Final Exam Study Guide

EPID 785R: Applied Foundations of Study Design and Regression Modeling, Fall 2026

Ashley I Naimi

The exam. In person, closed book, hand-written long-answer questions; 50% of the course grade; the exam will be on Monday December 14, 9am to noon in room RRR-R344. You will very likely not need the full three hours. The expectation is that you are able to explain each idea correctly in your own words and to compute quantities from simple equations that are given, such as a fitted regression model or a small table. This study guide is organized by topic area. Under each topic is listed what you should be able to explain or compute, and where to look ("W5 pp. 14–17" = pages of the Week 5 notes PDF on the course site, its own page numbers; "Lab 3 Step 2" = the lab handout; "A2 Q3" = Assignment 2, Question 3; "PQ2" = pop quiz 2).

Design: the target trial and the data it needs (Weeks 1–2, Lab 1, A1)

- **[1] Target trial emulation.**
  - Know: each protocol component (eligibility; strategies; assignment; blinding; time zero and timescale; outcomes, with censoring, truncation and competing events declared; the causal contrast) and its emulation counterpart (eligibility reconstructed; strategies inferred from records; assignment replaced by confounder measurement and adjustment; no blinding, partly absorbed by active-comparator and new-user designs; time zero imposed); what randomization gives beyond exchangeability (i.e., positivity and causal consistency).
  - Where: W1 "Randomized Controlled Trials and Target Trial Emulation" pp. 10–29; Lab 1 (TARGET table, Steps 1–3, Q11–Q13); PQ3 Q1.
- **[2] ITT vs. per-protocol.**
  - Know: ITT $E(Y \mid X{=}1) - E(Y \mid X{=}0)$, naive as-treated $E(Y \mid A{=}1) - E(Y \mid A{=}0)$, naive per-protocol $E(Y \mid X{=}1, A{=}1) - E(Y \mid X{=}0, A{=}0)$; from the trial DAG ($X \to A \to Y$, $C \to A$, $C \to Y$) why only the ITT is identified by randomization; what a principled per-protocol analysis needs (adjustment for the determinants of adherence; sequential exchangeability for sustained strategies); why an emulation's "modified ITT" is *not* an ITT.
  - Where: W1 "Causal Contrasts" and "Identifying Assumptions" pp. 26–29; A1 Q5; Lab 1 Synthesis.
- **[3] Time zero, left truncation, right censoring.**
  - Know: time zero as the start of follow-up on a declared timescale; the alignment rule (eligibility, classification and start of follow-up coincide) and its failures (immortal time; prevalent users; NHEFS quitting with a 1971 versus a 1982 time zero); censoring = in the study, but event status (right) or event timing (left, interval) unobserved; truncation = never in the data because of when the event happened (left, right, interval); a competing event is not censoring; left-censored timing leaves only end-of-follow-up risk and odds.
  - Where: W1 "Time Zero and Follow-up" pp. 18–20, "Outcomes" pp. 20–26; W2 "Exposure Timing" p. 22; A1 Q3–Q4; Lab 1 Steps 1 and 3.
- **[18] Exposure definition.**
  - Know: the implied intervention and its contrast; the referent group; timing relative to time zero (new versus prevalent user; nothing from after time zero); measurement consistent with the contrast; two codings of one variable align with two different interventions (per-unit shift, threshold) and should be guideline-relative (e.g., the vegetable-intake example). Consequently, coding is a scientific decision for variables in the target function; the checklist of questions on W2 p. 21, applied to a published exposure.
  - Where: W2 "Risk Function Contrasts: The Exposure and the Implied Intervention" pp. 17–24; W1 "Treatment Strategies" pp. 12–14; Lab 1 Step 2.

Outcome-dependent sampling (Week 3, Lab 2, A2)

- **[4] Case-control sampling schemes.**
  - Know: sampling depends on the outcome only; estimable (exposure given outcome) versus not (risks, the case fraction, any absolute measure without outside information); why the sampling fractions cancel in the exposure odds ratio; cumulative (survivor) sampling gives the cohort OR; case-base (inclusive) sampling gives the RR exactly, no rare-disease assumption; incidence-density (risk-set) sampling gives the rate ratio; OR to RR three ways (rare outcome; the exact conversion with an external baseline risk; change the scheme). Given a cohort 2×2 table and the sampling fractions, compute what each scheme returns.
  - Where: W3 "Case-Control Studies" pp. 4–10, "The Three Designs Side by Side" p. 27; A2 Q1–Q2.
- **[5] Case weights in outcome-dependent sampling.**
  - Know: case-cohort = a subcohort of fraction q drawn at baseline plus all cases; weights $1/q$ on subcohort person-time, 1 at every event, a subcohort case switching from $1/q$ to 1 just before its event, an outside case adding no comparison time; why this rebuilds the cohort's denominators and leaves the target unchanged; the design check (weighted person-time and events match the cohort); why model-based SEs are wrong (weighted records are not people; the subcohort draw varies), so the person-clustered robust SE is required. Given a case-cohort sample's counts and q, rebuild the weighted table and compute the risks, risk ratio and risk difference.
  - Where: W3 "Case-Cohort Studies" pp. 10–17 (the weights pp. 14–16; variance pp. 16–17; checklist p. 17), "Outcome-Dependent Sampling" pp. 19–27; Lab 2 Steps 3–7 and Synthesis; A2 Q3; PQ3 Q2.

