Cumulative Probability Models

Bryan Shepherd, PhD

Department of Biostatistics

Vanderbilt University

June 12-16, 2023

Continuous Response Data

- Common
- If two groups, we often think to use t-test
- Linear models extend t-test to adjust for covariates and/or to include non-binary covariates

Two-sample t-test

Compares mean between two independent groups

Assumes:

- Data are a random sample from a larger population
- Observations are independent between groups
- Data are approximately normally distributed
- The variance between the two groups is similar

Linear regression models

Normal linear model:

$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \epsilon_i$ where $\epsilon \sim^{i.i.d.} N(0, \sigma^2)$.

The expectation of Y given Z_1 and Z_2 is

$E(Y|Z_1, Z_2) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2$

β_0, β_1, and β_2 are estimated using least squares

- Finding the values that minimize $\sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i})]^2$.
- Equivalent to the maximum likelihood estimates from the normal linear model
- Although least squares estimation does not require normality, it performs better as data are closer to the normal linear model with constant variance.

Skewed Data

We often need to transform data prior to fitting linear regression model.

- $H(Y_i) = Y_i^* = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \epsilon_i$ where $\epsilon \sim^{i.i.d.} N(0, \sigma^2)$.
- May be difficult to select transformation.
  - We might try a log-transformation, or a square-root transformation, or if neither of those are good, then we might do a Box-Cox transformation. It is also sometimes challenging to decide what transformation is 'good'. And we tend to only consider a limited choice of transformations.
- May be difficult to interpret results after transforming data.

$E[H(Y|Z)] \neq H[E(Y|Z)]$

Example – CD4:CD8 ratio among people with HIV

- Low CD4:CD8 ratio is a marker of a weak immune system
- Low CD4:CD8 ratio has been associated with higher risks of co-morbidities
- Interest in assessing factors associated with CD4:CD8 ratio at initiation of antiretroviral therapy (ART)
- We will use data from adults starting ART for the first time in Middle Tennessee
- No standard transformation for analyzing CD4:CD8 ratio
    - Transformations in literature:
        - no transformation
        - log-transformed
        - square-root transformed
        - fifth-root transformed
        - dichotomized
        - categorized based on quantiles

Example – CD4:CD8 ratio

```
d<-read.csv("~/Library/CloudStorage/OneDrive-VUMC/data-files/jessie-cd4-cd8/cd4-cd8-small-analysis-datase
dim(d)
```

```
## [1] 2024    8
```

```
head(d)
```

```
##           y black female     age  route hcv hbv year
## 1 1.1818182     0      1 30.6037 Hetero   0   0 1999
## 2 0.2592795     1      0 40.0465    MSM   0   0 2010
## 3 0.1625000     0      0 50.1465    MSM   0   0 1999
## 4 0.8125000     0      0 53.4511    MSM   0   0 1999
## 5 0.2096774     0      0 40.3943    MSM   0   0 1999
## 6 0.6176471     1      0 40.2053    MSM   0   0 2004
```

```
length(unique(d$y))
```

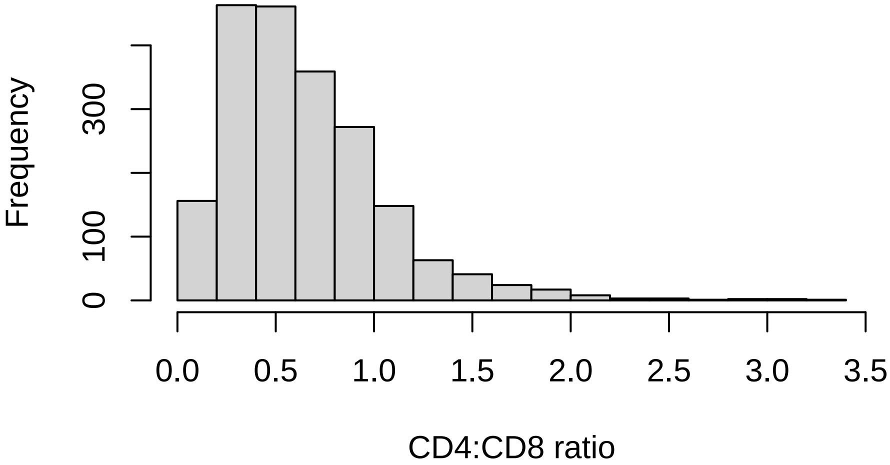
```
## [1] 1859
```

```
summary(d$y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04478 0.35726 0.57018 0.64803 0.84593 3.22222
```

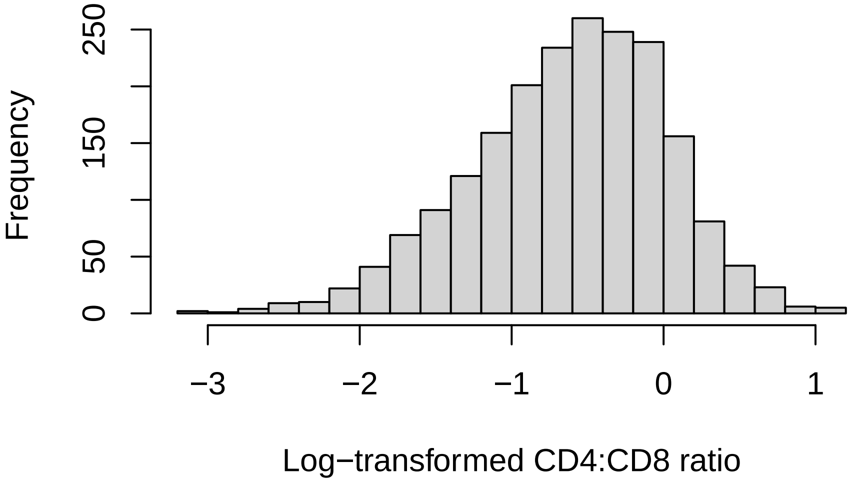
Example – CD4:CD8 ratio skewed

```
hist(d$y, main="", xlab="CD4:CD8 ratio", nclass=20)
```



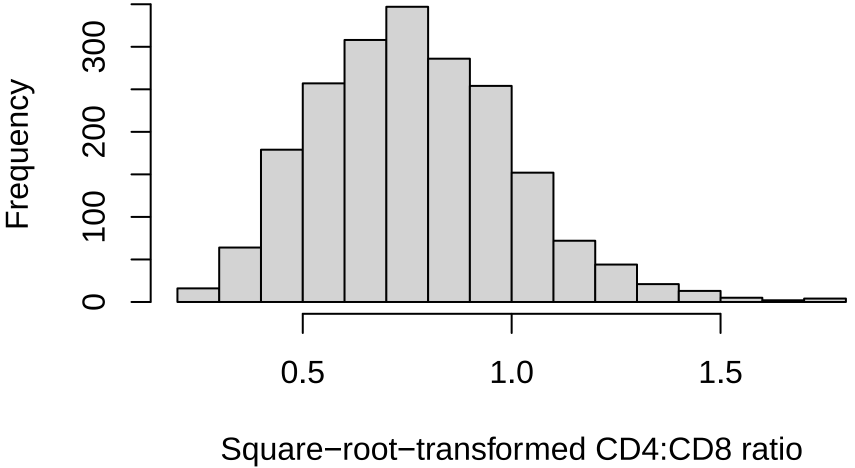
Example – Log-transformed CD4:CD8 ratio

```
hist(log(d$y), main="", xlab="Log-transformed CD4:CD8 ratio", nclass=20)
```



