Vedlegg A - Forventningsverdier og varians av y og β

Her skal vi vise utledningene for uttrykkene $\mathbb{E}[y]$, $\text{Var}(y)$,$\mathbb{E}[\hat{\boldsymbol{\beta}}]$ og $\text{Var}(\hat{\boldsymbol{\beta}})$ i tilfellene OLS og Ridge. Disse utledningen baserer seg på ukesoppgaver gjort i kurset FYS-STK4155 i uke 37. Vi tar utgangspunkt i en antagelse om at dataene våre kan beskrives av en kontinuerlig funksjon $f(\boldsymbol{x})$ og en normalfordelt feil $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon}$$

Deretter approksimerer vi denne funksjonen med modellen vår $\tilde{\boldsymbol{y}}$ som kommer fra løsningen av ligningene for minste kvadraters metode (OLS), slik at modellen vår approksimeres med $\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta}$.

Vi starter med å vise at modellen i det hele tatt er anvendbard, fordi $\mathbb{E}[y] = \mathbf{X}\boldsymbol{\beta}$. Vi må benytte oss av at $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, hvor $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$ og $\mathbf{X}$ og $\boldsymbol{\beta}$ er ikke-stokastiske.

Da blir

$$\mathbb{E}[y] = \mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbb{E}[\mathbf{X}\boldsymbol{\beta}] + \mathbb{E}[\boldsymbol{\varepsilon}].$$

Men fordi $\mathbf{X}$ og $\boldsymbol{\beta}$ er ikke-stokastiske er $\mathbb{E}[\mathbf{X}\boldsymbol{\beta}] = \mathbf{X}\boldsymbol{\beta}$, slik at

$$\mathbb{E}[y] = \mathbf{X}\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{X}\boldsymbol{\beta} + 0 = \mathbf{X}\boldsymbol{\beta},$$

siden $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$.

Videre skal vi vise at variansen til dataene våre er gitt ved $\text{Var}(y_i) = \sigma^2$. Vi begynner med å definere $\text{Var}(y)$ som

$$\text{Var}(y) = \mathbb{E}[(y - \mathbb{E}(y))^2].$$

Bruker at $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ og at $\mathbb{E}(y) = \mathbf{X}\boldsymbol{\beta}$, som gir

$$\text{Var}(y) = \mathbb{E}[(y - \mathbb{E}(y))^2] = \mathbb{E}[(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{\beta}))^2] = \mathbb{E}[(\boldsymbol{\varepsilon})^2] = \sigma^2$$

siden $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$. Videre skal vi se på forvendningsverdier for $\boldsymbol{\beta}$. Med $\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$, blir

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

Videre kan vi vise at $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1}$. Vi tar utgangspunkt i at

$$\text{Var}(\hat{\boldsymbol{\beta}_{OLS}}) = \mathbb{E}[(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}])^2]$$

Da blir

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

Videre skal vi vise at

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

Vi tar utgangspunkt i at

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

og

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Da blir

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

fordi $\boldsymbol{X}$ og $\boldsymbol{\beta}$ er ikke-stokastiske og $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$. For variansen til $\hat{\boldsymbol{\beta}}^{\text{Ridge}}$ tar vi utgangspunkt i at

$$\text{Var}(\hat{\boldsymbol{\beta}}^{\text{Ridge}}) = \mathbb{E}[(\hat{\boldsymbol{\beta}}^{\text{Ridge}} - \mathbb{E}[\hat{\boldsymbol{\beta}}^{\text{Ridge}}])^2]$$

Da får vi

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

Ser på faktoren foran $\mathbb{E}[\boldsymbol{\varepsilon})^2]$.

$$\begin{aligned}
&((\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I}_{pp})^{-1}\boldsymbol{X}^T)^2 \\
&= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I}_{pp})^{-1}\boldsymbol{X}^T((\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I}_{pp})^{-1}\boldsymbol{X}^T)^T \\
&= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I}_{pp})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\mathbf{I}_{pp})^{-1})^T
\end{aligned}$$

Da blir

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

Siden $\mathbb{E}[\boldsymbol{\varepsilon})^2] = \sigma^2$.