Models in Epidemiology And Biostatistics
Gordon Hilton Fick

The Proportional Probability Model

A dichotomous outcome linked to k explanatory variables

- The outcome is : 0 [absence of a condition]

  : 1 [presence of a condition]
- The condition might be a negative characteristic like a diagnosis of cancer : D
- The condition might be a positive characteristic like a successful surgery : S
- This outcome is to be linked to potential modifiers, potential confounders and other explanatory variables

Negative or Positive : Log Link : Two different models

$x_1, x_2, x_3 \ldots x_n$ are the explanatory variables : exposure, age, gender, weight...

$$p = Pr(D) \quad \log(p) = \sum_{i=0}^{k} \beta_i x_i$$

$$q = Pr(not\ D) \quad \log(q) = \sum_{i=0}^{k} \alpha_i x_i$$

Negative or Positive : Logit Link : One Model

$x_1, x_2, x_3 \ldots x_n$ are the explanatory variables :
exposure, age, gender, weight...

$$p = Pr(D) \quad \log\left(\frac{p}{1-p}\right) = \sum_{i=0}^{k} \beta_i x_i$$

$$q = Pr(not\ D) \quad \log\left(\frac{q}{1-q}\right) = \sum_{i=0}^{k} \alpha_i x_i$$

$$\frac{q}{1-q} = \frac{1-p}{p} = \frac{1}{\frac{p}{1-p}} \quad \text{so } \alpha_i = -\beta_i$$

A 2x2 Table from a cohort study When Probability is Risk

p_1 = Pr(D | E)

p_0 = Pr(D | not E)

$$\log(p) = \beta_0 + \beta_1 E$$

$$E = 0 : \log(p_0) = \beta_0$$

$$E = 1 : \log(p_1) = \beta_0 + \beta_1$$

$$\beta_1 = \log(p_1) - \log(p_0) = \log \frac{p_1}{p_0}$$

$$Risk\ Ratio = RR = \frac{p_1}{p_0} = \exp(\beta_1)$$

A 2x2 Table from a cohort study When Probability is Health

q_1 = Pr(S | E)
q_0 = Pr(S | not E)

$$\log(q) = \alpha_0 + \alpha_1 E$$

$$E = 0 : \log(p_0) = \alpha_0$$

$$E = 1 : \log(p_1) = \alpha_0 + \alpha_1$$

$$\alpha_1 = \log(p_1) - \log(p_0) = \log \frac{q_1}{q_0}$$

$$Health\ Ratio = HR = \frac{q_1}{q_0} = \exp(\beta_1)$$

2 2x2 Tables from a cohort study

p_{1j} = P(D | E and Strata j)

p_{0j} = P(D | not E and Strata j)

$$\log(p)=\beta_0+\beta_1 E+\beta_2 S+\beta_3 ES$$

$$S=0:\log(p)=\beta_0+\beta_1 E \quad S=1:\log(p)=\beta_0+\beta_2+(\beta_1+\beta_3)E$$

$$\beta_1=\log(p_{10})-\log(p_{00})=\log\frac{p_{10}}{p_{00}}$$

$$\beta_1+\beta_3=\log(p_{11})-\log(p_{01})=\log\frac{p_{11}}{p_{01}}$$

$$RR_0=\frac{p_{10}}{p_{00}}=\exp(\beta_1) \quad RR_1=\frac{p_{11}}{p_{01}}=\exp(\beta_1+\beta_3)$$

Compared with Logistic Regression

When constructing models with the log link, one uses the same processes as with models with the logit link. Now, with the log link, log odds are replaced with log probabilities and odds ratios are replaced with probability ratios. All of the interpretations are the same except the changes noted above.

Software Algorithm Trouble

- Error messages like:

  Failure to converge : yikes !

  Backing up : What's that ?

  Concave region : Who cares ?

- Fit gives nonsense like :

  Probabilities that are greater than one !

Likelihood Function Oddities? Additive Model Incorrect?

- Williamson, Eliasziw and Fick (2013) say no to both questions.
- Newton-Raphson and Fisher Scoring will sometimes fail even though the likelihood function is 'reasonable'.
- Up until 2017 : R, Stata and SAS could give incorrect results
- Boundaries must be identified.

Constrained Optimization offers a fix

- an adaptive barrier algorithm : Lange(1994,2004)
- implemented in R : constrOptim
- required conditions can be checked
- after the correct MLE is found, then 'standard' theory can be used to determine SE and Wald tests [sometimes]
- boundary check based on tolerance settings

lbreg and logbin

- lbreg : R package developed by Bernardo Andrade and Mateus Carbone Ananias (most recent version 1.2 released in January 2018)

  : built around constrOptim

  : paper published in Communications in Statistics

- logbin : R package developed by Mark Donoghoe and Ian Marschner (most recent version 2.0.4 released in August 2018)

  : offers constrOptim as an option

  : paper published in Journal of Statistical Software

A mystery

- A [possibly incomplete] work by Wedderburn[1976] was posthumously published. He left out the log link but discussed many other links in an important paper.
- He died from a reaction to a bee sting.
- In his doctoral dissertation, Gurbakhshash Singh [2017] makes an interesting contribution to this mystery. He also derives many closed form expressions that are rather surprising.
- We [Singh & Fick] hope to publish some of this material 'soon'.

