Stats 210A, Fall 2024
Homework 10

Due on: Friday, Nov. 15

Instructions: See the standing homework instructions on the course web page

Problem 1 (James-Stein estimator with regression-based shrinkage)**.** Consider estimating $\theta \in \mathbb{R}^n$ in the model $Y \sim N_n(\theta, \sigma^2 I_n)$. In the standard James-Stein estimator, we shrink all the estimates toward zero, but it might make more sense to shrink them towards the average value $\overline{Y}$ (as we explored in a previous problem) or towards some other value based on observed side information.

Suppose that we have side information about each parameter θ_i, represented by covariate vectors $x_1, \ldots, x_n \in \mathbb{R}^d$. Assume the design matrix $X \in \mathbb{R}^{n \times d}$, whose ith row is x_i', has full column rank. Suppose that we expect θ_i is not too far from $x_i'\beta$ for some $\beta \in \mathbb{R}^d$. But unlike the usual linear regression setup, we will not assume $\theta_i = x_i'\beta$ exactly, we just want to shrink our estimate toward $x_i'\beta$.

(a) Assume the error variance $\sigma^2 = 1$ is known. Find an estimator $\delta(Y)$ for θ that strictly dominates $\delta_0(Y) = Y$ whenever $n - d \geq 3$,
$$\text{MSE}(\theta; \delta) < \text{MSE}(\theta; \delta_0), \quad \text{for all } \theta \in \mathbb{R}^n,$$
and for which $\text{MSE}(X\beta; \delta) = d + 2$, for any $\beta \in \mathbb{R}^d$.

In the special case of "intercept-only" regression ($d = 1$ and $x_i = 1$ for all i), your estimator should reduce to the version of the James-Stein estimator that shrinks toward $\overline{Y}$ (but you do not have to show this).

Hint: The problem will become easier after an appropriate change of basis; think about how the estimator operates on different subspaces.

(b) Continue to assume the error variance $\sigma^2 = 1$ is known. Suppose we are unsure of whether $\theta = X\beta$ exactly. Suggest an appropriate test of the hypothesis $H_0 : \theta = X\beta$ vs $H_1 : \theta \neq X\beta$, treating $\beta \in \mathbb{R}^d$ as an unknown nuisance parameter.

(c) **Optional:** (Not graded, no extra points) Now suppose that the error variance $\sigma^2 > 0$ is unknown, but we have $r > 1$ replicates for each i; that is, we observe $Y_{i,k} \overset{\text{ind.}}{\sim} N(\theta_i, \sigma^2)$ for $i = 1, \ldots, n$ and $k = 1, \ldots, r$. Modify your test from the previous part for $H_0 : \theta = X\beta$ vs $H_1 : \theta \neq X\beta$.

Problem 2 (Confidence regions for regression)**.** Assume we observe $x_1, \ldots, x_n \in \mathbb{R}$, which are not all identical (for at least one pair i and j, $x_i \neq x_j$). We also observe
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for } \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$
$\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. Let $\bar{x}$ represent the mean value $\frac{1}{n}\sum_i x_i$.

(a) Give an explicit expression for the t-based confidence interval for β_1, in terms of a quantile of a Student's t distribution with an appropriate number of degrees of freedom (feel free to break up the expression, for example by first giving an expression for $\hat{\beta}_1$ and then using $\hat{\beta}_1$ in your final expression).

(b) Define the OLS estimator $\hat\beta = \binom{\hat\beta_0}{\hat\beta_1}$. Show that $\hat\beta \sim N_2\left(\beta,\ \sigma^2(X'X)^{-1}\right)$, for the design matrix $X = [1_n, x]$. Apply this fact to find an F-test for the hypothesis $H_0:\ \beta = 0$ vs $H_1:\ \beta \neq 0$.

