

MATHEMATIK-ARBEITSPAPIERE

**NORMAL GOODNESS-OF-FIT TESTS FOR
PARAMETRIC MULTINOMIAL MODELS
WITH LARGE DEGREES OF FREEDOM**

GERHARD OSIUS & DIETER ROJEK

PREPRINT NR. 36

DECEMBER 1989



MATHEMATIK-ARBEITSPAPIERE

A: MATHEMATISCHE FORSCHUNGSPAPIERE

**NORMAL GOODNESS-OF-FIT TESTS FOR
PARAMETRIC MULTINOMIAL MODELS
WITH LARGE DEGREES OF FREEDOM**

GERHARD OSIUS & DIETER ROJEK

PREPRINT NR. 36

DECEMBER 1989

**FACHBEREICH MATHEMATIK/INFORMATIK
UNIVERSITÄT BREMEN**

**Bibliothekstraße
28334 Bremen
Germany**

NORMAL GOODNESS-OF-FIT TESTS FOR PARAMETRIC MULTINOMIAL MODELS WITH LARGE DEGREES OF FREEDOM

G. OSIUS & D. ROJEK

Universität Bremen

Contents

Summary	1
1. Introduction	2
2. Sampling and Model	5
3. Estimation and Goodness-of-Fit	6
4. Fixed-Cells Asymptotics	6
5. Increasing-Cells Asymptotics	7
6. Binomial Sampling and Quantal-Response Models	13
7. Examples	19
8. Some Practical Aspects of the Test	22
Appendix A: Tables and Figures for the Example 1	24
Appendix B: Tables Example 2	35
References	36

Summary

Normal goodness-of-fit tests are proposed concerning models for independent multinomials with unknown parameters to be estimated. The tests are based on first-order normal approximations of the power-divergence statistics - including Pearson's X^2 and the likelihood ratio G^2 - and apply for large degrees of freedom. No restrictions are imposed on the multinomial sizes (admitting low as well as large expectations in each cell), but simplifications are given in cases where the harmonic resp. arithmetic mean of the sizes are large too. The underlying limit results are presented informally and without proofs, which are given elsewhere. Generalized linear models for binomial data are discussed in more detail, including illustrations of the methods using published data.

AMS 1980 subject classifications. Primary 62F03, 62F05, 62H10, 62H15, 62H17.
Secondary 62E20

Key words and phrases. Binomial data, Deviance, Generalized linear model, Goodness-of-fit, Large degrees of freedom, Likelihood ratio, Multinomial data, Pearson's Chi-square, Power-divergence statistic, Quantal response model, Small expectations, Sparse data.

1. Introduction

We are dealing with the problem of comparing *observed* with *expected* counts for a given model. The commonly used goodness-of-fit statistics are

$$\textbf{Pearson's statistic:} \quad X^2 = \sum_{\text{cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

and the competing

$$\textbf{Likelihood Ratio:} \quad G^2 = 2 \sum_{\text{cells}} \text{observed} \cdot \log\left(\frac{\text{observed}}{\text{expected}}\right) .$$

Cressie & Read(1984) have embedded these as well as other statistics in a family of power-divergence statistics which depend on a real parameter $\lambda \in \mathbb{R}$. Each of these statistics is a sum over all cells of deviations between observed and expected counts:

$$SD_{\lambda} = \sum_{\text{cells}} a_{\lambda}(\text{observed}, \text{expected}). \quad (\text{Sum of Deviations})$$

The deviation for a single cell is measured by some kind of "distance" a_{λ} which basically compares the ratio of observed to expected counts raised to a power λ with the unit 1, and multiplies the difference with the observed count and a constant:

$$a_{\lambda}(\text{observed}, \text{expected}) = \frac{2 \cdot \text{observed}}{\lambda(\lambda+1)} \cdot \left[\left(\frac{\text{observed}}{\text{expected}} \right)^{\lambda} - 1 \right] - \frac{2}{\lambda+1} \cdot [\text{observed} - \text{expected}] \geq 0 .$$

The cases $\lambda=0$, $\lambda=-1$ are defined by continuity as $\lambda \rightarrow 0$, $\lambda \rightarrow -1$. We introduced the second term in the definition of a_{λ} , not given by Cressie & Read(1984), to make it non-negative. The sum over all second terms will be zero provided the sum over all observed counts equals the sum over expected counts, as it is encountered in the situation considered here. Hence SD_{λ} coincides with the definition of Cressie & Read.

The values $\lambda=1$ resp. $\lambda=0$ are of particular interest

$$a_1(\text{observed}, \text{expected}) = \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$a_0(\text{observed}, \text{expected}) = 2 \left[\text{observed} \cdot \log\left(\frac{\text{observed}}{\text{expected}}\right) - [\text{observed} - \text{expected}] \right]$$

and yield Pearson's $X^2=SD_1$ resp. the likelihood-ratio $G^2=SD_0$, which coincides with the *deviance* for the multinomial data discussed here. Cressie & Read (1988) suggest $\lambda=2/3$ as a good compromise between these two. We will restrict ourselves here to values $\lambda>-1$ to allow for zero-observations in the definition of a_λ , which typically occur in sparse data.

Although the family SD_λ of statistics is certainly rich enough for most purposes, the results presented here are not restricted to this particular family since only a few common features of the power divergence statistics are exploited. In principle the subscript λ could be dropped all together thus allowing an even more general type of distance measure $a(-,-)$ in the definition of SD . For definiteness however, we restrict the presentation here to the power divergence family based on a_λ , but will omit the index λ to simplify notations whenever no confusion may arise.

Our aim is to derive the limiting normal distribution of the statistic SD_λ for an the "increasing-cells" asymptotic approach, where the number of cells increases with the total sample size. The expected counts in each cell may be small (sparse data), but need not be. We consider models for product-multinomial sampling with the *same* number K for each multinomial, where the cell probabilities depend on an unknown S -dimensional parameter vector Θ and assume, that the number K and the dimension S remain *fixed* for the asymptotics. Various authors have proved the asymptotic normality of Pearson's X^2 or the likelihood ratio statistic G^2 under particular assumptions for different kind of models. A review of their results is given by Cressie & Read (1988) in Sec. 4.3, 8.1. Since we concentrate here on models with parameters to be estimated, no attempt is made to summarize the work on completely specified models *without* parameters.

Köhler (1986) derives the asymptotic normality of G^2 for log-linear models admitting *closed form* maximum likelihood estimates under increasing-cells asymptotics, where the dimension S increases as well. In a series of papers McCullagh considers X^2 and G^2 under increasing-cells asymptotics in linear exponential family models (1985a, 1986) and generalized linear models for binomial and Poisson sampling (1985b). He argues that one should use the *conditional* distribution of the goodness-of-fit statistics, given the estimated parameter (resp. a sufficient statistic), and derives the conditional limiting distribution. It turns out however, that up to first order approximation, the *unconditional* limit distribution obtained here coincides with the *conditional* distribution given in McCullagh (1985b) under binomial sampling. Dale (1986) derives the asymptotic normality of X^2 and G^2 in the same setup as ours, but imposes the additional assumption that the expectations remain *bounded* for all cells (using different and partly stronger "regularity" conditions). The limit result for G^2 given there however differs from ours in a substantial way and we will comment on that later.

Reviewing finally our previous work, Osius (1985) derived the asymptotic normality of SD (for a general distance measure) under binomial sampling. These results were later extended in Osius (1986) to the case where the underlying model *fails*, and Rojek (1989) generalized and strengthened these arguments to multinomial sampling.

The purpose of this paper is to present goodness-of-fit tests based on the asymptotic normality of the power-divergence family SD_λ and show their performance for a concrete data set. We prefer to present the results rather informally emphasizing more practical aspects (like different approximations according to the distribution of the multinomial sample sizes involved) instead of deriving the limit results, which may be found in Rojek (1989).

