# Class 9: Chapter 6

## EPH 705

Min Lu

Division of Biostatistics
University of Miami

Spring 2017

# Overview

1. Object:
    - Logit Models for Ordinal Responses
    - Plot and Understand Model Specification
    - Cohen's Kappa Statistic
2. R Exercise
    - Logit Model for Ordinal Responses
    - Cohen's Kappa Statistic
    - In class exercise
    - Take home exercise

# Logit Models for Ordinal Responses



Proportional-odds cumulative logit model is possibly the most popular model for ordinal data. This model uses cumulative probabilities upto a threshold, thereby making the whole range of ordinal categories binary at that threshold. Let the response be $Y = 1, 2, ..., r$ where the ordering is natural. The associated probabilities are $\pi_1, \pi_2, ..., \pi_r$, and a cumulative probability of a response less than equal to $j$ is:

$$P(Y \leq j) = \pi_1 + \ldots + \pi_j$$

Then a cumulative logit is defined as

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \log\left(\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right)$$

This describes the log-odds of two cumulative probabilities, one less-than and the other greater-than type. This measures how likely the response is to be in category $j$ or below versus in a category higher than $j$.

# Logit Models for Ordinal Responses



The sequence of cumulative logits may be defined as:

**Cumulative logits**

$$L_1 = \log\left(\frac{\pi_1}{\pi_2 + \pi_3 + \cdots + \pi_r}\right)$$

$$L_2 = \log\left(\frac{\pi_1 + \pi_2}{\pi_3 + \pi_4 + \cdots + \pi_r}\right)$$

$$\vdots$$

$$L_{r-1} = \log\left(\frac{\pi_1 + \pi_2 + \cdots + \pi_{r+1}}{\pi_r}\right)$$

In this notation, $L_j$ is the log-odds of falling into or below category $j$ versus falling above it.

# Proportional-odds Cumulative Logit Model

Incorporate covariates into the model:

$$
\begin{aligned}
L_1 &= \beta_{10} + \beta_{11}X_1 + \cdots + \beta_{1p}X_p \\
L_2 &= \beta_{20} + \beta_{21}X_1 + \cdots + \beta_{2p}X_p \\
&\vdots \\
L_{r-1} &= \beta_{r-1,0} + \beta_{r-1,1}X_1 + \cdots + \beta_{r-1,p}X_p
\end{aligned}
$$

Now suppose that we simplify the model by requiring the coefficient of each X-variable to be identical across the $r-1$ logit equations:

**Proportional-odds cumulative logit model**

$$
\begin{aligned}
L_1 &= \alpha_1 + \beta_1X_1 + \cdots + \beta_pX_p \\
L_2 &= \alpha_2 + \beta_1X_1 + \cdots + \beta_pX_p \\
&\vdots \\
L_{r-1} &= \alpha_{r-1} + \beta_1X_1 + \cdots + \beta_pX_p
\end{aligned}
$$

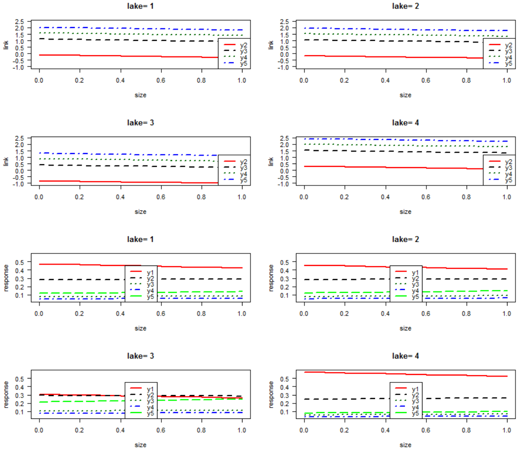
This model has $(r-1)$ intercepts plus $p$ slopes, for a total of $r+p-1$ parameters to be estimated.

