STAT5003 Computational Statistical Methods

THE UNIVERSITY OF SYDNEY · POSTGRADUATE · SEMESTER 2, 2026

Full guide + ask Sia about STAT5003

SIDE 1/2 HANDWRITTEN A4 PLANNER · priority methods, formulas, assumptions and output cues **COPY THIS SIDE BY HAND** 60% final · individual secure examination

01 · Exam map

COURSE.JSON + FINAL FAQ

5 | Project 35% | Final 60%

ASSESSMENT	WT	FORMAT
Weekly Workshop Contribution	5%	Participation monitored in allocated workshops
Group Project	35%	Collaborative analysis of a selected dataset
Invigilated Final Examination	60%	Individual secure examination

"Cheat sheet: One A4 sheet of handwritten notes double-sided. It should be handwritten!! Please note that printed sheets, including those written on the iPad and printed, are not allowed" *(STAT5003 Final Exam FAQ, Semester 2 2026)*

02 · Response spine

C1 · C12

assumption + inputs → compute + output → diagnose boundary → interpret in context

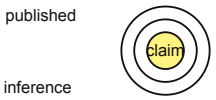
Every claim needs inputs, computation, diagnostic and interpretation.

LINE TO WRITE	WHAT IT PROVES
Target + population + units	You solved the requested object.
Rule before substitution	The operation is reconstructible.
Derived value + sign + scale	Arithmetic remains auditable.
Diagnostic / counter-case	The conclusion survives a check.
Context sentence + limitation	Output becomes analysis, not arithmetic.

THREE-ERROR AUDIT

translation error → calculation error → interpretation error. Repair the first failed move; then recheck a limit or independent route.

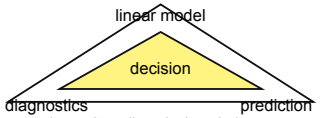
EXAM USE *Read the task verb first: estimate, predict, classify, compare, diagnose or interpret. Keep observed inputs separate from derived quantities.*



Do not let a valid general result become an overconfident case claim.

03 · Regression & smoothers

C3 · COURSE OVERVIEW



Fit → predict conditionally → check residual structure → qualify.

LINEAR MODEL

$E(Y|X)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$

$\hat{y}_i=\hat{\beta}_0+\sum_j\hat{\beta}_jx_{ij}$ · $e_i=y_i-\hat{y}_i$

Coefficient: expected change in the conditional mean for one-unit increase in that predictor, holding the others constant, under stated assumptions. Prediction belongs within the modelled range; attach uncertainty and units. Association from a fitted model is not automatically causal.

NEED	METHOD FAMILY	WHAT TO INSPECT
Conditional mean, interpretable slopes	Simple / multiple linear regression	Residual pattern, variance, range, uncertainty
Local pattern without one global line	Moving-average or kNN smoother	Neighbourhood width; edge behaviour
Flexible piecewise curve	Regression spline	Complexity and out-of-sample behaviour

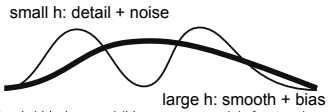
BIAS–VARIANCE DECOMPOSITION

Expected prediction error = irreducible noise + Bias² + Variance.

OUTPUT STEM *"Holding the other modelled predictors constant, the fitted conditional mean changes by … per …; this is an association within the model and data range."*

04 · Density & likelihood

C4 · GLOSSARY



Bandwidth changes visible structure; vary it before trusting a bump.

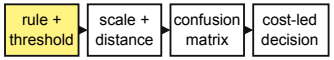
KERNEL DENSITY ESTIMATE

$\hat{f}_h(x)=(1/nh)\sum_iK((x-x_i)/h)$

small h → less smoothing · large h → more smoothing

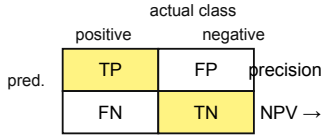
05 · Classification core

C5 · GLOSSARY



Thresholds create error trade-offs; stakeholder cost selects the operating point.

kNN: classify from nearby labelled cases. Scale predictors first or a large-unit variable dominates distance. Small neighbourhoods are flexible and sample-sensitive; larger neighbourhoods smooth the boundary. Tune inside resampling, then assess on independent evidence.



Read each metric from its own denominator; accuracy can hide imbalance.

