

COMPARISON OF SYMMETRY AND CONDITIONAL SYMMETRY MODELS AS TOOLS FOR RATER AGREEMENT MEASURES

A. O. Adejumo
Department of Statistics
University of Ilorin, Ilorin - Nigeria
ao123adejumo@yahoo.co.uk

Christian Heumann and Helge Toutenburg
Institute of Statistics
Ludwig-Maximilian-University
Munich, Germany.

Abstract.

In almost all life or social science researches, subjects are classified into categories by raters, interviewers or observers. Many approaches have been proposed by various authors for analyzing data or the results obtained from these raters. Symmetry and conditional symmetry models are models designed for square tables like the one arising from the raters results. Conditional symmetry model which possessed an extra parameter for the off-diagonal cells is a special case to symmetry system. In this paper, we examined these models on structure of agreement as well as their fittings. Generalized linear model (GLM) approach was used to model the loglinear model forms of these models with empirical examples. We observed that conditional symmetry based on extra parameter gave a tremendous improvement to the significant level of the test statistics over that of symmetry model. Hence when the off-diagonal cells are to be preserved as in the example, conditional symmetry model is preferred for raters agreement modelling.

1. Introduction

Symmetry and conditional symmetry models were originally proposed for any square table. For agreement measurement, square contingency table can be used to display joint ratings of two raters. Testing symmetry model is an important preliminary analysis for further analysis which requires symmetric table. Conditional symmetry model is a special case to symmetry which introduced a special parameter in order to preserve the off-diagonal cells. The main objective of this paper is to observe the effect of this extra parameter in conditional symmetry on the structure of agreement as well as the goodness of fit statistics over that of symmetry model.

In nearly all researches that involves ratings, measurements or diagnosis of subjects by various raters, observers or Pathologists, the researchers are already aware that the most important measurement of error or bias is the raters' variability involved in such studies (Fletcher and Oldham, 1962). At the initial stage of the experiment reliability test or studies has to be conducted among the raters, interviewers or observers involved to assess the level of raters' variability in the measurement procedure to be used in data acquisition. The data arising from such studies can be of any form - quantitative or qualitative, depending on the measuring scales involved. In a situation where the data arising from the studies are quantitative, tests for interobserver bias and measures of interobserver agreement are usually obtained from standard analysis of variance (ANOVA) mixed models or random effect model. Hypothesis testing of observers effects are used to investigate differences in the mean response among observers and estimates of intraclass correlation coefficients are used to measure interobserver reliability (Chinchilli et al. 1996). Improvements have been made on standard ANOVA model as proposed by Crowder and Hand (1990) and Toutenburg (2002). However, data arise which are qualitative/categorical in which the response variable is classified into either nominal or ordinal categories. For nominal data, as reviewed by Banerjee et al. (1999) and Shoukri (2004), a large numbers of estimation and testing procedures like the Cohen Kappa (Cohen, 1960), the weighted Kappa (Cohen, 1968), the

intraclass Kappa (Block and Kraemer, 1989), Tetrachoric correlation coefficient, weighted least squares (WLS) method for correlated kappa (Barnhart and Williamson 2002) and many other improved methods on kappa statistic were given. On the side of ordinal data, most of the medical diagnoses data often involve responses taken on an ordinal scale and many of which are very subjective. As it is pointed out by some authors with ordinal data, an intermediate category will often be subjective to more misclassification than an extreme category because there are two directions in which to err away from the extremes. Therefore a modified kappa statistic which accounts for severity of discordance or size of discrepancy is better suited for ordinal data. The weighted kappa (Cohen, 1968) offers such modification. Darroch and McCloud (1986); Tanner and Young (1985a); Adejumo (2005) pointed out the shortcomings of kappa-like statistics. These include; loss of information from summarizing the table by a single number, sensitivity of value to the form of the marginal distributions and subsequent dangers in comparing values of kappa between two tables. To this effect, some methods that involve modelling loglinear and latent class models for raters agreement are proposed by various authors such as Tanner and Young (1985a,1985b), Agresti (1988,1992); Agresti and Lang (1993); Graham (1995); Barnhart and Williamson (2002); Adejumo (2005) among others. A square table can be used to display joint ratings of two raters or observers. Two matters are usually considered for this type of table. Firstly, one can analyze differences in the marginal distributions. For ordered response categories, there is usually interest in whether classifications by one rater tend to be higher than those by the other rater. Secondly, one can analyze the extent of subject-wise agreement between raters, which involves investigating the frequency of main-diagonal occurrence within the joint distribution of the ratings. In this paper, we shall examine the two models under consideration on the assessment of the structure of raters agreement. We shall have the algorithms for modelling these models written in accordance with generalized linear modelling (GLM) techniques for loglinear model. These algorithms can be executed with Splus, R or SPSS programs. In sections 2 and 3 we present the approach for modelling raters agreement and description for GLM respectively while section 4 present empirical examples. And lastly in section 5, we summarize the results and make some conclusions.

