FYS-STK4155 - Project1

Gard, Are, David Andreas Bordvik

September 21, 2021

Motivation

In Project 1, we are tasked to study various regressions methods, such as Ordinary Least Squares, Ridge and Lasso. Our first area of study is how to fit polynomials to a specific two-dimensional function called Franke's Function. Our motivation behind fitting polynomials to Frank's function is to test the implementation of our regression algorithms, as well as studying various techniques such as bootstrapping and measurements such as the bias-variance tradeoff. Finally, we will move on to use real digital terrain data for our analysis.

Frank's function is given on the form

$$f(x,y) = \frac{3}{4}\exp\left(-\frac{(9x-2)^2}{4} - \frac{(9y-2)^2}{4}\right) + \frac{3}{4}\exp\left(-\frac{(9x+1)^2}{49} - \frac{(9y+1)}{10}\right) + \frac{1}{2}\exp\left(-\frac{(9x-7)^2}{4} - \frac{(9y-3)^2}{4}\right) - \frac{1}{5}\exp\left(-(9x-4)^2 - (9y-7)^2\right)$$

with a 3-dimensional plot given in Figure 1

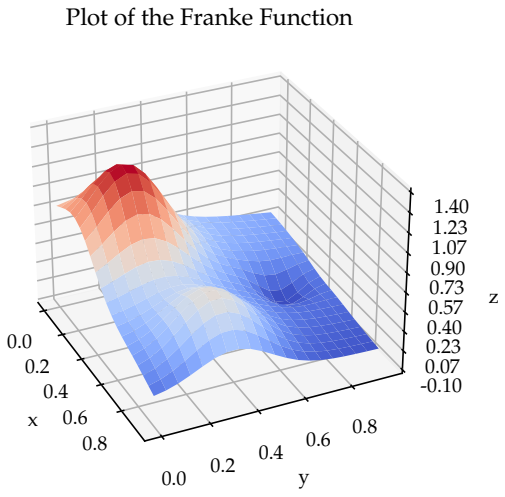

Figure 1: Plot of the Franke Function

Theory

MSE value between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ is defined as;

$$MSE(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2 \tag{1}$$

The R^2 value between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ is defined as;

$$R^2(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2} \tag{2}$$

The mean value of $\boldsymbol{y}$ is defined as;

$$\bar{y} = \frac{1}{n}\sum_{i=0}^{n-1} y_i \tag{3}$$

The variance of $\boldsymbol{y}$ is defined as;

$$\text{var}[\boldsymbol{x}] = \frac{1}{n}\sum_{i=0}^{n-1}(x_i - \overline{x})^2 = \mathbb{E}\left[(\boldsymbol{y} - \tilde{\boldsymbol{y}})^2\right] \tag{4}$$

