MAST30027: Modern Applied Statistics

Week 3 Lab

1. The binomial random variable $Y \sim \text{bin}(m, p)$ for known m (not a parameter) has mass function:

$$f(y;p) = \binom{m}{y} p^y (1-p)^{m-y} \text{ for } y = 0, 1, \ldots, m.$$

   Show that the binomial distribution is an exponential family.

   Solution:

$$\begin{aligned} f(y) &= \binom{m}{y} p^y (1-p)^{m-y} \text{ for } y = 0, 1, \ldots, m \\ &= \exp\left[y \boxed{\log \frac{p}{1-p}} + m \log(1-p) + \log \binom{m}{y} \right] \\ &= \exp\left[\frac{y\boxed{\theta} - b(\theta)}{\boxed{a(\phi)}} + c(y, \phi) \right] \end{aligned}$$

   where $\theta = \log \frac{p}{1-p}$, $\phi = 1$, and

$$\begin{aligned} b(\theta) &= -m \log(1-p) \\ &= -m \log\left(\frac{1}{1+e^\theta}\right) \\ &= m \log(1 + e^\theta) \\ a(\phi) &= \phi \\ c(y, \phi) &= \log \binom{m}{y} \end{aligned}$$

2. The `infert` dataset from the `survival` package presents data from a study of infertility after spontaneous and induced abortion. You can load the dataset using the following command.

```
library(survival)
data(infert)
?infert
str(infert)
```

   The response is `case`, with 1 indicating infertility and 0 fertility. The data comes from a case-control study, the aim of which was to estimate the effect of the number of prior induced and spontaneous abortions on the probability of becoming infertile. We will consider education, age and parity (something numeric, whatever it is) as other predictors.

   Fit a binomial regression model with `case` as a response variable and induced (the number of prior induced abortions), spontaneous (the number of prior spontaneous abortions), education, age and parity as predictors. Test whether the education is important to predict the probability of becoming infertile when all other predictors are in the model.

   Solution We will perform LRT using difference in scaled deviance.

```
> library(survival)
> data(infert)
> model1 <- glm(case ~ age+parity+education+spontaneous+induced,
+                data = infert, family = binomial())
> summary(model1)
```

```
Call:
glm(formula = case ~ age + parity + education + spontaneous +
    induced, family = binomial(), data = infert)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7603  -0.8162  -0.4956   0.8349   2.6536

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.14924    1.41220  -0.814   0.4158
age               0.03958    0.03120   1.269   0.2046
parity           -0.82828    0.19649  -4.215 2.49e-05 ***
education6-11yrs -1.04424    0.79255  -1.318   0.1876
education12+ yrs -1.40321    0.83416  -1.682   0.0925 .
spontaneous       2.04591    0.31016   6.596 4.21e-11 ***
induced           1.28876    0.30146   4.275 1.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 257.80  on 241  degrees of freedom
AIC: 271.8

Number of Fisher Scoring iterations: 4

> model2 <- glm(case ~ age + parity+spontaneous+induced,
+               data = infert, family = binomial())
> summary(model2)

Call:
glm(formula = case ~ age + parity + spontaneous + induced, family = binomial(),
    data = infert)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6281  -0.8055  -0.5299   0.8669   2.6141

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.85239    1.00428  -2.840  0.00451 **
age          0.05318    0.03014   1.764  0.07767 .
parity      -0.70883    0.18091  -3.918 8.92e-05 ***
spontaneous  1.92534    0.29863   6.447 1.14e-10 ***
induced      1.18966    0.28987   4.104 4.06e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 260.94  on 243  degrees of freedom
AIC: 270.94

Number of Fisher Scoring iterations: 4

> pchisq(deviance(model2) - deviance(model1), 2, lower.tail=FALSE)
```

```
[1] 0.2074555
```

The pvalue is bigger than 0.05, so the education is not important to predict the probability of becoming infertile when all other predictors are in the model.

3. The dataset `discoveries` lists the number of great scientific discoveries for the years 1860 to 1959, as chosen by "The World Almanac and Book of Facts", 1975 Edition. Has the discovery rate remained constant over time?

   To answer this question, fit a poisson regression model with a log link, and use the deviance to compare a null model with models including the year and year squared as predictors.

   Load the dataset using the command data(discoveries).

