Exercises week 37 - Markus Bjørklund

September 22, 2023

1 Exercise 1

Our original assumption is that $\mathbf{y}$ is given by

$$\mathbf{y} = f(x) + \epsilon, \tag{1}$$

where $f(x)$ is a continuous function and ϵ is Gaussian distributed. Now, we approximate this function $f(x)$ by $f(x) \approx \tilde{\mathbf{y}} = \mathbf{X}\beta$, such that

$$\mathbf{y} \approx \mathbf{X}\beta + \epsilon, \tag{2}$$

or

$$y_i \approx \sum_j X_{ij}\beta_j + \epsilon_i = \mathbf{X}_i \cdot \beta + \epsilon_i. \tag{3}$$

Now, if we take the expectation value of this, we get

$$\begin{aligned}
\mathbb{E}\left[y_i\right] &\approx \mathbb{E}\left[\mathbf{X}_i \cdot \beta\right] + \mathbb{E}\left[\epsilon_i\right] \\
&= \mathbf{X}_i \cdot \beta,
\end{aligned}$$

because $\mathbf{X}$ and β are assumed to be non-stochastic variables (enforced by design), and $\epsilon = 0$ by the properties of a normal distribution with zero mean.

For the variance, we have that

$$\begin{aligned}
Var\left[y_i\right] &= \mathbb{E}\left[\left(y_i - \mathbb{E}\left[y_i\right]\right)^2\right] \\
&= \mathbb{E}\left[y_i^2 - 2y_i\mathbb{E}\left[y_i\right] + \mathbb{E}\left[y_i\right]^2\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}_i \cdot \beta + \epsilon_i\right)^2 - 2\left(\mathbf{X}_i \cdot \beta + \epsilon_i\right)\mathbf{X}_i \cdot \beta + \left(\mathbf{X}_i \cdot \beta\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}_i \cdot \beta\right)^2 + 2\mathbf{X}_i \cdot \beta\epsilon + \epsilon^2 - 2\left(\mathbf{X}_i \cdot \beta\right)^2 - 2\mathbf{X}_i \cdot \beta\epsilon + \left(\mathbf{X}_i \cdot \beta\right)^2\right] \\
&= \mathbb{E}\left[\epsilon^2\right] \\
&= \sigma^2
\end{aligned}$$

For the expectation value of $\hat{\beta}$, we note that since $\mathbf{X}$ is non-stochastic, the inverse $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ and product with $\mathbf{X}^T$ is also non-stochastic, so

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}\right] &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}\left[\mathbf{y}\right] \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbb{E}\left[\mathbf{X}\beta\right] + \mathbb{E}\left[\epsilon\right]\right) \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\mathbf{y}} \\
&= \beta
\end{aligned}$$

Finally, the variance of $\hat{\beta}$, given by

$$\begin{aligned}
Var\left[\hat{\beta}\right] &= \mathbb{E}\left[\left(\hat{\beta} - \mathbb{E}\left[\hat{\beta}\right]\right)^2\right] \\
&= \mathbb{E}\left[\hat{\beta}^2 - 2\hat{\beta}\beta + \beta^2\right]
\end{aligned}$$

Now let's do some intermediate, simplifying calculations:

$$\begin{aligned}
\hat{\beta} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\beta + \epsilon\right) \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\mathbf{y}} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon \\
&= \beta + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon
\end{aligned}$$

$$\hat{\beta}^2 = \beta^2 + 2\beta\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon + \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\right]^2$$

$$-2\hat{\beta}\beta = -2\beta^2 - 2\beta\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon$$

Putting it all together we get

$$
\begin{aligned}
Var\left[\hat{\beta}\right] &= \mathbb{E}\left[\beta^2 + 2\beta\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon + \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\right]^2 - 2\beta^2 - 2\beta\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon + \beta^2\right] \\
&= \mathbb{E}\left[\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\right)^2\right] \\
&= \left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)^2\left(\mathbf{X}^T\mathbf{X}\right)\mathbb{E}\left[\epsilon^2\right] \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2 \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}
\end{aligned}
$$

assuming that all compounds of $\mathbf{X}$ are non-stochastic and the product $\left(\mathbf{X}^T\right)^2 = \mathbf{X}^T\mathbf{X}$ is properly defined (the notation might be a bit sloppy shorthand, but simplifications like $\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\mathbf{T}\right)^2 = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ using $\left(\mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_n\right)^T = \mathbf{A}_n^T\cdots\mathbf{A}_2^T\mathbf{A}_1^T$ and $\left(\mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_n\right)^{-1} = \mathbf{A}_n^{-1}\cdots\mathbf{A}_2^{-1}\mathbf{A}_1^{-1}$ are a recurring theme, which corresponds to “normal algebra”).

2 Exercise 2

Before we start, we note that the same arguments apply to non-stochastic variables, now including ridge parameter λ and the identity matrix.

$$
\begin{aligned}
\mathbb{E}\left[\hat{\beta}^{\text{Ridge}}\right] &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\
&= \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\mathbb{E}\left[\mathbf{y}\right] \\
&= \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\left(\mathbb{E}\left[\mathbf{X}\beta\right] + \mathbb{E}\left[\epsilon\right]\right) \\
&= \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\mathbf{X}\mathbb{E}\left[\beta\right] \\
&= \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta
\end{aligned}
$$

For the variance,

$$
\begin{aligned}
Var\left[\hat{\beta}_{\text{Ridge}}\right] &= \mathbb{E}\left[\left(\hat{\beta}_{\text{Ridge}} - \mathbb{E}\left[\hat{\beta}_{\text{Ridge}}\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\mathbf{y} - \left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\beta\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\epsilon\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\epsilon\right)\left(\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\epsilon\right)^T\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}\left\{\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\right\}^T\right] \\
&= \left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\mathbb{E}\left[\epsilon^2\right]\mathbf{X}\left\{\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\right\}^T \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\mathbf{X}^T\mathbf{X}\left\{\left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_{\text{pp}}\right)^{-1}\right\}^T
\end{aligned}
$$