REgression Risque

Muyao GUO

2025-10-10

```
library(MASS)
data(birthwt)
summary(birthwt)
```

```
##       low              age             lwt             race
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   1st Qu.:1.000
##  Median :0.0000   Median :23.00   Median :121.0   Median :1.000
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8   Mean   :1.847
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0   Max.   :3.000
##      smoke             ptl               ht                ui
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.3915   Mean   :0.1958   Mean   :0.06349   Mean   :0.1481
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :3.0000   Max.   :1.00000   Max.   :1.0000
##       ftv              bwt
##  Min.   :0.0000   Min.   : 709
##  1st Qu.:0.0000   1st Qu.:2414
##  Median :0.0000   Median :2977
##  Mean   :0.7937   Mean   :2945
##  3rd Qu.:1.0000   3rd Qu.:3487
##  Max.   :6.0000   Max.   :4990
```

```
birthwt = within(birthwt,{
  race = factor(race, labels=c("white", "black", "other"))
  smoke = factor(smoke,labels=c("No", "Yes"))
  ptl = factor(ptl > 0)
  ht= factor(ht>0)
  ui = factor(ui,labels=c("No", "Yes"))
  ftv = factor(ftv)
  levels(ftv)[-(1:2)] = "2+"
})

birthwt=birthwt[,-1]
summary(birthwt)
```

```
##       age              lwt           race    smoke        ptl          ht
##  Min.   :14.00   Min.   : 80.0   white:96   No :115   FALSE:159   FALSE:177
```

```
##  1st Qu.:19.00   1st Qu.:110.0   black:26   Yes: 74   TRUE : 30   TRUE : 12
##  Median :23.00   Median :121.0   other:67
##  Mean   :23.24   Mean   :129.8
##  3rd Qu.:26.00   3rd Qu.:140.0
##  Max.   :45.00   Max.   :250.0
##    ui       ftv          bwt
##  No :161   0 :100   Min.   : 709
##  Yes: 28   1 : 47   1st Qu.:2414
##            2+: 42   Median :2977
##                     Mean   :2945
##                     3rd Qu.:3487
##                     Max.   :4990
```

```
attach(birthwt)

# régression linéaire avec toutes les covariables
reg =lm(bwt~.,data=birthwt)
summary(reg)
```

```
##
## Call:
## lm(formula = bwt ~ ., data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1794.68  -444.72    46.29   495.95  1600.49
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2867.253    313.560   9.144  < 2e-16 ***
## age           -2.794      9.661  -0.289 0.772743
## lwt            4.363      1.724   2.531 0.012252 *
## raceblack   -465.821    149.542  -3.115 0.002145 **
## raceother   -316.223    117.416  -2.693 0.007754 **
## smokeYes    -300.541    109.628  -2.741 0.006741 **
## ptlTRUE     -230.904    137.983  -1.673 0.096001 .
## htTRUE      -591.228    201.175  -2.939 0.003731 **
## uiYes       -481.968    137.265  -3.511 0.000565 ***
## ftv1         111.302    122.972   0.905 0.366639
## ftv2+        -55.816    123.174  -0.453 0.650996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 646.1 on 178 degrees of freedom
## Multiple R-squared:  0.2568, Adjusted R-squared:  0.2151
## F-statistic: 6.151 on 10 and 178 DF,  p-value: 4.883e-08
```

```
# Prédiction
#    age lwt  race smoke   ptl    ht ui ftv
#     23  94  other  Yes FALSE FALSE No   0

Y0=data.frame(age=23, lwt=94, race="other", smoke="Yes",ptl="FALSE", ht="FALSE",ui="No",ftv="0")
predict(reg,newdata=Y0)
```

```
##        1
## 2596.384
```

```
# 2596.384

RSS=sum((bwt-predict(reg))^2)
RSE=sqrt(RSS/(189-11))  # 646.1 idem sortie R
TSS=sum((bwt-mean(bwt))^2)
R2=1-RSS/TSS # 0.2568224 idem sortie R

# erreur d'apprentissage
MSE_train = mean((bwt-predict(reg))^2)
RMSE_train=sqrt(MSE_train)
RMSE_train
```

```
## [1] 626.9739
```

```
# [1] 626.9739

## fonction de risque MSE
mse=function(y,ypred) round((mean((y-ypred)^2)),digits=2)
# plus interprétable, dans l'échalle de la variable à prédire:
rmse=function(y,ypred) sqrt(mse(y,ypred))

### Apprentissage / Test

set.seed(12345)  # calcul reproductible

test = sample(1:length(birthwt$bwt),60)
train = -test
train = birthwt[train, ] # 129 observations
test = birthwt[test,]

reg.train=lm(bwt~.,data=train)
y.pred=predict(reg.train, newdata=test)
rmse(test$bwt,y.pred)
```

```
## [1] 639.0716
```

```
# [1] 639.07

### Validation croisée

K=5
set.seed(123)
ind_fold=sample(1:K,nrow(birthwt),replace=TRUE)
table(ind_fold)
```

```
## ind_fold
##  1  2  3  4  5
## 42 38 36 32 41
```

```r
# 1  2  3  4  5
# 42 38 36 32 41

error=numeric()
for (j in 1:K)
{
  fit.lm=lm(bwt~.,data=birthwt[ind_fold!=j,])
  y.lm=predict(fit.lm,newdata=birthwt[ind_fold==j,])
  y.test=birthwt[ind_fold==j,"bwt"]
  error[j]=rmse(y.test,y.lm)
}

cv.error=mean(error)
cv.error
```

```
## [1] 672.0053
```

```r
# [1] 672.0053

sd(error)
```

```
## [1] 12.08386
```

```r
# [1] 12.08386
```