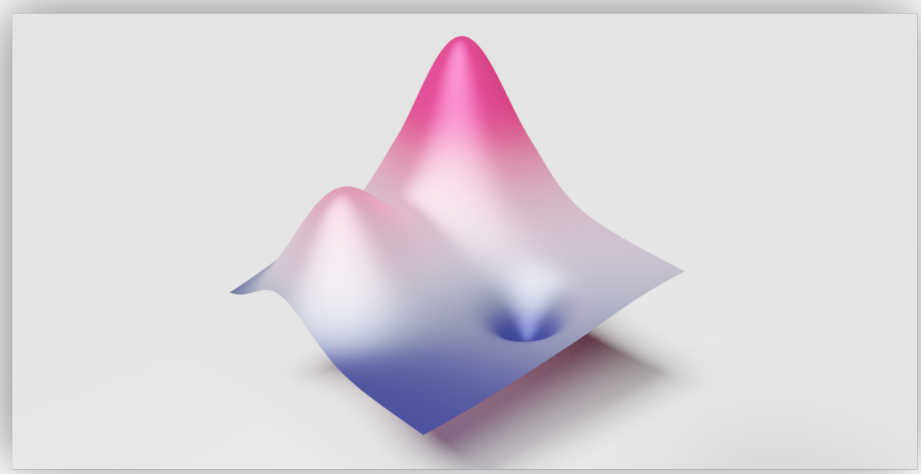

Regression And Resampling Techniques

FYS-STK4155

Assignment One

Ivar Haugaløkken Stangeby

September 10, 2019

Contents

1 Introduction

Regression analysis is the act of estimating relationships among variables. In this project, we study various regression methods in more detail. In particular, we compare the *ordinary least squares* (OLS) method with the *Ridge regression* and *Lasso regression* techniques. As we shall see, these three methods are all variations over the same theme. We start by testing the methods on noisy data sampled from a function known as *Franke's* bivariate test function (visualized on the title page), which has been in widespread use in the evaluation of interpolation and data fitting techniques. Finally, we run regression on real terrain data, comparing the aforementioned methods.

2 Regression Techniques

In general, the goal of regression analysis is to fit a *model function* $f(\boldsymbol{x}, \beta)$ to a set of n data points $\Omega = (\boldsymbol{x}_i, y_i)_{i=1}^n$. A simple example is that of a linear polynomial with two parameters:

$$f(x, \beta) = \beta_0 + \beta_1 x. \tag{1}$$

The *model parameters* β are determined in order to minimize a suitable *cost function* $C(\Omega, \beta)$ which measures to which extent the model function manages to capture trends in the data Ω. It is the choice of cost function $C(\Omega, \beta)$ which distinguishes the three regression techniques, OLS, Lasso regression, and Ridge regression.

Assumption 1. It is often assumed a priori that the data is infact generated from a noisy model, such that each y_i can be described as

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i \tag{2}$$

where each $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is normally distributed with zero mean and variance σ^2. This assumption on the error gives rise to what is known as *general linear models*.

The design matrix. We are often interested in finding the best model function in a specific function space. Assuming this space has a basis $\mathcal{B} = \left\{ \varphi_i \right\}_{i=1}^{M}$, we may write our model function in terms of the basis functions and the model parameters as

$$f(\boldsymbol{x}, \beta) := \sum_{i=1}^{M} \beta_i \varphi_i(\boldsymbol{x}). \tag{3}$$

As an example, if we were to use the space Π_2^2 of bi-quadratic polynomials, our basis functions would be

$$\mathcal{B} = \left\{ 1, x, x^2, y, y^2, xy \right\}. \tag{4}$$

With this formulation, we can represent the approximation $\hat{\boldsymbol{y}}$ to $\boldsymbol{y}$ as a matrix product

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\beta \tag{5}$$

where $\boldsymbol{X}$ is the *design matrix* defined by components $\boldsymbol{X}_{ij} = \varphi_j(\boldsymbol{x}_i)$, and $\beta = [\beta_1, \dots, \beta_M]$ is the model parameters.

Performance metrics. In order to evaluate the how accurately the model function $f(\Omega, \beta)$ captures trends in the target data $\boldsymbol{y}$, a few standard performance metrics are used. Firstly, the *mean squared error*:

$$\mathrm{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

which simply averages the squared error over all samples and estimates. Secondly, we have the *coefficient of determination* or R^2*-score*:

$$\mathrm{R}^2(\boldsymbol{y}, \hat{y}) := 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}. \tag{7}$$

The R^2-score measures much of the variation in $\boldsymbol{y}$ which can be attributed to a simple linear relation between $\boldsymbol{x}$ and $\boldsymbol{y}$. It is a ratio, thus a value of one tells us that all the variation in the data can be attributed to an approximate linear relationship between the data $\boldsymbol{x}$ and the response variable $\boldsymbol{y}$. A value of zero indicates that a non-linear model may be preferrable.

2.1 Ordinary Least Squares

One common cost function is the one which involves the sum of squared residuals (or squared errors):

$$C(\Omega, \beta) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

where $\hat{y}_i := f(\boldsymbol{x}_i, \beta)$. The method involving the minimization of this specific cost function is known as *least squares*. With the matrix notation from above in mind, we can also write the cost function as

$$C(\Omega, \beta) = \frac{1}{n}(\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}). \tag{9}$$

Optimizing the parameters in OLS.

