

Assignment 4

Variance Estimation and Conditional vs. Marginal Adjustment (Weeks 7–8)

Ashley I Naimi

Due: Monday, November 9, 2026

1 Overview

This assignment covers material from Week 7 (Variance Estimation) and Week 8 (Conditional vs. Marginal Adjustment). It is worth **5 points**, allocated for completion, and is due **Monday, November 9**. 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.

1.1 Submission

Submit a single R Markdown file (`.Rmd`) through Canvas. The file must render to PDF; knit it yourself before you upload it. A template that contains the questions with an answer block under each part, `assignment4_template.Rmd`, is available on the course website alongside this handout. Use it, and save it as `assignment4_lastname.Rmd`.

Write mathematics in LaTeX between dollar signs. Where you use R, put the code in a chunk so that your reader can see both the code and the output. If you include an image, upload it with the `.Rmd`, since your file must render from exactly what you submit. No R package is required; every number you need is in the question stems. If you load any, do so in the template’s setup chunk.

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 1–2 cover Week 7; Questions 3–4 cover Week 8.

2.1 Question 1: Two Patients per Practice

A cluster randomized trial assigned ten clinical practices to a treatment or a control arm, five practices each, and enrolled about two patients per practice, twenty patients in all. The outcome is BMI at one year. An analyst fits the linear model $E(Y | X) = \beta_0 + \beta_1 X$ by ordinary least squares, where X is the treatment indicator, and obtains $\hat{\beta}_1 = 0.42$. Five standard errors for $\hat{\beta}_1$ are then computed from the same fit. All five are the sandwich $V(\hat{\beta}) = (X^\top X)^{-1} X^\top \Sigma X (X^\top X)^{-1}$ with a different rule for Σ :

Estimator	Standard error of $\hat{\beta}_1$
Model-based	1.90
HC3	2.01
Cluster-robust, by practice (HC0)	2.61
Cluster-robust, by practice (HC1)	2.68
Cluster-robust, by practice (HC3)	3.12

(The data are Table 1 of the Mansournia et al. reading, if you want to reproduce the numbers; no computation is required.)

- Say what fills Σ under the model-based estimator and under HC3 (one sentence each), and name the assumption about the twenty observations that both of them share. Why does that assumption fail in this trial? Explain why HC3 moved the standard error so little (1.90 to 2.01) while the cluster-robust estimator moved it from 1.90 to 2.61.
- The analyst then fits a cluster robust HC3 variance estimator. State what happens to the coefficients in the linear model. State specifically how the cluster robust HC3 variance estimator differs from the standard HC3 variance estimator? Which component of the sandwich variance equation changes, and how does it change?
- The analyst writes: “With the cluster-robust standard error, the treatment estimate is now unbiased and its confidence interval is valid.” Say whether the cluster-robust estimator counts people or clusters, i.e., which of the trial’s two numbers (twenty versus ten), therefore matters more for the standard error?

2.2 Question 2: A Negative Risk Ratio

In a cohort of 1,000 people, 500 exposed and 500 unexposed, the outcome occurs in 25 exposed and 10 unexposed people, so the risk ratio is $(25/500)/(10/500) = 2.5$. An analyst bootstraps it: 2,000 resamples of 1,000 rows drawn with replacement, with the risk ratio recomputed in each. The 2,000 risk ratio replicates have mean 2.76 and standard deviation 1.22, and their 2.5th and 97.5th percentiles are 1.25 and 5.72. The same resamples give log risk ratio replicates with mean 0.94 and standard deviation 0.38. Passing the risk ratio replicates to `boot.ci(type = "norm")` returns the 95% interval $(-0.15, 4.62)$; `type = "bca"` returns $(1.19, 5.61)$. Of the 2,000 risk ratio replicates, 49.6% fall below 2.5.

- The lower limit is negative. Say what the normal-interval bootstrap assumes about the estimator and why that produced an impossible value here? Name the two properties from the notes that this interval lacks.
- The percentile interval $(1.25, 5.72)$ is the same whether it is computed on the risk ratio scale or on the log scale and then exponentiated. Explain why? Could the percentile interval have crossed zero here on the risk ratio scale? Why or why not?

2.3 Question 3: One Table, Three Measures

A cohort of 800 people is cross-classified by a binary exposure X , a binary covariate Z , and a binary outcome Y , with 200 people in each (X, Z) cell. Z affects Y but has no relation to X , so it is not a confounder; it is split 50:50 in both exposure groups. Cell entries are events / people:

	$X = 0$	$X = 1$
$Z = 0$	20 / 200	50 / 200
$Z = 1$	100 / 200	150 / 200

- Compute the risk difference, risk ratio, and odds ratio within each stratum of Z and in the whole cohort (collapsed over Z). Classify each of the three measures as strictly collapsible, collapsible, or noncollapsible, saying what each term means. For the odds ratio, say which direction the overall estimate departs from the stratum-specific estimates.
- The logistic model $\text{logit } P(Y = 1 | X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$ fit to these 800 people returns $\exp(\hat{\beta}_1) = 3.00$, and its fitted risks equal the four observed risks. Marginal standardization from the same fitted model, predicting every person’s risk with X set to 1 and then to 0 and averaging each set, returns an odds ratio of 2.33. Where does this 2.33 come from, and explain how every person’s odds ratio can be 3 while the odds ratio of the averaged risks is not?

- c) Fit to the same data, the log-link model $\log P(Y = 1 | X, Z) = \gamma_0 + \gamma_1 X + \gamma_2 Z$ returns $\exp(\hat{\gamma}_1) = 1.59$, and standardizing it returns a risk ratio of exactly 1.59; the identity-link model's coefficient and its standardized risk difference are both 0.189. Explain why standardization reproduces the coefficient under these two links but not under the logit.

2.4 Question 4: Which Nuisance Do You Model?

Twenty people have a binary confounder C (ten with $C = 0$, ten with $C = 1$) and a binary exposure X . Among those with $C = 0$, two are exposed and eight unexposed; among those with $C = 1$, nine are exposed and one unexposed. So $P(X = 1 | C = 0) = 0.2$, $P(X = 1 | C = 1) = 0.9$, and $P(X = 1) = 11/20 = 0.55$.

- For each of the four (C, X) groups, compute the unstabilized and stabilized inverse probability weights. Explain in what sense C no longer predicts X in the pseudo-population, and what the stabilized weights preserve that the unstabilized ones do not.
- Now suppose the study has 40,000 people, 150 events, and 25 measured confounders, and you must choose between marginal standardization from an outcome model and IP weighting from an exposure model. Say which nuisance function each route models and which model's misspecification each is vulnerable to. Additionally, a colleague proposes a third route: a logistic model of the outcome on the exposure, adjusting for the estimated propensity score, and reading off the exposure coefficient from the model. What kind of estimand is the coefficient that arises from this propensity score adjustment method?
- The weighted contrast is estimated with `glm(y ~ x, weights = sw, family = quasibinomial("identity"))`. Why `quasibinomial` rather than `binomial`? Why are the model-based standard errors from the `quasibinomial` fit not valid? How does the bootstrap fix this problem?