Statistics
in Medicine

Research Article

Received 5 October 2011, Accepted 2 November 2012 Published online 6 December 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5689

The trend odds model for ordinal data‡

Ana W. Capuano[a,b,c*†] **and Jeffrey D. Dawson**[a]

Ordinal data appear in a wide variety of scientific fields. These data are often analyzed using ordinal logistic regression models that assume proportional odds. When this assumption is not met, it may be possible to capture the lack of proportionality using a constrained structural relationship between the odds and the cut-points of the ordinal values. We consider a trend odds version of this constrained model, wherein the odds parameter increases or decreases in a monotonic manner across the cut-points. We demonstrate algebraically and graphically how this model is related to latent logistic, normal, and exponential distributions. In particular, we find that scale changes in these potential latent distributions are consistent with the trend odds assumption, with the logistic and exponential distributions having odds that increase in a linear or nearly linear fashion. We show how to fit this model using SAS Proc NLMIXED and perform simulations under proportional odds and trend odds processes. We find that the added complexity of the trend odds model gives improved power over the proportional odds model when there are moderate to severe departures from proportionality. A hypothetical data set is used to illustrate the interpretation of the trend odds model, and we apply this model to a swine influenza example wherein the proportional odds assumption appears to be violated. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: non-proportional odds; constrained cumulative odds; influenza; latent distributions; logistic distribution

1. Introduction

Ordinal data are often analyzed using the proportional odds model (POM). Conceptually introduced by Aitchison and Silvey in 1957, developed by Snell in 1964, and further developed by McCullagh in 1980, the POM is a popular extension of logistic regression to ordinal data [1, 2]. Odds are considered proportional when all possible dichotomizations of high versus low outcomes result in the same odds metric. This proportional odds assumption has been justified by a theoretical underlying logistic distribution with shift in location but constant scale [2]. McCullagh also introduced a linear 'multiplicative model' that relaxes the assumption of a constant scale of the underlying logistic distribution by introducing scale parameters to be estimated [2]. Peterson and Harrel suggested that the lack of proportionality might be relaxed for some of the covariates by introducing intermediate constrained structural relationship between the odds and the cut-points [3].

In this paper, we illustrate and investigate a monotone class of Peterson and Harrel's constrained non-proportional model, which we will term a trend odds model (TOM). We first review the notation of the cumulative odds model, emphasizing proportional and trend classes, and give a hypothetical data example where the linear TOM holds. We then show how latent distributions relate to the TOM, with particular attention given to the logistic distribution. We show how to fit the TOM using SAS Proc NLMIXED and present some simulation results when the data follow the POM and when they follow the TOM. Finally, we apply the TOM to a swine influenza data set that appears to come from a non-proportional odds process.

[a]*Department of Biostatistics, University of Iowa College of Public Health, Iowa City, IA 52240, U.S.A.*
[b]*Department of Occupational and Environmental Health, University of Iowa College of Public Health, Iowa City, IA 52240, U.S.A.*
[c]*Rush University Medical Center, RADC, Chicago, IL 60612, U.S.A.*
**Correspondence to: Ana W. Capuano, 600 South Paulina, Suite 1023B, Chicago, IL, 60612, U.S.A.*
†*E-mail: ana_capuano@rush.edu*
‡*Supporting information may be found in the online version of this article.*

Copyright © 2012 John Wiley & Sons, Ltd.

Statistics in Medicine

2. Cumulative odds models: proportional odds and trend odds models

The TOM belongs to the cumulative ordinal class of models. To better understand its structure, we revisit the generalized structures of cumulative odds models. For simplicity we will consider the single predictor case. The models are organized conceptually and not necessary chronologically.

Suppose there is an observed ordinal outcome Y with $I+1$ categories ($Y = 0, 1, \ldots, I$), and a single covariate X. We define Ψ_{ix} as the log cumulative odds, or the odds of Y having a level at least as high as each cut-point I, given X. That is, $\Psi_{ix} = \log(P(Y \geqslant i|X = x)/(1 - P(Y \geqslant i|X = x)))$. The general form of the cumulative odds model is known as unconstrained cumulative odds model represented as

$$\psi_{ix} = \alpha_i + \beta_i X, \quad i = 1, \ldots, I.$$

Note that this model allows a different odds ratio relating X to Y for each cut-point, because the β terms are indexed by i. For example, suppose that X represents an exposure status that is being investigated, with $X = 1$ for exposed and $X = 0$ for unexposed. Several different and unrelated odds ratio for exposed versus unexposed can be calculated depending on the cut-point (i.e., $\theta_1 = e^{\beta_1}, \theta_2 = e^{\beta_2}, \ldots, \theta_I = e^{\beta_I}$).

In contrast, a commonly used cumulative ordinal model known as the POM constrains the aforementioned model by assuming that all β_i are equal to a common β [2]. For the single predictor case, we will notate the POM [4] as

$$\psi_{ix} = \alpha_i + \beta X, \quad i = 1, \ldots, I. \tag{1}$$

With this model, the odds of being at or above any cut-point are assumed to be the same for all cut-points; this is known as the 'proportional odds' assumption. In other words, instead of having several different odds ratio for exposed versus unexposed, a single odds ratio is calculated ($\theta = e^{\beta}$). When the parameter β is greater than 0 and X increases, the response Y is consistently more likely to be in a higher end. Although the POM makes a strong assumption, it is easy to fit and interpret [5].

