Ordinal Regression Models

Bryan Shepherd, PhD

Department of Biostatistics

Vanderbilt University

June 12-16, 2023

Ordered Categorical Variables

Examples of ordinal variables:

1. How helpful is this workshop?
   - Waste of time
   - Poor
   - Fair
   - Good
   - Excellent
   - Outstanding
2. World Health Organization (WHO) Stage of HIV Disease
   - stage 1 (asymptomatic)
   - stage 2 (mildly symptomatic)
   - stage 3 (moderately symptomatic)
   - stage 4 (severely symptomatic)

Ordered Categorical Variables

1. What is stage of cervical intraepithelial neoplastic lesions based on cytology?
   - Normal
   - Atypical squamous cells of undetermined significance (ASCUS)
   - low grade intraepithelial lesions
   - high grade intraepithelial lesions
   - lesions suspicious for cancer

There is a natural ordering to these categories, but we do not want to assign numbers to them and treat them as continuous / interval values.

- For example, assigning the numbers 1-5 and performing analyses implicitly assumes that the difference between level 1 (normal) and level 2 (ASCUS) is the same as the difference between level 2 (ASCUS) and level 3 (low grade), which may not be the case.

Cervical Neoplastic Stage in HIV-infected women

Cervical specimens were collected for 150 non-pregnant HIV-infected women in Lusaka, Zambia (Parham et al., Gynecol Oncol, 2006; 103:1017-22). Variables collected:

Outcome: Cervical stage

Stage of Cervix	n
Normal	9
ASCUS	25
Low grade lesions	34
High grade lesions	49
Lesions suspicious for cancer	28

Predictors: age (years), CD4 cell count (cells/mm^3)

- median age=36, IQR=(31, 41), range=(23, 49)
- median CD4=165, IQR=(86, 299), range=(7, 942)

What is relationship between age, CD4 and cervical lesions?

Binary Logistic Regression

One approach for analyzing this data might be to fit a binary logistic regression model.

Dichotomize the outcome: lesion suspicious for cancer (yes/no, coded as 1/0)

- 28/150 cancerous

Logistic regression

Stage of Cervix Categorization	$\hat{\alpha}$	$\hat{\beta}$	OR	95% CI	p-value
cancerous	-1.13	-0.15	0.86	0.65, 1.13	0.270

estimates are in terms of 100-unit increase in CD4 count

But analyzing the cervical lesions data using binary logistic regression removes a lot of information.

Binary Logistic Regression (Review)

To understand methods for the analysis of ordinal data, it is important to understand logistic regression. Here is a quick review.

Suppose Y has levels 0 or 1, and there is a covariates, Z.

$$\text{logit}[P(Y = 1|Z)] = \alpha + \beta Z.$$

Interpretation of α?

Interpretation of β?

Binary Logistic Regression (Review)

$$\text{logit}[P(Y=1|Z)] = \alpha + \beta Z.$$

This means,

$$\begin{aligned}
&\log\left[\frac{P(Y=1|Z)}{P(Y=0|Z)}\right] = \alpha + \beta Z \\
&\Rightarrow \frac{P(Y=1|Z)}{P(Y=0|Z)} = \exp(\alpha + \beta Z), \\
&\Rightarrow \frac{P(Y=1|Z)}{1 - P(Y=1|Z)} = \exp(\alpha + \beta Z), \\
&\Rightarrow P(Y=1|Z) = \exp(\alpha + \beta Z)(1 - P(Y=1|Z)), \\
&\Rightarrow P(Y=1|Z)(1 + \exp(\alpha + \beta_1 Z)) = \exp(\alpha + \beta Z), \\
&\Rightarrow P(Y=1|Z) = \frac{\exp(\alpha + \beta Z)}{1 + \exp(\alpha + \beta Z)}, \\
&= \textit{expit}(\alpha + \beta Z)
\end{aligned}$$

Binary Logistic Regression (Review)

$$\frac{P(Y=1|Z)}{1-P(Y=1|Z)} = \exp(\alpha + \beta Z)$$

Thus,

$$\begin{aligned}\exp(\alpha) &= \frac{P(Y=1|Z=0)}{1-P(Y=1|Z=0)} \\ &= \text{odds when } Z=0.\end{aligned}$$

And,

$$\begin{aligned}\exp(\beta) &= \frac{\exp[\alpha+\beta(Z+1)]}{\exp[\alpha+\beta Z]} \\ &= \frac{P(Y=1|Z)/\left[1-P(Y=1|Z)\right]}{P(Y=1|Z)/\left[1-P(Y=1|Z)\right]} \\ &= \text{odds ratio for 1-unit increase in } Z.\end{aligned}$$

Multinomial Logistic Regression

Consider performing a series of logistic regressions, where each model has a common reference outcome level and you only fit the model with one of the other outcome levels – everyone else is excluded.

