Lab 14

Machine Learning 2021-2022 - UMONS
Souhaib Ben Taieb

1

Consider the problem of multiple linear regression, where the aim is to find $\hat{\beta^{LS}} \in \mathbb{R}^p$ such that:

$$\hat{\beta}^{LS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

We can show that the ordinary least squares estimate is given by $\hat{\beta}^{LS} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y}$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$. Now consider the problem of ridge regression, where the optimization problem is now formulated as finding $\hat{\beta}^R \in \mathbb{R}^p$ such that:

$$\hat{\beta}^{R} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

The solution is given by $\hat{\beta}^R = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y}$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix. Assuming that $p = n$ and that $\mathbf{X} = \mathbf{I}_p$,

- Prove that $\hat{\beta}^R = \frac{\hat{\beta}^{LS}}{\lambda+1}$.
- Assuming that $\text{Bias}(\hat{\beta}^{LS}) = 0$, compute the bias of the ridge estimator.
- Derive the covariance matrices of $\hat{\beta}^{LS}$ and $\hat{\beta}^R$ and show how they relate to one another. Assume that $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2 \mathbf{I}_n$, where $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a random noise vector (i.e. $\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$).

The covariance matrix of a vector $\boldsymbol{a} \in \mathbb{R}^p$ is given by $\text{Cov}(\boldsymbol{a}) = \mathbb{E}\left[(\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}])(\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}])^\mathsf{T}\right] \in \mathbb{R}^{p \times p}$

Solution :

$$\begin{aligned}
\hat{\beta}^{LS} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y} \\
&= (\mathbf{I}_p)^{-1}\mathbf{I}_p\boldsymbol{y} \\
&= \boldsymbol{y}
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}^{R} &= (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y} \\
&= (\mathbf{I}_p + \lambda\mathbf{I}_p)^{-1}\mathbf{I}_p\boldsymbol{y} \\
&= \Big((\lambda+1)\mathbf{I}_p\Big)^{-1}\mathbf{I}_p\boldsymbol{y} \\
&= \frac{1}{\lambda+1}\mathbf{I}_p\mathbf{I}_p\boldsymbol{y} \\
&= \frac{\boldsymbol{y}}{\lambda+1} \\
&= \frac{\hat{\beta}^{LS}}{\lambda+1}
\end{aligned}$$

The bias of the ridge estimator is obtained as:

$$\begin{aligned}
\text{Bias}(\hat{\beta}^{R}) &= \mathbb{E}[\hat{\beta}^{R}] - \beta \\
&= \mathbb{E}\left[\frac{\hat{\beta}^{LS}}{\lambda+1}\right] - \beta \\
&= \frac{1}{\lambda+1}\mathbb{E}[\hat{\beta}^{LS}] - \beta \\
&= \frac{\beta}{\lambda+1} - \beta
\end{aligned}$$

Thus $\hat{\beta}^{R}$ is a biased estimator of the true coefficients β if $\lambda \neq 0$.

The covariance matrix of the ordinary least squares estimates is obtained as:

$$\begin{aligned}
\text{Cov}(\hat{\beta}^{LS}) &= \mathbb{E}\Big[(\hat{\beta}^{LS} - \mathbb{E}[\hat{\beta}^{LS}])(\hat{\beta}^{LS} - \mathbb{E}[\hat{\beta}^{LS}])^{\intercal}\Big] \\
&= \mathbb{E}\Big[(\hat{\beta}^{LS} - \beta)(\hat{\beta}^{LS} - \beta)^{\intercal}\Big] \\
&= \mathbb{E}\Big[\hat{\beta}^{LS}(\hat{\beta}^{LS})^{\intercal} - \hat{\beta}^{LS}\beta^{\intercal} - \beta(\hat{\beta}^{LS})^{\intercal} + \beta\beta^{\intercal}\Big] \\
&= \mathbb{E}\Big[\hat{\beta}^{LS}(\hat{\beta}^{LS})^{\intercal}\Big] - \mathbb{E}\Big[\hat{\beta}^{LS}\beta^{\intercal}\Big] - \mathbb{E}\Big[\beta(\hat{\beta}^{LS})^{\intercal}\Big] + \mathbb{E}\Big[\beta\beta^{\intercal}\Big] \\
&= \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^{\intercal}] - \beta\beta^{\intercal} - \beta\beta^{\intercal} + \beta\beta^{\intercal} \\
&= \mathbb{E}\Big[(\mathbf{X}\beta + \boldsymbol{\varepsilon})(\mathbf{X}\beta + \boldsymbol{\varepsilon})^{\intercal}\Big] - \beta\beta^{\intercal} \\
&= \mathbb{E}[\mathbf{X}\beta\beta^{\intercal}\mathbf{X}^{\intercal}] + \mathbb{E}[\mathbf{X}\beta\boldsymbol{\varepsilon}^{\intercal}] + \mathbb{E}[\boldsymbol{\varepsilon}\beta^{\intercal}\mathbf{X}^{\intercal}] + \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{T}] - \beta\beta^{\intercal} \\
&= \beta\beta^{\intercal} + \sigma^{2}\mathbf{I}_{p} - \beta\beta^{\intercal} \\
&= \sigma^{2}\mathbf{I}_{p}
\end{aligned}$$