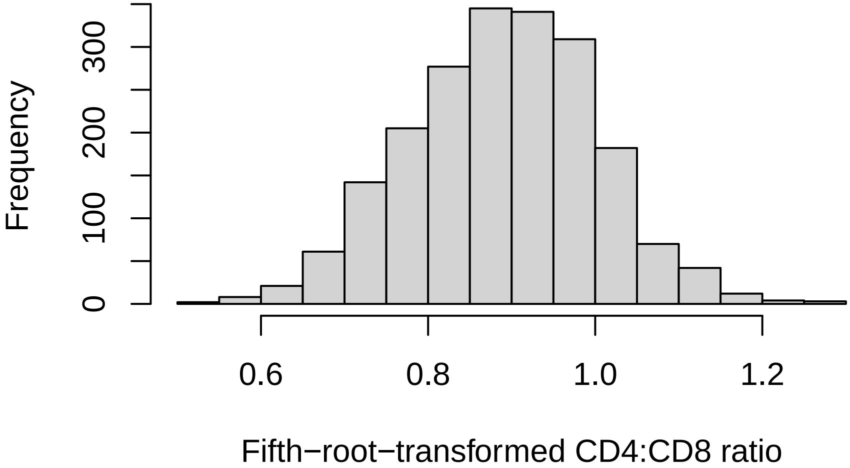
Example – Square-root-transformed CD4:CD8 ratio

```r
hist(sqrt(d$y), main="", xlab="Square-root-transformed CD4:CD8 ratio", nclass=20)
```



Example – Fifth-root-transformed CD4:CD8 ratio

```
hist(d$y^(0.2), main="", xlab="Fifth-root-transformed CD4:CD8 ratio", nclass=20)
```



Transformation can Impact Results

```
fit1<-lm(y~black, data=d)
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ black, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62613 -0.28993 -0.08076  0.19534  2.55132
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.63653    0.01099  57.914   <2e-16 ***
## black        0.03438    0.01900   1.809   0.0706 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4034 on 2022 degrees of freedom
## Multiple R-squared:  0.001616,   Adjusted R-squared:  0.001122
## F-statistic: 3.272 on 1 and 2022 DF,  p-value: 0.0706
```

Transformation can Impact Results

```r
fit2<-lm(log(y)~black, data=d)
summary(fit2)
```

```
##
## Call:
## lm(formula = log(y) ~ black, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.51034 -0.40006  0.05702  0.45322  1.77639
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.63730    0.01750 -36.410   <2e-16 ***
## black        0.04156    0.03026   1.373     0.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6424 on 2022 degrees of freedom
## Multiple R-squared:  0.0009318,  Adjusted R-squared:  0.0004377
## F-statistic: 1.886 on 1 and 2022 DF,  p-value: 0.1698
```

Transformation can Impact Results

```r
fit3<-lm(sqrt(y)~black, data=d)
summary(fit3)
```

```
##
## Call:
## lm(formula = sqrt(y) ~ black, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56939 -0.16887 -0.01578  0.14879  1.01406
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.763360   0.006462 118.124   <2e-16 ***
## black       0.017637   0.011174   1.578    0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2372 on 2022 degrees of freedom
## Multiple R-squared:  0.001231,   Adjusted R-squared:  0.0007367
## F-statistic: 2.491 on 1 and 2022 DF,  p-value: 0.1146
```

Transformation can Impact Results

```
fit4<-lm(y^0.2~black, data=d)
summary(fit4)
```

```
##
## Call:
## lm(formula = y^0.2 ~ black, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35777 -0.07477  0.00302  0.07645  0.36860
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.887404   0.003046 291.353   <2e-16 ***
## black       0.007658   0.005266   1.454    0.146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1118 on 2022 degrees of freedom
## Multiple R-squared:  0.001045,   Adjusted R-squared:  0.0005505
## F-statistic: 2.114 on 1 and 2022 DF,  p-value: 0.1461
```

Interpretation with Transformed Data can be Awkward

- From model fit to the untransformed data, $\hat{\beta} = 0.034 = \hat{E}(Y|Z_1 = 1) - \hat{E}(Y|Z_1 = 0)$ suggests that blacks have CD4:CD8 ratio that is on average 0.034 higher than non-blacks.
    - Easy to understand
- From model fit to fifth-root transformed data, $\hat{\beta} = 0.0077 = \hat{E}(Y^{1/5}|Z_1 = 1) - \hat{E}(Y^{1/5}|Z_1 = 0)$ suggests that blacks have fifth-root transformed CD4:CD8 ratio that is on average 0.0077 higher than non-blacks.
    - What does that mean? I have a hard time thinking on the fifth-root scale.
    - And we cannot simply back-transform the data
        - $E(Y^{1/5}|Z_1 = 1)^5 \neq E(Y|Z_1 = 1)$ because $E(Y^{1/5}|Z_1 = 1) \neq E(Y|Z_1 = 1)^{1/5}$

T-test

- Because black race is a dichotomous covariate, we could simply do a t-test and we will get very similar results to the linear model.
- The difference between means is equal to the linear model beta estimate with untransformed CD4:CD8 ratio.
- P-values are similar (0.083 vs. 0.071)

```
### t-test on original scale
with(d, t.test(y~black))
```

```
##
##  Welch Two Sample t-test
##
## data:  y by black
## t = -1.7301, df = 1204.9, p-value = 0.08386
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.073361833  0.004605746
## sample estimates:
## mean in group 0 mean in group 1
##       0.6365280       0.6709061
```

T-test

- Welch's t-test assumes (previous slide) unequal variances between blacks and non-blacks.
- If assume equal variances (not recommended), then we will get identical p-values to the linear model estimate (0.071).

```
### t-test on original scale with equal variances
with(d, t.test(y~black, var.equal=TRUE))
```

```
##
##  Two Sample t-test
##
## data:  y by black
## t = -1.809, df = 2022, p-value = 0.0706
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.071647526  0.002891439
## sample estimates:
## mean in group 0 mean in group 1
##       0.6365280       0.6709061
```

T-test

```r
### t-test on fifth-root transformed scale
with(d, t.test(y^0.2~black))
```

```
##
##  Welch Two Sample t-test
##
## data:  y^0.2 by black
## t = -1.4392, df = 1318, p-value = 0.1503
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.018095185  0.002780067
## sample estimates:
## mean in group 0 mean in group 1
##       0.8874043       0.8950619
```

