Stats305a Étude 1

Due: Monday, October 24 at 5:00pm on Gradescope.

Note: All data files available at https://web.stanford.edu/class/stats305a/Data/.

Question 1.1: Consider the typical linear regression model, where

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathsf{N}(0, \sigma^2 I). \tag{LM}$$

We consider some tests of the model (LM) based on splitting the data in half and fitting models on the two halves of the data. Assume $X \in \mathbb{R}^{2n \times d}$ with rows x_i^T and let $S_0, S_1 \subset [2n]$ be equal-sized partitions of $[2n] = \{1, \ldots, 2n\}$, so $\mathrm{card}(S_a) = n$ for $a \in \{0, 1\}$. Consider the data splits

$$X_a = \left[x_i^T\right]_{i \in S_a} \in \mathbb{R}^{n \times d}, \qquad Y_a = [Y_i]_{i \in S_a} \in \mathbb{R}^n$$

for $a \in \{0, 1\}$ so that (X_a, Y_a) corresponds to the subsets S_a. Define the estimators

$$\widehat{\beta}_0 = \operatorname*{argmin}_b \|X_0 b - Y_0\|_2^2, \quad \widehat{\beta}_1 = \operatorname*{argmin}_b \|X_1 b - Y_1\|_2^2$$

and the associated residual vectors $\widehat{\varepsilon}_a = Y_a - X_a \widehat{\beta}_a$ for $a \in \{0, 1\}$. You may assume that X_0, X_1 are both rank d matrices.

(a) Define the differences

$$\Delta := \widehat{\beta}_0 - \widehat{\beta}_1 \quad \text{and} \quad \delta := \widehat{\varepsilon}_0 - \widehat{\varepsilon}_1.$$

Give their distributions under the model (LM) and show that Δ and δ are independent.

(b) Let $B \in \mathbb{R}^{n \times d}$ be a rank d matrix with first column $\mathbf{1} \in \mathbb{R}^n$, the all-ones vector. Argue that $B(B^T B)^{-1} B^T = \frac{1}{n}\mathbf{1}\mathbf{1}^T + P$ where P is a projection matrix with $P\mathbf{1} = 0$. *Hint.* Consider the first column of Q in the QR factorization $B = QR$, $Q \in \mathbb{R}^{n \times d}$ with $Q^T Q = I_d$, $R \in \mathbb{R}^{d \times d}$.

(c) Give a symmetric PSD matrix $A \in \mathbb{R}^{d \times d}$ and any matrix $M \in \mathbb{R}^{n \times n}$ such that

$$A\Delta \sim \mathsf{N}(0, \sigma^2 I_d) \ \text{ and } \ M\delta \sim \mathsf{N}\left(0, \sigma^2 \begin{bmatrix} I_{n-r} & 0 \\ 0 & 0 \end{bmatrix}\right),$$

where r is the number of eigenvalues of $H_0 + H_1$ equal to 2, where $H_a = X_a(X_a^T X_a)^{-1} X_a^T$. *Hint.* The eigenvalues of H_a are in $\{0, 1\}$ and those of $H_0 + H_1$ are necessarily in $[0, 2]$. Consider the pseudo-inverse. It may be easier to first give a solution assuming $r = 0$.

Second hint. If you cannot find a matrix M satisfying the desired normality result, it is fine if you can give a matrix M such that $M\delta \sim \mathsf{N}(0, P)$ where P is an orthogonal projector of rank $n - r$, but do specify what P projects onto.

(d) Give a level α test for the model (LM) that uses $A\Delta$ and $M\delta$. Your test should be *pivotal* in the sense that it should work simultaneously for any $\sigma^2 > 0$ (i.e., it should *not* depend on σ^2).

(e) We are interested in departures from the model (LM) that involve heteroskedasticity—when the variance σ^2 depends on the index i—or nonlinearity, where

$$Y = X\beta + \eta + \varepsilon,$$

where $\eta \in \mathbb{R}^{2n}$ is the "nonlinear" effect.

One potential way to detect such issues is to find splits S_0, S_1 of the data that might plausibly identify them. (There are other ways to do this as well.) Consider two such splitting techniques:

i. Choose a vector $v \in \mathbb{R}^d$ uniformly on the sphere $\|v\|_2 = 1$, and then set S_0 to be the indices i satisfying $x_i^T v \leq \text{Median}(\{v^T x_j\}_{j=1}^{2n})$ and $S_1 = S_0^c$ to be its complement.

ii. Choose S_0 and S_1 uniformly at random (but of course with $S_1 = S_0^c$).

For the data file `maybe-its-nonlinear.csv`, compare the strategies i. and ii. above for your test in part (d). The last column is y and the rest of the columns form the X data matrix. Repeat the following 1000 times: construct the random splits S_0, S_1 that i. and ii. identify, then perform your test (you may assume that $r = 0$ in (c), so $H_0 + H_1$ has eigenvalues strictly smaller than 2). Give the fraction of rejections for each of the two splits for your method. Include your code.