

Practical session - Modèles de régression régularisée

Piseth KHENG, Borachhun YOU

03 October 2022

IV. Medical data

```
tab <- read.table("diabetes.txt", header=TRUE, sep="\t")
head(tab)
```

```
##   AGE SEX  BMI  BP  S1    S2 S3 S4    S5 S6  Y
## 1  59   2 32.1 101 157  93.2 38  4 4.8598 87 151
## 2  48   1 21.6  87 183 103.2 70  3 3.8918 69  75
## 3  72   2 30.5  93 156  93.6 41  4 4.6728 85 141
## 4  24   1 25.3  84 198 131.4 40  5 4.8903 89 206
## 5  50   1 23.0 101 192 125.4 52  4 4.2905 80 135
## 6  23   1 22.6  89 139  64.8 61  2 4.1897 68  97
```

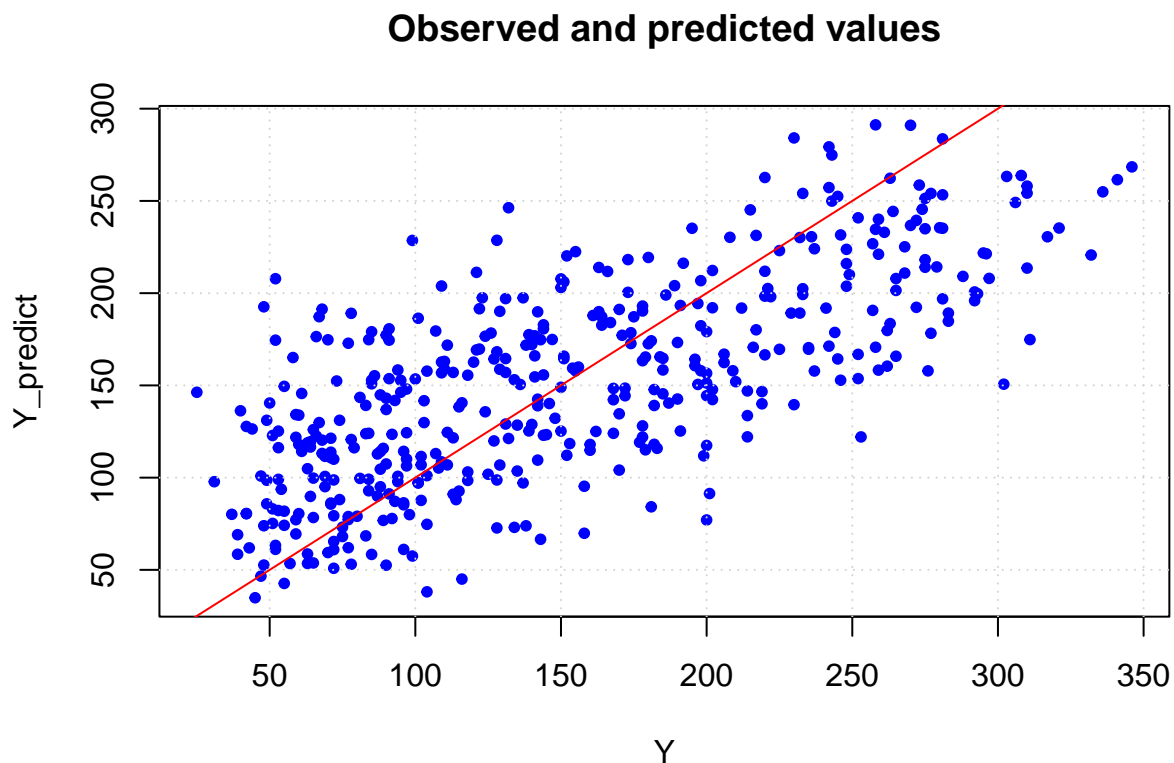
```
reg = lm(Y~., data=tab)
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ ., data = tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -155.827  -38.536   -0.228   37.806  151.353
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -334.56714    67.45462  -4.960 1.02e-06 ***
## AGE          -0.03636     0.21704  -0.168 0.867031
## SEX          -22.85965     5.83582  -3.917 0.000104 ***
## BMI           5.60296     0.71711   7.813 4.30e-14 ***
## BP            1.11681     0.22524   4.958 1.02e-06 ***
## S1           -1.09000     0.57333  -1.901 0.057948 .
## S2            0.74645     0.53083   1.406 0.160390
## S3            0.37200     0.78246   0.475 0.634723
## S4            6.53383     5.95864   1.097 0.273459
## S5           68.48312    15.66972   4.370 1.56e-05 ***
## S6            0.28012     0.27331   1.025 0.305990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 54.15 on 431 degrees of freedom
## Multiple R-squared:  0.5177, Adjusted R-squared:  0.5066
## F-statistic: 46.27 on 10 and 431 DF,  p-value: < 2.2e-16
```

```
Y <- tab$Y
Y_predict <- predict(reg, tab)

plot(Y, Y_predict, col="blue", pch=20, main="Observed and predicted values")
grid()
abline(a=0, b=1, col="red")
```



