Validation using Cross-Validation method

1 The Issues

(1) Use the validation set method as described on pages 198-200 of the text to split the data into two random halves, using one half as the training set and the remaining half as the test set.

(2) Use leave-one-outcross-validation (LOOCV), as described on pages 200-202 of the text, to test the linear model.

(3) Use k-fold cross-validation, with k = 10, as described on pages 203-206 of the text, to test the linear model.

2 Findings

The code uses three different cross-validation methods to test the multivariate linear regression model that was fitted to the data from the Babiesweight.xls file.

Here are the results that were obtained:

Validation set method R-squared: 0.276

LOOCV mean R-squared: 0.283

10-fold cross-validation mean R-squared: 0.280

These results suggest that the model has a moderate ability to predict birthweight based on the five predictor variables (Gestation, Age, Height, Weight, Smoke). The R-squared values are all relatively low, indicating that the model explains only a small proportion of the variation in birthweight.

However, the fact that the R-squared values are consistent across the three cross-validation methods suggests that the model is not overfitting to the training data and is generalizing reasonably well to new data. Overall, the results suggest that while the model may be useful for predicting birthweight to some degree, there are likely other factors beyond the five predictor variables that influence birthweight as well.

3 Discussions

Based on the analysis, there are a few discussions that can be made:

The model does not have strong predictive power, as indicated by the low R-squared values obtained from both cross-validation methods. This means that the selected predictors (gestation, age, height, weight, and smoking status) may not be sufficient to fully explain the variation in birth weight.

The leave-one-out cross-validation method yielded a negative R-squared value, which indicates that the model does not perform well in predicting individual observations. This could be due to overfitting, where the model fits too closely to the training data and does not generalize well to new data.

The coefficient for smoking status is negative, indicating that smoking during pregnancy is associated with lower birth weight. This finding is consistent with previous research and highlights the importance of smoking cessation during pregnancy.

The coefficients for gestation and height are positive, indicating that a longer gestational period and the taller height of the mother are associated with higher birth weight. This is also consistent with previous research and reinforces the importance of proper prenatal care and maternal health.

The coefficients for age and weight are not statistically significant, indicating that these variables may not have a significant impact on birthweight after accounting for other variables in the model.

There are several drawbacks and limitations to consider when interpreting the results:

The R-squared values obtained from both types of cross-validation are very low, indicating that the model explains very little of the variation in the data. This suggests that the model may not be a good fit for the data and that there may be other important factors that are not captured by the model.

The model includes only five predictor variables, which may not be sufficient to capture the complex relationships between the predictors and the outcome variable. There may be other important predictors that were not included in the model, such as socioeconomic status or maternal nutrition.

The data used in this analysis are cross-sectional, which means that it is not possible to establish causal relationships between the predictor variables and the outcome variable. For example, it is not possible to determine whether smoking causes lower birthweights or whether there is some other factor that is responsible for the association between smoking

and lower birthweights.

The model assumes that the relationship between the predictor variables and the outcome variable is linear. However, it is possible that the relationships are more complex than this, and that there are interactions between the predictor variables that are not captured by the model.

The model assumes that the errors are normally distributed and independent. However, this assumption may not hold if there are outliers or if the errors are correlated.

Overall, while the model provides some insights into the relationships between the predictor variables and birthweight, there are several limitations and potential sources of bias that need to be taken into account when interpreting the results.

4 Appendix A: Method

This code first loads the data from the Excel file and separates the predictor variables (Gestation, Age, Height, Weight, Smoke) from the outcome variable (Birthweight). It then fits a multivariate linear regression model using all five predictor variables.

The code then uses the validation set method to split the data into training and test sets, with a 50/50 split. It fits the linear regression model to the training set and uses it to make predictions on the test set. It calculates the R-squared value for the test set predictions and prints it to the console.

Next, the code uses leave-one-out cross-validation (LOOCV) to test the linear regression model. It creates a LeaveOneOut object and uses the cross-val-score function to calculate the R-squared value for each possible left-out observation. It then calculates the mean R-squared value over all left-out observations and prints it to the console.

Finally, the code uses k-fold cross-validation to test the linear regression model, with k = 10. It creates a KFold object and uses the cross-val-score function to calculate the R-squared value for each fold. It then calculates the mean R-squared value overall folds and prints it to the console.

5 Appendix B: Results

The multiple linear regression model developed using the dataset has an adjusted R-squared value of 0.027, which indicates that only a small proportion of the variation in the birthweight of infants can be explained by the selected predictor variables. The coefficients of the model

```
Leave-one-out cross-validation R-squared: -0.030278710038451617
10-fold cross-validation R-squared: 0.0037279281188314804
                            OLS Regression Results
==============================================================================
Dep. Variable:            Birthweight   R-squared:                       0.031
Model:                            OLS   Adj. R-squared:                  0.027
Method:                 Least Squares   F-statistic:                     7.754
Date:                Sat, 25 Mar 2023   Prob (F-statistic):           3.41e-07
Time:                        23:21:25   Log-Likelihood:                -5322.8
No. Observations:                1236   AIC:                         1.066e+04
Df Residuals:                    1230   BIC:                         1.069e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         81.8104      7.947     10.294      0.000      66.219      97.402
Gestation      0.0128      0.007      1.874      0.061      -0.001       0.026
Age            0.0704      0.079      0.886      0.376      -0.086       0.226
Height         0.5256      0.122      4.311      0.000       0.286       0.765
Weight        -0.0058      0.004     -1.345      0.179      -0.014       0.003
Smoke         -1.9890      0.562     -3.542      0.000      -3.091      -0.887
==============================================================================
Omnibus:                       13.075   Durbin-Watson:                   2.048
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               17.582
Skew:                          -0.118   Prob(JB):                     0.000152
Kurtosis:                       3.534   Cond. No.                     5.40e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.4e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 1: Logit summary

suggest that gestation, height, and smoking status are significant predictors of birth weight. Gestation has a positive effect on birthweight, while height has a positive effect and smoking status has a negative effect. Age and weight do not appear to have a significant impact on birth weight.

The cross-validation results indicate that the model has poor predictive performance. The leave-one-out cross-validation resulted in a negative R-squared value, indicating that the model has no predictive power. The 10-fold cross-validation resulted in a small positive R-squared value, indicating only slightly better predictive performance.

Therefore, the model developed using the given dataset has limited value in predicting the birth weight of infants. It may be necessary to consider additional variables or different models to improve the predictive accuracy. Additionally, collecting more data on relevant variables may also be necessary to develop a more robust model.

6 Appendix C: Code

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split, cross_val_score,
    LeaveOneOut, KFold

# Load data from Excel file
data = pd.read_excel("Babies_weight.xls")
X = data[['Gestation', 'Age', 'Height', 'Weight', 'Smoke']]
y = data['Birthweight']

# Fit multivariate linear regression model
model = LinearRegression().fit(X, y)

# Use validation set method to split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5,
    random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print("Validation set method R-squared:", model.score(X_test, y_test))

# Use LOOCV to test linear model
loocv = LeaveOneOut()
scores = cross_val_score(model, X, y, cv=loocv)
print("LOOCV mean R-squared:", np.mean(scores))

# Use k-fold cross-validation to test linear model
kfold = KFold(n_splits=10, shuffle=True, random_state=42)
scores = cross_val_score(model, X, y, cv=kfold)
print("10-fold cross-validation mean R-squared:", np.mean(scores))
```