

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

```
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.

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

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.

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

##	2.5 %	97.5 %
## (Intercept)	9.1726211883	4004.9580787
## FISH	0.0005135249	0.1712424
## TOTNITR	0.0001404314	0.2486600

Farm Pond Example: Predicted Probabilities

```
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
```

Classification Using Logistic Regression

If $P(y = 1) \geq 0.5 \rightarrow$ Classify as $y = 1$

If $P(y = 1) < 0.5 \rightarrow$ Classify as $y = 0$

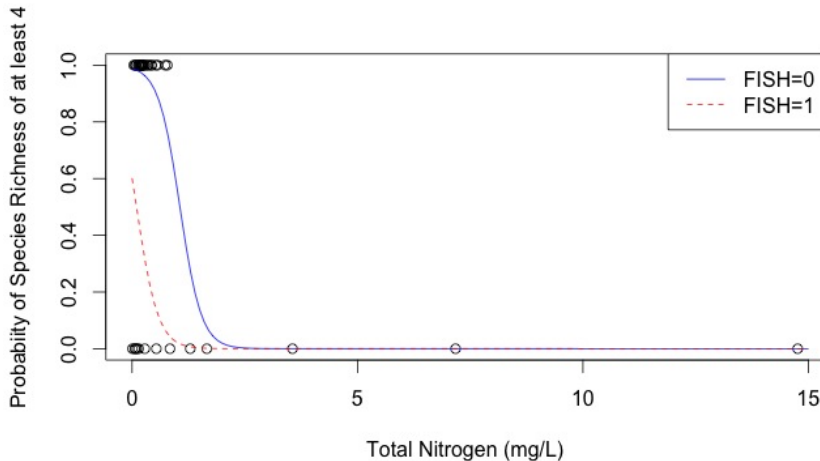
Since the estimated probability of a farm pond having a species richness of at least 4 for a pond with no fish and 0.8 mg/L of total nitrogen is 0.749, a pond with those characteristics would be classified as having species richness of at least 4.

Confidence Intervals for Predicted Probabilities

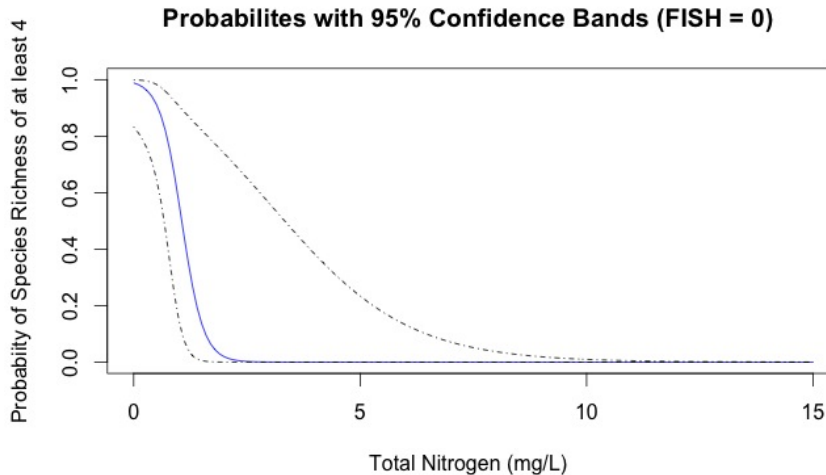
```
rich.out <- glm(RICH~FISH + TOTNITR,data=farmpond,family="binomial")
newdata <- data.frame(FISH=0,TOTNITR=0.8)
out <- predict(rich.out, newdata, se.fit=TRUE)
C = .95 # define the level of confidence
crit = qnorm(1-(1-C)/2) # get the appropriate critical value
lower = exp(out$fit-crit*out$se.fit)/(1+exp(out$fit-crit*out$se.fit))
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



Compare to Computing the Odds Ratio from a Table

The odds ratio computed from the coefficient in the logistic regression model with only that predictor in the model is the same as what is computed from the counts in the contingency table.

##		Richness < 4 Richness >= 4	
##			
##	No Fish	6	22
##	Fish Present	7	5

$$\frac{5/7}{22/6} = 0.1948052$$

Compare to Computing the Odds Ratio from a Table

The odds ratio computed from the model is:

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.2993     0.4606   2.821  0.00479 **
## FISH           -1.6358     0.7450  -2.196  0.02811 *
## ---
```

$$e^{(-1.635755)} = 0.1948052$$

Exactly the same as from the table!

What's What? One More Detail

```
data(Typing)
summary(Typing)
```

##	Method	Words
##	prior course:22	Min. :25.00
##	self-taught :26	1st Qu.:32.00
##		Median :35.00
##		Mean :35.00
##		3rd Qu.:37.25
##		Max. :51.00

What's What? One More Detail

```
## Call:
## glm(formula = Method ~ Words, family = "binomial", data = Typing)

##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   7.18763     2.93538   2.449   0.0143 *
## Words        -0.20059     0.08355  -2.401   0.0164 *
## ---
```

What's What? One More Detail

convert to 0's and 1's

```
Typing$Method2 <- ifelse(Typing$Method=="self-taught",1,0)
```

What's What? One More Detail

[video showing data table in R viewer]

What's What? One More Detail

```
## Call:
## glm(formula = Method2 ~ Words, family = "binomial", data = Typing)

##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   7.18763     2.93538   2.449   0.0143 *
## Words        -0.20059     0.08355  -2.401   0.0164 *
## ---
```

What's What? One More Detail

```
# convert to 0's and 1's
```

```
Typing$Method3 <- ifelse(Typing$Method=="prior course",1,0)
```


What's What? One More Detail

```
## Call:
## glm(formula = Method3 ~ Words, family = "binomial", data = Typing)

##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.18763    2.93538  -2.449   0.0143 *
## Words        0.20059    0.08355   2.401   0.0164 *
## ---
```