1. Model selection

A. Backward regression

```
regbackward = step(reg, direction="backward")

## Start:  AIC=3539.64
## Y ~ AGE + SEX + BMI + BP + S1 + S2 + S3 + S4 + S5 + S6
##
##           Df Sum of Sq    RSS   AIC
## - AGE      1      82 1264068 3537.7
```

```

## - S3      1      663 1264649 3537.9
## - S6      1     3080 1267066 3538.7
## - S4      1     3526 1267512 3538.9
## <none>                1263986 3539.6
## - S2      1     5799 1269785 3539.7
## - S1      1    10600 1274586 3541.3
## - SEX     1    44999 1308984 3553.1
## - S5      1    56016 1320001 3556.8
## - BP      1    72100 1336086 3562.2
## - BMI     1   179033 1443019 3596.2
##
## Step:  AIC=3537.67
## Y ~ SEX + BMI + BP + S1 + S2 + S3 + S4 + S5 + S6
##
##           Df Sum of Sq      RSS      AIC
## - S3      1         646 1264715 3535.9
## - S6      1        3001 1267069 3536.7
## - S4      1        3543 1267611 3536.9
## <none>                1264068 3537.7
## - S2      1        5751 1269820 3537.7
## - S1      1       10569 1274637 3539.4
## - SEX     1       45830 1309898 3551.4
## - S5      1       55964 1320032 3554.8
## - BP      1       73847 1337915 3560.8
## - BMI     1      179084 1443152 3594.2
##
## Step:  AIC=3535.9
## Y ~ SEX + BMI + BP + S1 + S2 + S4 + S5 + S6
##
##           Df Sum of Sq      RSS      AIC
## - S6      1        3093 1267808 3535.0
## - S4      1        3247 1267961 3535.0
## <none>                1264715 3535.9
## - S2      1        7505 1272219 3536.5
## - S1      1       26839 1291554 3543.2
## - SEX     1       46381 1311096 3549.8
## - BP      1       73533 1338248 3558.9
## - S5      1       97508 1362223 3566.7
## - BMI     1      178542 1443256 3592.3
##
## Step:  AIC=3534.98
## Y ~ SEX + BMI + BP + S1 + S2 + S4 + S5
##
##           Df Sum of Sq      RSS      AIC
## - S4      1        3686 1271494 3534.3
## <none>                1267808 3535.0
## - S2      1        7472 1275280 3535.6
## - S1      1       26378 1294186 3542.1
## - SEX     1       44684 1312492 3548.3
## - BP      1       82152 1349960 3560.7
## - S5      1      102520 1370328 3567.3
## - BMI     1      189976 1457784 3594.7
##
## Step:  AIC=3534.26