(c) Invert your F-test to give a *confidence ellipse* for $\beta = \binom{\beta_0}{\beta_1}$. It may be convenient to represent the set as an affine transformation of the unit ball in $\mathbb{R}^2$:

$$b + A\mathbb{B}_1(0) = \{b + Az:\ z \in \mathbb{R}^2, \|z\| \leq 1\}, \qquad \text{for } b \in \mathbb{R}^2, A \in \mathbb{R}^{2\times 2}.$$

Give explicit expressions for b and A in terms of a quantile of an appropriate F distribution.

Problem 3 (Confidence bands for regression)**.** The setup for this problem is the same as for the previous problem only now we are interested in giving *confidence bands* for the regression line $f(x) = \beta_0 + \beta_1 x$. In this problem you do not need to give explicit expressions for everything, but you should be explicit enough that someone could calculate the bands based on your description.

(a) For a fixed value $x_0 \in \mathbb{R}$ (not necessarily one of the observed x_i values) give a $1-\alpha$ t-based confidence interval for $f(x_0) = \beta_0 + \beta_1 x_0$. That is, we want to find $C_1^P(x_0), C_2^P(x_0)$ such that

$$\mathbb{P}\left(C_1^P(x_0) \leq f(x_0) \leq C_2^P(x_0)\right) = 1-\alpha.$$

For each x_0, the coverage should be exactly $1-\alpha$. The functions $C_1^P(x), C_2^P(x)$ that we get from performing this operation on all x values give a *pointwise confidence band* for the function $f(x)$.

(b) Now give a *simultaneous confidence band* around $f(x) = \beta_0 + \beta_1 x$. That is, give $C_1^S(x), C_2^S(x)$ with

$$\mathbb{P}\left(C_1^S(x) \leq f(x) \leq C_2^S(x),\ \text{ for all } x \in \mathbb{R}\right) \geq 1-\alpha,$$

and show that your confidence band has this property.

Hint: If all we know is that β is in the confidence ellipse from the previous problem, what can we deduce about $f(x)$?

(c) Download the data set in `hw10.csv` from the course web site and make a scatter plot of the data. Plot the OLS regression line as well as the two confidence bands. Describe what you see. What do the bands do as x goes away from the data set, and why does this make sense?

(d) **Optional:** (Not graded, no extra points) Show that the coverage of the simultaneous confidence band is *exactly* $1-\alpha$, not just greater than or equal to $1-\alpha$.

Problem 4 (Precision-weighted average)**.** Suppose that we observe two independent samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} (\mu, \sigma^2)$ and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} (\mu, \tau^2)$, with $n, m > 1$. The notation means that the expectation of a single X_i or Y_i is $\mu \in \mathbb{R}$, and the variance is $\sigma^2 > 0$ for a single X_i and $\tau^2 > 0$ for a single Y_i. All three parameters are unknown, but we are primarily interested in estimating the common expectation μ.

A natural estimator is to take a convex combination of the sample averages:

$$\delta_\gamma(X, Y) = \gamma\overline{X} + (1-\gamma)\overline{Y},$$

for $\gamma \in [0, 1]$.

(a) Show that the optimal (variance-minimizing) choice of γ is

$$\gamma^* = \frac{n\sigma^{-2}}{n\sigma^{-2} + m\tau^{-2}} = \frac{1}{1 + \rho m/n},$$

where $\rho = \sigma^2/\tau^2$. δ_{γ^*} is called the *precision-weighted average* because $n\sigma^{-2}$ and $m\tau^{-2}$ are the precisions (inverse variances) of $\overline{X}$ and $\overline{Y}$, respectively. Give the variance of $\delta_{\gamma^*}(X, Y)$.

(b) Since σ^2 and τ^2 are unknown, we must estimate them. Let S_X^2 and S_Y^2 denote the usual sample variances for the two samples. Show that $\hat{\rho} = S_X^2/S_Y^2$ is a consistent estimator for ρ as $m, n \to \infty$.

Hint: It may help to recall the identity $(n-1)S_X^2 = \sum_i X_i^2 - n\overline{X}^2$.

