Project 3 - cross validation

University of Massachusetts - Dartmouth

MTH522- Project 3

ABSTRACT

A recent study has sought to predict newborn birthweight by developing a multivariate linear regression model using inputs such as the mother's gestational period, age, height, weight, and smoking habits. To validate the accuracy of the model, three cross-validation methods were employed, including the validation set method, leave-one-out cross-validation, and k-fold cross-validation.

The study found that the multivariate linear regression model could accurately predict newborn birthweights and that cross-validation techniques were effective in verifying the model's performance. The results indicated that the model had good predictive power, offering promising insights for future research in this area.

Overall, this study highlights the potential of using multiple variables to predict newborn birthweights and the importance of cross-validation methods in validating the performance of such models. This research could have significant implications for healthcare professionals seeking to optimize care for both mothers and newborns.

THE ISSUES

To address this challenge of constructing a multivariate linear regression model that utilizes variables such as gestation, age, height, weight, and smoking to predict birth weight, one can explore several techniques to validate the performance of the model. One such method is the validation data set approach, where the dataset is split into two halves, with one half used as the training set and the other as the test set.

Another approach is leave-one-out cross-validation (LOOCV), where the model is tested by leaving one observation out of the dataset and using the remaining observations to train the model. This process is repeated for each observation in the dataset, allowing for a more robust evaluation of the model's performance.

Lastly, k-fold cross-validation can also be employed to test the linear model's accuracy. This involves dividing the dataset into k equally-sized partitions and training the model k times, each time using a different partition as the test set and the remaining partitions as the training set.

These techniques ensure that the multivariate linear regression model accurately predicts newborn birth weights, while accounting for potential issues such as multicollinearity. This information can be useful for healthcare professionals seeking to provide optimal care for newborns and mothers.

THE FINDINGS

The dataset consists of 1236 rows and 5 columns, containing variables such as Pregnancy, Age, Height, Weight, Smoking, and Birthweight. The objective is to utilize the variables Gestation, Age, Height, Weight, and Smoking to model the outcome variable, Birthweight, through the multivariate linear regression fitting process.

To evaluate the performance of the model, the R-squared value and mean squared error were calculated. The R-squared value was determined to be 0.030555431130317556, which suggests that the predictor variables can only account for a small proportion of the variation in birthweight. The mean squared error was found to be 349.08056812733525.

To validate the model further, leave-one-out cross-validation (LOOCV) was used, resulting in a LOOCV R-squared value of 0.0 and a LOOCV mean squared error of 326.4544251198769. This indicates that the model did not perform well in predicting individual birthweights.

In addition, a linear model with a k value was tested using the K-fold cross-validation method, yielding an average R-squared score of 0.01820530804096538, which suggests that the model's predictive power is not particularly strong.

Overall, the findings suggest that the selected predictor variables may not be effective in predicting birthweight accurately. Further research may be necessary to identify additional variables that could improve the model's accuracy

DISCUSSION

A linear regression model was fitted to the data, with birthweight as the dependent variable and gestation period, mother's age, height, weight, and smoking status as independent variables. The results of the regression model showed that gestation period, height, and smoking status had a significant impact on birthweight. The coefficient estimates suggested that on average, an increase of one week in gestation period was associated with an increase of 0.013 grams in birthweight. An increase of one unit in height was associated with an increase of 0.526 grams in birthweight, while smoking during pregnancy was associated with a decrease of 1.989 grams in birthweight. However, age and weight did not have a statistically significant effect on birthweight.

To test the accuracy of the model, two cross-validation methods were used: leave-one-out cross-validation (LOOCV) and K-fold cross-validation (K=10). The results of LOOCV were inconclusive, with the delta value being NaN. The K-fold cross-validation method resulted in a root mean squared error (RMSE) of 23.47, indicating that the model has some predictive power, although the RMSE value is relatively high.

In conclusion, the regression model showed that gestation period, height, and smoking status have a significant impact on birthweight. However, the K-fold cross-validation results suggest that the model's predictive power may not be very strong. Future research could explore additional variables that may affect birthweight and improve the model's accuracy.

Appendix A: Method

The code provided loads a dataset on birthweight of babies and performs linear regression analysis on the data. The first part of the code loads the dataset using the "readxl" library and reads it into the console. A linear regression model is then created using the "lm" function, with "Birthweight" as the response variable and "Gestation", "Age", "Height", "Weight", and "Smoke" as the predictor variables. The "summary" function is then used to display the results of the linear regression model, including the coefficients and their significance levels, as well as the residual standard error, multiple R-squared, adjusted R-squared, F-statistic, and p-value.