- e.g., only include those with $Y =$ ASCUS or normal, code as 1 and 0, respectively, and fit logistic regression
- repeat, but only include those with $Y =$ low or normal, code as 1 and 0, respectively, and fit logistic regression
- ...

Stage of Cervix Categorization	$\hat{\alpha}$	$\hat{\beta}$	OR	95% CI	p-value
ASCUS vs. normal	1.38	-0.14	0.87	0.62, 1.23	0.427
low vs. normal	1.56	-0.09	0.92	0.66, 1.27	0.605
high vs. normal	2.40	-0.31	0.73	0.51, 1.05	0.092
cancerous vs. normal	1.73	-0.27	0.77	0.53, 1.10	0.151

estimates are in terms of 100-unit increase in CD4 count

Multinomial Logistic Regression

Suppose there are $k = 1, \ldots, K$ categories.

Pick a reference category (e.g., $k = 1$)

$$\log \frac{P(Y = 2|Z)}{P(Y = 1|Z)} = \alpha_2 + \beta_1 Z$$
$$\log \frac{P(Y = 3|Z)}{P(Y = 1|Z)} = \alpha_3 + \beta_2 Z$$
$$\ldots$$
$$\log \frac{P(Y = K|Z)}{P(Y = 1|Z)} = \alpha_K + \beta_K Z$$

$$P(Y = j|Z) = \frac{\exp[\alpha_j + \beta_j Z]}{1 + \sum_{k=2}^{K} \exp[\alpha_k + \beta_k Z]}$$

Multinomial Logistic Regression

Fitting multinomial logistic regression model (`multinom` function of R package `nnet`).

(These are slightly different estimates than the brute force series of logistic regression approach I did before.)

Stage of Cervix Categorization	$\hat{\alpha}$	$\hat{\beta}$	OR	95% CI	p-value
ASCUS vs. normal	1.42	-0.16	0.86	0.60, 1.23	0.397
low vs. normal	1.57	-0.09	0.91	0.66, 1.27	0.593
high vs. normal	2.42	-0.32	0.72	0.51, 1.03	0.074
cancerous vs. normal	1.89	-0.34	0.71	0.48, 1.06	0.094

estimates are in terms of 100-unit increase in CD4 count

Interpretation?

Logistic Regression with Ordered Dichotomizations

Multinomial logistic regression ignored the order information in outcome and resulted in a lot of parameters.

Different way of dichotomizing outcomes that maintains the order

- normal **vs** (ASCUS, low, high, cancerous)
- (normal, ASCUS) **vs** (low, high, cancerous)
- (normal, ASCUS, low) **vs** (high, cancerous)
- (normal, ASCUS, low, high) **vs** cancerous

Logistic Regression after Dichotomizing Cervix Stages

Stage of Cervix Categorization	$\hat{\alpha}$	$\hat{\beta}$	OR	95% CI	p-value
ASCUS, low, high, cancerous	3.23	-0.21	0.81	0.60, 1.09	0.169
low, high, cancerous	1.46	-0.13	0.88	0.72, 1.08	0.227
high, cancerous	0.59	-0.23	0.80	0.65, 0.97	0.027
cancerous	-1.13	-0.15	0.86	0.65, 1.13	0.270

estimates are in terms of 100-unit increase in CD4 count

Ordinal Logistic Regression

Logistic Regression after Dichotomizing Cervix Stages

Stage of Cervix Categorization	$\hat{\alpha}$	$\hat{\beta}$	OR	95% CI	p-value
ASCUS, low, high, cancerous	3.23	-0.21	0.81	0.60, 1.09	0.169
low, high, cancerous	1.46	-0.13	0.88	0.72, 1.08	0.227
high, cancerous	0.59	-0.23	0.80	0.65, 0.97	0.027
cancerous	-1.13	-0.15	0.86	0.65, 1.13	0.270

estimates are in terms of 100-unit increase in CD4 count

Ordered Logistic Regression

No Cervix Stage Categorization	$\hat{\beta}$	OR	95% CI	p-value
Ordered logistic regression	-0.19	0.83	0.70, 0.98	0.030
$\alpha_j = 3.16, 1.60, 0.52, -1.07$				

Ordered Logistic Regression

Suppose Y has levels $k = 1, \ldots, K$, where $1 < \cdots < K$.