- Challenges with interpretation on this scale are similar with the t-test as they were with the linear model.
- e.g., $0.895^5 = 0.574 \neq 0.671 = \hat{E}(Y|Z = 1)$

Wilcoxon rank sum test (also known as Mann-Whitney U test)

- Rather than fit a t-test, which requires transforming data so that they are approximately normal with similar variances between groups, I typically prefer to perform a rank-based test.
- Wilcoxon rank sum test
  - Nonparametric test of the null hypothesis that for randomly selected values of Y_{black} and $Y_{nonblack}$ from two populations, the probability of Y_{black} being greater than $Y_{nonblack}$ is equal to the probability of $Y_{nonblack}$ being greater than Y_{black}.
    - Think of Y_{black} being the CD4:CD8 ratio among blacks and $Y_{nonblack}$ being the CD4:CD8 ratio among non-blacks.
  - This test is based on ranks, so it is invariant to a monotonic transformation of the data
    - In other words, you will get the same answer if you do not transform, log, square-root, or fifth-root transform the data
    - This is a nice property
    - This means I do not need to worry about transforming data

Wilcoxon rank sum test – CD4:CD8 data

```
### Wilcoxon rank sum test on original scale
with(d, wilcox.test(y ~ black))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  y by black
## W = 442962, p-value = 0.2948
## alternative hypothesis: true location shift is not equal to 0
```

```
### Wilcoxon rank sum test on fifth-root transformed scale
with(d, wilcox.test(y^(1/2) ~ black))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  y^(1/2) by black
## W = 442962, p-value = 0.2948
## alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon rank sum test

- Results in a p-value, but we often want something more
- Not a regression model
  - Cannot account for multiple covariates
- The Wilcoxon rank sum test is very closely related to the score test for β from ordered logistic regression
  - Ranked data can be thought of as ordered data
  - This is a direction for extending the rank sum test to account for multiple covariates (as will be seen later)

Logistic regression

We could dichotomize our skewed response data and analyze it with logistic regression

- Dichotomizing continuous data is a bad idea that we do not recommend!
- However, for sake of illustration, we are going to dichotomize.
- Logistic regression makes almost no assumptions on the outcome (only that it is binary), so some people dichotomize difficult continuous data. Some people also like the simple interpretation.
- Such a procedure results in a lot of information loss (as will be seen).

Logistic Regression with Dichotomized CD4:CD8 Ratio

```
### Dichotomizing at y<1 or y>=1 because 1 is used to denote healthy CD4:CD8 ratio in people without HIV
d$y2<-with(d,ifelse(y>=1,1,0))
table(d$y2)
```

```
##
##    0    1
## 1703  321
```

```
mod2<-lrm(y2~black, data=d)
mod2$coeff
```

```
##   Intercept       black
## -1.68341047  0.04353153
```

```
anova(mod2)
```

```
##                 Wald Statistics          Response: y2
##
##  Factor     Chi-Square d.f. P
##  black      0.12       1    0.7345
##  TOTAL      0.12       1    0.7345
```

Latent Variable Interpretation

The logistic regression model,

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

can alternatively be parameterized as

$$\text{logit}[P(Y = 0|Z)] = \alpha^* - \beta Z,$$

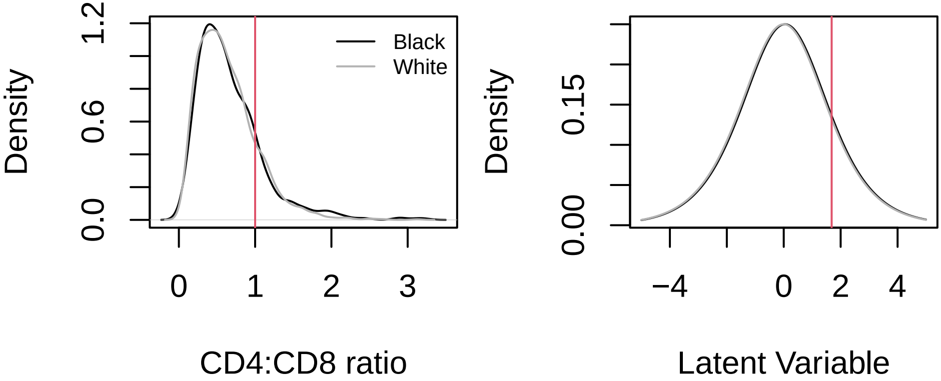
where $\alpha^* = -\alpha$.

This is equivalent to a latent variable model,

$$Y^* = \beta Z + \epsilon, \text{ where } \epsilon \sim \text{standard logistic distribution and } Y = 1 \text{ if } Y^* > \alpha^*.$$

Latent Variable Logistic Distribution, CD4:CD8 Ratio

```
par(mfrow=c(1,2),mar=c(4,4,.5,.5))
plot(density(d$y[d$black==1]), xlab="CD4:CD8 ratio", main=""); lines(density(d$y[d$black==0]),col="gray7
legend(x="topright",legend=c("Black", "White"),lty=c(1,1), col=c(1,"gray70"), bty="n",cex=.65)
yvals<-c(-500:500)/100; fy0<-dlogis(yvals,0); fy1<-dlogis(yvals,mod2$coeff[2])
plot(yvals,fy0, type="n",xlab="Latent Variable", ylab="Density")
lines(yvals,fy1,col=1); lines(yvals,fy0,col="gray70"); abline(v=-mod2$coeff[1], col=2)
```



Latent Variable Logistic Distribution, CD4:CD8 Ratio

```
mod2$coeff
```

```
##   Intercept       black
## -1.68341047  0.04353153
```

```
with(d, table(black, y2))
```

```
##      y2
## black    0    1
##     0 1136  211
##     1  567  110
```

```
### Probability of CD4:CD8>1 if white race
211/(1136+211)
```

```
## [1] 0.1566444
```

```
1-plogis(-mod2$coeff[1])
```

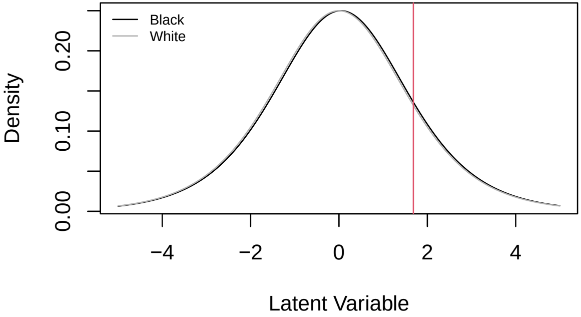
```
## Intercept
## 0.1566444
```

