Statistical Inference Projects

Project - 1. (Basketball Analytics using Statistical Inference)

The GitHub repository NBA Data contains the results of all 1230 NBA games from the 2015-2016 regular season. The 30 teams are encoded numerically from 1 to 30; the key for this encoding is provided in the file *NBA_teams.txt*. Each row of *NBA_record.csv* indicates the home team, away team, and outcome Y for one game, where $Y = 1$ if the home team won and $Y = 0$ otherwise. For parts (a) and (b), you may *not* use an existing software implementation of the Bradley-Terry or logistic regression model; however, you may use any generic optimization or equation-solving routine (or you may implement the Newton-Raphson iterations yourself, if you are brave).

(a) Fit the Bradley-Terry model, with an intercept term α for the home-court advantage, to this data set. What are the 8 teams (in ranked order) with the highest Bradley-Terry scores? How much greater are the log-odds of winning for the home team than for the away team?

One approach to do this in R is to use the generic optimization function optim. To do this, first define a function

```
loglik = function(theta,Home,Away,Y) {
         ...
   }
```

that returns the log-likelihood for the Bradley-Terry model given inputs $\theta = (\alpha, \beta_2, \ldots, \beta_k)$ (constraining $\beta_1 \equiv 0$), Home $= (i_1, \ldots, i_n)$, Away $= (j_1, \ldots, j_n)$, and Y $= (Y_1, \ldots, Y_n)$, where i_m and j_m are the home and away teams for game m. To then read the data file and maximize the log-likelihood:

```
table = read.csv("NBA_record.csv")
result = optim(theta0,loglik,Home=table$Home,Away=table$Away,Y=table$Y,
               method="BFGS",control=list("fnscale"=-1))
```

Here theta0 is any initialization for θ (for example the all 0's vector). The method will use the BFGS algorithm, and "fnscale" $= -1$ indicates that it should perform maximization rather than minimization.

(b) Fit the Bradley-Terry model without an intercept term. (You may do this in R by defining a new function

```
loglik_noalpha = function(theta,Home,Away,Y)
```

where now $\theta = (\beta_2, \ldots, \beta_k)$, and using optim as before.) Evaluate the log-likelihoods at the full model and sub-model MLEs, and conduct a generalized likelihood ratio test of the null hypothesis of no home court advantage, $H_0 : \alpha = 0$. What is the p-value that you obtain for your test?

(c) For the m^{th} game, suppose we define 30 covariates $x_{m,1}, \ldots, x_{m,30}$ in the following way: Let $x_{m,1} = 1$ always. Let $x_{m,i} = 1$ if team i is the home team of this game and $i \neq 1$, and let $x_{m,j} = -1$ if team j is the away team of this game and $j \neq 1$. Let $x_{m,k} = 0$ for all other k. Explain why logistic regression for Y_m using the covariates $x_{m,1}, \ldots, x_{m,30}$ is equivalent to the Bradley-Terry model, where we constrain the Bradley-Terry score of team 1 to be $\beta_1 \equiv 0$. If we were to run this logistic regression, what would be the interpretation of the fitted coefficient for the first covariate $x_{m,1}$? For the 10th covariate $x_{m,10}$?

(d) Fit the logistic regression in part (c) using any standard regression software, and verify that the fitted coefficients match (up to reasonable numerical accuracy) your estimated parameters from part (a).

To do this in R, you may construct a matrix X of size 1230×30 containing the covariates as defined in part (c), and then fit the regression using

```
model = glm.fit(X,table$Y,family=binomial())
coefs = model$coefficients
```

Important References:

1. https://www.r-bloggers.com/2022/02/what-is-the-bradley-terry-model/
2. https://squared2020.com/2017/11/09/bradley-terry-rankings-introduction-to-logistic-regression/
3. https://encyclopediaofmath.org/wiki/Bradley-Terry_model

Project - 2. (Analysis of the NASA Challenger Disaster Data)

The NASA space shuttle "Challenger" exploded shortly after its launch on 28 January 1986, with a loss of seven lives. The US Presidential Commission concluded that the accident was caused by a leakage of gas from one of the fuel tanks. Rubber insulating rings, so-called "O-rings", were not pliable enough after the overnight low temperature of 31°F, and did not plug the joint between the fuel in the tanks and the intense heat outside. Table 1 presents the data concerning previous flights which has been slightly modified, for illustration purposes, by only reporting the presence of failure of the O-rings. The last row corresponds to the conditions at which the Challenger was launched.

The description of this data can be found at: Challenger Explosion Data. A brief copy-paste description is as follows:

> The NASA space shuttle Challenger exploded on January 28, 1986, just 73 seconds after liftoff, bringing a devastating end to the spacecraft's 10th mission. The disaster claimed the lives of all seven astronauts aboard, including Christa McAuliffe, a teacher from New Hampshire who would have been the first civilian in space. It was later determined that two rubber O-rings, which had been designed to separate the sections of the rocket booster, had failed due to cold temperatures on the morning of the launch. The tragedy and its aftermath received extensive media coverage and prompted NASA to temporarily suspend all shuttle missions.

	Failure	Temperature	Pressure (psi)
1	0	66	50
2	1	70	50
3	0	69	50
4	0	68	50
5	0	67	50
6	0	72	50
7	0	73	100
8	0	70	100
9	1	57	200
10	1	63	200
11	1	70	200
12	0	78	200
13	0	67	200
14	1	53	200
15	0	67	200
16	0	75	200
17	0	70	200
18	0	81	200
19	0	76	200
20	0	79	200
21	1	75	200
22	0	76	200
23	1	58	200
C	-	31	200

Table 1: Challenger data. Failures out of 6 rings.

(a) Fit a logistic regression model to the data and find the values of the coefficients using MLE.

(b) Find out the Confidence Intervals (CIs) for each of the parameters of the fitted model.

(c) Find the probability of failure of the O-rings under the conditions of that day $\mathrm{x}_C = (1, 31, 200)^\top$.

(d) Interpret the results.

Project - 3. (Analysis of the Dosage Mortality Curve)

An experiment was conducted to evaluate the toxicity of gaseous carbon disulphide on flour beetle. Table 2.1 shows the numbers of beetles n_i that were exposed to gaseous carbon disulphide at concentrations x_i (Dose, expressed in $\log_{10}$ mg/L), as well as the numbers of beetles y_i dead after five hours of exposure. The data comes from an article by Bliss (1935), Thus, the responses y_i are binomial realizations with n_i trials at concentrations x_i. This kind of experiment can be modeled using a logistic regression model.

Dose x_i	n_i	y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Table 2: Beetle mortality data.

This real data example illustrates the method of maximum likelihood estimation for the parameters of the logistic regression model. Write R code to compare the results using the following methods:

(a) The Newton's method.

(b) A direct implementation. Implementing the log-likelihood function and maximizing it using the command optim().

(c) Using the command glm().

Also, as part of the project, provide a visualization of the fitted dose-response model.