Assignment on Logistic Regression

January 23, 2022

Do the following parts below. Data Analyses 2.1 (25pts), 2.3 (25pts), 2.4(35pts), Simulate and Analyze 3.2 (15pts)

Bonus 15% : Part 1 Problems

1 Problems

1. (a) Write the log likelihood for the logistic regression model

   $\text{logit}(\Pr[Y|X_1 = x_1, X_2 = x_2]) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

   (b) Differentiate with respect to β_0.

   (c) Let f_i be the linear combination $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ and $p_i = \exp[f_i]/(1 + \exp[f_i])$. Interpret p_i.

   (d) At the maximum likelihoood estimate the derivative above equals zero. Equate the derivative to zero and write in terms of p_i. What does the sum $\sum_{i=1}^{n} p_i$ equal, and what does the mean $\sum_{i=1}^{n} p_i/n$ equal ?

   (e) How would you describe $\sum(y_i - p_i)^2/n$?

2 Data Analyses

2.1 Analysis of Burn Data

1. Install and utilize the R library *aplore3*. Using the dataset *burn1000* develop a model for predicting death.

2. Report the C-index.

3. Is the effect of *inh_inj* on mortality modified by age?

4. Is the effect of age on mortality modified by *inh_inj*?

2.2 Analysis of University Admissions Data

Read in the data using the R code read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv"). The dependent variable is admit.

1. Report the univariable associations of GPA, GRE and Rank with Admit.

2. Develop a multivariable model.

3. Interpret the coefficients in your model.

4. Do any quadratic terms, or interactions add predictive ability?

5. Report the C-statistic.

2.3 Data With a Zero Cell

Create the following dataset consisting of a dependent variable, Success, and two co-variates, Treatment and Female, using the following 3 lines of code:

```
Treatment = rep(c(0,1,0,1),each=10)
Female = rep(c(0,1),each=20)
Success = rep(rep(0:1,4), times=c(8,2,5,5,5,5,0,10))
```

1. Calculate the success frequency for the 4 combinations of Treatment and Gender.

2. Estimate the odds ratio relating Success to Treatment.

3. Estimate the odds ratio relating Success to Gender.

4. Include in a logistic regression the interaction of Treatment and Gender and comment on its statistical significance and coefficient.

2.4 Concussion Data

Run the following code to read in and restructure a dataset that recorded concussions in college sports according to sex of athlete, sport and year. The columns in the matrix (data.frame) named Y are the number of athletes with and without concussions respectively.

```
DF <- read.delim("http://users.stat.ufl.edu/~winner/data/concussion.dat", sep="", h
names(DF) <- c("Sex","Sport","Year","Concussion","Count")
DF0 <- DF[DF$Concussion==0,]
DF1 <- DF[DF$Concussion==1,]
Cov <- data.frame(DF0[,1:3])
Y <- cbind(CountConc=DF1[,5], CountNoConc=DF0[,5])
```

1. Derive the contingency table of concussion by sex.
2. Calculate risk (frequency) of concussions by sex, and the risk ratio comparing males to females.
3. Apply Pearson's chi-square test to the contingency table.
4. Use logistic regression to test if concussions are equally likely between males and femakles.
5. Repeat the steps above substituting the variables sports for sex.
6. Run a multivariable logistic regression of concusions by sex, sports and year.
7. Report the adjusted odds ratios for sex and sports.
8. Test if there is an interaction of sex and sports.

3 Simulate and Analyze

1. Run the code below. Then try different arguments to the function, f, e.g. try f(N0=30,N1=30, mu0=0, mu1=0.5). What is this code illustrating?

```
f <- function(R=500, N0=30, N1=30, mu0=0, mu1=0, sd0=1, sd1=0) {
  ptt <- plinear <- plogistic <- rep(NA, R)
  for (r in 1:R) {
    Y0 <- rnorm(n=N0, mean=mu0, sd=sd0)
    Y1 <- rnorm(n=N1, mean=mu1, sd=sd1)
    ptt[r] <-  t.test(Y0, Y1)$p.value
    Y <- c(Y0, Y1)
    X <- rep(0:1, times=c(N0,N1))
    plinear[r] <- summary(lm(Y ~ X))$coef["X",4]
```

```
    plogistic[r] <- summary(glm(X ~ Y, family=binomial))$coef["Y",4]
    }
  par(mfrow=c(1,2))
  plot(plogistic, ptt)
  plot(plogistic, plinear)
  print(table(plogistic < 0.05, ptt <0.05))
  print(summary(plogistic - ptt))
  print(table(plogistic < 0.05, plinear <0.05))
  print(summary(plogistic - plinear))
  }
f()
```

2. Explain why the estimate of the coefficient for X in the logistic regression adjusting for covariate Z1 (see below) is significantly different from zero despite the causal effect being zero?

```
n = 2500
Z1 = rnorm(n)
Z2 = rnorm(n)
X = 0.7*rnorm(n) + 0.7*Z2
Lin = 0*X - 0.0*Z1 + 0.5*Z2 # causal model
Y = runif(n) < 1/(1+exp(-Lin))
summary(glm(Y ~ X + Z1, family=binomial))
```

3. (a) What does the following simulated data and analysis indicate about *probit* regression? (b) Comment on the simlarities and differences the probit and logistic regressions, such as the Z values for the three covariates in the model.

```
beta <- c(1, -1, +2)
cutoff <- 0.5
n <- 10^4
X <- cbind(runif(n) < 0.25, runif(n) < 0.50, rnorm(n))
Y <- X %*% beta + rnorm(n)
binary.Y <- ifelse(Y < cutoff, 1, 0)
summary(X)
summary(glm(binary.Y ~ X, family=binomial(probit))
summary(glm(binary.Y ~ X, family=binomial) # for comparison with probit
```