Regression Analysis and Resampling Methods

Hong Li

hong.li@geo.uio.no

October 7, 2018

Abstract

Regression is the most straightforward method to explore the relationships between a dependent variable and its independent variables. It helps us to understand and to predict how the dependent variable changes when the independent variables change. In this report, I perform three types of regression methods, i.e. Ordinary Least Squares (OLS), Ridge regression and Lasso regression to fit the Franke's function. I use Mean Squared Error (MSE) and R^2 score function to evaluate the regression models. Additionally, I also perform cross validation to test model stability. I compare the three regression methods based on two datasets. One is random generated from uniform distribution. Another dataset is Shuttle Radar Topography Mission (SRTM) digital terrain model (DTM). The results show that the three regression methods are comparable in fitting the Franke's function and in interpolating SRTM DTM. More specially, the Ridge regression gives slightly better results. However, it also gives the larges variation in estimate of parameters. The Lasso regression is just opposite. It gives the lowest performance, but highest stability in estimate of parameters. In interpolating the SRTM DTM, the model accuracy mainly depends on the terrain. When the terrain rises and falls, the model accuracy decreases. It is difficult to say which method is better. To comprise between model performance and parameter stability, the

OLS regression is recommended.

1 Introduction

Regression analysis is to estimate relationships among variables. It tells how the dependent variable changes with changes in one or more independent variables. The relationship can be linear or non-linear. The technique is widely used for forecasting and time series modeling in different fields. For example, linear regression is used in hydrology to find how runoff changes with change in precipitation and temperature. Regression is an important method and it can help us to understand how climate change affect water resources and flood.

There are many forms of regression and seven of them are most commonly used, i.e. Linear Regression, Logistic Regression, Polynomial Regression, stepwise Regression, Ridge Regression, Lasso Regression, ElasticNet Regression [3]. The main difference lies in data assumption and penalty function [1]. Therefore, the regression methods are suitable under different conditions and for different purposes.

The main aim of this project is to study three regression methods, including the Ordinary Least Squares (OLS) method, Ridge regression and Lasso regression. They are used to fit the Franke's function, which is a weighted sum of four exponential components. I use cross validation to evaluate the regression models. There are two datasets. One is random generated from uniform distribution and another is a patch of digital elevation model from Shuttle Radar Topography Mission (SRTM).

2 Methods

In this section, I introduce details of the Franke's function and the three regression methods.

2.1 Franke's function

The Franke's function is a weighted sum of four exponential as shown below. This function is widely used when testing various interpolation and fitting algorithms [2].

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

where $x, y \in [0, 1]$.

2.2 General linear models

Linear models can be written as

$$y = ax + b$$

The input x, parameters a and b, and y can be extended to vectors

$$\hat{y} = [y_0, y_1, y_2, \ldots, y_{n-1}]^T,$$

$$\hat{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_{n-1}]^T,$$

$$\hat{X} = \begin{bmatrix} 1 & x_0^1 & x_0^2 & \dots & \dots & x_0^{n-1} \\ 1 & x_1^1 & x_1^2 & \dots & \dots & x_1^{n-1} \\ 1 & x_2^1 & x_2^2 & \dots & \dots & x_2^{n-1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n-1}^1 & x_{n-1}^2 & \dots & \dots & x_{n-1}^{n-1} \end{bmatrix}$$

Therefore, the general model can be written as

$$\hat{y} = \hat{X}\hat{\beta}$$

In this model, $\hat{X}$ are inputs; β are parameters to estimate and $\hat{y}$ is used to estimate $\hat{\beta}$.

2.3 Ordinary Least Squares

The Ordinary Least Squares regression minimize the squared error to find the β. The objective function is the minimum value of

$$\left| y - \hat{X}\hat{\beta} \right|^2$$

If the matrix $\hat{X}^T\hat{X}$ is invertible, the parameter $\hat{\beta}$ is

$$\hat{\beta} = \left(\hat{X}^T\hat{X}\right)^{-1}\hat{X}^T y.$$

The estimate of β is only well-defined if $(\hat{X}^T\hat{X})^{-1}$ exits.

