Sample solutions

Stat 8051 Homework 6

Problem 1: ALR Exercise 12.5

12.5.1

```
> t = with(Donner, table(y,sex)); t
          sex
y          Female Male
  died         10   32
  survived     25   24
> print(fracs <- c(t[2,1]/sum(t[,1]),t[2,2]/sum(t[,2])))
[1] 0.7142857 0.4285714
> chisq.test(t,correct=F)

Pearson's Chi-squared test

data:  t
X-squared = 7.0748, df = 1, p-value = 0.007817
```

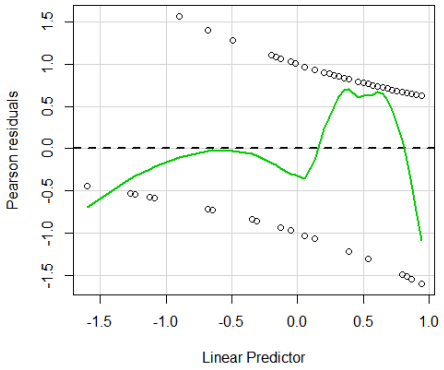
Thus the survival fractions for females and males are 0.714 and 0.429, respectively. The p-value obtained from uncorrected chi-squared test for homogeneity suggests strong evidence against the hypothesis that survival rates were same for females and males.

12.5.2

```
> summary(reg <- glm(y~age, family=binomial(),
+ data=Donner))

Call:
glm(formula = y ~ age, family = binomial(), data = Donner)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5946  -1.2017   0.8436   0.9882   1.5765

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.97917    0.37460   2.614  0.00895 **
```

```
age          -0.03689    0.01493  -2.471  0.01346 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.86  on 87  degrees of freedom
Residual deviance: 114.02  on 86  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 118.02

Number of Fisher Scoring iterations: 4

> exp(coef(reg))
(Intercept)         age
  2.6622535   0.9637838
```

The effect of Age is found significant. The sign of the coefficient suggests that survival probability decreases with increase in Age. Toe be more specific, the odds of survival is multiplied by a factor of about 0.964 for unit increase in Age.



12.5.3 The above residual plot is not very satisfactory, but the curve in the smoother does suggest the possibility that survival probability is overestimated for the older ages.

12.5.4 The modified model summary is given below:

```
> regnew = update(reg, ~.+I(age^2)+sex+status)
> summary(regnew)

Call:
glm(formula = y ~ age + I(age^2) + sex + status, family = binomial(),
    data = Donner)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0431  -1.0391   0.5120   0.8664   2.0797

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.986e-01  6.172e-01   0.322   0.7476
age           1.675e-01  7.107e-02   2.357   0.0184 *
I(age^2)     -3.889e-03  1.525e-03  -2.550   0.0108 *
sexMale      -6.637e-01  5.588e-01  -1.188   0.2349
statusHired  -1.625e+00  7.481e-01  -2.173   0.0298 *
statusSingle -1.852e+01  1.760e+03  -0.011   0.9916
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.855  on 87  degrees of freedom
Residual deviance:  92.363  on 82  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 104.36

Number of Fisher Scoring iterations: 16
```

There is significant quadratic effect due to age, as well as an effect of status. Note that after adjusting for age, there is no noticeable effect of sex on the outcome.

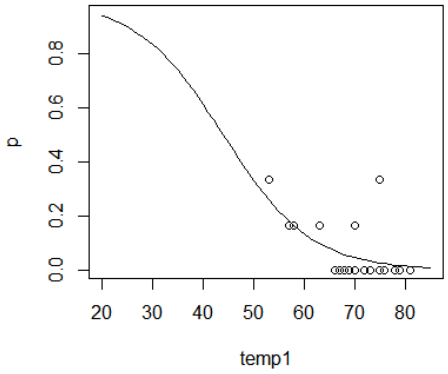
Problem 2: ALR Exercise 12.6

```
> summary(reg <-
+           glm(cbind(fail, n-fail)~temp,
+               family=binomial(),
+               data=Challeng))

Call:
glm(formula = cbind(fail, n - fail) ~ temp, family = binomial(),
    data = Challeng)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95227  -0.78299  -0.54117  -0.04379   2.65152

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
```

```
temp        -0.11560    0.04702  -2.458   0.0140 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 18.086  on 21  degrees of freedom
AIC: 35.647

Number of Fisher Scoring iterations: 5

>
> summary(reg1 <- glm(cbind(fail, n-fail)~temp+pres,
+                     family=binomial(),
+                     data=Challeng))

Call:
glm(formula = cbind(fail, n - fail) ~ temp + pres, family = binomial(),
    data = Challeng)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.05383  -0.65352  -0.56140  -0.03971   2.37171

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.409728   3.178539   1.073   0.2834
temp        -0.107747   0.044648  -2.413   0.0158 *
pres         0.007380   0.006447   1.145   0.2523
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 16.565  on 20  degrees of freedom
AIC: 36.125

Number of Fisher Scoring iterations: 5
```

