Generalized Linear Models

Goals:

Learn how to construct and solve GLMs with multiple explanatory variables

Outline:

Multiple glm (more than one independent variable)
Possibility for interactions

Assignments:

<table>
<thead>
<tr>
<th>Female Crab Width, cm</th>
<th>Young</th>
<th>Old</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of observations</td>
<td>Number having satellites</td>
</tr>
<tr>
<td>23</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>24</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>28</td>
<td>17</td>
</tr>
<tr>
<td>26</td>
<td>39</td>
<td>21</td>
</tr>
<tr>
<td>27</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>28</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>29</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

The table above refers to a study of nesting horseshoe crabs (the same we used in Lab 3). Female crabs are categorized here according to their width and age.

a) Construct and solve five logistic GLMs for different combinations of individual and interacting factors (see the Lab example for details.)

b) Discuss parameter significance (which parameter is significant, which is not? do you see collinearity effect?) and compare the model quality using the graphs, deviance and AIC.

c) Choose the best model and use it to predict the proportion of old female crabs with width 28 cm and having satellites.

Reports: Assignments require printed report, which will consist of R-results (do not print the entire session, only the necessary results!) and plots. Describe briefly the theoretical background for the methods you use, including necessary formulas, and make short statements about result interpretation. Consult instructor if you have any questions about the level of detail or formatting of your report.

Reports are due on Monday, November 5
Essential R commands:

Session management:
- help()
- ls()
- getwd()
- setwd()
- library()
- data()
- save()
- load()
- read.table()
- class()
- names()
- rm()

Vectors:
- c()
- seq()
- rep()
- factor()
- cbind()
- rbind()

Data summaries:
- mean()
- sd()
- median()
- quantile()
- summary()

Graphs:
- par()
- plot()
- points()
- lines()
- mosaicplot()
- text()

GLMs:
- glm()
- family()
- summary.glm()
- predict.glm()
GLMs with more than one independent variable

In case when we have more than one independent variable (say, $x$ and $y$), the analysis can take into account not only individual levels of those variables, but also their joint levels. As in regular regression, one should look for confounding and collinearities.

The sample code in program `GLM_2D.R` shows how to estimate parameters in a multiple GLM and plot the results for the example of the tobacco budworm *Heliothis virescens*’s reaction to doses of the pyrethroid *trans-cypermethrin*. In the experiment, batches of 20 moth of each sex were exposed for three days to the pyrethroid and the number in each batch that were dead or knocked down was recorded. The data are summarized in the table below:

<table>
<thead>
<tr>
<th>Sex</th>
<th>Dose, µg</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>13</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>Female</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>10</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>

We will use the following variables:

- Dose level: $x = 1, 2, 4, 8, 16, 32$
- Gender: $y = M, F$
- Proportion of knocked down moth: $\pi$

Our goal is to predict (describe) the proportion of knocked down moth from their gender and dose, that is find a proper function $\pi(x,y)$. 
Model 1

\[ \text{logit}[\pi(x,y)] = \alpha + \beta x \]

This model assumes the same dose-dependence for males and females.

Call:
\text{glm(formula = SF ~ ldose, family = binomial)}

Deviance Residuals:
\begin{array}{cccccc}
\text{Min} & 1\text{Q} & \text{Median} & 3\text{Q} & \text{Max} \\
-1.7989 & -0.8267 & -0.1871 & 0.8950 & 1.9850 \\
\end{array}

Coefficients:
\begin{array}{cccccc}
\text{Estimate} & \text{Std. Error} & \text{z value} & \text{Pr(>|z|)} \\
(\text{Intercept}) & -2.7661 & 0.3701 & -7.473 & 7.82e-14 *** \\
\text{ldose} & 1.0058 & 0.1236 & 8.147 & 3.74e-16 *** \\
\end{array}

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.876 on 11 degrees of freedom
Residual deviance: 16.984 on 10 degrees of freedom
AIC: 51.094

Effect is highly significant and explains most of the variation in the data. Still, we see that the gender information can improve the prediction.
Model 2

\[ \text{logit}[\pi(x,y)] = \alpha + \gamma y \]

