# Assignment (May 14)

*Saori Takaoka*

*18/05/20 15:24*

> Today, we discussed on three panels of beta coeddicients of three methods, stepwise, Lasso and Ridge. Describe the three panels. Also, write the formula for three methods and add description on them
>
> On May 21st, we will learn PCA. Check some brief focument on PCA method before the discussion on PCA in the class.

# Introduction

In machine learning, “good” models mean those with low bias and low variance. Bias represents how different the estimated values based on the model are from the real values. Variance stands for how stable the estimation is when the other conditions have changed.

However, it is difficult to obtain both the low bias and the low variance, represented by the essential concept of “bias-variance tradeoff”. The simple model accompanies low variance and high bias, in that the estimation does not change depending on the samples used for learning process, but the gap from the sample values is large. On the other hand, the complex model accompanies high variace and low bias, in that the sample choice greatly affect the model shape in spite of small gap. In short, bias and variance have inverse correlation.

The solution is to find as good a model as possible among several fitting models from simple to complex. In other words, you have to control the model complexity to achieve good estimation. A model in machine learning field corresponds to regression equation or regression line ($y = \alpha + \beta \times x$) in statistics, so you should determine $\beta$ coefficient through regression analysis. Here I will write the brief explanation of each methods; stepwise, Lasso and Ridge regression.

# 1.Stepwise regression

The stepwise method is to select the “strongest” variables with large $\beta$ first, then to append one by one, and finally to determine the appropriate set of $\beta$ coefficients considering the whole variables. Stepwise method is one of the simplest way of variable selection to perform multiple regression analysis. Stepwise regression requires 2 important step, one is to select and add the variables, and the other is to remove the selected ones. The cut-off value of each step is not specifically determined, but the overall goal is shown by;

$$\hat{\beta} = argmin_{\beta}(||y - \hat{y}||)$$

```
library(lars)
data(diabetes)
par(mfrow=c(1,2))
attach(diabetes)
```

```
## The following objects are masked from diabetes (pos = 3):
##
##     x, x2, y
```

```
## The following objects are masked from diabetes (pos = 4):
##
##     x, x2, y
```

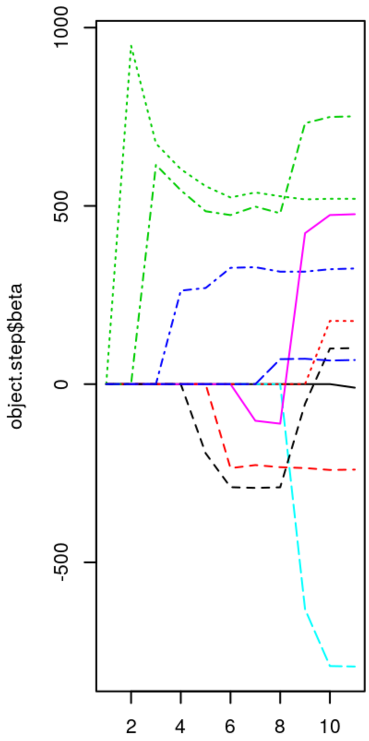
```
## The following objects are masked from diabetes (pos = 5):
##
##     x, x2, y
```

```
## The following objects are masked from diabetes (pos = 6):
##
##     x, x2, y
```

```
## The following objects are masked from diabetes (pos = 8):
##
##     x, x2, y
```

```
## The following objects are masked from diabetes (pos = 9):
##
##     x, x2, y
```

```
object.step <- lars(x,y,type="stepwise")
par(mfcol=c(1,3))
matplot(object.step$beta,type="l")
```




# 2. Lasso Regression

LASSO, Least absolute shrinkage and selection operator, is to detemine $\beta$ by;

$$\hat{\beta} = argmin_{\beta}(||y - X\beta||^2 + \lambda||\beta||)$$

$$||\beta|| = \sum_{i=1}^{d} \beta_i$$

Lasso regression is one of the extension of the least squred method to obtain $\beta$ with the minimum distance between the estimation and the samples. “Shrinkage” method, included in the name of Lasso, is one of the methods of multiple regression analysis. Also “selection” means that Lasso perform subset selection like stepwise method.

```
library(lars)
data(diabetes)
par(mfrow=c(1,2))
attach(diabetes)
```

```
## The following objects are masked from diabetes (pos = 3):
##
##     x, x2, y
```

```
## The following objects are masked from diabetes (pos = 4):
##
##     x, x2, y
```

```
## The following objects are masked from diabetes (pos = 5):
##
##     x, x2, y
```