```
### Probability of CD4:CD8>1 if black race
110/(567+110)
```

```
## [1] 0.1624815
```

```
1-plogis(-mod2$coeff[1],mod2$coeff[2])
```

```
## Intercept
## 0.1624815
```



Ordinal Logistic Regression Latent Variable Interpretation

An ordinal logistic regression model can be written as the following:

$$\text{logit}[P(Y \leq j|Z)] = \alpha_j - \beta Z,$$

for $j = 1, \ldots, K-1$ (which is how `polr` in `MASS` library formulates the model).

This is equivalent to a latent variable model,

$$Y^* = \beta Z + \epsilon, \text{ where } \epsilon \sim \text{standard logistic distribution and}$$

$$Y = \begin{cases} 1 \text{ if } Y^* \leq \alpha_1 \\ 2 \text{ if } \alpha_1 < Y^* \leq \alpha_2 \\ \ldots \\ K-1 \text{ if } \alpha_{K-2} < Y^* \leq \alpha_{K-1} \\ K \text{ if } Y > \alpha_{K-1}. \end{cases}$$

Ordered Logistic Regression with 3 Quantiles of CD4:CD8 Ratio

```
quants<-with(d,quantile(y,c(.33,.67)))
d$y3<-with(d,ifelse(y<quants[1],1,ifelse(y<quants[2],2,3)))
fit3<-polr(factor(y3)~black, data=d)
fit3
```

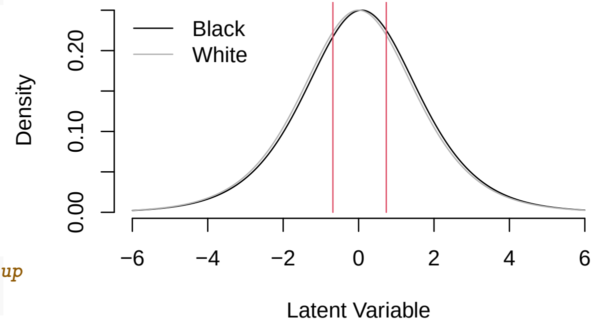
```
## Call:
## polr(formula = factor(y3) ~ black, data = d)
##
## Coefficients:
##      black
## 0.08045377
##
## Intercepts:
##        1|2        2|3
## -0.6814328  0.7350688
##
## Residual Deviance: 4445.926
## AIC: 4451.926
```

```
## Estimated probability CD4:CD8 ratio in lowest group
plogis(fit3$zeta["1|2"]) ## white race
```

```
##       1|2
## 0.3359416
```

```
plogis(fit3$zeta["1|2"],fit3$coefficients["black"]) ## black race
```

```
##       1|2
## 0.3182368
```



Ordered Logistic Regression with 3 Quantiles of CD4:CD8 Ratio

```
with(d, table(black,y3))
```

```
##      y3
## black   1   2   3
##     0 451 461 435
##     1 217 227 233
```

```
## Estimated probability that person will have
## CD4:CD8 ratio in first category (raw data)
451/(451+461+435) ## white race
```

```
## [1] 0.3348181
```

```
217/(217+227+233) ## black race
```

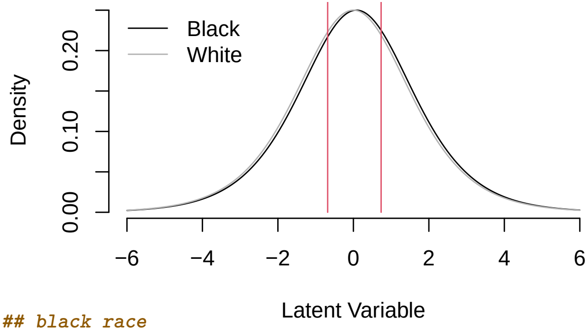
```
## [1] 0.3205318
```

```
## Estimated probability that person will have
## CD4:CD8 ratio in first category (model)
plogis(fit3$zeta["1|2"]) ## white race
```

```
##       1|2
## 0.3359416
```

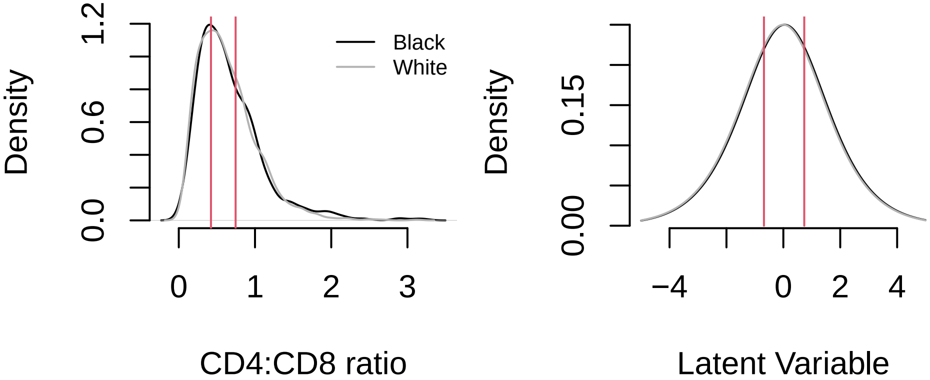
```
plogis(fit3$zeta["1|2"],fit3$coefficients["black"]) ## black race
```

```
##       1|2
## 0.3182368
```



Close, but not identical because ordered logistic regression assumes proportional odds.

Transformation of CD4:CD8 Ratio to Latent Logistic Distribution

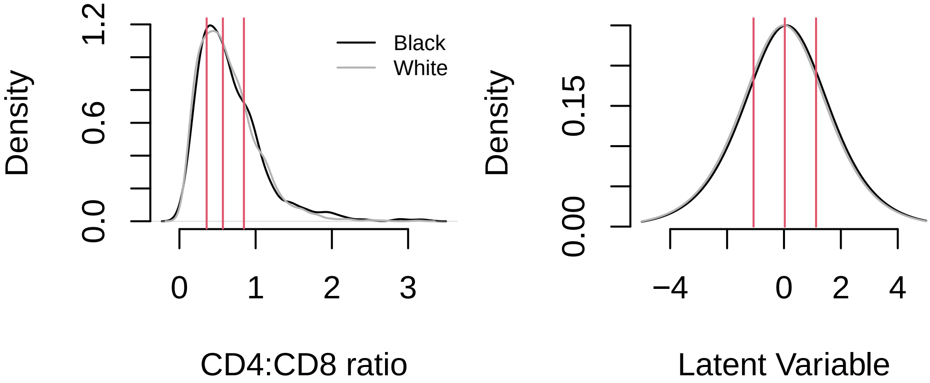


βZ shifts the location of the curve on the latent variable scale

Ordered Logistic Regression with 4 Quantiles of CD4:CD8 Ratio

