STAT 510 Homework 13

Ungraded

1. The solution to this problem was written by a former STAT 510 TA Chuanlong Du.

   First let's decide an appropriate model for this dataset:

```
#create a factor from age
donner$year=as.factor(donner$age)
#fit a model which treat age as factor
glm(status~year*sex,family=binomial(link=logit),data=donner)->glmout1
#fit a model which treat age as a quantity variable
glm(status~age*sex,family=binomial(link=logit),data=donner)->glmout2
#fit a additive model which treat age a quantity variable
glm(status~age+sex,family=binomial(link=logit),data=donner)->glmout3
#compare model glmout1 and glmout2
anova(glmout1,glmout2,test="Chisq")
Analysis of Deviance Table

Model 1: status ~ year * sex
Model 2: status ~ age * sex
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1        17     24.815
2        41     47.346 -24  -22.532    0.5476

#compare model glmout2 and glmout3
anova(glmout2,glmout3,test="Chisq")
Analysis of Deviance Table

Model 1: status ~ age * sex
Model 2: status ~ age + sex
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        41     47.346
2        42     51.256 -1  -3.9099     0.048 *
---
Signif. codes:  0 ł**0.001 ł*0.01 ł0.05 0.1 1
```

   From the above result, we can see that there's no significant difference between the first and the second model but a modestly significant difference between the second and the third model. Let's look at the AIC's of the three models:

```
> AIC(glmout1,glmout2,glmout3)
        df      AIC
glmout1 28 80.81486
glmout2  4 55.34637
glmout3  3 57.25628
```

   The above results show us that the AIC of the second model is the smallest. So based on the above analysis, the second model is preferred. Let's see what this model tells us about the relationship between the survival probability and age and sex.

```
#summary information about model glmout2
summary(glmout2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.24638    3.20517   2.261   0.0238 *
age         -0.19407    0.08742  -2.220   0.0264 *
sexMALE     -6.92805    3.39887  -2.038   0.0415 *
age:sexMALE  0.16160    0.09426   1.714   0.0865 .
```

From the above results we can see that generally speaking women had a survival advantage over men. Age also had a significant association with survival probability. With age increasing, the expected survival probability decreased both for female (the coefficient corresponding to age, i.e. $\hat{\beta}_2 = -0.194 < 0$) and for male ($\hat{\beta}_2 + \hat{\beta}_4 = -0.0325 < 0$). However, the expected survival probability for females decreased faster than for males ($\hat{\beta}_2 = -0.194 < -0.0325 = \hat{\beta}_2 + \hat{\beta}_4$), which can also be seen from Figure 1. Figure 1 also tells us that for people older than about 45, men had an estimated survival advantage over women.

Let's take a further step to do some quantitative analysis. I wrote a function $orint$ to calculate the odds ratios of female to male at different ages, confidence intervals for the odds ratios and also the p-values for tests of "$H_0 : \ odds\ ratio = 1$" at these ages. The code of function $orint$ is as follows:

```
orint=function(glmout,age,alpha=0.05)
{#calculate approximate confidence intervals for odds rate
#of the expected survival probability of woman to man at different ages
#+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
  cdiff=-c(0,0,1,age)
  temp=cdiff%*%coef(glmout)
  temp.var=sqrt(cdiff%*%vcov(glmout)%*%cdiff)
  oddsratio=exp(temp)
  orint=oddsratio*exp(c(-1,1)*temp.var*qnorm(1-alpha/2))
  pvalue=pnorm(abs(temp/temp.var),lower.tail=FALSE)*2
  list(oddsratio=oddsratio,orint=orint,pvalue)
}
```

The results of applying the function to model $glmout2$ at different ages are given below:

```
> orint(glmout2,20)
$oddsratio
         [,1]
[1,] 40.29036
$ci
[1]    1.543675 1051.590136
$pvalue
          [,1]
[1,] 0.0263621
```

The above results tells us that the odds ratio of female at age 20 to male at age 20 is about 40 (pretty big) and is significant different from 1, i.e., the odds of survival for 20-year-old females were estimated to be about 40 times greater than the odds of survival for 20-year-old males.

```
> orint(glmout2,30)
$oddsratio
         [,1]
[1,] 8.005618
$ci
[1]  1.080386 59.321322
$pvalue
           [,1]
[1,] 0.04178775
```