2. Sampling and Model

We now turn to the sampling schemes and models involved. The observed data consists of a $J \times K$ contingency table where each row represents a group $j=1, \dots, J$ usually characterized by an additional S -dimensional vector $x_j \in \mathbb{R}^S$ of covariables. The N_j individuals at risk in group j are classified into K categories thus giving the observed cell counts Y_{jk} :

Observed data

Group	S-vector of covariables	observed counts in category			Size (at risk)
		1	...	k	
1	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
j	$x_j = (x_{j1}, \dots, x_{jS})$	Y_{j1}	Y_{jk}	Y_{jK}	N_j
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
J	⋮	⋮	⋮	⋮	⋮
Total sample size					$n = N_+$

We assume *product-multinomial sampling* throughout, i.e. the vector of row counts $Y_j = (Y_{j1}, \dots, Y_{jK})$ are independent for all $j = 1, \dots, J$, each having multinomial distribution $\mathcal{R}(Y_j) = \text{Multi}_K(N_j, \pi_j)$ with a vector $\pi_j = (\pi_{j1}, \dots, \pi_{jK})$ of *positive* cell probabilities.

In the parametric models considered here, each cell probability π_{jk} depends on a common S -dimensional parameter $\theta = (\theta_1, \dots, \theta_S) \in \mathbb{R}^S$ through a *known* function G_{jk} , assumed to be sufficiently "smooth":

$$\pi_{jk} = G_{jk}(\theta) > 0, \quad \text{for all } j \text{ and } k.$$

In the presence of covariables $x_j \in \mathbb{R}^S$, the functions G_{jk} typically depend on the covariables only through the linear combination $x_j^T \theta$, i.e.

$$G_{jk}(\theta) = G_k(x_j^T \theta)$$

with known functions G_k .

The *binomial* case $K=2$ is of particular interest and will be treated in more detail later.

3. Estimation and Goodness-of-Fit

Taking the Maximum-Likelihood estimator $\hat{\theta}$ of θ (or some asymptotically equivalent estimate) we get the expected cell probabilities $\hat{\pi}_{jk} = G_{jk}(\hat{\theta})$ and expected cell counts $\hat{m}_{jk} = N_j \cdot \hat{\pi}_{jk}$, both being positive.

The *power divergence goodness-of-fit statistic* then is the sum over all cells of deviations between observed and expected *counts* Y_{jk} and \hat{m}_{jk} , or equivalently, the weighted deviations of observed and expected *frequencies* \hat{P}_{jk} and $\hat{\pi}_{jk}$, the weight being the size N_j of the group:

$$\begin{aligned} SD_{\lambda}(\hat{\theta}) &= \sum_j \sum_k a_{\lambda}(Y_{jk}, \hat{m}_{jk}) && \text{(Sum of Deviations)} \\ &= \sum_j \sum_k N_j \cdot a_{\lambda}(\hat{P}_{jk}, \hat{\pi}_{jk}) \quad , \end{aligned}$$

with observed cell frequencies $\hat{P}_{jk} = Y_{jk}/N_j$.

Large values of the statistic $SD_{\lambda}(\hat{\theta})$ indicate a *lack of fit*.

4. Fixed-Cells Asymptotics

Since the exact distribution of the power divergence statistics are untractable in the general setting one has to rely on asymptotic results. The general assumptions for the asymptotics discussed here are

- the total sample size *increases* to infinity: $n = N_+ \rightarrow \infty$,
- the dimension S of parameter remains *fixed* ,
- the number K of categories remains *fixed* .

The commonly assumed *classical "fixed cells" asymptotic* approach requires further:

- the number J of groups (and hence of all cells) remains *fixed*,
- *all* group sizes increase (suitably): $N_j \rightarrow \infty$ for all j , and hence *all* expectations increase: $m_{jk} \rightarrow \infty$.

The asymptotic results for "fixed cells" are given in Cressie & Read and may be summarized as follows. Under the null hypothesis that the given models holds, the statistics of the family SD_{λ} are asymptotically equivalent for *all* $\lambda \in \mathbb{R}$, each having the same (central) χ^2 -distribution with the appropriate degrees of freedom $J(K-1)-S$. The asymptotic equivalence for the family still holds under *local* alternatives to the null hypothesis, the limit distribution now being a noncentral χ^2 .

5. Increasing - Cells Asymptotics

The classical *fixed-cells* asymptotics are not always appropriate, in particular not for sparse data. In bigger data sets one often has a large number of groups with a considerable amount of small group sizes (sometimes as low as 1). To deal with these situations we consider here an *increasing-cells asymptotics*, for which (in addition to the general assumptions above) the number J of groups (and hence the total number $J \times K$ of cells) *increases* : $J \rightarrow \infty$.

The basic asymptotic result for increasing cells is that the statistic SD_λ has an asymptotic *normal* distribution (under the nullhypothesis and arbitrary alternatives). This is not surprising since SD_λ is an increasing sum of components, which are "nearly" independent, except for their dependence through the common estimate $\hat{\theta}$ (having a fixed dimension). However the power-divergence statistics need no longer be asymptotically equivalent for increasing cells, and in fact explicit models may be given where X^2 and G^2 are *not* equivalent.

Here we will only be concerned with the asymptotic distribution of SD_λ under the *nullhypothesis*, i.e. the model *holds*, although similar results are available for *arbitrary alternatives*, i.e. the given model *fails*, see Rojek(1989). Under assumptions given below, we can show that SD_λ has an asymptotic normal distribution with expectation μ_λ and variance σ_λ^2 , i.e. the normalized sum SD_λ converges in distribution to the standard normal:

Asymptotic Normality of SD_λ under the model (and assumptions given below):

$$T_\lambda = \frac{SD_\lambda - \mu_\lambda}{\sigma_\lambda} \xrightarrow{D} N(0,1) \quad \text{resp.} \quad SD_\lambda \underset{\text{as.}}{\sim} N(\mu_\lambda, \sigma_\lambda^2),$$

provided the denominator is not 0 (see the variance condition later).

The resulting one-sided *normal test* of level α rejects the model if the *test statistic* T_λ exceeds the upper α -quantile z_α of the standard normal distribution. In extreme situations to be discussed later (under "individual groups") a *two-sided* test may be more appropriate.

Collecting the essential assumptions used for deriving the limit result, we first list somewhat informally - for details see Rojek (1989) - what can be called the "standard" assumptions:

- The cell probabilities $\pi_{jk} = G_{jk}(\theta)$ are sufficiently smooth functions of θ (in terms of differentiability) and the matrix of partial derivatives has full rank S .
- The information increases, e.g. the scaled information matrix $\frac{1}{n} I(\theta)$ converges to a positive-definite limit.
- The cell probabilities $\pi_{jk} = G_{jk}(\theta)$ evaluated at the true value of θ satisfy certain bounding conditions, e.g. are all bounded away from 0 and 1.
- The estimator $\hat{\theta}$ is consistent and has an asymptotic normal distribution (which in turn may be derived from further assumptions).

Of course this set of assumptions may be weakened to some extent, usually giving rather technical conditions which are often difficult to interpret.

The additional specific assumptions for the increasing-cells asymptotics involve some characteristics of the group sizes N_1, \dots, N_J , namely their

- harmonic mean: $HM = \left[\frac{1}{J} (N_1^{-1} + \dots + N_J^{-1}) \right]^{-1}$,
- arithmetic mean: $AM = \frac{1}{J} (N_1 + \dots + N_J)$,
- quadratic mean: $QM = \left[\frac{1}{J} (N_1^2 + \dots + N_J^2) \right]^{1/2}$

which satisfy the relation: $1 \leq HM \leq AM \leq QM$.

The asymptotic expectation μ_λ and variance σ_λ^2 of the sum SD_λ of deviations simplify considerably if the above means tend suitably to infinity. Although this is not necessary for the general increasing-cell asymptotics, it will be instructive to discuss some important special cases before turning to the general case.

Special Case 1: "Fast increasing Harmonic Mean" $HM / \sqrt{J} \xrightarrow{J \rightarrow \infty} \infty$

Here μ_λ and σ_λ^2 are given independently of the parameter λ by

$$\begin{aligned} \mu_\lambda &= J(K-1) = DF + S \quad \text{with degrees of freedom } DF = J(K-1) - S, \\ \sigma_\lambda^2 &= 2J(K-1) = 2\mu_\lambda \quad . \end{aligned}$$

The goodness-of-fit test statistic is thus

$$T_\lambda^{(1)} = \frac{SD_\lambda(\hat{\theta}) - J(K-1)}{\sqrt{2J(K-1)}}$$

For large degrees of freedom the expectation is approximately the degree of freedom, $\mu_\lambda \approx DF$. Hence $T_\lambda^{(1)}$ may be viewed as a crude normal approximation to the limiting χ^2 -distribution of $SD_\lambda(\hat{\theta})$ under the classical fixed-cells asymptotics.

