FYS-STK4155 Project 1

Bendik Steinsvåg Dalen & Gabriel Sigurd Cabrera

October 4, 2019

Abstract

1 Introduction

Things we need to write about:

Background on regression analysis and resampling methods. Mention: OLS Ridge Lasso k-fold cross-validation Bias-Variance trade of?

We will first study how they preform for the two dimensional Franke-function. (A bit about the Franke-function, maybe a tldr for the method).

We will then implement them for some real terrain data for Møsvatn Austfjell in Norway. mm. (biggest lake in Norway)

2 Data

We will use real terrain data for Møsvatn Austfjell in Telemark, Norway, collected from https://earthexplorer.usgs.gov.

3 Method

4 Results

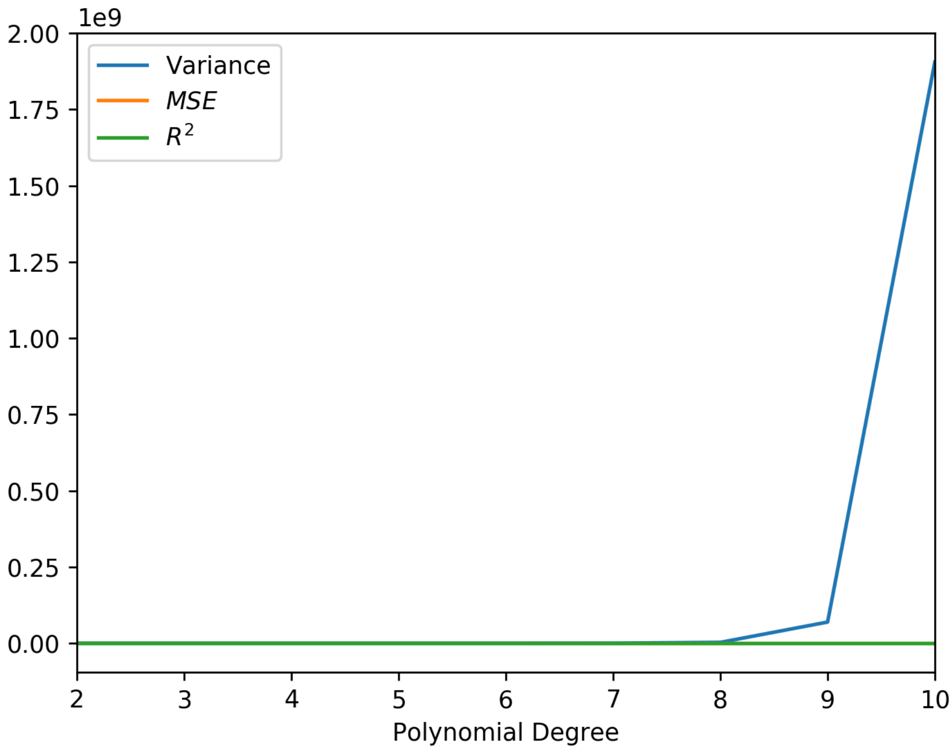

Figure 1: A plot of the mean square error, the R^2-score and the σ variance of the β-values against the polynomial degree after performing a standard least square regression analysis on the Franke-function