```

```
## Y ~ SEX + BMI + BP + S1 + S2 + S5
##
##           Df Sum of Sq      RSS      AIC
## <none>                1271494 3534.3
## - S2      1      39377 1310871 3545.7
## - SEX     1      41856 1313350 3546.6
## - S1      1      65236 1336730 3554.4
## - BP      1      79625 1351119 3559.1
## - BMI     1     190592 1462086 3594.0
## - S5      1     294092 1565586 3624.2
```

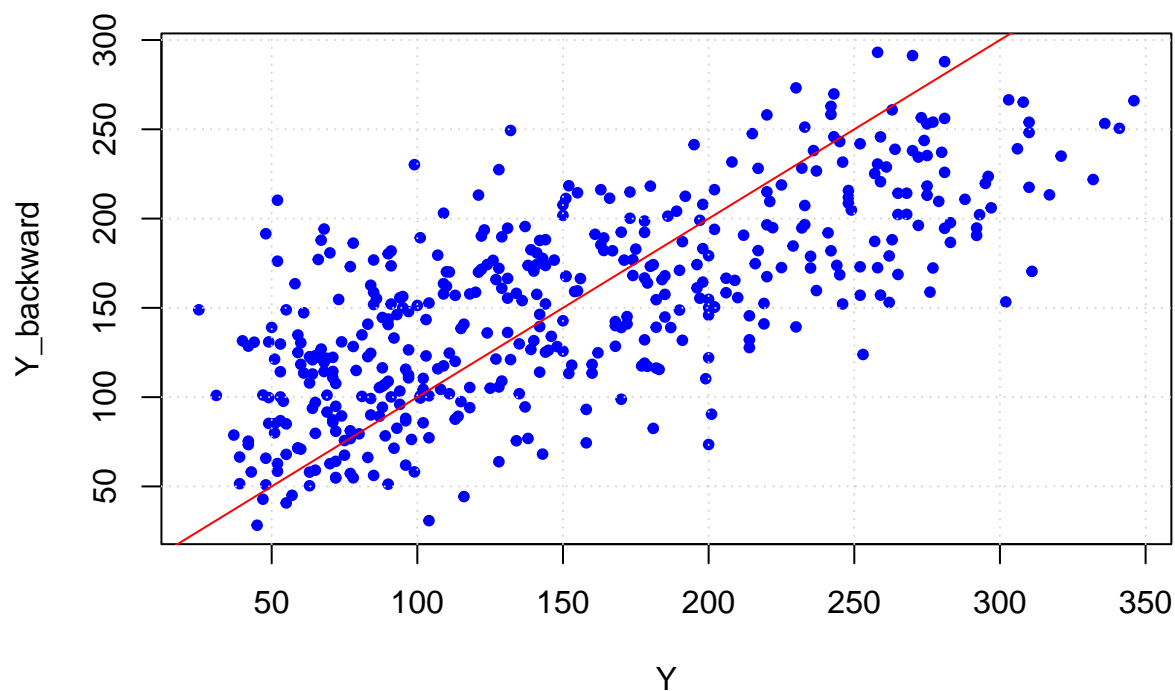
```
summary(regbackward)
```

```
##
## Call:
## lm(formula = Y ~ SEX + BMI + BP + S1 + S2 + S5, data = tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -158.275  -39.476   -2.065    37.219   148.690
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -313.7666    25.3848  -12.360  < 2e-16 ***
## SEX          -21.5910     5.7056   -3.784  0.000176 ***
## BMI           5.7111     0.7073    8.075  6.69e-15 ***
## BP            1.1266     0.2158    5.219  2.79e-07 ***
## S1           -1.0429     0.2208   -4.724  3.12e-06 ***
## S2            0.8433     0.2298    3.670  0.000272 ***
## S5           73.3065     7.3083   10.031  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.06 on 435 degrees of freedom
## Multiple R-squared:  0.5149, Adjusted R-squared:  0.5082
## F-statistic: 76.95 on 6 and 435 DF, p-value: < 2.2e-16
```

```
Y_backward <- predict(regbackward, tab)
```

```
plot(Y, Y_backward, col="blue", pch=20, main="Observed and predicted values of backward regression")
grid()
abline(a=0, b=1, col="red")
```