The above results tells us that the odds ratio of female at age 30 to male at age 30 is about 8 (kind of big) and is still significantly different from 1, i.e., the odds of survival for 30-year-old females were estimated to be about 8 times greater than the odds of survival for 30-year-old males.

```
> orint(glmout2,45)
$oddsratio
          [,1]
[1,] 0.7090641
$ci
[1] 0.05159103 9.74533511
$pvalue
          [,1]
[1,] 0.7970726
```

The above results tells us that the odds ratio of female at age 45 to male at age 45 is about 0.71 (close to 1) but is not significantly different from 1, i.e. at age 45 neither female nor male had a significant advantage to survive over the other.

```
> orint(glmout2,50)
$oddsratio
          [,1]
[1,] 0.3160693
$ci
[1] 0.01107195 9.02278308
$pvalue
          [,1]
[1,] 0.5005899
```

The above results tells us that the odds ratio of female at age 50 to male at age 50 is about 0.32 (smaller than 1) but is not significantly different from 1, i.e. at age 50 neither female nor male has significant advantage to survive over the other.

```
> orint(glmout2,70)
$oddsratio
           [,1]
[1,] 0.01247871
$ci
[1] 1.439862e-05 1.081481e+01
```

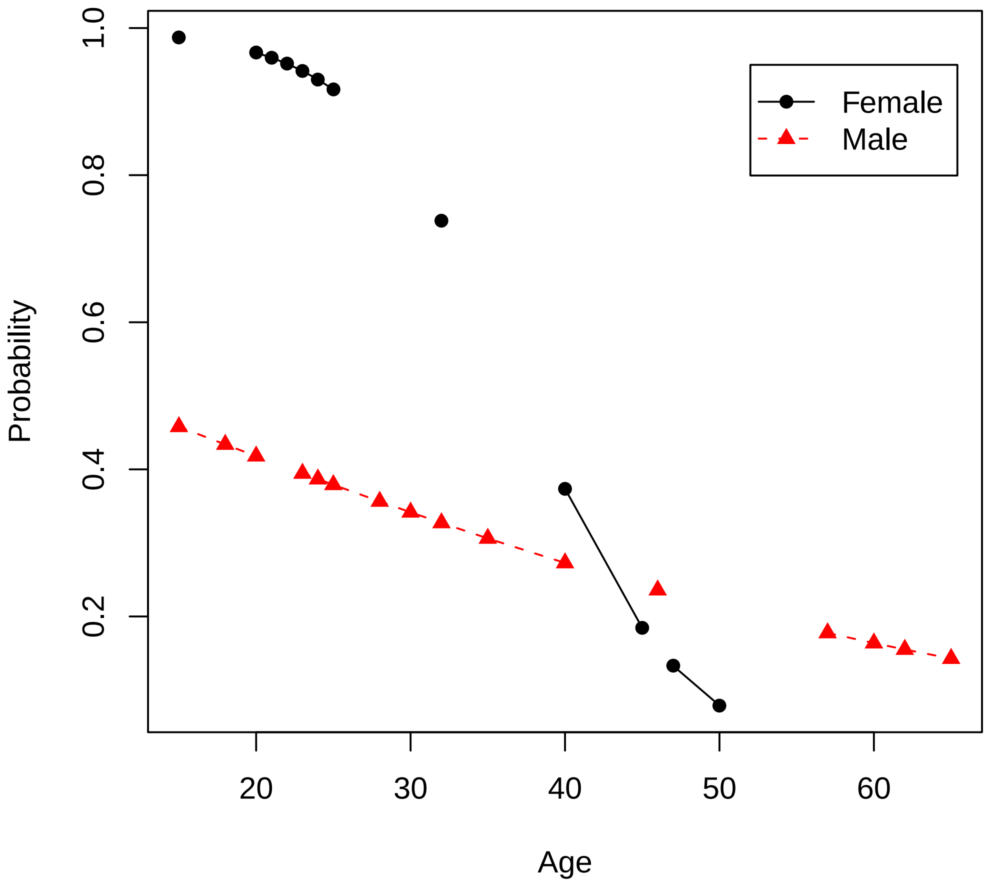
```
$pvalue
          [,1]
[1,] 0.2040397

> orint(glmout2,100)
$oddsratio
             [,1]
[1,] 9.789287e-05
$ci
[1] 4.942524e-10 1.938891e+01
$pvalue
          [,1]
[1,] 0.1379327

> orint(glmout2,150)
$oddsratio
             [,1]
[1,] 3.031934e-08
$ci
[1] 1.578567e-17 5.823397e+01
$pvalue
          [,1]
[1,] 0.1124465
```

The above results tells us that for really old people (with $age \geq 70$), the odds ratio for male is much higher than for female. However, it's still not significantly different from 1. Furthermore, the oldest people in the data set are males age 65 or less. The oldest female was only age 50. Thus, it would be dangerous to extrapolate out to ages like 70.