None of the models do not seem to fit to well. This becomes apparent from the plot of values with the plot on only `Temp` below:

12.5.3




```
> predict(reg, data.frame(temp=31), type="response")
        1
0.8177744
```

The predicted probability comes out to be about 0.82, thus according to the model built after analyzing the data, there is a high risk of failure at this temperature. Although the reliability of this estimate is questionable since we do not have any observation near the new value where the prediction is being made.

Problem 3: ALR Exercise 12.7

```
> summary(m1 <- glm(cbind(surv, m - surv) ~
+                     class + age + sex,
+                   binomial, data=Whitestar))

Call:
glm(formula = cbind(surv, m - surv) ~ class + age + sex, family = binomial,
    data = Whitestar)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1356  -1.7126   0.7812   2.6800   4.3833

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1862     0.1586   7.480 7.40e-14 ***
classfirst    0.8577     0.1573   5.451 5.00e-08 ***
classsecond  -0.1604     0.1738  -0.923    0.356
classthird   -0.9201     0.1486  -6.192 5.93e-10 ***
agechild      1.0615     0.2440   4.350 1.36e-05 ***
sexmale      -2.4201     0.1404 -17.236  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

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

    Null deviance: 671.96  on 13  degrees of freedom
Residual deviance: 112.57  on  8  degrees of freedom
AIC: 171.19

Number of Fisher Scoring iterations: 5
```

12.7.1 From Table 12.8, nearly all females survived, except in third class, where female survival was much lower. This implies a `class`×`sex` interaction. Other interactions might exist as well.

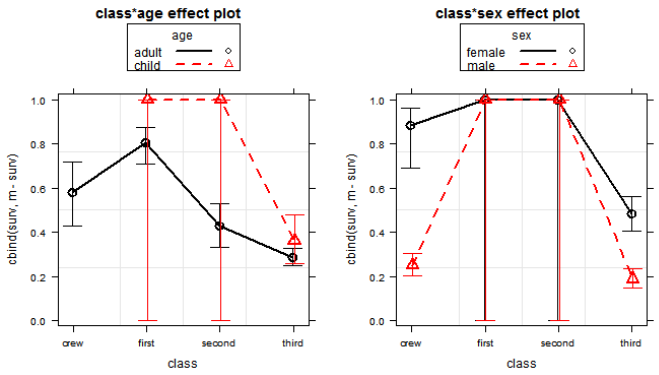
12.7.2

```
> m2 <- update(m1, ~(class + age + sex)^2)
> Anova(m2)
Analysis of Deviance Table (Type II tests)

Response: cbind(surv, m - surv)
          LR Chisq Df Pr(>Chisq)
class       120.73  3  < 2.2e-16 ***
age          20.34  1  6.486e-06 ***
sex         359.37  1  < 2.2e-16 ***
class:age    37.26  2  8.101e-09 ***
class:sex    65.01  3  4.984e-14 ***
age:sex       1.69  1     0.1942
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The `age`×`sex` interaction can apparently be dropped, but the other two interactions are required.

```
> m3 <- update(m2, ~ . - age:sex)
> plot(Effect(c("age", "sex"), m3),
+      rescale.axis=FALSE, grid=TRUE,
+      multiline=TRUE, ci.style="bars")
> plot(allEffects(m3),
+ rescale.axis=FALSE, grid=TRUE,
+ multiline=TRUE, ci.style="bars")
```

(plot on next page)


class*age effect plot
age
adult
child
cbind(surv, m - surv)
1.0
0.8
0.6
0.4
0.2
0.0
crew
first
second
third
class
class*sex effect plot
sex
female
male
cbind(surv, m - surv)
1.0
0.8
0.6
0.4
0.2
0.0
crew
first
second
third
class