For $j = 2, \cdots, K$,

$$\text{logit}[P(Y \geq j|Z)] = \alpha_j + \beta Z.$$

- For each level j, it is a logistic regression model with intercept α_j and coefficient β.
- Proportional odds assumption.

Proportional Odds Assumption

Assumed model:

$$\text{logit}[P(Y \geq j|Z)] = \alpha_j + \beta Z.$$

More flexible (multinomial) model:

$$\begin{aligned}\text{logit}[P(Y \geq 1|Z)] &= \alpha_1 + \beta_1 Z. \\ &\dots \\ \text{logit}[P(Y \geq K-1|Z)] &= \alpha_{K-1} + \beta_{K-1} Z.\end{aligned}$$

Proportional Odds Assumes $\beta_j = \beta$ for all $j \in \{1, \dots K-1\}$.

Odds

$$\frac{P(Y \geq j|Z)}{P(Y < j|Z)} = \exp[\alpha_j + \beta_j Z]$$

Proportional Odds

$$\begin{aligned}\frac{P(Y \geq j|Z)}{P(Y < j|Z)} \Big/ \frac{P(Y \geq i|Z)}{P(Y < i|Z)} &= \frac{\exp[\alpha_j + \beta_j Z]}{\exp[\alpha_i + \beta_i Z]} \\ &= \frac{\exp[\alpha_j]}{\exp[\alpha_i]}.\end{aligned}$$

Ordered Logistic Regression

$$\text{logit}[P(Y \geq j|Z)] = \alpha_j + \beta Z.$$

Interpretation:

- $exp(\beta)$ has an odds ratio interpretation.
- For a 1-unit increase in Z, the odds of being in a higher category of Y increase $exp(\beta)$.

Cumulative "Link" Regression Models

The logit function, is often referred to as a link function.

There are other link functions, G, including probit and complementary log-log.

This general class of models is sometimes referred to as cumulative link models:

$$G[P(Y \geq j|Z)] = \alpha_j + \beta Z.$$

Interpretation of β parameters changes depends on the link function that is used.

We will focus mostly on the logit link function (ordered logistic regression) in this workshop, but it is good to know that there are other possible link functions.

Cervical Lesions Example

```
d1<-read.csv("data/Zambia dataset with condom variables.csv")
d1$age<-d1$AgeRAW
d1$cd4<-d1$CD4RAW
d1$y<-d1$Cytology
d1$exclude<-ifelse(is.na(d1$cd4),1,0)
d<-d1[d1$exclude==0, c("y","age","cd4")]

table(d$y)
```

```
##
##  1  2  3  4  5
##  9 25 34 49 28
```

```
summary(d$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.00   31.00   36.00   36.19   41.00   49.00
```

```
summary(d$cd4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       7      86     165     208     299     942
```

Ordered Logistic Regression

Like all regression models, easily handles more than one covariate.

$$\text{logit}[P(Y \geq j|Z_1, Z_2)] = \alpha_j + \beta_1 Z_1 + \beta_2 Z_2.$$

Interpretation:

- $exp(\beta)$ has an odds ratio interpretation.
- Holding Z_2 constant, for a 1-unit increase in Z_1, the odds of being in a higher category of Y increase $exp(\beta_1)$.

Cervical Lesions Example

```
library(rms)
dd<-with(d, datadist(age,cd4,y))
options(datadist='dd')
mod1<-lrm(y~cd4, data=d, x=TRUE, y=TRUE)
mod1
```

```
## Logistic Regression Model
##
##  lrm(formula = y ~ cd4, data = d, x = TRUE, y = TRUE)
##
##
##  Frequencies of Responses
##
##   1  2  3  4  5
##   9 25 34 49 28
##
##                         Model Likelihood      Discrimination    Rank Discrim.
##                               Ratio Test             Indexes          Indexes
##  Obs            145    LR chi2      4.71      R2       0.034    C       0.562
##  max |deriv| 2e-05    d.f.            1      R2(1,145)0.025    Dxy     0.123
##                       Pr(> chi2) 0.0299    R2(1,135.7)0.027    gamma   0.123
##                                              Brier    0.178    tau-a   0.094
##
##       Coef    S.E.   Wald Z Pr(>|Z|)
##  y>=2  3.1552 0.4070  7.75  <0.0001
##  y>=3  1.6028 0.2806  5.71  <0.0001
##  y>=4  0.5166 0.2476  2.09  0.0369
##  y>=5 -1.0733 0.2655 -4.04  <0.0001
##  cd4  -0.0019 0.0009 -2.17  0.0304
```

