

# Regression analysis and re-sampling methods

**Johan Nereng**

Department of Physics, University of Oslo, Norway

Sep 30, 2019

**Abstract**

# 1 Introduction

The main aim of this project is to study in more detail various regression methods, including the Ordinary Least Squares (OLS) method, Ridge regression and finally Lasso regression, including use of resampling. Start by fitting polynomials to Franke's function (see Methods).

Having established the model and method, the models are further evaluated using resampling techniques such as cross-validation.

choice of modelling based on insight. Terrain possibly polynomial ? use franke function to test algorithms.

**Project flow:** This project starts by introducing

**Main findings**

## 2 Methods

### 2.1 Multiple Linear Regression

A multiple variable polynomial of the form

The linearity in multiple linear regression relates to the linear relationship between the regression coefficients  $\beta_i$ .

Linear regression models the relationship between the response or target and the predictors linearly. In this project, three different methods are applied to fit the model to the data. Ordinary Least Squares (OLS), Ridge, and Lasso. Since this project assumes that the

In order to apply these methods effectively, the regression problem is represented by a matrix vector multipl

Using a polynomial fit of degree  $p$ , the

Later in this project,

$$x_1, x_2 \quad (1)$$

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{m-1} \\ 2 & x_2 & x_2^2 & \dots & x_2^{m-1} \\ & & \dots & & \\ n & x_n & x_n^2 & \dots & x_n^{m-1} \end{bmatrix} \quad (2)$$

corresponding to the choice of  $m$  may be constructed, along with a vector  $\mathbf{x} = [c_1, c_2, \dots, c_m]$ , containing the coefficients of the polynomial. If  $\mathbf{b} = [y_1, y_2, \dots, y_n]$ , then  $S = \sum_{i=1}^n (p(x) - y_i)^2$  may be expressed as  $S = \|\mathbf{Ax} - \mathbf{b}\|_2^2$ . In order to minimize  $S$ , this assignment utilizes two methods: a QR factorization of  $A$ , and solving the normal equations  $A^T A \mathbf{x} = A^T \mathbf{b}$  using Cholesky factorization and forward substitution.

Cost function e define a function which gives a measure of the spread between the values  $y_i$  (which represent hopefully the exact values) and the parameterized values, namely

hva er tilde y, epsilon etc.

### 2.1.1 Ordinary Least Squares (OLS)

OLS fits a function by minimizing the sum of the squares of the errors between a model and a data set. This results in the cost function

$$C(\beta) = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2 = \frac{1}{n} (\mathbf{y} - \mathbf{X}^T \beta)^T (\mathbf{y} - \mathbf{X}^T \beta) \quad (3)$$

Which, if A has full column rank, has a unique solution [1][p.144] which satisfies the normal equations:

$$(\mathbf{X}^T \mathbf{X}) \mathbf{b} = \mathbf{X}^T \mathbf{y} \quad (4)$$

This corresponds to finding the minima of the derivative of the cost function with respect to  $\beta$ . To which the solution is

$$\beta^{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{A}^\dagger \mathbf{y}. \quad (5)$$

This matrix inversion may then be carried out using for example SVD or LU decomposition. Alternatively, minimizing (3) may be accomplished using orthogonal transformation with QR decomposition, the details of which may be found in chapter one of Uri M. Ascher and Chen Grief's book [1]. The choice between the two approaches hinges on computational stability versus computational speed.

**Computational Speed:** The two approaches are as mentioned dissimilar in computational cost, measured in floating point operations (flops) - especially in overdetermined cases, where the matrix  $\mathbf{X} \in \mathbb{R}^{n,m}$ , has  $m < n$ . The normal equation solution uses  $mn^2 + (1/3)n^3$  flops [1][p.146], while the QR decomposition requires  $2mn^2 - (2/3)n^3$  flops [1][p.155]. Which means that the normal equations method is significantly faster than the QR approach.

**Computational Stability:** The conditioning of the two approaches, or stability with respect to input perturbation, are measured in conditioning number of the matrix representing the problem,  $K(\mathbf{X})$ . Higher  $K(\mathbf{X})$  means less stable.

$$K(\mathbf{X}) = \|\mathbf{X}^{-1}\| \times \|\mathbf{X}\| \quad (6)$$

In order to acquire the normal equations, the product of  $\mathbf{X}$  and it's transpose is used. Hence, it's conditioning number goes as  $K(\mathbf{X}^T \mathbf{X}) = K(\mathbf{X})^2$ . This squaring of the conditioning number is not required for the QR approach, which means that the QR method has a lower conditioning number, and is thus the more stable method.

confidence interval for beta hastie et all p.47

choice of model - regression type - linear etc. polynomials - understanding of the problem at hand. choice of minimization evaluation of betas type of regression method, OLS etc.

### 2.1.2 Ridge Regression

Alterations to the OLS cost function (3) may be done in order to assert control of the [bias variance trade off](#). One such alteration is known as Ridge Regression, with the following cost function [3][p.22]:

$$C(\boldsymbol{\beta}) = \frac{1}{n}(\mathbf{y} - \mathbf{X}^T \boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}^T \boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \quad (7)$$

### 2.1.3 Lasso Regression

## 2.2 Re-sampling techniques

## 2.3 Bias Variance trade-off

## 2.4 Algorithms

## 2.5 Franke's function

Describe Franke's function.

Polynomial degree	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_6$
0	$0.44 \pm 0.13$	-	-	-	-	-	-
1	$0.44 \pm 0.13$	$6.06 \pm 1.08$	$4.44 \pm 0.99$	-	-	-	-
2	$0.44 \pm 0.13$	$6.06 \pm 1.08$	$4.44 \pm 0.99$	$-27.72 \pm 3.38$	$-3.67 \pm 2.71$	$-16.86 \pm 3.24$	-
3	$0.44 \pm 0.13$	$6.06 \pm 1.08$	$4.44 \pm 0.99$	$-27.72 \pm 3.38$	$-3.67 \pm 2.71$	$-16.86 \pm 3.24$	37.79
4	$0.44 \pm 0.13$	$6.06 \pm 1.08$	$4.44 \pm 0.99$	$-27.72 \pm 3.38$	$-3.67 \pm 2.71$	$-16.86 \pm 3.24$	37.79

# 3 Discussion of results

[2]

## 3.1 Conclusions

## References

- [1] Uri M. Ascher and Chen Greif. Linear least squares problems. In *A First Course on Numerical Methods*, pages 141–166. Society for Industrial and Applied Mathematics, 2011.
- [2] Morten Hjorth-Jensen. Lectures notes in fys-stk4155. data analysis and machine learning: Linear regression and more advanced regression analysis, Sep 13 2019.
- [3] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810:1–124, 2019.