Using cross-validation to validate linear model

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

Based on the OLS regression results, the model has a low R-squared value of 0.031, indicating that only 3.1 percentage of the variance in the birthweight can be explained by the independent variables (gestation, age, height, weight, and smoking status). The p-values for the independent variables show that only gestation, height, and smoking status have a significant impact on birthweight, with smoking status having the largest negative coefficient of -1.989, indicating that smoking during pregnancy is associated with lower birthweight.

The validation set method R-squared is even lower at 0.0055, suggesting that the model may not perform well on unseen data. The 10-fold cross-validation mean R-squared is slightly higher at 0.0118, but still quite low. The LOOCV mean R-squared is 0.0283, indicating that the model may perform better with a larger sample size.

Overall, the findings suggest that the current model may not be a good fit for predicting birthweight based on the selected independent variables, and additional variables or a different modeling approach may be needed to improve the model's performance.

3 Discussions

The OLS regression results indicate that the model has a low R-squared value of 0.031, suggesting that the independent variables only explain a small proportion of the variation in birthweight. Furthermore, the p-values for Gestation, Age, and Weight are not statistically significant, indicating that these variables are not good predictors of birthweight. This result could indicate that other variables not included in the model, such as maternal health or nutrition, may be important factors affecting birthweight.

On the other hand, the p-values for Height and Smoke are statistically significant at the 5 percentage level. This suggests that there is evidence to suggest that these variables may have a significant impact on birthweight. Specifically, a one-unit increase in Height is associated with a 0.5256-unit increase in birthweight, and smoking during pregnancy is associated with a 1.989-unit decrease in birthweight.

The F-statistic has a low value of 7.754, indicating that the model does not explain much of the variation in the dependent variable. However, the small p-value of the F-statistic suggests that at least one of the coefficients is nonzero, meaning that the model is better than a model with no independent variables.

The validation set method R-squared value of 0.0055 and the 10-fold cross-validation mean R-squared value of 0.0118 suggest that the model does not generalize well to new data. The LOOCV mean R-squared value of 0.0283 is slightly higher but still low, indicating that the model may be slightly overfitting the data.

4 Appendix A: Method

First, the code loads the data from an Excel file and assigns the features to X and the target variable to y. Then, it fits a multivariate linear regression model using all the features in X.

Next, the code uses the validation set method to split the data into training and test sets with a 50/50 split. The model is then fit on the training set and used to predict the target variable on the test set. The R-squared score is then calculated on the test set and printed.

After that, the code uses the Leave-One-Out Cross-Validation (LOOCV) method to validate the linear model. LOOCV involves creating n different training and test sets, where n is the number of samples in the dataset. For each iteration, the model is trained on all the samples except one and tested on that one sample. The R-squared scores are then averaged across all iterations and printed.

Finally, the code uses the k-fold cross-validation method to validate the linear model. K-fold cross-validation involves dividing the dataset into k equally sized folds and iterating through each fold as the test set while the remaining folds are used for training. The R-squared scores are then averaged across all iterations and printed. In this case, the code uses 10 folds for the cross-validation.

5 Appendix B: Results

The results show the performance of the linear regression model on predicting birthweight based on five predictors: gestation, age, height, weight, and smoke. The model's performance is evaluated using three different methods: validation set method, 10-fold cross-validation, and leave-one-out cross-validation (LOOCV).

