Kristina White
ACMS 40878
Final Project
November 20, 2020

Iteratively Reweighted Least Squares on Heart Data

This project uses a dataset on heart disease from the University of California Irvine. The dataset contains the medical history from patients in Hungary and Switzerland. Its dimensions are 303 by 14, or 303 observations for 13 predictor variables and one response variable. The heart dataset predicts whether a patient has heart disease based on the factors age, sex, type of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic measurement, maximum heart rate achieved, exercise-induced angina, ST depression, slope of the peak exercise ST segment, number of major blood vessels colored by flourosopy, and the blood disorder thalassemia. The response variable is "target." This is a binary variable that is 1 if the patient has been diagnosed with heart disease and 0 if not. I have summarized the predictor variables in **Table 1.1** below.

Table 1.1	
ï..age	age in years
sex	1=male, 0=female
cp	Level of chest pain experienced (1=typical angina; 2=atypical angina; 3=non-anginal pain, 4=asymptomatic
trestbps	Resting blood pressure (mm hg) on admission to the hospital
chol	Serum cholesterol (mg/dl)
fbs	Fasting blood sugar > 120 mg/dl (1=true, 0=false).
restecg	Resting electrocardiographic measurement (0=normal; 1=st-t wave abnormality; 2=probable or definite left ventricular hypertrophy)
thalach	Maximum heart rate
exang	Exercise-induced agina (1=yes, 0=no).
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise st segment (1=upsloping, 2=flat, 3=downsloping)
ca	Number of major blood vessels (0-3) colored by fluoroscopy

thal	Blood disorder thalassemia (3=normal; 6=fixed defect; 7=reversable defect)

I performed logistic regression on the heart data using the iteratively reweighted least squares algorithm, which is a multivariate application of Newton's method. Logistic regression models the response variable as $y_i|x_i \sim \text{Bernoulli}(\pi_i)$ where:

$$\pi_i = \exp(\beta^T x_i)/(1+ \exp(\beta^T x_i))$$

and

$$\beta = (\beta_0, \beta_1,\ldots,\beta_p)^T$$

x is a vector of p covariate values and y is a vector of n binary responses. π_i is the probability of success. β is a multivariate vector of parameters we want to estimate using Newton's method. The log-odds or regression function is:

$$\log(\pi_i/(1- \pi_i)) = \beta^T x_i$$

The likelihood and log-likelihood of β functions can be worked out to:

$$L(\beta) = \Pi_i^n (\exp(\beta^T x_i)/(1+ \exp(\beta^T x_i))) \text{ ^} y_i * (1/(1+ \exp(\beta^T x_i))\text{^}(1-y_i)$$

and

$$\log(L(\beta)) = y^T X\beta - b^T 1$$

where

$$b = (\log(1+ \exp(\beta^T x_1))\ldots\log(1+ \exp(\beta^T x_p)))$$

and

X is a n x (p+1) matrix.

The score function is found by taking the derivative of the log-likelihood function with respect to β, which simplifies to:

$$\frac{d}{d\beta}\log(L(\beta)) = X^T(y-\pi)$$

And the Hessian matrix is found by taking the second derivative of the log-likelihood function with respect to β, which simplifies to:

$$\frac{d}{d\beta}(X^T(y-\pi)) = -X^T W X$$

where W is a diagonal matrix with the ith entry equal to $\pi_i(1-\pi_i)$. So Newton's update is:

$$\beta^{(t+1)} = \beta^{(t)} - (\log(L(\beta^{(t)}))'')^{-1} \log(L(\beta^{(t)}))' = \beta^{(t)} - (-X^T W X)^{-1}(X^T(y-\pi))$$

where $\pi^{(t)}$ is the value of π corresponding to $\beta^{(t)}$, and W is the diagonal weight matrix evaluated at $\pi^{(t)}$.

I implemented this iteratively reweighted least squares algorithm in R to perform logistic regression on the heart data. The goal is to estimate β. I used 5 of the 13 predictor variables (ï..age , trestbps, chol, thalach, oldpeak) to predict whether a patient has heart disease.

I first defined n, X, and Y:

```
n = dim(heart)[1]

X = as.matrix(heart[,c("ï..age", "trestbps", "chol", "thalach", "oldpeak")])

X = cbind(rep(1,n), X)

Y = as.vector(heart[,"target"])
```

where n is the number of observations for each variable, X is a matrix of the observations for each of the predictor variables with an added vector of constants, and Y is the binary response variable "target."