2. Modelling raters agreement measure

Tanner and Young (1985a) proposed a modelling structure of agreement, by considering loglinear models to express agreement in terms of components, such as chance agreement and beyond chance agreement. Using the loglinear model approach, one can display patterns of agreement among several observers, or compare patterns of agreement when subjects are stratified by values of a covariate. Assuming there are n subjects who are rated by the same k raters ($k > 2$) into m nominal categories, Tanner and Young (1985a) express chance agreement, or statistical independence of the ratings, using the following loglinear model representation as:

$$\log(v_{i_1 i_2 \dots i_k}) = u + u_{i_1}^{R_1} + u_{i_2}^{R_2} + \dots + u_{i_k}^{R_k}, \quad (1)$$

$$i_1, i_2, \dots, i_k = 1, \dots, m$$

where $v_{i_1 i_2 \dots i_k}$ is the expected cell counts in the $i_1, i_2, \dots, i_k$ th cell of the joint k -dimensional cross classification of the ratings, u is the overall effect (the intercept), $u_{i_k}^{R_k}$ is the effect due to categorization by the k th rater in the i_k th category ($k=1, \dots, K$; $i_k=1, \dots, m$), and. They also presented a useful generalization of the independence model that incorporated agreement beyond chance in the following fashion:

$$\log(v_{i_1 i_2 \dots i_k}) = u + u_{i_1}^{R_1} + u_{i_2}^{R_2} + \dots + u_{i_k}^{R_k} + \delta_{i_1 i_2 \dots i_k}, \quad (2)$$

$$i_1, i_2, \dots, i_k = 1, \dots, m$$

The additional term $\delta_{i_1 i_2 \dots i_k}$ represents agreement beyond chance for the $i_1, i_2, \dots, i_k$ th cell.

Agresti (1988) also proposed a model in form of loglinear model for agreement plus linear-by-linear association, which is the combination of Tanner and Young's model and the uniform

association model of Goodman (1979) for bivariate cross-classification of ordinal variables. The model is

$$\log(v_{ij}) = u + u_i^{R_1} + u_j^{R_2} + \beta\lambda_i\lambda_j + \delta_{ij}, \quad (3)$$

$$i, j = 1, \dots, m$$

where $\lambda_1 < \dots < \lambda_m$ are fixed scores assigned to the response categories, and the u 's and v 's are as defined in equation (1).

Following the same structure of loglinear model as presented above by Tanner and Young (1985a); Agresti (1988) in their respective works, we shall also use the loglinear model forms of symmetry and conditional symmetry models to model the structure of agreement by presenting the structures and properties of the models. These two models are originally designed for any square tables. Also due to the fact that testing symmetry model is an important preliminary analysis for further analysis which requires symmetric table, we shall observe how conditional symmetry model stands as a better alternative to ordinary symmetry model.

For a given $m \times m$ contingency table, let π_{ij} be the probability of cell i, j . There exist symmetry if

$$\pi_{ij} = \pi_{ji} \quad (4)$$

whenever $i \neq j$. Let v_{ij} be the expected value of the cell i, j as defined before, such that

$$v_{ij} = n\pi_{ij} \quad (5)$$

In a 2×2 table, the hypothesis of marginal homogeneity ($\pi_{i+} = \pi_{+i}$) is equivalent to the hypothesis of symmetry ($\pi_{ij} = \pi_{ji}$). However for more general square ($m \times m$) tables, symmetry is a much stronger hypothesis for agreement than marginal homogeneity. Symmetry model as loglinear model (Agresti, 1990), Adejumo (2005) is

$$\log(v_{ij}) = u + u_i^{R_1} + u_j^{R_2} + u_{ij}^{R_1R_2}, i, j = 1, \dots, m \quad (6)$$

where

$$u \text{ is the intercept, } u_{ij}^{R_1R_2} = u_{ji}^{R_1R_2}, \quad \sum_{i=1}^m u_i^{R_1} = 0 \quad \sum_{j=1}^m u_j^{R_2} = 0,$$

and

$$\sum_{j=1}^m u_{ij}^{R_1R_2} = 0 \quad \text{for } i = 1, 2, \dots, m$$

There are no superscripts on the main or marginal effect terms because they are the same for rows and columns, that is, $u_i^{R_1} = u_j^{R_2}$ when $i=j$. In other words, the row and column margins are equal, that is, $v_{i+} = v_{+i}$. The likelihood equations are

$$\hat{v}_{ij} + \hat{v}_{ji} = n_{ij} + n_{ji} \text{ for all } i \text{ and } j$$

The kernel of the log-likelihood is

$$\sum_{ij} \log(v_{ij}) = n_{++}u + \sum_i (n_{i+} + n_{+i})u_i + \sum_{ij} \left\{ \frac{n_{ij} + n_{ji}}{2} \right\} u_{ij}. \quad (7)$$

Maximizing this equation yields the following expected cell values

$$\hat{v}_{ij} = \begin{cases} \frac{n_{ij} + n_{ji}}{2} & \text{if } i \neq j \\ n_{ii} & \text{if } i = j, \end{cases} \quad (8)$$

The goodness of fit statistics for symmetry model are: Pearson's chi-square statistic given as

$$\chi^2 = \sum_{i>j} \frac{(n_{ij} - n_{ji})^2}{(n_{ij} + n_{ji})} \quad (9)$$

and the likelihood ratio statistic is

$$G^2 = \sum_{i \neq j} n_{ij} \log \left(\frac{2n_{ij}}{n_{ij} + n_{ji}} \right) \quad (10)$$

