# FYS-STK4155: Project 1

Simen Nyhus Bastnes

5. October 2020

**Abstract**

In this project, we will study various methods of linear regression, namely the Ordinary Least Squares method, Ridge, and Lasso regression, as well as investigate resampling the data via Bootstrapping and $k$-fold Cross-Validation. The data we will be looking at is the so-called Franke function, as well as terrain elevation data from a region in Norway. The Franke function gives us a way of testing our implementation before moving on to the more complex terrain data.

## 1 Introduction

With the emergence of more powerful computers, the field of machine learning is steadily becoming an integral part of both business and many fields of science. While many of the concepts and algorithms used in machine learning today has been known for a long time, some of them have simply been too computationally expensive to do efficiently. While not typically too computationally heavy **do something about this**, linear regression is one of the simplest and most-studied forms of machine learning, and provides a good introduction to concepts commonly used in machine learning.

In this project, we will look at three different methods of regression analysis and compare how they fare against each other. The methods we will be using is the Ordinary Least Squares method, Ridge regression, and Lasso regression. We will also see how resampling the data affects the results from the regression methods, by implementing the Bootstrap algorithm and the $k$-fold Cross-Validation.

There are two different data sets that will be studied in this project. The first, is the Franke function from [4], as well as terrain data for a region in Norway taken from [1]. First, in Chapter 2 we will introduce the theory behind linear regression, as well as the the regression methods and resampling methods employed later in the project. In Chapter 3 we go through the implementation of the methods, explaining how the code is structured and used. Then, in Chapter 4 we go through the results of both the Franke function and the terrain data, discussing them in more detail in Chapter 5. Lastly, we conclude our findings in Chapter 6.

## 2 Theory

### 2.1 Linear regression

Linear regression is a method of fitting a set of $p$ *predictors* $\boldsymbol{X}$ to a data set $\boldsymbol{y}$, while minimizing the error between the *response* $\tilde{\boldsymbol{y}}$ and the actual data $\boldsymbol{y}$. For each of the $n$ samples $y_i$ in the data set the

relationship between the response and the predictors $\boldsymbol{X}_i$ is modeled in a linear fashion, giving us the following matrix equation

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

where $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})^\mathsf{T}$ are the regression parameters we are trying to estimate, one for each predictor, and $\boldsymbol{\epsilon}$ is the error in our approximation. The matrix $\mathbf{X}$ is often called the design matrix, and the equation can be written a bit more explicitly as

$$y_i = \beta_0 + X_{i,1}\beta_1 + ... + X_{i,p-1}\beta_{p-1} + \epsilon_i$$

Exactly what each predictor is can vary a lot from case to case, and how the design matrix is set up is important for the accuracy of the fit. In our case, we will focus on a form of linear regression where the predictors is on the form of a polynomial in the input parameters. In the case where we have a data set $\boldsymbol{y}(\boldsymbol{x})$, the design matrix can for example be written on the form of

$$\mathbf{X} = (\boldsymbol{x}^0, \boldsymbol{x}^1, ..., \boldsymbol{x}^{p-1})$$

With that said, we still need some way to find the $\beta$'s that fit the data best, and we will now look at three ways to try to do this.

### 2.1.1 Ordinary least squares

Following Chapter 2.3 of Hastie et al. [3], in order to find the optimal regression parameters $\beta$, the OLS method minimizes the residual sum of squares

$$\text{RSS}(\beta) = \sum_{i=1}^{N}(y_i - x_i^T\beta)^2$$

With $\boldsymbol{y}$ as the vector containing all $N$ $y_i$, and $X$ an $N \times p$ matrix as shown in section 2.1, this can be written as

$$\text{RSS}(\lambda) = (\boldsymbol{y} - \mathbf{X}\beta)^\mathsf{T}(\boldsymbol{y} - \mathbf{X}\beta)$$

Differentiating with respect to $\beta$ we get

$$\frac{\partial \text{RSS}}{\partial \beta} = X^\mathsf{T}(\boldsymbol{y} - \mathbf{X}\beta)$$

In order to find an optimal $\beta$, this has to be zero

$$\begin{aligned}\mathbf{X}^\mathsf{T}(\boldsymbol{y} - \mathbf{X}\beta) &= 0 \\ \mathbf{X}^\mathsf{T}\mathbf{X}\beta &= \mathbf{X}^T\boldsymbol{y}\end{aligned}$$

Finally giving us the expression for the optimal regression parameters

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

assuming that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is invertible.

