Project 1

Edvard B. Rørnes,* Anton A. Brekke,† and Isak O. Rukan‡
Institute of Physics, University of Oslo,
0371 Oslo, Norway
(Dated: September 3, 2024)

Abstracting very cool

CONTENTS

1. INTRODUCTION

The methods used in this project are Ordinary Least Squares (OLS), Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) regression.

2. THEORY

The general structure of all our models is that we have some data set $\{x_i, y_i\}$ where $i \in \{0, 1..., n-1\}$ where x_i are independent variables whilst y_i are dependent variables. The data is assumed to be described by

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon} \tag{1}$$

where f is some continuous function which takes $\boldsymbol{x}$ as input and $\boldsymbol{\varepsilon}$ is a normal distributed error $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$.

The function f will then be approximated with a model $\tilde{\boldsymbol{y}}$ in which we will consider a polynomial expansion with coefficients β_i:

$$\tilde{y}_i = \sum_{j=0}^{p-1} \beta_j x_i^j \tag{2}$$

defining the $n \times p$ design matrix $(\boldsymbol{X})_{ij} = (x_i)^j$ we can rewrite this as

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} \tag{3}$$

Further, each model will be defined with a different cost function $C(\boldsymbol{\beta})$ which we minimize to find the coefficients for each respective model.

2.1. OLS

OLS is a primitive method used in linear regression to estimate coefficients of a linear model. The cost function in OLS is simply defined as the residual sum of squares (RSS)

$$C_{\text{OLS}}(\boldsymbol{\beta}) = \text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \tilde{\boldsymbol{y}})^2 = (y_i - X_{ij}\beta_j)^2$$

where we employ the summation notation where repeated indices are summed over. As mentioned prior, the coefficients $\boldsymbol{\beta}$ are found by minimizing the cost function, i.e. taking the derivative w.r.t. $\boldsymbol{\beta}$. This results in

$$\boldsymbol{\beta}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

which yields the model

$$\tilde{\boldsymbol{y}}_{\text{OLS}} = \boldsymbol{X}\boldsymbol{\beta}_{\text{OLS}} \tag{4}$$

Assuming our data takes the form of (1) then the expectation value $\boldsymbol{y}$ is

$$\mathbb{E}(y_i) = \mathbb{E}(f(x_i)) = \boldsymbol{X}_{i,*}\beta$$

since $\mathbb{E}(\varepsilon_i) = 0$ follows from its definition. The variance of $\boldsymbol{y}$ is given by

$$\begin{aligned}\text{Var}(y_i) &= \mathbb{E}\{[y_i - \mathbb{E}(y_i)]^2\} = \mathbb{E}\{(\boldsymbol{X}_{i,*}\beta + \varepsilon_i)^2\} - (\boldsymbol{X}_{i,*}\boldsymbol{\beta})^2 \\ &= (\boldsymbol{X}_{i,*}\boldsymbol{\beta})^2 + \mathbb{E}(\varepsilon_i^2) + 2\mathbb{E}(\varepsilon_i)\boldsymbol{X}_{i,*}\boldsymbol{\beta} - (\boldsymbol{X}_{i,*}\boldsymbol{\beta})^2 \\ &= \text{Var}(\varepsilon_i^2) = \sigma^2\end{aligned}$$

* e.b.rornes@fys.uio.no
† asdf
‡ asdfafs

which shows that $y_i \sim \mathcal{N}(\boldsymbol{X}_{i,*}\boldsymbol{\beta}, \sigma^2)$. The expectation value of the optimal parameters $\hat{\boldsymbol{\beta}}$ can be found to be

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) &= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}[\boldsymbol{y}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.\end{aligned}$$

with the variance

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) &= \mathbb{E}\{[\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta})][\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta})]^T\} \\ &= \mathbb{E}\{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{\beta}] \\ &\quad \times [(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{\beta}]^T\} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}\{\boldsymbol{y}\boldsymbol{y}^T\}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &\quad - \boldsymbol{\beta}\boldsymbol{\beta}^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T + \sigma^2]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &\quad - \boldsymbol{\beta}\boldsymbol{\beta}^T \\ &= \boldsymbol{\beta}\boldsymbol{\beta}^T + \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^T \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