The degrees of freedom (df) for the residuals of these statistics can be obtained as (number of cells) minus (number of non-redundant parameters) or (number of off-diagonal cells) minus (number of unique parameters), mathematically

$$\begin{aligned} df &= m(m-1) - \frac{m(m-1)}{2} \\ &= \frac{m(m-1)}{2} \end{aligned}$$

The two statistics have asymptotic χ^2 distribution with the above degrees of freedom under the null hypothesis of independence in the symmetry model. Now consider conditional symmetry model which was proposed by McCullagh (1978) and has an additional parameter over symmetry, is given as

$$\pi_{ij} = \begin{cases} \theta\phi_{ji} & \text{if } i < j, \\ \phi_{ii} & \text{if } i = j, \\ (2-\theta)\phi_{ij} & \text{if } i > j, \end{cases} \quad (11)$$

with $\phi_{ij} = \phi_{ji}$ and $\sum_i \sum_j \phi_{ij} = 1$ for $i, j = 1, 2, \dots, m$.

Conditional symmetry model is a palindromic invariant and not a permutation invariant Adejumo (2005). However, the reverse permutation applied to conditional symmetry model (11) yields

$$\pi_{ij} = \begin{cases} (2-\theta)\phi_{m-i+1, m-j+1} & \text{if } i < j, \\ \phi_{m-i+1, m-i+1} & \text{if } i = j, \\ \theta\phi_{m-i+1, m-j+1} & \text{if } i > j, \end{cases} \quad (12)$$

with $\phi_{m-i+1, m-i+1} = \phi_{m-j+1, m-j+1}$, $\sum_{i=1}^m \sum_{j=1}^m \phi_{m-i+1, m-j+1} = 1 \quad \forall i = j = 1, 2, \dots, m$.

The loglinear version of conditional symmetry model (Agresti 1990, Adejumo 2005) is defined as

$$\log(v_{ij}) = u + u_i^{R_1} + u_j^{R_2} + u_{ij}^{R_1 R_2} + d(i < j), \quad (13)$$

where u is the intercept, $u_{ij}^{R_1 R_2} = u_{ji}^{R_1 R_2}$, $\sum_{j=1}^m u_j^{R_1} = 0$, $\sum_{j=1}^m u_{ij}^{R_1 R_2} = 0$ for $i = 1, 2, \dots, m$ and $I(i < j)$ is the indicator function defined as

$$I(i < j) = \begin{cases} 1 & \text{if } i < j, \\ 0 & \text{if } i \geq j. \end{cases} \quad (14)$$

Based on this additional parameter, the model is mainly for ordered classification when symmetry may not hold, often either

$$\pi_{ij} > \pi_{ji} \quad \forall i < j \quad \text{or} \quad \pi_{ij} < \pi_{ji} \quad \forall i < j$$

The model (11) implies that for all $i < j$, if R_1 denotes the row number and R_2 the column number of an observation made according to distribution $\{\pi_{ij}\}$, then the conditional interpretation of model (13) is

$$P(R_1 = i, R_2 = j / R_1 < R_2) = P(R_1 = j, R_2 = i / R_1 > R_2) = \phi_{ij}.$$

This means the cells probabilities above the main diagonal are mirror image of the cells probabilities below it. Based on this property, the model is called Conditional symmetry model.

The likelihood equations for the model are

$$\hat{v}_{ij} + \hat{v}_{ji} = n_{ij} + n_{ji} \text{ for all } i \text{ and } j$$

and

$$\sum_{i < j} \hat{v}_{ij} = \sum_{i < j} n_{ij} \text{ or } \sum_{i > j} \hat{v}_{ij} = \sum_{i > j} n_{ij}$$

The solution of these equations that satisfy the model is

$$\hat{\tau} = \log \frac{\sum_{i < j} n_{ij}}{\sum_{i > j} n_{ij}}$$

$$\hat{v}_{ij} = \begin{cases} \frac{\exp(\hat{\tau})(n_{ij} + n_{ji})}{\exp(\hat{\tau}) + 1} & \text{if } i < j, \\ \frac{(n_{ij} + n_{ji})}{\exp(\hat{\tau}) + 1} & \text{if } i > j, \\ n_{ii} & \text{if } i = j, \end{cases} \quad (15)$$

for $i = j = 1, 2, \dots, m$.

The residual degrees of freedom (df) is

$$df = \frac{(m+1)(m-2)}{2},$$

which is one less than that of the symmetry because of the additional parameter $\tau(i < j)$. Symmetry model is a special case to Conditional symmetry if $\tau = 0$.

3. Fittings of the models

Generalized linear model (GLM) procedure would be used to fit these models. GLM has three important components, the random, the systematic and the link components, for detail see McCullagh and Nelder (1989); Adejumo (2005). Poisson sampling is mostly assumed when fitting GLM to categorical data with $m \geq 2$. The log likelihood function is

$$l(\theta, \phi) = \sum_{i=1}^n \left(\frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right) \quad (16)$$

where θ subsumes all of the θ_i . It could also be written as a function of β and ϕ because (given the x_i), β determines all the θ_i . The main way of maximizing β is by maximizing (16). The fact that $G(\mu_i) = x_i \beta$ suggests a crude approximation estimate: regress $G(y_i)$ on x_i , perhaps modifying y_i in order to avoid violating range restrictions (such as taking $\log(0)$), and accounting for the differing variances of the observations.