2.4 Ridge regression

If $\hat{X}^T\hat{X}$ is not invertible, the parameters β cannot be found [2]. This is very likely, when the matrix $\hat{X}$ is high-dimensional. To solve this problem, a diagonal component can be added to the matrix to make it invertible. Therefore, $\hat{X}^T\hat{X}$

is changed to $\hat{X}^T\hat{X} + \lambda\hat{I}$, where $\hat{I}$ is the identity matrix. The objective function becomes

$$\left|y - \hat{X}\hat{\beta}\right|^2 + \lambda\left|\hat{\beta}\right|^2$$

The estimate of $\hat{\beta}$ becomes

$$\hat{\beta} = \left(\hat{X}^T\hat{X} + \lambda I\right)^{-1}\hat{X}^T y$$

2.5 Lasso regression

When there are many features in models, we only want to keep the most important features. Therefore, we conduct a process called regularization. Lasso regression is a common modeling technique to do regularization. Lasso regression performs L1 regularization, i.e. it adds a factor of sum of absolute value of coefficients in the optimization objective [3]. This is done by introducing a parameter α and the optimization function becomes

$$\left|y - \hat{X}\hat{\beta}\right|^2 + \alpha\left|\hat{\beta}\right|$$

When α is 0, Lasso regression is the same as OLS. When α is high, the most feature coefficients are close to zero.

2.6 Criteria

In total, there are three criteria used to evaluate the regression models. The criteria are Mean Squared error (MSE), R^2 score function and Variance. They are calculated by the following formulas. The smaller they are, the more accurate the models are.

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

$$R^2(\hat{y}, \tilde{\hat{y}}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2},$$

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

3 Implementations

The scripts are written in python and R. The calculation for each task is written in a stand alone python script and the figures are plotted by R. Table 1 gives an overview of scripts and their functions. Scripts can be downloaded from `https://github.com/hongliosloo/FYS-STK4155_P1` .

Table 1: Overview of scripts and their functions.

Name	Function
P1_OLS.ipynb	Perform the OLS regression based on generated data and do cross validation
P1_Ridge.ipynb	Find λ for the Ridge regression based on generated data and do cross validation
P1_Ridge-0.01.ipynb	Perform Ridge regression with λ is 0.01 based on generated data
P1_Lasso.ipynb	Find α for the Lasso regression based on generated data and do cross validation
P1_Lasso-10e-20.ipynb	Perform the Lasso regression with α is $10e^{20}$ based on generated data
real_dataOLS.ipynb	Perform the OLS regression based on SRTM DTM
real_dataRidge.ipynb	Perform the Ridge regression based on SRTM DTM
real_dataLasso.ipynb	Perform the Lasso regression based on SRTM DTM
plot_fig1_RidgeLambda.R	Plot Figure 1 and show the sensitivity of λ for the Ridge regression
plot_fig2_LassoAlpha.R	Plot Figure 2 and show the sensitivity of α for the Lasso regression
plot_fig3_MSE_R2.R	Plot Figure 3 and show performance of the OLS, Ridge and Lasso regression methods
plot_fig4_beta_stability.R	Plot Figure 4 and show stability of β
plot_fig5_realdata.R	Plot Figure 5 and show regression results based on SRTM DTM

4 Results and Analysis

4.1 Determine λ and α

To use the Ridge and Lasso regression methods, I have to determine λ and α first. This is done by manual calibration. Criteria Mean Squared error (MSE)

and R^2 score function. Results are shown in Figure 1 and Figure 2 respectively for the Ridge and Lasso regression methods.