Cervical Lesions Example

```
mod2<-lrm(y~cd4+age, data=d, x=TRUE, y=TRUE)
mod2
```

```
## Logistic Regression Model
##
##  lrm(formula = y ~ cd4 + age, data = d, x = TRUE, y = TRUE)
##
##
##  Frequencies of Responses
##
##   1  2  3  4  5
##   9 25 34 49 28
##
##                         Model Likelihood      Discrimination    Rank Discrim.
##                               Ratio Test             Indexes          Indexes
##  Obs            145    LR chi2      6.52     R2       0.046    C        0.569
##  max |deriv| 5e-12    d.f.            2     R2(2,145)0.031    Dxy      0.137
##                       Pr(> chi2) 0.0383    R2(2,135.7)0.033    gamma    0.137
##                                             Brier    0.176    tau-a    0.105
##
##       Coef    S.E.   Wald Z Pr(>|Z|)
##  y>=2  1.9441 0.9828  1.98  0.0479
##  y>=3  0.3826 0.9458  0.40  0.6858
##  y>=4 -0.7121 0.9449 -0.75  0.4511
##  y>=5 -2.3137 0.9623 -2.40  0.0162
##  cd4  -0.0016 0.0009 -1.87  0.0611
##  age   0.0328 0.0244  1.34  0.1790
##
```

Lesions Example – other software (for the advanced learner)

```
library(MASS)
d$y1<-with(d, factor(y))
mod3<-polr(y1~cd4+age, data=d)
mod3
```

```
## Call:
## polr(formula = y1 ~ cd4 + age, data = d)
##
## Coefficients:
##          cd4          age
## -0.001640688  0.032833647
##
## Intercepts:
##        1|2        2|3        3|4        4|5
## -1.9440489 -0.3825956  0.7120437  2.3136700
##
## Residual Deviance: 428.4398
## AIC: 440.4398
```

Wait? Aren't those different answers from `lrm`?

`lrm` vs. `polr` (for the advanced learner)

- `lrm` in `rms` package:

$$logit[P(Y \geq k|Z)] = \alpha_{k-1} + \beta Z \text{ for } k = 2, \ldots, K$$

- `polr` in `MASS` package:

$$logit[P(Y \leq k|Z)] = \alpha_k^* - \beta^* Z \text{ for } k = 1, \ldots, K-1$$

- Note that `rms` is parameterized with exceedance probabilities, whereas `polr` is parameterized with cumulative probabilities.
- Beta coefficients are the same (because `polr` subtracts β^*.

$$\beta = \beta^*$$

- Alpha coefficients have opposite sign.

$$\alpha_k = -\alpha_k^*$$

Checking Proportional Odds Assumption

- Compare proportional odds model:
  - $\text{logit}[P(Y \geq j|Z_1, Z_2)] = \alpha_j + \beta_1 Z_1 + \beta_2 Z_2$
- with multinomial model (equivalent):
  - $\text{logit}[P(Y \geq j|Z_1, Z_2)] = \alpha_j + \beta_{j,1} Z_1 + \beta_{j,2} Z_2$
- Smaller AIC (Deviance + 2df) is preferable
- Can do likelihood ratio test also
- Using `impactPO` function in `rms` library
- Brant test
  - `brant` function of `brant` package
  - However, apparently not very reliable
    - (personal communication with Frank Harrell)
- Diagnostic tests reject with big N, but model fit may be OK.

```
impactPO(y~age+cd4, data=d, newdata=d[1,])
```

```
##                              PO     Multinomial
## Deviance                     428.44 426.73
## d.f.                          6     12
## AIC                          440.44 450.73
## p                            2      8
## LR chi^2                     6.52   8.23
## LR - p                       4.52   0.23
## LR chi^2 test for PO                1.71
##   d.f.                              6
##   Pr(>chi^2)                        0.9444
## MCS R2                       0.044  0.055
## MCS R2 adj                   0.031  0.002
## McFadden R2                  0.015  0.019
## McFadden R2 adj               0.006 -0.018
## Mean |difference| from PO           0.015
##
##          method y age cd4 y1 Probability id
## 1.1           PO 1  28 386  3      0.0971  1
## 2.1 Multinomial 1  28 386  3      0.1040  2
## 1.2           PO 2  28 386  3      0.2417  1
## 2.2 Multinomial 2  28 386  3      0.2121  2
## 1.3           PO 3  28 386  3      0.2661  1
## 2.3 Multinomial 3  28 386  3      0.2977  2
## 1.4           PO 4  28 386  3      0.2787  1
## 2.4 Multinomial 4  28 386  3      0.2749  2
## 1.5           PO 5  28 386  3      0.1163  1
## 2.5 Multinomial 5  28 386  3      0.1113  2
##
## Covariate combination-specific mean |difference| in pr
```

Partial Proportional Odds Models

Suppose proportional odds model was a poor fit. It is possible to relax the proportional odds assumption for a specific variable or set of variables.

