Module 6 - Logistic regression. Poisson regression. Log-linear models.

Pawel Chilinski

December 13, 2013

Exercise 1.

Data set kyphosis in library rpart represents data on children who have had corrective spinal surgery. The data contains the following columns:
Kyphosis - a factor with levels: absent, present, indicating if a kyphosis (a type of deformation) was present after the operation,
Age - age in months,
Number - the number of vertebrae involved,
Start - the number of the first (topmost) vertebra operated on.

```
> library(rpart)
> data(kyphosis)
```

- Fit a logistic regression model to explain how probability of presence of kyphosis depends on variables Age, Number and Start.

  ```
  > kyphosis.lr.model <- glm(Kyphosis~Age+Number+Start,data = kyphosis,family = "binomial")
  > summary(kyphosis.lr.model)

  Call:
  glm(formula = Kyphosis ~ Age + Number + Start, family = "binomial",
      data = kyphosis)

  Deviance Residuals:
      Min       1Q   Median       3Q      Max
  -2.3124  -0.5484  -0.3632  -0.1659   2.1613

  Coefficients:
               Estimate Std. Error z value Pr(>|z|)
  (Intercept) -2.036934   1.449575  -1.405  0.15996
  Age          0.010930   0.006446   1.696  0.08996 .
  Number       0.410601   0.224861   1.826  0.06785 .
  Start       -0.206510   0.067699  -3.050  0.00229 **
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  (Dispersion parameter for binomial family taken to be 1)

      Null deviance: 83.234  on 80  degrees of freedom
  Residual deviance: 61.380  on 77  degrees of freedom
  AIC: 69.38

  Number of Fisher Scoring iterations: 5

  > #Converting coefficients to odds ratio:
  > exp(kyphosis.lr.model$coefficients)

  (Intercept)         Age      Number       Start
    0.1304281   1.0109904   1.5077239   0.8134181
  ```

  So we can see that Number is the biggest factor increasing probability of kyphosis (not significant at 0.05 critical value), then increase in Age increases slightly probability of kyphosis and increase in Start decreases this probability.

- What are the calculated null deviance and residual deviance for this model? What are their asymptotic distributions and what conditions are needed for them to hold?

  Calculated null deviance and residual deviance:

  ```
  > kyphosis.lr.model$null.deviance
  ```

```
[1] 83.23447
```

```
> kyphosis.lr.model$deviance
```

```
[1] 61.37993
```

The asymptotic distribution of deviances is chi-squared with n-p degrees of freedom (because deviance is calculated from the saturated model with n parameters).
Null deviance df is:

```
> kyphosis.lr.model$df.null
```

```
[1] 80
```

```
> #which is the same as (n-1, because for null model we have 1 parameter)
> nrow(kyphosis)-1
```

```
[1] 80
```

Residual deviance df is:

```
> kyphosis.lr.model$df.residual
```

```
[1] 77
```

```
> #which is the same as (n-4, because for model we have 4 parameters)
> nrow(kyphosis)-4
```

```
[1] 77
```

The chi-squared are asymptotic distributions provided model (null or fitted by us) is adequate (does not differ significantly from the saturated model) and we have large sample.

- Perform likelihood ratio test to decide whether any of the predictors is significant in the model.

  ```
  > 1-pchisq(kyphosis.lr.model$null.deviance-kyphosis.lr.model$deviance,4)
  ```

  ```
  [1] 0.0002142331
  ```

  So we can see that null model is not adequate and we can reject hypothesis (with critical value of 0.05) that all predictors are insignificant (so model contains some significant predictor(s)).

- Can we perform a goodness of fit test to check adequacy of the model?

  We cannot perform a goodness of fit test because we do not have more than one observation for every object in the data set.

- Calculate the proportion of explained deviance for this model

  ```
  > 1-kyphosis.lr.model$deviance/kyphosis.lr.model$null.deviance
  ```

  ```
  [1] 0.2625661
  ```

  So we can see that only 26% of the deviance is explained by this model.