### 2.1.2 Ridge regression

Ridge regression is an example of a so-called shrinkage method, which shrinks the regression coefficients by adding a small penalty proportional to their size.

$$\beta^{\text{Ridge}} = \min_{\beta \in \mathbb{R}^p} \frac{1}{n} ||\boldsymbol{X}\beta - \boldsymbol{y}||_2^2 + \lambda ||\beta||_2^2$$

where $\lambda$ is a regularization parameter that controls the amount of shrinkage, and we $||\beta||_2^2 \leq t$ where $t$ is a finite number larger than zero. The higher $\lambda$, the more shrinkage occurs. This can be shown to give the Ridge solution

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

The aim with Ridge regression is to limit the potential problems with singularities when computing the inverse of $\mathbf{X}^\mathsf{T}\mathbf{X}$, which can be a problem when there are many correlated variables.

### 2.1.3 Lasso regression

Lasso regression is another shrinkage method, with a slightly different optimization equation compared to Ridge regression

$$\beta^{\text{Lasso}} = \min_{\beta \in \mathbb{R}^p} \frac{1}{n} ||\boldsymbol{X}\beta - \boldsymbol{y}||_2^2 + \lambda ||\beta||_1$$

where $||\beta||_1$ is the $L_1$ norm.

## 2.2 Bias-Variance decomposition

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

## 2.3 R2????

## 2.4 Confidence intervals

For the OLS and Ridge regression cases, it is possible to derive the variance of $\beta$ (a proper derivation is given in [5]), and thus the confidence intervals as well. For the OLS method, the variance is given by

$$\text{Var}(\beta) = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$$

where $\sigma^2$ is the estimated variance of $y$ given by

$$\sigma^2 = \frac{1}{N-p-1}\sigma_{i=1}^{N}(y_i - \tilde{y}_i)^2$$

Taking the square root of the diagonal of $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ gives us an estimate of the variance of the $j$-th regression coefficient

$$\sigma^2(\beta_j) = \sigma^2\sqrt{[\mathbf{X}^\mathsf{T}\mathbf{X}]_{jj}^{-1}}$$

Letting us construct the 95% confidence intervals by

$$\text{CI}(\beta_j) = \left[\beta_j - 2\sqrt{\sigma_2(\beta_j)}\,,\ \ \beta_j + 2\sqrt{\sigma_2(\beta_j)}\right]$$

Similarly, the variance for $\beta$ for Ridge regression can be found to be

$$\text{Var}[beta^{\text{Ridge}}] = \sigma^2[\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}]^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}[(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}]^\mathsf{T}$$

and confidence interval can be constructed following the same steps as done above for OLS.

### 2.5 Resampling methods

#### 2.5.1 Bootstrap

#### 2.5.2 Cross-validation

## 3 Implementation

The heart. [2]

## 4 Results

## 5 Discussion

## 6 Conclusion

Introduce why we set out, then explain results

## References


[1] Earthexplorer. `https://earthexplorer.usgs.gov/`.

[2] Github repository, project 1. `https://github.com/simennb/fysstk4155-project1`.

[3] Hastie et al. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction.* Springer, second edition edition, 2017.

[4] Richard Franke. A critical comparison of some methods for interpolation of scattered data, 1979.

[5] Wessel N. van Wieringen. Lecture notes on ridge regression, 2020.


## A Appendix