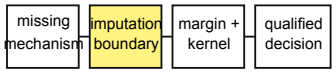
CONFUSION METRICS

Accuracy=(TP+TN)/N
Sensitivity/recall=TP/(TP+FN)
Specificity=TN/(TN+FP)
Precision=TP/(TP+FP)

Metric choice: actual positives → sensitivity; actual negatives → specificity; positive calls → precision. Under class imbalance, inspect errors by actual and predicted class rather than reporting accuracy alone.

06 · SVM + missing data

C6 · GLOSSARY

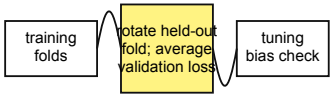


Data-loss assumptions remain separate from the classifier fitted afterwards.

ITEM	BOUNDARY
MCAR / MAR / MNAR	Name the assumed missingness mechanism; sensitivity matters when it depends on unobserved information.
Single / multiple / MICE	Carry imputation uncertainty; completed values are not observed facts.
SVM	Maximum-margin boundary uses support vectors; kernels represent nonlinear boundaries in a transformed feature space.

08 · CV & model selection

C7 · GLOSSARY

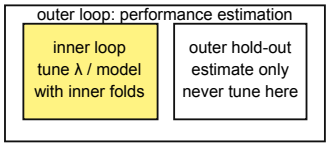


K-fold rotates validation; repeated CV reduces split sensitivity.

K-FOLD ESTIMATE

CV error = (1/k)Σⱼ Loss on fold j, after fitting on the other folds.

Split roles: training fits parameters; validation compares or tunes choices; independent test evidence supports the final out-of-sample claim. Once you repeatedly tune to one validation result, it has become training information.



Nested CV separates inner tuning from outer performance estimation.

SELECTOR	JOB / CAUTION
Best subset	Search predictor sets; expensive as choices grow.
Stepwise	Sequential search; path can miss another subset.
Cp / AIC / BIC / adjusted R^2	Balance fit against complexity with different penalties.
Repeated CV	Reduce split sensitivity; still keep tuning separate from final estimation.

09 · Regularisation

GLOSSARY

RIDGE

$\min\beta\ \text{Loss}(\beta)+\lambda\sum_j\beta_j^2$

LASSO

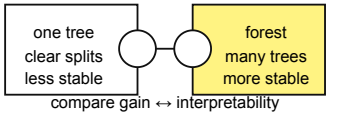
$\min\beta\ \text{Loss}(\beta)+\lambda\sum_j|\beta_j|$

λ controls shrinkage. Ridge shrinks correlated coefficients; lasso can set coefficients exactly to zero. Tune λ inside the resampling scheme and report the performance estimate from evidence not used to choose it.

OUTPUT CUE	WRITE NEXT
Training error falls as complexity rises	Do not choose from training fit; compare held-out loss.
Repeated CV varies across runs	Report split sensitivity rather than one fortunate split.
λ selected	Name the tuning loss and the resampling level that selected it.
Final model refitted	Keep final performance evidence independent of the choices made.

10 · Trees & ensembles

C8 · OVERVIEW



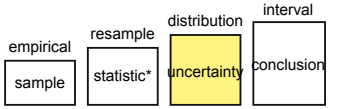
A forest aggregates trees; its importance is not a causal ranking.

METHOD	COURSE-LEVEL DISTINCTION
Decision tree	Interpretable splits; compare predictive gain with stability.
Bagging / random forest	Aggregate many trees; compare performance and interpretability.
AdaBoost / stochastic gradient boosting / XGBoost	Ensemble methods named in the course; tune and validate out of sample.
Importance	Name fitted model, data split and metric. Ranking can change with any of them.

BOUNDARY *Predictive importance is model- and metric-dependent. Never turn a ranking into a causal effect.*

11 · Bootstrap

C9



Resample under a stated scheme; retain the sample's defects.

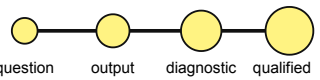
1. Define the observed empirical data and target statistic.
2. Resample according to the design; recompute the statistic each time.
3. Use the bootstrap distribution for uncertainty or interval construction.
4. State dependence, sampling and bias limitations.

A bootstrap inherits bias and dependence problems in the observed sample. Naive row resampling is not justified merely because software makes it easy; match the resampling unit to the data structure.

SAME GOAL?	KEEP DISTINCT
Forest aggregation	Targets predictive stability across many trees.

