STAT 542: Homework 1

Due: Feb. 9 midnight on Canvas

Please make sure that your solutions are readable and the file size is reasonable. Typing the answers is highly encouraged.

Problem 1.

Suppose that the true observation model is given by

$$Y = X\beta + \epsilon$$

where $X \in \mathbb{R}^{n\times 2}$, and ϵ satisfies $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}\left[\epsilon\epsilon^\top\right] = \sigma^2 I$. Further assume that the $X_1, X_2 \in \mathbb{R}^n$ are the two columns of X, $\|X_1\|_2 = \|X_2\|_2 = 1$, and the inner product $\langle X_1, X_2\rangle = r$. Denote by

$$\hat{\beta} := \left(X^\top X\right)^{-1} X^\top Y$$

and OLS estimator using the full model, and

$$\hat{\beta}^r := \left(X_1^\top X_1\right)^{-1} X_1^\top Y$$

the OLS estimator using the reduced model.

- [1 pts] Suppose that we are only interested in estimating the first coordinate, β_1. Compute $\mathbb{E}\left[\hat{\beta}_1\right]$ and $\operatorname{var}\left(\hat{\beta}_1\right)$ (express the answers using β, σ and r).
- [2 pts] Compute $\mathbb{E}\left[\hat{\beta}_1^r\right]$ and $\operatorname{var}\left(\hat{\beta}_1^r\right)$.
- [2pts] Use the bias-variance tradeoffs to compute the mean square errors of $\hat{\beta}_1$ and $\hat{\beta}_1^r$ (defined as $\mathbb{E}\left[\left|\hat{\beta}_1 - \beta_1\right|^2\right]$ and $\mathbb{E}\left[\left|\hat{\beta}_1^r - \beta_1\right|^2\right]$). Find the range of β_2 for which the reduced model has a smaller mean square error than the full model. Hint: note that when $|\beta_2|$ is large, you would expect that the full model is "more correct" and hence having a smaller error.

Solution to part 1

Given the model $Y = X\beta + \epsilon$, the OLS estimator is $\hat{\beta} = (X^T X)^{-1} X^T Y$. We want to compute $E[\hat{\beta}_1]$ and $\text{var}(\hat{\beta}_1)$.

First, for the expected value $E[\hat{\beta}_1]$:

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ &= \beta + (X^T X)^{-1} X^T \epsilon\end{aligned}$$

Taking the expectation on both sides, we get:

$$\begin{aligned}E[\hat{\beta}] &= E[\beta + (X^T X)^{-1} X^T \epsilon] \\ &= \beta + (X^T X)^{-1} X^T E[\epsilon]\end{aligned}$$

Since $E[\epsilon] = 0$, it follows that:

$$E[\hat{\beta}] = \beta$$

Therefore, $E[\hat{\beta}_1] = \beta_1$.

Next, for the variance $\text{var}(\hat{\beta}_1)$:

$$\begin{aligned}\text{var}(\hat{\beta}) &= \text{var}((X^T X)^{-1} X^T \epsilon) \\ &= (X^T X)^{-1} X^T \text{var}(\epsilon) X (X^T X)^{-T} \\ &= \sigma^2 (X^T X)^{-1}\end{aligned}$$

Given $X^T X = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$, its inverse is:

$$(X^T X)^{-1} = \frac{1}{1 - r^2} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}$$

Thus, $\text{var}(\hat{\beta}_1)$ is the first element of $\sigma^2 (X^T X)^{-1}$:

$$\text{var}(\hat{\beta}_1) = \sigma^2 \cdot \frac{1}{1 - r^2}$$

In conclusion:

- $E[\hat{\beta}_1] = \beta_1$
- $\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{1 - r^2}$

Part Two

For the reduced model estimator $\hat{\beta}_{r1}$, given by:

$$\hat{\beta}_{r1} = X_1^T Y$$

we compute $E[\hat{\beta}_{r1}]$ and $var(\hat{\beta}_{r1})$ using the provided assumptions and given that $Y = X_1 \beta_1 + X_2 \beta_2 + \epsilon$.