- Proportional odds model:
  - $\text{logit}[P(Y \geq j|Z_1, Z_2)] = \alpha_j + \beta_1 Z_1 + \beta_2 Z_2$
- Multinomial model:
  - $\text{logit}[P(Y \geq j|Z_1, Z_2)] = \alpha_j + \beta_{j,1} Z_1 + \beta_{j,2} Z_2$
- Partial proportional odds model:
  - $\text{logit}[P(Y \geq j|Z_1, Z_2)] = \alpha_j + \beta_1 Z_1 + \beta_{j,2} Z_2$

Notice that the partial proportional odds model is a compromise between the multinomial model and the proportional odds model:

- It assumes a common β_1 for Z_1 (similar to proportional odds model)
- It allows β_2 to vary by the category (similar to a multinomial model)

Fitting Partial Proportional Odds Models in R

This code relaxes the proportional odds assumption for the CD4 variable using the `vglm` function in the `VGAM` library.

```
d$y1<-ordered(d$y)
mod6<-vglm(y1~age+cd4, family=cumulative(parallel=FALSE~cd4), data=d)
summary(mod6)
```

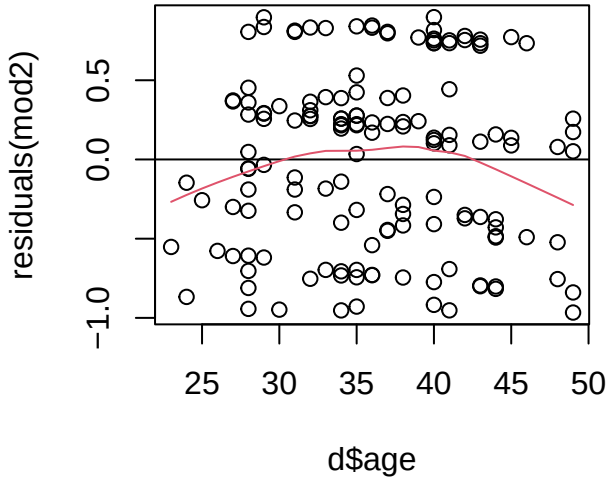
```
##
## Call:
## vglm(formula = y1 ~ age + cd4, family = cumulative(parallel = FALSE ~
##     cd4), data = d)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -1.946219   1.060092  -1.836   0.0664 .
## (Intercept):2 -0.182202   0.963621  -0.189   0.8500
## (Intercept):3  0.652515   0.955098   0.683   0.4945
## (Intercept):4  2.399998   0.987956   2.429   0.0151 *
## age           -0.034269   0.024523  -1.397   0.1623
## cd4:1          0.001818   0.001529   1.189   0.2344
## cd4:2          0.001009   0.001051   0.960   0.3369
## cd4:3          0.002222   0.001038   2.141   0.0322 *
## cd4:4          0.001446   0.001427   1.013   0.3108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3]), logitlink(P[Y<=4])
##
## Residual deviance: 426.5921 on 571 degrees of freedom
```

Proportional Odds Models (Continued)

Since the proportional odds assumption was reasonable for the dataset, we will assume it for the remainder of these slides.

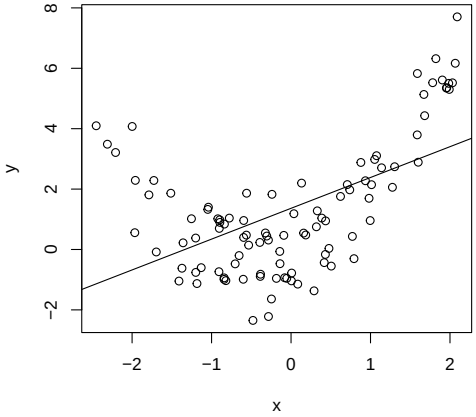
Checking covariate functional form with probability-scale residuals