This model assumes the same gender dependence at each dose level. It can be rewritten for males and females as:

- **Males:** \( \text{logit}[\pi(x)] = \alpha \)
- **Females:** \( \text{logit}[\pi(x)] = \alpha + \gamma \)

**Call:**
```
glm(formula = SF ~ sex, family = binomial)
```

**Deviance Residuals:**
```
Min 1Q Median 3Q Max
-4.7887 -2.9371 0.1015 2.3400 4.9522
```

**Coefficients:**
```
            Estimate Std. Error   z value  Pr(>|z|)  
(Intercept)   -0.4754      0.1878   -2.532   0.0113 *  
sexM           0.6425      0.2623    2.449   0.0143 *  
```

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

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

Null deviance: 124.88 on 11 degrees of freedom
Residual deviance: 110.80 on 10 degrees of freedom
**AIC:** 152.91

Effect is significant but can't explain the variation in data: the residual deviance is almost the same as the null deviance. **Collinearity:** Notice that significance of parameters decreased wrt Model 1. This is because both parameters here have the same effect (proportion level) and R cannot decide which one to choose.
Model 3

$$\text{logit}(\pi(x,y)) = \alpha + \beta x + \gamma y$$

This model considers separate effects of gender and dose. It has the same dose-dependence and different intercepts for males and females:

Males: $\text{logit}(\pi(x)) = (\alpha + \gamma) + \beta x$
Females: $\text{logit}(\pi(x)) = \alpha + \beta x$

Call:
```
glm(formula = SF ~ sex + ldose, family = binomial)
```
Deviance Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.10540</td>
<td>-0.65343</td>
<td>-0.02225</td>
<td>0.48471</td>
<td>1.42944</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|
| (Intercept) | -3.4732 | 0.4685 | -7.413 | 1.23e-13 *** |
| sexM | 1.1007 | 0.3558 | 3.093 | 0.00198 ** |
| ldose | 1.0642 | 0.1311 | 8.119 | 4.70e-16 *** |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.876 on 11 degrees of freedom
Residual deviance: 6.757 on 9 degrees of freedom
AIC: 42.867

![Graph](image1.png)

![Graph](image2.png)

The same dose-slope $\beta$ for males and females, but different intercepts: $(\alpha + \gamma)$ and $\alpha$. Effects of both factors are significant and explain really well the data variance.
Model 4

\[
\logit(\pi(x,y)) = \alpha + \beta_1 x I_{y=M} + \beta_2 x I_{y=F}
\]

This model considers joint effects of gender and dose (so-called interaction). It has the same intercept but different dose-dependencies for males and females:

**Males:** \[\logit(\pi(x)) = \alpha + \beta_1 x\]

**Females:** \[\logit(\pi(x)) = \alpha + \beta_2 x\]

**Call:**
```r
glm(formula = SF ~ sex:dose, family = binomial)
```

**Deviance Residuals:**
```
  Min 1Q Median 3Q Max
-1.45854 -0.29503 -0.05339 0.44888 1.06990
```

**Coefficients:**
```
             Estimate Std. Error t value Pr(>|z|)
(Intercept) -2.9073     0.3893 -7.468  8.12e-14  ***
sexF:dose    0.8823     0.1275   6.920   4.52e-12  ***
sexM:dose    1.2893     0.1669   7.723   1.13e-14  ***
```

---

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

(Dispersion parameter for binomial family taken to be 1)

**Null deviance:** 124.8756 on 11 degrees of freedom
**Residual deviance:** 5.0443 on 9 degrees of freedom
**AIC:** 41.155

Different dose-slopes \(\beta\) for males and females, intercept \(\alpha\) is the same. All parameters are significant and explain really well the data variation.
Model 5

\[
\text{logit}[\pi(x,y)] = \alpha + \beta x + \gamma y + \beta_1 x_{(y=M)}
\]

This model also considers joint effects of gender and dose (so-called interaction). It has gender-dependent slope and gender-dependent intercept:

Males: \( \text{logit}[\pi(x)] = (\alpha + \gamma + (\beta + \beta_1)x \)

Females: \( \text{logit}[\pi(x)] = \alpha + \beta x \)