Special Case 2: "Increasing Harmonic Mean" $HM \xrightarrow{J \rightarrow \infty} \infty$

Passing from the above to this weaker case, the variance remains unchanged, but the expectation has to be "corrected" to the actual expectation of $SD_\lambda(\theta)$, evaluated at the estimate $\hat{\theta}$:

$$\begin{aligned} \mu_\lambda &= \mu_\lambda(\hat{\theta}) \quad \text{with } \mu_\lambda(\theta) := E_\theta\{SD_\lambda(\theta)\}, \\ \sigma_\lambda^2 &= 2J(K-1) \quad \text{(as in Case 1)} . \end{aligned}$$

This gives the following goodness-of-fit test statistic

$$T_\lambda^{(2)} = \frac{SD_\lambda(\hat{\theta}) - \mu_\lambda(\hat{\theta})}{\sqrt{2J(K-1)}}$$

However for Pearson's statistic, i.e. $\lambda = 1$, one gets the same expectation as in case 1:

$$\mu_1(\theta) = J(K-1) .$$

For general λ no simple computational formula for $\mu_\lambda(\theta)$ exists.

Special Case 3: "Increasing Arithmetic Mean" $AM = \bar{N} \xrightarrow{J \rightarrow \infty} \infty$

If we pass from the above to this weaker case, now the expectation remains the same, but the variance has to be "corrected" to the actual variance of $SD_\lambda(\theta)$, evaluated at the estimate $\hat{\theta}$:

$$\begin{aligned} \mu_\lambda &= \mu_\lambda(\hat{\theta}) && \text{as in Case 2,} \\ \sigma_\lambda^2 &= v_\lambda^2(\hat{\theta}) && \text{with } v_\lambda^2(\theta) := \text{Var}_\theta\{SD_\lambda(\theta)\} . \end{aligned}$$

This yields the test statistic

$$T_\lambda^{(3)} = \frac{SD_\lambda(\hat{\theta}) - \mu_\lambda(\hat{\theta})}{\sqrt{v_\lambda^2(\hat{\theta})}} .$$

For Pearson's statistic, i.e. $\lambda = 1$, the variance simplifies to:

$$v_1^2(\theta) = 2J(K-1) + \sum_j \frac{1}{N_j} \left[\sum_k \frac{1}{\pi_{jk}(\theta)} - K^2 - 2(K-1) \right] .$$

In comparison to case 2, this variance contains in the sum corrections for small group-sizes as well as for small cell probabilities. Unfortunately, for general λ no simple expression for the variance like above is available.

Finally, we look at the *general case*, where the (arithmetic) mean of the group sizes need not to be large. The expectation remains as in the last cases 2-3, but the variance is reduced (as compared to case 3) by a quadratic form $Q_\lambda(\hat{\theta})$ which represents a correction for using the estimate $\hat{\theta}$ instead of the true parameter θ :

$$\begin{aligned} \mu_\lambda &= \mu_\lambda(\hat{\theta}) && \text{as in Case 2-3} \\ \sigma_\lambda^2 &= \sigma_\lambda^2(\hat{\theta}) = v_\lambda^2(\hat{\theta}) - Q_\lambda(\hat{\theta}) && \text{with } v_\lambda^2 \text{ as in Case 3.} \end{aligned}$$

The quadratic form

$$Q_\lambda(\theta) := c_\lambda^T(\theta) \cdot I^{-1}(\theta) \cdot c_\lambda(\theta)$$

involves the S -dimensional covariance vector of the sum $SD_\lambda(\theta)$ and the score vector, i.e. the derivative $D\ell(\theta)$ of the log-likelihood $\ell(\theta)$,

$$c_\lambda(\theta) = \text{Cov}_\theta\{SD_\lambda(\theta), D\ell(\theta)\} ,$$

as well as the inverse of the $S \times S$ information matrix

$$I(\theta) = \text{Cov}_\theta\{D\ell(\theta)\} .$$

The final goodness-of-fit statistic as a normalized power-divergence statistic looks in the *general case* like this:

$$T_\lambda = \frac{SD_\lambda(\hat{\theta}) - \mu_\lambda(\hat{\theta})}{\sqrt{v_\lambda^2(\hat{\theta}) - Q_\lambda(\hat{\theta})}} \quad (\text{normalized test statistic}).$$

Of course, T_λ is asymptotically equivalent to $T_\lambda^{(i)}$ in the special cases $i=1,2,3$ above.

For Pearson's statistic, i.e. $\lambda = 1$, the covariance vector is given by

$$c_1(\theta) = \left(\sum_j \sum_k \frac{1}{\pi_{jk}(\theta)} \cdot \frac{\partial}{\partial \theta_s} \pi_{jk}(\theta) \right)_{s=1, \dots, S},$$

and involves only the cell probabilities and their derivatives, but not the group sizes N_j .

Unfortunately, for general λ no simple expression is available for the quantity $c_\lambda(\theta)$, as well as for $\mu_\lambda(\theta)$ and $\sigma_\lambda^2(\theta)$ above, except for integer values $\lambda=1,2,3,\dots$ which are (besides $\lambda=1$) not of primary interest. Nevertheless these quantities may in principle be computed as a sum over all possible outcomes of each multinomial group. This causes no problem in the binomial case ($K=2$), but the computational effort increases heavily with the number K of classes. Except for small values of K or in other special cases (like above) only Pearson's statistic ($\lambda=1$) can be evaluated exactly for a general pattern of group sizes. The use of Pearson's statistic has - besides its easy computation and interpretation - a further advantage, that the expectation $\mu_\lambda(\theta)$ is *independent* of the parameter θ , and the same holds in case 2 above for the variance.

We now discuss the additional assumptions used to derive the asymptotic normality of the test statistic T_λ in the general case. First, we need that the variance $\sigma_\lambda^2(\theta)$ increases fast enough, i.e. a

$$\text{Variance-Condition: } J / \sigma_\lambda^2(\theta) \quad \text{is bounded (for "true" } \theta \text{)}.$$

This always holds for *increasing* harmonic means (Case 2 above) and - at least in the binomial case discussed later - under weaker conditions for the group sizes.

Secondly, we still need (at least up to now) a condition on the group sizes, namely that the arithmetic mean is increasing (case 3 above) *or* that the quadratic mean is of lower order than the square root of the total sample size $N_+ := \sum N_j$:

$$\text{Size-Condition: } AM \rightarrow \infty \quad \text{or} \quad QM / \sqrt{N_+} \rightarrow 0$$

Let us now look at other special cases concerning smaller group sizes encountered in sparse data.

Special Case 4: Bounded arithmetic mean

Suppose (in contrast to case 3 above) that the arithmetic mean $AM = \bar{N}$ is *bounded*. The size-condition then equivalently states that the empirical *variance* of the group sizes is of lower order than the total sample size N_+ :

$$\frac{1}{J} \sum_j (N_j - \bar{N})^2 / N_+ \longrightarrow 0$$

i.e. the variance of the group sizes does not increase too fast. This trivially holds for *balanced* group sizes: $N_j = \bar{N}$ for all j . The variance condition has to be checked and may fail under certain circumstances (see case 5 below).

Special Case 5: Individual groups (No grouping)

We now pass from the case above to the smallest possible groups, which arise when each individual forms a group of its own, i.e. $N_j = 1$ for all j , and hence $J = N_+$. This can always be achieved by splitting each *original* group of size $N_j > 1$ into N_j *new individual* groups. The size condition is always satisfied (see case 4 above), but the variance condition *may fail* for some model-dependent values of λ (see the binomial case later).

For individual groups, the observed counts are either 0 or 1, and hence the test statistic uses the distance function a_λ only through the partial functions $a_\lambda(0, -)$ and $a_\lambda(1, -)$. This slightly disturbs the conventional interpretation of the sum SD_λ , so that *small* values of SD_λ may also indicate a lack of fit, suggesting a *two-sided* test. More details - for the binomial sampling - are given later.

