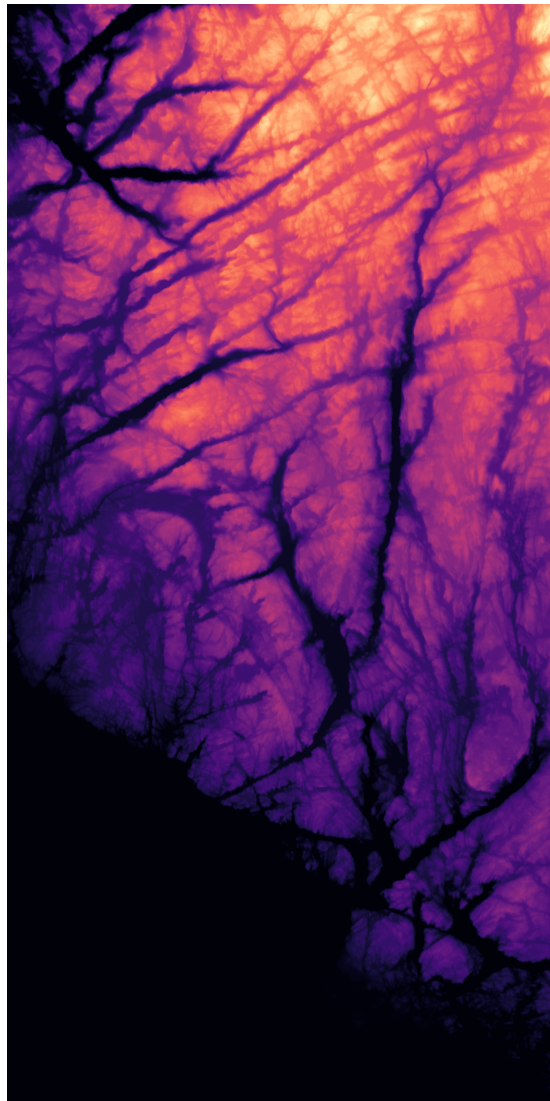



# Regression analysis and resampling methods

FYS-STK3155 - Project 1

Didrik Spanne Reilstad



**Github**

 [github.com/dreilstad/FYS-STK3155/Project1](https://github.com/dreilstad/FYS-STK3155/Project1)

## **Abstract**

Using regression methods

# **1 Introduction**

Regression analysis is the process of using statistical methods to examine the relationship between a number of independent variables and a dependent variable. The goal of regression analysis is to exploit the resulting model from our regression analysis to make predictions and fit a function to the independent variables. The ability to infer information from the relationship between a set of independent variables and a dependent variable to predict and forecast is an incredibly powerful tool.

Regression analysis and modeling is widely used in disciplines such as social sciences, epidemiology, finance and economics, and also plays an important role in machine learning. Especially relevant at the time of writing, is regression analysis in the field of epidemiology. Considering the current COVID-19 pandemic, using regression analysis to track the spread of a virus in a population and being able to predict the development of spread is an incredibly powerful tool. In general regression analysis is an important tool used in analytical epidemiology.[1]

The focus of this project was to study various regression methods such as Ordinary Least Squares(OLS), Ridge regression and Lasso regression. Furthermore using resampling methods like k-fold cross validation and bootstrap in order to evaluate the resulting model from the various regression methods. Code written for OLS and Ridge regression, and also for both of the aforementioned resampling methods is located in the linked Github repository. For Lasso regression, functionality from Scikit-Learn was used.

The subsequent sections of the project starts with an outline of the theory and method used to produce the results in the present work. Also information about the datasets used in the project is provided. Thereafter the results are presented and discussed. The report is ended with a conclusion.

# **2 Method**

This sections outlines the theory and method behind the regression analysis presented in the report. Including regression methods, measurements for assessing the models and resampling techniques.

## 2.1 Linear regression

Given a set of  $p$  independent variables, also called characteristics or features, of  $n$  samples organized in a so called design matrix  $\mathbf{X}$ . In addition where  $\mathbf{y}$  is the observable response or outcome. If we assume a linear relationship between the independent variables of  $\mathbf{X}$  and the response  $\mathbf{y}$ , we can fit a model using linear regression. Meaning  $\mathbf{y}$  can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

The design matrix  $\mathbf{X}$  is a matrix of size  $n \times p$ . The response  $\mathbf{y}$  is a vector of size  $n$  and the parameter vector  $\boldsymbol{\beta}$  containing the linear regression coefficients  $\beta_i$  of our model is of size  $p$ . The coefficients of  $\boldsymbol{\beta}$  are unknown to us. Lastly,  $\boldsymbol{\varepsilon}$  is a vector of size  $n$  and represents the error term or disturbance term or sometimes noise of our model.  $\boldsymbol{\varepsilon}$  contains the factors which influence the response  $\mathbf{y}$  other than the independent variables  $\mathbf{x}_i$ , which are the columns of  $\mathbf{X}$ .  $\boldsymbol{\varepsilon}$  is also assumed to be normally distributed,  $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$ .