- One of the methods of proving inadequacy of a model for individual observations is to fit a larger model and perform likelihood ratio test to compare these two models. Enlarge the initial model with squared predictors and compare values of residual deviances, proportions of explained deviance and AIC for these two models. What is the result of LRT test?

  Fitting larger model:

  ```
  > kyphosis.lr.model.squares <- glm(Kyphosis~Age+Number+Start+I(Age^2)+I(Number^2)+I(Start^2),
  +                 data = kyphosis,family = "binomial")
  > 1-pchisq(kyphosis.lr.model$deviance-kyphosis.lr.model.squares$deviance,3)
  ```

  ```
  [1] 0.003590371
  ```

  We can conclude (using 0.05 critical value) that adding squares of predictors significantly improves the model (result of LRT test).

  Explained deviance for the bigger model:

  ```
  > 1-kyphosis.lr.model.squares$deviance/kyphosis.lr.model.squares$null.deviance
  ```

  ```
  [1] 0.42533
  ```

The explained deviance increased to 42%.

Comparing AIC values:

```
> kyphosis.lr.model$aic

[1] 69.37993

> kyphosis.lr.model.squares$aic

[1] 61.83235
```

Akaike Information Criterion is smaller for the bigger model.

All above tests vote in favour of the bigger model.

- Choose the best subset of predictors using:
  - AIC criterion

```
> step(glm(Kyphosis~1,data = kyphosis,family = "binomial"), direction = c("forward"), k=2,
+                  scope=list(upper=.~.+Age+Number+Start+I(Age^2)+I(Number^2)+I(Start^2)))
Start:  AIC=85.23
Kyphosis ~ 1

              Df Deviance    AIC
+ I(Start^2)   1   64.345 68.345
+ Start        1   68.072 72.072
+ Number       1   73.357 77.357
+ I(Number^2)  1   74.668 78.668
<none>             83.234 85.234
+ Age          1   81.932 85.932
+ I(Age^2)     1   83.139 87.139

Step:  AIC=68.35
Kyphosis ~ I(Start^2)

              Df Deviance    AIC
+ Age          1   61.386 67.386
+ Start        1   61.699 67.699
+ Number       1   61.944 67.944
<none>             64.345 68.345
+ I(Number^2)  1   62.492 68.492
+ I(Age^2)     1   63.451 69.451

Step:  AIC=67.39
Kyphosis ~ I(Start^2) + Age

              Df Deviance    AIC
+ I(Age^2)     1   54.608 62.608
+ Number       1   58.691 66.691
+ Start        1   58.845 66.845
<none>             61.386 67.386
+ I(Number^2)  1   59.564 67.564

Step:  AIC=62.61
Kyphosis ~ I(Start^2) + Age + I(Age^2)

              Df Deviance    AIC
+ Start        1   51.298 61.298
+ Number       1   51.776 61.776
<none>             54.608 62.608
+ I(Number^2)  1   52.692 62.692

Step:  AIC=61.3
Kyphosis ~ I(Start^2) + Age + I(Age^2) + Start

              Df Deviance    AIC
<none>             51.298 61.298
+ Number       1   49.455 61.455
+ I(Number^2)  1   50.203 62.203

Call:  glm(formula = Kyphosis ~ I(Start^2) + Age + I(Age^2) + Start, 
    family = "binomial", data = kyphosis)

Coefficients:
(Intercept)   I(Start^2)          Age     I(Age^2)  
 -4.1855978   -0.0481986    0.0816004   -0.0004092  
      Start  
  0.5619041  

Degrees of Freedom: 80 Total (i.e. Null);  76 Residual
Null Deviance:	    83.23 
Residual Deviance: 51.3 	AIC: 61.3
```

So the sub-model selected based on AIC criterion is Kyphosis $\sim$ I(Start^2) + Age + I(Age^2) + Start