Dale (1986) derives the asymptotic normality of X^2 (Thm.3) and G^2 (Thm.4) for *bounded* group sizes N_j , using different regularity conditions. For Pearson's X^2 the asymptotic expectation and variance in Thm.3 agree with μ_λ and σ_λ^2 given above for the general case. However Thm.4 on the likelihood ratio G^2 gives μ_λ and σ_λ^2 as in special case 3 above which does not apply for bounded sizes, as will be shown later using a constant model.

6. Binomial Sampling and Quantal-Response Models

In the presence of only $K=2$ categories, usually termed "response" and "non-response", the above model reduces to a general type of *quantal response model*. The observed data is usually represented as a $J \times 2$ table in the following way:

Group	S-vector of covariables	observed counts in category		Size (at risk)
		Response	No Response	
1	\cdot	\cdot	\cdot	\cdot
\vdots	\vdots	\vdots	\vdots	\vdots
j	$x_j = (x_{j1}, \dots, x_{jS})$	Y_j	$N_j - Y_j$	N_j
\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots
J	\cdot	\cdot	\cdot	\cdot
Total sample size				$n = N_+$

We then have *product-binomial sampling* with independent counts Y_1, \dots, Y_J for the responses, each having a binomial distribution $\mathcal{B}(Y_j) = B(N_j, \pi_j)$ with response probability π_j .

The parametric models given above include the generalized linear models considered in McCullagh & Nelder (1989), where the response probability depends on the linear combination $x_j^T \theta$ through a fixed distribution function G (e.g. the standard normal or logistic)

$$\pi_j = G(x_j^T \theta) \quad \text{resp.} \quad \eta_j := g(\pi_j) = x_j^T \theta \quad \text{for all } j=1, \dots, J,$$

with link function $g=G^{-1}$ (e.g. the probit or logit transformation).

The power-divergence statistic reduces to the following sum over all binomial groups

$$SD_\lambda(\hat{\theta}) = \sum_j N_j \cdot A_\lambda(\hat{P}_j, \hat{\pi}_j),$$

where $\hat{P}_j = Y_j/N_j$ denotes the observed rate of response in group j , and

$$A_\lambda(p, \pi) := a_\lambda(p, \pi) + a_\lambda(1-p, 1-\pi)$$

serves as a new "distance measure" which is symmetric about $1/2$ in both arguments.

For computing μ_λ and σ_λ^2 we introduce for any binomial(N, π) variable Y the notations

$$\begin{aligned} e_\lambda(N, \pi) &:= E\{A_\lambda(Y, N\pi)\} & , & & v_\lambda^2(N, \pi) &:= \text{Var}\{A_\lambda(Y, N\pi)\} , \\ c_\lambda(N, \pi) &:= \text{Cov}\{Y, A_\lambda(Y, N\pi)\} . \end{aligned}$$

For Pearson's case $\lambda=1$ these expressions reduce to

$$\begin{aligned} e_\lambda(N, \pi) &:= 1 & , & & v_\lambda^2(N, \pi) &:= 2 + (\pi^{-1}[1-\pi]^{-1} - 6) / N , \\ c_\lambda(N, \pi) &:= 1-2\pi . \end{aligned}$$

But for general λ the computation of e_λ , v_λ and c_λ requires a summation over all possible outcomes of the underlying binomial distribution. With this notation we get

$$\begin{aligned} \mu_\lambda(\theta) &= \sum_j e_\lambda(N_j, \pi_j) & , & & v_\lambda^2(\theta) &= \sum_j v_\lambda^2(N_j, \pi_j) \\ c_\lambda(\theta) &= X^T \cdot \text{Diag}\{G'(\eta_j)/(\pi_j[1-\pi_j])\} \cdot c_\lambda(N, \pi) , \end{aligned}$$

where $c_\lambda(N, \pi)$ denotes the J -vector with components $c_\lambda(N_j, \pi_j)$.

Let us now give a decomposition of the asymptotic variance σ_λ^2 which is of theoretical and of computational interest. We perform an ordinary weighted linear regression using the same model matrix X , but the J -vector of "observations"

$$u_\lambda(\theta) := \text{Diag}^{-1}\{N_j G'(\eta_j)\} \cdot c_\lambda(N, \pi)$$

and the $J \times J$ -diagonal matrix of weights

$$W(\theta) = \text{Diag}\{N_j G'(\eta_j)^2 / (\pi_j[1-\pi_j])\} .$$

The residual sum of squares for this regression may be written as

$$\text{RSS}_\lambda(\theta) = u_\lambda^T(\theta) \left[W(\theta) - W(\theta) X I^{-1}(\theta) X^T W(\theta) \right] u_\lambda(\theta) ,$$

where $I(\theta) = X^T W(\theta) X$ is the information matrix of the generalized linear model. This leads to the decomposition

$$\begin{aligned} \sigma_\lambda^2(\theta) &= V_\lambda(N, \pi) + \text{RSS}_\lambda(\theta) & & \text{with} \\ V_\lambda(N, \pi) &:= \sum_j \left[v_\lambda^2(N_j, \pi_j) - c_\lambda^2(N_j, \pi_j) / (N_j \pi_j [1-\pi_j]) \right] \geq 0 , \end{aligned}$$

which can be used for computing $s_\lambda^2(\hat{\theta})$. Again, for Pearson's statistic a simple formula is available (HM is the harmonic mean of the group sizes):

$$V_1(N, \pi) = 2 \left[J - \sum_j N_j^{-1} \right] = 2J \left[1 - \text{HM}^{-1} \right] .$$

Each term of the sum V_λ is non-negative, and zero if and only if $N_j=1$. Hence we get $V_\lambda > 0$ except in the case of individual grouping. More generally, $J/V_\lambda(N, \pi)$ may be shown to be bounded (and consequently the variance condition holds), provided the harmonic mean of the group sizes is bounded away from 1, i.e. the proportion of non-individual group sizes $N_j > 1$ stays away from 1. If this is not the case, we are basically left with individual grouping which we now take up again.

Special Case 5: Individual groups under binomial sampling

The numerator of the statistic T_λ may now be written as a weighted sum

$$SD_\lambda(\hat{\theta}) - \mu(\hat{\theta}) = \sum_j w_\lambda(\hat{\pi}_j) [Y_j - \hat{\pi}_j]$$

using the weight function

$$w_\lambda(\pi) := A_\lambda(1, \pi) - A_\lambda(0, \pi),$$

which reduces for Pearson's resp. the likelihood ratio statistic to

$$w_1(\pi) = (1-2\pi)/(\pi[1-\pi]) \quad \text{resp.} \quad w_0(\pi) = -2 \logit(\pi).$$

Looking at the denominator of T_λ we have $s_\lambda^2(\theta) = \text{RSS}_\lambda(\theta)$ since $V_\lambda(N, \pi) = 0$. Clearly $s_\lambda^2(\theta) = 0$ if and only if the vector $u_\lambda(\theta)$ lies in the linear subspace $\mathfrak{M} \subset \mathbb{R}^J$ generated by the columns of X . We provide two examples where this happens.

Example 1 (constant model for individual groups): The model with constant cell probabilities $\pi_j = \pi$ for all j is a one-dimensional linear logistic model with the $S \times 1$ model matrix $X = (1, \dots, 1)^T$. Here $\eta = G^{-1}(\pi) \in \mathfrak{M}$ always implies $u_\lambda(\theta) \in \mathfrak{M}$. \square

Example 2 (logistic model): In the linear logistic model the components of $u_\lambda(\theta)$ are $c_\lambda(1, \pi_j)/(\pi_j[1-\pi_j])$ since $N_j = 1$. For the likelihood ratio statistic we get $u_0(\theta) = -2\eta = -2 \logit(\pi) \in \mathfrak{M}$, and hence G^2 has no diagnostic power in this case, as already noted by McCullagh (1985a). Furthermore, the ML-estimate satisfies $SD_0(\hat{\theta}) = \mu_0(\hat{\theta})$ which produces an undefined statistic $T_0 = 0/0$. \square

The normal test based on the asymptotic normal distribution of T_λ with individual grouping may be viewed as a score test, as pointed out by K. Drescher (1984, private communication). For this purpose we consider the enlarged (usually non-linear) model with an additional parameter $\phi \in \mathbb{R}$ given by

$$\pi_j = H_\lambda(x_j^T \theta | \phi) := G(x_j^T \theta + \phi \cdot h_\lambda(x_j^T \theta)),$$

where the function h_λ is defined as

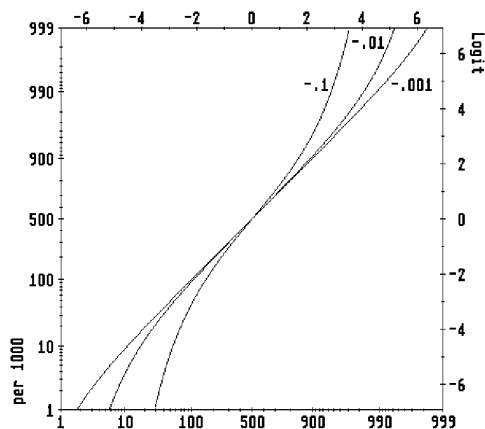
$$h_\lambda(y) = w_\lambda(G(y)) \cdot G(y) \cdot [1-G(y)] / G'(y) \quad \text{for } y \in \mathbb{R}.$$

For the logistic model, this function is given by $h_\lambda(\logit(\pi)) = w_\lambda(\pi)$ for any $0 < \pi < 1$. Plots of the function H_λ for some values of λ and ϕ are given in Fig.6.1 (logit model) and Fig.6.2 (probit model).

