HA 1
Due on Thursday, February 18, 15:00

Details:

(a) Students can work in **groups** of maximum **2** people. You need to submit one HA per group.

(b) All computation should be done in R. If you want to use other languages (Matlab or Python), you need to contact me and get an approval for that.

(c) You can use built-in functions, unless I explicitly ask you to write your own functions.

(d) I encourage you to discuss all the problems with your classmates (although, for the most part they are quite trivial), but you should write your own write-up. Cheating won't be tolerated.

(e) You can either submit your answers electronically or just hand them in before the lecture on Thursday.

Problem 1 (Simulation problem)

Consider setup from the first lecture: we have N numbers $Y = \{y_1, \ldots, y_N\}$ and observe n of them $\tilde{Y} = \{y_{I(1)}, \ldots, y_{I(n)}\}$. Our goal is to say something about $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$. We know that all the numbers belong to $[0, 100]$.

(a) Let $N = 100$ and let $n = 20$. Generate N random numbers from $U[0, 100]$ and use them as set Y, then take first n of them and use them for the set $\tilde{Y}$. Report μ that you have for you set Y.

(b) Build the worst case interval from the first lecture. Report its width and center. Compare the center with μ from (1).

(c) Assume that numbers from $Y \setminus \tilde{Y}$ are distributed as $100 * \text{Beta}(2,2)$ i.i.d. random variables. Calculate $\mathbb{E}[\mu]$ and $\mathbb{V}[\mu]$ under this assumption. Calculate probability that μ belongs to the following interval:

$$\left[\mathbb{E}[\mu] - 2\sqrt{\mathbb{V}[\mu]}, \mathbb{E}[\mu] + 2\sqrt{\mathbb{V}[\mu]}\right] \tag{1}$$

You can use simulations to answer this question (e.g., take $B = 100$ random samples) or you can give some analytical bounds. *Bonus question*: Estimate standard error of your answer.

(d) Consider three probability models: $100 * \text{Beta}(0.5.0.5)$, $100 * \text{Beta}(1,1)$ and $100 * \text{Beta}(2,2)$. Using the algorithm described in the lecture calculate likelihood for each of these models and construct probabilities. Report these probabilities. Calculate the mean and the variance of μ under the assumption that we first select the model at random and then data in $Y \setminus \tilde{Y}$ are generated from this model. Build the same interval as in (c) and estimate probability that μ belongs to this interval.

Problem 2 (Simple theoretical problems)

(a) Draw a picture with two risk functions such that the following conditions hold (simultaneously):

(1) Neither of them dominates each other on the whole domain.

(2) The first one dominates the second one on the restricted domain.

(3) The second one has lower maximal risk over the whole domain.

(b) Take the fact that $f(X) := \mathbb{E}[Y|X] = \arg\min_{h \in F} \mathbb{E}[(Y - h(X))^2]$ as given and prove that $\mathbb{E}[(Y - f(X))g(X)] = 0$ for any function g.*

(c) Prove that for any set of functions $\mathcal{G}$, $g(X) := \arg\min_{h \in \mathcal{G}} \mathbb{E}[(Y - h(X))^2]$ also solves the problem $\min_{h \in \mathcal{G}} \mathbb{E}[(f(X) - h(X))^2]$, where $f(X) := \mathbb{E}[Y|X]$.

---

*Assume any integrability conditions you need to guarantee that all expectations are well defined.

(d) Let $g(X) := \arg\min_{h \in \mathcal{G}} \mathbb{E}[(Y - h(X))^2]$ and let $f(X) := \mathbb{E}[Y|X]$. Let $\|h_1 - h_2\|_2^2 := \mathbb{E}[(h_1(X) - h_2(X))^2]$. Find reasonable conditions on $\mathcal{G}$ under which we have the following identity:

$$\|f - \hat{f}\|_2^2 = \|f - g\|_2^2 + \|g - \hat{f}\|_2^2 \tag{2}$$

for any (fixed) function $\hat{f} \in \mathcal{G}$

(e) Assume that we have data vector Y and matrix X (n-dimensional vector and $n \times p$ matrix, $p < n$). Assume that $X^T X = \mathcal{I}_p$, let $\hat{\beta}$ be the OLS estimator. Prove the following:

(1) $\hat{\beta}_k^{l_0} = \{|\hat{\beta}_k^{l_0}| > \lambda\}\hat{\beta}_k$

(2) $\hat{\beta}_k^{l_1} = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_k| - \lambda)_+$

(3) $\hat{\beta}_k^{l_2} = \frac{\hat{\beta}_k}{1+\lambda}$

where $\hat{\beta}^{l_q} = (\hat{\beta}_1^{l_q}, \ldots \hat{\beta}_p^{l_q})$ is defined as the solution to the following problem:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2 + \lambda P_q(\beta) \to \min_{\beta} \tag{3}$$

where P_q corresponds to a penalty function for l_q norm (see lecture 3).

Problem 3 (Cross-validation and C_p)

(a) Assume that $\hat{y} = Sy$ and prove that $\sum_{i=1}^{n} \text{cov}(y_i, \hat{y}_i) = \text{trace}(S)$

(b) Assume that we are using linear fitting procedure: $\hat{y} = Sy$. Let $\hat{y}^{(i)} = S^{(i)}y^{(i)}$ be the results of this procedure if we drop i-th observation.

(1) Assume that $\hat{y}_i^{(i)} = S_{ii}\hat{y}_i^{(i)} + \sum_{j \neq i} S_{ij}y_j$ and prove that $y_i - \hat{y}_i^{(i)} = \frac{y_i - \hat{y}_i}{1 - S_{ii}}$. Explain why this is useful for LOOCV.

(2) Prove that assumption in (1) is valid for S arising from OLS.

(3) Prove that assumption in (1) is valid for S arising from ridge regression.

(4) Can you characterize linear procedures (at least informally) for which this assumption should hold?

(c) Assume that $S_{ii} \approx$ const and that n is large, so that S_{ii} is small. Using first order Tailor expansion show that LOOCV is similar to C_p statistics. What is the difference between them?

Problem 4 (Computational exercise)

Download the file *ha_1.txt*: it contains $N = 1000$ observations on $p = 200$ covariates and one outcome variable. In this exercise you need to do the following (in R):

(a) Randomly separate data into two pieces: training set (approximately 70% of observations) and test set (the rest). Standardize (separately) both sets.

(b) Write a function that will estimate OLS (on the training set). Estimate residual variance, using unbiased estimator. Estimate prediction risk using LOOCV and using training error. Compare and comment.

(c) Construct a sequence $\lambda = (\lambda_1, \ldots, \lambda_{100})$ such that degrees of freedom for ridge regression decrease from 200 to 1.

(d) Write a unction that will estimate ridge regression for a given value of λ_k. Estimate it for each λ_j that you constructed in (c).

(e) For each λ_k estimate prediction risk using LOOCV (efficiently!). Estimate prediction risk with C_p statistics, using estimate of variance from (b). Plot both estimated risks (on one graph) as a function of degrees of freedom. Comment on similarities and differences. Which λ_i is the best?

(f) Install package *lars* and use it to estimate lasso. Use built-in function for 10-fold CV to find the best model and report non-zero coefficients and number of zeros. Plot the path for lasso and estimated prediction risk.

(g) Based on your results from (b)-(f) what model would you use? Estimate prediction risk for the selected model using the test set. Compare is with the estimated prediction risk you obtained on the training set.