When the proportional odds assumption is not met, it may be possible to capture the lack of proportionality using intermediate constrained structural relationship between the odds and the cut-points [1, 3, 6–8]. In general, this can be accomplished by adding a parameter, say γ, and a multiplicative scalar that varies with the cut-point, say t_i. For the single predictor case, we will define the constrained cumulative odds model as

$$\psi_{ix} = \alpha_i + (\beta + \gamma t_i)X, \quad i = 1, \ldots, I. \tag{2}$$

In general, the constrained cumulative odds models could potentially have any set of scalars, as long as the odds ratios from the resulting model adequately fit the data. Consider again the investigation of exposure status. If this generalized model is used with a set of scalars that includes consecutive zero values, then there would be proportionality of odds ratios across some of the cut-points within the same covariate. For example, the scalar values of $t_1 = 1$, $t_2 = -1$, $t_3 = 0$, and $t_4 = 0$ would yield the odd ratios of $\theta_1 = e^{\beta}e^{\gamma}, \theta_2 = e^{\beta}e^{-\gamma}$, and $\theta_3 = \theta_4 = e^{\beta}$. In other words, under the family of constrained non-proportional odds models, the scalar set can be unstructured.

The TOM is a structured constrained cumulative odds model. Although many types of trends may be possible, we define the TOM to have a monotonic structure, such that $t_1 \leqslant t_2 \leqslant \ldots \leqslant t_I$. Hence, for the single predictor case, the TOM is

$$\psi_{ix} = \alpha_i + (\beta + \gamma t_i)X, \quad i = 1, \ldots, I, \quad t_i < t_{i+1}. \tag{3}$$

The parameter space of α_i and β is similar to the POM, being comprised of any real number from minus infinity to infinity. In contrast, on the basis of the underlying theory that is going to be presented in the following section, the parameter space of $\gamma \cdot x$ depends on the scale of the theoretical underlying distribution of X. A trend in log odds ratios will result when the γ is non-zero. If exposure status is being used to predict the ordinal outcome, and if the proportional odds assumption does not appear reasonable, then the researcher can test for a trend in the cumulative odds ratio as the ordinal outcome (or the cut-points thereof) increases. Depending on the values of t_1, β, and γ, the odds ratios may all be below 1, all be above 1, or span the value of 1, but their trend must be monotonic.

Copyright © 2012 John Wiley & Sons, Ltd.

3. A latent logistic motivation for the TOM

Suppose that a response variable, Y^*, follows a logistic distribution with location and scale parameters (m and s) being functions of a predictor variable, X, such that

$$Y^*|X = x \sim \text{Logistic}(m_x, s_x). \tag{4}$$

Hence, $E(Y^*|X = x) = m_x$ and $\text{Var}(Y^*|X = x) = \pi^2 s_x^2/3$. For I-specific values of interest, say c_i, for $i = 1, 2, \ldots, I$, we would have the CDF of Y^*

$$F(Y^* = c_i|X = x) = P(Y^* < c_i|X = x) = 1/(1 + \exp(-(c_i - m_x)/s_x)). \tag{5}$$

By taking the logit function and rearranging terms, we have

$$\log(P(Y^* \geqslant c_i|X = x)/(1 - P(Y^* \geqslant c_i|X = x))) = (m_x - c_i)/s_x. \tag{6}$$

Suppose that the ordinal variable, Y, takes on values 0, 1, 2, …, I, based on the latent variable, Y^*, according to the c_i values. Specifically, let

$$Y = i \leftrightarrow Y^* \in (c_i, c_{i+1}], \tag{7}$$

and

$$Y >= i \leftrightarrow Y^* \geqslant c_i, \tag{8}$$

where $c_0 = -\infty$ and $c_{I+1} = \infty$.

So, in terms of the ordinal variable from a latent logistic distribution, Equation (6) becomes

$$\Psi_{ix} = (m_x - c_i)/s_x. \tag{9}$$

The logistic distribution can be related to the TOM through known cut-points. In the TOM introduced in the previous section, (Equation (2)), we have, for $i = 1, 2, \ldots, I$,

$$\Psi_{ix} = \alpha_i + (\beta + \gamma t_i)x, \tag{10}$$

where the t_i are monotonic values specified when fitting the model and α, β, and γ are parameters estimated by the fitting procedure. If the cut-points, c_1, c_2, …c_I, are known, then we can choose $t_i = c_i$ and then we use the superscript * to indicate the parameters fit in this situation, such that $\Psi_{ix} = \alpha_i^* + (\beta^* + \gamma^* c_i)x$.

Hence, combining (9) and (10) when $t_i = c_i$, we obtain

$$\alpha_i^* + (\beta^* + \gamma^* c_i)x = (m_x - c_i)/s_x = m_x/s_x - c_i/s_x. \tag{11}$$

To obtain interpretations of the parameters (α_i^*, β^*, and γ^*), we consider what happens under various values of X. When $X = 0$, we have $(m_0 - c_i)/s_0 = \alpha_i^* + (\beta^* + \gamma^* c_i)0$, implying

$$\alpha_i^* = (m_0 - c_i)/s_{0.} \tag{12}$$

For $X = x$, we have

$$(m_x - c_i)/s_x = \alpha_i^* + (\beta^* + \gamma^* c_i)x. \tag{13}$$

Combining (12) and (13) and subtracting $(m_0 - c_i)/s_0$ from both sides, we obtain

$$(m_x - c_i)/s_x - (m_0 - c_i)/s_0 = (\beta^* + \gamma^* c_i)x. \tag{14}$$

Gathering the terms of the index i together, we obtain

$$m_x/s_x - m_0/s_0 + (1/s_0 - 1/s_x)c_i = \beta^* x + \gamma^* x c_i. \tag{15}$$

Hence, for $x \neq 0$, the parameters are defined as

$$\beta^* = (m_x/s_x - m_0/s_0)/x \tag{16}$$

Copyright © 2012 John Wiley & Sons, Ltd.