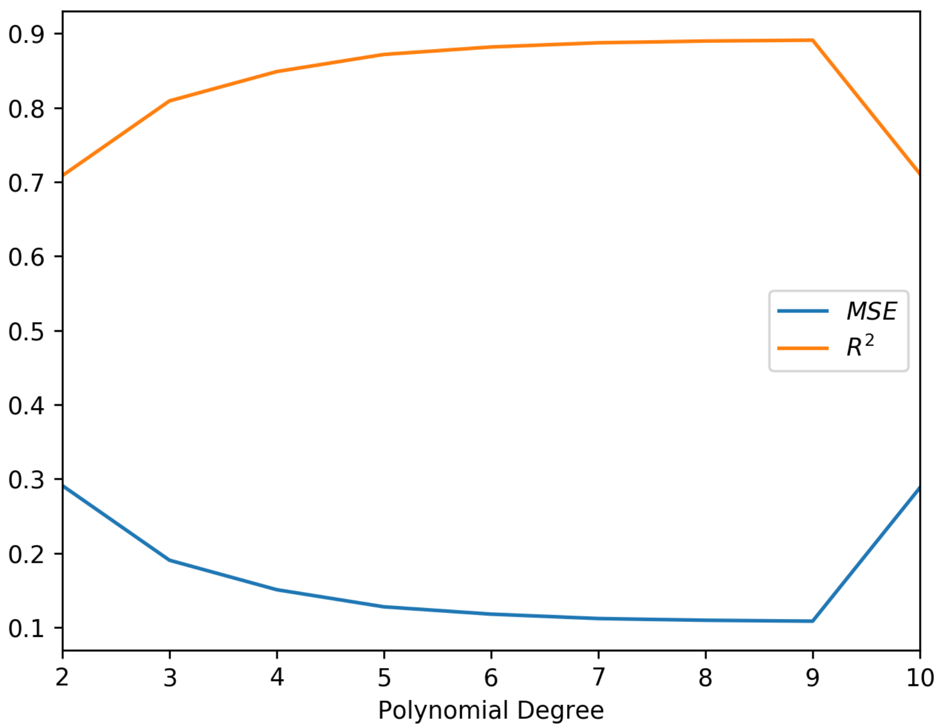

Figure 2: A plot of the mean square error and the R^2-score against the polynomial degree after performing a standard least square regression analysis on the Franke-function and and performing a k-fold cross-validation

