with unequal costs.
**5** · penalising unstandardised predictors · standardising or screening before the CV split · the lasso's selected set called "the variables that matter" · post-selection p-values · a dropped variable read as "no effect".
**6** · a quadratic read from $\beta_1$ alone · a curve extrapolated beyond the data.

## Writing a Part D answer

Part D rewards a **well-argued refusal**. A strong answer does four things, in order:

1. **Say what the number is.** "Seven variables that jointly predict well at one $\lambda$, on one sample" — not "the drivers".
2. **Say what it does not license**, and why: no causal design, a constructed measure, selection that invalidates the p-values.
3. **Name the evidence that would change your answer**: a design that isolates the cause, a validation of the measure, a stability check across resamples, an out-of-sample test.
4. **Recommend what to do meanwhile**: report the association with its uncertainty, run the named diagnostic first, and hold any decision that assumes causation.

**Ideas worth carrying in your head.** Mean against median under skew · a slope is a sentence with units, holding the others fixed · $R^2$ is a share of variance, not proof of cause · adjusted $R^2$ penalises useless variables · not significant is not no effect · odds $= p/(1-p)$ · $e^{\beta}$ multiplies the odds, not the probability · a probability needs the whole linear predictor · accuracy means nothing without the base rate · ridge shrinks, the lasso zeroes, the elastic net keeps groups together · $\lambda$ is chosen by cross-validation, inside the folds · with a squared term, the effect depends on $x$ · association is not causation.