Exercise 1

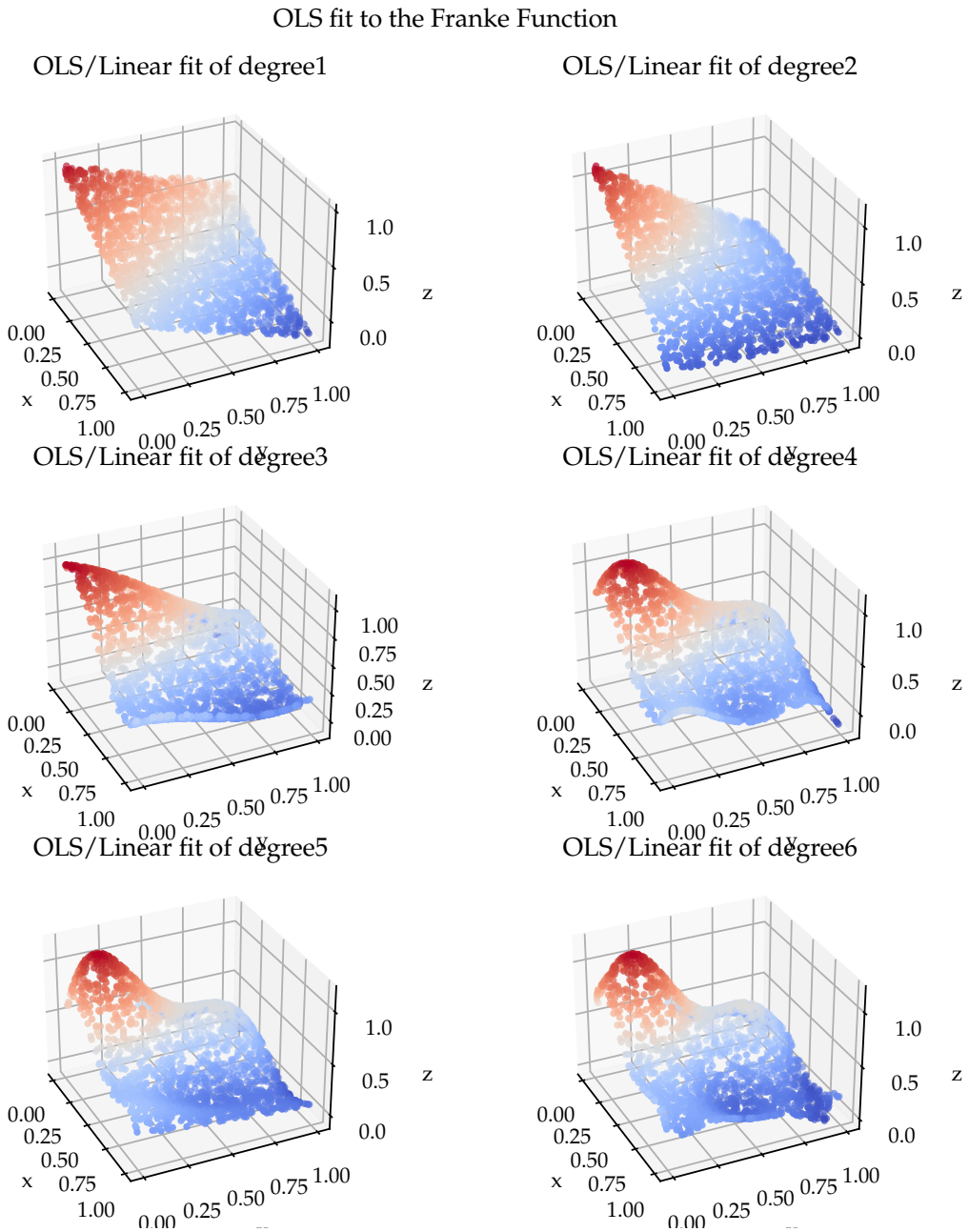
OLS fit to the Franke Function
OLS/Linear fit of degree1
OLS/Linear fit of degree2
OLS/Linear fit of degree3
OLS/Linear fit of degree4
OLS/Linear fit of degree5
OLS/Linear fit of degree6
x
y
z

Figure 2: ?