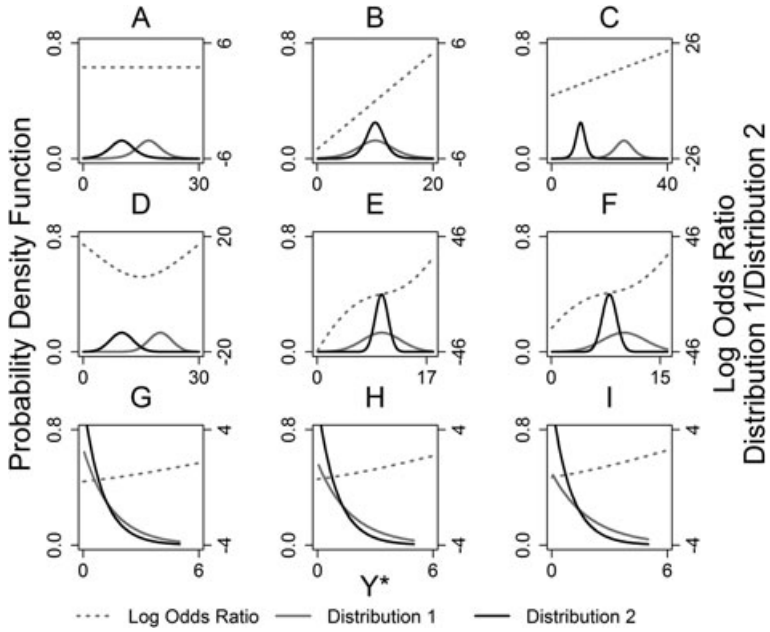


Figure 1. Panel of underlying logistic, normal, and respective log ratios. (A) Logistic(17,2) versus (10,2), (B) logistic(10,2) versus (10,1), (C) logistic(25,2) versus (10,1), (D) N(20,3) versus (10,3), (E) N(10,3) versus (10,1), (F) N(10,3) versus (8,1), (G) exponential(1.5) versus (1), (H) exponential(1.75) versus (1), and (I) exponential(2) versus (1).

and

$$\gamma^* = (1/s_0 - 1/s_x)/x. \tag{17}$$

There is no restriction on the parameter space of α_i^* and β^* as there is no restriction on the space of the logistic distribution mean. In contrast, the parameter space of γ^* is restricted as the logistic distribution scale can only assume positive values (i.e., $-1/s_x < \gamma^* x < 1/s_0$).

Figure 1A–C illustrates the trend in log odds ratio with underlying latent logistic distributions with location shift only (Figure 1A), scale shift only (Figure 1B), and scale and location shift (Figure 1C) for a simple binary covariate X.

Conversely, we can also find expressions for the s_x and m_x functions in terms of the TOM parameters. This is particularly relevant to better understand the association with continuous covariates. Solving for s_x in (17), we obtain

$$s_x = s_0/(1 - \gamma^* x\, s_0). \tag{18}$$

Note that TOM holds for x values satisfying $\gamma^* x < 1/s_0$. The parameter γ^* dictates the nonlinearity in the location shifts and simultaneously dictates the heteroskedasticity, associated with changing x values. In contrast, when γ^* is 0 or close to 0, the latent logistic has a mean that shifts linearly with changes in x, with constant variance ($s_x \approx s_0$).

Solving for m_x in (16) and s_x defined in (18), we obtain

$$m_x = (m_0 + \beta^* x\, s_0)/(1 - \gamma^* x\, s_0). \tag{19}$$

Note that (8) and (9) have respectively the form of

$$s_x = s_0/(1 + bx) \tag{20}$$

and

$$m_x = m_0(1 + ax)/(1 + bx), \tag{21}$$

where $a = -\beta^* s_0/m_0$ and $b = -\gamma^* s_0$.

Copyright © 2012 John Wiley & Sons, Ltd.

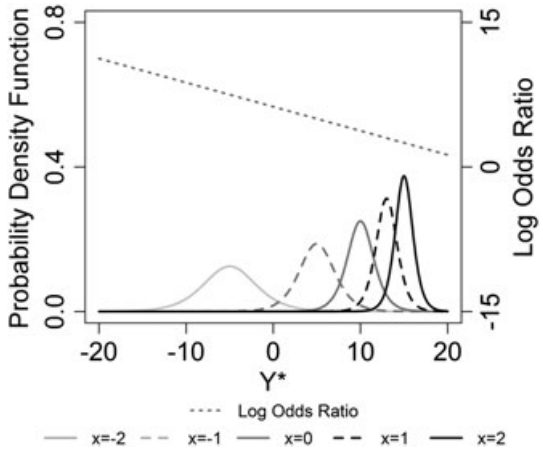


Figure 2. Underlying logistic and respective log odds ratio for continuous X. Logistic(m_x, s_x) where m_x and s_x are defined by Equations (19) and (20), and $m_0 = 10, s_0 = 1, a = 0.625$ and $b = 0.25$.

Hence, Equations (20) and (21) show how the parameters of a latent logistic distribution must depend on the covariate X in order for the TOM to hold, when the cut-points (c_i) used for categorizing the data are known. For example, a negative trend in log odds ratio is obtained with underlying latent logistic distributions with scale and location shift as a function of a continuous covariate X, when $m_0 = 10, s_0 = 1, a = 0.625,$ and $b = 0.25$ (Figure 2; Supporting Information).

4. Other underlying distributions

The logistic distribution provides a theoretical basis for the POM and for the TOM. It is feasible, however, that the trend odds assumption could hold when the underlying latent variable Y^* follows other distributions (Figure 1D–I). To investigate this, we calculated log odds ratios from two group situations with Y^* following normal and exponential distributions. Figure 1D–F illustrates the log odds ratio with underlying latent normal distributions with location shift only (Figure 1D), scale shift only (Figure 1E), and scale and location shift (Figure 1F) for a simple binary predictor X. Note that when a scale shift occurs, the trend is monotonic nonlinear, with the log odds ratio spanning the value of 0. For the location shift only (Figure 1D), the trend is non-monotonic, but the log odds ratio is always positive. Figure 1G–I illustrates the log odds ratio with underlying latent exponential distribution with increasing shift in location scale for a simple binary predictor X. Note that these log odds ratio trends are all monotonic, and nearly linear.