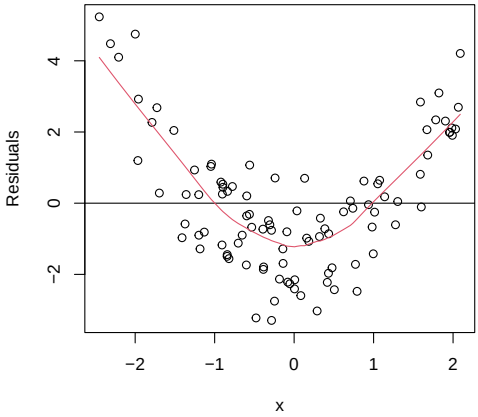
```
plot(d$age, residuals(mod2))
abline(h=0)
lines(lowess(d$age, residuals(mod2)), col=2)
```



A Detour Discussiing Residuals (for the advanced learner)

Linear regression residuals from a toy data analysis





Probability-Scale Residual (for the advanced learner)

Observed-minus-expected residual (OMER):

$$y - \hat{y},$$

where y is an observed value and $\hat{y}$ is a 'fitted value,' typically defined as the estimated conditional expectation.

Let Y^* be a random variable from the fitted distribution.

$$OMER = E(y - Y^*) = y - E(Y^*) = y - \hat{y}.$$

Probability-scale residual (PSR):

$$PSR = E(sign(y, Y^*)) = P(Y^* < y) - P(Y^* > y),$$

where sign(a, b)=1 if $a > b$, -1 if $a < b$, or 0 if $a = b$.

Probability-Scale Residual (for the advanced learner) (Continued)

Fit a model of Y on $\boldsymbol{Z}$ with parameter estimate $\hat{\boldsymbol{\theta}}$.

Example:

	normal	ASCUS	low grade	high grade	cancer
$P(Y^* = y \mid \boldsymbol{Z} = \boldsymbol{z}; \hat{\boldsymbol{\theta}})$	0.3	0.2	0.2	0.2	0.1

If y= 'ASCUS', then

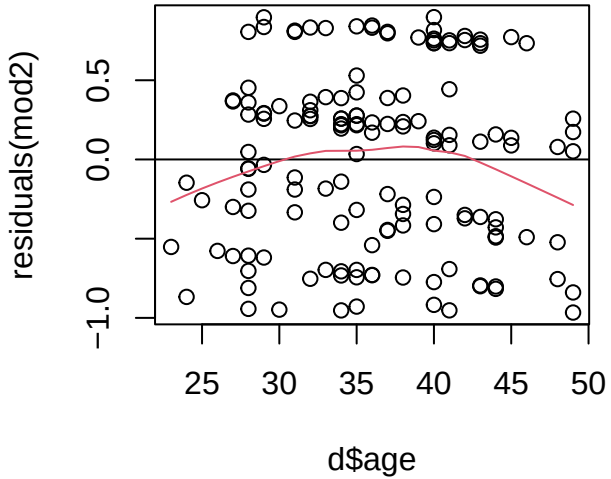
$$\begin{aligned} r(y, F^*) &= P(Y^* < y) - P(Y^* > y) \\ &= 0.3 - 0.2 - 0.2 - 0.1 \\ &= -0.2. \end{aligned}$$

If y= 'cancer', then

$$\begin{aligned} r(y, F^*) &= 0.3 + 0.2 + 0.2 + 0.2 \\ &= 0.9. \end{aligned}$$

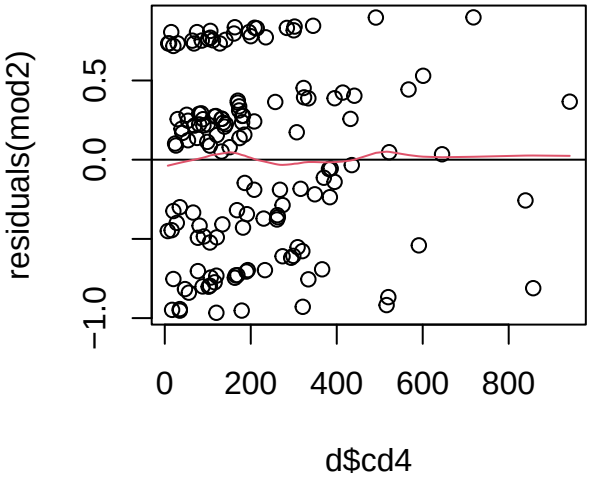
Checking covariate functional form with probability-scale residuals

```
plot(d$age, residuals(mod2))
abline(h=0)
lines(lowess(d$age, residuals(mod2)), col=2)
```



Checking covariate functional form with probability-scale residuals

```
plot(d$cd4, residuals(mod2))
abline(h=0)
lines(lowess(d$cd4, residuals(mod2)), col=2)
```