An ordinal outcome linked to k explanatory variables

- The outcome has J ordered levels
- There are J-1 ways to 'cut' the outcome
- One can order the levels from best(1) to worst(J)

  then Pr(of being below the jth cut) is [sorta like]:

  the probability of doing better.
- One can code the levels from worst(1) to best(J)

  then Pr (of being below the jth cut) is [sorta like]:

  the probability of doing worse.

Four Levels : Three Cuts : Two Orders

	Complete Remission	Partial Remission	No Change	Progression Of Disease
B to W	1	2	3	4
Cut	1	2	3	

	Progression Of Disease	No Change	Partial Remission	Complete Remission
W to B	1	2	3	4
Cut	1	2	3	

The Proportional Probability Model

- Analogous to the proportional odds model

$$p_j = Pr(\text{below the jth cut})$$

$$\log(p_j) = \kappa_j + \sum_{i=1}^{k} \beta_i x_i \quad j = 1,2,\ldots J-1$$

β_i assumed common to cuts κ_i

κ_i assumed common to the β_i

- There are two [different] models based on the ordering [coding] of the levels

4x2 table from a cohort study

			Exposure	
	Cut		Yes	No
Progression 4				
	3	$\log(p_3)$	$\kappa_3+\beta_1$	κ_3
No Change 3				
	2	$\log(p_2)$	$\kappa_2+\beta_1$	κ_2
Partial 2				
	1	$\log(p_1)$	$\kappa_1+\beta_1$	κ_1
Complete 1				

For each cut, compare exposed with unexposed.

The difference is always

$$(\kappa_j + \beta_1) - \kappa_j = \beta_1$$

In other words, this difference is assumed common to the cuts.

The exponent is an assumed common probability ratio.

The probability of 'doing better' for those exposed divided by the probability of 'doing better' for those unexposed.

Reverse coding the ordinal outcome

- We again get the exponent being the assumed common probability ratio.
- Now - the probability of 'doing worse'.
- Here, many would call this ratio a 'risk ratio';

  assumed common to the cuts.
- The two ratios are NOT the same and are not functionally related.

The Log Cumulative Probability Model

- Analogous to the Generalized Ordered Logit Model

$$\log(p_j) = \sum_{i=0}^{k} \beta_{ij} x_i \quad j = 1,2,\ldots J-1$$

  Fits J-1 cut specific sets of regression coefficients

- Not quite the same as fitting J-1 marginal models
- Can be used to assess the proportional probability assumption

4x2 table from a cohort study

			Exposure	
	Cut		Yes	No
Progression 4				
	3	$\log(p_3)$	$\beta_{03}+\beta_{13}$	β_{03}
No Change 3				
	2	$\log(p_2)$	$\beta_{02}+\beta_{12}$	β_{02}
Partial 2				
	1	$\log(p_1)$	$\beta_{01}+\beta_{11}$	β_{01}
Complete 1				

For each cut, compare exposed with unexposed.

The difference is now

$$(\beta_{0j}+\beta_{1j})-\beta_{0j}=\beta_{1j}$$

In other words, this difference is now cut specific.

The exponent is a cut specific probability ratio.

The probability of 'doing better' for those exposed divided by the probability of 'doing better' for those unexposed.

Reverse coding the ordinal outcome

- We again get the exponent being a probability ratio now specific to each cut.
- Now - the probability of 'doing worse'.
- Here, these ratios are a 'risk ratios'; one ratio for each cut
- The sets of ratios [based on the ordering of the outcome] are NOT the same and are not functionally related.

lcpm and ppm

lcpm and ppm : R package developed by Gurbakhshash Singh and Gordon Hilton Fick

built using constrOptim

cran.rproject.org/web/packages/lcpm/index.html

Singh G and Fick GH [2020] 'Ordinal outcomes: A cumulative probability model with the log link and an assumption of proportionality' Statistics in Medicine 39(9) p1343 - 1361

Example with R : Proportional Odds Model

```
summary(polr(factor(outc)~gender+therapy,data=tumor))


Coefficients:

          Value Std. Error t value

gender  -0.5414     0.2872  -1.885

therapy -0.5807     0.2121  -2.737


Intercepts:

    Value   Std. Error t value

1|2 -1.3180  0.1798    -7.3315

2|3  0.2492  0.1614     1.5443

3|4  1.3001  0.1850     7.0276
```

The Two Proportional Probability Models

```
summary(ppm(outcr~gender+therapy,data=tumor))

          Estimate    StdErr  z.value

cut_1    -1.705491  0.132597 -12.8623

cut_2    -0.933873  0.081287 -11.4885

cut_3    -0.231336  0.042972  -5.3834

gender   -0.068861  0.112131  -0.6141

therapy -0.198156  0.075178  -2.6358

summary(ppm(outc~gender+therapy,data=tumor))

          Estimate    StdErr  z.value

cut_1    -1.304604  0.096628 -13.5013

cut_2    -0.484566  0.052475  -9.2342

cut_3    -0.225795  0.039652  -5.6944

gender    0.132104  0.047115   2.8039

therapy   0.050464  0.047426   1.0641
```