STA 561: Midterm I (Due March 8 at midnight)

Reminder: unlike your homeworks and projects, this one is on your own. You are allowed to use any materials you manage to dig up from class, the internet, tea leaves, or the library. However, you cannot ask living humans, sentient robots, undead humans, non-sentient humans, or your Uncle Mo. If you're unsure if a source is acceptable, the rule you should is use is that a source is acceptable if the information in the source have been available when you acquired it if all humans but you vanished on Feb 28, i.e., you cannot post a question on stack exchange, create an anonymous website where people can post solutions (so that technically you found the info on the internet), or tattoo the answer on your dog and then upload photos of your dog to instagram[1].

Your exam should be submitted as pdf file generated using latex. Late submissions will not be accepted (special pre-arranged accommodations notwithstanding). Good luck and science.

[1] also, don't tattoo dogs, what's the matter with you?

1. (Zabka has problems.) Suppose that we observe n i.i.d. input-output pairs $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ drawn from fixed but unknown distribution P. Let $(\boldsymbol{X}, Y)$ denote a generic input-output pair and assume $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. Suppose that you friend[2] Zabka believes that the data follow a heteroskedastic model of the form

$$Y = \boldsymbol{X}^\intercal \boldsymbol{\beta}^* + \sigma(\boldsymbol{X})\epsilon,$$

where ϵ is an independent error with mean zero and variance one, and $\sigma(\boldsymbol{x}) > 0$ is a function of $\boldsymbol{x}$. Suppose that Zabka's model is correct. He makes the following claims. State whether the claim is true or false and justify your answer (e.g., by providing a short proof or derivation).

Zabka claim 1: Because of the heteroskedastic variance, the least squares estimator $\widehat{\boldsymbol{\beta}}_n = \arg\min_\beta \mathbb{P}_n(Y - \boldsymbol{X}^\intercal \boldsymbol{\beta})^2$ is biased.

Zabka claim 2: It is not possible to construct a valid Wald-type confidence interval for $\boldsymbol{\beta}^*$ if we do not know or cannot estimate the function $\sigma(\boldsymbol{x})$.

Zabka claim 3: If we used weighted least squares, say

$$\widehat{\boldsymbol{\beta}}_n^\omega = \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n(Y - \boldsymbol{X}^\intercal \boldsymbol{\beta})^2 \omega(\boldsymbol{X}),$$

for some positive real-valued function $\omega(\boldsymbol{x}) > 0$ then $\widehat{\boldsymbol{\beta}}_n^\omega$ will be biased unless $\omega(\boldsymbol{x}) \equiv 1/\sigma^2(\boldsymbol{x})$.

Extra credit: What is the optimal choice of $\omega(\boldsymbol{x})$ in this model in terms of minimizing the limiting covariance of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^\omega - \boldsymbol{\beta}^*)$? Recall that we say one a matrix $\Sigma \leq \Omega$ for matrices Σ and Ω if $\Omega - \Sigma$ is nonnegative definite.

[2] To use the term loosely.

2. (McStreamy.) Assume the same regression setup as in the previous problem. The elastic-net estimator is given by

$$\widehat{\boldsymbol{\beta}}_n^{\tau,\lambda} = \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n \left(Y - \boldsymbol{X}^\intercal \boldsymbol{\beta}\right)^2 + \tau ||\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1.$$

Show that this depends on the data only through the sufficient statistics $\mathbb{P}_n \boldsymbol{X} Y$ and $\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^\intercal$ and suggest an estimator that can be used with datasets that are too large to read into memory (you can assume that n is large and p is modest). How would you modify your algorithm to make it adaptive? (I.e., you have and adaptive ridge and adaptive lasso penalties.)

3. (It all goes wrong.) Construct a regression example in which $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^\intercal \boldsymbol{\beta}^*$ and the following hold: (i) $p = 100$, (ii) $\beta_j^* = 0$ for $j \geq 3$, (ii) the theoretical R^2 is 0.9, (iii) $n = 1000$, and (iv) running sure independent screening (SIS) to select the top 10 variables and then fitting a model to these variables has R^2 less than 0.10 more than 50% of the time in repeated samples. That is, if one were to repeat the following many times:

   - randomly draw a data set of size $n = 1000$ from your model
   - run SIS to select 10 variables
   - fit a linear model with the selected variables
   - compute the the R^2 of the fitted model,

   they would find that the R^2 is less than 0.10 more than 50% of the time.