Re-fitting using Restricted Cubic Splines

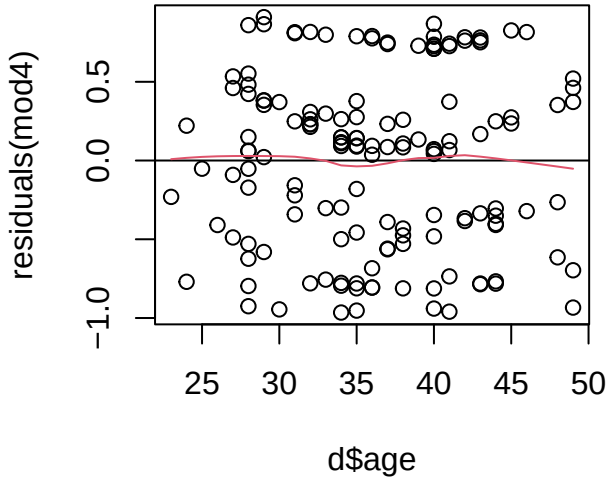
We will use 3 knots because the relationship from the residuals looks to be quadratic.

```
mod4<-lrm(y~cd4+rcs(age,3), data=d, x=TRUE, y=TRUE)
anova(mod4)
```

```
##                 Wald Statistics          Response: y
##
##  Factor     Chi-Square d.f. P
##  cd4          2.49      1    0.1149
##  age          6.92      2    0.0314
##   Nonlinear   5.08      1    0.0242
##  TOTAL       11.61      3    0.0089
```

Checking covariate functional form with probability-scale residuals

```
plot(d$age, residuals(mod4))
abline(h=0)
lines(lowess(d$age, residuals(mod4)), col=2)
```



Model output

```
mod4
```

```
## Logistic Regression Model
##
## lrm(formula = y ~ cd4 + rcs(age, 3), data = d, x = TRUE, y = TRUE)
##
##
## Frequencies of Responses
##
##  1  2  3  4  5
##  9 25 34 49 28
##
##                        Model Likelihood      Discrimination    Rank Discrim.
##                              Ratio Test             Indexes          Indexes
## Obs            145    LR chi2     11.64      R2       0.081    C       0.595
## max |deriv| 2e-09    d.f.            3      R2(3,145)0.058    Dxy     0.190
##                        Pr(> chi2) 0.0087    R2(3,135.7)0.062    gamma   0.190
##                                             Brier    0.172    tau-a   0.145
##
##       Coef    S.E.   Wald Z Pr(>|Z|)
## y>=2 -1.8436 1.9312 -0.95  0.3398
## y>=3 -3.4407 1.9329 -1.78  0.0751
## y>=4 -4.5672 1.9492 -2.34  0.0191
## y>=5 -6.1990 1.9732 -3.14  0.0017
## cd4  -0.0014 0.0009 -1.58  0.1149
## age   0.1574 0.0603  2.61  0.0091
## age' -0.1567 0.0695 -2.25  0.0242
##
```

Interpretation

Compare specific levels of age.

```
summary(mod4, cd4=c(100,200), age=c(35,25))
```

```
##              Effects              Response : y
##
##  Factor      Low High Diff. Effect   S.E.     Lower 0.95 Upper 0.95
##  cd4         100 200  100   -0.13970 0.088606 -0.313360   0.033964
##   Odds Ratio 100 200  100    0.86962       NA  0.730980   1.034500
##  age          35  25  -10   -1.36430 0.519360 -2.382200  -0.346350
##   Odds Ratio  35  25  -10    0.25557       NA  0.092347   0.707270
```

```
summary(mod4, cd4=c(300,200), age=c(35,30))
```

```
##              Effects              Response : y
##
##  Factor      Low High Diff. Effect   S.E.     Lower 0.95 Upper 0.95
##  cd4         300 200  -100   0.13970 0.088606 -0.033964   0.31336
##   Odds Ratio 300 200  -100   1.14990       NA  0.966610   1.36800
##  age          35  30    -5  -0.58207 0.221410 -1.016000  -0.14811
##   Odds Ratio  35  30    -5   0.55874       NA  0.362030   0.86233
```

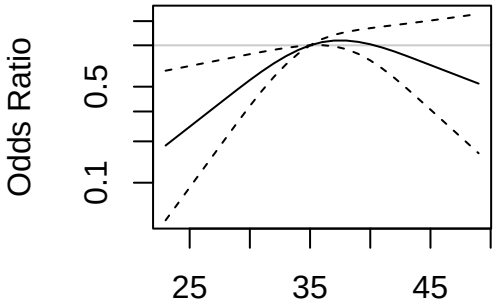
Interpretation

A table that I might include in a biomedical manuscript:

	Odds ratio	95% CI	p-value
CD4 (per 100 cells/mm^3)	0.87	(0.73, 1.03)	0.11
Age (years)			0.031
25	0.26	(0.09, 0.71)	
30	0.56	(0.36, 0.86)	
35 (reference)	1		
40	1.02	(0.78, 1.33)	
45	0.72	(0.34, 1.51)	

Interpretation

```
ages<-c(min(d$age):max(d$age)); ests<-matrix(NA,length(ages),3)
for(i in 1:length(ages)){
  ests[i,]<-summary(mod4, age=c(35, ages[i]))[4,c(4,6,7)]
}
plot(rep(ages,3),log(c(ests[,1],ests[,2],ests[,3])),xlab="Age", ylab="Odds Ratio", axes=FALSE, type="n")
abline(h=0,lwd=1,col=gray(.8)); axis(1); box()
axis(2, at=log(c(1.5, 1, .5, .33, .2, .1)), labels=c("1.5", "1", "0.5", "0.33", "0.2", "0.1"))
lines(ages,log(ests[,1])); lines(ages,log(ests[,2]), lty=2); lines(ages,log(ests[,3]), lty=2)
```



Predicted probabilities

```
newdata=data.frame(cd4=165,age=ages)
pred.probs<-predict(mod4, type="fitted", se.fit=TRUE, newdata=newdata)
```

```
## Warning in predict.lrm(mod4, type = "fitted", se.fit = TRUE, newdata = newdata):
## se.fit not supported with type="fitted" or type="mean"
```

```
cbind(ages, pred.probs)
```

```
##    ages      y>=2      y>=3      y>=4       y>=5
## 1    23 0.8244176 0.4873807 0.2355809 0.05684817
## 2    24 0.8460550 0.5267077 0.2650972 0.06590137
## 3    25 0.8654611 0.5657060 0.2968755 0.07627971
## 4    26 0.8827600 0.6039075 0.3307488 0.08813825
## 5    27 0.8980964 0.6408819 0.3664725 0.10163748
## 6    28 0.9116275 0.6762551 0.4037271 0.11693910
## 7    29 0.9234717 0.7095960 0.4419749 0.13412918
## 8    30 0.9336164 0.7401136 0.4800070 0.15293089
## 9    31 0.9421007 0.7671613 0.5164383 0.17278676
## 10   32 0.9490290 0.7903649 0.5499707 0.19290725
## 11   33 0.9545345 0.8095691 0.5794799 0.21229568
## 12   34 0.9587506 0.8247603 0.6040488 0.22980415
## 13   35 0.9617895 0.8359819 0.6229424 0.24421178
## 14   36 0.9637249 0.8432507 0.6355364 0.25431326
## 15   37 0.9646172 0.8466351 0.6414989 0.25924319
## 16   38 0.9646216 0.8466516 0.6415282 0.25926768
## 17   39 0.9638663 0.8437859 0.6364751 0.25508296
## 18   40 0.9624374 0.8384045 0.6271081 0.24750737
## 19   41 0.9603994 0.8308198 0.6141700 0.23741461
## 20   42 0.9578127 0.8213435 0.5984240 0.22567798
## 21   43 0.9547487 0.8103306 0.5806850 0.21312418
```

Predicted probabilities (continued)

The predicted probability of being in a specific category for given covariate values can be easily computed.

```
newdata=data.frame(cd4=165,age=ages)
pred.probs<-predict(mod4, type="fitted", se.fit=TRUE, newdata=newdata)
```

```
## Warning in predict.lrm(mod4, type = "fitted", se.fit = TRUE, newdata = newdata):
## se.fit not supported with type="fitted" or type="mean"
```

```
head(cbind(ages, pred.probs))
```

```
##   ages      y>=2      y>=3      y>=4       y>=5
## 1   23 0.8244176 0.4873807 0.2355809 0.05684817
## 2   24 0.8460550 0.5267077 0.2650972 0.06590137
## 3   25 0.8654611 0.5657060 0.2968755 0.07627971
## 4   26 0.8827600 0.6039075 0.3307488 0.08813825
## 5   27 0.8980964 0.6408819 0.3664725 0.10163748
## 6   28 0.9116275 0.6762551 0.4037271 0.11693910
```

For example, for a 23 year old woman with CD4=165 cells/mm^3, the predicted probability of being normal is $1 - 0.824 = 0.176$; the predicted probability of having an ASCUS lesion ($y = 2$) is $0.824 - 0.487 = 0.337$; the predicted probability of having cancerous lesion ($y = 5$) is 0.057.

End of Slides