Table I shows the log odds ratio as a function of c_i that follows logistic, normal, or exponential distributions. The existence of a closed form solution makes it possible to show theoretically when the proportional odds or the trend odds assumption holds. For the logistic distribution, using Equation (9), the log odds ratio would be

$$\Psi_{ix} - \Psi_{i0} = (m_x - c_i)/s_x - (m_0 - c_i)/s_0. \tag{22}$$

When the scale is the same ($s = s_x = s_0$) but there is a location shift, the log odds can be represented by $(m_x - m_0)$/s, which is independent of c_i (proportional odds assumption holds). That is, for a POM, the common log odds ratio is proportional to the distance in mean underlying response given two values of a predictor X [9, 10]. In contrast, if there is a shift in scale with or without a shift in location the log odds ratio, then the trend in odds has a slope proportional to the inverse difference in scale (trend odds assumption holds). The first derivative

$$\mathrm{d}(\Psi_{ix} - \Psi_{i0})/\mathrm{d}c_i = 1/s_0 - 1/s_x \tag{23}$$

is a constant that can be shown to correspond to $c_i x$ or $-(b/s_0)x$ by using respectively Equation (17) or (20).

Copyright © 2012 John Wiley & Sons, Ltd.

Table I. Some examples of distributions for the hypothetical underlying latent variable Y^* and holding relation with log odds ratios.

Latent distribution	Parameter shift from $X = 0$ to $X = x$	Log odds ratio $(\Psi_{ix} - \Psi_{i0})$	Holds
Logistic	Location (m)	$(m_x - m_0)/s$	Proportional odds
Logistic	Scale (s_x)	$(m - c_i)(1/s_x - 1/s_0)$	Trend odds—linear monotonic
Logistic	Location (m_x) and scale (s_x)	$(m_x - c_i)/s_x - (m_0 - c_i)/s_0$	Trend odds—linear monotonic
Normal	Mean (μ_x) and variance (σ_x)	$\log((1 - \Phi(c_i/\sigma_x - \mu_x/\sigma_x))/\Phi(c_i/\sigma_x - \mu_x/\sigma_x)) - \log((1 - \Phi(c_i/\sigma_0 - \mu_0/\sigma_0))/\Phi(c_i/\sigma_0 - \mu_0/\sigma_0))$	Trend odds—nonlinear monotonic
Normal	Variance (σ_x)	$\log((1 - \Phi(c_i/\sigma_x - \mu/\sigma_x))/\Phi(c_i/\sigma_x - \mu/\sigma_x)) - \log((1 - \Phi(c_i/\sigma_0 - \mu/\sigma_0))/\Phi(c_i/\sigma_0 - \mu/\sigma_0))$	Trend odds—nonlinear monotonic
Normal	Mean (μ_x)	$\log((1 - \Phi(c_i/\sigma - \mu_x/\sigma_x))/\Phi(c_i/\sigma - \mu_x/\sigma)) - \log((1 - \Phi(c_i/\sigma - \mu_0/\sigma_0))/\Phi(c_i/\sigma - \mu_0/\sigma))$	Quadratic
Exponential	Location scale (λ_0)	$\log(e^{c_i\lambda 0} - 1) - \log(e^{c_i\lambda x} - 1)$	Trend odds—nonlinear monotonic

For the exponential distribution, with parameter λ_x (i.e., with a mean of $1/\lambda_x$), the log odds ratio can be represented by

$$\Psi_{ix} = \log(\exp(-\lambda_x c_i)/(1 - (\exp(-\lambda_x c_i))) = -\lambda_x c_i - \log(1 - (\exp(-\lambda_x c_i)). \tag{24}$$

Shifts in location scale result in log odds ratio (provided in Table I) that are a monotonic nonlinear function of c_i (trend odds assumption holds). The first derivative

$$\mathrm{d}(\Psi_{ix} - \Psi_{i0})/\mathrm{d}c_i = \lambda_0 - \lambda_x + \lambda_0/(\exp(\lambda_0 c_i) - 1) - \lambda_x/(\exp(\lambda_x c_i) - 1) \tag{25}$$

has a limit of

$$\lim_{y*\to\infty}(\mathrm{d}(\Psi_{ix} - \Psi_{i0})/\mathrm{d}c_i) = \lambda_0 - \lambda_x. \tag{26}$$

That is, although the finite trend is nonlinear, it approaches linearity with increasing y^*.

5. Simulation study

The properties of the TOM were compared with the binary logistic model at different cut-points and the POM. Simulation was performed in SAS considering the simulation steps suggested by Burton *et al.* [11]. Simulated scenarios included different sample sizes (50–1000 per group), different reference multinomial distribution, and combinations of β and γ to represent holding proportional odds assumption or trend odds assumption. A total of 1000 data sets were randomly generated from a multinomial distribution using different seeds. The data were modeled with the binary logistic model, the POM, and the TOM. This was accomplished using Proc Logistic and Proc NLMIXED in SAS. Bias, coverage, accuracy performance, test size, and power were evaluated in the statistical methods for different scenarios.

Table II shows the simulation results of modeling using TOM and the POM. The POM was performed in SAS Proc Logistic and Proc NLMIXED, with practically identical results. In the scenarios where the proportional odds hold ($\gamma = 0$; flat trend), mean and median common estimated odds ratios are very close to the true odds ratio (3.0). The observed coverage for the 95% confidence intervals was between 93% and 96%. In the scenarios where the trend odds hold, mean and median common estimated odds ratios are close to the odds ratio of the first cut-point due to higher frequency of lower titers. Not surprisingly, coverage was poor for this situation, as the confidence interval for the estimate common odds does not cover three levels of true odds ratios.