Regression: one machine, three uses, and its anatomy (Weeks 4–6, Labs 3–5, A2 Q4–Q5, A3)

- **[6] The descriptive-predictive-causal trichotomy.**
  - Know: the three question types and the operational test (what will be done with the answer); why the machinery is descriptive at the model level (the question appears nowhere in least squares or maximum likelihood; the fit is a projection, the best approximation within the family, which under misspecification depends on the covariate distribution; it describes the data generator, not the target population); the true-model myth; the three roadmaps (estimand, identification, estimation) and

their distinct middle steps: sampling and measurement (descriptive), transportability and out-of-sample loss (predictive), exchangeability, positivity and consistency plus model-layer assumptions (causal); the Table 2 fallacy; backwards selection judged under each of the three questions.
  - Where: W4 "Three Questions, One Machine" pp. 2–5, "Regression Machinery Is Always Descriptive" pp. 14–18, "Three Roadmaps for One Machine" pp. 21–25; Lab 3 (Q1–Q7); A2 Q4–Q5.
- **[7] Anatomy of a GLM.**
  - Know: left-hand side (outcome coding; the central-tendency function); right-hand side (functional form, coding, interactions; target function coded for interpretation, nuisance function for performance); link (the scale on which the model is additive, hence the contrast the exposure coefficient reports: identity gives a risk difference, log a risk ratio, logit an odds ratio); distribution (the variance behind the model-based SE); offset (coefficient fixed at 1; log person-time makes a rate model). Given a fitted model, label its parts, compute a predicted outcome and the exposure contrast on the model's scale and on the risk scale, and state what the exposure coefficient is under logistic, log-binomial, Poisson (with and without an offset) and linear-probability fits.
  - Where: W5 pp. 9–26 (Target and Nuisance p. 14; Links p. 17; Distributions p. 20; Offsets p. 21), Conclusion table p. 33; W2 note pp. 23–24; Lab 4 Steps 1–2; Lab 5 Step 1; A3 Q1, Q3.
- **[8] The parametric → semiparametric → nonparametric spectrum.**
  - Know: a saturated model has one parameter per covariate pattern and reproduces every cell mean (lossless; the link only changes coordinates); fewer parameters impose a restriction the data cannot contradict (lossy; the link decides which pattern survives); nonparametric imposes no form, parametric fixes it, semiparametric fixes a finite target and leaves the nuisance unrestricted (the Cox model); a right restriction buys precision, a wrong one is misspecification. Given a $2 \times 2 \times 2$ table, show that the saturated logistic model returns the cells and the three-parameter model does not; "which curve is right?" answered for a descriptive, a predictive and a causal question.
  - Where: W5 "The Parametric-to-Nonparametric Spectrum" pp. 26–33; W4 pp. 14–16; Lab 4 Steps 1, 3–4 (Q1–Q2, Q7–Q10); Lab 5 Step 3 (Q6–Q7); A3 Q2.
- **[19] The exponential family as organizing concept.**
  - Know: the exponential family is why one routine fits every GLM: naming the distribution fixes the variance function (Gaussian constant; binomial $\mu(1-\mu)$; Poisson μ; gamma μ^2) and a natural scale for its parameter; the canonical link maps the mean onto that scale (identity, logit, log), which is why logit is the default for a binary outcome; canonical is not "correct": a non-canonical link returns the contrast the question asks for (log gives a RR, identity a RD) at the risk of fitted means out of range (the log-binomial failure and its remedies); the family barely moves the coefficients but fully shapes the model-based SE. Given a family-link pair, name the contrast the exposure coefficient reports and the variance function behind its SE.
  - Where: W6 "The Exponential Family" pp. 7–12, Binomial pp. 15–22, "One Mean Model but Many Distributions" pp. 32–36, Conclusion table p. 36; Lab 5 Q1–Q3; A3 Q3–Q4.