Expectation of $\hat{\beta}_{r1}$

We have:

$$\begin{aligned}\hat{\beta}_{r1} &= X_1^T(X_1\beta_1 + X_2\beta_2 + \epsilon) \\ &= X_1^T X_1\beta_1 + X_1^T X_2\beta_2 + X_1^T\epsilon \\ &= \beta_1 + r\beta_2 + X_1^T\epsilon\end{aligned}$$

Taking the expectation of both sides, given $E[\epsilon] = 0$, we obtain:

$$E[\hat{\beta}_{r1}] = \beta_1 + r\beta_2$$

Variance of $\hat{\beta}_{r1}$

The variance is computed as follows:

$$\begin{aligned}var(\hat{\beta}_{r1}) &= var(\beta_1 + r\beta_2 + X_1^T\epsilon) \\ &= var(X_1^T\epsilon) \\ &= X_1^T var(\epsilon) X_1 \\ &= \sigma^2 X_1^T X_1 \\ &= \sigma^2\end{aligned}$$

Therefore, the variance of $\hat{\beta}_{r1}$ is σ^2.

In conclusion, for the reduced model we have:

- $E[\hat{\beta}_{r1}] = \beta_1 + r\beta_2$
- $var(\hat{\beta}_{r1}) = \sigma^2$

Part Three

To determine the range of β_2 for which the reduced model has a smaller mean square error (MSE) than the full model, we compare the MSEs of $\hat{\beta}_1$ and $\hat{\beta}_{r1}$.

The mean square error of an estimator $\hat{\theta}$ is given by:

$$MSE(\hat{\theta}) = \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta}, \theta)$$

For the full model estimator $\hat{\beta}_1$, the MSE is:

$$MSE(\hat{\beta}_1) = \text{var}(\hat{\beta}_1) + \text{bias}^2(\hat{\beta}_1, \beta_1)$$

$$MSE(\hat{\beta}_1) = \frac{\sigma^2}{1 - r^2}$$

since the estimator is unbiased.

For the reduced model estimator $\hat{\beta}_{r1}$, the MSE is:

$$MSE(\hat{\beta}_{r1}) = \text{var}(\hat{\beta}_{r1}) + \text{bias}^2(\hat{\beta}_{r1}, \beta_1)$$

$$MSE(\hat{\beta}_{r1}) = \sigma^2 + (r\beta_2)^2$$

We want to find the range of β_2 such that $MSE(\hat{\beta}_{r1}) < MSE(\hat{\beta}_1)$:

$$\sigma^2 + (r\beta_2)^2 < \frac{\sigma^2}{1-r^2}$$

$$(r\beta_2)^2 < \frac{\sigma^2}{1-r^2} - \sigma^2$$

$$r^2\beta_2^2 < \frac{\sigma^2 r^2}{1-r^2}$$

$$\beta_2^2 < \frac{\sigma^2}{1-r^2}$$

Taking the square root of both sides gives the range for β_2:

$$-\frac{\sigma}{\sqrt{1-r^2}} < \beta_2 < \frac{\sigma}{\sqrt{1-r^2}}$$

Thus, the reduced model has a smaller MSE than the full model when β_2 lies within this range.

Problem 2.

Use R or Python to perform the following experiment: you pick arbitrary numbers $\rho \in (0,1)$ and $r \in (0,1)$ satisfying

$$\frac{6\rho}{1+\rho^2} > \frac{1}{r} + 2r$$

Set $X = \begin{pmatrix} 1 & \rho r \\ \rho & r \end{pmatrix}$ and $Y = X \begin{pmatrix} -1 \\ 2 \end{pmatrix}$. For any $\lambda > 0$, define

$$\hat{\beta}_\lambda := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

Plot the coefficients of $\hat{\beta}_\lambda$ as a function of $\left\|\hat{\beta}_\lambda\right\|_1$, and repeat the experiments with different ρ and r satisfying (4). Include the plots in your solution. Do you find $\left\|\hat{\beta}_\lambda\right\|_0$ to be a monotonic function of the ℓ_1 norm or not? What is the implication of this phenomenon for implementing the LARS algorithm?

Hint: lasso.R in Canvas contains most of the ingredients of the code. Note that using R code you can easily plot the lasso coefficients with the L_1 norm (see slides). Also, beware that the default options for the intercept and feature normalizations of the R function may not be what you want.

Problem 3.

Generate a design matrix $X \in \mathbb{R}^{100\times 200}$ and let $\beta \in \mathbb{R}^{200}$ be defined as

$$\beta_j = 1\{j \leq 30\}, \quad j = 1, \ldots, 200$$

where 1{} denotes the indicator function. In the model $Y = X\beta + \epsilon, \epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$, compute the optimal lasso regularization parameter λ_{opt} using cross-validation by R (Caution: no intercept and column normalization). Study the trend of λ_{opt} as σ varies, by plotting a figure showing their dependence.