12 · Output sentence

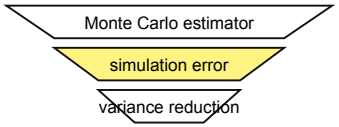
C12



Question → decisive output → diagnostic → assumption-sensitive conclusion.

13 · Monte Carlo

C10 · OVERVIEW



Improve efficiency without changing the target quantity.

MONTE CARLO MEAN

$\hat{\mu}^{MC}=(1/M)\sum_mg(X_m)$

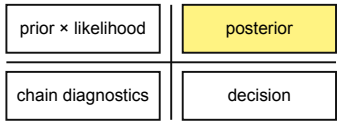
Monte Carlo SE ≈ s[g(X)]/√M

Define the target and sampling mechanism; simulate reproducibly; compute the estimator; report simulation error. Increasing iterations shrinks simulation error at roughly the square-root rate. It does not repair biased inputs or model misspecification.

TECHNIQUE NAMED IN COURSE	ROLE
Inverse transform	Generate from a target via a transformed uniform draw where available.
Acceptance–rejection	Generate candidates, then accept according to the target/proposal relationship.
Variance reduction	Improve precision for the same simulation budget without changing the estimand.

14 · Bayes & MCMC

C11 · OVERVIEW



Posterior updating is distinct from the computation used to approximate it.

BAYES RULE, PROPORTIONAL FORM

$p(\theta|y)\propto p(y|\theta)p(\theta)$

posterior ∝ likelihood × prior

METHOD / CHECK	WHAT EARNS THE LINE
Metropolis–Hastings	Name proposal, acceptance step and target posterior.
Gibbs	Update from conditional distributions in turn.
Diagnostics	Inspect stability and mixing across the chain; compare behaviour rather than citing length alone.
Sensitivity	Keep prior information and observed-data likelihood identifiable in the posterior interpretation.

TRAP *A long chain is not evidence of convergence or adequate mixing. More draws cannot rescue the wrong target or an unidentified model.*

Semester 2 formal examination period, from 16 November 2026 (the Exams Office publishes the exact date).

Your exam allows one HANDWRITTEN A4 (both sides): copy side 1 by hand; do not bring this printout.

STAT5003 Computational Statistical Methods · Drills & Fast Recall

THE UNIVERSITY OF SYDNEY · POSTGRADUATE · SEMESTER 2, 2026

Full guide + ask Sia about STAT5003

SIDE 2/2 DRILLS BEFORE THE EXAM · five 4-mark response patterns · traps · fast recall

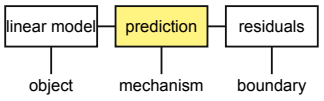
PRACTICE, THEN RETRIEVE UNAIDED

source-grounded · no invented datasets

D1 · Regression response [4]

C3 WORKED EXAMPLE

Prompt pattern: translate coefficient output into a conditional prediction and uncertainty statement; restore the missing link between the linear model and residual diagnostics.



Do not jump from model name straight to diagnostics; make prediction explicit.

1. **[1]** Mark the starting condition or object represented by the linear model.
2. **[1]** Write the change, rule or mechanism supplied by prediction as a verb-led link.
3. **[1]** Show how that link reaches residual diagnostics; do not skip an intermediate quantity or stage.
4. **[1]** Answer with the completed chain and preserve the limit: association from a fitted model is not automatically causal.

ANSWER FRAME

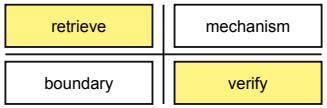
```
For [population], holding [other predictors] constant, [coefficient/prediction] means … . The residual check shows … . This supports [narrow conclusion], subject to [assumption/range/uncertainty]; it is not automatically causal.
```

LOSE THE MARK	REPAIR
Coefficient without units/reference	Name outcome, predictor, unit and held-constant condition.
"Good fit" from one statistic	Read residual pattern and uncertainty.
Extrapolation	State the modelled range.

SELF-MARK LINE	EVIDENCE VISIBLE?
Object and condition stated before arithmetic	Outcome, predictors, population and units are named.
Prediction link is verb-led	The coefficient/output changes a specified conditional mean.
Residuals test the fitted structure	A pattern, spread or boundary is interpreted rather than merely named.

R3 · Drill protocol

STUDY STRATEGY

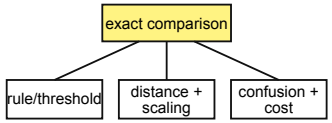


Retrieve → apply to a changed case → test a boundary → verify the first failed move.