Fisher scoring iteration is the widely used technique for maximizing the GLM likelihood over β . The basic step is

$$\beta^{(k+1)} = \beta^k - \left\{ E \left(\frac{\partial^2 l}{\partial \beta \partial \beta'} \right) \right\}^{-1} \frac{\partial l}{\partial \beta} \quad (17)$$

which can also be written as,

$$\beta^{(k+1)} = \beta^k + \left\{ -E \left(l''(\beta^{(k)}) \right) \right\}^{-1} l'(\beta^{(k)}) \quad (18)$$

where l is the loglikelihood function for the entire sample $y_1, \dots, y_N$ and the expectations are taken with $\beta = \beta^{(k)}$. This is the same as Newton step, except that Hessian of l is replaced by its expectation. Fisher scoring simplifies to

$$\beta^{(k+1)} = (X'WX)^{-1} X'WY \quad (19)$$

where W is a diagonal matrix with

$$W_{ii} = (G'(\mu_i)^2 b''(\theta_i))^{-1} \quad (20)$$

and

$$Z_i = (Y_i - \mu_i)G'(\mu_i) + x_i\beta \quad (21)$$

Both equations (20) and (21) use $\beta = \beta^{(k)}$ to derive values of $\theta_i^{(k)}$ and $\mu_i^{(k)}$. The iteration (19) is known as "iteration reweighted least squares", or IRLS. The weights W_{ii} have the usual interpretation as reciprocal variances: $b''(\theta_i)$ is proportional to the variance of Y_i and the $G'(\mu)$ factor in Z_i is squared in W_{ii} . Fisher scoring may also be written as

$$\beta^{(k+1)} = \beta^{(k)} + (X'WX)^{-1} X'WZ^* \quad (22)$$

where

$$Z_u^* = (Y_i - \mu_i)G'(\mu_i)$$

We need to describe the structure of variables involved in the modification of symmetry model to obtain their estimates as stated in the model. To this effect, we need to create variable that takes on a unique value for each diagonal cell and a unique value of each pair of cells. So for example if $l=5$, the symmetry variables, which are also parameters in the model, we shall have the following,

$$u = \begin{cases} u_1 & \text{if } i = j = 1 \\ u_2 & \text{if } i = j = 2 \\ u_3 & \text{if } i = j = 3 \\ u_4 & \text{if } i = j = 4 \\ u_5 & \text{if } i = j = 5 \\ u_{12} & \text{if } (i, j) = (1, 2)(2, 1) \\ u_{13} & \text{if } (i, j) = (1, 3)(3, 1) \\ \vdots & \vdots \\ u_{45} & \text{if } (i, j) = (4, 5)(5, 4) \end{cases} \quad (23)$$

where u_1 will be zero. All these variables are treated as nominal variables.

In the case of conditional symmetry model in equation (13) which adds parameter $\pi(i < j)$ to the symmetry model, the $\pi(i < j)$ which is defined by the indicator variable in equation (14) above can be expressed as a matrix of dummies as follows

$$\pi(i < j) = \tau = \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 0 & 0 & 1 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

For example, if $l=5$ we have,

$$\pi(i < j) = \tau = \begin{Bmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{Bmatrix}$$

The Fisher scoring iteration procedure stated above was used to fit each of these models under generalized linear model (GLM) procedure.

For further justification of this research, we incorporated Cohen Kappa statistic (Cohen, 1960) which measures raw agreement in terms of the proportion of the subjects classified into the same category by two raters. The statistic was given as

$$\hat{k}_c = \frac{\hat{\pi}_o - \hat{\pi}_e}{1 - \hat{\pi}_e} \quad (24)$$

where

$$\hat{\pi}_o = \sum_{i=1}^I \pi_{ii} \quad (25)$$

and

$$\hat{\pi}_e = \sum_{i=1}^I \pi_{i+} \pi_{+i} \quad (26)$$

this statistic we used the asymptotic variance formulae of $\hat{k}_c$ given by Fleiss et al. (1969) for the general $I \times I$ tables which is given as

$$\text{var}_n(\hat{k}_c) = \frac{\pi_e + \pi_e^2 - \sum_{i=1}^I \pi_{i+} \pi_{+i} (\pi_{i+} + \pi_{+i})}{n(1 - \pi_e)^2} \quad (27)$$

By assuming a normal distribution, the confidence interval (CI) of size

$$100(1 - \alpha)\% = \hat{k} \pm Z_{1 - \frac{\alpha}{2}} SE(\hat{k})$$

can be obtained for $\hat{k}_c$, where $SE(\hat{k}) = \sqrt{\text{Var}(\hat{k})}$ is the standard error.

In order to justify further the comparison of these two models above, we considered some reasonable and appropriate data sets on this subject to demonstrate implementation.

