Exercises week 37

Exercise 1: Expectation values for ordinary least squares expressions

There exists a continuous function $f(\mathbf{x})$ and a normal distributed error $\epsilon \sim N(0, \sigma^2)$ which describes our data $\mathbf{y} = f(\mathbf{x}) + \epsilon$. We approximate the function f with our model $\tilde{\mathbf{y}} = \mathbf{X}\beta$, minimized by $(\mathbf{y} - \tilde{\mathbf{y}})^2$.

Show that the expectation value of $\mathbf{y}$ for a given element in i:

$$\mathbb{E}(y_i) = \sum_j x_{ij}\beta_j = \mathbf{X}_{i,*}\beta$$

and its variance is:

$$\mathrm{Var}(y_i) = \sigma^2$$

Given $\mathbf{y} = f(\mathbf{x}) + \epsilon$ and $\epsilon \sim N(0, \sigma^2)$

Expectation value:

$$\mathbb{E}(\mathbf{y}) = \mathbb{E}(f(\mathbf{x}) + \epsilon) = \mathbb{E}(f(\mathbf{x}))$$

Since $\mathbb{E}(\epsilon) = 0$, because we assume ϵ to be normally distributed with a mean value of 0, and a variance of σ^2.

We need to keep in mind our model for $\mathbf{y}$ is $\tilde{\mathbf{y}} = \mathbf{X}\beta$

From there we can look at it element-wise:

$$\mathbb{E}(y_i) = \sum_j x_{ij}\beta_j = \mathbf{X}_{i,*}\beta$$

Variance: The variance lies in the normal distributed error $\epsilon \sim N(0, \sigma^2)$. For each point in $\mathbf{y}$ the variance $\mathrm{Var}(y_i) = \sigma^2$. And therefor $y_i \sim N(\mathbf{X}_{i,*}\beta, \sigma^2)$ with the mean value $\mathbf{X}_{i,*}\beta$ and variance σ^2

Show that $\mathbb{E}(\hat{\beta}) = \beta$ using the (OLS) expression for the optimal parameters $\hat{\beta}$,

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Then we need the expression for $\mathbf{y}$, $\mathbf{y} = \mathbf{X}\beta + \epsilon$ and we can substitute into the OLS expression:

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

Then we can find the expectation value:

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

This is because $\mathbb{E}(\beta) = \beta$ and $\mathbb{E}(\epsilon) = 0$

Thus, we have shown that: $\mathbb{E}(\hat{\beta}) = \beta$

To show that $\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, we can start the same way as before:

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

Since β is a constant vector, its variance is zero: $\mathrm{Var}(\beta) = 0$

Variance of linear transformation: $\mathrm{Var}(\mathbf{A}\epsilon) = \mathbf{A}\mathrm{Var}(\epsilon)\mathbf{A}^\mathbf{T}$.

Set $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$:

$$\mathrm{Var}(\mathbf{A}\epsilon) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Var}(\epsilon)((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T$$

$\mathrm{Var}(\epsilon) = \sigma^2$.

$$\begin{aligned}&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T\sigma^2 \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\end{aligned}$$

Exercise 2: Expectation values for Ridge regression

Show that $\mathbb{E}\left[\hat{\beta}^{Ridge}\right] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}(\mathbf{X}^T\mathbf{X})\beta$

First of all $\hat{\beta}^{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{y}$ Substitute for $\mathbf{y}$:

$$\hat{\beta}^{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon)$$

$$\hat{\beta}^{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}(\mathbf{X}^T\mathbf{X}\beta + \mathbf{X}^T\epsilon)$$

Then we find the expectation value of the expression:

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

And again, since $\mathbb{E}(\epsilon) = 0$

$$= \mathbb{E}\left[\hat{\beta}^{Ridge}\right] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbb{E}\left[\mathbf{X}^T\mathbf{X}\beta\right]$$

Since $\mathbb{E}\left[\mathbf{X}^T\mathbf{X}\beta\right]$ is a constant term:

$$\mathbb{E}\left[\mathbf{X}^T\mathbf{X}\beta\right] = \left[\mathbf{X}^T\mathbf{X}\beta\right]$$

And therefor

$$\mathbb{E}\left[\hat{\beta}^{Ridge}\right] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}(\mathbf{X}^T\mathbf{X})\beta$$

And lastly, show that the variance is:

$$\mathrm{Var}\left[\hat{\beta}^{Ridge}\right] = \sigma^2\left[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right]^{-1}\mathbf{X}^T\mathbf{X}\left(\left[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right]^{-1}\right)^T$$

Again we start with the expression of $\hat{\beta}^{Ridge}$ and substitute for $\mathbf{y}$:

$$\begin{aligned}\hat{\beta}^{Ridge} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}(\mathbf{X}^T\mathbf{X}\beta + \mathbf{X}^T\epsilon)\end{aligned}$$

When looking at the variance we can ignore the β term, because $\mathrm{Var}(\beta) = 0$

$$\mathrm{Var}\left[\hat{\beta}^{Ridge}\right] = \mathrm{Var}\left((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T\epsilon\right)$$

And again we need to remember that $\mathrm{Var}(\mathbf{A}\epsilon) = \mathbf{A}\mathrm{Var}(\epsilon)\mathbf{A}^T$ and set $\mathbf{A} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T$

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