# Interpretation



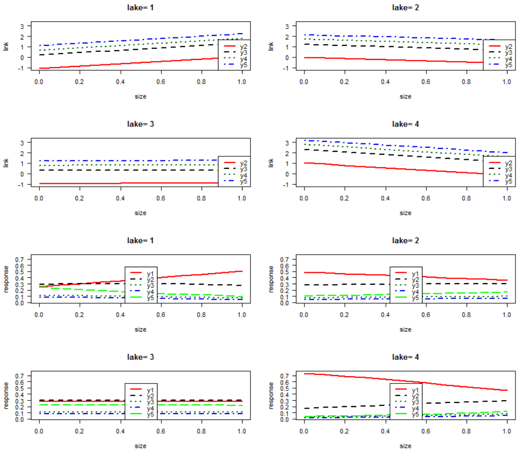
- In the above model, intercept $\alpha_j$ is the log-odds of falling into or below category $j$ when $X_1 = X_2 = ... = 0$.
- A single parameter $\beta_k$ describes the effect of $x_k$ on $Y$ such that $\beta_k$ is the increase in log-odds of falling into or below any category associated with a one-unit increase in $X_k$, holding all the other X-variables constant.
- Constant sloped $\beta_k$: The effect of $X_k$, is the same for all $r-1$ ways to collapse $Y$ into dichotomous outcomes. For simplicity, let's consider only one predictor: logit$[P(Y \leq j)] = \alpha_j + \beta x$. Then the cumulative probabilities are given by: $P(Y \leq j) = \exp(\alpha_j + \beta x)/(1 + \exp(\alpha_j + \beta x))$ and since $\beta$ is constant, the curves of cumulative probabilities plotted against $x$ are parallel.
- The odds-ratio is proportional to the difference between $x_1$ and $x_2$ where $\beta$ is the constant of proportionality: $\exp[\beta(x_1 - x_2)]$, and thus the name "proportional odds model".

## Model Specification Example: alligators SAS example

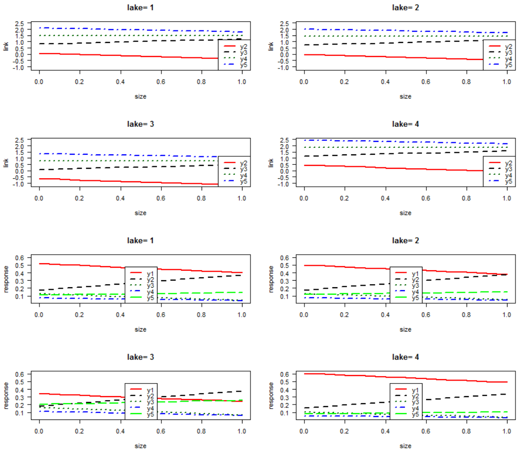


# Proportional-Odds Cumulative Logit Model with Interaction

## Model Specification Example: alligators SAS example



## Model Specification Example: alligators SAS example



# Cohen's Kappa Statistic



Cohen's kappa coefficient, $\kappa$, is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, since $\kappa$ takes into account the possibility of the agreement occurring by chance. Cohen's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories.
The definition of $\kappa$ is:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where $p_o$ is the relative observed agreement among raters (identical to accuracy), and $p_e$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as given by $p_e$), $\kappa \leq 0$. For categories $k$, number of items $N$ and $n_{ki}$ the number of times rater i predicted category $k$:

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

# Cohen's Kappa Statistic



Suppose that you were analyzing data related to a group of 94 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Suppose the disagreement count data were as follows, where A and B are readers, data on the main diagonal of the matrix (top left-bottom right) the count of agreements and the data off the main diagonal, disagreements:

| | | **B** | |
|---|---|---|---|
| | | Yes | No |
| **A** | Yes | a | b |
| | No | c | d |

e.g.

| | | **B** | |
|---|---|---|---|
| | | Yes | No |
| **A** | Yes | 20 | 5 |
| | No | 10 | 15 |