```
quants<-with(d,quantile(y,c(.25,.5,.75)))
d$y4<-with(d,ifelse(y<quants[1],1,ifelse(y<quants[2],2,ifelse(y<quants[3],3,4))))
fit4<-polr(factor(y4)~black, data=d)
fit4$coeff["black"]
```

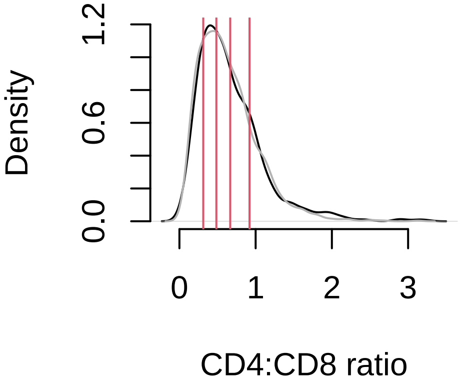
```
##      black
## 0.08751324
```

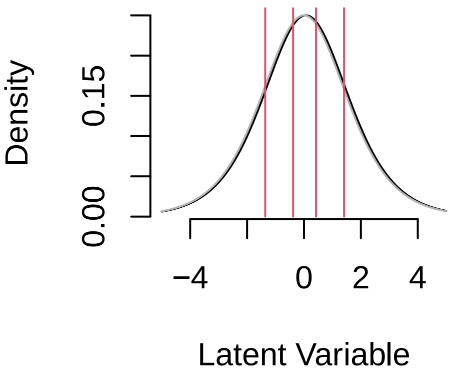


Ordered Logistic Regression with 5 Quantiles of CD4:CD8 Ratio

```
quants<-with(d,quantile(y,c(.2,.4,.6,.8)))
d$y5<-with(d,ifelse(y<quants[1],1,ifelse(y<quants[2],2,ifelse(y<quants[3],3,ifelse(y<quants[4],4,5)))))
fit5<-polr(factor(y5)~black, data=d)
fit5$coeff["black"]
```

```
##      black
## 0.06681472
```

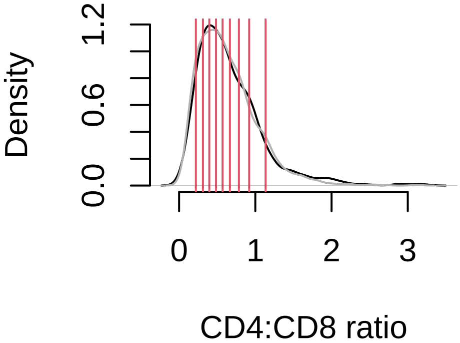


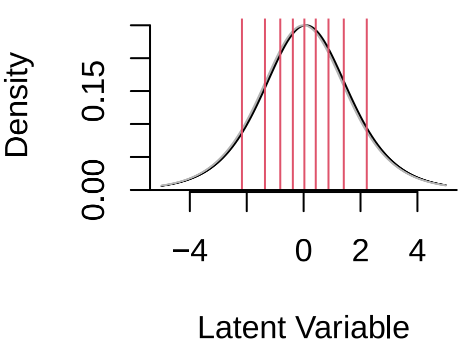


Ordered Logistic Regression with 10 Quantiles of CD4:CD8 Ratio

```
quants<-with(d,quantile(y,c(1:9)/10))
d$y10<-with(d,ifelse(y<quants[1],1,ifelse(y<quants[2],2,ifelse(y<quants[3],3,ifelse(y<quants[4],4,
               ifelse(y<quants[5],5,ifelse(y<quants[6],6,ifelse(y<quants[7],7,ifelse(y<quants[8],8,
               ifelse(y<quants[9],9,10))))))))))
fit10<-polr(factor(y10)~black, data=d)
fit10$coeff["black"]
```

```
##      black
## 0.07888959
```





Ordered Logistic Regression with increasing Categorizations of CD4:CD8 Ratio

- All of the estimated beta coefficients for black race are estimating the same population parameter
  - The shift in the latent variable distribution due to race
- Notice that the beta estimates are all fairly close
- Notice that the standard deviation of the estimates decreases with more categories
  - Quite a bit of information is lost if one simply dichotomizes CD4:CD8 ratio

```
##   categories       beta    beta.SD
##            2 0.04353153 0.12835184
##            3 0.08045377 0.08664500
##            4 0.08751324 0.08453354
##            5 0.06681472 0.08324597
##           10 0.07888959 0.08196221
```

Ordered Logistic Regression with Every Value its own Category

- What if we do ordered logistic regression but treating every value as its own category?
- In the CD4:CD8 ratio example, this is corresponds with 1859 categories ($n = 2024$).
- Requires a new function; we will use the `orm` function in the `rms` library.

```
modN<-orm(y~black, data=d)
modN$coeff["black"]
```

```
##      black
## 0.08553269
```

Ordered Logistic Regression with Every Value its own Category

- Again, this estimates the same beta parameter as the other categorizations.
- All categorizations yield similar, but slightly different beta parameter estimates.
  - With more categorizations, eventually beta coefficient estimate will converge to the estimate using every value as its own category.
  - It is kind of nice not to have to select the number of categorizations, as this is arbitrary and results in information loss.
- Notice the slightly decreased standard deviation of the estimate using every value as its own category.
- The alpha parameters (“intercepts”) can be thought of as the values that map the original data to the latent variable scale.

```
##  categories       beta    beta.SD
##           2 0.04353153 0.12835184
##           3 0.08045377 0.08664500
##           4 0.08751324 0.08453354
##           5 0.06681472 0.08324597
##          10 0.07888959 0.08196221
##        1859 0.08553269 0.08161979
```

Ordered Logistic Regression with Every Value its own Category

- The p-value from ordered logistic regression letting every value be its own category is approximately equal to the p-value from the Wilcoxon rank-sum test.

```
anova(modN)
```

```
##                 Wald Statistics          Response: y
##
##  Factor     Chi-Square d.f. P
##  black      1.1        1    0.2947
##  TOTAL      1.1        1    0.2947
```

```
wilcox.test(y~black, data=d)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  y by black
## W = 442962, p-value = 0.2948
## alternative hypothesis: true location shift is not equal to 0
```

Ordered Logistic Regression with Every Value its own Category

```
modN2<-orm(y~black + age, data=d)
anova(modN2)
```

```
##                 Wald Statistics          Response: y
##
##  Factor     Chi-Square d.f. P
##  black        0.12      1    0.7313
##  age        104.00      1    <.0001
##  TOTAL      105.07      2    <.0001
```