– likelihood ratio test (LRT). Use function step(model,test=”Chisq”)

```
> step(glm(Kyphosis~Age+Number+Start+I(Age^2)+I(Number^2)+I(Start^2),data = kyphosis,
+                                  family = "binomial"), direction = c("backward"),
+                                  scope=list(upper=.~1),test="Chisq")
Start:  AIC=61.83
Kyphosis ~ Age + Number + Start + I(Age^2) + I(Number^2) + I(Start^2)

              Df Deviance    AIC     LRT Pr(>Chi)   
- I(Number^2)  1   49.455 61.455  1.6226 0.202727   
<none>             47.832 61.832                    
- Number       1   50.203 62.203  2.3708 0.123626   
- Start        1   50.357 62.357  2.5243 0.112106   
- I(Start^2)   1   53.221 65.221  5.3886 0.020268 * 
- I(Age^2)     1   55.504 67.504  7.6719 0.005609 **
- Age          1   58.190 70.190 10.3575 0.001289 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Step:  AIC=61.45
Kyphosis ~ Age + Number + Start + I(Age^2) + I(Start^2)

             Df Deviance    AIC    LRT Pr(>Chi)   
- Number      1   51.298 61.298 1.8426 0.174652   
<none>            49.455 61.455                   
- Start       1   51.776 61.776 2.3209 0.127649   
- I(Start^2)  1   54.428 64.428 4.9728 0.025749 * 
- I(Age^2)    1   56.942 66.942 7.4872 0.006214 **
- Age         1   59.129 69.129 9.6739 0.001869 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Step:  AIC=61.3
Kyphosis ~ Age + Start + I(Age^2) + I(Start^2)

             Df Deviance    AIC    LRT Pr(>Chi)   
<none>            51.298 61.298                   
- Start       1   54.608 62.608 3.3100 0.068860 . 
- I(Start^2)  1   58.414 66.414 7.1166 0.007637 **
- I(Age^2)    1   58.845 66.845 7.5479 0.006008 **
- Age         1   60.925 68.925 9.6276 0.001917 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Call:  glm(formula = Kyphosis ~ Age + Start + I(Age^2) + I(Start^2), 
    family = "binomial", data = kyphosis)

Coefficients:
(Intercept)          Age        Start     I(Age^2)  
 -4.1855978    0.0816004    0.5619041   -0.0004092  
 I(Start^2)  
 -0.0481986  
```

```
Degrees of Freedom: 80 Total (i.e. Null);  76 Residual
Null Deviance:           83.23
Residual Deviance: 51.3        AIC: 61.3
```

So this method gives the same result as the one based on AIC criterion. So model selected based on AIC and LRT:

```
> kyphosis.lr.model.selected<-glm(Kyphosis~Age+Start+I(Age^2)+I(Start^2),
+                   data = kyphosis,family = "binomial")
```

- Compare values of residual deviances, proportions of explained deviance and AIC for the initial and resulting two models.

  By selecting the model based on AIC and LRT:
  Residual deviance decreased from 61 to 51:

  ```
  > kyphosis.lr.model$deviance

  [1] 61.37993

  > kyphosis.lr.model.selected$deviance

  [1] 51.29752
  ```

  Proportions of explained deviance increased from 26% to 38%:

  ```
  > 1-kyphosis.lr.model$deviance/kyphosis.lr.model$null.deviance

  [1] 0.2625661

  > 1-kyphosis.lr.model.selected$deviance/kyphosis.lr.model.selected$null.deviance

  [1] 0.3836986
  ```

  AIC criterion decreased from 69 to 61:

  ```
  > kyphosis.lr.model$aic

  [1] 69.37993

  > kyphosis.lr.model.selected$aic

  [1] 61.29752
  ```

- What is the estimated probability of presence of kyphosis for a child which is 20 months old and for which variable Start is equal to 10?

  The probability is 13%:

  ```
  > predict(kyphosis.lr.model.selected,data.frame(Age=20,Start=10),type="response")

          1
  0.1280725
  ```

Exercise 2.