Variance (Week 7, Lab 6, A4 Q1–Q2, PQ2)

- **[11] Model-based vs. sandwich vs. bootstrap.**
  - Know: what each estimator builds the SE from: model-based, the family's variance function at each fitted mean, so it inherits that assumption (constant variance for a linear model; μ for Poisson; $\mu(1-\mu)$ for binomial); robust (HC0–HC3), each person's observed squared residual, so it survives a wrong variance function or heteroscedasticity but not a wrong mean model; cluster-robust, residuals pooled within cluster (what GEE reports; needs many clusters); bootstrap, resampling (normal-interval or percentile intervals; resample clusters for clustered data). The consequences of a model-based SE under a wrong variance function (Poisson model of a binary outcome: SE 60% too large; heteroscedastic linear model: coverage 0.83 instead of 0.95); a large gap between model-based and robust SEs is a diagnostic of the model. Given a coefficient and its SEs from several estimators, say which to report and why, and compute the confidence interval.
  - Where: W7 "Model-Based Standard Errors" pp. 5–10, "Robust (Sandwich) Standard Errors" pp. 10–21 (cluster-robust and GEE pp. 18–21), "The Bootstrap" pp. 21–34, Conclusion table p. 34; W6 pp. 32–36; Lab 6 (Q1–Q8); A4 Q1–Q2; PQ2 Q1.

Conditional versus marginal adjustment (Weeks 8–9, A4 Q3–Q4)

- **[9] Collapsibility and the CCF.**
  - Know: the Greenland table (Z a cause of Y, unrelated to X): the RD is constant across strata (strictly collapsible), the RR collapses as a weighted average (collapsible), the OR does not (2.67 in both strata, 2.25 overall); the characteristic collapsibility function formalizes this: for a given link, it maps each person's unexposed risk to their exposed risk, is a straight line for the identity and log links (so averaging predictions returns the coefficient) and bends for the logit (so the OR of the averaged risks is not the conditional OR); hence conditional OR $\neq$ marginal OR in a logistic model even with no confounding; interactions as a second, separate source of difference; conditional/marginal describe the estimand, adjusted/unadjusted the analysis. Given a stratified 2×2 table, compute the stratum and overall RD, RR and OR and say which collapse.
  - Where: W8 "Collapsibility and Noncollapsibility: Basic Concepts" pp. 2–5, "Noncollapsibility in a Regression Context" pp. 5–8, "the CCF at work" pp. 16–17, Conclusion p. 38; A4 Q3.
- **[10] G-computation vs. IP weighting.**
  - Know: marginal standardization (g-computation; the point-exposure g-formula): fit the outcome model, predict for everyone at X=1 and at X=0, average, contrast; the stratified outcome model for full exposure-covariate interaction; bootstrap the whole procedure. IP weighting: propensity score $P(X{=}1 \mid C)$; weights $1/P(X{=}x \mid C)$ or stabilized $P(X{=}x)/P(X{=}x \mid C)$; the pseudo-population in which C no longer predicts X; robust SE with weights treated as known (conservative), or bootstrap including the propensity fit. Both need exchangeability, positivity and consistency and target the same marginal estimand; they differ in the nuisance function modeled (outcome versus exposure), hence in which misspecification has effects and when each is preferred (rare outcome, many confounders: weighting). Standardize by hand from a given model; compute four-cell weights and the weighted contrast.
  - Where: W8 "Marginal Adjustment" pp. 12–18, interactions and stratified models pp. 18–27, "Propensity Scores and Inverse Probability Weighting" pp. 27–38, Conclusion table p. 38; W4 "The causal roadmap" p. 24; A4 Q4.

Flexible and penalized regression (Weeks 10–12; FR = the Flexible Regression notes, PR = the Penalized Regression notes)