```
modN2$coeff["black"]
```

```
##      black
## 0.02808078
```

```
sqrt(modN2$var["black","black"])
```

```
## [1] 0.08178529
```

- There are substantial benefits of ordinal logistic regression over the Wilcoxon rank-sum test:
  - One can adjust for other variables
  - Interpretable regression coefficients
- e.g., Association between CD4:CD8 ratio and black race after adjusting for age.
  - After adjusting for age, blacks have similar odds of having a higher CD4:CD8 ratio than whites.
  - Odds ratio $= \exp(-0.0281) = 0.97$; 95% confidence interval: $\exp(-0.0281 \pm 1.96 \times 0.0818) = (0.83, 1.14)$; p=0.73

Summary So Far

- Skewed data often needs to be transformed
- Difficult to choose the transformation
- One could dichotomize the skewed data and fit logistic regression (with information loss)
- One could categorize the skewed data and fit ordered logistic regression
- One can simply fit ordered logistic regression to the skewed data without categorizing
    - This estimates the same beta coefficient as logistic / ordered logistic regression with categorizing (shift in the latent logistic variable due to covariates)
    - This is more efficient than categorizing
    - It does not require arbitrary selection of the number of categories
    - The alpha parameters can be thought of as the values that map the original data to the latent variable scale.
    - With binary predictors it results in nearly an identical p-value to Wilcoxon rank sum test

- Let's now think about this from another direction

Linear Transformation Models and Cumulative Probability Models

Y is continuous outcome, X is vector of covariates

Let $Y^* = h(Y)$ where $h(\cdot)$ is a monotonic transformation.

Linear transformation model:

$$\begin{aligned} h(Y) = Y^* = \beta^T X + \epsilon, &\text{ where } \epsilon \sim F_\epsilon, \text{ a specified distribution.} \\ \Rightarrow Y = H(\beta^T X + \epsilon), &\text{ where } H(\cdot) \equiv h(\cdot)^{-1}. \end{aligned}$$

Cumulative probability model:

$$\begin{aligned} P(Y \le y|X) &= P[H(\beta^T X + \epsilon) \le y|X] \\ &= P[\epsilon \le H^{-1}(y) - \beta^T X|X] \\ &= F_\epsilon[\alpha(y) - \beta^T X]. \\ \Rightarrow G[P(Y \le y|X)] &= \alpha(y) - \beta^T X, \end{aligned}$$

where $G = F_\epsilon^{-1}$ is a link function and $\alpha(\cdot)$ is an intercept function.

Cumulative Probability Models

$$G[P(Y \leq y|X)] = \alpha(y) - \beta^T X.$$

$Y = H(\beta^T X + \epsilon)$ implies $\alpha(Y) = H^{-1}(Y) = \beta^T X + \epsilon$, or that $\alpha(\cdot)$ is the transformation needed for Y to be fit with a linear regression model with error term $\epsilon \sim F_\epsilon$.

- Example: Normal linear model with square-root transformed Y.

$$\begin{aligned} &\sqrt{Y} = \gamma_0 + \gamma^T X + \delta, \text{ where } \delta \sim N(0, \sigma^2). \\ \Rightarrow &\alpha(Y) = (\sqrt{Y} - \gamma_0)/\sigma = \beta^T X + \epsilon, \text{ where } \beta = \gamma/\sigma \text{ and } \epsilon \sim N(0,1). \\ \Rightarrow &\Phi^{-1}[P(Y \leq y|X)] = \alpha(y) - \beta^T X. \end{aligned}$$

Semiparametric Linear Transformation Model

Instead of assuming $\alpha(\cdot)$, let's estimate it!

$$G[P(Y \leq y|X)] = \alpha(y) - \beta^T X.$$

We could put a parametric form on $\alpha(y)$ and estimate it, but that may limit our options. In the same spirit as the Wilcoxon rank sum test, we might want to estimate $\alpha(y)$ non-parametrically with a step function.

With the observed values $y_{(1)} < \cdots < y_{(J)}$ for J unique values, the CPM can be expressed as

$$G[P(Y \leq y_{(j)}|X)] = \alpha_j - \beta^T X,$$

where $\alpha_j = \alpha(y_{(j)})$. Here the parameters are $(\beta, \alpha_1, \ldots, \alpha_{J-1}, \alpha_J)$, where $\alpha_1 \leq \cdots \leq \alpha_{J-1} \leq \alpha_J \equiv \infty$.

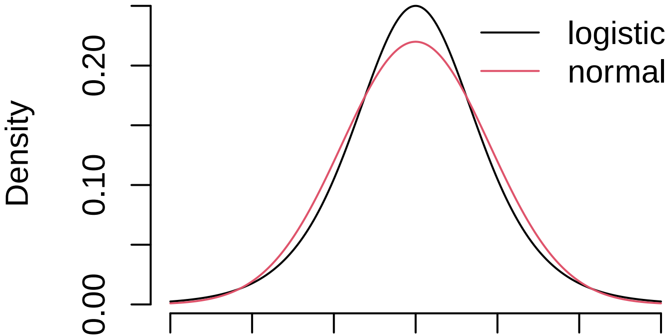
Note that his looks identical to the CPM for ordinal outcome Y with K categories:

$$G[P(Y \leq C_k|X)] = \alpha_k - \beta^T X \quad (k = 1, ..., K-1).$$

Link Functions

Table 1: Commonly used link functions and their corresponding error distributions.

Name	Link Function	Error Distribution	CDF (F_ϵ)
logit	$\log\left[p/(1-p)\right]$	logistic	$\exp(\epsilon)/[1+\exp(\epsilon)]$
probit	$\Phi^{-1}(p)$	normal	$\Phi(\epsilon)$
loglog	$-\log\left[-\log(p)\right]$	extreme value type II (Gumbel Max)	$\exp\left[-\exp(-\epsilon)\right]$
cloglog	$\log\left[-\log(1-p)\right]$	extreme value type I (Gumbel Min)	$1-\exp\left[-\exp(\epsilon)\right]$



Semiparametric cumulative probability models

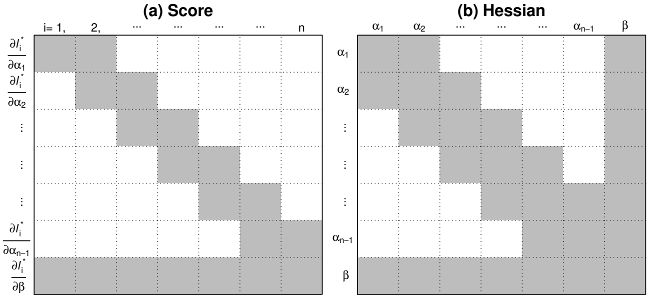
The nonparametric likelihood is identical to the multinomial likelihood used for ‘cumulative link models’ for ordinal data, such as ordered logistic regression:

$$\begin{aligned} L(\beta, \boldsymbol{\alpha}) &= \prod_{j=1}^{J} \prod_{i:y_i=y_{(j)}} \left[F(y_i|\boldsymbol{x}_i) - F(y_i^-|\boldsymbol{x}_i) \right], \\ &= \prod_{j=1}^{J} \prod_{i:y_i=y_{(j)}} \left[G^{-1}(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) - G^{-1}(\alpha_{j-1} - \boldsymbol{\beta}^T \boldsymbol{x}_i) \right], \end{aligned}$$

where $\alpha_0 = -\infty, \alpha_n = \infty$.
where $-\infty \equiv \alpha_0 < \alpha_1 < \cdots < \alpha_{J-1} < \alpha_J \equiv \infty$.

Equivalent to fitting ordinal regression model and treating each unique outcome as its own category.

Sparse Structure of Score Function and Hessian Matrix


(a) Score
i= 1,
2,
…
…
…
…
n
$\frac{\partial l_i^*}{\partial \alpha_1}$
$\frac{\partial l_i^*}{\partial \alpha_2}$
⋮
⋮
⋮
$\frac{\partial l_i^*}{\partial \alpha_{n-1}}$
$\frac{\partial l_i^*}{\partial \beta}$
(b) Hessian
α_1
α_2
…
…
…
α_{n-1}
β
α_1
α_2
⋮
⋮
⋮
α_{n-1}
β

- Computation can be performed with thousands of unique outcomes using R package `rms`, the function `orm`.
- This software takes advantage of the sparse structure of the score and hessian matrix.
- Other software for ordinal outcomes typically has problems with this many unique outcomes.

Estimation of Expectations and Distributions Conditional on Covariates

Cumulative distribution function conditional on covariates is estimated as

$$\hat{P}(Y \leq y|X) = \hat{F}(y|X) = G^{-1}(\hat{\alpha}_j - \hat{\beta}X),$$

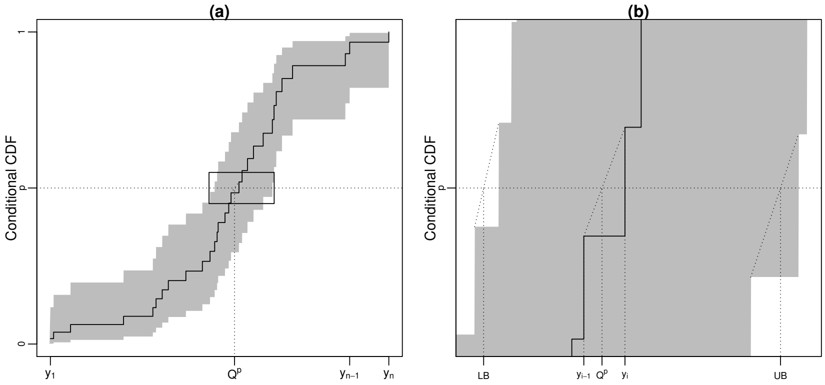
where $y_{(j)} = \max\{y_i : y_i \leq y\}$.

Expectation conditional on covariates is estimated as

$$\hat{E}(Y|X) = \sum_{j=1}^{n} y_{(j)} \left\{ \hat{F}(y_{(j)}|X) - \hat{F}(y_{(j-1)}|X) \right\}.$$

Delta method used to compute variance of $\hat{F}(y|X)$ and $\hat{E}(Y|X)$.

Estimation of CDF and Quantiles Conditional on Covariates



- Estimation of quantiles conditional on covariates is done by inverting the conditional distribution function.
- Linear interpolation can account for the discreteness.

Returning to CD4:CD8 Ratio Example

Fit a regression model with several covariates including age, which will be included using splines.

