Cross-validation Methods

Issues

Data set consists of numerical characteristics of several mothers of gestion, age, height, weight, smoke birthweight in which row represent to particular mother.

Here we developed a multivariate linear regression model and applied different cross validation methods to find the test errors.

i. Used the validation set method to split the data into two random halves, using one half as the training set and the remaining half as the test set.
ii. Used leave-one-outcross-validation (LOOCV), to test the linear model.
iii. Used k-fold cross-validation, with k = 10, to test the linear model.

Findings

In this Multivariate linear regression model, the findings indicate that the model has a moderate level of accuracy in predicting birthweight using the five predictor variables.

The R-squared values obtained from the predictions are 0.03423, 0.03056, 0.03056 for validation set method, LOOCV, K-fold cross-validation methods respectively.

Despite the moderate ability of the model to predict birthweight using the five predictor variables, the low R-squared values suggest that the model can only account for a small portion of the variation in birthweight. This

implies that there are likely other factors beyond the five predictor variables that also play a role in determining birthweight. As a result, while the model may be helpful in predicting birthweight to a certain extent, it cannot be fully relied upon to provide accurate predictions.

Discussions

The R-squared values obtained by the model are notably inadequate, indicating that the model may not be appropriate for the given data and may have overlooked important factors. Due to the model's utilization of only five predictor variables, it may not fully capture the complex associations between the predictors and the outcome variable. It is conceivable that there exist other crucial predictors that were not accounted for in the model.

Appendix A: Method

This code reads data from an Excel file and separates predictor variables from the outcome variable. It uses a multivariate linear regression model with all five predictors to predict the outcome.

i. In validation set method, we split the data into two parts (70% data for training and 30% data for testing). And we fitted model to training data set and calculated the R-squared value-0.03423.
ii. In Leave-one-out cross-validation (LOOCV) to test linear regression model R-squared value -0.03056.
iii. The code ultimately utilizes 10-fold cross-validation to assess the predictive capability of the model. It establishes a KFold object and calculates the R-squared value, which is 0.03056.

Appendix B: Results

i. For the multivariate regression model, R squared value is 0.03423

```
> summary(model2)

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

Residuals:
    Min      1Q  Median      3Q     Max
-65.171 -10.909   0.661  11.368  57.138

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 76.987319  10.212881   7.538 1.20e-13 ***
Gestation    0.010469   0.008287   1.263  0.20680
Age          0.020699   0.098568   0.210  0.83372
Height       0.645193   0.158343   4.075 5.03e-05 ***
Weight      -0.011526   0.006033  -1.910  0.05642 .
Smoke       -2.308129   0.680060  -3.394  0.00072 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.25 on 866 degrees of freedom
Multiple R-squared:  0.03423,  Adjusted R-squared:  0.02866
F-statistic: 6.139 on 5 and 866 DF,  p-value: 1.337e-05
```

ii. For leave-one-outcross-validation (LOOCV), R squared value is 0.03056.

```
> summary(model3)

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
```

iii. For k-fold cross-validation, R squared value is 0.03056.

```
> summary(model4)

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 <-"E:\\Assignments\\MTH 522\\Project 3\\babies_weight.xls"

data <- read_excel(file, sheet = 1)

data
```

```
str(data)
summary(data)
colnames(data)
ncol(data)
nrow(data)

#linear model
model1 <- lm(Birthweight~., data=data)
summary(model1)

#spilting the data
set.seed(222)
div<-sample(2,nrow(data),replace=T,prob=c(0.7,0.3))
training<-data[div==1,]
testing<-data[div==2,]

#linear model for training data
model2 <- lm(Birthweight~., data=training)
summary(model2)

print(paste( "R-squared:", summary(model2)$r.squared))
```

```
#Predicting the testing data using the model
pred <- predict(model2, testing)
pred

#Using leave-one-outcross-validation (LOOCV)
lo_model <- trainControl(method="LOOCV")
model3 <- train(Birthweight ~ ., data = data, method = "lm", trControl = lo_model)
summary(model3)

#Using k-fold cross-validation with k=10
kfold <- trainControl(method = "cv", number = 10,summaryFunction = defaultSummary)
model4 <- train(Birthweight ~ ., data = data, method = "lm", trControl = kfold)
summary(model4)
```