Generalized Linear Models: One explanatory variable

Goals:

Learn how to construct and solve GLMs with a single explanatory variable

Assignments:

Horseshoe crab study

1a. Construct and solve GLMs with logit and probit links for predicting the proportion of females having satellites from the female width using the following data:

<table>
<thead>
<tr>
<th>Female Crab Width, cm</th>
<th>Number of observations</th>
<th>Number having satellites</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>14</td>
<td>5</td>
<td>0.36</td>
</tr>
<tr>
<td>24</td>
<td>14</td>
<td>4</td>
<td>0.29</td>
</tr>
<tr>
<td>25</td>
<td>28</td>
<td>17</td>
<td>0.61</td>
</tr>
<tr>
<td>26</td>
<td>39</td>
<td>21</td>
<td>0.54</td>
</tr>
<tr>
<td>27</td>
<td>22</td>
<td>15</td>
<td>0.68</td>
</tr>
<tr>
<td>28</td>
<td>24</td>
<td>20</td>
<td>0.83</td>
</tr>
<tr>
<td>29</td>
<td>18</td>
<td>15</td>
<td>0.83</td>
</tr>
<tr>
<td>30</td>
<td>14</td>
<td>14</td>
<td>1.00</td>
</tr>
</tbody>
</table>

This table comes from a study of nesting horseshoe crabs (J. Brockmann, *Ethology*, 102: 1-21, 1996). Each female horseshoe crab in the study had a male crab attached to her in her nest. The study investigated whether the female crab width has any effect on whether the female had any other males, called satellites, residing nearby her.

1b. Plot the observed sample proportions and the predicted values from logit and probit models.

1c. Find the predicted number of satellites for the female width of 25.5 cm according to each model.

Snoring study (see below)

2. Use different numerical representation of snoring levels to construct logit and probit models for snoring data. Find such snoring levels that your models have the best fit to the data (by trial-and-error). Compare the best levels and discuss.

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.

The deadline for reports is Monday, October 29, 2012
### 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()`
1. One independent variable

We consider here an example based on the data from an epidemiological survey of 2484 subjects described in the table below:

<table>
<thead>
<tr>
<th>Snoring</th>
<th>Heart Disease</th>
<th>Proportion Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Never</td>
<td>24</td>
<td>1355</td>
</tr>
<tr>
<td>Occasional</td>
<td>35</td>
<td>603</td>
</tr>
<tr>
<td>Nearly every night</td>
<td>21</td>
<td>192</td>
</tr>
<tr>
<td>Every Night</td>
<td>30</td>
<td>224</td>
</tr>
</tbody>
</table>

The subjects here are classified according to their snoring level, as reported by their spouses (see P. G. Norton and E. V. Dunn, *Br. Med. J.*, 291, 630-632, 1985.)

Our goal is to construct a glm that will predict the probability to have a heart disease from the snoring level. We will consider logit and probit link functions that lead to the following models:

\[
\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \alpha + \beta x \quad \text{(logit)}
\]

\[
\text{probit} \left( \pi(x) \right) = \alpha + \beta x \quad \text{(probit)}
\]

We need to choose the numerical values for snoring levels. For example, for snoring levels \( x = [0, 2, 3, 4] \) we find, using the MLE approach:

\[
\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = -3.95 + 0.52x \\
\text{probit} \left( \pi(x) \right) = -2.09 + 0.24x
\]

The sample code in program glm_1D.R shows how to estimate parameters in those models and plot the results. The model fitting is done using the command `glm()`, the detailed results can be seen using the command `summary()`, the prediction using the fitted model is done using the command `predict()`.
Details of `glm()`:

```
gl <- glm(SF ~ s, family = binomial(link = 'logit'))
```

- `link` = function of the mean response
- `Explanatory variable`
- `Family` = distribution of response
- `Response variable`; for binomial link can be a two-column matrix with success/failure counts

Details of `predict()`:

```
pl <- predict(gl, data.frame(s = x), type = 'response')
```

- `“response”` = compute the response variable
- `“link”` = compute the link function values
- `Result of the glm()`
- `“x”` is the grid on which the predicted values will be calculated
- `“s”` is the name that we used in our call to glm()
Sample R session (file glm.R)

#=========================================  
#              STAT 453/653  
#        Generalized Linear Models  
#=========================================  

# One independent (explanatory) variable  
#=========================================

# Set-up  
#=========================================  

data <- c(0, 2, 3, 4) # snoring levels  
hd <- c(24, 35, 21, 30) # heart disease counts  
n <- c(1355, 603, 192, 224) # healthy patient counts  
SF <- cbind(hd, n) # matrix of "successes" and "failures"  
p <- hd/(n+hd) # probabilities of "success"  

# GLM estimation  
#=========================================  

gl <- glm(SF~s, family=binomial(link='logit'))  
summary(gl)  
summary(gp)  

# Response function plot  
#=========================================  

cats <- 0:5  
probs <- seq(0, 1, by=0.1)  
pl <- predict(gl, data.frame(s=x), type='response')  
lines(x, pl, col=3) # green line  
pp <- predict(gp, data.frame(s=x), type='response')  
lines(x, pp, col=4) # blue line  
grid()  

# Link function plot  
#=========================================  

cats <- 0:5  
pl <- predict(gl, data.frame(s=x), type='link')  
lines(x, pl, col=3)