D2 · Classification response [4]

C5 WORKED EXAMPLE

Prompt pattern: connect a classification threshold to errors and stakeholder cost; keep classification rule, distance/scaling and confusion matrix separate.



The matrix is a discriminator, not an afterthought.

1. **[1]** State the exact comparison required.
2. **[1]** Define the classification rule and attach the observation that belongs to it.
3. **[1]** Define distance and scaling separately; state why the concepts cannot be collapsed.
4. **[1]** Apply the confusion matrix to the same evidence and conclude within the limit that accuracy alone can mislead under class imbalance.

ANSWER FRAME

```
The rule classifies … at threshold … . Scaling matters because … . The confusion matrix shows [error type/metric]; given stakeholder cost, choose … rather than relying on accuracy alone.
```

SIGNAL	SAY
Minority positives costly to miss	Inspect FN and sensitivity.
False alarms costly	Inspect FP and precision/specificity.
Units differ	Scale before neighbour distance.
Threshold changes	Expect a sensitivity/specificity trade-off.



Moving a threshold trades error types; cost determines which trade is acceptable.

DENOMINATOR TEST	FAST INTERPRETATION
TP+FN	All actual positives: sensitivity asks how many were found.
TN+FP	All actual negatives: specificity asks how many were rejected.
TP+FP	All positive predictions: precision asks how many were correct.
N	Accuracy averages both classes and can conceal minority failure.

METRIC RECALL

```
Sensitivity=TP/(TP+FN) · Specificity=TN/(TN+FP)
Precision=TP/(TP+FP) · Accuracy=(TP+TN)/N
```

D3 · Cross-validation [4]

C7 WORKED EXAMPLE

Prompt pattern: estimate out-of-sample performance without leaking validation information; use k-fold CV and tuning bias to test the choice.

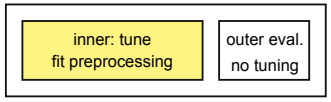


Validation used repeatedly for choices is no longer independent evidence.

1. **[1]** Extract the outcome or operation the task requires.
2. **[1]** State the precondition under which training/validation is relevant.
3. **[1]** Use k-fold CV to reject the nearest alternative; run a failure-path check for tuning bias.
4. **[1]** Choose the response and state when it must be withdrawn: repeated tuning against one validation result makes it part of training.

BAD WORKFLOW	WHY	REPAIR
Tune and quote same CV score	Optimistic selection	Nested CV or independent final evidence
Preprocess before folds	Held-out information enters training	Fit preprocessing within each training fold
Try many models on one split	Split becomes training information	Repeat CV; retain outer/test evidence

FAST CHECK *For every transformation ask: "Was this fitted using only the training portion available at that stage?"*



Inner choices; outer estimation. Keep those jobs distinct.

CHOICE MADE	WHERE IT MUST OCCUR
Scaling / imputation	Fit within the available training portion; apply to its hold-out.
Features / model family / λ	Compare inside the tuning loop.
Reported performance	Use the outer or independent evidence untouched by those choices.
Final refit	Refit only after the assessment protocol and model are fixed.

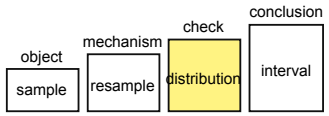
ONE-SENTENCE AUDIT

```
If a held-out value influenced preprocessing, selection or tuning, it is no longer held out for that decision.
```

D4 · Bootstrap response [4]

C9 WORKED EXAMPLE

Prompt pattern: approximate sampling uncertainty from a reproducible resampling scheme; assign empirical resampling, bootstrap distribution and interval construction separate jobs.



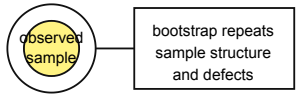
Observed sample → resampling rule → distribution → bounded conclusion.

1. **[1]** Use empirical resampling to fix the object or condition being analysed.
2. **[1]** Use the bootstrap distribution to state the mechanism.
3. **[1]** Use interval construction for a consequence, counter-case or check.
4. **[1]** Conclude without crossing the limit that the bootstrap inherits bias and dependence problems in the observed sample.

ANSWER FRAME

```
Resample [unit] from the observed empirical data under [scheme]; recompute [statistic] each time. The bootstrap distribution approximates sampling uncertainty, giving [interval/SE]. This relies on [sampling/dependence condition].
```

TRAP	CORRECTION
Resample individual rows despite dependence	Match the resampling unit to dependence.
Treat completed values as observed	Keep imputation assumptions visible.
Call a narrow interval unbiased	Precision does not erase sample bias.
Unrecorded randomness	Preserve the reproducible scheme.



