STAT GR5206 Lab 6

Hanao Li hl3202

Instructions

Make sure that you upload an RMarkdown file to the canvas page (this should have a .Rmd extension) as well as the PDF output after you have knitted the file (this will have a .pdf extension). The files you upload to the Canvas page should be updated with commands you provide to answer each of the questions below. You can edit this file directly to produce your final solutions. The lab is due by the end of the semester.

Goal

The goal of Lab 6 is to estimate a logistic regression model via maximum likelihood. The optimization procedure used is Newton's method, which is the most common (and amazing) form of the gradient decent algorithm. A description of the dataset and logistic statistical model follow below:

Data Description

A university medical center urology group was interested in the association between prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostatectomies. The 8 variables are:

Variable	Variable Name	Description
X_1	PSA level	Serum prostate-specific antigen level (mg/ml)
X_2	Cancer volume	Estimate of prostate cancer volume (cc)
X_3	Weight	Prostate weight (gm)
X_4	Age	Age of patient (years)
X_5	Benign prostatic hyperplasia	Amount of benign prostatic hyperplasia (cm2)
X_6	Seminal vesicle invasion	Presence or absence of seminal vesicle invasion
X_7	Capsular penetration	Degree of capsular penetration (cm)
Y	Gleason score	Pathologically determined grade of disease

Below we read in the dataset and name it **prostate**.

```r
setwd("C:/users/36576/Desktop")
prostate <- read.table("Lab6.txt")
head(prostate)
```

```
##      X1     X2     X3 X4 X5 X6 X7 Y
## 1 0.651 0.5599 15.959 50  0  0  0 6
## 2 0.852 0.3716 27.660 58  0  0  0 7
## 3 0.852 0.6005 14.732 74  0  0  0 7
## 4 0.852 0.3012 26.576 58  0  0  0 6
## 5 1.448 2.1170 30.877 62  0  0  0 6
## 6 2.160 0.3499 25.280 50  0  0  0 6
```

In our setting we create a new binary response variable Y, called high-grade cancer by letting $Y = 1$ if Gleason score equals 8, and $Y = 0$ otherwise (i.e., if Gleason score equals 6 or 7). The goal is to carry out a logistic regression analysis, where the response of interest is high-grade cancer (Y).

```
prostate$Y <- ifelse(prostate$Y==8,1,0)
head(prostate)
```

```
##      X1     X2     X3 X4 X5 X6 X7 Y
## 1 0.651 0.5599 15.959 50  0  0  0 0
## 2 0.852 0.3716 27.660 58  0  0  0 0
## 3 0.852 0.6005 14.732 74  0  0  0 0
## 4 0.852 0.3012 26.576 58  0  0  0 0
## 5 1.448 2.1170 30.877 62  0  0  0 0
## 6 2.160 0.3499 25.280 50  0  0  0 0
```

```
nrow(prostate)
```

```
## [1] 97
```

Logistic Model

Let $Y_1, Y_2, \ldots, Y_{97}$ be independent Bernoulli random variables with expected values

$$E[Y_i] = p_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i})}, \quad i = 1, 2, \ldots, 97. \tag{1}$$

Part 1: Fit Logistic Model Using glm()

Problem 1)

Use the base **R** function **glm()** to fit the logistic model. Run the code below:

```
model <- glm(Y~X1+X2+X3+X4+X5+X6+X7,
                data=prostate,
                family=binomial(link = "logit")
                )
model
```

```
##
## Call:  glm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, family = binomial(link = "logit"),
##     data = prostate)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##  -10.276610     0.055228     0.086869     0.001615     0.106633
##          X5           X6           X7
##   -0.071827    -1.154137     0.123324
##
## Degrees of Freedom: 96 Total (i.e. Null);  89 Residual
## Null Deviance:       101.4
## Residual Deviance: 61.57     AIC: 77.57
```

Problem 2)

Consider a respondent with the following characteristics:

