Homework 06

Due Wednesday, March 26, 2025, 11:59 PM

YOUR NAME

Today's Date: 2025-03-19

Problem 1

The textbook describes that the `cv.glm()` function can be used in order to compute the Leave-One Out Cross Validation (LOOCV) test error estimate. Alternatively, one could compute those quantities using just the `glm()` and `predict.glm()` functions, and a for loop. You will now take this approach in order to compute the LOOCV error for a simple logistic regression model on the `Weekly` data set.

```
> library(ISLR2)
> data("Weekly")
>
> head(Weekly)
```

```
  Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

(a) Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2`. Report and comment on the result.

(b) Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2` using all but the first observation. Report and comment on the result.

(c) Use the model from (b) to predict the direction of the first observation. You can do this by predicting that the first observation will go up if $\Pr(\texttt{Direction} = \text{“}Up\text{”}|\texttt{Lag1}, \texttt{Lag2}) > 0.5$. Was this observation correctly classified?

(d) Write a for loop from $i = 1$ to $i = n$, where n is the number of observations in the data set, that performs each of the following steps:

i. Fit a logistic regression model using all but the ith observation to predict Direction using `Lag1` and `Lag2`.
ii. Compute the posterior probability of the market moving up for the ith observation.

iii. Use the posterior probability for the ith observation in order to predict whether or not the market moves up.
iv. Determine whether or not an error was made in predicting the direction for the ith observation. If an error was made, then indicate this as a 1, and otherwise indicate it as a 0.

(e) Take the average of the n numbers obtained in part **(d) iv.** in order to obtain the LOOCV estimate for the test error. Comment on the results.

Problem 2

This question uses the variables `dis` (the weighted mean of distances to five Boston employment centers) and `nox` (nitrogen oxides concentration in parts per 10 million) from the `Boston` data. We will treat `dis` as the predictor and `nox` as the response.

```
> data("Boston")
> head(Boston[, c('dis', 'nox')])
```

```
     dis   nox
1 4.0900 0.538
2 4.9671 0.469
3 4.9671 0.469
4 6.0622 0.458
5 6.0622 0.458
6 6.0622 0.458
```

(a) Use the `poly()` function to fit a cubic polynomial regression to predict `nox` using `dis`. Report and comment on the regression output, and plot the resulting data and polynomial fits.

(b) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10); report and comment on the associated residual sum of squares.

(c) Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.

(d) Use the `bs()` function to fit a regression spline to predict `nox` using `dis`. Report and comment on the output for the fit using four degrees of freedom. How did you choose the knots? Plot the resulting fit.

(e) Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits; report and comment on the resulting RSS. Describe the results obtained.

(f) Perform cross-validation or another approach in order to select the best degrees of freedom for a regression spline on this data. Describe your results.