Problem Set #4

Ian Bach

2024-09-13

1. (10 points) This question involves the Default dataset, which is included in the package ISLR2.

(a) Fit a logistic regression that uses income and balance to predict default (Model 1). Report the estimates. [Hint: it should include an intercept.]
(b) Suppose we classify an individual to the default category with a threshold probability of 0.5. Compute the LOOCV test error estimate for Model 1.
(c) Now consider a logistic regression that predicts default using income, balance, and a dummy variable for student (Model 2). Compute the LOOCV test error estimate for this model. Should we include a dummy variable for student?

1A.

$$\text{logit}(P(\text{default} = 1)) = \beta_0 + \beta_1 \times \text{income} + \beta_2 \times \text{balance}$$

```
# Load the necessary package and data
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.4.1
```

```
# Fit a logistic regression model (Model 1)
model1 <- glm(default ~ income + balance, data = Default, family = "binomial")

# View the model summary to see the estimates
summary(model1)
```

```
##
## Call:
## glm(formula = default ~ income + balance, family = "binomial",
##     data = Default)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

1B.

```
# Load the boot package
library(boot)

# Define a function to calculate the prediction error for LOOCV
loocv_error <- function(model, data) {
  # LOOCV using cv.glm from the boot package
  loocv_result <- cv.glm(data, model, K = nrow(data))
  return(loocv_result$delta[1])  # LOOCV estimate
}

# Calculate LOOCV error for Model 1
loocv_model1 <- loocv_error(model1, Default)
loocv_model1
```

```
## [1] 0.02146706
```

1C.

$$\text{logit}(P(\text{default} = 1)) = \beta_0 + \beta_1 \times \text{income} + \beta_2 \times \text{balance} + \beta_3 \times \text{student}$$

```
# Fit a logistic regression model (Model 2) with student dummy variable
model2 <- glm(default ~ income + balance + student, data = Default, family = "binomial")

# Calculate LOOCV error for Model 2
loocv_model2 <- loocv_error(model2, Default)
loocv_model2
```

```
## [1] 0.02139653
```

2. (10 points) This question uses the bootstrap to estimate the standard errors under Model 2 in Exercise 1.

(a) Write a function, boot.estimates, that takes as input the Default dataset and an index of the observations, and outputs the coefficient estimates of income, balance and student in a logistic regression (with an intercept).
(b) Generate 1000 bootstrapped samples and use boot.estimates to estimate the standard errors of the coefficients corresponding to income, balance and student.

2A.

```r
# Define the boot.estimates function
boot.estimates <- function(data, index) {
  # Subset the data using the provided index
  boot_sample <- data[index, ]

  # Fit a logistic regression model using the bootstrapped data
  model <- glm(default ~ income + balance + student, data = boot_sample, family = "binomial")

  # Return the coefficients of income, balance, and student
  return(coef(model)[c("income", "balance", "studentYes")])  # studentYes corresponds to the dummy for
}
```

2B.

```r
# Load necessary library
library(boot)

# Define the bootstrap function to calculate estimates
boot.estimates <- function(data, index) {
  boot_sample <- data[index, ]
  model <- glm(default ~ income + balance + student, data = boot_sample, family = "binomial")
  return(coef(model)[c("income", "balance", "studentYes")])
}

# Perform the bootstrap with 1000 replications
set.seed(123)  # For reproducibility
boot_results <- boot(Default, boot.estimates, R = 1000)

# View the bootstrap results
boot_results
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Default, statistic = boot.estimates, R = 1000)
##
##
## Bootstrap Statistics :
##          original        bias     std. error
## t1*  3.033450e-06  1.160860e-07 8.249317e-06
## t2*  5.736505e-03  1.681596e-05 2.294104e-04
## t3* -6.467758e-01 -3.118653e-03 2.427747e-01
```

```r
# Calculate the standard errors of the coefficients
boot_se <- apply(boot_results$t, 2, sd)
boot_se
```

```
## [1] 8.249317e-06 2.294104e-04 2.427747e-01
```