The data set Brands.txt contains information on 735 subjects who were asked their preference on three brands of some product (e.g., car or TV). Included in the data set are the following variables:
brand - number of a preferred brand (one of three),
female - coded as 0 for male and 1 for female,
age - subject's age.

```
> brands <- read.table(file="Brands.txt",header=T)
```

Our goal is to associate the brand choices with age and gender. We assume the following relationship between probabilities ratio and our predictor variables female and age:
logit(P (brand = 2)/P (brand = 1)) = $\beta_{10} + \beta_{11}$ female + β_{12} age,
logit(P (brand = 3)/P (brand = 1)) = $\beta_{20} + \beta_{21}$ female + β_{22} age.

- Fit a multinomial logit model to the data. Use function multinom().

  ```
  > library(nnet)
  > brand.mn.model<-multinom(brand~female+age,data=brands)

  # weights:  12 (6 variable)
  initial  value 807.480032
  iter  10 value 702.976983
  final  value 702.970704
  converged
  ```

- Interpret fitted coefficients. Coefficients as odds ratios:

```
> exp(coef(brand.mn.model))

   (Intercept)   female      age
2 7.696912e-06 1.688465 1.445142
3 1.355862e-10 1.593525 1.985575
```

$\beta_{11} = 1.688465$ means that one unit of change in female will multiply the odds of the of the brand = 2 (compared to the brand = 1) by 1.688465
$\beta_{12} = 1.445142$ means that one unit of change in age will multiply the odds of the of the brand = 2 (compared to the brand = 1) by 1.445142
$\beta_{21} = 1.593525$ means that one unit of change in female will multiply the odds of the of the brand = 3 (compared to the brand = 1) by 1.593525
$\beta_{22} = 1.985575$ means that one unit of change in age will multiply the odds of the of the brand = 3 (compared to the brand = 1) by 1.985575
$\beta_{10} and \beta_{20}$ are increase in odds for brand = 2 and 3 relative to brand = 1 when age and female are 0.

- Calculate predicted probabilities of preference for every brand for a group of 15 males and 15 females aged between 24 and 38 (by 1 year) (i.e. age=rep(24:38,2), female=c(rep(0,15),rep(1,15))). Plot predicted probabilities as a function of age (one plot for each gender) for each brand separately.

  Probabilities:

```
> (probs<-predict(brand.mn.model,data.frame(age=rep(24:38,2),female=c(rep(0,15),rep(1,15))),type="probs")

            1          2           3
1  0.94795822 0.05022928 0.001812497
2  0.92560893 0.07087707 0.003514002
3  0.89429634 0.09896238 0.006741279
4  0.85114634 0.13611419 0.012739473
5  0.79313204 0.18329690 0.023571058
6  0.71788078 0.23975762 0.042361592
7  0.62507214 0.30168986 0.073237996
8  0.51809731 0.36137023 0.120532456
9  0.40487271 0.40810321 0.187024082
10 0.29639505 0.43175057 0.271854380
11 0.20299318 0.42732066 0.369686162
12 0.13057819 0.39724053 0.472181272
13 0.07951425 0.34957338 0.570912364
14 0.04627531 0.29400395 0.659720747
15 0.02598163 0.23855071 0.735467663
16 0.91532076 0.08189042 0.002788820
17 0.88079243 0.11387905 0.005328526
18 0.83412798 0.15585237 0.010019650
19 0.77287635 0.20868979 0.018433857
20 0.69561789 0.27143910 0.032943012
21 0.60315605 0.34012755 0.056716394
22 0.49958951 0.40713262 0.093277864
23 0.39239993 0.46212768 0.145472386
24 0.29086347 0.49503135 0.214105181
25 0.20320551 0.49979219 0.297002309
26 0.13411165 0.47668442 0.389203929
27 0.08404134 0.43168592 0.484272746
28 0.05034077 0.37368466 0.575974569
29 0.02903144 0.31143283 0.659535730
30 0.01623089 0.25162197 0.732147148
```

```
> par(mfrow=c(2,3))
> plot(24:38,probs[1:15,1],main="Female=0,brand=1",xlab="age",ylab="Pr")
> plot(24:38,probs[1:15,2],main="Female=0,brand=2",xlab="age",ylab="Pr")
> plot(24:38,probs[1:15,3],main="Female=0,brand=3",xlab="age",ylab="Pr")
> plot(24:38,probs[16:30,1],main="Female=1,brand=1",xlab="age",ylab="Pr")
> plot(24:38,probs[16:30,2],main="Female=1,brand=2",xlab="age",ylab="Pr")
> plot(24:38,probs[16:30,3],main="Female=1,brand=3",xlab="age",ylab="Pr")
```