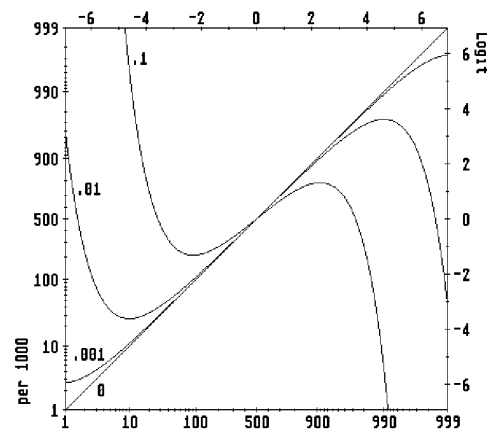
Testing the nullhypothesis $H_0: \psi=0$ (i.e. the original model holds) using the score test produces exactly T_λ as a test statistic. From this and the above representation of $SD_\lambda - \mu_\lambda$ we suggest the *two-sided* normal test, which also rejects the model for large negative values of SD_λ , to be more appropriate for individual grouping. Furthermore the test based on T_λ no longer appears as a universal "goodness-of-fit test" against *all* possible alternatives but rather as a powerful test against particular alternatives concerning the link between the "linear predictor" $x^T\theta$ and the response probability π . In the two examples above, the enlarged model above actually coincides with the original model, which again explains the breakdown of the test.

Figure 6.1: The function H_λ in the logit model for $\lambda=1$ (Pearson), $\lambda=2/3$ and different values of ψ . The plots show $H_\lambda(\text{logit}(\pi)|\psi)$ as a function of the response probability π (per 1000) on logit scales (lower and left axes) resp. $\text{logit}(H_\lambda(y|\psi))$ as a function of $y=\text{logit}(\pi)$ (upper and right axes). The original model ($\psi=0$) yields a straight line, while departures from this model correspond to curvature. The case $\lambda=0$ (Likelihood ratio) is not included, since it gives straight lines for any ψ .

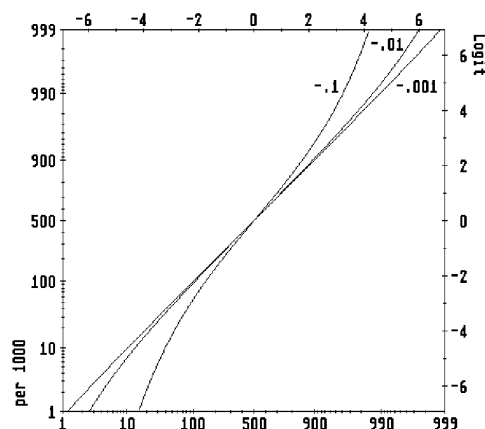
a) $\lambda=1: \psi = -0.1, -0.01, -0.001$



b) $\lambda=1: \psi = 0, 0.001, 0.01, 0.1$



c) $\lambda=2/3: \psi = -0.1, -0.01, -0.001$



d) $\lambda=2/3: \psi = 0.001, 0.01, 0.1$

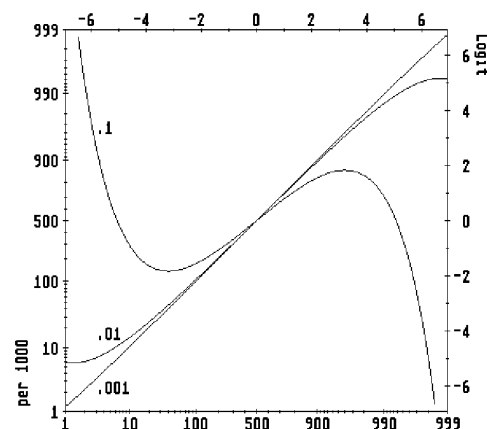
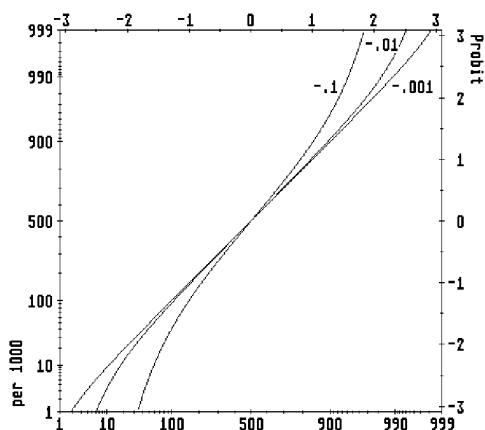
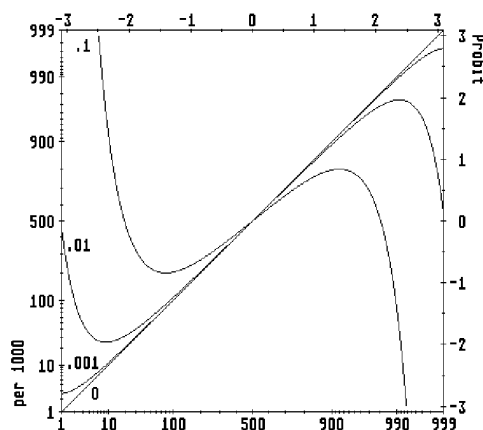


Figure 6.2: The function H_λ in the probit model for $\lambda=1$ (Pearson), $\lambda=2/3$, $\lambda=0$ (Likelihood ratio) and different values of ψ . The plots show $H_\lambda(\text{probit}(\pi)|\psi)$ as a function of the response probability π (per 1000) on probit scales (lower and left axes) resp. $\text{probit}(H_\lambda(y|\psi))$ as a function of $y=\text{probit}(\pi)$ (upper and right axes). The original model ($\psi=0$) yields a straight line, while departures from this model correspond to curvature. Note the somewhat different type of curves for $\lambda=0$.

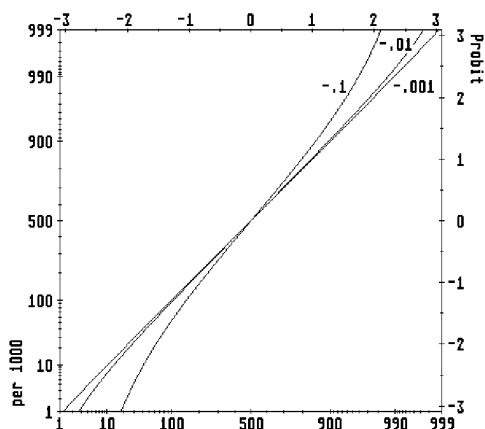
a) $\lambda = 1$: $\psi = -0.1, -0.01, -0.001$



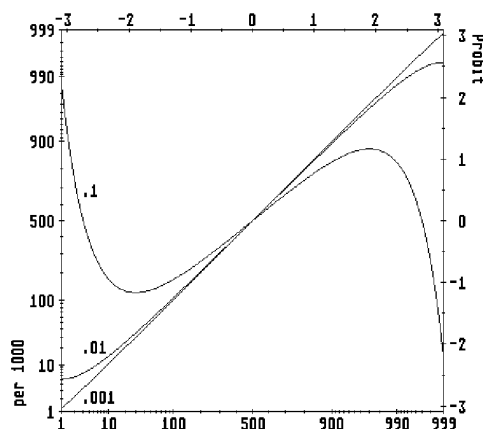
b) $\lambda = 1$: $\psi = 0, 0.001, 0.01, 0.1$