The covariance matrix of the ridge estimates are given by:

$$\begin{aligned}
\text{Cov}(\hat{\beta}^{R}) &= \mathbb{E}\Big[(\hat{\beta}^{R} - \mathbb{E}[\hat{\beta}^{R}])(\hat{\beta}^{R} - \mathbb{E}[\hat{\beta}^{R}])^{\intercal}\Big] \\
&= \mathbb{E}\Big[\Big(\frac{\hat{\beta}^{LS}}{\lambda+1} - \frac{\beta}{\lambda+1}\Big)\Big(\frac{\hat{\beta}^{LS}}{\lambda+1} - \frac{\beta}{\lambda+1}\Big)^{\intercal}\Big] \\
&= \frac{1}{(\lambda+1)^{2}}\Big(\mathbb{E}[\hat{\beta}^{LS}(\hat{\beta}^{LS})^{\intercal}] - \mathbb{E}[\hat{\beta}^{LS}\beta^{\intercal}] - \mathbb{E}[\beta(\hat{\beta}^{LS})^{\intercal}] + \mathbb{E}[\beta\beta^{\intercal}]\Big) \\
&= \frac{1}{(\lambda+1)^{2}}(\sigma^{2}\mathbf{I}_{p} + \beta\beta^{\intercal} - \beta\beta^{\intercal} - \beta\beta^{\intercal} + \beta\beta^{\intercal}) \\
&= \frac{\sigma^{2}\mathbf{I}_{p}}{(\lambda+1)^{2}} \\
&= \frac{\text{Cov}(\hat{\beta}^{LS})}{(\lambda+1)^{2}}
\end{aligned}$$

2

Suppose that the columns of $\mathbf{X}_1$ are orthonormal, and that $\mathbf{X}_2 = 10\mathbf{X}_1$. Show that the ordinary least squares estimates are equivariant, meaning that multiplying $\mathbf{X}$ by a constant c scales the coefficients estimates by a factor $\frac{1}{c}$.

Is it also the case for the ridge estimates ?

Solution :

$$\begin{aligned}
\hat{\beta}^{LS} \mid \mathbf{X}_2 &= (\mathbf{X}_2^\intercal \mathbf{X}_2)^{-1} \mathbf{X}_2^\intercal \boldsymbol{y} \\
&= (100\mathbf{X}_1^\intercal \mathbf{X}_1)^{-1} 10\mathbf{X}_1^\intercal \boldsymbol{y} \\
&= \frac{10}{100} (\mathbf{X}_1^\intercal \mathbf{X}_1)^{-1} \mathbf{X}_1^\intercal \boldsymbol{y} \\
&= \frac{1}{10} \hat{\beta}^{LS} \mid \mathbf{X}_1
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}^{R} \mid \mathbf{X}_2 &= (\mathbf{X}_2^\intercal \mathbf{X}_2 + \lambda \mathbf{I}_p)^{-1} \mathbf{X}_2^\intercal \boldsymbol{y} \\
&= (100\mathbf{X}_1^\intercal \mathbf{X}_1 + \lambda \mathbf{I}_p)^{-1} 10\mathbf{X}_1^\intercal \boldsymbol{y} \\
&= (100\mathbf{I}_p + \lambda \mathbf{I}_p)^{-1} 10\mathbf{X}_1^\intercal \boldsymbol{y} \\
&= \frac{10}{100+\lambda} \mathbf{I}_p^{-1} \mathbf{X}_1^\intercal \boldsymbol{y} \\
&= \frac{10}{100+\lambda} \mathbf{X}_1^\intercal \boldsymbol{y} \\
&\neq \frac{1}{1+\lambda} \mathbf{X}_1^\intercal \boldsymbol{y} = \hat{\beta}^{R} \mid \mathbf{X}_1
\end{aligned}$$

The ridge estimates are not equivariant, meaning that they are sensitive to the scale of the data !