Observed and predicted values of backward regression



B. Forward regression

```
regforward = step(lm(Y~1, data=tab), list(upper=reg), direction="forward")
```

```
## Start:  AIC=3841.99
## Y ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + BMI      1   901427 1719582 3657.7
## + S5        1   839308 1781701 3673.4
## + BP        1   510851 2110158 3748.2
## + S4        1   485646 2135363 3753.4
## + S3        1   408507 2212502 3769.1
## + S6        1   383437 2237572 3774.1
## + S1        1   117824 2503186 3823.7
## + AGE       1    92527 2528482 3828.1
## + S2        1    79403 2541607 3830.4
## <none>                2621009 3842.0
## + SEX       1     4860 2616149 3843.2
##
## Step:  AIC=3657.7
## Y ~ BMI
##
##           Df Sum of Sq    RSS    AIC
```

```

## + S5      1      302888 1416694 3574.1
## + BP      1      136477 1583105 3623.1
## + S4      1      111511 1608071 3630.1
## + S3      1       97767 1621815 3633.8
## + S6      1       73738 1645844 3640.3
## + AGE     1       17087 1702495 3655.3
## + S1      1      12008 1707574 3656.6
## <none>                1719582 3657.7
## + S2      1       1228 1718354 3659.4
## + SEX     1        197 1719385 3659.6
##
## Step:  AIC=3574.06
## Y ~ BMI + S5
##
##           Df Sum of Sq      RSS      AIC
## + BP      1      53985 1362709 3558.9
## + S1      1      27624 1389070 3567.4
## + S3      1      26914 1389781 3567.6
## + S2      1       9256 1407438 3573.2
## + SEX     1       6881 1409813 3573.9
## + S6      1       6801 1409893 3573.9
## <none>                1416694 3574.1
## + S4      1       2376 1414318 3575.3
## + AGE     1        176 1416518 3576.0
##
## Step:  AIC=3558.88
## Y ~ BMI + S5 + BP
##
##           Df Sum of Sq      RSS      AIC
## + S1      1    31277.3 1331431 3550.6
## + S3      1    29921.2 1332787 3551.1
## + SEX     1    17532.1 1345177 3555.2
## + S2      1    10809.8 1351899 3557.4
## <none>                1362709 3558.9
## + S4      1     3218.7 1359490 3559.8
## + AGE     1     2106.4 1360602 3560.2
## + S6      1     1240.1 1361469 3560.5
##
## Step:  AIC=3550.62
## Y ~ BMI + S5 + BP + S1
##
##           Df Sum of Sq      RSS      AIC
## + SEX     1    20560.5 1310871 3545.7
## + S2      1    18080.9 1313350 3546.6
## + S4      1    15188.0 1316243 3547.6
## + S3      1    14360.4 1317071 3547.8
## <none>                1331431 3550.6
## + S6      1     2898.8 1328533 3551.7
## + AGE     1      472.0 1330959 3552.5
##
## Step:  AIC=3545.74
## Y ~ BMI + S5 + BP + S1 + SEX
##
##           Df Sum of Sq      RSS      AIC

```

```
## + S2      1      39377 1271494 3534.3
## + S4      1      35591 1275280 3535.6
## + S3      1      35001 1275870 3535.8
## <none>                1310871 3545.7
## + S6      1       5288 1305583 3546.0
## + AGE     1        49 1310822 3547.7
##
## Step:  AIC=3534.26
## Y ~ BMI + S5 + BP + S1 + SEX + S2
##
##           Df Sum of Sq      RSS      AIC
## <none>                1271494 3534.3
## + S4      1      3686.2 1267808 3535.0
## + S6      1      3532.6 1267961 3535.0
## + S3      1       394.8 1271099 3536.1
## + AGE     1        10.9 1271483 3536.3
```