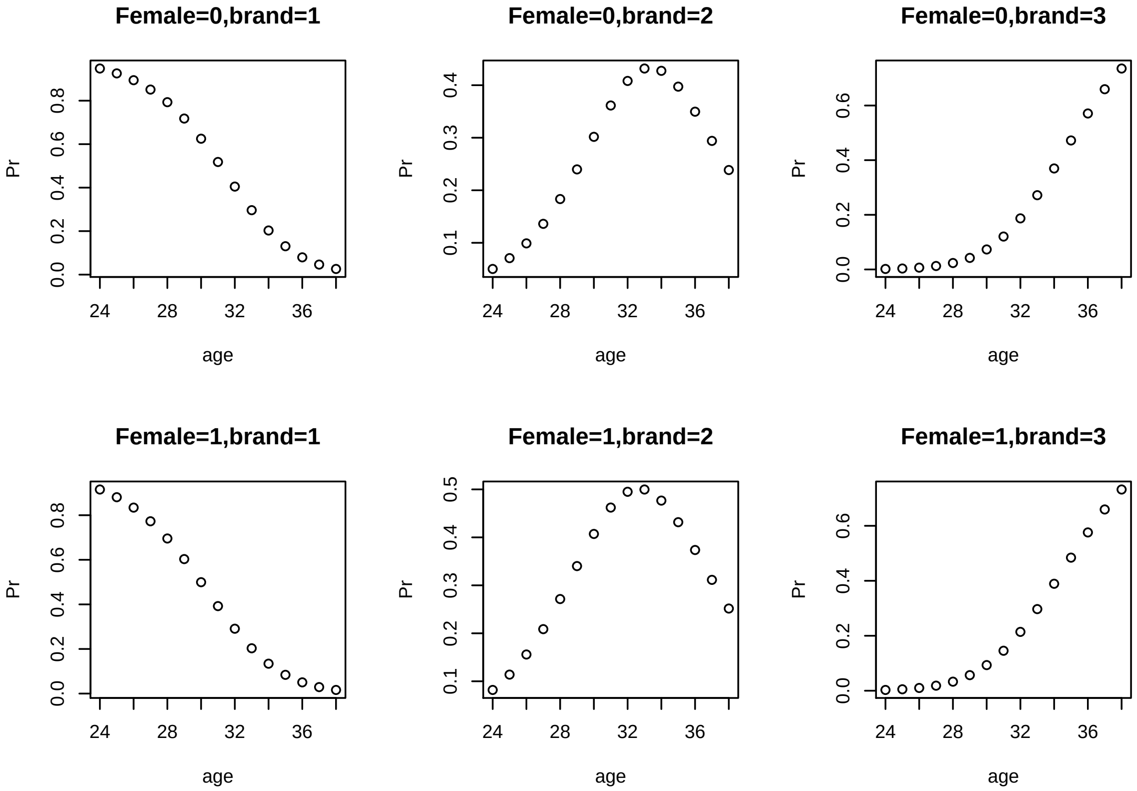

Figure 1: Probabilities as a function of age (one plot for each gender) for each brand separately

Exercise 3.