Assume now that our model is of the form given in Equation (3) and that our cost function is the mean squared error defined in Equation (8). We are interested in finding the parameters β that minimize the cost function $C(\Omega, \beta)$. Since $C(\Omega, \beta)$ is convex, it suffices to differentiate with respect to β and equating to zero.

We have that

$$\frac{\partial C(\Omega, \beta)}{\partial \beta} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta), \tag{10}$$

and equating this to zero yields the familiar *normal equations*

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\beta. \tag{11}$$

If the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, we may obtain the solution by direct numerical matrix inversion. In this case, the optimal model parameters are found directly by

$$\beta_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{12}$$

However, these matrices may be ill-conditioned when the number of equations are very large, and it is therefore common to apply approximate solvers for the inverse, for instance using low-rank approximation based on SVD-decomposition, which we will briefly turn to in the following.

Singular value decomposition.

Recall that any matrix $\boldsymbol{A} \in \mathbb{C}^{n,m}$ can be decomposed as

$$\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^T, \tag{13}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are comprised of the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$ respectively. As these eigenvectors are orthonormal, it follows that both $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices. Furthermore, Σ is a matrix

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{14}$$

where Σ_1 is a square diagonal matrix of size $r \times r$ with the non-zero singular values of $\boldsymbol{A}$. The integer r is the *rank* of $\boldsymbol{A}$. As the matrix Σ is mostly containing zeros, the information stored in $\boldsymbol{A}$ is attributed to only some parts of $\boldsymbol{U}$ and $\boldsymbol{V}$. We can remove the redundant parts, and more compactly express $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{U}_1\Sigma_1\boldsymbol{V}_1^T$ without loss of accuracy in the decomposition. Here, $\boldsymbol{U}_1$ is $m \times r$ and $\boldsymbol{V}_1$ is $n \times r$.

Applying the singular value decomposition to the matrix $\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ we can analyze the expression for the prediction $\hat{\boldsymbol{y}}$ in terms of the matrix $\boldsymbol{X}^T\boldsymbol{X}$. First of all, we have that

$$\begin{aligned} \boldsymbol{X}^T\boldsymbol{X} &= \boldsymbol{V}\Sigma^T\boldsymbol{U}^T\boldsymbol{U}\Sigma\boldsymbol{V}^T \\ &= \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T, \end{aligned} \tag{15}$$

where we have used that $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ and defined $\boldsymbol{D} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$. Thus, plugging this into Equations (5) and (12), we obtain the expression

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\beta = \boldsymbol{X}(\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{16}$$

Finally, substituting $\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ and noting that diagonal matrices always commute, we end up with

$$\hat{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{y}. \tag{17}$$

The bias-variance tradeoff

Since we are attempting to estimate a true distribution based on finite samples, there are certain statistical properties it is worthwile discussing. We follow the derivation in.

Recall Assumption 1 which states that our true data is generated from a noisy model

$$y = f(\boldsymbol{x}) + \varepsilon \tag{18}$$

with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We select a finite sample from the true data, which constitutes our finite data-set Ω. Furthermore, we have a model function $f(\Omega, \beta)$ which depends on the selected data-set and the model parameters β. Thus, our cost-function C will depend on the specific sample taken from the true distribution. It is of interest to examine the expected out-of-sample error $\mathbb{E}_{\Omega,\varepsilon}[C(\Omega, \beta)]$ of the model function over the sampled data set Ω and the noise ε. This can be decomposed into three distinct parts, *variance*, *bias*, and *noise*:

$$\mathbb{E}_{\Omega,\varepsilon}[C(\Omega, \beta)] = \frac{1}{n}\sum_{i=1}^{n}\left(f(\boldsymbol{x}_i) - \mathbb{E}_\Omega\left[\hat{y}_i\right]\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\Omega\left[\left(\hat{y}_i - \mathbb{E}_\Omega[\hat{y}_i]\right)^2\right] + \sigma^2. \tag{19}$$

See Appendix A for the full derivation. The bias-term measures the deviation of the expectation of the model function $f(\Omega, \beta)$ from the true value f, while the variance-term measures the fluctuations arising due to using finite samples.

2.2 Ridge Regression

When the dimensionality of the matrix $\boldsymbol{X}$ is high, the problem of near-collinear columns often arise, as mentioned earlier. Thus, the model parameters β_{OLS} of the ordinary least squares regression cannot be determined. The propsed fix to this problem is to slightly perturb the matrix $\boldsymbol{X}^T\boldsymbol{X}$ by adding small positive values along the diagonal to ensure its invertability. This is known as *ridge regression* and yields a model parameter estimator

$$\beta_{\text{ridge}}(\lambda) = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{20}$$

where $\lambda \geq 0$ is a hyper parameter. To see why the resulting modified matrix is invertible, one has to look at the singular values. It turns out that the model parameters $\beta_{\text{ridge}}(\lambda)$ directly minimize the *ridge loss function* defined as

$$C_{\text{ridge}}(\Omega, \beta; \lambda) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{i=1}^{p}\beta_i^2 = C(\Omega, \beta) + \lambda\beta^T\beta, \tag{21}$$

which we recognize as standard penalized least squares.

Ridge regregression carries several benefits. One of the more important is its resilience to overfitting. In a high-dimensional setting, if many of the data-features can be attributed to a few covariates, then the remaining covariates do nothing more than fit to the noise. Large values of β_i then a good indication that your model might be overfitting as some covariates are given more weight than others. The Ridge regression solves this by directly penalizing large model parameters in the cost function.

A Derivation of the bias-variance-noise relationship

Combining the two previous equations, we obtain