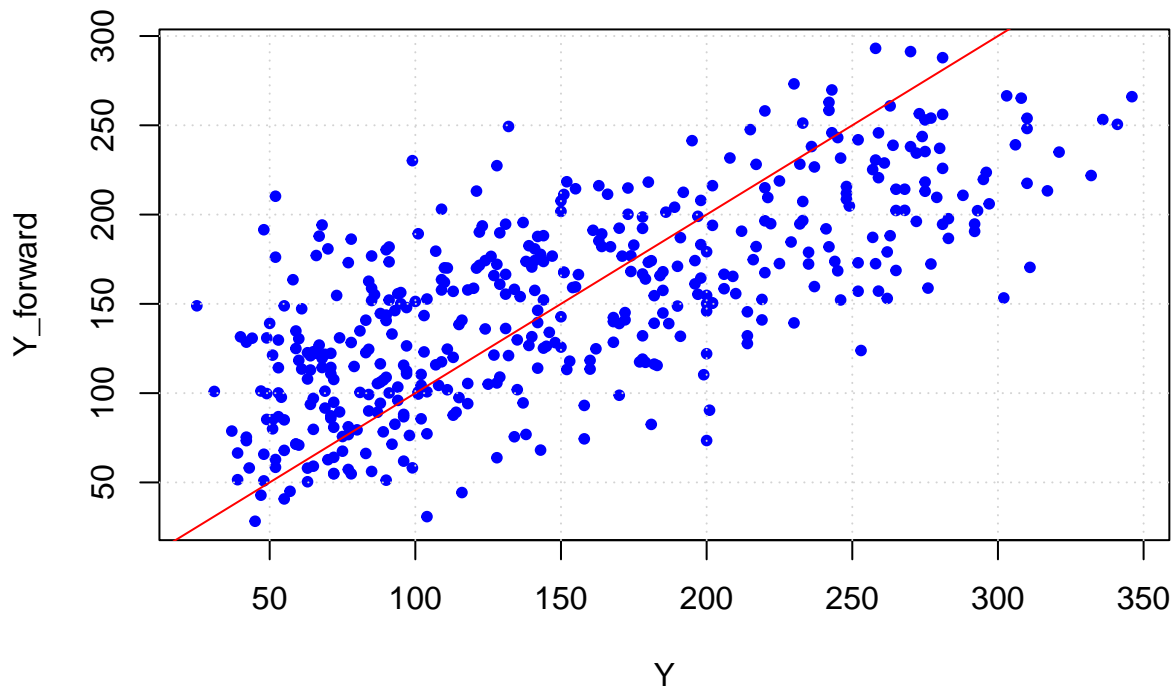
```
summary(regforward)
```

```
##
## Call:
## lm(formula = Y ~ BMI + S5 + BP + S1 + SEX + S2, data = tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -158.275  -39.476   -2.065    37.219   148.690
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -313.7666    25.3848  -12.360 < 2e-16 ***
## BMI           5.7111     0.7073   8.075 6.69e-15 ***
## S5           73.3065     7.3083  10.031 < 2e-16 ***
## BP            1.1266     0.2158   5.219 2.79e-07 ***
## S1           -1.0429     0.2208  -4.724 3.12e-06 ***
## SEX          -21.5910     5.7056  -3.784 0.000176 ***
## S2            0.8433     0.2298   3.670 0.000272 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.06 on 435 degrees of freedom
## Multiple R-squared:  0.5149, Adjusted R-squared:  0.5082
## F-statistic: 76.95 on 6 and 435 DF, p-value: < 2.2e-16
```

```
Y_forward <- predict(regforward, tab)
```

```
plot(Y, Y_forward, col="blue", pch=20, main="Observed and predicted values of forward regression")
grid()
abline(a=0, b=1, col="red")
```

