Stats305a Étude 3

Due: Monday, November 28 at 5:00pm on Gradescope.

Question 3.1 (Stability, instability, and robustness)**:** In this question, we develop a leave-one-out robustness heuristic and investigate the (in)stability of non-discoveries in regression.

(a) Let A, B be square matrices where A is invertible.

(i) **2 pts.** Show that for any $k \in \mathbb{N}$,

$$(A - B)\left(A^{-1} + A^{-1}BA^{-1} + \cdots + A^{-1}(BA^{-1})^k\right) = I - (BA^{-1})^{k+1}.$$

Hint. Use induction.

(ii) **2 pts.** Recall that the operator norm $\|A\|_{\text{op}} = \sup_{u,v}\{u^T A v \mid \|u\| = \|v\| = 1\}$. Assume that $\|B\|_{\text{op}} < 1/\|A^{-1}\|_{\text{op}}$. Show that

$$(A - B)^{-1} = \sum_{i=0}^{\infty} A^{-1}(BA^{-1})^i \quad \text{and} \quad (A + B)^{-1} = \sum_{i=0}^{\infty}(-1)^i A^{-1}(BA^{-1})^i.$$

Consider a typical regression setting with data $X \in \mathbb{R}^{n\times d}$ (whose rows are x_i^T as usual) and $y \in \mathbb{R}^n$, and for a vector $\boldsymbol{\delta} \in [0,1]^n$, define the estimator

$$\widehat{\beta}_{\boldsymbol{\delta}} := \operatorname*{argmin}_b \frac{1}{n}\sum_{i=1}^{n}(1 - \delta_i)(x_i^T b - y_i)^2,$$

which omits examples with $\delta_i = 1$ and downweights those with $\delta_i > 0$. Then the standard OLS estimator is $\widehat{\beta} = \widehat{\beta}_{\mathbf{0}}$, and we let $\widehat{y} = X\widehat{\beta}$ be the usual prediction for y.

(b) **2 pts.** Let $D(\boldsymbol{\delta}) = \text{diag}(\boldsymbol{\delta})$ be the diagonal matrix with ith diagonal entry δ_i. Show that

$$\widehat{\beta}_{\boldsymbol{\delta}} = (X^T(I - D(\boldsymbol{\delta}))X)^{-1}X^T(I - D(\boldsymbol{\delta}))Y.$$

We say a matrix $A = O(r)$ (read "A is big-O of r") if $\|A\|_{\text{op}}/r$ is bounded as $r \to 0$. For example, in part (a) above, the matrix $(A - B)\sum_{i=0}^{k} A^{-1}(BA^{-1})^k = O(\|B\|^{k+1})$ for B near 0. The matrix $X^T D(\boldsymbol{\delta})X$ is $O(\|\boldsymbol{\delta}\|)$, and $AD(\boldsymbol{\delta})BD(\boldsymbol{\delta})C$ is $O(\|\boldsymbol{\delta}\|^2)$ for any matrices A, B, C as $\boldsymbol{\delta} \to 0$.

(c) **3 pts.** Let $\widehat{C} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T = \frac{1}{n}X^T X$ be the "covariance" of the x_i. Show that

$$\widehat{\beta}_{\boldsymbol{\delta}} = \widehat{\beta} + \frac{1}{n}\widehat{C}^{-1}X^T D(\boldsymbol{\delta})(\widehat{y} - y) + O(\|\boldsymbol{\delta}\|^2)$$

(d) **2 pts.** Show that if e_i is the ith standard basis vector, then

$$\lim_{\delta\to 0}\frac{\widehat{\beta}_{\delta e_i} - \widehat{\beta}}{\delta} = \frac{1}{n}\widehat{C}^{-1}x_i(\widehat{y}_i - y_i).$$

We define $\mathsf{iinf}(\widehat{\beta}, i) := \frac{1}{n}\widehat{C}^{-1}x_i(\widehat{y}_i - y_i)$ to be the *instantaneous influence* of example i, as it (roughly) captures the effect of removing example i from the dataset.

(e) **6 pts.** Download the UCI Abalone dataset (https://archive.ics.uci.edu/ml/datasets/Abalone). It consists of 9 features (one of which, `Sex`, is nominal and takes values `M`, `F`, and `I` for infant, so you should transform it into a one-hot encoding, but make sure you don't accidentally make your design low rank), and the last of which (`rings`) is the attribute to predict (i.e., y). We would like to investigate *how* significant we can make a p-value by removing just a few (in this case, k) datapoints. You should do this by the following, which we describe for a fixed k.

i. Construct the design $X \in \mathbb{R}^{n \times d}$ and response y from the dataset. Standardize X so that the columns have ℓ_2-norm $\sqrt{n}$ and are (except for the intercept) mean 0.

ii. For each coordinate $j \in \{1, \ldots, d\}$ and each datapoint $i \in \{1, \ldots, n\}$, compute the instantaneous influence of example i on parameter j, which is

$$e_j^T \mathsf{iinf}(\widehat{\beta}, i).$$

iii. For each coordinate $j \in \{1, \ldots, d\}$, choose the index set $\mathcal{I} \subset \{1, \ldots, n\}$ of cardinality k maximizing the (estimated) $\widehat{\beta}$ that results after removing those $\mathcal{I}$ examples, that is, maximizing

$$\left| e_j^T \left(\widehat{\beta} + \sum_{i \in \mathcal{I}} \mathsf{iinf}(\widehat{\beta}, i) \right) \right|.$$

iv. For each of these index sets (you should compute per coordinate j of $\widehat{\beta}$), set $\boldsymbol{\delta} = \sum_{i \in \mathcal{I}} e_i$, that is, an n vector with a 1 in each position corresponding to $\mathcal{I}$, and compute $\widehat{\beta}_{\boldsymbol{\delta}}$ and the corresponding p-value that a T-test yields for the null

$$H_{0,j} : \beta_j = 0$$

in the model $y_i = x_i^T \beta + \varepsilon_i$, $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$ for $i \notin \mathcal{I}$, that is, the linear model *without* examples in $\mathcal{I}$. (You do not need to do any multiplicity correction.)

Repeat steps i–iv above for index set sizes $k = 1, \ldots, 20$, and for each of your coordinates j, plot the resulting p values (that is, plot k on the horizontal axis, beginning from $k = 0$, against p on the vertical axis; and yes, $k = 0$ corresponds to doing things with the initial data). How much can you modify the "significance" of your results by removing a few datapoints? You may want to plot your results on a logarithmic scale. Include your code.