(Poisson regression) Data set discoveries in base package in R contains numbers of great discoveries in a given year within years 1860-1959. We assume that the number Yi of great discoveries in each year is a random variable pertaining to Poisson distribution: Yi $\sim$ Poiss(μ_i). We are interested in deciding whether the mean value of discoveries is constant in time.

```
> data(discoveries)
```

- Plot the data.

```
> plot(1860:1959,discoveries,ylab="discoveries",xlab="year")
```

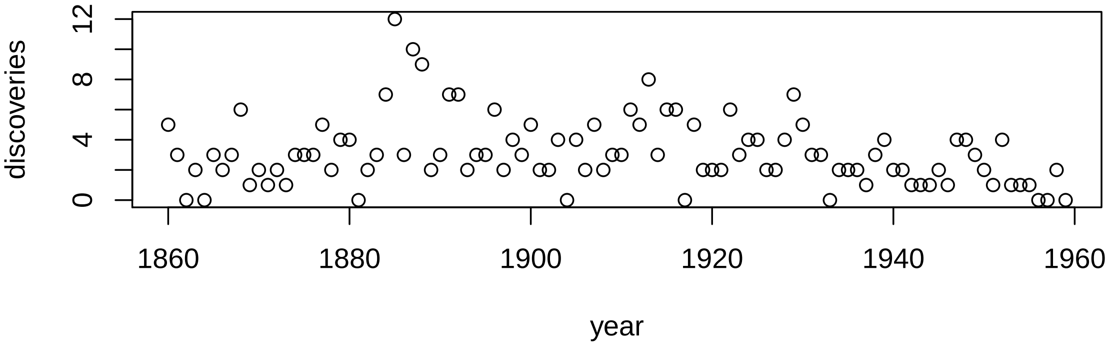

Figure 2: Great discoveries 1860-1959

- Calculate the mean value of discoveries in the whole considered period.

  ```
  > mean(discoveries)
  ```

  ```
  [1] 3.1
  ```

- Fit a poisson regression model taking constant value as the predictor: glm(discoveries∼1,poisson).

  ```
  > (discoveries.poisson.model.const <- glm(discoveries~1,family = "poisson"))
  ```

  ```
  Call:  glm(formula = discoveries ~ 1, family = "poisson")

  Coefficients:
  (Intercept)
        1.131

  Degrees of Freedom: 99 Total (i.e. Null);  99 Residual
  Null Deviance:            164.7
  Residual Deviance: 164.7         AIC: 435.7
  ```

- Perform a goodness-of-fit test (based on residual deviance) of this model.

  ```
  > 1-pchisq(discoveries.poisson.model.const$deviance,length(discoveries)-1)
  ```

  ```
  [1] 3.79455e-05
  ```

  So the model is not adequate and we can conclude that mean value of discoveries is not constant in time. We can compare the result to the Pearson X^2 statistic:

  ```
  > (X_square<-sum((discoveries-discoveries.poisson.model.const$fitted.values)^2/
  +                                                        discoveries.poisson.model.const$fitted.values))
  ```

  ```
  [1] 162.2581
  ```

  ```
  > 1-pchisq(X_square,length(discoveries)-1)
  ```

  ```
  [1] 6.334777e-05
  ```

  And again we receive the same result.

- A different way to check adequacy of this model is fitting a larger model and performing a comparison:

  - fit a quadratic model taking year and year^2 as predictors

    ```
    > (discoveries.poisson.model.quad <- glm(discoveries~years+I(years^2),family = "poisson",
    +                                       data = data.frame(discoveries=discoveries,years=1:100)))
    ```

    ```
    Call:  glm(formula = discoveries ~ years + I(years^2), family = "poisson",
        data = data.frame(discoveries = discoveries, years = 1:100))

    Coefficients:
    (Intercept)        years   I(years^2)
      0.7252798    0.0343781   -0.0004106

    Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
    Null Deviance:            164.7
    Residual Deviance: 132.8         AIC: 407.8
    ```

  - perform a likelihood ratio test to compare these two models. Is the larger model better fitted than the smaller one?