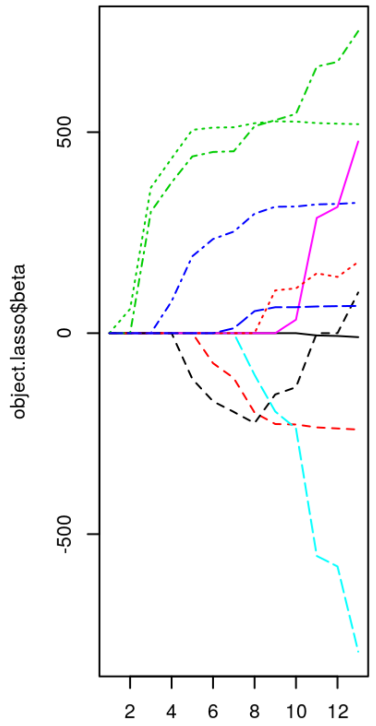
```
## The following objects are masked from diabetes (pos = 6):
##
##     x, x2, y
```

```
## The following objects are masked from diabetes (pos = 7):
##
##     x, x2, y
```

```
## The following objects are masked from diabetes (pos = 9):
##
##     x, x2, y
```

```
## The following objects are masked from diabetes (pos = 10):
##
##     x, x2, y
```

```
object.lasso <- lars(x,y,type="lasso")
par(mfcol=c(1,3))
matplot(object.lasso$beta,type="l")
```



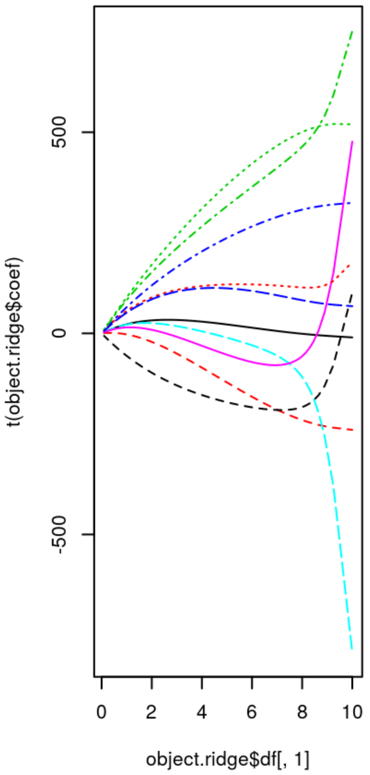

# 3. Ridge Regression

Ridge regression is to determine $\beta$ by;

$$\hat{\beta} = argmin_{\beta}(||y - X\beta||^2 + \lambda||\beta||^2)$$

Like Lasso regression, Ridge regression is based on the least squred method, which returns $\beta$ with the least distance between the estimation and the sample values. The difference from Lasso is just $\lambda||\beta||^2$, which means the “penalty” (model complexity) in L2 regularization. Bot Ridge regression and the least squred method is a kind of fitting, to find the most appropriate prediction of the given variables.

```
library(lars)
library(ridge)
lambda <- seq(from=100,to=0,length=10000)
object.ridge <- linearRidge(y ~ x,lambda=lambda)
par(mfcol=c(1,3))
matplot(object.ridge$df[,1],t(object.ridge$coef),type="l")
```




# Questions

from https://github.com/ryamada22/JD_lectures/blob/44ee7f393046e09d8815d72d6cdbbf4c4aa66a52/cells/Regularization_Shrinkage_Non-Full_Model.ipynb (https://github.com/ryamada22/JD_lectures/blob/44ee7f393046e09d8815d72d6cdbbf4c4aa66a52/cells/Regularization_Shrinkage_Non-Full_Model.ipynb)

> Three methods, stepwise, LASSO, and Ridge, are performed below.

I have done above.

> The last step of stepwise method returns the same result with LASSO and Ridge with their regularization parameter $/lambda = 0$. How do the plots show this phenomenon? This "goal" is "Full Model".

All the 3 methods have the same goal to perform the best fitting of the model to the samples. The $\lambda$ in Lasso and Ridge regression means "penulty" in other words the model complexity, which should be reduced during the process. Thus if the above 3 regression reach the ideal "full" model, all the $\beta$ coefficients must be the same.

> How do the plots show “subset selection” of stepwise and LASSO methods ? How do the plots show “no subset selection” of Ridge method?

When subset selection is performed, the graph is not smooth but shows discontinuous pattern. I guess softwares automatically connect the each points so that the results can be recognized easily. On the other hand, “no subset selection” shows the smooth and continuous graph pattern.