4. EMPIRICAL EXAMPLES

4.1. Example 1.

Consider the data obtained by two pathologist that assessed 27 patients twice for the presence (Y) or absence (N) of dysplasia as presented by Barnhart and Williamson (2002). The categories used are 1=NN (dysplasia absence on both times), 2=NY (dysplasia absence in the first time but presence in the second time), 3=YN (dysplasia presence in the first time but absence in the second time), 4=YY (dysplasia presence on both times). See the data in Table 1 in the Appendix together with the tables for the parameter estimates with the goodness of fit statistics under the two models in Tables 2 and 3.

4.2. Example 2.

Consider the data arising from the study reported in Holmquist et al. (1967) that investigated the variability in the classification of carcinoma in situ of the uterine cervix in which seven pathologists were requested to separately evaluate and classify 118 slides into one of the the following five categorical scales based on the most involved lesion: 1=negative; 2=Atypical squamous hyperplasia; 3=carcinoma in situ; 4=squamous carcinoma with early stromal invasion; 5=invase carcinoma. These pathologists are labelled with letters A, B, C, D, E, F

and G. We have the cross-classification for pathologists A and G as presented in Table 4 in the Appendix together with the tables for the parameter estimates with the goodness of fit statistics under the two models also in Tables 5 and 6.

5. Result and Conclusion

5.1. Result

We present the summary of results in Table 7 in the Appendix by including column for differences between the respective goodness of fits statistics of symmetry and conditional symmetry models with their degrees of freedom. From the results in Table 7 and figures 1 and 2, due to the inclusion of extra parameter τ we observed that there was a great improvement on the goodness-of-fit statistics. The columns of difference also in Table 7, showed the major contribution of this parameter being included in conditional symmetry model over ordinary symmetry model. In Table 8, we present the estimates as well as the 95% confidence bounds of Cohen Kappa statistic for the two examples. We observed that the degrees of agreement for the two examples under Cohen Kappa statistic in Tables 8 were equally reflected under modelling with the two models in Tables 7. We also present the expected frequencies under each of the models in the Appendix. We also tested these with two other different data on cervical ectopy ratings of two different methods, Visual assessment and Computerized planimetry as given by Barnhart and Williamson (2002) and we have the same conclusion.

5.2. Conclusion

In order to preserve the total in each triangular array but not the marginal distributions in the resulted two-way cross-classified table of the ratings of the raters, we need to fit conditional symmetry model to the data. From the result above we observed that inclusion of the extra parameter τ by the conditional symmetry had a great improvement on the goodness of fit statistics for the rater agreement measure. Conditional symmetry is not only testing the pattern of association, but also whether the distribution between categories is the same whichever rater is better. Therefore, introducing conditional symmetry model which preserves the off-diagonal cells, leads to significant improvement on the results over the symmetry model for raters agreement modelling in symmetric table.

6. References

1. Adejumo, A. O (2005). *Modelling Generalized Linear (Loglinear) Models for Raters Agreement measures*; Peter Lang, Frankfurt am Main, Germany.
2. Agresti, A. (1988). A model for agreement between ratings on an ordinal scales, *Biometrics*, 44, 539- 548.
3. Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.
4. Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statist. Methods Med. Res.*, 1, 201-218.
5. Agresti, A. and Lang, J. B. (1993). Quasi-symmetry latent class models, with application to rater agreement. *Biometrics*, 49, 131-139.
6. Banerjee, M., Capozzoli, M., Mcsweeney, L., and Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measure. *The Cana. J. of. Statist.*, 27(1), 03-23.
7. Barnhart, H. X, and Williamson, J. M. (2002). Weighted least-squares approach for comparing correlated kappa. *Biometrics*, 58, 1012-1019.
8. Block, D. A., and Kraemar, H. C. (1989). 2 x 2 kappa coefficients: Measures of agreement or association. *Biometrics*, 45, 269-287.
9. Chinchilli, V. M., Martel, J. K., Kumanyika, S., and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, 52, 341-353.
10. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Edu. and Psych. Meas.*, 20, 37-46.
11. Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psych. Bull.*, 70, 213-220.
12. Crowder, M. J., and Hand, D. J. (1990). *Analysis of repeated measures*, Chapman and Hall.
13. Darroch, J. N., and McCloud, P. I. (1986). Category distinguishability and observer agreement. *Australian J. Statist.*, 28, 371-388.
14. Fleiss, J. L., Cohen J. and Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psych. Bull.*, 72, 323-327.
15. Fletcher, C. M., and Oldham, P. D. (1962). *Diagnosis in group resaerch*. Chapter 2 of Med. Surv and Clini. Trials, ed. L. J. Witts. 2nd edit. Oxford univ. Press, London.
16. Goodman, L. A. (1979). Multiplicative Models for Square contingency tables with ordered categories. *Biometrika*, 66, 413-418.
17. Graham, P. (1995). Modeling covariate effects in observer agreement studies: The case of nominal scale agreement. *Statist. Med.*, 14, 299-310.
18. Holmquist, N. S., McMahon, C. A. and Williams, O. D. (1967). Variability in classification of carcinoma in situ of the uterine cervix. *Archieves of Pathology*, 84, 334-345.
19. Maclure, M., and Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *Amer. J. Epidemiol.*, 126, 161- 169.
20. McCullagh, P. (1978). A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika*, 65(2), 413-418.
21. McCullagh, P., and Nelder, J. A. (1989). *Generalized linear models*, 2nd Edit. Chapman and Hall.
22. Shoukri, M. M. (2004). *Measures of interobserver agreement*. Chapman and Hall.
23. Tanner, M. A., and Young, M. A. (1985a). Modelling agreement among raters. *J. Amer. Statist. Assoc.*, 80, 175-180.
24. Tanner, M. A., and Young, M. A. (1985b). Modelling ordinal scale agreement . *Psych. Bull.*, 98, 408-415.
25. Toutenburg, H. (2002). *Statistical analysis of designed experiments*. Springer.

