## Logistic Regressions

The goal of logistic regression is to estimate the probability $p_i$ of a binary event *i* given predictor variables. For example, is success (1) or failure (0) of an animal to reproduce a function of its age? Or other factors too? Many outcomes can be described in these terms. Logistic regression is thus flexible, widely used, and fairly simple to run, though some thought is required to express outcomes as probabilities.

A logistic relationship between $p_i$ and the predictor variables is S-shaped, like the population growth model, where a switch from $p_i = 0$ to $p_i = 1$ takes place somewhere in the middle. 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 again use Generalized Linear Model (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) as predictors.
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 first try a linear model that assumes a Gaussian (i.e., normal) error variance. Enter:

```
liniso <- glm(incidence ~ isolation, family=gaussian)
summary(liniso)
par(mfrow=c(2,2))
plot(liniso)
```

   See any problems? *You should!* Thus the problem for analyzing binary data with tools used so far this semester.

4. 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 `liniso` but use `family=binomial` instead, and get a summary. I assume you call that model `logiiso`.

   Notice how much the coefficients changed simply by assuming a binomial distribution for the binary data? Coefficients in logistic regression indicate the effect of a one-unit change in the predictor variable on the *log odds of 'success'*.

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

6. Now also compute a similar logistic function for incidence as a function of area. I assume

you call that model `logiarea`.

7. Now `plot logiiso` and `logiarea` functions *easy peasy*, 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")
```

8. See what you did there? Play with the code above a little. See other options in the `popbio` Help screen to customize that plot – for example, the width of histogram bars, etc. etc.
9. Which model (`logiiso` or `logiarea`) best explains incidence of the bird on the islands?
10. Both isolation and area are central to the Theory of Island Biogeography, and both look logistic (though in opposite directions), so let's make a multiple logistic regression:

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

11. What does this tell you? **Note**: Coefficients in the model are in *logit* units, which represent the log(odds of 'success'). For a decent description of odds ratios, check out http://en.wikipedia.org/wiki/Odds_ratio. So the odds ratio for area = exp(0.5807) = 1.7873 and for isolation = exp(-1.3719) = 0.2536. In English: for a one-unit increase in area, the odds of incidence increase 1.78-fold. And a one-unit increase in isolation means the odds of incidence are about ¼ of the previous value.
12. Now how to plot this? Try this:

```
# make an empty dataframe called preddata
preddata <- expand.grid(
   isolation = pretty(islandbird$isolation, 20),
   area = pretty(islandbird$area, 20))
# fill that dataframe with predicted values
preddata$predicted_incidence <- predict(logimult, newdata=preddata, type="response")

#plot the predicted curves with data
library(ggplot2)
library(cowplot)
# first with predictor = isolation and covariate = area
ggplot(preddata, aes(x = isolation, y = predicted_incidence, colour = area, group =
area)) +
   geom_line() +
   geom_point(data=islandbird, aes(y=incidence, x=isolation), alpha=0.5, size=2)
# then with predictor = area and covariate = isolation
ggplot(preddata, aes(x = area, y = predicted_incidence, colour = isolation, group =
isolation)) +
   geom_line() +
   geom_point(data=islandbird, aes(y=incidence, x=area), alpha=0.5, size=2)
```

13. Now say out loud: Oooooooh. Ahhhhhh.
14. Which most clearly predicts incidence: isolation or area?
15. Does a CART analysis (e.g., using the `tree` package) help "see" the pattern?