Take-Home Final Exam

Advanced Regression Methods for Independent Data, Autumn 2019

Instructor: Mauricio Sadinle, Department of Biostatistics, U. of Washington – Seattle

Submit your solutions via Canvas. Due by 12:00pm (noon) on December 13, 2019.

From this exam you can get a maximum of 75 points. **Hand-written solutions will not be accepted**, so you should use some typesetting software to prepare your solution (if you haven't done so, consider learning LaTeX and R Markdown). Submit a pdf document with your solutions via Canvas.

You should not discuss questions or answers with anyone except the instructor and TAs. If you have a question about what is being asked (after reading the question carefully) please email Mauricio (msadinle@uw.edu), copying both Hongjian (hongshi@uw.edu) and Kun (yuek@uw.edu) on your message.

Bonus points: you can earn up to 15 extra points that will count towards your final grade (a potential increase of 0.1 in your final grade). The details are at the end of this document.

Model Misspecification and Biased Sampling

Consider a response variable Y, and a parametric family of distributions conditional on covariates X with conditional density

$$p^A(y \mid x, \theta) := p_\theta(y \mid x),$$

where θ are some parameters to be estimated. Let F be the true distribution of (Y, X), and $p_\star(y \mid x)$ represent the true conditional density. Assume that you have independent data $\{(Y_i, X_i)\}_{i=1}^n$. Based on this sample and the parametric model, the log-likelihood (divided by n) is given by

$$\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log p_\theta(Y_i \mid X_i),$$

and the MLE $\hat{\theta}$ maximizes $\ell_n(\theta)$.

1. (5 points) Assume your data are generated $\{(Y_i, X_i)\}_{i=1}^n \overset{i.i.d.}{\sim} F$. Show that the MLE asymp-

totically minimizes the mean conditional Kullback-Leibler divergence of p_θ from $p_\star$,

$$\mathrm{E}_X[\mathrm{KL}(p_\star, p_\theta \mid X)], \tag{1}$$

where E_X indicates an expectation with respect to the true marginal distribution of X, and

$$\mathrm{KL}(p_\star, p_\theta \mid x) = \int \log \left[\frac{p_\star(y \mid x)}{p_\theta(y \mid x)}\right] p_\star(y \mid x)\ \mathrm{d}y.$$

Call the minimizer of (1) $\theta^\star$.

2. (5 points) For this and all the problems below assume that data are generated $\{(Y_i, X_i, Z_i, S_i)\}_{i=1}^n \overset{i.i.d.}{\sim} G$, where F is the Y, X marginal of G. The new indicator variables $S_i \in \{0, 1\}$ will complicate your life. Say S_i is a selection indicator, such that if $S_i = 1$ then the draw (Y_i, X_i) actually appears in your sample, otherwise (Y_i, X_i) is not included in your sample. The Z_i are extra variables that you did not necessarily think about when you formulated your model, but they are observed for all i. Explain how working with the estimator $\hat{\theta}_1$ that maximizes

$$\ell_n^1(\theta) = \frac{1}{n_1} \sum_{i=1}^n S_i \log p_\theta(Y_i \mid X_i),$$

changes the result of problem 1., where $n_1 = \sum_{i=1}^n S_i$.

3. (5 points) Explain how working with the estimator $\hat{\theta}_\pi$ that maximizes

$$\ell_n^\pi(\theta) = \frac{1}{n_1} \sum_{i=1}^n \frac{S_i \log p_\theta(Y_i \mid X_i)}{\pi(Y_i, X_i, Z_i)},$$

asymptotically minimizes (1), where $\pi(y, x, z) = P(S = 1 \mid y, x, z)$ is known.

4. (20 points) Derive a general estimation strategy based on the estimator $\hat{\theta}_\pi$ that maximizes $\ell_n^\pi(\theta)$. Your strategy should include: an estimating equation, how to solve for $\hat{\theta}_\pi$ in that estimating equation, the asymptotic distribution of $\hat{\theta}_\pi$ based only on the assumption of i.i.d. data, and confidence intervals for the elements of $\theta^\star$.