APPENDIX

Table 1: Two-way table of dysplasia assessment for 27 patients twice by two pathologists.

Category	Pathologist 2				
Pathologist 1	1	2	3	4	Total
1	9	4	1	6	20
2	0	1	0	0	1
3	0	0	0	0	0
4	1	1	0	4	6
Total	10	6	1	10	27

Table 2: Parameter estimates under symmetry model on dysplasia assessment.

Coefficients	value	standard error	t-value
Intercept	-2.523	11.462	-0.220
u_2	-0.752	0.300	-2.503
u_3	-0.713	0.348	-2.048
u_4	0.130	0.136	0.960
u_{12}	-0.172	0.209	-0.825
u_{13}	-1.747	9.550	-0.183
u_{14}	0.052	1.372	0.038
u_{23}	-1.098	10.181	-0.108
u_{24}	-0.854	6.514	-0.131
u_{34}	0.434	1.274	0.341
Goodness of fit statistics			
G^2	12.281		
X^2	9.572		
Df	6		

Table 3: Parameter estimates under conditional symmetry model on dysplasia assessment.

Coefficients	value	standard. error	t-value
Intercept	-2.455	17.781	-0.138
u_2	-0.489	0.306	-1.597
u_3	-0.625	0.349	-1.793
u_4	0.174	0.136	1.280
u_{12}	-0.251	0.210	-1.198
u_{13}	-1.835	13.826	-0.133
u_{14}	-1.257	1.981	0.048
u_{23}	-0.890	16.766	-0.075
u_{24}	-0.427	9.472	-0.094
u_{34}	0.434	1.976	0.216
τ	-1.705	0.769	-2.218
Goodness of fit statistics			
G^2	5.421		
X^2	6.416		
Df	5		

Table 4: Cross-classification of pathologists A and G on carcinoma in situ of the uterine cervix of 118 slides.

Category Pathologist A	Pathologist G					Total
	1	2	3	4	5	
1	24	2	0	0	0	26
2	7	13	6	0	0	26
3	1	4	32	1	0	38
4	0	1	20	1	0	22
5	0	0	3	1	2	6
Total	32	20	61	3	2	118

Table 5: Parameter estimates under symmetry model of pathologists A and G on carcinoma in situ of the uterine cervix.

Coefficients	value	standard. error	t-value
Intercept	-0.845	4.016	-0.210
u_2	-0.837	0.195	-4.283
u_3	-1.011	0.340	-2.978
u_4	-2.530	8.689	-0.291
u_5	-1.518	7.165	-0.212
u_{12}	0.881	1.639	0.537
u_{13}	0.492	1.171	0.420
u_{14}	0.082	0.887	0.092
u_{15}	-0.836	3.922	-0.213
u_{23}	0.557	0.669	0.832
u_{24}	0.354	0.548	0.646
u_{25}	0.133	0.459	0.290
u_{34}	0.080	0.394	0.206
u_{35}	0.020	0.339	0.059
u_{45}	0.110	0.291	0.378
Goodness of fit statistics			
G^2	32.736		
X^2	26.369		
Df	10		

Table 6: Parameter estimates under symmetry model of pathologists A and G on carcinoma in situ of the uterine cervix.

Coefficients	value	standard. error	t-value
Intercept	-1.645	6.214	-0.265
u_2	-1.306	0.246	-5.309
u_3	-1.168	0.343	-3.402
u_4	-2.826	13.444	-0.210
u_5	-1.696	11.086	-0.153
u_{12}	1.064	2.536	0.419
u_{13}	0.489	1.811	0.270
u_{14}	0.079	1.364	0.058
u_{15}	-0.935	6.068	-0.154
u_{23}	0.659	1.035	0.636
u_{24}	0.352	0.847	0.416
u_{25}	0.131	0.707	0.186
u_{34}	0.152	0.602	0.253
u_{35}	0.014	0.517	0.027
u_{45}	0.167	0.446	0.374
τ	1.414	0.372	3.804
Goodness of fit statistics			
G^2	14.442		
X^2	14.814		
Df	9		

Table 7: Summary of the goodness of fit statistics for the two examples with column for differences between symmetry (SM) and conditional symmetry (CS).

Model		Symmetry model (SM)		Conditional symmetry (CS)			Difference (SM-CS)	
Example	n	G_{SM}^2	X_{SM}^2	G_{CS}^2	X_{CS}^2	df	G_d^2	df_d
1	27	12.281	9.571	5.421	6.416	5	6.860	1
		6						
2	118	32.736	26.369	14.442	14.814	9	18.294	1
		10						

Table 8: Cohen Kappa statistic estimates for the two examples.