```r
dd <- datadist(d)
options(datadist='dd')
mod <- orm(y ~ female + black + rcs(age, 4) + route + hcv + hbv + year,
           data=d, x=TRUE, y=TRUE)
anova(mod)
```

```
##                 Wald Statistics          Response: y
##
##  Factor     Chi-Square d.f. P
##  female      26.07      1   <.0001
##  black        1.36      1   0.2438
##  age        103.23      3   <.0001
##   Nonlinear   8.80      2   0.0123
##  route        0.31      3   0.9579
##  hcv          0.14      1   0.7113
##  hbv          0.84      1   0.3594
##  year         0.51      1   0.4737
##  TOTAL      162.49     11   <.0001
```

CD4:CD8 Ratio Example – Odds Ratios

```r
options(width=200)
summary(mod)
```

```
##              Effects              Response : y
##
##  Factor                    Low      High    Diff.  Effect     S.E.     Lower 0.95 Upper 0.95
##  female                       0.000    1.00  1.000  0.6134300 0.120130  0.377980   0.848880
##   Odds Ratio                  0.000    1.00  1.000  1.8467000       NA  1.459300   2.337000
##  black                        0.000    1.00  1.000 -0.1000700 0.085853 -0.268340   0.068193
##   Odds Ratio                  0.000    1.00  1.000  0.9047700       NA  0.764650   1.070600
##  age                         34.389   47.66 13.271 -0.3493800 0.108010 -0.561070  -0.137690
##   Odds Ratio                 34.389   47.66 13.271  0.7051300       NA  0.570600   0.871370
##  hcv                          0.000    1.00  1.000 -0.0526120 0.142150 -0.331220   0.225990
##   Odds Ratio                  0.000    1.00  1.000  0.9487500       NA  0.718050   1.253600
##  hbv                          0.000    1.00  1.000 -0.1482000 0.161710 -0.465150   0.168740
##   Odds Ratio                  0.000    1.00  1.000  0.8622600       NA  0.628040   1.183800
##  year                      2004.000 2010.00  6.000  0.0466040 0.065043 -0.080877   0.174090
##   Odds Ratio               2004.000 2010.00  6.000  1.0477000       NA  0.922310   1.190200
##  route - Hetero:MSM           3.000    1.00     NA  0.0436010 0.111390 -0.174730   0.261930
##   Odds Ratio                  3.000    1.00     NA  1.0446000       NA  0.839690   1.299400
##  route - IDU:MSM              3.000    2.00     NA  0.0059968 0.173200 -0.333470   0.345460
##   Odds Ratio                  3.000    2.00     NA  1.0060000       NA  0.716430   1.412600
##  route - Other/Unknown:MSM    3.000    4.00     NA  0.0999920 0.216220 -0.323790   0.523780
##   Odds Ratio                  3.000    4.00     NA  1.1052000       NA  0.723400   1.688400
```

CD4:CD8 Ratio Example – Odds Ratios