In the scenarios where the proportional odds hold (flat trend), mean and median common estimated odds ratios are very close to the true odds ratio (3.0), with acceptable coverage for the parameters 94%–96% ($\gamma = 0$ and $\beta = 1.1$). In the scenarios where the trend odds hold, the mean estimated odds ratio per cut-point is higher than expected for smaller samples but approaches the true odds ratio as the

Copyright © 2012 John Wiley & Sons, Ltd.

Table II. Result of simulations for the trend odds model with SAS Proc NLMIXED and proportional odds model with SAS Proc Logistic or Proc NLMIXED.

	Trend odds hold				Proportional odds hold			
Parameters	1	2	3	4	1	2	3	4
Sample size per group	50	100	400	1000	50	100	400	1000
True odds ratio								
Cut-point 1	1.5	1.5	1.5	1.5	3.0	3.0	3.0	3.0
Cut-point 2	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
Cut-point 3	6.0	6.0	6.0	6.0	3.0	3.0	3.0	3.0
True parameters								
β	0.4	0.4	0.4	0.4	1.1	1.1	1.1	1.1
γ	0.7	0.7	0.7	0.7	0.0	0.0	0.0	0.0
Estimated mean odds ratio								
POM	2.52	2.44	2.35	2.34	3.30	3.14	3.04	3.01
TOM, cut-point 1	1.63	1.58	1.52	1.51	3.93	3.35	3.10	3.04
TOM, cut-point 2	3.37	3.17	3.03	3.01	3.42	3.18	3.05	3.02
TOM, cut-point 3	7.99	6.81	6.14	6.07	3.38	3.17	3.03	3.01
Estimated median odds ratio								
POM	2.34	2.32	2.34	2.34	3.07	2.99	3.01	3.01
TOM, cut-point 1	1.48	1.50	1.50	1.51	3.11	2.96	3.04	3.02
TOM, cut-point 2	3.08	3.01	3.00	3.00	3.10	2.99	3.02	3.00
TOM, cut-point 3	6.17	6.06	5.97	5.99	3.06	3.02	2.99	2.99
Coverage (%)								
POM, cut-point 1	79	60	9	0	95	93	95	96
POM, cut-point 2	88	83	53	13	95	93	95	96
POM, cut-point 3	30	7	0	0	95	93	95	96
TOM, cut-point 1	96	95	94	96	95	95	95	96
TOM, cut-point 2	95	94	95	96	95	93	95	96
TOM, cut-point 3	96	94	95	95	96	94	94	95
LRT (% significant)								
POM	65	83	100	100	85	99	100	100
TOM	83	98	100	100	78	97	100	100

Simulated scenarios considering four levels of antibody titers and 1000 repeats. Non-exposed group outcome follows a multinomial (n (0.50,0.40,0.092,0.008)).

sample increases. The median estimated odds ratio is close to the true odds ratio regardless of the sample size. Coverage for the parameters is acceptable, from 95% to 96% ($\gamma = 0.7$ and $\beta = 0.4$).

Test size and power were investigated in the simulations by varying sample size, β, and γ. Overall, when using the likelihood ratio test with the intercept-only model as a reference, the models performed well. The percentage of simulated models that was significant using this test is reported in Table II. Initial simulation of power were obtained with a fixed sample size of 150 per group, a fixed β of log (1.5), and varying values of e^{γ}, which can be thought of a multiplicative factor. As e^{γ} becomes larger, the TOM becomes more powerful than POM, with the cross-over point being when e^{γ} is around 1.25 (Figure 3; Supporting Information). In this example, that would represent a series of odds ratios of 1.5, 1.875 (or 1.5 times 1.25), and 2.34 (or 1.875 times 1.25). Additional simulation of power was obtained with a fixed sample size of 150 per group, a fixed $e^{\gamma} = 1.0$ (proportional odds is true; flat trend), and varying β. The type I error rate of the models stayed in an expected level around 0.05. The level of the type I error for TOM was reasonable. The POM presented a superior simulated power in this setting where there is a real proportionality of odds with no trend (Figure 4; Supporting Information).

6. Application to swine influenza data

Zoonotic influenza A viruses, such as avian and swine influenza, have caused many human pandemics. Introduction of avian influenza to humans was thought to be the cause of pandemics in 1918–1919, 1957–1958, and 1968–1969. These human pandemics left more than 20 million people dead. A swine influenza H1N1 outbreak in 1976 resulted in thousands of infected humans [12, 13]. Influenza virus

Copyright © 2012 John Wiley & Sons, Ltd.

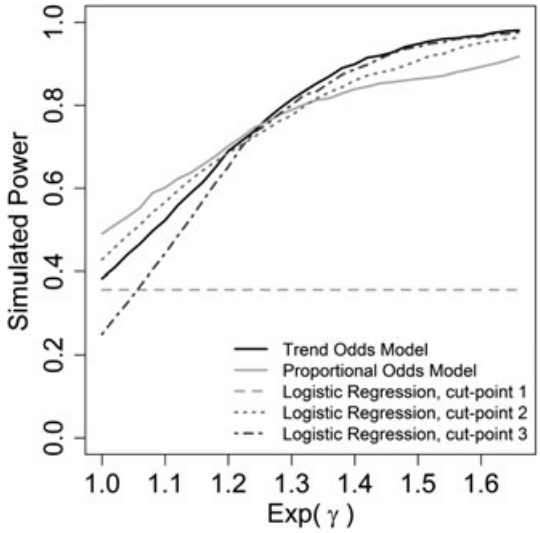


Figure 3. Simulation of power with the non-exposed group outcome following a multinomial($n = 150$, (0.30, 0.26, 0.24, 0.20)), β of log(1.5), sample size of 150 per group, and 1000 replications.