We assume the non-trivial and more realistic situation where the  $\boldsymbol{\varepsilon}$  vector contains non-zero values. . Meaning there will be a deviation between our model, denoted by  $\tilde{\mathbf{y}}$ , and the observed values  $\mathbf{y}$ . We then write our model as

$$\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} \quad (2)$$

and the error or deviation between our model and the observed values  $\mathbf{y}$  can be written as

$$\begin{aligned} \boldsymbol{\varepsilon} &= \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y} - \tilde{\mathbf{y}} \end{aligned} \quad (3)$$

We want to our error to be as small as possible, meaning that our model most accurately models our data. Estimating the coefficients of  $\boldsymbol{\beta}$  such that  $\boldsymbol{\varepsilon}$  is minimized, is the goal of linear regression.

## 2.2 Ordinary Least Squares

To estimate the unknown coefficients of  $\boldsymbol{\beta}$  such that  $\boldsymbol{\varepsilon}$  is minimized, we require an expression for  $\boldsymbol{\beta}$ . Using the Mean Squared Error (MSE) with the Euclidean  $L^2$  norm for a vector space, we can write the cost function as

$$\begin{aligned} C(\boldsymbol{\beta}) &= \|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2 \\ &= \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \tilde{y}_i)^2 \\ &= \frac{1}{n} \{(\mathbf{y} - \tilde{\mathbf{y}})^T (\mathbf{y} - \tilde{\mathbf{y}})\} \\ &= \frac{1}{n} \{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\} \end{aligned} \quad (4)$$

The optimal coefficients of  $\beta$ , which we denote as  $\hat{\beta}$ , is therefore

$$\hat{\beta}^{OLS} = \underset{\beta}{\operatorname{argmin}} C(\beta) \quad (5)$$

In order to minimize  $C(\beta)$ , we can differentiate the function with respect to  $\beta$  and set the result equal zero. We can omit the constant  $\frac{1}{n}$ , since it will ultimately be multiplied away. Differentiating results in

$$\begin{aligned} \frac{\partial C(\beta)}{\partial \beta} &= \frac{\partial}{\partial \beta} ((\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)) \\ &= \mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0 \end{aligned}$$

and rewriting the equation yields the ordinary least squares estimator

$$\begin{aligned} \mathbf{X}^T \mathbf{y} &= \mathbf{X}^T \mathbf{X} \beta \\ \hat{\beta}^{OLS} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned} \quad (6)$$

Equation (6), if the product  $\mathbf{X}^T \mathbf{X}$  is invertible, provides the optimal coefficients. In section 2.1, the design matrix was defined as  $\mathbf{X} \in \mathbb{R}^{n \times p}$  where  $n$  samples can become quite large in our dataset because it is usually the case that  $n \gg p$ . Thankfully and important to note is that the product  $\mathbf{X}^T \mathbf{X} \in \mathbb{R}^{p \times p}$ , is much less computationally intensive to invert seeing as  $p$  normally is relatively small.

Using the result from equation (6), we can calculate our model with equation (2)

$$\tilde{\mathbf{y}} = \mathbf{X} \hat{\beta} \quad (7)$$

$\tilde{\mathbf{y}}$  is our model of predicted values.

### 2.3 Ridge regression

Ridge regression is similar to OLS by way of obtaining the coefficients of  $\beta$ , but for ridge regression a regularization parameter  $\lambda$  is added. The cost function for ridge regression then becomes

$$\begin{aligned} C(\beta) &= \|\mathbf{y} - \tilde{\mathbf{y}}\|_2^2 + \lambda \|\beta\|_2^2 \\ &= \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \\ &= \frac{1}{n} \{(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)\} + \lambda \beta^T \beta \end{aligned} \quad (8)$$

Similarly to OLS, we differentiate with respect to  $\beta$ . Resulting in an expression for the optimal values of  $\beta^{Ridge}$ .

$$\begin{aligned} \frac{\partial C(\beta)}{\partial \beta} &= \frac{\partial}{\partial \beta} ((\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta) \\ &= \mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta = 0 \end{aligned}$$

and rewriting yields the ridge estimator

$$\begin{aligned}\mathbf{X}^T \mathbf{y} &= \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \lambda \boldsymbol{\beta} \\ \mathbf{X}^T \mathbf{y} &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}}^{Ridge} &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}\end{aligned}\tag{9}$$

where  $\mathbf{I}$  is the identity matrix. With the constraints

$$\begin{aligned}\lambda &\geq 0 \\ \sum_{i=0}^{p-1} \beta_i^2 &\leq t\end{aligned}$$

where  $t$  is a finite positive number. Also notice that for  $\lambda = 0$ , the estimator becomes an OLS estimator.