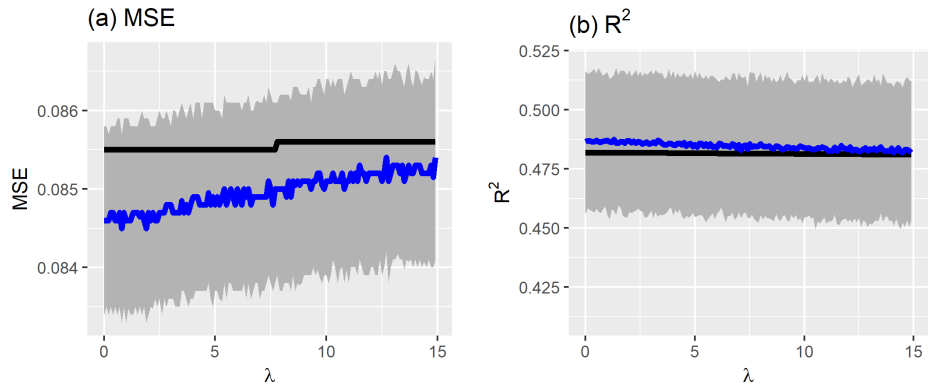


Figure 1: Changes of MSE and R^2 with λ. The blue line shows the median of cross validation results. The gray shade shows the 95% confidence of cross validation, in which 80% data is used to train model and 20% is used to test model for 1000 times.

As shown in Figure 1, the Ridge regression performance slightly decreases with λ. The decrease is not obvious for the whole dataset, but is noticeable for the cross validation. It is fair to say that model performance is not sensitive to λ in this case. When λ is 0, the Ridge regression produces the same results as the OLS regression. Therefore, a small λ value, 0.01 is used in future analysis.

As shown in Figure 2, the Lasso regression performance decreases with α. The decreasing trend is noticeable when α is larger than 10^{-8}. When α is larger than 10^{-8}, the decreasing trend is very small. Therefore, a small α value, 1e-20 is used in future analysis.

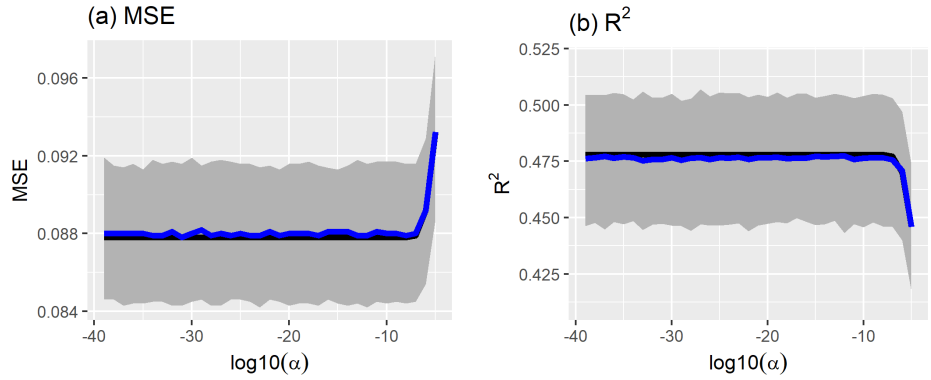


Figure 2: Changes of MSE and R^2 with α. The blue line shows the median of cross validation results. The gray shade shows the 95% confidence of cross validation, in which 80% data is used to train model and 20% is used to test model for 1000 times.

4.2 Model performance on generated data

As shown in Figure 3, the three regression methods show comparable performance. The R^2 for the whole dataset is 0.48, as shown by the black dots. The blue dots, the cross validation results distributed around. There are more blue dots close to the black dot and less blue dots far away from the black one. The Ridge regression is slightly better than other two other methods. However, there are also several blue dots which are far from the center of the dots cloud. It means that the Ridge regression gives better results with reduced predictability.