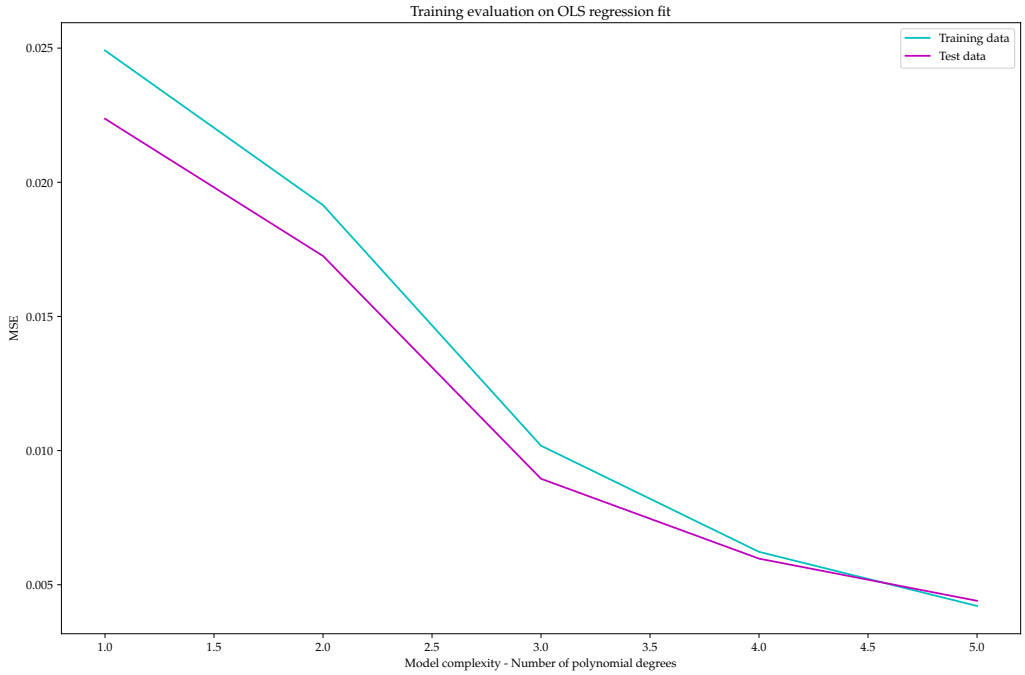

Figure 3: ?

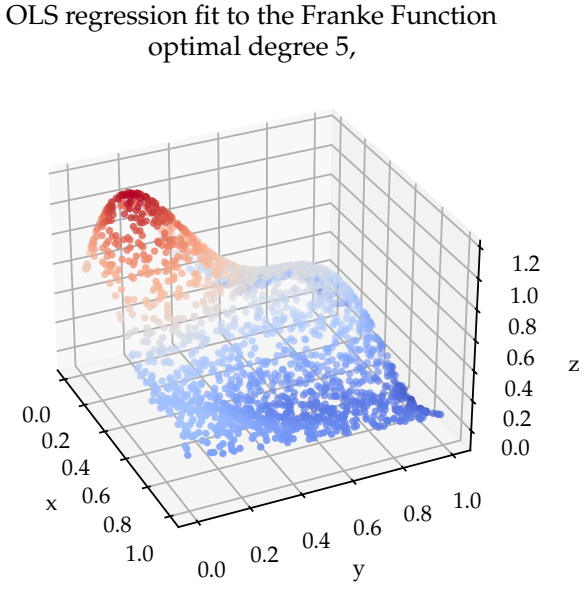

Figure 4: ?

Exercise 2

Moving on to the bias-variance trade-off analysis, we start off by showing that

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

Can be rewritten as

$$\mathbb{E}\left[(\boldsymbol{y} - \tilde{\boldsymbol{y}})^2\right] = \frac{1}{n}\sum_{i}(f_i - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2 + \frac{1}{n}\sum_{i}(\tilde{y}_i - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2 + \sigma^2.$$

where the terms are respectively (bias)^2, variance and noise. For simplicity, assume that we have a dataset where the data is generated from a noisy model

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

Furthermore, we will assume that the residuals ϵ are independant and normally distributed $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Finally, $\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta}$ is our approximation to the functions f. Start off by adding and subtracting $\mathbb{E}\left[\tilde{\boldsymbol{y}}\right]$ inside the expectation value.