OLS simply finds the unbiased coefficients of  $\boldsymbol{\beta}$ , meaning every feature (column) of the design matrix  $\mathbf{X}$  is evaluated equally. The method does not consider if a number of the features are more important than the others. For ridge regression on the other hand you are able to tune the regularization parameter  $\lambda$ , which will lead to different coefficients. Ridge regression shrinks the coordinates  $\mathbf{y}$ . To better understand why this is the case, we must first introduce Singular Value Decomposition (SVD).

### 2.3.1 Singular Value Decomposition

In the situation where the product  $\mathbf{X}^T \mathbf{X}$  is non-invertible, meaning that the columns of  $\mathbf{X}$  are linearly dependent and therefore we have no solution to equation (6), we have the option of computing the pseudo-inverse matrix using the Singular Value Decomposition algorithm.

From the textbook *Linear Algebra and its Applications* by David C. Lays, et al. [2]:

#### Singular Value Decomposition

Let  $A$  be an  $m \times n$  matrix with rank  $r$ . Then there exist an  $m \times n$  matrix  $\Sigma$  for which the first  $r$  diagonal entries are the singular values of  $A$ ,  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ , and there exists an  $m \times m$  orthogonal matrix  $U$  and an  $n \times n$  orthogonal matrix  $V$  such that

$$A = U \Sigma V^T$$

The singular values of  $A$  are the square root of the eigenvalues of  $A^T A$ . The columns of matrix  $U$  are the eigenvectors of the product  $A^T A$ . While the columns of matrix  $V$  are the eigenvectors of the product  $A A^T$ . The columns of  $U$  forms an orthonormal basis for  $\text{Col}(A)$  and the columns of  $V$  form an orthonormal basis for  $\text{Row}(A)$ . Because  $U$  and  $V$  are orthogonal we can use in the following section the properties,  $U^T U = I$  and  $V V^T = I$ .

### 2.3.2 OLS and ridge regression with SVD

Applying SVD to  $\mathbf{X}^T \mathbf{X}$ , where  $\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T$ , we get the following

$$\begin{aligned} \mathbf{X}^T \mathbf{X} &= (\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T)^T (\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T) \\ &= \mathbf{V} \mathbf{\Sigma}^T \mathbf{U}^T \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T \\ &= \mathbf{V} \mathbf{\Sigma}^2 \mathbf{V}^T \end{aligned} \quad (10)$$

Combining equation (6), (7) and (10), we get for OLS

$$\begin{aligned} \tilde{\mathbf{y}}^{OLS} &= \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ \tilde{\mathbf{y}}^{OLS} &= \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T (\mathbf{V} \mathbf{\Sigma}^2 \mathbf{V}^T)^{-1} (\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T)^T \mathbf{y} \\ \tilde{\mathbf{y}}^{OLS} &= \mathbf{U} \mathbf{U}^T \mathbf{y} = \sum_{j=1}^p \mathbf{u}_j \mathbf{u}_j^T \mathbf{y} \end{aligned} \quad (11)$$

Now doing the same for ridge regression, we get

$$\begin{aligned} \tilde{\mathbf{y}}^{Ridge} &= \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \\ \tilde{\mathbf{y}}^{Ridge} &= \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T (\mathbf{V} \mathbf{\Sigma}^2 \mathbf{V}^T + \lambda \mathbf{I})^{-1} (\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T)^T \mathbf{y} \\ \tilde{\mathbf{y}}^{Ridge} &= \mathbf{U} \mathbf{U}^T \mathbf{y} = \sum_{j=1}^p \mathbf{u}_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \mathbf{u}_j^T \mathbf{y} \end{aligned} \quad (12)$$

Since  $\lambda \geq 0$ , we must have

$$\frac{\sigma_j^2}{\sigma_j^2 + \lambda} \leq 1 \quad (13)$$

As mentioned previously, ridge regression shrinks the coordinates of  $\mathbf{y}$ . Comparing equation (11) and (12), we can see that ridge regression shrinks  $\mathbf{y}$  with respect to the orthonormal basis  $\mathbf{U}$  by a factor of  $\frac{\sigma_j^2}{\sigma_j^2 + \lambda}$ . Recalling from the section about SVD, that the square root of the eigenvalues are ordered diagonally in a descending order such that  $\sigma_i \geq \sigma_{i+1}$ .  $\sigma_i$  with smaller values are less important because they contribute less to  $\mathbf{y}$  and are therefore applied more shrinkage than  $\sigma_i$  with larger values. Essentially this means that ridge regression shrinks the coordinates of  $\mathbf{y}$ , but shrinks the features that are less important than the features which are more important.

## **2.4 Lasso regression**

## **2.5 Design Matrix**

## **3 Datasets**

## **4 Results**

## **5 Discussion**

## **6 Conclusion**

## **References**

- [1] Bender, R. (2009). Introduction to the use of regression models in epidemiology. *Methods in molecular biology* (Clifton, N.J.), 471, 179–195. <https://pubmed.ncbi.nlm.nih.gov/19109780/>
- [2] Lay, David C., et al. *Linear Algebra and its Applications*, Global Edition. Pearson. (2018).