    ```
    >          1-pchisq(discoveries.poisson.model.const$deviance-
    +                                       discoveries.poisson.model.quad$deviance,2)
    ```

    ```
    [1] 1.21532e-07
    ```

    The test result can be interpreted that the simple model is not adequate and we get better model by adding predictors year and year^2

  - Can we say that the mean value of discoveries is constant in time?

    Based on the results we can conclude that the mean value of discoveries is not constant in time. First the constant model is not adequate (by failing goodness of fit test) and the model with additional parameters is better fitted.

- Is the larger model well-fitted? Calculate the percent of explained deviance and perform a goodness-of-fit test for this model.

  Performing goodness of fit for larger model:

```
> 1-pchisq(discoveries.poisson.model.quad$deviance,length(discoveries)-3)
```

```
[1] 0.009204575
```

We can compare the result to the Pearson X^2 statistic:

```
> (X_square<-sum((discoveries-discoveries.poisson.model.quad$fitted.values)^2/
+                                                         discoveries.poisson.model.quad$fitted.values))
```

```
[1] 126.648
```

```
> 1-pchisq(X_square,length(discoveries)-3)
```

```
[1] 0.02326856
```

Even the larger model is not adequate so it is not well fitted.

Percent of explained deviance:

```
> 1-discoveries.poisson.model.quad$deviance/discoveries.poisson.model.quad$null.deviance
```

```
[1] 0.1933768
```

So 19% of deviance is explained by the model.

Exercise 4.

(Log-linear model) File gator.data contais data on alligators. We are interested in testing independence of variables lake and food.

```
> gator <- read.table(file="gator.data",header=T)
```

- Aggregate data to contingency table

  ```
  > gator1 <- aggregate(gator$count,list(food=gator$food,lake=gator$lake),FUN=sum)
  ```

- We want to test hypothesis $p_{ij} = p_i * p_j$

  ```
  > (gator.model <- glm(x~as.factor(food)+as.factor(lake),poisson,gator1))
  ```

  ```
  Call:  glm(formula = x ~ as.factor(food) + as.factor(lake), family = poisson,
      data = gator1)

  Coefficients:
       (Intercept)  as.factor(food)2  as.factor(food)3
           3.16156          -0.43242          -1.59886
  as.factor(food)4  as.factor(food)5  as.factor(lake)2
          -1.97835          -1.07756          -0.13613
  as.factor(lake)3  as.factor(lake)4
          -0.03704           0.13580

  Degrees of Freedom: 19 Total (i.e. Null);  12 Residual
  Null Deviance:          145.9
  Residual Deviance: 43.2         AIC: 136.3
  ```

  ```
  > 1-pchisq(gator.model$deviance,gator.model$df.residual)
  ```

  ```
  [1] 2.092387e-05
  ```

  Model does not fit so we cannot say that the variables of lake and food are independent.

- Note that fitted coefficients in this model reflect margin counts in rows and columns. To see this calculate the fraction of alligators that come from lake 1 and calculate the estimated probability that an alligator comes from lake 1. Compare these two numbers. (The expected number of observations in each cell is equal to the observed count.)

  ```
  > #all alligators
  > n<-sum(gator1$x)
  > #the fraction of alligators that come from lake 1
  > sum(gator1[gator1$lake==1,'x'])/n
  ```

  ```
  [1] 0.2511416
  ```

```
> #the expected fraction
> (exp(sum(gator.model$coefficients[1]))+
+                         exp(sum(gator.model$coefficients[1:2]))+
+                         exp(sum(gator.model$coefficients[1:3]))+
+                         exp(sum(gator.model$coefficients[1:4]))+
+                         exp(sum(gator.model$coefficients[1:5])))/n

[1] 0.1945091
```