Variable	Variable Name	Actual value
X_1	PSA level	21.3 (mg/ml)
X_2	Cancer volume	8.4 (cc)
X_3	Weight	48.4 (gm)
X_4	Age	68 (years)
X_5	Benign prostatic hyperplasia	4.7 (cm2)
X_6	Seminal vesicle invasion	0
X_7	Capsular penetration	3.2 (cm)

Estimate the probability that this respondent is among the high-grade cancer group. Based on this estimated probability, do you believe that a respondent with these characteristics belongs to the high-grade cancer group? The **R** code is explicitly given below.

```
## Solution goes here
predict(model, data.frame(X1 = 21.3, X2 = 8.4, X3 = 48.4, X4 = 68,
                          X5 = 4.7, X6 = 0, X7 = 3.2), type = 'response')
```

```
##         1
## 0.2720329
```

```
# From the result, we do not believe that the respondent belongs to the high-grade cancer group.
```

Part 1: Maximum Likelihood Estimation and Newton's Method

In Model (1), the response values represent high-grade cancer and the features $X_1, X_2, \ldots, X_7$ are outlined in the data description from earlier in this document. To estimate Model (1), we use the method of *Maximum Likelihood.* The objective function of interest (log-likelihood) is

$$\begin{aligned}\ell(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7) = &\sum_{i=1}^{n} y_i\big(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+\beta_5X_{5i}+\beta_6X_{6i}+\beta_7X_{7i}\big)\\ &-\sum_{i=1}^{n}\log\big(1+\exp(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+\beta_5X_{5i}+\beta_6X_{6i}+\beta_7X_{7i})\big)\end{aligned}$$

The above log-likelihood is the same function derived in class except we have several more parameters to estimate, i.e., $\beta_0, \beta_1, \ldots, \beta_7$. In class we only considered *simple logistic regression* (one feature).

Problem 3)

Create a function in **R** called **logistic.NLL** with inputs **beta** and **data**, where **beta** is a vector of β coefficients and **data** is a dataframe defaulted by **data=prostate**. The function **logistic.NLL** should output the negative of the log-likelihood, i.e.,

$$-\ell(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7).$$

Recall that maximizing the log-likelihood is equivalent to minimizing the negative log-likelihood. Also evaluate the function using the vector **beta=rep(0,8)**, i.e., run the code **logistic.NLL(beta=rep(0,8))**.

```
## Solution goes here
logistic.NLL <- function(beta, data = prostate){
  return(-(sum(data$Y * (beta[1] + beta[2]*data$X1 + beta[3]*data$X2 + beta[4]*data$X3 +
                         beta[5]*data$X4 + beta[6]*data$X5 + beta[7]*data$X6 + beta[8]*data$X7)) -
           sum(log(1 + exp(beta[1] + beta[2]*data$X1 + beta[3]*data$X2 + beta[4]*data$X3 +
                             beta[5]*data$X4 + beta[6]*data$X5 + beta[7]*data$X6 + beta[8]*data$X7)))))
}
logistic.NLL(beta = rep(0, 8))
```

```
## [1] 67.23528
```

Problem 4)

Write a **R** function called **Newtons.Method** that performs *Newton's Optimization Method* on a generic function **f**. The function should have inputs **f**, **x0**, **max.iter**, **stopping.deriv** and **...** . The input **f** is the generic function we wish to minimize, **x0** is the starting point in the Newton's Method algorithm, **max.iter** is the maximum number of iterations (defaulted at 200), **stopping.deriv** is the gradient's threshold at which the algorithm terminates (defaulted at 0.001) and **...** allows you to pass additional arguments, based on **f**, into the **Newtons.Method** function. The output of **Newtons.Method** should be a list giving all updates of our minimizer and the number of iterations required to perform the procedure. You are welcome to add additional outputs if you would like. Hint: just copy and paste the **grad.descent** function from class and change the update step.

```
## Solution goes here
library(numDeriv)
library(MASS)
Newtons.Method <- function(f, x0, max.iter = 200, stopping.deriv = 0.001, ...){
  n <- length(x0)
  xmat <- matrix(0, n, max.iter)
  xmat[, 1] <- x0

  for (i in 2:max.iter){
    gradient <- grad(f, xmat[, i - 1])
    hess <- hessian(f, xmat[, i - 1])

    if(all(abs(gradient) < stopping.deriv)){
      i <- i - 1
      break
    }
    xmat[, i] <- xmat[, i-1] - ginv(hess) %*% gradient
  }
  xmat <- xmat[, 1:i]
  return(list(x = xmat[, i], xmat = xmat, i = i))
}
```

