University of Oslo Fall, 2023
Department of Mathematics Weekly Exercises: Week 37

Author: Jonas Semprini Næss

FYS-STK4155: Applied Data Analysis and Machine Learning

Analytical Exercises

Expectation values for ordinary least squares expressions:

(**I**.) *Show that the expectation value of* $\mathbf{y}$ *for a given element* i *is*

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

Solution:

Recall that we can describe our model $\mathbf{y}$ by a function $f(\mathbf{x}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$. The function $f(\mathbf{x})$ can be interpreted as some matrix $\mathbf{X}$ times a non-random scalar $\boldsymbol{\beta}$. Thus is the expectation value of y_i [1]

$$\begin{aligned}
\mathbb{E}(y_i) &= \mathbb{E}(\mathbf{X}_{i,*}\boldsymbol{\beta} + \epsilon_i) \\
&= \mathbb{E}(\mathbf{X}_{i,*}\boldsymbol{\beta}) + \underbrace{\mathbb{E}(\epsilon_i)}_{=0} \\
&= \mathbf{X}_{i,*}\boldsymbol{\beta}
\end{aligned}$$

Which is what we wanted to show. ■

(**II**.) *Show that*

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

Solution:

By direct calculation of the variance we have that

$$\begin{aligned}
\mathrm{Var}(y_i) = \mathbb{E}\left[(y_i^2 - \mathbb{E}(y_i))^2\right] &= \mathbb{E}(y_i^2) - (\mathbb{E}(y_i))^2 \\
&= \mathbb{E}((\mathbf{X}_{i,*}\boldsymbol{\beta} + \epsilon_i)^2) - (\mathbf{X}_{i,*}\boldsymbol{\beta})^2 \\
&= \mathbb{E}((\mathbf{X}_{i,*}\boldsymbol{\beta})^2) + \mathbb{E}(2\epsilon_i\mathbf{X}_{i,*}\boldsymbol{\beta}) + \mathbb{E}(\epsilon_i^2) - (\mathbf{X}_{i,*}\boldsymbol{\beta})^2 \\
&= \mathbb{E}(\epsilon_i^2) = \sigma^2.
\end{aligned}$$

Which is what we wanted to show. ■

[1] By convention of notation used in the description of the exercise $\mathbf{X}_{i,*}$ is supposed to define the sum over all values k in row i of the matrix $\mathbf{X}$

(**III.**) *Show that for the optimal parameters $\hat{\boldsymbol{\beta}}$ in OLS that*

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

Solution:

By defintion we have that the optimal parameters $\hat{\boldsymbol{\beta}}$ for OLS is given by

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

which then yields an expectation value of

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right) \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{Y}) \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}.
\end{aligned}$$

Where we have used the fact that $\mathbf{X}$ is a non-stochastic variable and that the $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$. Hence can we observe that the OLS estimator is unbiased. ■

(**IV.**) *Show that the variance for $\hat{\boldsymbol{\beta}}$ is*

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1}.$$

Solution:

Let $\phi = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ such that we can write $\hat{\boldsymbol{\beta}} = \phi\mathbf{Y}$. Then by calculating the variance we have that

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}(\phi\mathbf{Y}) && (1) \\
&= \phi\mathrm{Var}(\mathbf{Y})\phi^T && (2) \\
&= \phi\mathrm{Var}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\phi^T && (3) \\
&= \phi\sigma^2\phi^T && (4) \\
&= \sigma^2\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T\right) && (5) \\
&= \sigma^2\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right) && (6) \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1} && (7)
\end{aligned}$$

which is what we wanted to show. ■

Expectation values for Ridge regression

(**I.**) *Show that*

$$\mathbb{E}\Big[\boldsymbol{\beta}^{\text{Ridge}}\Big] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}(\mathbf{X}^\top\mathbf{X})\ \boldsymbol{\beta}^{\text{OLS}}.$$

By the definition of ridge regression we know that the optimal parameters are given by

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Hence would accordingly the expectation value yield

$$\begin{aligned}
\mathbb{E}(\tilde{\boldsymbol{\beta}}) &= \mathbb{E}\left((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top\mathbf{Y}\right) \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top\mathbb{E}(\mathbf{Y}) \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top\mathbb{E}(\mathbf{X}\boldsymbol{\beta} + \epsilon_i) \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\
&= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}^{\text{OLS}}. \qquad \boxed{\text{By results of Ordinary Least Squares}}
\end{aligned}$$