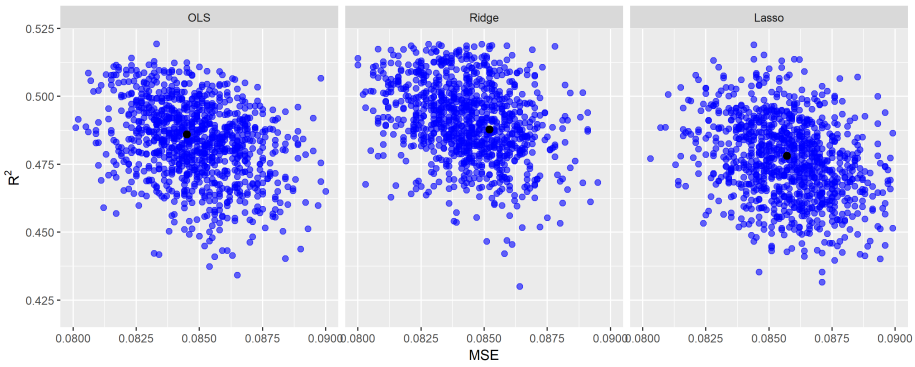


Figure 3: Performance of the three methods, shown by MSE and R^2. The blue points show results by cross validation and the black point shows results for the whole data.

4.3 Stability of β

The stability of β is also an important aspect to evaluate models. Figure 4 shows estimate of β and their 95% confidence intervals. Though all the methods are linear models, but the values of β are not comparable, especially by the Lasso method. Most of the β for the Lasso regression are close to zero, which is a sign of the parameter α is too large. As seen from Figure 3, the Ridge regression is the best in terms of MSE and R^2. This method is also the most unstable shown by the large error bars in Figure 4. It may cause problems when we apply this method.

