Advanced Data Analysis

Haoyang Chen | hc2812 | Assignment 4

1. Consider a multiple linear regression model

a). Investigate whether there is any multicollinearity:

There is multicollinearity. Although there does not exist a VIF larger than 10, the mean VIF is greater than 1 which indicates a serious multicollinearity

> vif(multiLinearModel)

age lwt race smoke ptl ht ui

1.125945 1.177116 1.224579 1.206096 1.124835 1.087378 1.087593

ftv

1.076820

> mean(vif(multiLinearModel))

[1] 1.138795

b). Run a ridge regression analysis and compare the results with (i):

The coefficients from ridge regression model are somewhat shrunken comparing to linear regression

Comparison:

|  |  |  |
| --- | --- | --- |
|  | Linear Regression | Ridge Regression |
| (Intercept) | 3129.46 | 3125.3122 |
| age | -0.2658 | -0.1828 |
| lwt | 3.4351 | 3.4173 |
| race | -188.4895 | -187.0416 |
| smoke | -358.4552 | -355.6267 |
| ptl | -51.1526 | -52.0323 |
| ht | -600.6465 | -596.6093 |
| ui | -511.2512 | -508.5071 |
| fty | -15.5358 | -15.2083 |

> summary(multiLinearModel)$coef

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3129.459388 344.242352 9.09086104 1.783264e-16

age -0.265810 9.594740 -0.02770372 9.779291e-01

lwt 3.435131 1.699899 2.02078565 4.478380e-02

race -188.489514 57.733892 -3.26479832 1.311221e-03

smoke -358.455188 107.517228 -3.33393256 1.039609e-03

ptl -51.152559 103.000275 -0.49662546 6.200592e-01

ht -600.646526 204.345418 -2.93936870 3.720106e-03

ui -511.251254 140.279187 -3.64452677 3.503426e-04

ftv -15.535798 46.935377 -0.33100402 7.410265e-01

> lm.ridge(bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv, data = birthwt, lambda = 1)

age lwt race smoke

3125.3122243 -0.1827917 3.4173081 -187.0415935 -355.6267387

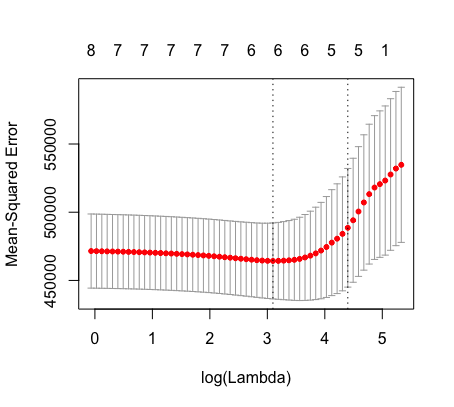
ptl ht ui ftv

-52.0323232 -596.6093106 -508.5071287 -15.2082711

1. Compare models selected using LASSO and a stepwise procedure

Lasso:

According to the result of cross validation, choose lambda:



The coefficients of lasso are:

> coef

9 x 1 sparse Matrix of class "dgCMatrix"

1

(Intercept) 3104.597034

age .

lwt 2.725628

race -157.669973

smoke -301.555511

ptl -29.212313

ht -479.296453

ui -462.170874

ftv .

Stepwise Procedure:

Coefficients:

(Intercept) lwt race smoke ht

3104.438 3.434 -187.849 -366.135 -595.820

ui

-523.419

Comparison:

|  |  |  |
| --- | --- | --- |
|  | Lasso | Stepwise |
| (Intercept) | 3104.5970 | 3104.438 |
| age | 0 | 0 |
| lwt | 2.7256 | 3.434 |
| race | -157.6700 | -187.849 |
| smoke | -301.5555 | -366.135 |
| ptl | -29.2123 | 0 |
| ht | -479.2965 | -595.820 |
| ui | -462.1709 | -523.419 |
| fty | 0 | 0 |

1. For the procedures listed in Table 1 next page, give appropriate ranks with respect to the listed attributes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | OLS | Ridge | Lasso | Elastic Net |
| Performance when p >> n | 3 | 2 | 1 | 1 |
| Performance under multicollinearity | 3 | 1 | 2 | 1 |
| Unbiased estimators | 1 | 3 | 3 | 3 |
| Model selection capability | 3 | 3 | 1 | 1 |
| Simplicity Computation, Inference, Interpretation | 1 | 2 | 3 | 3 |