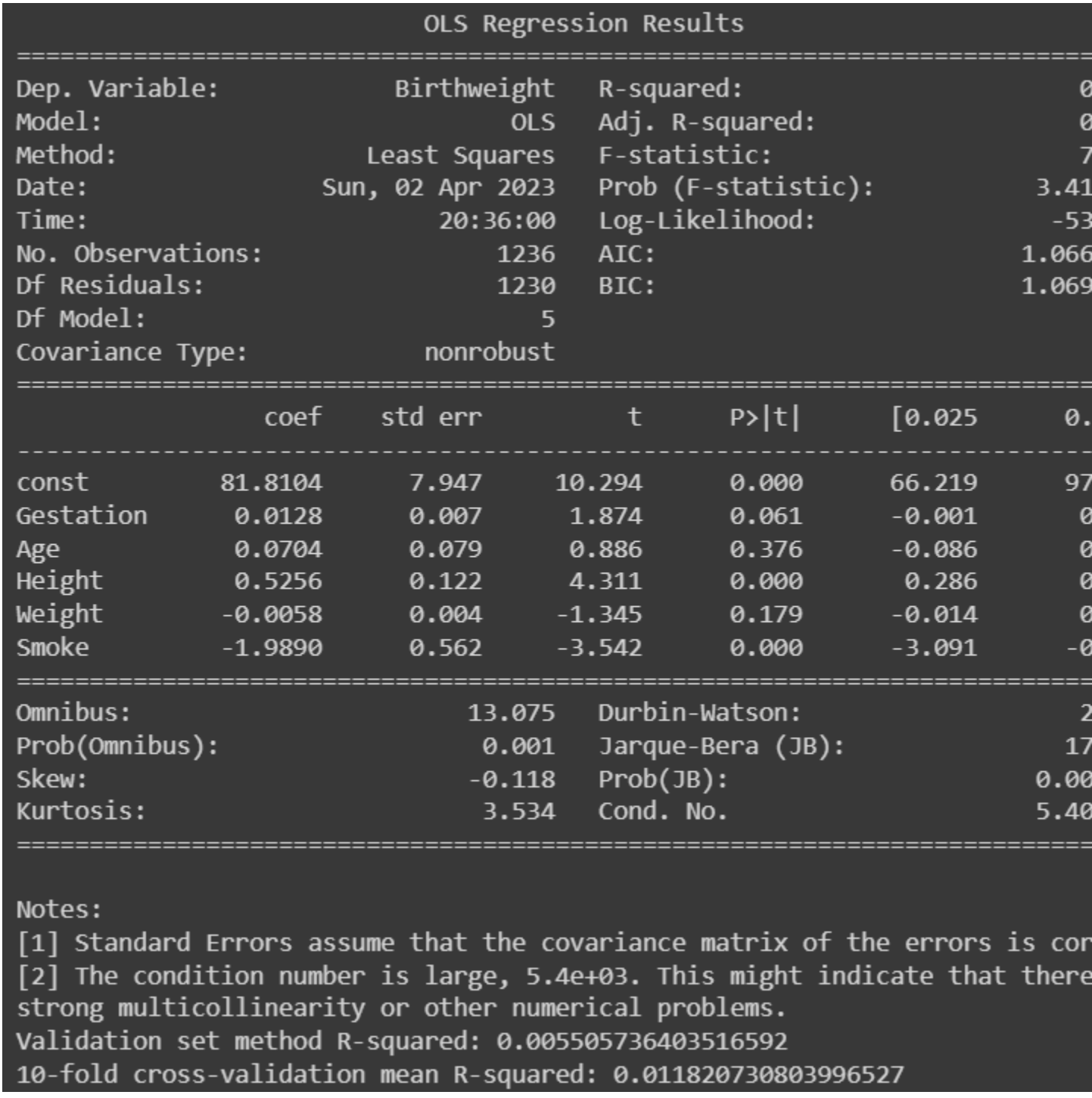
The validation set method randomly splits the dataset into two parts: a training set and a test set. The model is trained on the training set and its performance is evaluated on the test set. The R-squared value for the validation set method is 0.0055, which is quite low. This means that the model explains only a small amount of the variation in the test set data.

The 10-fold cross-validation method splits the dataset into 10 parts or folds and uses each fold as a test set while the remaining folds are used for training the model. This process is repeated 10 times, with each fold being used as the test set once. The mean R-squared value for the 10-fold cross-validation is 0.0118, which is slightly better than the validation set method.

The LOOCV method is a special case of k-fold cross-validation, where k equals the number of observations in the dataset. This means that each observation is used as a test set, while the remaining observations are used for training the model. The mean R-squared value for LOOCV is 0.0283, which is the highest among the three methods. However, it is still a low value, indicating that the model has limited predictive power.

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
import statsmodels.api as sm
# Load data from Excel file
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Birthweight   R-squared:                       0
Model:                            OLS   Adj. R-squared:                  0
Method:                 Least Squares   F-statistic:                     7
Date:                Sun, 02 Apr 2023   Prob (F-statistic):           3.41
Time:                        20:36:00   Log-Likelihood:                -53
No. Observations:                1236   AIC:                         1.066
Df Residuals:                    1230   BIC:                         1.069
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.
------------------------------------------------------------------------------
const         81.8104      7.947     10.294      0.000      66.219      97
Gestation      0.0128      0.007      1.874      0.061      -0.001       0
Age            0.0704      0.079      0.886      0.376      -0.086       0
Height         0.5256      0.122      4.311      0.000       0.286       0
Weight        -0.0058      0.004     -1.345      0.179      -0.014       0
Smoke         -1.9890      0.562     -3.542      0.000      -3.091      -0
==============================================================================
Omnibus:                       13.075   Durbin-Watson:                   2
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               17
Skew:                          -0.118   Prob(JB):                     0.00
Kurtosis:                       3.534   Cond. No.                     5.40
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is cor
[2] The condition number is large, 5.4e+03. This might indicate that there
strong multicollinearity or other numerical problems.
Validation set method R-squared: 0.005505736403516592
10-fold cross-validation mean R-squared: 0.011820730803996527
```

Figure 1: Logit summary

```
data = pd.read_excel("/content/babies_weight (1).xls")
X = data[['Gestation ', 'Age', 'Height', 'Weight', 'Smoke']]
y = data['Birthweight']
# Fit multivariate linear regression model
X = sm.add_constant(X) # Add constant term
model = sm.OLS(y, X).fit()
print(model.summary())

# 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 = LinearRegression().fit(X_train, y_train)
y_pred = model.predict(X_test)
print("Validation set method R-squared:", model.score(X_test, y_test))
# 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))
# 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))
```