2.2. Ridge

Ridge regression is an extension of OLS where define the cost function as a modified version of the OLS cost function with an added penalty term which is proportional to the coefficients β_i^2:

$$C_{\text{Ridge}}(\boldsymbol{\beta}) = C_{\text{OLS}}(\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^2 \tag{5}$$

Here $\lambda \geq 0$ is a regularization parameter which controls the strength of this additional penalty. This regulator essentially drives the magnitude of these coefficients allowing for more tweaking in the parameter space. This parametrization of course includes the constraint that $\boldsymbol{\beta}^2 \leq t$ for some $t < \infty$ such that we can choose our arbitrary parameter $\lambda \geq 0$ to be sufficiently small s.t. the cost function (5) does not diverge. The optimal parameters for Ridge regressions can again be found by the same process as for OLS:

$$\begin{aligned}0 = \frac{\partial C_{\text{Ridge}}}{\partial\boldsymbol{\beta}} &= -\frac{2}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X} + 2\lambda\boldsymbol{\beta}^T \\ &= \frac{2}{n}(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{y}^T\boldsymbol{X}) + 2\lambda\boldsymbol{\beta}^T \\ 0 &= \boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X} + \tilde{\lambda}\boldsymbol{I}) - \boldsymbol{y}^T\boldsymbol{X} \\ \boldsymbol{\beta}^T &= \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \tilde{\lambda}\boldsymbol{I})^{-1} \\ \boldsymbol{\beta} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

where we defined $\tilde{\lambda} \equiv n\lambda$, renamed $\tilde{\lambda} \to \lambda$ and used that the matrix in the parenthesis is a symmetric matrix and thus its inverse must also be symmetric. Here we can see that the effect of adding this pentalty term is essentially taking $(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$ when compared to the OLS case. In the past this was generally the starting point for Ridge regression in the cases where the matrix $\boldsymbol{X}^T\boldsymbol{X}$ was not invertible. (If we want to keep this next part we need to mention SVD) A direct way of seeing the effect of the regulator is by considering

$$\begin{aligned}\tilde{\boldsymbol{y}}_{\text{Ridge}} &= \boldsymbol{X}\boldsymbol{\beta}_{\text{Ridge}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T((\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T + \lambda\boldsymbol{I})^{-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\boldsymbol{I})\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y} \\ &= \sum_{j=0}^{p-1}\boldsymbol{u}_j\boldsymbol{u}_j^T\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\boldsymbol{y}\end{aligned}$$

where the last step is valid due to the orthogonality of $\boldsymbol{U}$ and σ_j are the elements on the diagonal of $\boldsymbol{\Sigma}$. Since $\lambda \geq 0$ then this added factor compared to OSL is ≤ 1. The larger λ is the smaller this factor becomes and is the so-called a "shrinkage" factor.

2.3. LASSO

Similarly to Ridge, LASSO also includes a penalty factor. The cost function in this case is instead defined to be

$$C_{\text{Ridge}}(\boldsymbol{\beta}) = C_{\text{OLS}}(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1 \tag{6}$$

where

$$||\boldsymbol{\beta}||_k \equiv \sum_{i=0}^{n-1}|\beta_i|^k$$

is the L^k norm of $\boldsymbol{\beta}$. This has the added benefit of being able to set certain parameters to be 0 instead of suppressing them, at the cost of losing analytical expressions in non-trivial cases.

Show that you can rewrite

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

as

$$\mathbb{E}[(\boldsymbol{y} - \tilde{\boldsymbol{y}})^2] = \text{Bias}[\tilde{y}] + \text{Var}[\tilde{y}] + \sigma^2$$

where

$$\text{Bias}[\tilde{y}] = \mathbb{E}\big[(\boldsymbol{y} - \mathbb{E}[\tilde{\boldsymbol{y}}])^2\big]$$

and

$$\text{Var}[\tilde{y}] = \mathbb{E}\big[(\tilde{\boldsymbol{y}} - \mathbb{E}[\tilde{\boldsymbol{y}}])^2\big] = \frac{1}{n}\sum_{i}(\tilde{y}_i - \mathbb{E}[\tilde{\boldsymbol{y}}])^2$$