# Cohen's Kappa Statistic



The observed proportionate agreement is:

$$p_o = \frac{a+d}{a+b+c+d} = \frac{20+15}{50} \approx 0.70$$

To calculate $p_e$ (the probability of random agreement), the expected probability that both would say yes at random is:

$$p_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = 0.5 * 0.6 = 0.3$$

Similarly:

$$p_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d} = 0.5 * 0.4 = 0.2$$

Overall random agreement probability is the probability that they agreed on either Yes or No, i.e.:

$$p_e = p_{\text{Yes}} + p_{\text{No}} = 0.3 + 0.2 = 0.5$$

So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.70 - 0.50}{1 - 0.50} = 0.40$$

# R Exercise

## Proportional-odds Cumulative Logit Model

```
library("VGAM")
alligators <- read.table('http://www.stat.ufl.edu/~aa/glm/data/Alligators.dat',header = T)
fit <- vglm(cbind(y1,y2,y3,y4,y5)~ size + factor(lake)
          # relevel(factor(lake), ref = "1")
           ,family=cumulative(parallel=T), data=alligators)
#summary(fit)
round(coef(fit,matrix = T), 2)
```

```
##                 logitlink(P[Y<=1]) logitlink(P[Y<=2]) logitlink(P[Y<=3])
## (Intercept)                  -0.10               1.12               1.60
## size                         -0.21              -0.21              -0.21
## factor(lake)2                -0.06              -0.06              -0.06
## factor(lake)3                -0.71              -0.71              -0.71
## factor(lake)4                 0.41               0.41               0.41
##                 logitlink(P[Y<=4])
## (Intercept)                   2.02
## size                         -0.21
## factor(lake)2                -0.06
## factor(lake)3                -0.71
## factor(lake)4                 0.41
```

```
round(depvar(fit), 3)## predicted proportion
```

```
##      y1    y2    y3    y4    y5
## 1 0.590 0.103 0.051 0.051 0.205
## 2 0.438 0.000 0.062 0.188 0.312
## 3 0.250 0.550 0.050 0.000 0.150
## 4 0.464 0.286 0.214 0.036 0.000
## 5 0.208 0.458 0.083 0.042 0.208
## 6 0.276 0.241 0.207 0.103 0.172
## 7 0.390 0.463 0.024 0.049 0.073
## 8 0.773 0.045 0.000 0.045 0.136
```

