Cross-validation

Issues

In this analysis, we used a dataset containing information on 1236 mothers, where each row represents a single mother. The dataset includes variables such as Gesta on, Age, Height, Weight, Smoke, and Birthweight. We applied multivariate linear regression to model the outcome variable, Birthweight, using the forementioned predictor variables. To validate the model's performance, we employed cross-validation techniques.

- 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 on (LOOCV) is 0.03056.
- The R-squared value for k-fold cross-validation, with k = 10 is 0.03056.

Discussions:

According to the analysis, the multivariate linear regression model using predictor variables has a moderate level of accuracy in predicting birthweight. However, the R-squared values obtained from the validation set technique, LOOCV, and K-fold cross-validation were low, indicating that the model can only explain a small portion of the variation in birthweight. This suggests that there may be other factors besides the predictor variables that play a role in determining birthweight. Despite this, the model can still be useful in predicting birthweight to some extent.

Appendix A: Method

The Babies weight data, which has 1236 rows and 6 columns, was uploaded into R Studio. The read xl and caret packages were installed to perform cross-validation methods. Firstly, a linear model was developed using the provided data, and the summary of the model was analyzed to determine its usefulness. Next, the data was split into two parts with 80% of the data in the training set and 20% of the data in the testing set. The linear model was developed using the training dataset, and its summary was obtained. The model was then used to predict the testing data. To further evaluate the model's performance, leave-one-out cross-validation (LOOCV) and k-fold cross-validation with k = 10 were performed, and the results were obtained.

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)

data <- read_excel(file, sheet = 1)

mod <- lm(Birthweight ~ ., data = data)

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

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

pred <- predict(mod1, testing)

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

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