Logistic Regression

The Logistic Regression Model

$$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$y = 1$ if the event of interest occurs; $y = 0$ otherwise

Example: Farm Ponds

In a study of small, constructed agricultural ponds in southeastern Minnesota, pond and the surrounding landscape features were used to assess their value as amphibian breeding sites. One measure of this was when the amphibian species richness was at least four.

Species richness is the number of different species observed at each pond.

Example: Farm Pond Variables

Dependent Variable

RICH = 1 if species richness is at least 4; RICH = 0 otherwise

Independent Variables

FISH = 1 if fish are present; FISH = 0 otherwise

TOTNITR = total nitrogen in mg/L

R Code for Farm Pond Example

```r
rich.out <- glm(RICH~FISH + TOTNITR,data=farmpond,family="binomial")
summary(rich.out)
```

Farm Ponds: Logistic Regression Output

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)   
## (Intercept)    4.451      1.452   3.066  0.00217 **
## FISH          -4.039      1.387  -2.912  0.00359 **
## TOTNITR       -4.195      1.794  -2.338  0.01937 * 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 50.446  on 39  degrees of freedom
## Residual deviance: 25.591  on 37  degrees of freedom
```

Farm Pond - The Estimated Model

Here is the estimated model, with the coefficients from the R output.

$$\ln\left(\frac{\widehat{P(y=1)}}{1-P(y=1)}\right) = 4.451 - 4.039 \cdot \text{FISH} - 4.195 \cdot \text{TOTNITR}$$

Farm Pond Example - Interpreting Coefficients of Categorical Predictors

Exponentiating the coefficients yields **odds ratios**. Using the estimated coefficient of FISH, we have

$$e^{-4.039} = 0.018$$

The odds of having species richness of at least 4 at a pond where fish are present are only 1.8% as large as the odds of species richness being at least 4 at a pond where fish are not present, given that total nitrogen is held constant. That is, they are 98.2% less!

Farm Pond Example - Interpreting Coefficients of Categorical Predictors (again)

When coefficients are negative, the interpretation sometimes has more impact when we switch the perspective and use the reciprocal of the exponentiated coefficient.

$$\frac{1}{e^{-4.039}} = e^{4.039} = 56.8$$

The odds of having species richness of at least 4 at a pond where fish are **not** present are 56.8 times as large as the odds of species richness being at least 4 at a pond where fish are present.

Farm Pond Example - Interpreting Coefficients of Quantitative Predictors

Using the estimated coefficient of TOTNITR, we have

$$e^{-4.195} = 0.015$$

The odds of having species richness of at least 4 decrease by 98.5% for each additional mg/L of total nitrogen.

More R Code for Farm Pond Example

Confidence intervals for the odds ratios are available via the following R code.

```r
rich.out <- glm(RICH~FISH + TOTNITR,data=farmpond,family="binomial")
exp(confint(rich.out))
```

```
##                    2.5 %       97.5 %
## (Intercept) 9.1726211926 4004.9580926
## FISH        0.0005135249    0.1712424
## TOTNITR     0.0001404314    0.2486600
```

Farm Pond Example: Predicted Probabilities