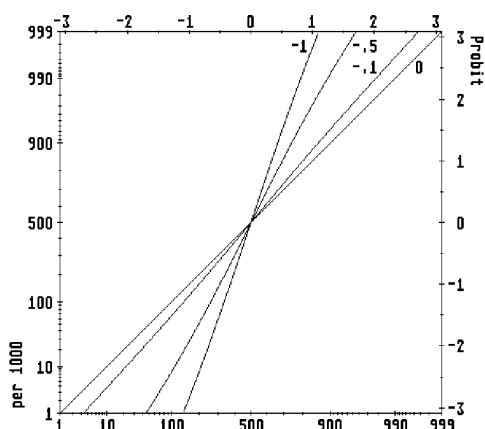
c) $\lambda = 2/3$: $\psi = -0.1, -0.01, -0.001$



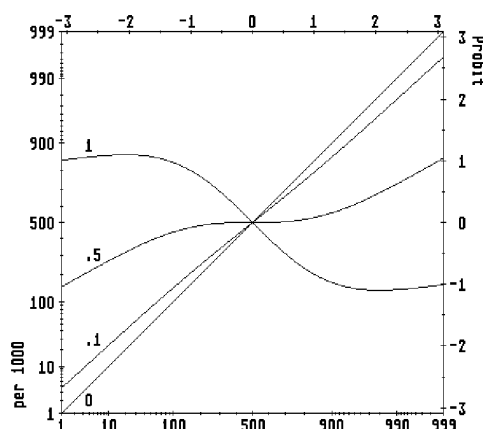
d) $\lambda = 2/3$: $\psi = 0.001, 0.01, 0.1$



e) $\lambda = 0$: $\psi = -1, -0.5, -0.1$



f) $\lambda = 0$: $\psi = 0, 0.1, 0.5, 1$



Returning back to the general discussion, McCullagh (1985b) argues that one should use the *conditional* distribution of goodness-of-fit statistics, given the estimated parameter. For the generalized linear model however he shows that Pearson's statistic X^2 is asymptotically independent of the ML-estimate $\hat{\theta}$, at least to first order (in J). Hence the conditional asymptotic expectation and variance of X^2 agree (to order J) with the corresponding unconditional cumulants $\mu_1(\hat{\theta})$ and $\sigma_1^2(\hat{\theta})$ given above (both being of order J). Somewhat surprisingly, the conditional asymptotic expectation and variance of the likelihood ratio statistic G^2 given by McCullagh(1985b) for the linear logistic model may be written as

$$E\{X^2 | \hat{\theta} = \theta\} = (1 - S/J) \mu_0(\hat{\theta}) + O(1), \quad \text{Var}\{X^2 | \hat{\theta} = \theta\} = (1 - S/J) \sigma_0^2(\hat{\theta}) + O(1),$$

and hence coincide to first order with the unconditional $\mu_0(\hat{\theta})$ and $\sigma_0^2(\hat{\theta})$ derived here. More generally, Osius(1986) has shown that the conditional asymptotic distribution of SD_λ (for *any* λ) given $\hat{\theta} = \theta$ agrees to first order with the unconditional distribution. Apart from this coincidence to first order, McCullagh improves the first order approximation by computing higher conditional cumulants to use Edgeworth and Cornish-Fisher expansion to determine P-values $P\{T_\lambda \geq z\}$ resp. quantiles for the standardized goodness-of-fit statistics T_0 and T_1 .

Finally, we give an example showing that the simplified statistics given in the special cases 1-3 are not appropriate in the general case, and furthermore that X^2 and G^2 are not asymptotically equivalent for increasing-cells with bounded group sizes.

Example 3 (constant model for balanced sizes):

We take up again the constant model of example 1 with balanced group sizes $N_j = N$ for all j and probability $\pi \neq 1/2$. The expected probability $\hat{\pi}_j = Y_+ / N_+$ is the overall observed rate of response, and the variance reduces to

$$s_\lambda^2(\theta) = J \left[v_\lambda^2(N, \pi) - c_\lambda^2(N, \pi) / (N\pi[1-\pi]) \right].$$

The simplified statistic $T_\lambda^{(3)}$ from case 3 above (which is *not* appropriate here) still has a limiting normal distribution with mean 0 and variance

$$1 - v_\lambda^{-2}(N, \pi) \cdot c_\lambda^2(N, \pi) / (N\pi[1-\pi])$$

which is < 1 , at least in the interesting cases $\lambda = 0$ and $\lambda = 1$, so that $T_\lambda^{(3)}$ and T_λ are *not* asymptotically equivalent here. In the special case $N = 2$, the observed counts Y_j may only take the values 0,1,2 and therefore one can establish the following relation between the standardizations of X^2 and G^2 :

$$T_1 = 2 \hat{\pi} [1 - \hat{\pi}] \log(4) \cdot T_0 \quad (\text{for } N=2),$$

and hence X^2 and G^2 are not asymptotically equivalent in this situation.

7. Examples

We illustrate some features of the goodness-of-fit tests described above using two sets of published data. The tables and figures are given in the corresponding appendices.

Example 1: Study on Infant Mortality

The data are taken from an analysis of infant mortality given by Karn & Penrose (1951–52). The large body of data assembled from records of U.C.H. Obstetric Hospital for the years 1935–46 contains information about 13 730 infants (7037 male, 6693 female, no twins) and their mothers. We apply a linear logistic model to parts of the data (taken from Table 1 in Karn & Penrose), regarding non-survival at 28 days (including stillbirth) as a response, which is to be related to the following variables:

- birth weight W , recorded in 25 classes with range: 1.0 (0.5) 13.5 pound (lb.),
- gestation time G , recorded in 41 classes with range: 155 (5) 355 days,
- sex S of infant, recorded as a factor: 1=male, 2=female .

For the grouped variables birth weight and gestation time we have chosen the *lower* limit of its corresponding class to represent its actual value. This is just a matter of scale and does not affect the goodness of fit we are interested in.

Karn & Penrose originally fitted the models separately for the males and females, using the linear logit model $1 + W + W^2 + G + G^2 + W \cdot G$ and some submodels thereof (for the notation of models see McCullagh & Nelder).

We first fit a joint model $1 + S + W + W^2 + G + G^2$ for both sexes (Table A.1). The power-divergence statistics SD_λ given in Table A.2 differ dramatically for the three values $\lambda=1$ (Pearson), $\lambda=0$ (Likelihood ratio), $\lambda=2/3$ (Cressie-Read), and so do the conclusions of the goodness-of-fit tests based on the classical χ^2 -approximation. This approximation however is not justified, due to a large number of individual groups (i.e. $N_j=1$, see Table A.3, Fig. A.1) and consequently very low expected values. Looking at the normalized statistics T_λ in Tables A.5 the situation changes completely, and the fit appears quite satisfactory for all 3 values of λ , taking the large sample size into account. Passing from the χ^2 -approximation, which corresponds roughly to $T_\lambda^{(1)}$, over $T_\lambda^{(2)}$ and $T_\lambda^{(3)}$ to the general T_λ in Table A.5, we find the most dramatic changes between $T_\lambda^{(1)}$ and $T_\lambda^{(3)}$, but the final step to T_λ provides only a slight correction. This could have been expected from the theory since the arithmetic mean of the group sizes is not small (see Table A.3).

A different aspect of judging the fit is provided by residual plots. We use the *scaled deviance residual* [McCullagh & Nelder, Sec. 2.3–4] which may be written as

$$r(\hat{P}, \hat{\pi}) := \text{sign}(\hat{P} - \hat{\pi}) \cdot \left[A_0(\hat{P}, \hat{\pi}) / \hat{\Phi} \right]^{1/2},$$

with a scale factor $\hat{\Phi}$ is chosen such that the sum of squared residuals equals the degrees

of freedom: $\sum r^2(\hat{P}_j, \hat{\pi}_j) = DF$. The plots of the residual r against the linear predictor $\hat{\eta} = \text{logit}(\hat{\pi})$ and the covariables (Fig. A.2) confirm a satisfactory fit of the model. The curved lines appearing in the plot of the residual r against $\hat{\eta}$ are due to the large amount of *small* group sizes. For example, in all individual groups (i.e. $N_j=1$) the observed rate \hat{P} is either 0 or 1 and hence all these residuals must lie on one of the two curves $r(0, G(\eta))$ and $r(1, G(\eta))$ viewed as functions in η (see Fig. A.3). More generally the residuals of all groups of a given size n will appear on the corresponding $n+1$ different curves $r(k/n, G(\eta))$ for $0 \leq k \leq n$.