Problem 5)

Run the function **Newtons.Method** to minimize the function **logistic.NLL** using initial value **x0=rep(0,8)**, maximum iterations **max.iter=200** and **stopping.deriv=.001**. Display the estimated parameters for $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$. How many iterations did the algorithm take to converge?

```
## Solution goes here
Newtons.Method(logistic.NLL, x0 = rep(0, 8))
```

```
## $x
## [1] -10.276609180   0.055227995   0.086868612   0.001614891   0.106632573
## [6]  -0.071827473  -1.154137141   0.123324153
##
## $xmat
##      [,1]         [,2]         [,3]         [,4]          [,5]
## [1,]    0 -3.941963881 -7.217539800 -9.591222110 -10.200491143
## [2,]    0  0.009541978  0.021695042  0.040141203   0.053379730
## [3,]    0  0.076053924  0.100434682  0.096543310   0.087617222
## [4,]    0  0.001019644  0.001662889  0.001722042   0.001602292
## [5,]    0  0.031815490  0.070238188  0.099821580   0.105979354
## [6,]    0 -0.034591871 -0.053208836 -0.062939591  -0.070151466
## [7,]    0 -0.074343552 -0.470163351 -0.879590818  -1.123077802
## [8,]    0  0.033269981  0.076085713  0.102292971   0.120950183
##                [,6]          [,7]
## [1,] -10.275165030 -10.276609180
## [2,]   0.055204314   0.055227995
## [3,]   0.086867172   0.086868612
## [4,]   0.001613849   0.001614891
## [5,]   0.106618955   0.106632573
## [6,]  -0.071796030  -0.071827473
## [7,]  -1.153724670  -1.154137141
## [8,]   0.123299046   0.123324153
##
## $i
## [1] 7
```

Problem 6)

Use the base **R** function **nlm()** to minimize the negative log-likelihood using initial value **x0=rep(0,8)**. Display the estimated parameters for $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$. How many iterations did the algorithm take to converge?

```
## Solution goes here
logistic.nlm <- nlm(logistic.NLL,p=rep(0,8),data=prostate)
logistic.nlm$estimate
```

```
## [1] -10.276279062   0.055227384   0.086867770   0.001614479   0.106627782
## [6]  -0.071821442  -1.154123414   0.123325485
```

```
logistic.nlm$iterations
```

```
## [1] 82
```

Problem 7)

Check that the parameter estimates from the logistic model are reasonably close to the estimates coming from the **glm()** function.

```
model$coefficients
```

```
##   (Intercept)            X1            X2            X3            X4
## -10.276609553   0.055228000   0.086868614   0.001614891   0.106632576
##            X5            X6            X7
##  -0.071827481  -1.154137232   0.123324157
```

```
logistic.nlm$estimate
```

```
## [1] -10.276279062   0.055227384   0.086867770   0.001614479   0.106627782
## [6]  -0.071821442  -1.154123414   0.123325485
```

```
Newtons.Method(logistic.NLL, x0 = rep(0, 8))$x
```

```
## [1] -10.276609180   0.055227995   0.086868612   0.001614891   0.106632573
## [6]  -0.071827473  -1.154137141   0.123324153
```

```
# As we can see, the parameter estimates are reasonably close.
```