		Cohen Kappa			95% CI for k_c	
Example	n	$\hat{k}_c$	standard. Error	z-value	Lower bound	upper bound
1	27	0.242	0.114	2.116	0.018	0.466
2	118	0.467	0.050	9.349	0.369	0.564

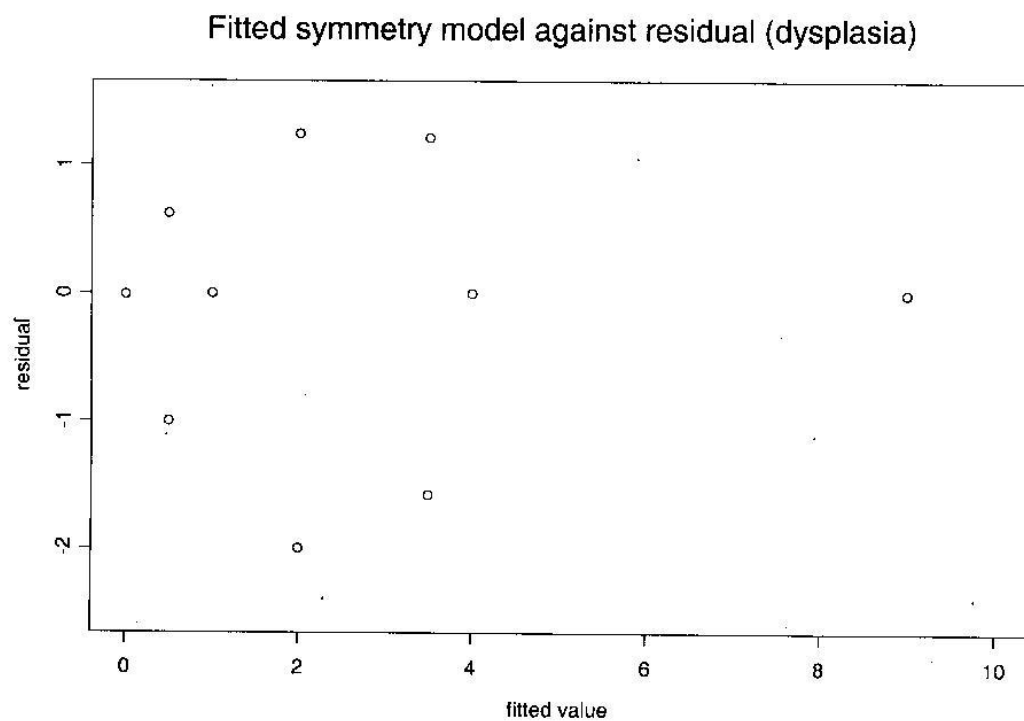


Figure 1: Plot of fitted symmetry model against their respective residuals on dysplasia assessment.

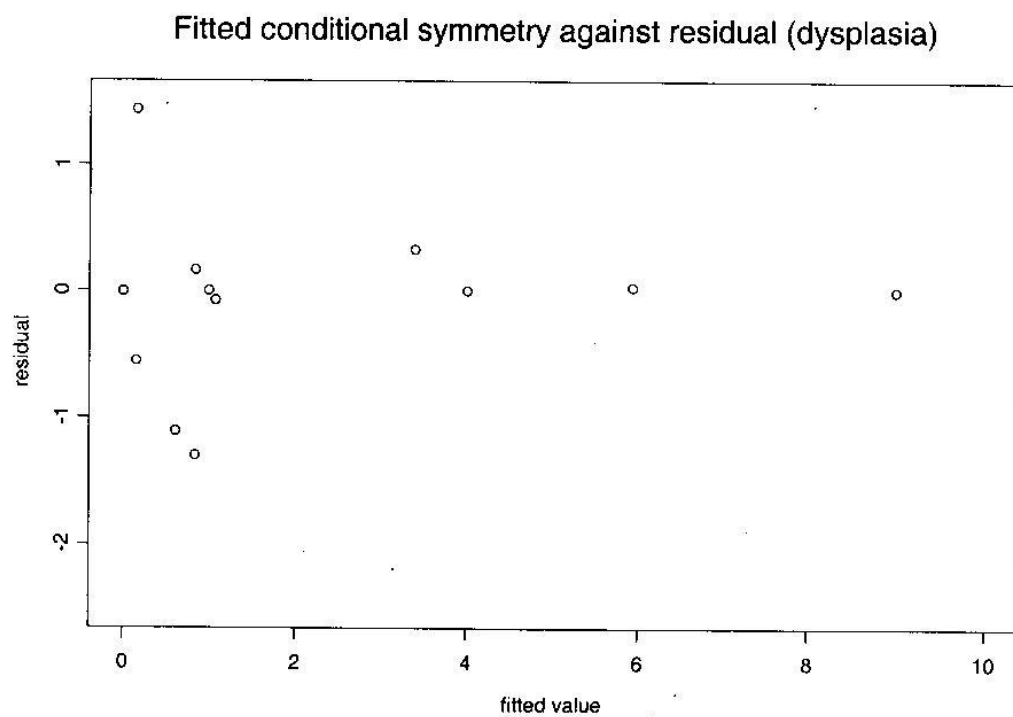


Figure 2: Plot of fitted conditional symmetry model against their respective residuals on dysplasia assessment.