It is interesting to compare the performance of the normal test with other tests against specific alternative models. One such test is the normal test based on individual grouping (given in Table A.6), which does not reject the model here on the 5%-level for Pearson's X^2 and $\lambda=2/3$ (G^2 is not applicable).

Another possibility is to use the deviance test with respect to a specified enlarged model. Passing to the model $S*(1+W+W^2+W^3+G+G^2+G^3+W \cdot G)$ with several additional interactions the decrease in deviance is not significant (Table A.7), and hence the basic model above not rejected.

On the other hand, we can eliminate significant variables from our model and see whether the normal test rejects the simpler submodel. Deleting birth weight, we get the model $1+S+G+G^2$, which is clearly rejected by all three normal tests based on the original grouping (which still uses birth weight), and this conforms with the deviance test (Table A.8). Collapsing the groups over birth weight, the picture changes and the normal test for Pearson's X^2 (but not the others) now accepts the model (Table A.9). Refining the groups on the other hand to individual grouping produces similar results but smaller changes with respect to λ (Table A.10).

The second submodel $1+S+W+W^2$ with gestation time excluded is also clearly rejected by the normal tests based on the original groups, in accordance with the deviance test (Table A.11). However for collapsed groups (Table A.12) or individual grouping (Table A.13) the normal tests do not reject the submodel. This is not surprising, since the grouping clearly influences the teststatistic and its power.

For the third submodel $1+W+W^2+G+G^2$ without sex, the normal tests for $\lambda=0, 2/3$ have roughly the same P-level as the deviance test (about 3.7%), but Pearson's X^2 clearly accepts the submodel (Table A.14). And for collapsed or individual groups the submodel is not rejected by any of the 3 normal tests (Tables A.15-16).

In all examples above the normal tests of submodels based on the *original* grouping are in accordance with the deviance test (which also needs the original groups in an essential way). However collapsing or refining the grouping may change the results. Further investigations on the power will be necessary to clarify these points for a general setting.

Example 2: Ile-et-Villaine Study on Oesophageal Cancer

This example is included here for two reasons. First, the data is analysed under various aspects by Breslow & Day (1980), and McCullagh (1985ab) uses the data to illustrate his unconditional test, which allows comparison with our method. And second, the body of data is of a rather moderate size (in contrast to example 1). Although the sampling scheme of this study is *not* of the binomial type discussed here, we still apply the normal goodness-of-fit tests above to the data (assuming the data *were* binomial) only for the purpose of illustration and numerical comparison with McCullagh (1985ab).

The data are taken from the Ile-et-Villaine study on oesophageal cancer as given in Appendix I of Breslow & Day (1980). This is a retrospective case-control study with 975 individuals (200 cases and 775 controls), classified according to the 3 covariables: age (6 classes), alcohol and tobacco consumption (4 classes each). Of the 96 possible combinations only J=88 different groups contained at least one individual at risk.

We only look at the model containing the main effects of the covariables viewed as *factors*, denoted by AgeGrp (6 levels), AlcGrp and TobGrp (4 levels each), where Grp stands for 'Group'. The power divergence statistics and their P-levels based on the χ^2 -distribution are given in Table B.1. Although these P-values vary with λ , none of them is small enough to reject the model. Looking at the distribution of the 88 group sizes in Table B.2 the use of the χ^2 -approximation appears questionable. The P-values based on the normal distribution given in table Table B.3 are much higher than the corresponding χ^2 -values and change only slightly with λ . Looking at the simpler approximations for the special cases 1-3 above, we observe a substantial disagreement with T only for T(1).

McCullagh (1985ab) gives an approximate upper 5% point of 121.60 for Pearson's X^2 , which is based on the Cornish-Fisher approximation $\kappa_1 + (z_\alpha + [z_\alpha^2 - 1] \rho_3 / 6) / \sqrt{\kappa_2}$ for the *conditional* limiting distribution and the conditional cumulants $\kappa_1 = 77.38$, $\kappa_2 = 401.1$ and $\rho_3 = 1.98$. The corresponding *conditional P-level* based on the Edgeworth approximation $P\{Z \geq z\} = 1 - \Phi(z) + \varphi(z)(z^2 - 1)\rho_3 / 6$ for the standardization $Z = [X^2 - \kappa_1] / \sqrt{\kappa_2}$ is $P = 23.1\%$. Using our asymptotic cumulants $\mu_1 = 88$, $\sigma_1^2 = 464.44$ of $X^2 = SD_1$, we obtain as an upper 5% point based on the first-order normal approximation the value 123.45, which is very close to McCullagh's value, but far from the corresponding value 97.35 based on the χ^2_{76} -approximation. Although the use of the χ^2 -approximation is not justified in this situation, it does lead to the same decision concerning the fit as the normal approximation (in contrast to example 1 above). Other considerations by Breslow & Day (1980, Sec.6.5) confirm a satisfactory fit of this model as well.

8. Some Practical Aspects of the Test

We conclude the theoretical discussion with some recommendations and comments which reflect our experience with goodness-of-fit tests based on the power-divergence family SD_λ . It must be pointed out however, that any such test is just *one* aspect of judging the fit of a model and should therefore always be accompanied by other diagnostic tools of fit, e.g. examination of residuals, analysis of deviances with respect to enlarged models of particular interest.

- 1) Compute SD_λ for *different* values of λ like:

1	(Pearson)
0	(Likelihood ratio)
$\frac{2}{3}$	(Cressie-Read)

Considerable variations of SD_λ with respect to λ may be due to *lack of fit* or *small expected values* (indicating that the classical χ^2 -approximation may be poor).

- 2) Look at the *distribution* of the group sizes N_j and compute their *harmonic, arithmetic, quadratic mean* and *variance* to get an idea which asymptotic may be appropriate.

- 3) The χ^2 -approximation of SD_λ based on fixed-cells asymptotics only applies if *all* group sizes (resp. expectations) are *reasonably large*. This will typically be *not* the case for *sparse* data sets.

- 4) For *large* degrees of freedom the normal approximation based on increasing-cells asymptotics is appropriate and the normalized statistic T_λ (for the *general* case) is recommended.

- 5) A *two-sided* normal test based on the statistic T_λ under *individual grouping* is always applicable (for a large total sample size N_+) and may be used to detect departures of the model against specific alternatives concerning the link function.

The remaining points apply only for *large* degrees of freedom.

- 6) It is instructive to look at the expectation $\mu_\lambda(\hat{\theta})$ and variance $\sigma_\lambda^2(\hat{\theta})$ of SD_λ , and to compare T_λ with the the different standardizations $T_\lambda^{(i)}$ given for the special cases $i=1,2,3$. This will usually explain possibly different conclusions based on the normal or the χ^2 -approximation.

- 7) For *small* arithmetic mean AM of the group sizes (special case 4 above), the *variance condition* should be checked by comparing the number J of groups with the variance $\sigma_\lambda^2(\hat{\theta})$, the latter should not be small.

8) The exact computation of $\mu_\lambda(\theta)$, $v_\lambda^2(\theta)$, $c_\lambda(\theta)$ in the general case is time-consuming, unless the number K of categories is small (e.g. $K=2$ for binomial sampling). Explicit expressions (in terms of multinomial moments) are available only for $\lambda \in \mathbb{N}$ (e.g. $\lambda=1$) or in special cases, like individual grouping (special case 5). The great advantage of Pearson's normalized statistic T_1 over the other T_λ is its ease of computation and interpretability.

Clearly, further research and simulation studies are needed to clarify the meaning of "*small*" and "*large*" in the above recommendations.