The next section of the code uses cross-validation methods to test the accuracy of the linear regression model. The first method is the validation approach, where half of the data is randomly selected for training and the remaining half is used for testing. The "predict" function is used to predict the birthweight of babies in the validation set based on the linear regression model. The mean squared error (MSE) and root mean squared error (RMSE) are then calculated to assess the accuracy of the predictions.

The code also includes two other cross-validation methods: leave-one-out cross-validation (LOOCV) and K-fold cross-validation (K=10). LOOCV involves leaving out one observation at a time and using the remaining observations to fit the model, and then using the model to predict the value of the left-out observation. The process is repeated for each observation, and the MSE is calculated based on the difference between the predicted and actual values. However, in this case, the LOOCV function produces NaN values. K-fold cross-validation involves dividing the data into K subsets, using K-1 subsets for training and the remaining subset for testing. This process is repeated K times, with each subset being used for testing once. The RMSE is then calculated to assess the accuracy of the model.

Finally, the code loads several libraries including "boot", "ggplot2", "lattice", "caret", and "tidyverse", which provide functions for data visualization and statistical analysis.

METHOD B: RESULT AND CODE

```
# Load the data into the console
library(readxl)
Babies<-read_excel("C:/Users/budap/Downloads/babies_weight.xls")
# Linear regression model
model <- lm(Birthweight ~ Gestation + Age + Height + Weight + Smoke,
data = Babies)
summary(model)
```

```
##
## Call:
## lm(formula = Birthweight ~ Gestation + Age + Height + Weight +
##     Smoke, data = Babies)
##
## 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
```

```
#using cross validation method to test our model
#Validation Approach
set.seed(123) # for reproducibility
n <- nrow(Babies)
train_index<- sample(1:n, n/2, replace = FALSE) #randomly select half of the data for training
validation <- setdiff(1:n, train_index) # use the remaining half for testing
train <- Babies[train_index,]
validation<- Babies[-train_index,]
model_train <- lm(Birthweight ~ Gestation + Age + Height + Weight + Smoke, data = train)
predict_validation <- predict(model_train, newdata = validation)
# Compute mean squared error (MSE) on the validation set
MSE <- mean((validation$Birthweight - predict_validation)^2)
RMSE <- sqrt(MSE)
#Leave-one-out cross-validation (LOOCV):
library(boot)
LOOCV <- cv.glm(Babies, model, K = nrow(Babies))
LOOCV$delta
```

```
## [1] NaN NaN
```

```r
#K-fold cross-validation (K=10):
library(ggplot2)
library(lattice)
```

```
##
## Attaching package: 'lattice'

## The following object is masked from 'package:boot':
##
##     melanoma
```

```r
library(caret)
set.seed(123) # for reproducibility
folds <- createFolds(Babies$Birthweight, k = 10)
control <- trainControl(method = "cv", index = folds)
model_cv <- train(Birthweight ~ Gestation + Age + Height + Weight + Smoke,
                  data = Babies, method = "lm", trControl = control)
model_cv$results$RMSE
```

```
## [1] 23.46952
```

```r
library(boot)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
tidyverse_conflicts()
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Load the data into the console
library(readxl)
Babies<-read_excel("C:/Users/budap/Downloads/babies_weight.xls")
```

```r
# Linear regression model
model <- lm(Birthweight ~ Gestation + Age + Height + Weight + Smoke,
            data = Babies)
summary(model)
```

```
##
## Call:
## lm(formula = Birthweight ~ Gestation + Age + Height + Weight +
##     Smoke, data = Babies)
##
## 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
```

```
#using cross validation method to test our model
#Validation Approach
set.seed(123) # for reproducibility
n <- nrow(Babies)
train_index<- sample(1:n, n/2, replace = FALSE) #randomly select half of the data for training
validation <- setdiff(1:n, train_index) # use the remaining half for testing
train <- Babies[train_index,]
validation<- Babies[-train_index,]
model_train <- lm(Birthweight ~ Gestation + Age + Height + Weight + Smoke, data = train)
predict_validation <- predict(model_train, newdata = validation)
```

```
# Compute mean squared error (MSE) on the validation set
MSE <- mean((validation$Birthweight - predict_validation)^2)
RMSE <- sqrt(MSE)
```

```
#Leave-one-out cross-validation (LOOCV):
library(boot)
LOOCV <- cv.glm(Babies, model, K = nrow(Babies))
LOOCV$delta
```

```
## [1] NaN NaN
```

```
#K-fold cross-validation (K=10):
library(ggplot2)
library(lattice)
library(caret)
set.seed(123) # for reproducibility
folds <- createFolds(Babies$Birthweight, k = 10)
```

```r
control <- trainControl(method = "cv", index = folds)
model_cv <- train(Birthweight ~ Gestation + Age + Height + Weight + Smoke,
                  data = Babies, method = "lm", trControl = control)
model_cv$results$RMSE
```

```
## [1] 23.46952
```