Meaning $\mathbb{E}\Big[\tilde{\boldsymbol{\beta}}\Big] \neq \boldsymbol{\beta}^{\text{OLS}}$ for any $\lambda > 0$ and concludes what we wanted to show. ■

(**II.**) *Show also that the variance is*

$$\text{Var}[\boldsymbol{\beta}^{\text{Ridge}}] = \sigma^2[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}]^{-1}\mathbf{X}^T\mathbf{X}\{[\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}]^{-1}\}^T.$$

Solution:

By defintion of the variance for a random stochastic variable we have that

$$\text{Var}(\tilde{\boldsymbol{\beta}}) = \mathbf{A}\text{Var}(\mathbf{Y})\mathbf{A}^T$$

where $\mathbf{A} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^T$. Hence

$$\begin{aligned}
\text{Var}(\tilde{\boldsymbol{\beta}}) &= \mathbf{A}\text{Var}(\mathbf{X}\boldsymbol{\beta} + \epsilon_i)\mathbf{A}^T && (8)\\
&= \mathbf{A}\sigma^2\mathbf{A}^T && (9)\\
&= \sigma^2\left((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1})\mathbf{X}^T((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1})\mathbf{X}^T)^T\right) && (10)\\
&= \sigma^2\left((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1})\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1})^T\right) && (11)\\
&= \sigma^2[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}]^{-1}\mathbf{X}^T\mathbf{X}\{[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}]^{-1}\}^T && (12)
\end{aligned}$$

which is what we wanted to show. ■

Appendix:

More detailed calculations:

<u>Transpose of Matrix product:</u>

Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ then

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T.$$

Used in (6) and (11).

Variance Identity:

Let $\boldsymbol{\phi} \in \mathbb{R}^{m \times n}$ and $\mathbf{X} \in \mathbb{R}^{n \times 1}$. Then

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\phi}\mathbf{X}) &= \mathbb{E}\{[(\boldsymbol{\phi}\mathbf{X} - \mathbb{E}(\boldsymbol{\phi})(\mathbf{X}\boldsymbol{\phi}\mathbf{X} - \mathbb{E}(\boldsymbol{\phi}\mathbf{X})]^T\} \\
&= \mathbb{E}\{[\boldsymbol{\phi}\mathbf{X} - \boldsymbol{\phi}\mathbb{E}(\mathbf{X})][\boldsymbol{\phi}\mathbf{X} - \boldsymbol{\phi}\mathbb{E}(\mathbf{X})]^T\} \\
&= \mathbb{E}\{[\boldsymbol{\phi}(\mathbf{X} - \mathbb{E}(\mathbf{X}))][\boldsymbol{\phi}(\mathbf{X} - \mathbb{E}(\mathbf{X}))]^T\} \\
&= \boldsymbol{\phi}\mathbb{E}\{[\mathbf{X} - \mathbb{E}(\mathbf{X})][\mathbf{X} - \mathbb{E}(\mathbf{X})]^T\}\boldsymbol{\phi}^T \\
&= \boldsymbol{\phi}\mathbf{X}\boldsymbol{\phi}^T
\end{aligned}$$

Used at (2) and (9).

(4) Want to show that $\mathbb{E}(\mathbf{Y}\mathbf{Y}^T) = \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T + \sigma^2 I_{n\times n}$.

Remember that we can model $\mathbf{y}$ by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$. This implies that for each component y_i we have that $y_i = X_{i,*}\beta_i + \epsilon_i$ where each ϵ_i has variance σ^2. Thus for the full model the $\boldsymbol{\epsilon}$ is simply a diagonal matrix with its variance along the main diagonal, hence $\sigma^2 I_{n\times n}$ by factorisation. By utilising this fact we then have that

$$\begin{aligned}
\mathbb{E}(\mathbf{Y}\mathbf{Y}^T) &= \mathbb{E}\left((\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})^T\right) \\
&= \mathbb{E}\left(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\epsilon}^T + \boldsymbol{\epsilon}\boldsymbol{\beta}^T\mathbf{X}^T + \boldsymbol{\epsilon}^2\right) \\
&= \mathbb{E}\left(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T\right) + \mathbb{E}\left(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\epsilon}^T\right) + \mathbb{E}\left(\boldsymbol{\epsilon}\boldsymbol{\beta}^T\mathbf{X}^T\right) + \mathbb{E}\left(\boldsymbol{\epsilon}^2\right) \\
&= \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T + 0 + 0 + \sigma^2 I_{n\times n} \\
&= \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T + \sigma^2 I_{n\times n}.
\end{aligned}$$

Which is what we wanted to show. $\blacksquare$