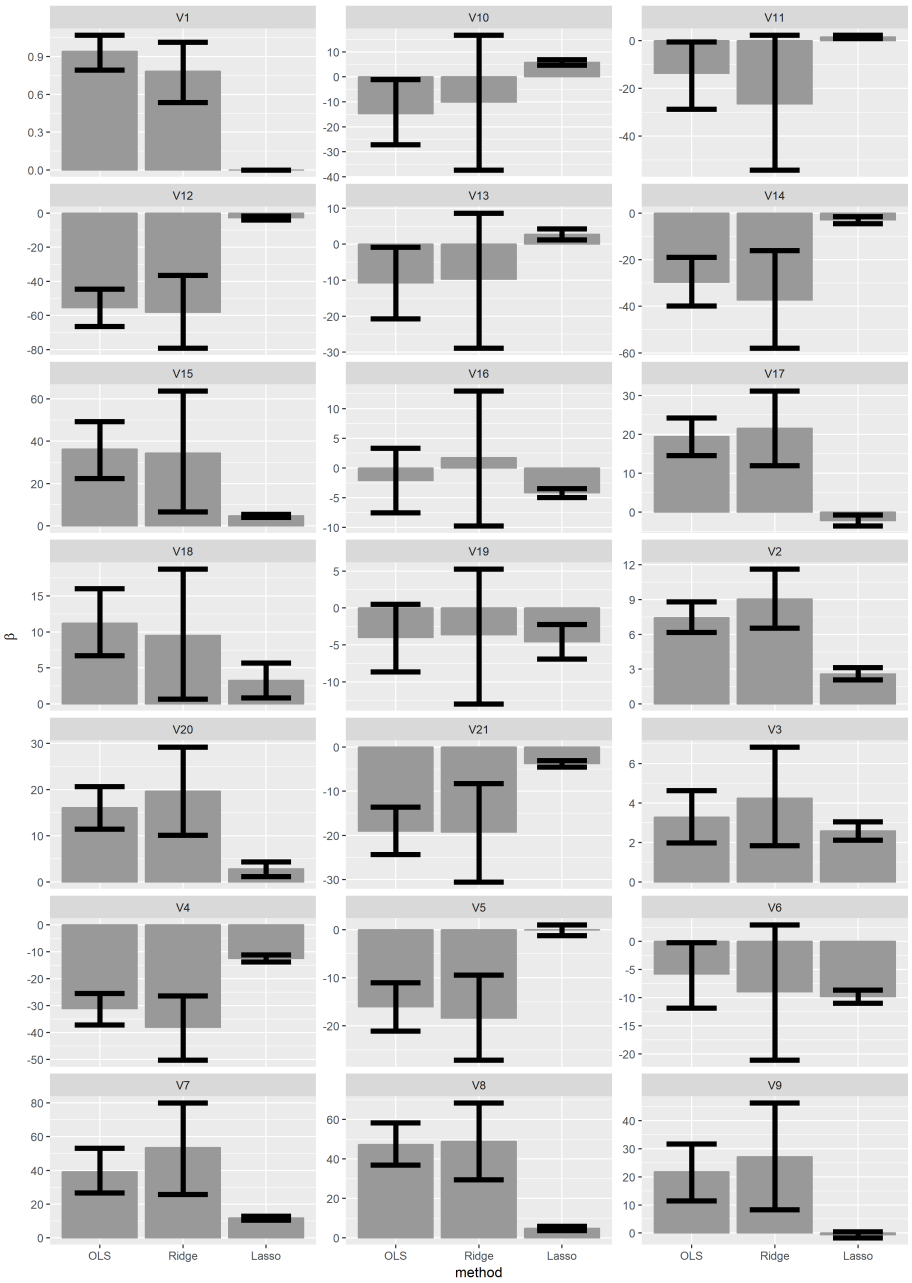


Figure 4: Estimate of β and their 95% confidence intervals.

4.4 Performance on STRM

The three regression methods are future used to interpolate SRTM DTM. The true and fitted heights for the four pitches are shown in Figure 5. As we can see, the three models are also almost the same. The results for the whole SRTM DTM are summarized in Table 2 and it shows three methods are almost the same.

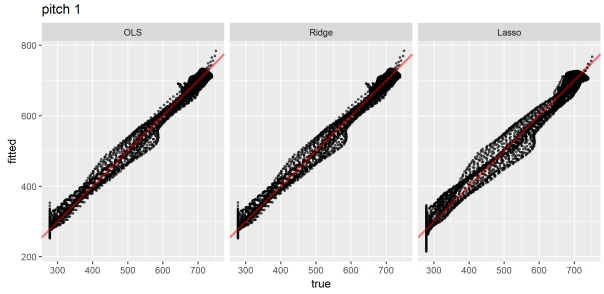
The regression methods work well on Pitch 1 and Pitch 3 and less on Pitch 2 and Pitch 4. To explore why the regression methods have different performance on the four pitches, I plot the true terrain surface in Figure 6. In Pitch 2 and 4, the terrain rises and falls. However, in Pitch 1 and Pitch 3, the terrain is not so complex. The complexity of terrain is the reason for the varied performance of the regression methods.

Table 2: Error of the regression method when interpolating the whole SRTM DTM.

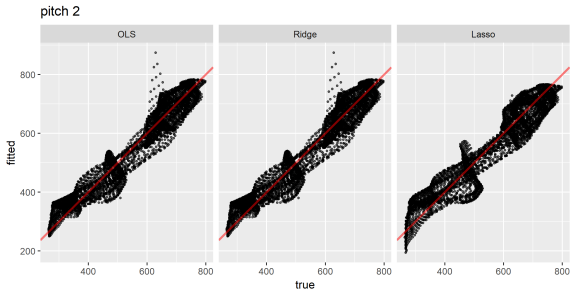
Method	MSE	R^2	Variance
OLS	42617	0.5373	49491
Ridge	42617	0.5373	49491
Lasso	42950	0.5337	48925

5 Conclusions

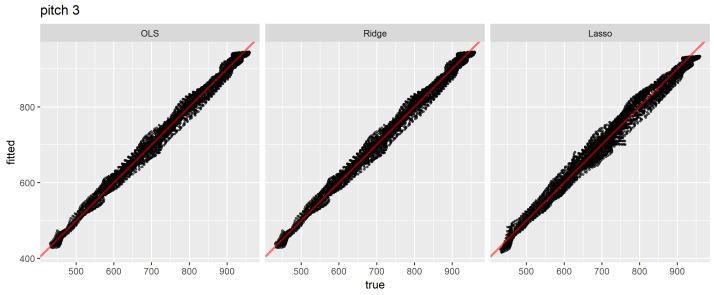
The main objective is to see details of three regression methods, i.e. the OLS regression, Ridge regression and Lasso regression. This report uses the regression methods to fit a weighted sum of four exponential components at the fifth orders of two independent variables. First, the regression methods are compared based a random generated data. The results show that the three regression methods are very similar. The Ridge regression method is slightly better in terms of MSE and R^2. However, this regression method is also worse in terms of param-



