Bradly Terry Test

1. Stigler (1994 Statistical Sciences) presents the data in http://stat.rutgers.edu/home/kolassa/stigler.dat reflecting 16843 journal citations for eight statistical journals, the Annals of Statistics (AnnStat), Biometrics (Biocs), Biometrika (Bioka), Communications in Statistics (ComSt), the Journal of the American Statistical Association (JASA), the Journal of the Royal Statistical Society Series B (JRSSB), the Journal of the Royal Statistical Society Series C (JRSSC), and Technometrics (Tech). The first line of this file gives column headings. The remaining lines give the number of times that articles in the citer journal cite articles in the cited journal. For example, the line Biocs,AnnSt,42 indicates that 42 times involve the Annals of Statistics citing Biometrics. More important articles will be cited more, and better journals will have more important citations, and so each citation incidence represents a victory of the cited journal over the citing journal. Fit the Bradley–Terry model, and determine the most prestigious journal. Keep in mind that you need to indicate to proc logistic or glm that each line in the input file represents multiple observations. There is an assumption of the model that is violated by the fact that multiple citations come from a single article. Explain why this is an issue.

The following code reads in the data:

```r
stig<-read.table("stigler.dat", head=T, sep=" ",
```

The data set stig now has 64 lines. Eight of these represent a journal citing itself, which we will ignore. Build a design matrix with 56=64-8 rows and 8 columns, with each column representing a journal. Also, create a new response variable \( y \), containing those elements of the third column of the data set not representing self-citations.

```r
x<-array(0,c(56,8))
y<-rep(NA,56)
dimnames(x)<-list(NULL,
  c("AnnSt","Biocs","Bioka","ComSt","JASA","JRSSB",
    "JRSSC","Tech"))
```

Now copy responses into the correct row of \( y \), and write the correct covariates for \( x \). Do this by cycling through all 64 lines of the data set, and each time you hit a line that is not a self citation (which is most lines), write to the design matrix and response vector. Note that the line of the data set reduced by removing self-citations will be different from the line in the original data set; index the original line by \( j \), and the line in the reduced data set by count.
```r
count<-0
for(j in 1:64){
  if(stig[j,1]!=stig[j,2]){  
    count<-count+1
    x[count,stig[j,1]]<-1
    x[count,stig[j,2]]<-1
    y[count]<-stig[j,3]  
  }
}
```

Now fit the model. Fit a logistic regression. I used the input format with the response variable a matrix with two columns, representing numbers of successes and failures. The data are formatted in such a way that every line represents a success; failures for the line are reflected as successes on the line with covariates negated. For example, the first row of \( x \) is \((-1,1,0,0,0,0,0,0,0,0,0)\), and the first row of the response matrix is \((42,0)\). These represent the 42 times Biometrics cites the Annals of Statistics. The eighth row of \( x \) is \((1,-1,0,0,0,0,0,0,0,0,0)\), and the eighth row of the response matrix is \((155,0)\). These represent the 155 times Annals of Statistics cites Biometrics. These two lines could have been combined as a row of \( x \) with \((-1,1,0,0,0,0,0,0,0,0,0)\), and a row of the response matrix is \((42,155)\). This would have been more concise, but not necessary. Fit the model using

```r
glm(cbind(y,0)~x-1,family="binomial")
```

Note that the intercept is suppressed here. Observe the results:

```
Call: glm(formula = cbind(y, 0) ~ x - 1, family = "binomial")
Coefficients:
xAnnSt xBiocs xBioka xComSt xJASA xJRSSB xJRSSC xTech
 0.9807 -0.2109  0.6271 -2.2920  0.1680  0.9173 -0.3209 NA
Degrees of Freedom: 56 Total (i.e. Null); 49 Residual
Null Deviance: 15880
Residual Deviance: 11530  AIC: 11540
```

Hence the Annals of Statistics is most prestigious. Since multiple citations come from the same article, one might argue that these are not independent.

2. Matched pairs via conditional logistic regression

   Horan, Francis, Falsey, Kolassa, Smith, and Hall (2001) present results of a case–control study explaining the respiratory tract infections as a function of a variety of variables. One of these variables is fibrinogen level, a chemical found in the blood. The data set at http://stat.rutgers.edu/home/kolassa/Data/fibrinogen.dat contains a pair number, the change in a variable fibrinogen, whether the individual is a case or control, and whether the individual is from an older (1) or younger (0) group. Case–control pairs were found by finding two individuals, one with infection and one without, who had similar ages and genders.

   a. A few values of the pairing variable are only represented once in the data set. These represent individuals with no match. I really shouldn’t call these pairs, but use the more
generic word strata. What impact to these individuals have on a conditional logistic regression analysis accounting for strata?

*Strata with a single member contribute nothing to the analysis. The conditional likelihood is the product of contributions from each of the strata, and each contribution is the product of the exponential of the product of the response times the covariate vector times the parameter, divided by the sum of these over all rearrangements of the observations, or in symbols, \[ \prod_j \exp(\sum_k x_{jk} \beta \theta_{jk}) / \prod \sum \exp(\sum_k x_{jk} \beta \theta_{jk}) \] where \( j \) indexes stratum, \( k \) represents individual within stratum, and \( \sum \) represents summation over all rearrangements of the \( y_{j1}, y_{j2}, \ldots \). When \( k = 1 \), then there is only one arrangement, and the numerator and denominator in that factor are the same, and cancel.*

b. Estimate the effect of age group on disease status, accounting for pair. Use conditional logistic regression. Compare your answer to that obtained by the ratio of discordant pairs. I suggest doing this last part by hand.

*Note that all pairs are concordant, and hence the estimator is \( 0/0 \), or undefined.*

c. Now estimate the effect of a change in fibrinogen on disease status, accounting for matching.

*Here are the SAS commands to do the job:*

```sas
data fib; infile 'fibrinogen.dat';
  input st fd caco $ a;
  cacon=1; if caco='control' then cacon=0;
run;
```

*Here’s an older way to do the exact logistic regression:*

```sas
proc phreg data=fib; strata st;
  model cacon*cacon(1)=fd/ties=exact; run;
```

*Here’s a newer way:*

```sas
proc logistic data=fib; strata st; model cacon=fd; run;
```

*The following code will do this in R:*

```r
fib<-read.table("fibrinogen.dat",stringsAsFactors=F)
names(fib)<-c("pair","fd","casecontrol","agegp")
fib$fd<-'as.numeric(fib$fd)
fib$cc<-(fib$casecontrol=='case')+0
library(survival)
clogit(cc~fd+strata(pair),data=fib)
```

*Here are the results:*

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Standard</th>
<th>Hazard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>DF</td>
<td>Estimate</td>
</tr>
<tr>
<td>fd</td>
<td>1</td>
<td>-0.00754</td>
</tr>
</tbody>
</table>

*Hence as the change in fibrinogen increases, the chance of being a case increases, and for each one unit increase in fibrinogen change, the log odds ratio changes by -0.00754.*