Note: If you are wondering what it means for both m and n to go to ∞, you may assume that we have a sequence of problems indexed by $k = 1, 2, \ldots$ and $\min\{m_k, n_k\} \to \infty$ as $k \to \infty$. You should feel free to work more informally than this.

(c) Let $\hat{\gamma} = 1/(1 + \hat{\rho}m/n)$ and assume that $m, n \to \infty$ with $m/n \to c \in (0, \infty)$. Show that the adaptive estimator

$$\delta_{\hat{\gamma}}(X, Y) = \hat{\gamma}\overline{X} + (1 - \hat{\gamma})\overline{Y}$$

has an asymptotic normal distribution as $n, m \to \infty$, and give its asymptotic distribution after appropriately centering and scaling it. Compare the asymptotic distribution of the adaptive estimator $\delta_{\hat{\gamma}}(X, Y)$ to the asymptotic distribution of the oracle estimator $\delta_{\gamma^*}(X, Y)$.

Hint: Start by considering the asymptotic distribution of $(\overline{X}, \overline{Y})$. You may use without proof the result that if $Z_n \Rightarrow P$ and $W_n \Rightarrow Q$, and Z_n and W_n are independent for each n, then $(Z_n, W_n) \to P \times Q$ (meaning the product measure between the distributions P and Q).

Note: Again, if we want to set up a formal sequence of problems in which the distribution converges, we could assume the ratio $c_k = m_k/n_k$ is converging to $c \in (0, \infty)$, in addition to our previous assumption that $\min\{m_k, n_k\} \to \infty$. As before, you can also work more informally.

Problem 5 (Probabilistic big-O notation)**.** Let $X_1, X_2, \ldots$ denote a sequence of random vectors (with $\|X_n\| < \infty$ almost surely for each n). We say the sequence is *bounded in probability* (or sometimes *tight*) if for every $\varepsilon > 0$ there exists a constant $M_\varepsilon > 0$ for which

$$\mathbb{P}(\|X_n\| > M_\varepsilon) < \varepsilon, \quad \forall n.$$

Informally, there is "no mass escaping to infinity" as n grows. Like regular big-O notation, these symbols can help to make rigorous asymptotic proofs look clean and intuitive.

For a fixed sequence a_n, we say $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{p} 0$ as $n \to \infty$, and $X_n = O_p(a_n)$ if the sequence $(X_n/a_n)_{n \geq 1}$ is bounded in probability.

Prove the following facts for $X_n, Y_n \in \mathbb{R}^d$:

(a) If $X_n \Rightarrow X$ for any random vector X, then $X_n = O_p(1)$.

(b) If $X_n = o_p(a_n)$ then $X_n = O_p(a_n)$.

(c) If $X_n = O_p(a_n)$ and $Y_n = o_p(b_n)$, then $X_n'Y_n = o_p(a_n b_n)$. If $X_n = O_p(a_n)$ and $Y_n = O_p(b_n)$, then $X_n'Y_n = O_p(a_n b_n)$.

(d) If $X_n = O_p(1)$ and $g : \mathbb{R}^d \to \mathbb{R}^k$ is continuous then $g(X_n) = O_p(1)$.

(e) For $d = 1$, if $X_n = O_p(a_n)$ with $a_n \to 0$ and $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable with $g(0) = \dot{g}(0) = 0$, then $g(X_n) = o_p(a_n)$. Show further that if g is twice continuously differentiable then $g(X_n) = O_p(a_n^2)$. (**Hint:** Use the mean value theorem and apply a previous part of this problem.)

(f) For $d = 1$, if $\text{Var}(X_n) = a_n^2 < \infty$ and $\mathbb{E}X_n = 0$ then $X_n = O_p(a_n)$. (**Hint:** Use Chebyshev's inequality.)

(g) If $\text{Var}(X_n) = a_n^2 < \infty$, is it impossible to have $X_n = o_p(a_n)$? Prove or give a counterexample.