Cross-validation

Issues

Babies weight dataset consists of information about 1236 mothers in which each row represents information about a single mother – Gestion, Age, Height, Weight, Smoke, Birthweight. By using multivariate linear regression to model the outcome variable Birthweight on the basis of the variables Gestation, Age, Height, Weight, Smok**e** and validated the model by using cross-validation methods.

- The validation set method.
- Leave-one-outcross-validation (LOOCV).
- K-fold cross-validation, with k = 10.

Findings

By developing the model by considering birthweight as dependent variable on the other variables, the finding indicates that the model has a moderate level of accuracy in predicting.

- The R-squared value for validation set method is 0.02968.
- The R-squared value for leave-one-outcross-validation (LOOCV) is 0.03056.
- The R-squared value for k-fold cross-validation, with k = 10 is 0.03056.

Discussions

The results show that the model has a moderate level of accuracy in predicting birthweight utilizing the predictor variables in this multivariate linear regression model. The R-squared values derived from the predictions are 0.02968, 0.03056, and 0.03056 for the validation set technique, LOOCV, and K-fold cross-validation.

Low R-squared values indicate that the model can only adequately explain a tiny portion of the variation in birthweight, despite the model's moderate ability to predict birthweight using the predictor variables. This suggests that factors other than the predictor variables are probably involved in determining birthweight as well. Therefore, although the model may be useful in predicting birthweight to some extent.

Appendix A: Method

Babies weight data was uploaded into the R studio in which it contains the 1236 rows and 6 columns and installed the packages readxl, caret to perform the cross-validation methods.

Firstly, developed the linear model using the given data and analyzed the summary of model whether it useful or not. Then split the data into two parts with 80 percent data in the training set and 20 percent data in the testing set. By using the training dataset developed the linear model and obtained the summary and by using the model predicted the testing data.

Performed leave-one-outcross-validation (LOOCV) and k-fold cross-validation, with k = 10 and obtained results.

Appendix B: Results

```
> data <- read_excel(file, sheet = 1)
> mod<- lm(Birthweight~., data=data)
> summary(mod)

Call:
lm(formula = Birthweight ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-65.231 -11.317   0.325  11.284  55.745

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.810363   7.947180  10.294  < 2e-16 ***
Gestation    0.012800   0.006830   1.874 0.061131 .
Age          0.070370   0.079456   0.886 0.375981
Height       0.525584   0.121922   4.311 1.76e-05 ***
Weight      -0.005831   0.004336  -1.345 0.178946
Smoke       -1.989031   0.561626  -3.542 0.000413 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.99 on 1230 degrees of freedom
Multiple R-squared:  0.03056,  Adjusted R-squared:  0.02661
F-statistic: 7.754 on 5 and 1230 DF,  p-value: 3.415e-07
```

1. For Multivariate regression model

```
> set.seed(222)
> spilting<-sample(2,nrow(data),replace=T,prob=c(0.8,0.2))
> training<-data[spilting==1,]
> testing<-data[spilting==2,]
> mod1 <- lm(Birthweight~., data=training)
> summary(mod1)

Call:
lm(formula = Birthweight ~ ., data = training)

Residuals:
    Min      1Q  Median      3Q     Max
-64.652 -10.818   0.531  10.919  56.777

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 87.422077   8.812284   9.920  < 2e-16 ***
Gestation    0.011944   0.007464   1.600 0.109862
Age         -0.017971   0.092611  -0.194 0.846176
Height       0.476701   0.135452   3.519 0.000452 ***
Weight      -0.004025   0.004773  -0.843 0.399337
Smoke       -2.236639   0.628656  -3.558 0.000391 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.06 on 1002 degrees of freedom
Multiple R-squared:  0.02968,  Adjusted R-squared:  0.02484
F-statistic: 6.129 on 5 and 1002 DF,  p-value: 1.328e-05
```

2. For leave-one-outcross-validation (LOOCV) method

```
> #LOOCV
> loocv_model <- trainControl(method="LOOCV")
> mod2 <- train(Birthweight ~ ., data = data, method = "lm", trControl = l
oocv_model)
> summary(mod2)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-65.231 -11.317   0.325  11.284  55.745

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.810363   7.947180  10.294  < 2e-16 ***
Gestation    0.012800   0.006830   1.874 0.061131 .
Age          0.070370   0.079456   0.886 0.375981
Height       0.525584   0.121922   4.311 1.76e-05 ***
Weight      -0.005831   0.004336  -1.345 0.178946
Smoke       -1.989031   0.561626  -3.542 0.000413 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.99 on 1230 degrees of freedom
Multiple R-squared:  0.03056,  Adjusted R-squared:  0.02661
F-statistic: 7.754 on 5 and 1230 DF,  p-value: 3.415e-07
```

3. For k-fold cross-validation, with k = 10 method

```
> k_fold <- trainControl(method = "cv", number = 10,summaryFunction = defa
ultSummary)
> mod3 <- train(Birthweight ~ ., data = data, method = "lm", trControl = k
_fold)
> summary(mod3)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-65.231 -11.317   0.325  11.284  55.745

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.810363   7.947180  10.294  < 2e-16 ***
Gestation    0.012800   0.006830   1.874 0.061131 .
Age          0.070370   0.079456   0.886 0.375981
Height       0.525584   0.121922   4.311 1.76e-05 ***
Weight      -0.005831   0.004336  -1.345 0.178946
Smoke       -1.989031   0.561626  -3.542 0.000413 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.99 on 1230 degrees of freedom
Multiple R-squared:  0.03056,  Adjusted R-squared:  0.02661
F-statistic: 7.754 on 5 and 1230 DF,  p-value: 3.415e-07
```

Appendix C: Code

```
install.packages('readxl')
library(readxl)
install.packages('pROC')
library(pROC)
install.packages("caret")
library(caret)
file<-"C:\\Users\\sasik\\OneDrive\\Desktop\\babies_weight.xls"
data <- read_excel(file, sheet = 1)
str(data)
mod<- lm(Birthweight~., data=data)
summary(mod)

set.seed(222)
spilting<-sample(2,nrow(data),replace=T,prob=c(0.8,0.2))
training<-data[spilting==1,]
testing<-data[spilting==2,]

mod1 <- lm(Birthweight~., data=training)
summary(mod1)

pred <- predict(mod1, testing)
pred
```

```
#LOOCV

loocv_model <- trainControl(method="LOOCV")

mod2 <- train(Birthweight ~ ., data = data, method = "lm", trControl = loocv_model)

summary(mod2)


# k-fold cross-validation

k_fold <- trainControl(method = "cv", number = 10,summaryFunction = defaultSummary)

mod3 <- train(Birthweight ~ ., data = data, method = "lm", trControl = k_fold)

summary(mod3)
```