Observed and predicted values of forward regression



C. Stepwise regression

```
regboth = step(reg, direction="both")
```

```
## Start: AIC=3539.64
## Y ~ AGE + SEX + BMI + BP + S1 + S2 + S3 + S4 + S5 + S6
##
##      Df Sum of Sq    RSS    AIC
## - AGE   1      82 1264068 3537.7
## - S3     1     663 1264649 3537.9
## - S6     1    3080 1267066 3538.7
## - S4     1    3526 1267512 3538.9
## <none>          1263986 3539.6
## - S2     1    5799 1269785 3539.7
## - S1     1   10600 1274586 3541.3
## - SEX    1   44999 1308984 3553.1
## - S5     1   56016 1320001 3556.8
## - BP     1    72100 1336086 3562.2
## - BMI    1   179033 1443019 3596.2
##
## Step: AIC=3537.67
## Y ~ SEX + BMI + BP + S1 + S2 + S3 + S4 + S5 + S6
##
##      Df Sum of Sq    RSS    AIC
```



```

## - S3      1      646 1264715 3535.9
## - S6      1     3001 1267069 3536.7
## - S4      1     3543 1267611 3536.9
## <none>           1264068 3537.7
## - S2      1     5751 1269820 3537.7
## - S1      1    10569 1274637 3539.4
## + AGE     1       82 1263986 3539.6
## - SEX     1    45830 1309898 3551.4
## - S5      1    55964 1320032 3554.8
## - BP      1    73847 1337915 3560.8
## - BMI     1   179084 1443152 3594.2
##
## Step:  AIC=3535.9
## Y ~ SEX + BMI + BP + S1 + S2 + S4 + S5 + S6
##
##           Df Sum of Sq      RSS      AIC
## - S6      1      3093 1267808 3535.0
## - S4      1      3247 1267961 3535.0
## <none>           1264715 3535.9
## - S2      1      7505 1272219 3536.5
## + S3      1       646 1264068 3537.7
## + AGE     1        66 1264649 3537.9
## - S1      1     26839 1291554 3543.2
## - SEX     1     46381 1311096 3549.8
## - BP      1     73533 1338248 3558.9
## - S5      1     97508 1362223 3566.7
## - BMI     1    178542 1443256 3592.3
##
## Step:  AIC=3534.98
## Y ~ SEX + BMI + BP + S1 + S2 + S4 + S5
##
##           Df Sum of Sq      RSS      AIC
## - S4      1      3686 1271494 3534.3
## <none>           1267808 3535.0
## - S2      1      7472 1275280 3535.6
## + S6      1      3093 1264715 3535.9
## + S3      1       738 1267069 3536.7
## + AGE     1         0 1267807 3537.0
## - S1      1     26378 1294186 3542.1
## - SEX     1     44684 1312492 3548.3
## - BP      1     82152 1349960 3560.7
## - S5      1    102520 1370328 3567.3
## - BMI     1    189976 1457784 3594.7
##
## Step:  AIC=3534.26
## Y ~ SEX + BMI + BP + S1 + S2 + S5
##
##           Df Sum of Sq      RSS      AIC
## <none>           1271494 3534.3
## + S4      1      3686 1267808 3535.0
## + S6      1      3533 1267961 3535.0
## + S3      1       395 1271099 3536.1
## + AGE     1        11 1271483 3536.3
## - S2      1     39377 1310871 3545.7

```

```
## - SEX    1      41856 1313350 3546.6
## - S1     1      65236 1336730 3554.4
## - BP     1      79625 1351119 3559.1
## - BMI    1     190592 1462086 3594.0
## - S5     1     294092 1565586 3624.2
```

