Assignment 2

Outcome-Dependent Sampling and Regression as a Toolkit (Weeks 3–4)

Ashley I Naimi

Due: Tuesday, September 22, 2026

1 Overview

This assignment covers material from Week 3 (Outcome-Dependent Sampling) and Week 4 (Regression as a Toolkit). It is worth **5 points**, allocated for completion, and is due **Tuesday, September 22**. You may resubmit as many times as you like over the course of the semester. You are encouraged to work in groups, but each of you must submit your own work.

The assignment consists of the free response questions below.

2 Free Response Questions

Answer each of the following. Where a question has parts, answer every part. Show your reasoning; a correct conclusion with no argument behind it will not receive useful feedback from your peer reviewer. Questions 1–3 cover Week 3; Questions 4–5 cover Week 4.

2.1 Question 1: Cancellation in the CACO Odds Ratio

A closed cohort is followed over a fixed risk period and cross-classified by binary exposure A and outcome Y, with cohort cell counts N_{ay} (so N_{11} is the number of exposed cases). A case-control study samples cases with probability $\pi_1 = P(S = 1 \mid Y = 1)$ and controls from the non-cases with probability $\pi_0 = P(S = 1 \mid Y = 0)$, where S indicates selection.

a) Write the expected cells of the sampled two-by-two table in terms of N_{ay}, π_1, and π_0; form the sample exposure odds ratio; and show it equals the cohort cross-product ratio $N_{11}N_{00}/(N_{10}N_{01})$. Identify the step that uses $P(S = 1 \mid Y, A) = P(S = 1 \mid Y)$.

b) Explain why "X% of our study participants experienced the outcome" describes a design decision rather than the population, and give two downstream casualties for a predictive model fit to case-control data as if randomly sampled (consider the intercept and the positive predictive value).

2.2 Question 2: The CACO OR as an Estimator of Three Targets

A closed cohort of 4,000 (2,000 exposed, 2,000 unexposed) is followed for exactly 5 years with no loss to follow-up and no competing events; 500 exposed and 200 unexposed become cases. Where person-time is needed, assume events occur uniformly, so each case contributes 2.5 years at risk on average.

a) Compute the 5-year risks, the risk ratio, the outcome odds, and the cohort odds ratio. In one sentence (foreshadowing a subsequent lecture): why is the odds ratio farther from the null here?

b) **Design 1 (cumulative sampling).** Use all 700 cases and 700 controls drawn by cumulative sampling. Compute the CACO odds ratio, and name the cohort parameter it estimates.

c) **Design 2 (case-base sampling).** Use all 700 cases and 700 controls drawn by case-base sampling. Compute the CACO odds ratio, and name the cohort parameter it estimates.

d) **Design 3 (incidence density sampling).** One control is drawn from the risk set at each case's event time. Name the cohort parameter estimated, and explain what makes the control series represent person-time. No computation needed.

2.3 Question 3: Case-cohort Weights and Variance

A prospective cohort of 30,000 adults has banked baseline serum and complete 8-year follow-up for incident stroke; 900 become cases. A biomarker costs $200 per assay. The team draws a 10% subcohort at baseline ($q = 0.10$), assays the subcohort plus all cases, and fits a weighted Cox model.

a) One subcohort member (i.e., randomly sampled into the subcohort at baseline) has a stroke in year 3. Describe what her analytic record(s) look like (write out an example of her data) and describe the weight(s) for her records. What would be over-counted if she instead carried a weight of $1/q = 10$ through to the end of her person-time, her event included?

b) Give two distinct reasons the default model-based standard errors are wrong.

2.4 Question 4: Three Different Research Questions

Suppose three teams in one health system all plan a logistic regression.

- **Team 1** wants to flag admitted patients with a high probability of one-year mortality, to offer a palliative care consultation.
- **Team 2** wants the proportion of adults with untreated hypertension in each of 40 service areas, to place two new mobile clinics.
- **Team 3** wants to know whether switching from drug A to drug B would lower one-year fracture risk, to inform a formulary change.

a) Classify each question as descriptive, predictive, or causal using the operational framework in week 4. State one assumption outside the model that each team needs for its regression output to answer their question.

b) The three `glm()` calls for each question may be line-for-line identical. Why can nothing in the fitted output reveal which team it belongs to, and where does the descriptive, predictive, or causal status of a result come from?

2.5 Question 5: Regression Machinery

a) Suppose a cohort yields the following data on a binary exposure and outcome:

	Cases ($Y = 1$)	Non-cases ($Y = 0$)	Total
Unexposed ($X = 0$)	4	16	20
Exposed ($X = 1$)	12	8	20

For the saturated logistic model logit $P(Y = 1 \mid X = x) = \beta_0 + \beta_1 x$, use the closed-form score-equation solutions from the Week 4 notes to compute $\hat{\beta}_0$ and $\hat{\beta}_1$; interpret $\exp(\hat{\beta}_1)$.

b) A colleague fits a linear model to a strongly curvilinear conditional mean; it returns a slope with a small standard error. Per the Week 4 notes on model projections: what does the slope estimate; why can a second study over a different covariate range legitimately report a different slope; and when is the projection still a scientifically useful summary?