# Plot and Understand Model Specification

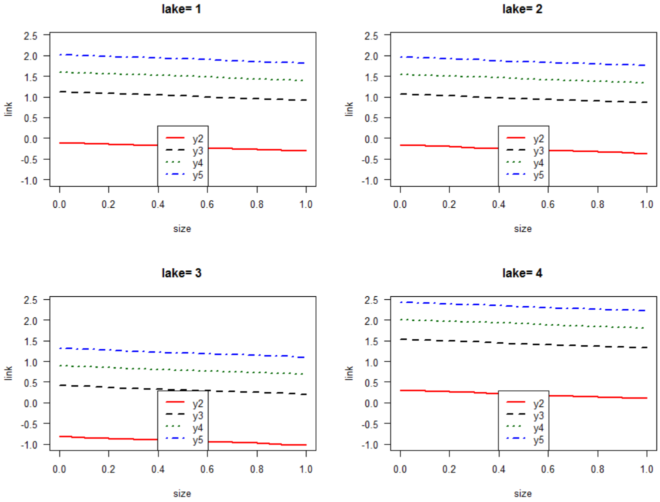


## Plot function

```
plot.alligators <- function(type = "link", model) {
    xx <- data.frame(lake = rep(as.factor(1:4), each = 11), size = rep(c(0:10)/10, 4))
    pp.xx <- cbind(predict(model, type = type, newdata = xx))
    for (i in unique(alligators$lake)) {
        part = which(xx$lake == i)
        mycol <- c("red", "black", "darkgreen", "blue", "green")
        # ooo <- with(alligators, order(size))
        with(alligators, matplot(xx$size[part], pp.xx[part, ], type = "l", las = 1, lwd = 2,
            ylim = c(min(pp.xx), max(pp.xx)), ylab = type, xlab = "size", main = paste("lake=",
                i), col = mycol))
        if (type == "link") {
            legend("bottom", col = mycol, lty = 1:4, legend = colnames(fit@y)[-1], lwd = 2)
        } else {
            legend("bottom", col = mycol, lty = 1:4, legend = colnames(fit@y), lwd = 2)
        }
    }
}
```

# Plot and Understand Model Specification





# Western Collaborative Group Study data

Data:The Western Collaborative Group Study (WCGS), a prospective cohort studye, recruited middle-aged men (ages 39 to 59) who were employees of 10 California companies and collected data on 3154 individuals during the years 1960-1961. These subjects were primarily selected to study the relationship between behavior pattern and the risk of coronary hearth disease (CHD). A number of other risk factors were also measured.

| variable name | discreption |
|---|---|
| id | Subject ID: |
| age0 | Age: age in years |
| height0 | Height: height in inches |
| weight0 | Weight: weight in pounds |
| sbp0 | Systolic blood pressure: mm Hg |
| dbp0 | Diastolic blood pressure: mm Hg |
| chol0 | Cholesterol: mg/100 ml |
| behpat0 | Behavior pattern: |
| ncigs0 | Smoking: Cigarettes/day |
| dibpat0 | Dichotomous behavior pattern: 0 = Type B; 1 = Type A |
| chd69 | Coronary heart disease event: 0 = none; 1 = yes |
| typechd | to be done |
| time169 | Observation (follow up) time: Days |
| arcus0 | Corneal arcus: 0 = none; 1 = yes |

## Logistic regression for Ordinal outcome

```
wcgs <- read.csv("wcgs.csv")[, -1]
# head(wcgs)
library("VGAM")
```

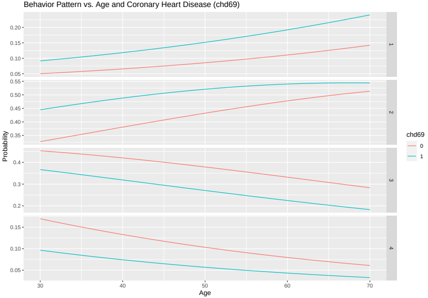
```
## Loading required package: stats4

## Loading required package: splines
```

```
wcgs$behpat1 <- factor(wcgs$behpat0, order = T)
wcgs$chd69 <- as.factor(wcgs$chd69)
model <- vglm(behpat1 ~ age0 + chd69, family = cumulative(parallel = T), data = wcgs)
# summary(model)
beta <- coef(summary(model), matrix = T)
round(beta, 2)
```

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1     -3.80       0.29  -13.04     0.00
## (Intercept):2     -1.36       0.28   -4.81     0.00
## (Intercept):3      0.73       0.28    2.58     0.01
## age0               0.03       0.01    4.70     0.00
## chd691             0.65       0.12    5.20     0.00
```

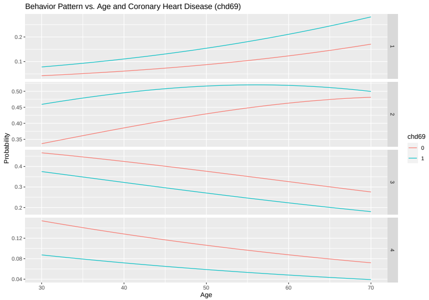
## GGplot Logistic regression for Ordinal outcome