The following is the code for plotting Figure 1.

```
#calculate expect survival prob for each group
tapply(fitted(glmout2),list(donner$age,donner$sex),mean)->probmean
#profile plot of the expected survival prob
age.uni=sort(unique(donner$age))
matplot(age.uni,probmean,type='l',lty=c(1,2))
matpoints(age.uni,probmean, pch=c(16,17))
legend(52,0.95,legend=c("Female",'Male'),lty=c(1,2),col=1:2,pch=c(16,17))
```

Figure 1: Estimated Survival Probability



2.

```
> y=c(15,9,15,23,14,18,5,7,12,11)
>
> o=glm(y~1,family=poisson(link=log))
>
> 1-pchisq(deviance(o),df.residual(o))
[1] 0.01685265
>
> #The residual deviance statistic suggests
> #that there is significant lack of fit.
> #The p-value is 0.01685.
>
> #The Pearson statistic is
>
> P=sum((y-mean(y))^2/mean(y))
> P
[1] 19.75969
>
> 2*(1-pchisq(P,length(y)-1))
[1] 0.03890952
>
> #The Pearson statistic also suggests
> #that there is significant lack of fit.
>
> #We should conclude that the data are not
> #an independent and identically distributed
> #sample from one Poisson distribution.
```

3.

```
> y=c(39, 31, 43, 31, 34, 36, 34, 24,
+ 23, 28, 24, 19, 16, 20, 25, 12,
+ 36, 38, 33, 22, 23, 17, 29, 16)
>
> g=as.factor(rep(c("A","B","C"),each=8))
>
> o=glm(y~g,family=poisson(link=log))
>
> o

Call:  glm(formula = y ~ g, family = poisson(link = log))

Coefficients:
(Intercept)           gB           gC
     3.5264      -0.4878      -0.2398

Degrees of Freedom: 23 Total (i.e. Null);  21 Residual
Null Deviance:      61.02
Residual Deviance: 35.57        AIC: 163.9
>
> anova(o,test="Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    23     61.017
g     2   25.452        21     35.565 2.973e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> #The test above suggests that there are
> #significant differences among genotypes.
>
> #Before going further with analysis,
> #let’s check for overdispersion.
>
> 1-pchisq(deviance(o),df.residual(o))
[1] 0.02445231
>
> #The test suggests a lack of fit that
> #could be caused by over dispersion.
>
> #Let’s look at a residual plot to make sure
> #the lack of fit is not due to extreme outliers.
>
> plot(fitted(o),resid(o,type="deviance"))
>
> #No extreme outliers noted. Thus, it seems
> #reasonable to blame the lack of fit on
> #overdispersion.
>
> #Let’s estimate overdispersion parameter.
>
> phihat=deviance(o)/df.residual(o)
> phihat
[1] 1.693594
>
> #Let’s test again for a difference among
> #genotypes, but this time we will account
> #for overdispersion
>
> oq=glm(y˜g,family=quasipoisson(link=log))
>
```

```
> anova(oq,test="F")
Analysis of Deviance Table

Model: quasipoisson, link: log

Response: y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev      F   Pr(>F)
NULL                    23     61.017
g     2   25.452        21     35.565 7.7292 0.003051 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> #There is significant evidence of differences
> #among genotypes.
>
> #Let’s compare pairs of genotypes.
>
> v=vcov(oq)
> b=coef(oq)
>
> C=matrix(c(
+ 0,1,0,
+ 0,0,1,
+ 0,1,-1),byrow=T,nrow=3)
>
> Cb=C%*%b
> se=sqrt(diag(C%*%v%*%t(C)))
> tt=drop(Cb/se)
> 2*(1-pt(abs(tt),df.residual(o)))
[1] 0.0008923671 0.0535405334 0.0752392327
>
> #Based on the p-values above, all pairwise
> #comparisons are significant at the .10 level.
> #Only A vs. B is significant at the .05 level.
>
> coef(oq)
(Intercept)          gB          gC
  3.5263605  -0.4878083  -0.2398261
>
> #Genotype A seems significantly more susceptible
> #then genotype B.
>
> #Now let’s address overdispersion by fitting a
> #GLMM that allows for overdispersion in the data.
```

```
>
> library(lme4)
Loading required package: lattice
Loading required package: Matrix
Warning message:
package lme4 was built under R version 2.15.3
>
> leaf=factor(1:24)
> oglmm=glmer(y˜g+(1|leaf),family=poisson(link="log"))
> oglmmreduced=glmer(y˜1+(1|leaf),family=poisson(link="log"))
> anova(oglmmreduced,oglmm)
Data:
Models:
oglmmreduced: y ˜ 1 + (1 | leaf)
oglmm: y ˜ g + (1 | leaf)
             Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
oglmmreduced  2 173.52 175.88 -84.761   169.52
oglmm         4 164.26 168.97 -78.129   156.26 13.264      2   0.001318 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> #From the above, we see that the likelihood ratio test
> #statistic for comparing the null model with only an
> #intercept parameter and a leaf variance component
> #to the alternative model with one parameter for each
> #genotype and a leaf variance component is 13.264.
> #Comparing to a chi-square distribution with 2 df
> #results in a p-value of 0.001318.
```