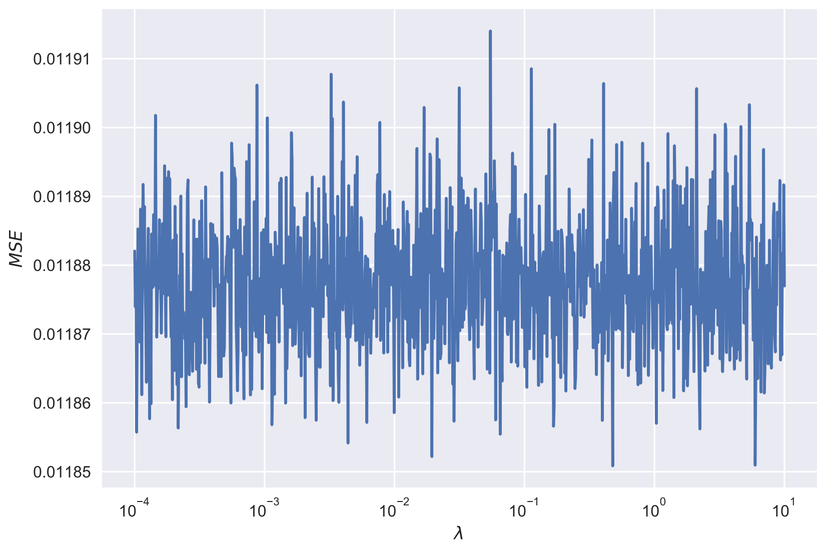

Figure 3: The Mean Squared Error for the Ridge method for different values of λ

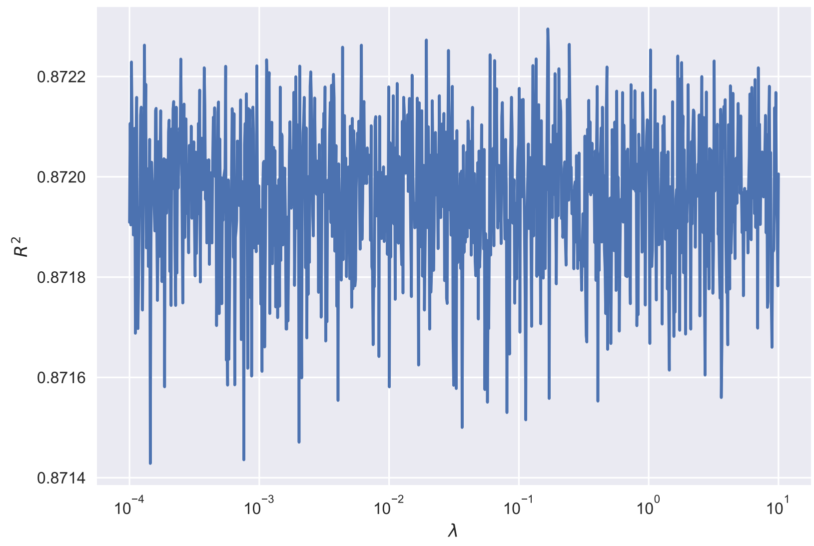

Figure 4: R^2-score for the Ridge method for different values of λ

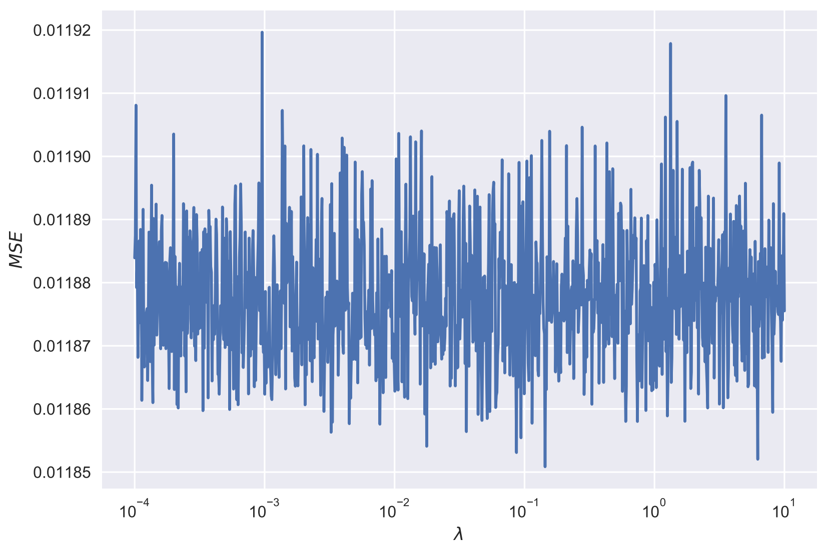

Figure 5: The Mean Squared Error for the Lasso method for different values of λ

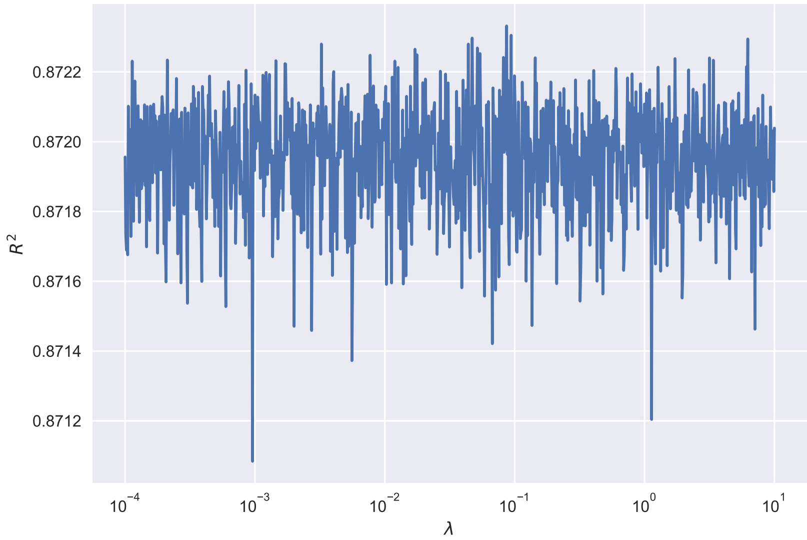

Figure 6: R^2-score for the Lasso method for different values of λ

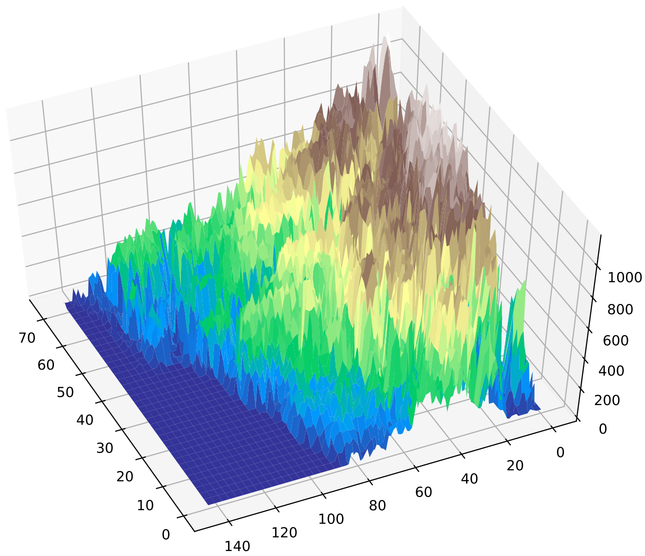

Figure 7: The terrain-data we are studying, from Møsvatn Austfjell in Norway

5 Discussion

6 Appendix

6.1 Part c math

We are too show that

$$C(\boldsymbol{X}, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2 = \mathbf{E}\left[(\boldsymbol{y} - \tilde{\boldsymbol{y}})^2\right] = \frac{1}{n}\sum_i (f_i - \mathbf{E}[\tilde{\boldsymbol{y}}])^2 + \frac{1}{n}\sum_i (\tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}])^2 + \sigma^2 \tag{1}$$

$$\begin{aligned}
\mathbf{E}\left[(\boldsymbol{y} - \tilde{\boldsymbol{y}})^2\right] =& \frac{1}{n}\sum_i (y_i - \tilde{y}_i)^2 = \frac{1}{n}\sum_i (f_i + \varepsilon - \tilde{y}_i)^2 && (2)\\
=& \frac{1}{n}\sum_i (f_i + \varepsilon - \tilde{y}_i + \mathbf{E}[\tilde{\boldsymbol{y}}] - \mathbf{E}[\tilde{\boldsymbol{y}}])^2 \qquad | \text{ introduce } a = f_i - \mathbf{E}[\tilde{\boldsymbol{y}}] \text{ and } b = \tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}] && (3)\\