```r
rich.out <- glm(RICH~FISH + TOTNITR,data=farmpond,family="binomial")
newdata <- data.frame(FISH=0,TOTNITR=0.8)
predict(rich.out,newdata,type="response")
```

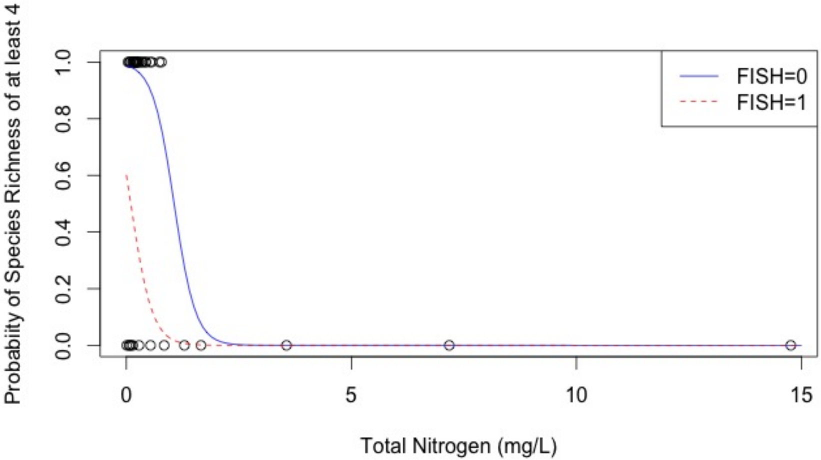
```
##         1
## 0.7492629
```

Confidence Intervals for Predicted Probabilities

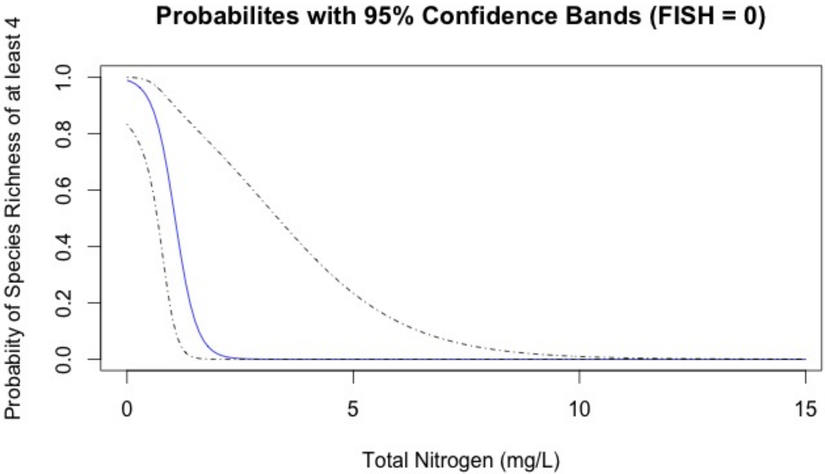
```r
rich.out <- glm(RICH~FISH + TOTNITR,data=farmpond,family="binomial")
out <- predict(rich.out, newdata, se.fit=TRUE)
C = .95  # define the level of confidence
crit = qnorm(1-(1-C)/2)  # get the appropriate critical value
lower = exp(out$fit-crit*out$se.fit)/(1+exp(out$fit-crit*out$se.fit))
upper = exp(out$fit+crit*out$se.fit)/(1+exp(out$fit+crit*out$se.fit))
c(lower,upper)
```

```
##         1         1
## 0.3523169 0.9425807
```

Farm Pond Example: Plotting Predicted Probabilities



Confidence Bands for Predicted Probabilities



Model selection using AIC with “step”

Suppose the variable POND_AREA in the farmpond data set was suspected of having an effect on the species richness being at least 4. Consider fitting the logistic regression model containing FISH, TOTNITR, and POND_AREA and all possible interactions among them.

```r
rich.out.full <- glm(RICH~FISH*TOTNITR*POND_AREA,data=farmpond,
                     family="binomial")
```

Model selection using AIC with “step”

The R function **step** is used as follows to evaluate

```
rich.out.full <- glm(RICH~FISH*TOTNITR*POND_AREA,data=farmpond,
                     family="binomial")
step(rich.out.full)
```

First step from the output of “step”

```
## Start:  AIC=30.9
## RICH ~ FISH * TOTNITR * POND_AREA
##
##                           Df Deviance    AIC
## - FISH:TOTNITR:POND_AREA  1   14.897 28.897
## <none>                        14.897 30.897
```

Last step from the output of “step”

```
## Step:  AIC=23.92
## RICH ~ FISH + TOTNITR + FISH:TOTNITR
##
##                 Df Deviance    AIC
## <none>              15.924 23.924
## - FISH:TOTNITR  1   25.591 31.591
```

Testing subsets of coefficients with “anova”

```
rich.out <- glm(RICH~FISH + TOTNITR,data=farmpond,family="binomial")
rich.out2 <- glm(RICH~FISH*TOTNITR + POND_AREA,data=farmpond,family="b
anova(rich.out,rich.out2,test="Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: RICH ~ FISH + TOTNITR
## Model 2: RICH ~ FISH * TOTNITR + POND_AREA
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        37     25.591
## 2        35     15.437  2   10.154 0.006238 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```