Logistic Regressions

The goal of logistic regression is to estimate the probability p_i of a binary event (0,1) given predictor variables. For example, is success or failure of an animal to reproduce a function of its age? Or other factors too? Many outcomes can be described in these terms, and logistic regression is widely used. The OpenIntro book provides an effective overview – read it! Meanwhile, here comes some practice doing logistic regression in R.

For a logistic regression, we expect a logistic relationship relationship between pi and the predictor variables that is S-shaped, like the population growth model, where a switch from p_i = 0 to p_i = 1 takes place somewhere in the pattern. Logistic regression is based on the logit function, which is a log transformation of p_i:

$$logit(p_i) = \log_e \left(\frac{p_i}{1 - p_i} \right)$$

To compute a logistic regression, we must introduce a new form of regression - the **Generalized** Linear Model, or **glm**. A glm is able to deal with a big problem for lm: error variance that is not evenly distributed across the model. Here is an example of such a variance problem for lm:

1. Import and attach the `islandbird.txt` data set. This includes incidence (presence = 1, absence = 0) for a bird species on islands in an archipelago, with given area (km^2) and isolation (km from the nearest island).
2. Make simple plots of incidence as function of isolation and of area to see the data. Do you think both predictor factors affect incidence?
3. Let's verify that `lm` and `glm` will yield the same answer with an assumption of a Gaussian (i.e., normal) error variance. Enter:

```
liniso1 <- lm(incidence ~ isolation)
liniso2 <- glm(incidence ~ isolation, family=gaussian)
summary(liniso1)
summary(liniso2)
```

Do we get the same coefficients? Notice that we walk away from R^2 with `glm` – instead we use AIC to evaluate alternative models made with `glm`.

4. Examine the lm residuals by entering

```
par(mfrow=c(2,2))
plot(liniso1)
```

   See any problems? You should!

5. Now we try a logistic glm, where we can specify that binomial errors are to be expected with the binary data. Logistic regression simply assumes response variable observations are independent. That's it - no need not sweat residual distributions. Now make a new glm model, with all as in liniso2 but use family=binomial instead, and get a summary.

6. Notice how much the coefficients changed simply by assuming a binomial distribution

for the binary data?

7. How much better is your logistic model than the linear glm (`liniso2`)? Load the `bbmle` package and compute an `AICctab` with `weights=TRUE` to find out.

8. Now also compute a similar logistic function for incidence as a function of area.

9. Now let's `plot` our logiiso and logiarea functions easy & pretty, in the popbio package. `Install the popbio package` and then:

```
library(popbio)
logi.hist.plot(isolation,incidence,boxp=FALSE,type="hist",col="gray")
logi.hist.plot(area,incidence,boxp=FALSE,type="dit",col="gray")
```

10. See what you did there? Play with it a little. Other options in the `popbio` Help screen can customize that plot – for example, the width of histogram bars, etc. etc.

11. Which model (`logiiso` or `logiarea`) best explains incidence of the bird on the islands?

12. Given that both isolation and area are central to the Theory of Island Biogeography (TIB), and that both look logistic (though in opposite directions), let's make a multiple logistic regression:

```
logimult <- glm(incidence ~ area + isolation, family=binomial)
summary(logimult)
```

13. What does this tell you? **Note**: Coefficients in the model are in logit units. So the odds ratio for area = exp(0.5807) = 1.7873 and for isolation = exp(-1.3719) = 0.2536. For a decent description of odds ratios, check out http://en.wikipedia.org/wiki/Odds_ratio.

14. Now how to plot this? Try this:

```
newdata <- expand.grid(
    isolation = pretty(islandbird$isolation, 20),
    area = pretty(islandbird$area, 20))
newdata$predicted_incidence <- predict(logimult, newdata = newdata, type = "response")
library(ggplot2)
library(cowplot)
ggplot(newdata, aes(x = isolation, y = predicted_incidence, colour = area, group =
    area)) + geom_line()
ggplot(newdata, aes(x = area, y = predicted_incidence, colour = isolation, group =
    isolation)) + geom_line()
```

15. Which most efficiently predicts incidence: the isolation- or the area-based model?

16. Does a CART analysis (e.g., using the `tree` package) help "see" the pattern?