4. Because the response is binomial with m trials, the variance for each treatment group is as follows:

Treatment Group	Variance
A	$m \cdot 0.5 \cdot (1 - 0.5) = 0.25 \cdot m$
B	$m \cdot 0.5 \cdot (1 - 0.5) = 0.25 \cdot m$
C	$m \cdot 0.95 \cdot (1 - 0.95) = 0.0475 \cdot m$

Thus, the variance for each of treatment groups A and B is more than five times greater than the variance for treatment group C. The ANOVA approach assumes that the variance is the same for all treatment groups. The variance will be estimated by MSE. The MSE is obtained by pooling the variance estimates for each treatment group with weights proportional to the degrees of freedom for each treatment group. Because the degrees of freedom are 9, 9, and 49 for treatment groups A, B, and C, respectively; the MSE will be pulled strongly towards the variance for treatment group C. As a result, the MSE form the ANOVA analysis will tend to underestimate the variance for treatment groups A and B. This will lead to a standard error for the difference between treatment A and B means that is too small. This will cause the test statistic to be too far from zero and the p-value to be too small. The resulting analysis will have a much higher type I error rate than advertised.

Note that nonconstant variance is often a much bigger problem for analysis of variance than lack of normality. ANOVA often works quite well even with nonnormal data. However, nonconstnt variance can be a serious problem, especially when comparing a pair of treatments using a variance estimate obtained by pooling across many treatments that do not all have the same error variance.

5.

```
> d=read.delim(
+ "http://www.public.iastate.edu/~dnett/S510/PlaneCrashes.txt")
> d
   index crashes
1    376       8
2    347       5
3    322       8
4    104       4
5    103       6
6     98       4
7     96       8
8     85       6
9     82       4
10    63       2
11    44       7
12    40       4
13     5       3
14     5       2
15     0       4
16     0       3
17     0       2
>
> plot(d)
>
> o=glm(crashes~index,family=poisson(link=log),data=d)
>
> summary(o)

Call:
glm(formula = crashes ~ index, family = poisson(link = log),
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1974  -0.3978  -0.1766   0.3537   1.4919

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.3098588  0.1582327   8.278   <2e-16 ***
index       0.0019933  0.0008166   2.441   0.0146 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 15.295  on 16  degrees of freedom
Residual deviance:  9.794  on 15  degrees of freedom
AIC: 70.365

Number of Fisher Scoring iterations: 4

>
> anova(o,test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: crashes

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     16     15.295
index  1   5.5013        15      9.794    0.019 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> #It looks like there is significant evidence
> #of association between the news coverage index
> #and the number of crashes. This might be evidence
> #in favor of these sociologists’ theory.
>
> #Check for lack of fit.
>
> 1-pchisq(deviance(o),df.residual(o))
[1] 0.8324938
>
> #There is no evidence of lack of fit.
> #However, it’s not clear how good the
> #asymptotic chi-square approximation
> #will be in this case since n is low
> #and the counts are small.
>
> exp(100*coef(o)[2])
  index
1.22059
>
> #A 100 unit increase in news coverage index
> #is associated with an estimated 22% increase
> #in the mean number of crashes that occur in the
> #subsequent week.
```