```
options(width=200)
summary(mod, age=c(35,45), year=c(2004,2005))
```

```
##              Effects              Response : y
##
##  Factor                    Low  High Diff. Effect     S.E.     Lower 0.95 Upper 0.95
##  female                       0    1  1     0.6134300 0.120130  0.37798    0.848880
##   Odds Ratio                  0    1  1     1.8467000       NA  1.45930    2.337000
##  black                        0    1  1    -0.1000700 0.085853 -0.26834    0.068193
##   Odds Ratio                  0    1  1     0.9047700       NA  0.76465    1.070600
##  age                         35   45 10    -0.2539100 0.093774 -0.43770   -0.070113
##   Odds Ratio                 35   45 10     0.7757600       NA  0.64552    0.932290
##  hcv                          0    1  1    -0.0526120 0.142150 -0.33122    0.225990
##   Odds Ratio                  0    1  1     0.9487500       NA  0.71805    1.253600
##  hbv                          0    1  1    -0.1482000 0.161710 -0.46515    0.168740
##   Odds Ratio                  0    1  1     0.8622600       NA  0.62804    1.183800
##  year                      2004 2005  1     0.0077674 0.010840 -0.01348    0.029014
##   Odds Ratio               2004 2005  1     1.0078000       NA  0.98661    1.029400
##  route - Hetero:MSM           3    1 NA     0.0436010 0.111390 -0.17473    0.261930
##   Odds Ratio                  3    1 NA     1.0446000       NA  0.83969    1.299400
##  route - IDU:MSM              3    2 NA     0.0059968 0.173200 -0.33347    0.345460
##   Odds Ratio                  3    2 NA     1.0060000       NA  0.71643    1.412600
##  route - Other/Unknown:MSM    3    4 NA     0.0999920 0.216220 -0.32379    0.523780
##   Odds Ratio                  3    4 NA     1.1052000       NA  0.72340    1.688400
```

CD4:CD8 Ratio Example – Exceedance Probabilities

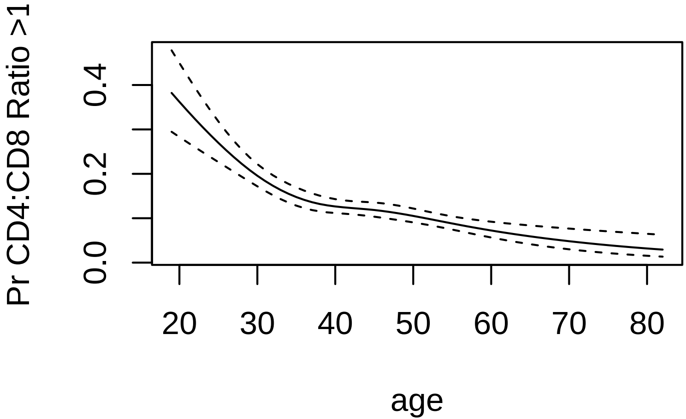
Computing predicted probabilities that CD4:CD8 ratio is greater than 1 for various ages and holding all other covariates constant at their medians or modes.

```
ages<-with(d,c(round(min(age)):round(max(age))))
P<-ExProb(mod)
Predict(mod, age=c(20,30,40,50,60,70,80), fun= function(x) P(x, y=1))
```

```
##   female black age route hcv hbv year       yhat      lower      upper
## 1      0     0  20   MSM   0   0 2007 0.36208425 0.28287766 0.44956542
## 2      0     0  30   MSM   0   0 2007 0.19564243 0.17184453 0.22185313
## 3      0     0  40   MSM   0   0 2007 0.12649325 0.11167048 0.14296689
## 4      0     0  50   MSM   0   0 2007 0.10518448 0.09056089 0.12185306
## 5      0     0  60   MSM   0   0 2007 0.07236862 0.05656322 0.09215910
## 6      0     0  70   MSM   0   0 2007 0.04826056 0.03005026 0.07663431
## 7      0     0  80   MSM   0   0 2007 0.03190734 0.01542414 0.06484537
##
## Response variable (y):
##
## Adjust to: female=0 black=0 route=MSM hcv=0 hbv=0 year=2007
##
## Limits are 0.95 confidence limits
```

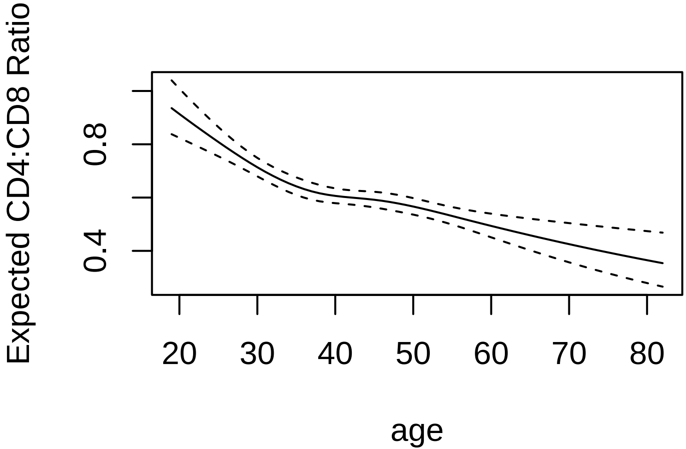
Predicted Probability CD4:CD8 Ratio > 1 as Function of Age

```
ages<-with(d,c(round(min(age)):round(max(age)))); P<-ExProb(mod)
pred.probs<-Predict(mod, age=ages, fun= function(x) P(x, y=1))
plot(c(ages,ages),c(pred.probs$lower,pred.probs$upper),type="n",xlab="age",ylab="Pr CD4:CD8 Ratio >1")
lines(ages,pred.probs$yhat); lines(ages,pred.probs$lower,lty=2); lines(ages,pred.probs$upper,lty=2)
```



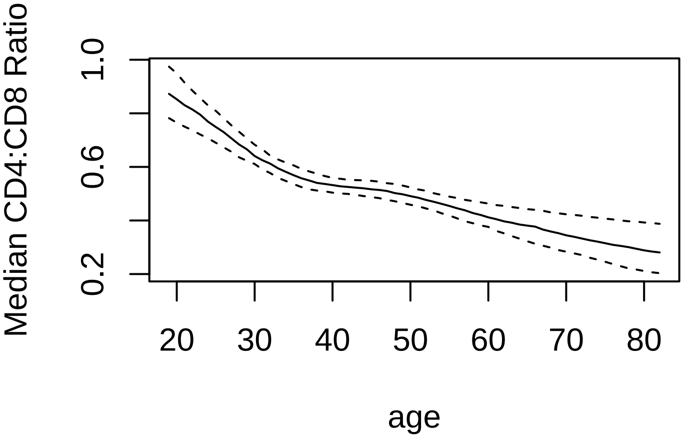
Expectation (Mean) of CD4:CD8 Ratio as Function of Age

```
mean.fun<-Mean(mod)
pred.means<-Predict(mod, age=ages, fun= function(x) mean.fun(x))
plot(c(ages,ages),c(pred.means$lower,pred.means$upper),type="n",xlab="age",ylab="Expected CD4:CD8 Ratio")
lines(ages,pred.means$yhat); lines(ages,pred.means$lower,lty=2); lines(ages,pred.means$upper,lty=2)
```



Median CD4:CD8 Ratio as Function of Age

```
quants.fun<-Quantile(mod)
pred.medians<-Predict(mod, age=ages, fun= function(x) quants.fun(0.5, x))
plot(c(ages,ages),c(pred.medians$lower,pred.medians$upper),type="n",xlab="age",ylab="Median CD4:CD8 Ratio
lines(ages,pred.medians$yhat); lines(ages,pred.medians$lower,lty=2); lines(ages,pred.medians$upper,lty=2)
```



Conclusions

Continuous data can be analyzed using models for ordinal data

- Strengths
  - No need to transform data
  - Directly models CDF, from which other statistics can be derived
    - conditional expectation, quantiles, probabilities, probability indices
  - Detection limits easily handled
  - Can handle cluster data
  - Unbiased estimation, proper confidence interval coverage for moderately sized n
- Limitations
  - Requires specification of a link function
    - Fairly robust to moderate misspecification (e.g., wrong link function)
  - Some bias with small sample sizes
  - Not as fast as linear regression