I next prepared to iterate using Newton's method over 10 iterations:

```
iter = 10

beta = matrix(NA, iter, dim(X)[2])

W = matrix(0, n, n)


i = 1

beta[i,] = rep(0.0,dim(X)[2])
```

I initialized the diagonal weight matrix W and set all values to 0, and I chose the starting values for β to be 0 for all variables. I used the strategy of starting with small values to find the initial values for β and using 0s caused the estimated parameter to converge.

Next, I implemented Newton's method:

```
for (i in 1:(iter-1)) {
  pi = exp(X %*% beta[i,])/(1 + exp(X %*% beta[i,]))
  for (j in 1:n) {
    W[j,j] = pi[j] * (1 - pi[j])
  }
  cov.matrix = solve(t(X) %*% W %*% X)
  beta[i+1,] = beta[i,] + cov.matrix %*% t(X) %*% (Y - pi)
```

}

beta

where, in terms of my discussion of logistic regression and the iteratively reweighted least squares method:

$$pi = \pi_i ,$$

W[j, j] is the ith entry in W,

$$\text{cov.matrix} = \frac{d}{d\beta}(X^T(y-\pi)) ,$$

and

$$beta[i+1,] = \beta^{(t+1)}$$

The results for the estimated parameters of β are:

β_0	β_1	β_2	β_3	β_4	β_5
0.000000	0.000000000	0.000000000	0.000000000	0.00000000	0.0000000
-1.743238	-0.001845576	-0.007031514	-0.002162190	0.02692533	-0.5285797
-1.947571	-0.001419914	-0.010477991	-0.002946262	0.03358064	-0.6883319
-1.963853	-0.001229776	-0.011108580	-0.003083403	0.03454798	-0.7126341
-1.964169	-0.001223795	-0.011122260	-0.003086398	0.03456763	-0.7131361
-1.964169	-0.001223791	-0.011122266	-0.003086400	0.03456764	-0.7131363
-1.964169	-0.001223791	-0.011122266	-0.003086400	0.03456764	-0.7131363
-1.964169	-0.001223791	-0.011122266	-0.003086400	0.03456764	-0.7131363
-1.964169	-0.001223791	-0.011122266	-0.003086400	0.03456764	-0.7131363
-1.964169	-0.001223791	-0.011122266	-0.003086400	0.03456764	-0.7131363

Where β_0 is the intercept and $\beta_1,\ldots,\beta_5$ are the coefficients for ï..age, trestbps, chol, thalach, and olpeak, respectively. The results obtained from using the iteratively reweighted least squares method closely match the results from performing logistic regression using glm:

logit <- glm(target ~ ï..age + trestbps + chol + thalach + oldpeak, data = heart, family = "binomial")

Coefficients

Intercept	-1.964169
ï..age	-0.001224
trestbps	-0.011122
chol	-0.003086
thalach	0.034568
oldpeak	-0.713136

The results of logistic regression using iteratively reweighted least squares on the heart data can be interpreted as follows:

For every one-unit increase in ï..age, the log odds of heart disease decrease by 0.001223791.
For every one-unit increase in trestbps, the log odds of heart disease decrease by 0.011122266.
For every one-unit increase in chol, the log odds of heart disease decrease by 0.003086400.
For every one-unit increase in thalach, the log odds of heart disease increase by 0.03456764.
For every one-unit increase in oldpeak, the log odds of heart disease decrease by 0.7131363.

So, according to this regression, a patient with higher age, resting blood pressure on admission to the hospital, serum cholesterol, or ST depression induced by exercise relative to rest, considering each factor individually, has a lower log odds of having heart disease, while a patient with a higher maximum heart rate has a higher log odds of having heart disease.

R Code:

```
heart <- read.csv("heart.csv")
head(heart)
dim(heart)

n = dim(heart)[1]
X = as.matrix(heart[,c("ï..age", "trestbps", "chol", "thalach", "oldpeak")])
X = cbind(rep(1,n), X)
Y = as.vector(heart[,"target"])

iter = 10
beta = matrix(NA, iter, dim(X)[2])
W = matrix(0, n, n)
i = 1
beta[i,] = rep(0.0,dim(X)[2])

for (i in 1:(iter-1)) {

pi = exp(X %*% beta[i,])/(1 + exp(X %*% beta[i,]))

  for (j in 1:n) {

    W[j,j] = pi[j] * (1 - pi[j])

  }

  cov.matrix = solve(t(X) %*% W %*% X)

  beta[i+1,] = beta[i,] + cov.matrix %*% t(X) %*% (Y - pi)

}
beta

logit <- glm(target ~ ï..age + trestbps + chol + thalach + oldpeak, data = heart, family =
"binomial")
summary(logit)
```