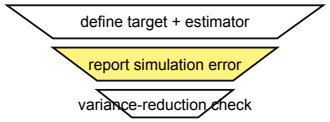
Resampling cannot manufacture information absent from the original sample.

ANSWER COMPONENT	REQUIRED DETAIL
Object	Observed empirical data, sampling unit and target statistic.
Mechanism	What is resampled, with replacement, and what is recomputed.
Distribution	What feature represents variability or supports the interval.

D5 · Monte Carlo + MCMC [4]

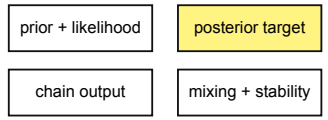
C10–C11 WORKED PATTERNS

Monte Carlo prompt: quantify approximation error and improve efficiency without changing the target.
MCMC prompt: separate posterior updating from the computation used to approximate it.



More iterations reduce simulation error, not model misspecification.

1. **[1]** Define the target quantity represented by the Monte Carlo estimator.
2. **[1]** Write the simulation-error relationship before substituting or simplifying.
3. **[1]** Carry the computation through; use variance reduction as the interpretation check.
4. **[1]** Report result, unit and scope; preserve the model-misspecification boundary.



Target distribution and approximation diagnostics answer different questions.

MCMC LINE	CHECK
Prior + likelihood → posterior	Keep sources of information identifiable.
Metropolis–Hastings / Gibbs	Name the update used to approximate the target.
Long chain	Not proof of convergence or mixing.
Posterior conclusion	Attach diagnostic and assumption limits.

SQUARE-ROOT RECALL *Multiplying iterations substantially gives only square-root improvement in Monte Carlo error; do not confuse computational precision with model validity.*

COMPARE	MONTE CARLO	MCMC
Target	Expectation / integral	Posterior approximation
Error cue	Simulation SE and square-root rate	Stability and mixing across a Markov chain
More iterations	Reduce simulation error	Do not alone establish convergence
Shared boundary	Computation cannot repair a misspecified target or unsupported assumptions.	

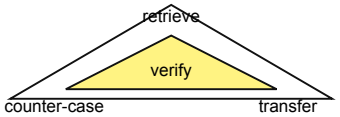
FINAL DRILL LINE

```
Target → sampling/update mechanism → numerical summary → diagnostic → scope-limited interpretation.
```

R1 · Method chooser

COURSE OVERVIEW

IF ASKED TO…	CANDIDATE FAMILY	DECISIVE CHECK
Predict a continuous response	Linear / multiple regression; local smoothers; splines	Residuals, range, uncertainty, CV
Estimate distribution shape	Histogram, KDE, maximum likelihood	Bins/bandwidth; model and sampling noise
Classify	Logistic regression, LDA, kNN, SVM, tree/ensemble	Scaling, threshold, confusion errors, cost
Visualise high-dimensional structure	PCA, t-SNE, MDS	Scale, selection and interpretation limits
Cluster	k-means, hierarchical	Representation, distance and stability
Select / tune	Best subset, stepwise, information criteria, ridge/lasso, CV	Complexity penalty and leakage
Quantify sampling uncertainty	Bootstrap	Resampling unit, bias, dependence
Approximate expectation/integral	Monte Carlo; variance reduction	Target, simulation SE, reproducibility
Approximate posterior	Metropolis–Hastings, Gibbs	Mixing, stability, prior/likelihood sensitivity

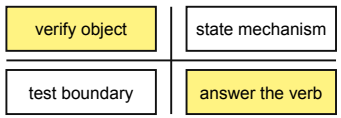


Retrieve unaided → transfer to a changed case → test a counter-case → verify.

R2 · Trap register

C2–C12 LIMITS

TEMPTING CLAIM	BOUNDARY TO WRITE
"The plot proves structure."	Scale, selection and overplotting can mislead.
"The bump is real."	Vary bins/bandwidth; it may be sampling or MC noise.
"Accuracy is high."	Inspect class-specific errors and stakeholder cost.
"Imputation fixed the data."	It did not restore unobserved information.
"CV is unbiased because k-fold was used."	Repeated tuning can leak validation information.



Final 20-second audit: object, mechanism, boundary, requested conclusion.