5. The steps below represent an idealized story of how a dataset ends up in the hands of an analyst.

   - *Nature step.* First, nature generates a population of size N from a superpopulation as follows, $\{(Y_i, X_i, Z_{i1}, Z_{i2})\}_{i=1}^N \overset{i.i.d.}{\sim} H$, where $Z_{i1} \in \{0, 1\}, P(Z_{i1} = 1) = 0.5, \quad Z_{i2} \in \{0, 1\}, P(Z_{i2} = 1 \mid Z_{i1} = z) = 0.1 + 0.8z, \quad X_i \perp\!\!\!\perp Z_{i2} \mid Z_{i1}$,
   $X_i \mid Z_{i1} = z \sim \mathrm{Normal}(-1 + 2z, 1), \quad Y_i \mid \mu_i, \sigma_i \sim \mathrm{Normal}(\mu_i, \sigma_i^2)$, where the mean is $\mu_i = X_i I(Z_{i1} = Z_{i2}) - X_i I(Z_{i1} \neq Z_{i2})$, and the standard deviation is $\sigma_i = \exp(0.2|X_i|)$.

- *Data collection step.* Then, a statistical agency collects a sample from this population. They are interested in having enough data for each stratum formed by combinations of the categories of Z_1 and Z_2. Since the statistical agency has access to all Z_{i1} and Z_{i2} in the population, they know that the subpopulation where $Z_1 = Z_2$ is roughly 9 times the size of the subpopulation where $Z_1 \neq Z_2$, so they proceed to collect a sample by drawing a selection indicator S_i independently for each individual in the population using the selection mechanism $P(S_i = 1 \mid Y_i = y, X_i = x, Z_{i1} = z_1, Z_{i2} = z_2) = \tau I(z_1 = z_2) + \rho I(z_1 \neq z_2)$, $\tau = 0.01$, $\rho = 9\tau$, so that the strata formed by combinations of the categories of Z_1 and Z_2 are equally represented in the sample. They record (Y_i, X_i) for those included in the sample.
- *Naive analyst step.* Now an analyst finds out that there is this dataset made publicly available by the statistical agency, and it contains information on Y and X which are variables of interest to them. They only learned about regression under normality in their basic statistics class, so they proceed to work with the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ for their analysis, and estimate β_0 and β_1 via MLE. Furthermore, they treat the data as a simple random sample from the population.

Now you will try to bring some light into what's happening here.

(a) (5 points) Derive theoretically (give a formula) and compute or approximate numerically the value to which the naive analyst's MLE converges to, as $N \to \infty$, when computed on the data collected by the statistical agency. Call this value $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1)^T$.

(b) (5 points) Derive the true distribution of $Y_i \mid X_i$ (at the superpopulation level).

(c) (5 points) Derive theoretically (give a formula) and compute or approximate numerically the value to which the naive analyst's MLE converges to, as $N \to \infty$, when using data collected as a simple random sample from the population ($\tau = \rho$ in the selection mechanism). Call this value $\beta^\star = (\beta_0^\star, \beta_1^\star)^T$.

(d) (25 points) Now conduct a simulation study to explore the performance of the estimation strategy that you devised in Problem 4. using the data collected by the statistical agency and the model assumed by the naive analyst. In your simulation, for each population size $N \in \{10^4, 3 \times 10^4, 10^5, 3 \times 10^5, 10^6\}$, repeat the data generation process outlined in the steps above 1000 times (theoretically, a repetition includes both the generation of the population and the collection of the sample, but you can see that here we don't actually

need to generate values that won't make it into the sample), and for each repetition compute your point estimate and 95% confidence interval for $\beta_1^\star$. Report two plots: one containing the estimated actual coverage probability of your confidence interval as a function of N (plot a horizontal line at 0.95 for reference), and another plot containing the average and 2.5 and 97.5 percentiles of your point estimates as a function of N (draw a horizontal line at the true value of $\beta_1^\star$ and lines for the 2.5 and 97.5 percentiles of the asymptotic distribution of your estimator). Submit your code as an appendix, and make sure your code is well organized and commented.

Bonus Points

You have the opportunity of earning a maximum of 15 extra points that will be added to your total number of points for this class (so, a potential increase of 0.1 in your final grade, which could mean going from A- to A). Here's what you need to do: report all the typos that you have found in Wakefield's textbook or in the slides used for this class. The points will be assigned as follows: each unreported typo in chapters 1–7 in the book or in the slides is worth 15 points. The worth of each typo will be divided among the people who report it. So, if you are the only one who reports a typo, you will get those 15 points. If all 46 of you report a single typo, each of you will get 15/46 points. Typos that have been corrected in the book's online errata or by me in the slides will not count. When reporting a typo, you should also say what the correction should be, for example:

"In page X, line Y, instead of '$\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{Y}$' it should be '$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$'."

Report as many typos as you can to increase your chances of earning the extra 15 points!