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{y} - \tilde{\boldsymbol{y}})^2\right] &= \mathbb{E}\left[(\boldsymbol{y} - \tilde{\boldsymbol{y}} + \mathbb{E}\left[\tilde{\boldsymbol{y}}\right] - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2\right] \\ &= \mathbb{E}\left[((\boldsymbol{y} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right]) - (\tilde{\boldsymbol{y}} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right]))^2\right]\end{aligned}$$

By using the fact that $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$ we can rewrite this as

$$= \mathbb{E}\left[((f(\boldsymbol{x}) - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right]) - (\tilde{\boldsymbol{y}} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right]) + \epsilon)^2\right]$$

Computing the square inside the expectation value gives us

$$= \mathbb{E}\left[(f(\boldsymbol{x}) - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2 + (\tilde{\boldsymbol{y}} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2 + \epsilon^2\right]$$

$$+2\left(\mathbb{E}\left[\epsilon(f(\boldsymbol{x}) - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])\right] - \mathbb{E}\left[\epsilon(\tilde{\boldsymbol{y}} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])\right] - \mathbb{E}\left[(f(\boldsymbol{x}) - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])(\tilde{\boldsymbol{y}} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])\right]\right)$$

Moreover, as ϵ are independant variables, the expectation value involving them as a product can be written as a product of expectation values. Knowing that $\mathbb{E}\left[\epsilon\right] = 0$, the third and second to last term is equal to zero. Also, knowing that $\mathbb{E}\left[\tilde{\boldsymbol{y}}\right] = \tilde{\boldsymbol{y}}$, the last t

$$= \mathbb{E}\left[(f(\boldsymbol{x}) - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2\right] + \mathbb{E}\left[(\tilde{\boldsymbol{y}} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2\right] + \mathbb{E}\left[\epsilon^2\right] \tag{5}$$

The first term in (5) can be discretized as

$$\mathbb{E}\left[(f(\boldsymbol{x}) - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2\right] = \frac{1}{n}\sum_{i}(f_i - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2$$

Which is the bias squared as we were to show.

The second term in (5) is also discretized, yielding

$$\mathbb{E}\left[(\tilde{\boldsymbol{y}} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2\right] = \frac{1}{n}\sum_{i}(\tilde{\boldsymbol{y}} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2$$

Which takes form of the variance, as was set out to show.

Finally, it can be shown that $\text{var}(\boldsymbol{y}) = \text{var}(f + \epsilon) = \mathbb{E}\left[(f+\epsilon)^2\right] - (\mathbb{E}\left[(f+\epsilon)\right])^2 = \mathbb{E}\left[\epsilon^2\right]$ As such, we can use that

$$\text{var}(y) = \sigma^2 = \mathbb{E}\left[\epsilon^2\right]$$

to see that the final term in (5) is equal to the noise. Thus we have shown that

$$\mathbb{E}\left[(\boldsymbol{y} - \tilde{\boldsymbol{y}})^2\right] = \mathbb{E}\left[(f(\boldsymbol{x}) - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2\right] + \mathbb{E}\left[(\tilde{\boldsymbol{y}} - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2\right] + \mathbb{E}\left[\epsilon^2\right]$$

$$= \frac{1}{n}\sum_i (f_i - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2 + \frac{1}{n}\sum_i (\tilde{y}_i - \mathbb{E}\left[\tilde{\boldsymbol{y}}\right])^2 + \sigma^2.$$

The final expression consists of three terms. The first term is a sum of the squared bias, the second term is the variance and the final term is a constant noise term. The sum of squared bias and variance make up the mean square error of our model. [1]

The bias component of the mean square error measures the difference from the true mean to the desired regression model. As function of model complexity, the bias will decrease as complexity increases. The second term, the variance, gives a measurement of the variation of the model values around their average value. The variance will increase with model complexity. The constant noise term σ^2 is an irreducible error, and as such does not contribute when analysing the bias-variance tradeoff.

The bias-variance tradeoff can be used as a method for model selection. As has been alluded to in the previous paragraph, the variance is inverse proportional to the bias. As such, there is a trade-off between bias and variance. As the bias-variance is directly related to the mean square error for both the training data but also test and production data used for making new predictions. When selecting a model we wish to balance and minimize both quantities, as that leads to the model with the best predicitve capabilities. [2]

Furthermore, the problem of overfitting can also be discussed in light of the bias variance tradeoff. Overfitting is proportional to the model variance, as such a high variance leads to an overfitted model. An overfitted model is one that has learned the noise of the training data, resulting in a perfect fit for the training data but a high mean square error when predicting using new data. Hence, a higher bias can be considered more useful to circumvent overfitting.