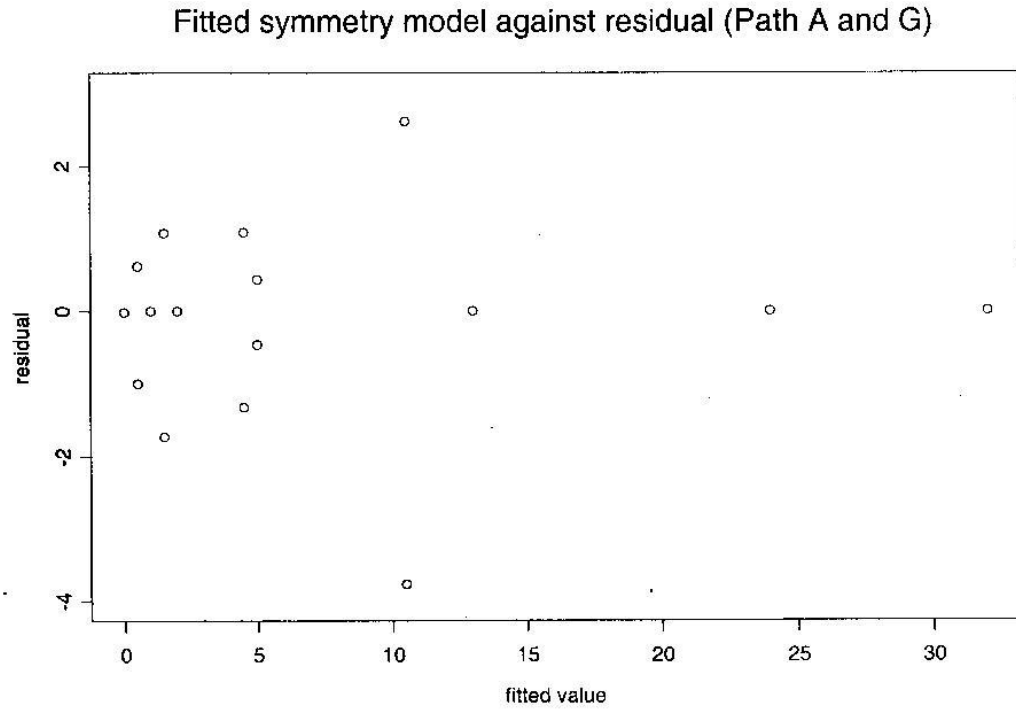


Figure 3: Plot of fitted symmetry model against their respective residuals for Pathologists A and G

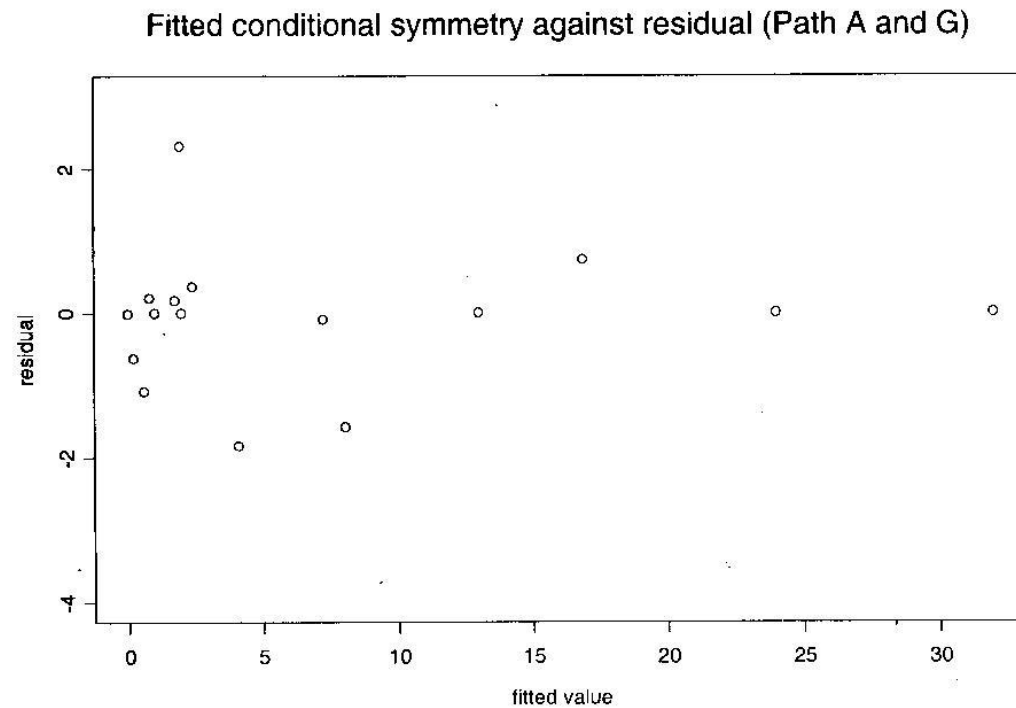


Figure 4: Plot of fitted conditional symmetry model against their respective residuals for Pathologists A and G.