Regression analysis and resampling methods

FYS-STK3155: Project 1

Kim-Roger Grønmo
http://github.com/kimgronmo/

October 10, 2020

Abstract

We investigate the suitability of several regression methods for building models to reproduce data from the Franke function and also real world terrain data. The methods used include Ordinary Least Squares (OLS) regression, Ridge regression and Lasso regression. Varying degrees of polynomials are used to fit the data in our models. We also use techniques such as bootstrap and cross-validation to further assess these models. In particular we investigate the bias-variance trade off as a function of model complexity. We find that the Franke function is best represented using OLS regression and the real world data also by the OLS regression model. Our investigation reveal that this is most likely due to our predicted data being situated in a region where bias dominates the mean square error and we would need to use more complex models than the ones currently being investigated to better fit the data sets to the models.

1 Introduction

We are living in a world with abundant access to data. As our ability to collect and visualize this data increases, our inherent need to understand which causes and effects lies behind what we observe in the real world become more apparent. With the advent of both modern computers and mobile technology we have the tools needed to both collect and analyze the data generated by the real world.

Imagine that we wanted to calculate the average annual salary for a person. It is quite apparent that there a great many factor that influence what a person earns. A few example can be age, gender, level of education, degree earned, years of work experience, the general job market within the given field and so on. In order to try to model these kind of situations with statistics one often chooses linear regression methods.

A basic premise of regression modelling is that the observations we collect, the response, is dependent on some predictor variables. For linear regression models it is assumed that there is a linear relationship between these predictor variables and the responses they generate. We then have the ability to make models which uses this linearity to make a design matrix to predict the effect these factors have on the response variable. We can use linear algebra to use such a design matrix in an optimal combination with the given factors to generate responses that predict real world obseervations. The ability to minimize the error between these predictions and observations using modern computers make Linear Regression Models popular.

2 Regression Models

The general reference for this section is reference [1].

For the least squares method we are evaluating a situation in which p characteristics of n samples are measured. The *response* is denoted $\mathbf{y}$: a vector with size n. The measured characteristics, denoted the predictors, we will organize in a matrix $\mathbf{X}$ of size $n \times p$. This matrix is referred to as the *design matrix*.

In order to explain the relatioship between the response and the predictor variables we will use a function $\mathbf{y}(\mathbf{X})$. When there is a linear relationship between $\mathbf{X}$ and $\mathbf{y}$ then the single response can be written as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{1}$$

where ε is the deviation of the linear model $\mathbf{X}\beta$ and the response $\mathbf{y}$. β is a vector containing the linear regression coefficients β_i. In our model we denote the generated response variables as $\tilde{\mathbf{y}}$, Then

$$\tilde{\mathbf{y}} = \mathbf{X}\beta \qquad\qquad = \mathbf{y} - \varepsilon.$$

For our model we want to calculate β in such a way that the error ε gets minimized.

2.1 The design Matrix

In our models we considered different sizes of the design matrix. For a model with polynomial 2 we use two predictors—we will denote them x and y—with the response y. With the intercept included this can be written as:

$$\mathrm{X} = \begin{bmatrix} 1 & x_1 & y_1 & x_1 y_1 \\ 1 & x_2 & y_2 & x_2 y_2 \\ \vdots & \vdots & \vdots & \\ 1 & x_n & x_n & x_n y_n \end{bmatrix}$$

2.2 Fitting the models

It can be shown (see reference[1]) that the best fits are given by:

$$\beta_{\mathrm{OLS}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

For Ridge regression:

$$\beta_{\mathrm{Ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{1}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{3}$$

For LASSO regression we used the build in functions of Scikit-Learn and thus did not solve for the optimal β manually

2.3 Resampling methods

The bootstrap method is a statistical technique for estimating quantities about a population by averaging estimates from multiple smaller data samples. These samples are constructed by drawing observations with replacement from our training set. The cross validation method splits up a data set into several mutually exclusive subsets. One of these subsets will in turn take the role of training set and the remaining union of subsets constitutes the training set. These methods was used to give better estimates to the mean square error, bias and variance for several model complexities. Further results are given in the github adress: http://github.com/kimgronmo

2.4 The Franke function

The Franke function $f(x, y)$ is defined by:

$$\begin{aligned} f(x,y) = &\frac{3}{4}\exp\left\{\frac{-1}{4}\left[(9x-2)^2-(9y-2)^2\right]\right\} \\ &+\frac{3}{4}\exp\left\{\frac{-1}{49}[(9x+1)^2+\frac{1}{10}(9y+1)^2\right]\} \\ &+\frac{1}{2}\exp\left\{\frac{-1}{4}\left[(9x-7)^2+(9y-3)^2\right]\right\} \\ &-\frac{1}{5}\exp\left\{-1\left[(9x+4)^2+(9y-7)^2\right]\right\}. \end{aligned} \tag{4}$$

2.5 Error Analysis

A common way to check how close a predicted response value is to the observed value is by calculating the Mean Square Error (MSE). It is defined by:

$$\mathrm{MSE}(\hat{y},\hat{\tilde{y}}) = \frac{1}{n}\sum_{i=0}^{n-1}(y_i-\tilde{y}_i)^2, \tag{5}$$

The parameters β were found by optimizing the mean squared error.