Lastly, it should be noted that the bias-variance tradeoff measuremnet is performed for a single dataset of limited size. As such, if there were more datapoints to train the model with, the model would attain a better overall fit. Moreover, a greater amount of training data would reduce the level of overfitting for a given model complexity. [2] Though there are some limitations to the measurement, the bias-variance tradeoff can be used to estimate where the trained model is general enough to both avoid under -and overfitting (i.e. too high bias or too high variance).

	MSE	R2-score
Training data	0.01636495753709963	0.8053964213340778
Test data	0.01636495753709963	0.8053964213340778

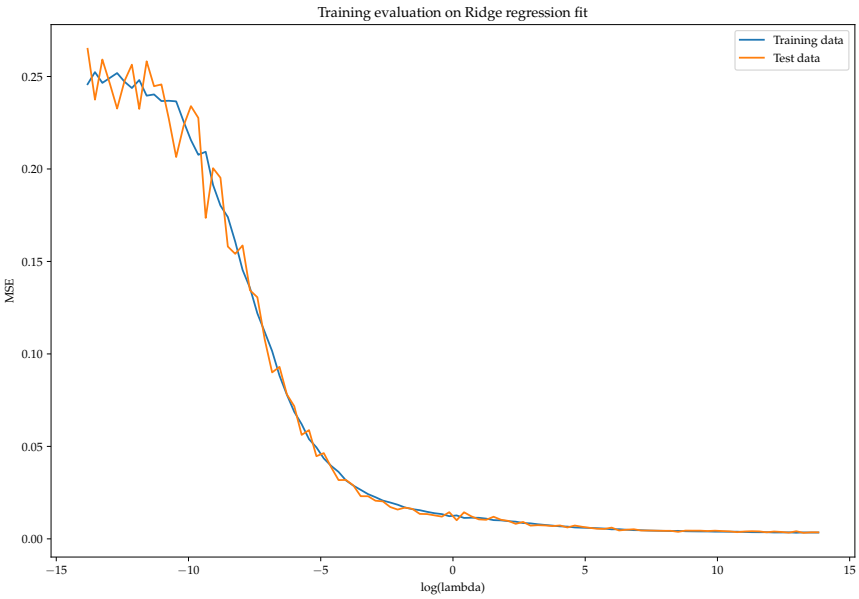

Figure 5: The best fit using Ridge

Ridge regression fit to the Franke Function
Degree 6,λ:-1.1

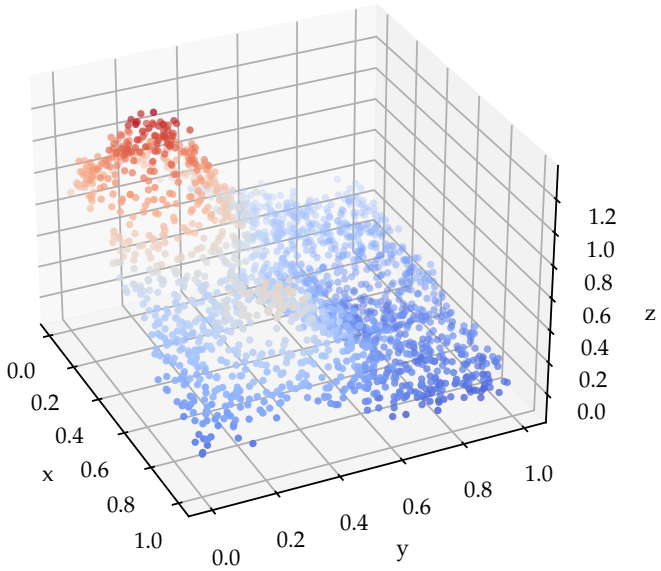

Figure 6: The best fit using Ridge

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. SPRINGER NATURE, Feb. 2009.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer New York, Aug. 2016.