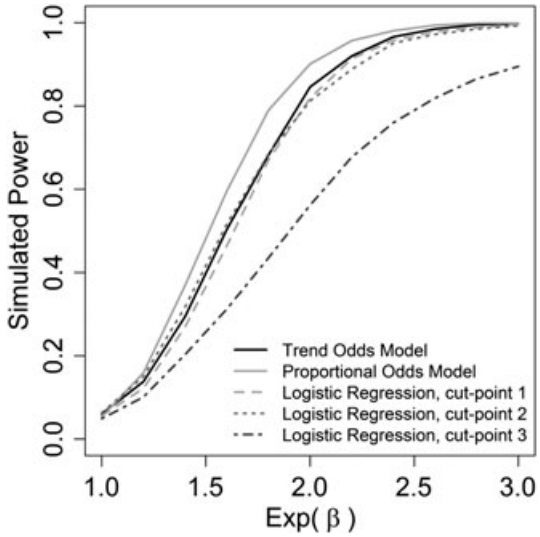


Figure 4. Simulated power with the non-exposed group outcome following a multinomial(n(0.30,0.26, 0.24,0.20)), γ of 0 (no trend in odds ratios), sample size of 150 per group, and 1000 replications.

transmission is complex. Viral shedding depends on several factors. For example, children and individuals who are immunocompromised and symptomatic may shed higher titers than other individuals, whereas people with subclinical or asymptomatic infection may shed lower titers than others. In addition to the viral shedding variation, there are several possibilities for contact with the virus. Transmission may occur by contacting aerosolized respiratory droplets or by touching respiratory secretions-contaminated objects [14]. Influenza can also be transmitted interspecies. Swine are thought to have a major role in interspecies transmission because they have receptors for both human and avian influenza viruses. For this reason, swine are considered a 'mixing vessel host for creation of novel reassortant progeny virus' [15, 16].

In some epidemiological studies of zoonotic influenza, researchers try to understand the etiology of the disease by retrospectively studying populations considered at risk. These researchers try to link infections to certain exposures or risk factors. Human response to influenza virus infection involves antibody response that can be used as evidence of previous infection. Using antibody titers is often an important diagnostic tool in epidemiological assessment of infections, especially in studies where infection may be subclinical. The use of some laboratory methods based on titers often produces data that are categorized

Copyright © 2012 John Wiley & Sons, Ltd.

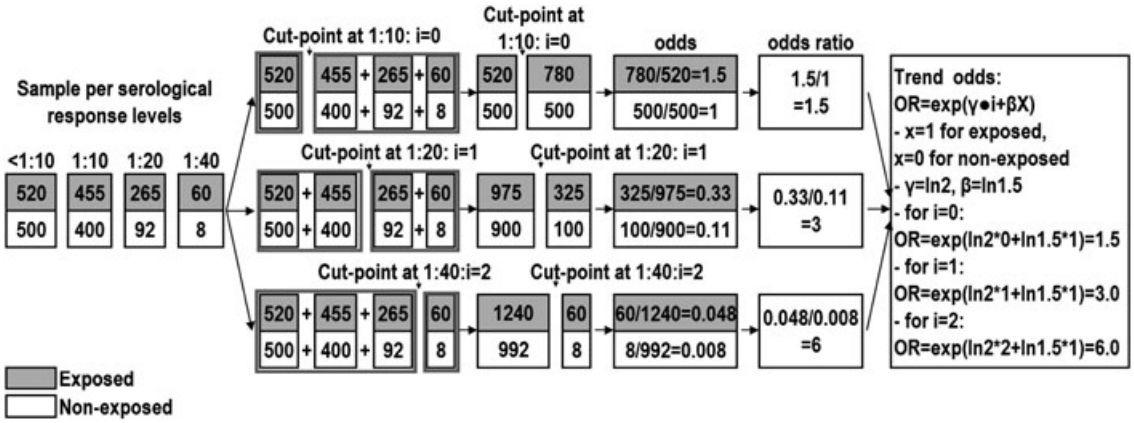


Figure 5. An illustration of the trend odds concept.

into ordinal levels. Examples of these methods are microneutralization and hemagglutination inhibition [17] where dilutions are reported. The ordinal titer categories are recorded on the basis of the sequence of dilution used (e.g., '$< 1:10$', '$1:10$', '$1:20$', and '$1:40$'). The inverse of the dilutions is usually equal spaced in the log scale. The first recorded level represents antibody concentration that cannot be detected at the first dilution (e.g., '$< 1:10$'). The second recorded level represents antibody concentration that is detected at the first dilution but cannot be detected at the second dilution (e.g., '$1:10$'), and so on. In general, the data would fall into $K+1$ categories based on K cut-points.

For didactical reasons, we start by illustrating the trend odds concept with simple hypothetical frequencies across titer levels (Figure 5). Suppose a cross-sectional study obtained exposure status (exposure versus non-exposure) and levels of antibody titers ($< 1:10, 1:10, 1:20$, and $1:40$). In the general TOM, any monotonic scalar set, t_i, could potentially be adopted. It is, however, of practical use to consider structures that eases the model interpretation, provided that they adequately fit the data. Scalar sets starting at 0 are convenient so that β is interpreted as the baseline odds ratio (the odds ratio at the first cut-point, $\theta = e^{\beta}$). That is,

$$\beta x = \Psi_{ix} - \Psi_{i0}. \tag{27}$$

In addition, one may consider a set such that $t_i = i-1$ (where $t_i = 0, 1, 2, \ldots I$), for a model interpretation based on increments in the ordinal observed variable Y. When data are collected in equal spacing, it may be reasonable to say that $i \cdot k = c_i - c_0$, where k is an unknown constant. Related odds ratio can then be calculated from the baseline odds ratio using $\delta = e^{\gamma}$, so that, for $i = 1, 2, \ldots, I$,

$$\theta_i = \theta\delta^{i-1}. \tag{28}$$

In this simple example, there are three cut-points and three possible odds ratios. In our hypothetical example, the odds ratios present an increasing trend with a positive γ giving a $\delta > 1$ (in this case, $\delta = 2$). The baseline θ is 1.5, meaning that exposed participants have at least 50% higher odds of having higher antibody titers. The odds ratio doubled with increasing ordinal titer levels. The next odds ratios can be calculated as $1.5 \cdot 2 = 3.0$ and $3.0 \cdot 2 = 6.0$.