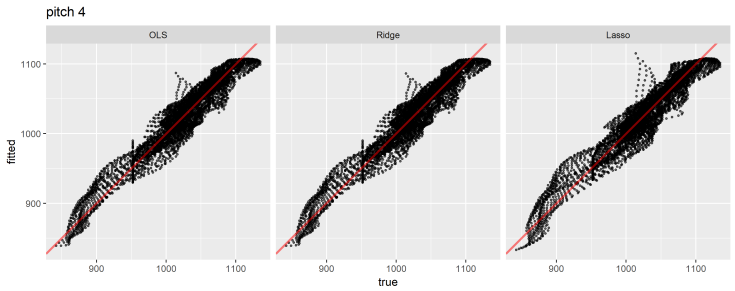
(a)



(b)

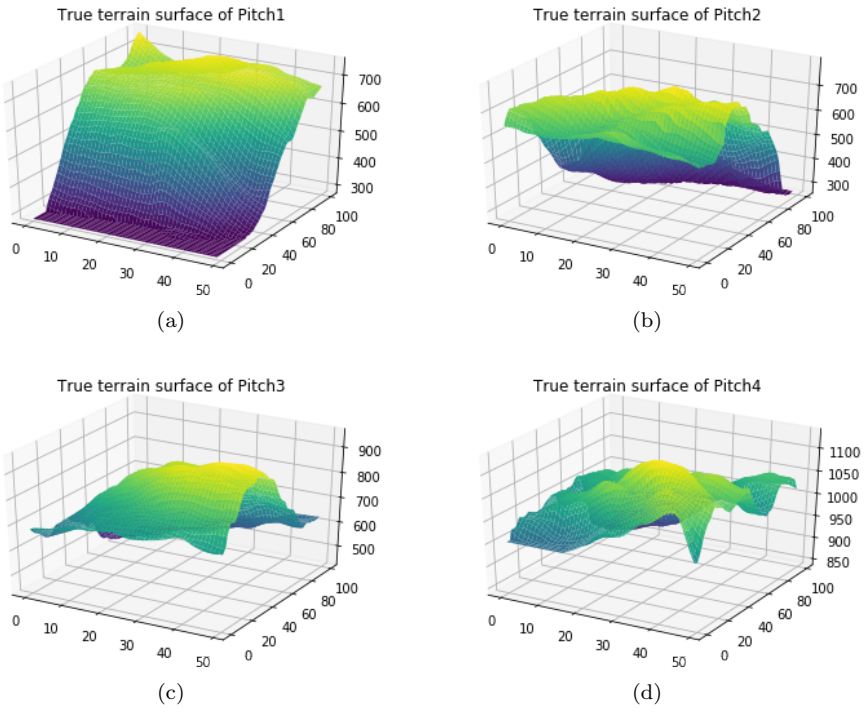


(c)



(d)

Figure 5: Scatter plots of true and fitted DTM for four pitches. The red line is $y = x$.



(a) (b) (c) (d)

Figure 6: True DTM of the four pitches.

eter's stability. Second, the three regression methods are applied to interpolate a digital terrain model. The results confirm that the three regression methods are very similar and model accuracy mainly depends on the terrain. When the terrain rises and falls, the model accuracy decreases.

6 Reference

References

[1] Jain Aarshay. A Complete Tutorial on Ridge and Lasso Regression in Python.

[2] Morten Hjorth_Jensen. Overview of course material: Data Analysis and Machine Learning.

[3] Ray Sunil. 7 Types of Regression Techniques you should know.