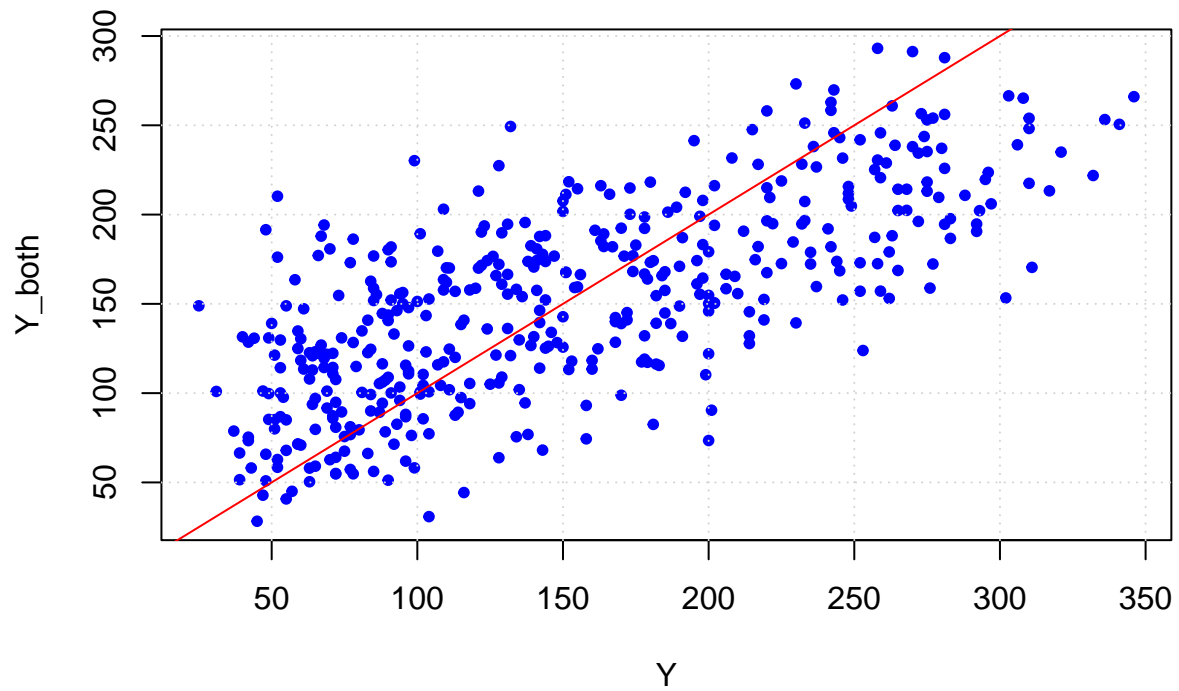
```
summary(regboth)
```

```
##
## Call:
## lm(formula = Y ~ SEX + BMI + BP + S1 + S2 + S5, data = tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -158.275  -39.476   -2.065    37.219   148.690
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -313.7666    25.3848  -12.360  < 2e-16 ***
## SEX          -21.5910     5.7056   -3.784  0.000176 ***
## BMI           5.7111     0.7073    8.075  6.69e-15 ***
## BP            1.1266     0.2158    5.219  2.79e-07 ***
## S1           -1.0429     0.2208   -4.724  3.12e-06 ***
## S2            0.8433     0.2298    3.670  0.000272 ***
## S5           73.3065     7.3083   10.031  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.06 on 435 degrees of freedom
## Multiple R-squared:  0.5149, Adjusted R-squared:  0.5082
## F-statistic: 76.95 on 6 and 435 DF, p-value: < 2.2e-16
```

```
Y_both <- predict(regboth, tab)
```

```
plot(Y, Y_both, col="blue", pch=20, main="Observed and predicted values of stepwise regression")
grid()
abline(a=0, b=1, col="red")
```

Observed and predicted values of stepwise regression



2. Ridge

```
scaled_tab <- scale(tab)
```

3. Lasso