   Solution First we fit two models, the first including the year and the second the year and the year squared. The plot gives the fitted rates in each case.

```
> data(discoveries)
> disc.df <- data.frame(year=1860:1959, disc=discoveries)
> model1 <- glm(disc ~ year, family=poisson, disc.df)
> summary(model1)

Call:
glm(formula = disc ~ year, family = poisson, data = disc.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8112  -0.9482  -0.3533   0.6637   3.5504

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.354807   3.775677   3.007  0.00264 **
year        -0.005360   0.001982  -2.705  0.00683 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 164.68  on 99  degrees of freedom
Residual deviance: 157.32  on 98  degrees of freedom
AIC: 430.32

Number of Fisher Scoring iterations: 5

> model2 <- glm(disc ~ year + I(year^2), family=poisson, disc.df)
> summary(model2)

Call:
glm(formula = disc ~ year + I(year^2), family = poisson, data = disc.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9066  -0.8397  -0.2544   0.4776   3.3303

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.482e+03  3.163e+02  -4.685 2.79e-06 ***
year         1.561e+00  3.318e-01   4.705 2.54e-06 ***
I(year^2)   -4.106e-04  8.699e-05  -4.720 2.35e-06 ***
---
```

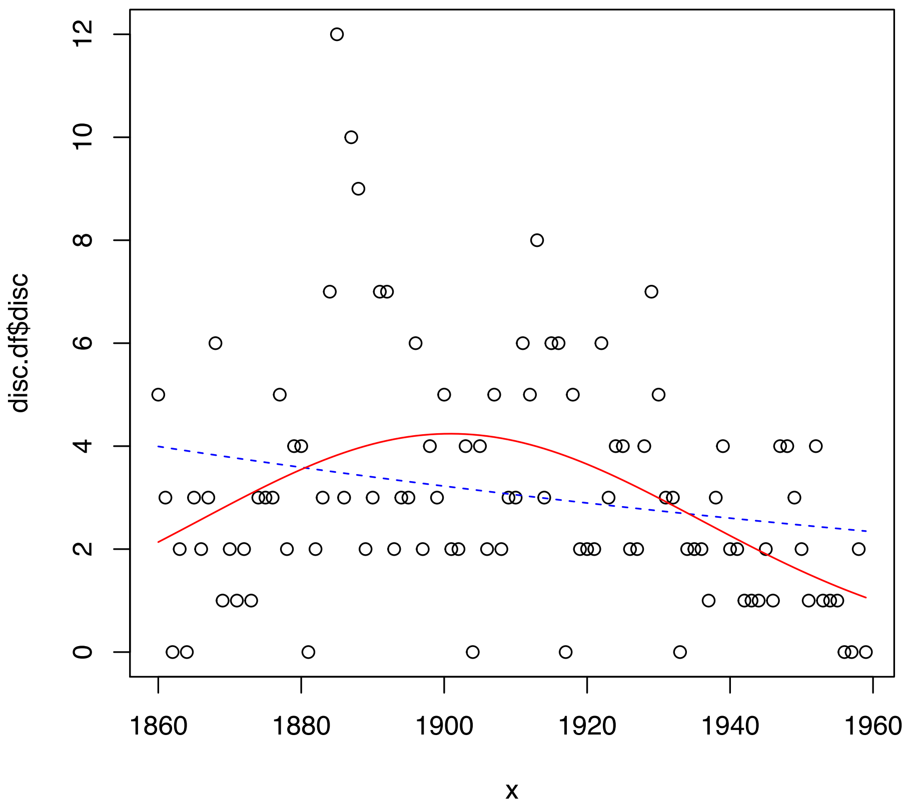
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 164.68  on 99  degrees of freedom
Residual deviance: 132.84  on 97  degrees of freedom
AIC: 407.85

Number of Fisher Scoring iterations: 5
```

```
> x <- disc.df$year
> plot(x, disc.df$disc)
> beta1 <- model1$coefficients
> lines(x, exp(beta1[1] + beta1[2]*x), col="blue", lty=2)
> beta2 <- model2$coefficients
> lines(x, exp(beta2[1] + beta2[2]*x + beta2[3]*x^2), col="red")
```



We will use deviance differences to perform likelihood ratio tests. From the above, the null model has deviance 164.68, the model with just year has deviance 157.32, and the model with year and year squared has deviance 132.84. We compare the null model with the model with year, and the model with year and year squared. We also compare the model with year with the model with year and year squared.

```
> pchisq(164.68-157.32, 1, lower.tail=FALSE)

[1] 0.006669079

> pchisq(164.68-132.84, 2, lower.tail=FALSE)
```

```
[1] 1.219079e-07

> pchisq(157.32-132.84, 1, lower.tail=FALSE)

[1] 7.508521e-07
```

There is strong evidence that year improves the model, and very strong evidence that year squared has something to add. We conclude that there is strong evidence that the discovery rate has changed over time.