   Suggest a new variant of SIS that would not have this problem (at least in your data set).

4. (Short and noisy.) Write adaptive ridge as a noise addition method. I.e., show that

$$\widehat{\boldsymbol{\beta}}_n^{\lambda} = \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n (Y - \boldsymbol{X}^{\intercal}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 / \left(\widehat{\beta}_{n,j}^{\mathrm{OLS}}\right)^2,$$

where $\boldsymbol{\beta}_n^{\mathrm{OLS}}$ is the OLS estimator, can be written as

$$\arg\min_{\boldsymbol{\beta}} \mathbb{P}_n \mathbb{E}_{\boldsymbol{Z}} \left\{Y - (\boldsymbol{X} + \boldsymbol{Z})^{\intercal}\boldsymbol{\beta}\right\}^2,$$

for some distribution $\boldsymbol{Z}$.

Extra credit: Consider the following iterative procedure:

(a) $\widehat{\boldsymbol{\beta}}_n^{(0)} = \widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}$

(b) $\widehat{\boldsymbol{\beta}}_n^{(k)} = \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n (Y - \boldsymbol{X}^{\intercal}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 / |\widehat{\beta}_j^{(k-1)}|$ for $k = 1, \ldots$

show that $\widehat{\boldsymbol{\beta}}_n^{(k)}$ converges to the lasso solution as $k \to \infty$.

Extra credit: Implement the preceding algorithm and compare with the output of a standard lasso package (e.g., sklearn, lars, glmnet, etc.).

5. (Thresh metal.) Suppose we observe $X_1, X_2, \ldots, X_n \sim_{i.i.d.} \text{Normal}(\mu, \sigma^2)$ and consider the adaptive hard-thresholding estimator $\widehat{\mu}_n^{\lambda_n} = \text{sgn}(\overline{X}_n)\left\{|\overline{X}_n| - \lambda_n/|\overline{X}_n|\right\}_+$ of μ. For any estimator $\widetilde{\mu}_n$, where λ_n is a tuning parameter. Define $R(\widetilde{\mu}_n, \mu) = \mathbb{E}(\mu - \widetilde{\mu}_n)^2$. Note* you do not need to compute any integrals.

   (a) Show that $R(\widehat{\mu}_n^{\lambda_n}, 0) < R(\overline{\boldsymbol{X}}_n, 0)$ provided $\lambda_n > 0$, i.e., $\widehat{\mu}_n^{\lambda_n}$ has lower expected loss at $\mu = 0$.

   (b) Suppose $\mu \neq 0$. Show that if $\lambda_n\sqrt{n} \to c$ as $n \to \infty$, where $c > 0$ is a finite constant then $\limsup_{n\to\infty} R(\widehat{\mu}_n^{\lambda_n}, \mu)/R(\overline{X}_n, \mu) \leq 1$, i.e., $\widehat{\mu}_n^{\lambda_n}$ has expected loss that is no worse than $\overline{X}_n$.

   Hint: under the condition on λ_n we have

$$\frac{\lambda_n}{|\overline{X}_n|} \quad = \quad \frac{\lambda_n\sqrt{n}}{\sqrt{n}|\overline{X}_n|}$$

   where the numerator is going to the constant c and the denominator is going to ∞ almost surely. Also recall that for any v we can write $v = \text{sgn}(v)|v|$.

   (c) From the first two parts of this problem it seems as though we have found an estimator that is superior to the mean![3] However now suppose that for each n we draw $X_{n,1}, X_{n,2}, \ldots, X_{n,n} \sim_{i.i.d.} \text{Normal}(\mu_n, \sigma^2)$ where $\mu_n = \kappa/\sqrt{n}$. Note that $\overline{X}_n = n^{-1}\sum_{i=1}^n X_{n,i} \sim \text{Normal}(\mu_n, \sigma^2/n)$ for all n. The reason for considering such sequences is that they allow use to study what happens in cases where it is difficult to distinguish the true μ from zero. Show what happens to $R(\widehat{\mu}_n^{\lambda_n}, \mu_n)/R(\overline{X}_n, \mu_n)$ as $n \to \infty$. Your answer will depend on c and κ. What does this tell you about our new estimator?

[3] In fact, as the mean attains the Cramer-Rao efficiency bound we say that $\widehat{\mu}_n^{\lambda_n}$ is super efficient. So, that's uh, super, right?