=& \frac{1}{n}\sum_i (a - b + \varepsilon)^2 = \frac{1}{n}\sum_i \left(a^2 - 2ab + b^2 - 2b\varepsilon + \varepsilon^2 + 2a\varepsilon\right) && (4)\\
=& \frac{1}{n}\sum_i (f_i - \mathbf{E}[\tilde{\boldsymbol{y}}])^2 + \frac{1}{n}\sum_i \left(\varepsilon^2\right) + \frac{1}{n}\sum_i (\tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}])^2 - 2\frac{1}{n}\sum_i \varepsilon\,(\tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}]) + 2\frac{1}{n}\sum_i \varepsilon\,(f_i - \mathbf{E}[\tilde{\boldsymbol{y}}]) && (5)\\
&- 2\frac{1}{n}\sum_i (f_i - \mathbf{E}[\tilde{\boldsymbol{y}}])(\tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}]) && (6)\\
=& \frac{1}{n}\sum_i (f_i - \mathbf{E}[\tilde{\boldsymbol{y}}])^2 + \frac{1}{n}\sum_i (\tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}])^2 + \sigma^2 - 2\mathbf{E}[\varepsilon]\frac{1}{n}\sum_i (\tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}]) + 2\mathbf{E}[\varepsilon]\frac{1}{n}\sum_i (f_i - \mathbf{E}[\tilde{\boldsymbol{y}}]) && (7)\\
&- 2\frac{1}{n}\sum_i (f_i - \mathbf{E}[\tilde{\boldsymbol{y}}])(\tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}]) && (8)\\
=& \frac{1}{n}\sum_i (f_i - \mathbf{E}[\tilde{\boldsymbol{y}}])^2 + \frac{1}{n}\sum_i (\tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}])^2 + \sigma^2 \quad \blacksquare && (9)
\end{aligned}$$

Where $\frac{1}{n}\sum_i (f_i - \mathbf{E}[\tilde{\boldsymbol{y}}])$ is the bias and $\frac{1}{n}\sum_i (\tilde{y}_i - \mathbf{E}[\tilde{\boldsymbol{y}}])^2$ is the variance.
(skal vi gjøre noe annet og?)