To illustrate the TOM and its interpretation in practice, we now use actual real data from a cross-sectional occupational epidemiology study among farmers, meat processing workers, veterinarians, and healthy controls. The data were collected from 2002 to 2004 in the Center for Emergency Infectious Diseases at the University of Iowa [15]. Age was collected at enrolment using a questionnaire, and sera were collected and tested according to the Centers for Disease Control and Prevention hemagglutination inhibition serologic protocol. In this example, we use age (in decades) to predict the hemagglutination inhibition titers against swine H1N2 (A/Swine/WI/R33F/01) virus. Modeling was performed in SAS with Proc NLMIXED. A reference code is provided (Appendix A; Supporting Information). The TOM indicates a significant negative trend in odds ratios ($\gamma = 0.515$, $SE(\gamma) = 0.052$, p-value $= 0.01$). Odds of having increased antibody titers against swine H1N2 are significantly higher for older participants but the effect decreases with increasing antibody titers. As seen in Table III, estimated odds ratios are 1.67 at baseline (first cut-point, $1:10$, β=0.133, $SE(\beta) = 0.085$, p-value < 0.01) and decreases as titers go

Copyright © 2012 John Wiley & Sons, Ltd.

Table III. Data on antibody titers against swine H1N2 influenza, modeled as a function of age (in decades).

		Antibody titer level				
Odds ratio (per decade age)	Total ($n = 349$)	$< 1:10$ ($n = 214$)	$1:10$ ($n = 56$)	$1:20$ ($n = 38$)	$1:40$ ($n = 28$)	$\geqslant 1:80$ ($n = 13$)
Estimated from individual logistic regression models	—	—	1.68	1.39	1.33	1.25
Estimated from Trend Odds model (TOM)	—	—	1.67	1.47	1.28	1.12

Example of univariate comparison using the trend odds model fit in SAS (model has proportional odds assumption violated and there is a trend in odds ratio).

up ($e^{\gamma} = \delta = 0.88$). That is, the increase in one level in antibody titer decreases the odds ratio in 12%. In practice such information can raise several hypotheses about the transmission of the virus, as mentioned in the succeeding section.

7. Discussion

When fitting data with statistical models, there is often a tradeoff between model simplicity and adequate fit. Although the POM has a somewhat simple interpretation, it may give inadequate fit when its proportionality assumption is violated. In this case, the TOM is a type of constrained cumulative odds model that may be a good compromise when dealing with this tradeoff. Specifically, it may give improved model fit at the added complexity of only one additional parameter.

This paper examines the relationship between the TOM and underlying latent distributions, especially the logistic distribution. The TOM can do an adequate job in describing non-proportional odds even when the latent variable distribution is unknown. Although we did not explore this issue in this paper, it is likely that the choice of the t_i could make a difference in the adequacy of model fit, depending on how different the t_i is from the c_i and what the underlying distribution is. If the t_i happens to be chosen to be proportional to, or a linear transformation of, the c_i, then the incorrect specification could be accommodated by the model through the β and γ parameters, in order to provide meaningful odds ratios. Similarly, if the t_i is highly correlated with the c_i, the model may still give reasonable estimates. In future work we hope to examine the robustness of the choice of the t_i.

As a matter of convenience in practice, it may work well to assign $t_1 = 0$ so that the estimate of β will correspond to the log odds ratio at the lowest cut-point. This is why we set $t_i = i - 1$ in our two data examples. It is possible, however, to use other parameterizations. For example, in some applications with clear neutral central level, it may make sense to set a middle t_i value to 0, so that β would be a central cumulative odds ratio.

Note that the TOM, with $\beta > 0$, does not guarantee that the probability of high values of Y increases with X at all cut-points. For the distribution shown in Figure 1B, for example, the odds ratio would be 0.08 at $Y^* = 5$, 1.0 at $Y^* = 10$, and 12.18 at $Y^* = 15$ (or respectively log odds of -2.5, 0, and 2.5). Hence, if the variances are different but the means are the same at two different values of X, then one is likely to see odds ratios that cross the value of 1 as cut-points go from smallest to largest.

Analyses based on the entire spectrum of the ordinal outcome are particularly important in studies that seek the etiological understanding of the disease spread and identification of high-risk groups. These analyses serve as justification and motivation for the development of the TOM. Whereas the POM allows researchers to identify populations at a uniform higher risk of worse outcomes, the TOM allows the identification of populations whose wider variability results in a non-uniform but increasingly higher risk of worse outcomes. The interpretation of both the POM and TOM is based on odds ratios, a statistic that is familiar to most health professionals. The identification of a trend has public health implications. It is important not only to find populations with constantly higher odds of worse outcomes, but also to identify populations whose wider variability yields higher odds of worse outcomes for certain individuals. Trends in odds can also indicate an underlying higher variability that may be caused by lurking variables, generating further hypothesis to be tested. For example, in our influenza real data example, we saw a decreasing risk of higher titers with increasing age. That may be associated to the

Copyright © 2012 John Wiley & Sons, Ltd.

fact that antibody titers tend to vary with time from infection. That may also indicate that younger populations have other risk factors that increase the variability by yielding higher odds of worse outcomes for certain individuals. Short term, this information is important in the implementation of preventive measures such as the use of protective equipment [18]. Long term, it can increase knowledge about the disease by hypothesis generation for future research.