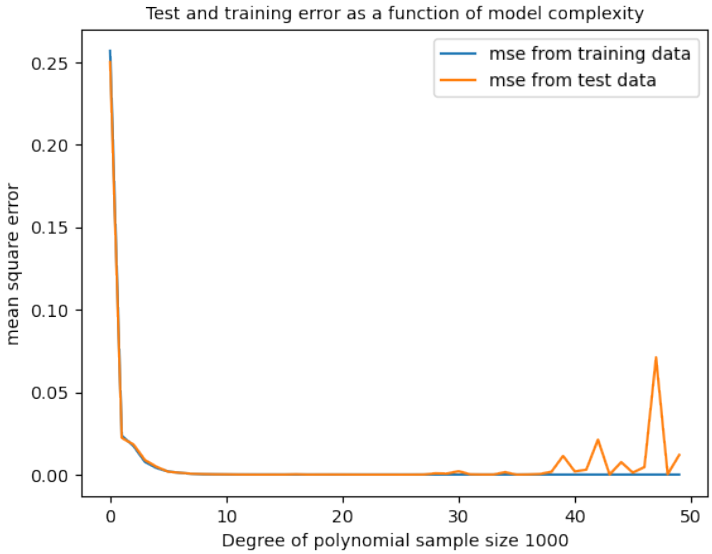

Figure 1:

The expected prediction error consist of the sum of the irreducible error (variance of the new test target) and the mean square error. The mean square error is broken up into a bias component and variance component.

As the model complexity increases the variance tend to increase and squared bias decreases. We can observe this effect in the given figure 1 and Figure 2. Note that when the sample size increases this effect is lowered. As we can see in Figure 3, Figure 4 and Figure 5 when the model complexity increases (higher degrees of polynomials are used to fit data), we get a bias-variance trade off. The mean square error which is largely dominated by the bias decreases as the complexity of the model increases. This comes at a cost of increased variance.

3 Results and Discussion

Further results for The Franke Function and the terrain data are included in the github folder. As we can see in Table 1 and Table 2 the results for the chosen regression models both show the same thing. The model that provides the best fit for the data sets generated is the Ordinary Least Square method, with the Ridge method second and LASSO in third. Since the complexity of the model was quite low this was to be expected.

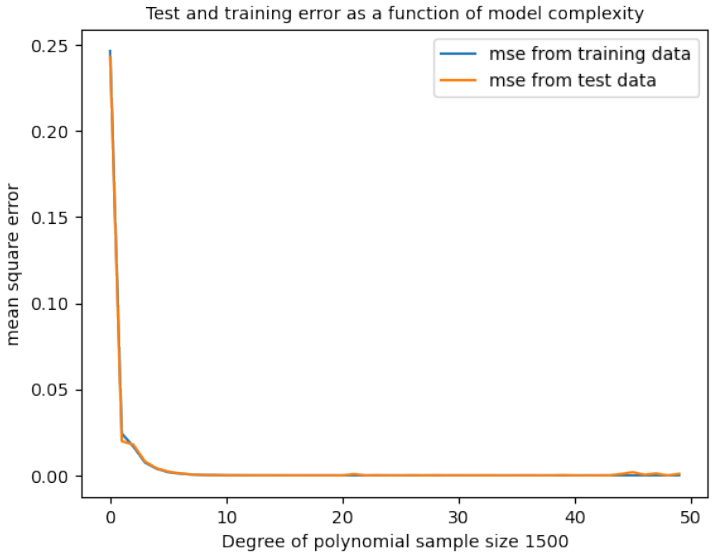

Figure 2:

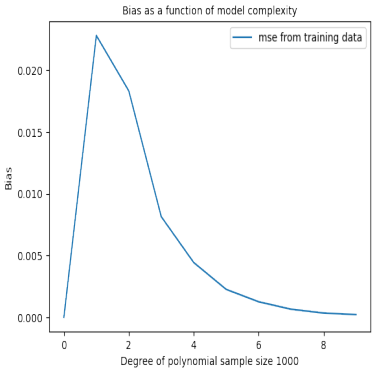

Figure 3:

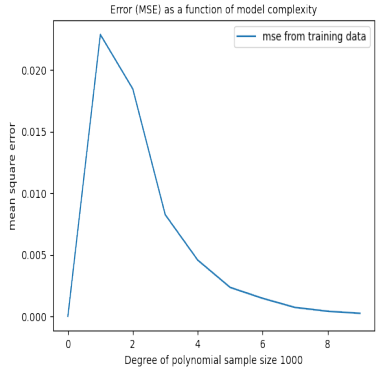

Figure 4:

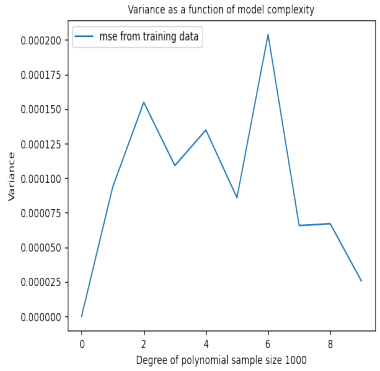

Figure 5:

Table 1: Minimum MSE for Polynomial 5 Franke Function

Regression method	**MSE Values**	**lambda**
LASSO	0.00775365	0.001
Ridge	0.00207766	0.001
OLS	0.00204132	Na

Table 2: Minimum MSE for Polynomial 5 Terrain Data

Regression method	**MSE Values**	**lambda**
LASSO	0.00095964	0.001
Ridge	0.00083987	0.001
OLS	0.00075617	Na

4 Conclusion

We have have performed linear regression fits for both the Franke function as well as real world terrain data. These fits indicate that ordinary least squares regression are better suited at fitting data than Ridge- and Lasso regression for both of our data sets. The reason for this seems to be that for the polynomial fit chosen the MSE are in a region where bias dominates the results and the effect of variance is very small.

We would need to use more complex models than the ones currently being investigated to better fit the data sets to the models. Then variance would matter more for the total error, and methods that aim at reducing this error such as Ridge- and LASSO regression would perform better.

5 Bibliography

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference, and prediction. 2009.