Call:
\text{glm(formula = SF - sex * ldose, family = binomial)}

Deviance Residuals:
\[
\begin{array}{ccccc}
\text{Min} & \text{Q} & \text{Median} & \text{3Q} & \text{Max} \\
-1.39849 & -0.32094 & -0.07592 & 0.38220 & 1.10375 \\
\end{array}
\]

Coefficients:
\[
\begin{array}{cccccc}
\text{Estimate} & \text{Std. Error} & \text{z value} & \text{Pr}(>|z|) \\
(\text{Intercept}) & -2.9935 & 0.5527 & -5.416 & 6.09e-08 \text{ ***} \\
\text{sexM} & 0.1750 & 0.7733 & 0.225 & 0.822 \\
\text{ldose} & 0.9060 & 0.1671 & 5.422 & 5.89e-08 \text{ ***} \\
\text{sexM:ldose} & 0.3529 & 0.2790 & 1.207 & 0.226 \\
\end{array}
\]

---
Signif. codes: 0 ** *** 0.001 *** 0.01 ** 0.05 * 0.1 . 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.8756 on 11 degrees of freedom
Residual deviance: 4.9937 on 8 degrees of freedom
AIC: 43.104

Different dose-slopes \((\beta + \beta_1)\) and \(\beta\) for males and females, different intercepts \((\alpha + \gamma)\) and \(\alpha\). Effects of both factors are significant and explain really well the data variation. Collinearity affects the significance of gender-related coefficients. The AIC suggests that this model is inferior to Model 4: additional parameter is not worth the fit improvement.
Sample R session (file GLM_2D.R)

# ==============================================================
# STAT 453/653
# Generalized Linear Models:
# Several independent variables
# ==============================================================

# Set-up
# ==============================================================
ldose<-rep(0:5,2)
umdead<-c(1,4,9,13,18,20,0,2,6,10,12,16)
sex<-factor(rep(c("M","F"),c(6,6)))
SF<-cbind(numdead,numalive=20-numdead)

# GLM
# ==============================================================
g<-glm(SF~ldose,family=binomial) # model 1
g<-glm(SF~sex,family=binomial) # model 2
g<-glm(SF~sex+ldose,family=binomial) # model 3
g<-glm(SF~sex*ldose,family=binomial) # model 4
g<-glm(SF~sex*ldose,family=binomial) # model 5

# Proportion: Data plot
# ==============================================================
windows()
par(bg='yellow')
plot(c(1,32),c(0,1),type='n',
 xlab='Dose, mg',ylab='Proportion of knocked/dead',log='x')
text(2^ldose,numdead[1:6]/20,labels=as.character(sex[1:6]),col='blue')
text(2^ldose,numdead[7:12]/20,labels=as.character(sex[7:12]),col='red')

# Proportion: Prediction plot
# ==============================================================
ld<-seq(0,5,0.1)
lines(2^ld,predict(g,
data.frame(ldose=ld,sex=factor(rep("M",length(ld)),levels=levels(sex))),
type='response'),col='blue')
lines(2^ld,predict(g,
data.frame(ldose=ld,sex=factor(rep("F",length(ld)),levels=levels(sex))),
type='response'),col='red')

# Logit: Data plot
# ==============================================================
plot(c(1,32),c(-5,5),type='n',
 xlab='Dose, mg',ylab='Logit function',log='x')
text(2^ldose,log(numdead[1:6]/20/(1-numdead[1:6]/20)),
 labels=as.character(sex[1:6]),col='blue')
text(2^ldose,log(numdead[7:12]/20/(1-numdead[7:12]/20)),
 labels=as.character(sex[7:12]),col='red')

# Logit: Prediction plot
# ==============================================================
ld<-seq(0,5,0.1)
p<-predict(g,
data.frame(ldose=ld,sex=factor(rep("M",length(ld)),levels=levels(sex))),
type='response')
lines(2^ld,log(p/(1-p)),col='blue')
p<-predict(g,
data.frame(ldose=ld,sex=factor(rep("F",length(ld)),levels=levels(sex))),
type='response')
lines(2^ld,log(p/(1-p)),col='red')