```
xx <- data.frame(chd69 = rep(as.factor(c(0, 1)), each = 41), age0 = rep(c(30:70),2))
pp.xx <- cbind(xx, predict(model, type = "response", newdata =xx)) #
library("reshape")
lpp <- melt(pp.xx, id.vars = c("age0","chd69"), value.name = "probability")

library("ggplot2")
ggplot(lpp, aes(x = age0, y = value, colour = chd69)) + geom_line() + facet_grid(variable ~., scales = "free")+ labs(lines= "Nitrogen",
       x = "Age",
       y = "Probability",
       title = "Behavior Pattern vs. Age and Coronary Heart Disease (chd69)")
```



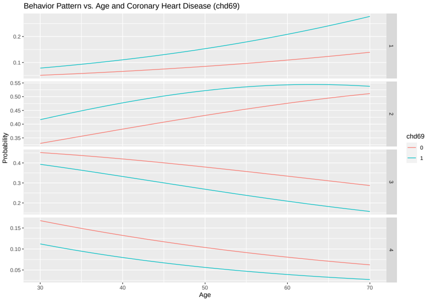
## Relaxing Proportional-odds Assumption

```r
model <- vglm(behpat1 ~ age0 + chd69, family = cumulative(parallel = F ~ age0), data = wcgs)
xx <- data.frame(chd69 = rep(as.factor(c(0, 1)), each = 41), age0 = rep(c(30:70), 2))
pp.xx <- cbind(xx, predict(model, type = "response", newdata = xx))  #
library("reshape")
lpp <- melt(pp.xx, id.vars = c("age0", "chd69"), value.name = "probability")

library("ggplot2")
ggplot(lpp, aes(x = age0, y = value, colour = chd69)) + geom_line() + facet_grid(variable ~
    ., scales = "free") + labs(lines = "Nitrogen", x = "Age", y = "Probability", title = "Behavior Pattern vs. Age and Coronary Heart Disease (chd69)")
```



## Proportional-odds Assumption with interaction

```
model <- vglm(behpat1 ~ age0 + chd69 + age0 * chd69, family = cumulative(parallel = T),
    data = wcgs)
xx <- data.frame(chd69 = rep(as.factor(c(0, 1)), each = 41), age0 = rep(c(30:70), 2))
pp.xx <- cbind(xx, predict(model, type = "response", newdata = xx))  #
library("reshape")
lpp <- melt(pp.xx, id.vars = c("age0", "chd69"), value.name = "probability")

library("ggplot2")
ggplot(lpp, aes(x = age0, y = value, colour = chd69)) + geom_line() + facet_grid(variable ~
    ., scales = "free") + labs(lines = "Nitrogen", x = "Age", y = "Probability", title = "Behavior Pattern vs. Age and Coronary Heart Disease (chd69)")
```



# Cohen's kappa

```
library(psych)
#rating data (with thanks to Tim Bates)
rater1 <- c(1, 2, 3, 4, 5, 6, 7, 8, 9)  # rater one's ratings
rater2 <- c(1, 3, 1, 6, 1, 5, 5, 6, 7)  # rater one's ratings
cohen.kappa(x = cbind(rater1, rater2))
```

```
## Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha, levels = levels)
##
## Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundaries
##                    lower estimate upper
## unweighted kappa -0.18     0.00  0.18
## weighted kappa    0.43     0.68  0.93
##
##  Number of subjects = 9
```

```
# data matrix taken from Cohen
cohen <- matrix(c(0.44, 0.07, 0.09, 0.05, 0.2, 0.05, 0.01, 0.03, 0.06), ncol = 3, byrow = TRUE)

cohen.kappa(cohen, n.obs = 200)
```

```
## Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha, levels = levels)
##
## Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundaries
##                    lower estimate upper
## unweighted kappa  0.39     0.49  0.59
## weighted kappa    0.33     0.45  0.58
##
##  Number of subjects = 200
```

# In class exercise



Use the wcgs data and make ggplot but relax the proportional-odds assumption.

# Take home exercise

- Use the wcgs data and make ggplot as in class with the proportional-odds assumption with interaction but substitute variable chd69 with arcus0.
- Agresti 6.8 and 6.20 using R.

# Class over