- **[12] Nuisance-function misspecification as bias.**
  - Know: a confounder entered on the wrong form leaves part of its relation with the outcome unexplained, and that leftover moves into the exposure coefficient because exposure depends on the confounder: residual confounding is a property of the population that no sample size removes and that vanishes only if the form is right or there is no confounding; a flexible nuisance term (such as a spline) removes it, so this kind of flexibility is an attempt at bias control; splines as extra columns in the design matrix (basis expansion) joined at knots (natural, restricted cubic, B-splines; report predictions and contrasts, not coefficients); the price of letting the data choose the form (printed SEs ignore the choice; a bootstrap does not repair a biased center); the partially linear model (parametric exposure term, flexible nuisance) as the way to keep a reportable coefficient with an honest SE. Given a spline fit's predictions at two doses, compute and interpret the contrast.
  - Where: FR "Why Linearity Fails" pp. 2–5, "Splines" pp. 5–11, "What We Lose: Inference" pp. 16–17, "Partially Linear Models" pp. 17–22; W5 pp. 14–17; PQ2 Q1; Lab 5 Q7.
- **[13] Penalized regression and causal inference.**
  - Know: ridge adds a penalty on squared coefficients and shrinks them all; the lasso penalizes absolute values and can set coefficients exactly to zero; the elastic net keeps a duplicated pair together; Firth penalizes the likelihood and keeps a separated coefficient finite (the second remedy in A3 Q4c); shrinkage is bias-variance tradeoff by construction; why the lasso is inappropriate for causal estimation: it can shrink the exposure coefficient toward zero, and exempting the exposure makes it delete confounders with small outcome coefficients, leaving the residual confounding of [12]; the interval printed after selection ignores the selection (a spike at zero plus a displaced hump; coverage below nominal; the bootstrap inherits it); a penalty belongs on the nuisance functions inside an estimator built to tolerate it, never on the exposure coefficient.
  - Where: PR "When the Criterion Cannot Decide" pp. 2–5, "One Criterion, Three Penalties" pp. 5–9, "Penalization and Causal Estimation" pp. 11–14 (the "exam topic 13" box); A3 Q4(c).
- **[14] Cross-validation for penalty selection.**
  - Know: the risk of a rule is its expected loss on new data; in-sample loss is optimistic, more so with flexibility, so it can't be used to choose; K-fold cross-validation (fit on $K-1$ folds, predict the held-out fold, average) estimates out-of-sample prediction risk at each λ or degrees of freedom, for new data from the same distribution; $\lambda_{\min}$ versus the one-standard-error rule as an explicit bias-variance choice; the chosen λ is itself a noisy estimate; not guaranteed: that the selected columns are the true ones, valid SEs for the selected model, calibration, or an honest error at the chosen setting (that needs a nested outer loop).
  - Where: PR "Cross-Validation: Tuning the Penalty and Honest Evaluation" pp. 9–11 (the "exam topic 14" box); FR "Cross-Validation and the Discrete Super Learner", "The Super Learner" pp. 11–16; W4 "The predictive roadmap" p. 23.

Survival analysis (Week 2 foundations; the Dec 1 lecture to come)

- **[15] Competing events and cumulative incidence.**
  - Know: why simple counting breaks under right censoring; Kaplan-Meier: at each event time, survival is multiplied by one minus the share of those at risk who fail, and censored people simply leave the risk set, which pushes their risk mass onto later events (redistribution to the right; ten-person example: $\hat{F}(11) = 0.475$ and $\hat{F}(16) = 1$ with seven deaths observed); cause-specific risk (competing events coded as censored; the switched-off world) versus sub-distribution risk (the world as it is; 0.7 and 0.3, summing to the all-cause risk); Aalen-Johansen: each event's jump is the all-cause survival just before it times the share failing from that cause, so $1-\text{KM}$ overstates the sub-distribution risk by exactly the competing-event mass; the sub-distribution risk is the decision-relevant, more causal quantity; coding a competing event as censored changes the estimand. Given a small cohort's event times and types, compute $1-\text{KM}$ and the Aalen-Johansen risk by hand.
  - Where: W2 "Risk" through "Cause Specific, Sub-Distribution, and Causality" pp. 3–10; the survival notes when posted; A1 Q4(b); PQ1.
- **[16] The hazard ratio as a summary measure.**
  - Know: the hazard as the instantaneous rate among those still at risk; a good causal estimand is (A) a contrast of the two counterfactual outcome distributions, (B) taken within a common set of units; the risk difference meets both (same denominator for both arms); the hazard ratio fails (B): the two hazards condition on different survivor sets (region H), Hernán's "built-in selection bias", which grows over follow-up in a heterogeneous population even with correctly estimated hazards; why "the HR is biased" is imprecise (bias belongs to estimators, not estimands); the hierarchy of information: risk at the top, with rate (events per accumulated person-time), odds and hazard as functions of it, and what data each needs (A1 Q6).
  - Where: W2 "Functions of Cumulative Risk" pp. 10–17 ("Hazard" pp. 11–15); the survival notes when posted; Hernán 2010.
- **[17] Parametric AFT vs. semiparametric Cox.**
  - Know, from material in hand: the Cox model as the flagship semiparametric model: a finite set of hazard-ratio coefficients and a baseline hazard left unspecified (a nuisance function), so the fitted model alone reports hazard ratios, not risks; what a parametric baseline (Weibull, generalized gamma AFT) buys, a complete risk curve and extrapolation, at the price of a shape assumption; how a risk curve is built from each (the AFT distribution evaluated at the covariates; the Cox baseline survival estimated after the fit and raised to $\exp(\beta^\top x)$, or standardized over covariates). Given fitted coefficients and a baseline survival value, compute the risk at t for a covariate pattern.
  - Where: W5 technical-definitions note p. 30; the survival notes when posted.