For certain applications, the use of dichotomization has scientific justification. In this case, optimal cut-point methods such as the ROC curve technique [19] may be an alternative. When there is no strong reason to dichotomize the data, fitting multiple binary logistic models for different cut-points may be useful as an exploratory step [20], before fitting the TOM.

Although not the scope of this work, it is important to note that there are some alternative analyses to the model proposed. One can, for example, explore different link families to obtain parallelism or proportionality, or seek to estimate the underlying latent distribution parameters (i.e., predicting location and scale) [2, 21].

In this paper, we have focused on models with a single predictor variable. With multiple covariates, Peterson and Harrell [3] presented models to accommodate a mixture of covariates that do and do not adhere to the proportional odds assumption. Extending this idea, it is feasible to allow the effect of some of the covariates to adhere to the proportional odds assumption, some to adhere to the TOM assumptions, some to require non-trend constraints, and some to be completely unconstrained. Furthermore, it is unclear whether all covariates that are fit with the TOM assumption should have the same set of scalars or whether they should vary, and, if the latter, whether this can be data dependent. Further study on these issues is needed.

Acknowledgements

This work was supported by the NIOSH Occupational Epidemiology Training Program within the Heartland Center for Occupational Health and Safety (T42 OH008491; R. William Field, PI) and US Armed Forces Health Surveillance Center–Global Emerging Infections Surveillance Operations and National Institute of Allergy and Infectious Diseases (R01 AI068803; Gregory C Gray, PI)). The authors thank Dr. Joseph Lang and Dr. Peter McCullagh for their comments and insights. We also thank Dr. Gregory Gray and Dr. Kendall Myers for the use of the swine influenza data.

References

1. Aitchison J, Silvey SD. The generalization of probit analysis to the case of multiple responses. *Biometrika* 1957; **44**(1/2):131–140. DOI: 10.1093/biomet/44.1-2.131.
2. McCullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)* 1980; **42**(2):109–142.
3. Peterson B, Harrell FE Jr. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1990; **39**(2):205–217.
4. O'Connell AA. *Logistic Regression Models for Ordinal Response Variables. Quantitative Applications in the Social Sciences*. SAGE Publications: Thousand Oaks, California, 2006.
5. Capuano AW, Dawson JD, Gray GC. Maximizing power in seroepidemiological studies through the use of the proportional odds model. *Influenza and Other Respiratory Viruses* 2007; **1**(3):87–93. DOI: 10.1111/j.1750-2659.2007.00014.x.
6. Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology* 1997; **26**(6):1323–1333. DOI: 10.1093/ije/26.6.1323.
7. Greenland S. Alternative models for ordinal logistic regression. *Statistics in Medicine* 1994; **13**(16):1665–1677.
8. Lall R, Campbell MJ, Walters SJ, Morgan K. A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods Medical Research* 2002; **11**(1):49–67. DOI: 10.1191/0962280202sm271ra.
9. Agresti A. *Analysis of Ordinal Categorical Data*, (2nd edn). Wiley: Hoboken, New Jersey, 2010.
10. Agresti A. *Categorical Data Analysis*, (2nd edn). Wiley-Interscience: New York, 2002.
11. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006; **25**(24):4279–4292. DOI: 10.1002/sim.2673.
12. Kawaoka Y, Krauss S, Webster RG. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. *Journal of Virology* 1989; **63**(11):4603–4608.
13. Reid AH, Fanning TG, Janczewski TA, Taubenberger JK. Characterization of the 1918 'Spanish' influenza virus neuraminidase gene. *Proceedings of the National Academy of Sciences of the United States of America* 2000; **97**(12): 6785–6790. DOI: 10.1073/pnas.100140097.
14. Van-Tam J, Sellwood C. *Introduction to Pandemic Influenza*. CABI: Wallingford, Oxfordshire; Cambridge, MA, 2010.
15. Myers KP, Olsen CW, Setterquist SF, Capuano AW, Donham KJ, Thacker EL, Merchant JA, Gray GC. Are swine workers in the United States at increased risk of infection with zoonotic influenza virus? *Clinical Infectious Diseases* 2006; **42**(1):14–20. DOI: 10.1086/498977.

Copyright © 2012 John Wiley & Sons, Ltd.

16. Webster RG, Sharp GB, Claas EC. Interspecies transmission of influenza viruses. *American Journal of Respiratory and Critical Care Medicine* 1995; **152**(4 Pt 2):S25–S30.
17. Meijer A, Bosman A, van de Kamp EE, Wilbrink B, Du Ry van Beest Holle M, Koopmans M. Measurement of antibodies to avian influenza virus A(H7N7) in humans by hemagglutination inhibition test. *Journal of Virological Methods* 2006; **132**(1-2):113–120. DOI: 10.1016/j.jviromet.2005.10.001.
18. Ramirez A, Capuano AW, Wellman DA, Lesher KA, Setterquist SF, Gray GC. Preventing zoonotic influenza virus infection. *Emerging Infectious Diseases* 2006; **12**(6):996–1000.
19. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* 2004; **159**(9):882–890. DOI: 10.1093/aje/kwh101.
20. Bender R, Grouven U. Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology* 1998; **51**(10):809–816. DOI: 10.1016/S0895-4356(98)00066-3.
21. Poon WY. A latent normal distribution model for analysing ordinal responses with applications in meta-analysis. *Statistics in Medicine* 2004; **23**(14):2155–2172. DOI: 10.1002/sim.1814.

Copyright © 2012 John Wiley & Sons, Ltd.