Appendix A: Tables and Figures for Example 1 (Infant Mortality Data)

Table A.1: Model parameters with standard error and 2-sided normal P-level
Logit model 1+S+W+W²+G+G²

Parameter	S.E.	Parameter/SE	2s-P-Level	Variable
31.194477	3.909628	7.978886	0.0000 %	1
-0.203377	0.097691	-2.081843	3.7357 %	SEX#2
-2.820764	0.162318	-17.378020	0.0000 %	WEIGHT
-0.173454	0.030268	-5.730619	0.0000 %	GESTATION
0.182237	0.012198	14.939396	0.0000 %	WEIGHT^2
0.000311	0.000057	5.484005	0.0000 %	GEST^2

Table A.2: Power divergence statistics and P-levels (with resp. to χ^2)
Logit model 1+S+W+W²+G+G²

Distance	λ	SD	D.F.	χ^2 -P-Level
Likelihood Ratio	0	658.5107	702	87.8522 %
Cressie-Read	2/3	681.5320	702	70.3265 %
Pearson	1	789.5946	702	1.1768 %

Table A.3: The distribution (upper part) and characteristics (lower part) of the group sizes N_j for the original and other groupings (see text).

Size N_j	<i>original</i> groups	<i>pooled</i> groups collapsed over:		
		Sex	Birth Weight	Gestation Time
1	27.26%	23.88%	5.26%	4.44%
2	12.85%	11.35%	10.53%	2.22%
3	9.60%	7.33%	5.26%	0.00%
4	6.21%	5.91%	3.95%	2.22%
5	3.95%	4.49%	1.32%	2.22%
6 - 9	10.03%	11.35%	7.89%	2.22%
10 - 19	10.17%	11.35%	10.53%	13.33%
20 - 39	7.06%	7.09%	10.53%	13.33%
40 - 79	5.79%	5.91%	10.53%	13.33%
80 - 159	4.38%	5.44%	9.21%	6.67%
160 -	2.68%	5.91%	25.00%	40.00%
Number of Groups	708	423	76	45
Minimum of Sizes	1	1	1	1
Maximum of Sizes	275	540	1249	1403
Harmonic Mean HM	2.4	2.7	6.2	11.5
Arithmetic Mean AM	19.4	32.5	180.7	305.1
Quadratic Mean QM	45.8	83.1	365.5	518.7
Variance	1717.4	5849.5	100960.5	175977.9

Table A.4: Expectation μ , variance v^2 and quadratic form Q for the statistics SD

Distance	λ	SD	$\mu(\hat{\theta})$	$v^2(\hat{\theta})$	$Q(\hat{\theta})$
Likelihood Ratio	0	658.5107	603.9573	925.2193	101.3733
Cressie-Read	2/3	681.5320	604.8386	2631.0166	450.9972
Pearson	1	789.5946	708.0000	32154.2294	1023.6666

Table A.5: Different normalizations of SD (according to special cases 1-3) with 1-sided normal P-level

Distance	λ	T	$T^{(3)}$	$T^{(2)}$	$T^{(1)}$
<i>Normalized static</i>					
Likelihood Ratio	0	1.9006	1.7935	1.4497	-1.3152
Cressie-Read	2/3	1.6426	1.4952	2.0381	-0.7034
Pearson	1	0.4625	0.4550	2.1684	2.1684
<i>Normal P-level</i>					
Likelihood Ratio	0	2.8675 %	3.6447 %	7.3566 %	90.5772 %
Cressie-Read	2/3	5.0235 %	6.7433 %	2.0770 %	75.9091 %
Pearson	1	32.1878 %	32.4543 %	1.5066 %	1.5066 %

Table A.6: Normal goodness-of-fit statistic with 2-sided P-level based on all 13730 *individual* groups.

Distance	λ	Statistic T	P-Level 2-sided
Likelihood Ratio	0	not defined	
Cressie-Read	2/3	-1.7047	8.8246 %
Pearson	1	-1.2451	21.3078 %

Table A.7: Analysis of deviance for the enlarged and basic model

Source of variation	Deviance	D.F.	P-Level
Difference due to submodel	5.0957	10	88.4695 %
Model : $S*(1+W+W^2+W^3+G+G^2+G^3+W*G)$	653.4150	692	
Submodel: $1+S+W+W^2+G+G^2$	658.5107	702	

Figure A.1: The group sizes represented as areas of circles classified by sex (males above, females below), birth weight and gestation time. The minimal size is 1, see also Table A.3.

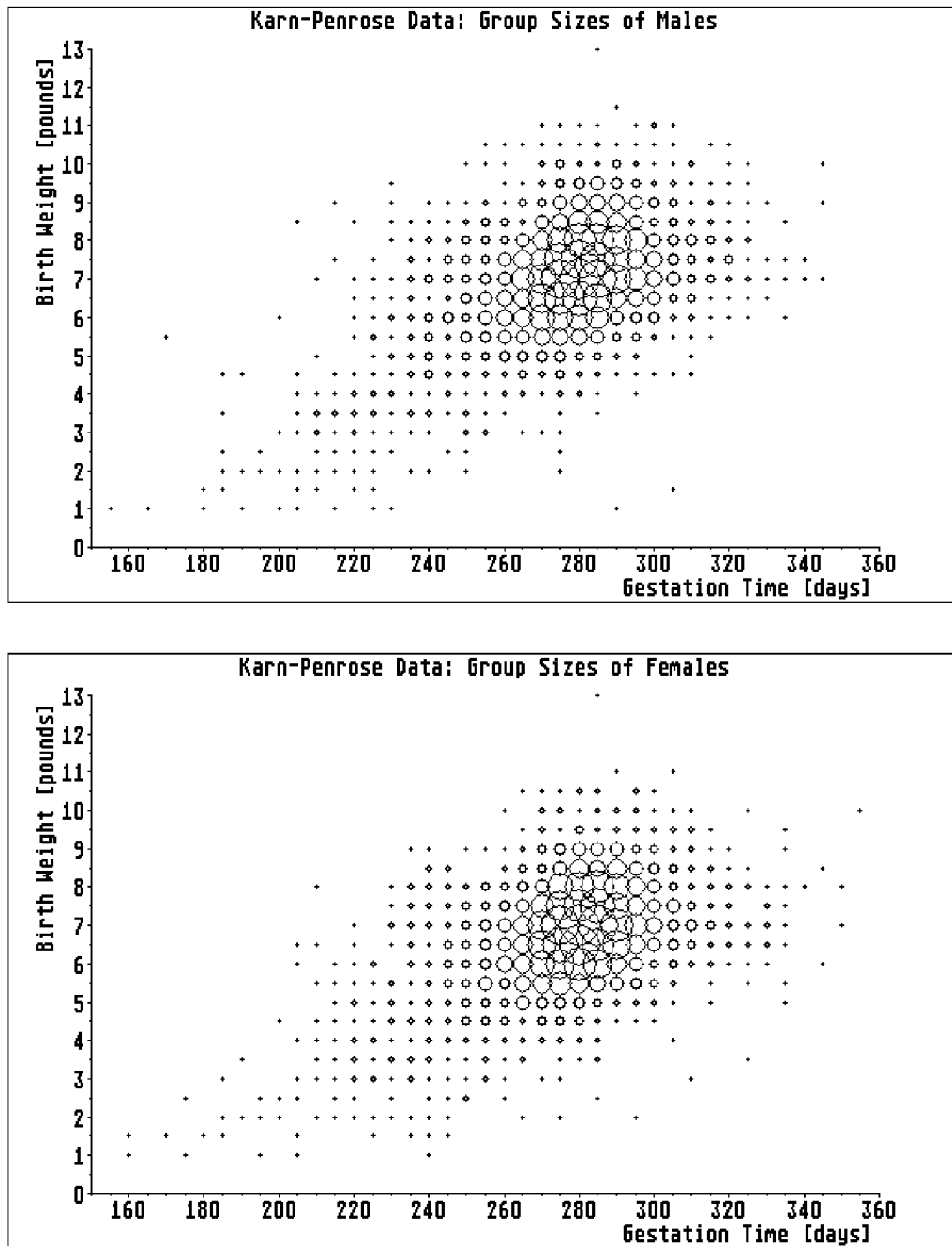


Figure A.2: Plots of the scaled deviance residual against the mortality rate resp. linear predictor, birth weight and gestation time for the logit model $1+S+W+W^2+G+G^2$.

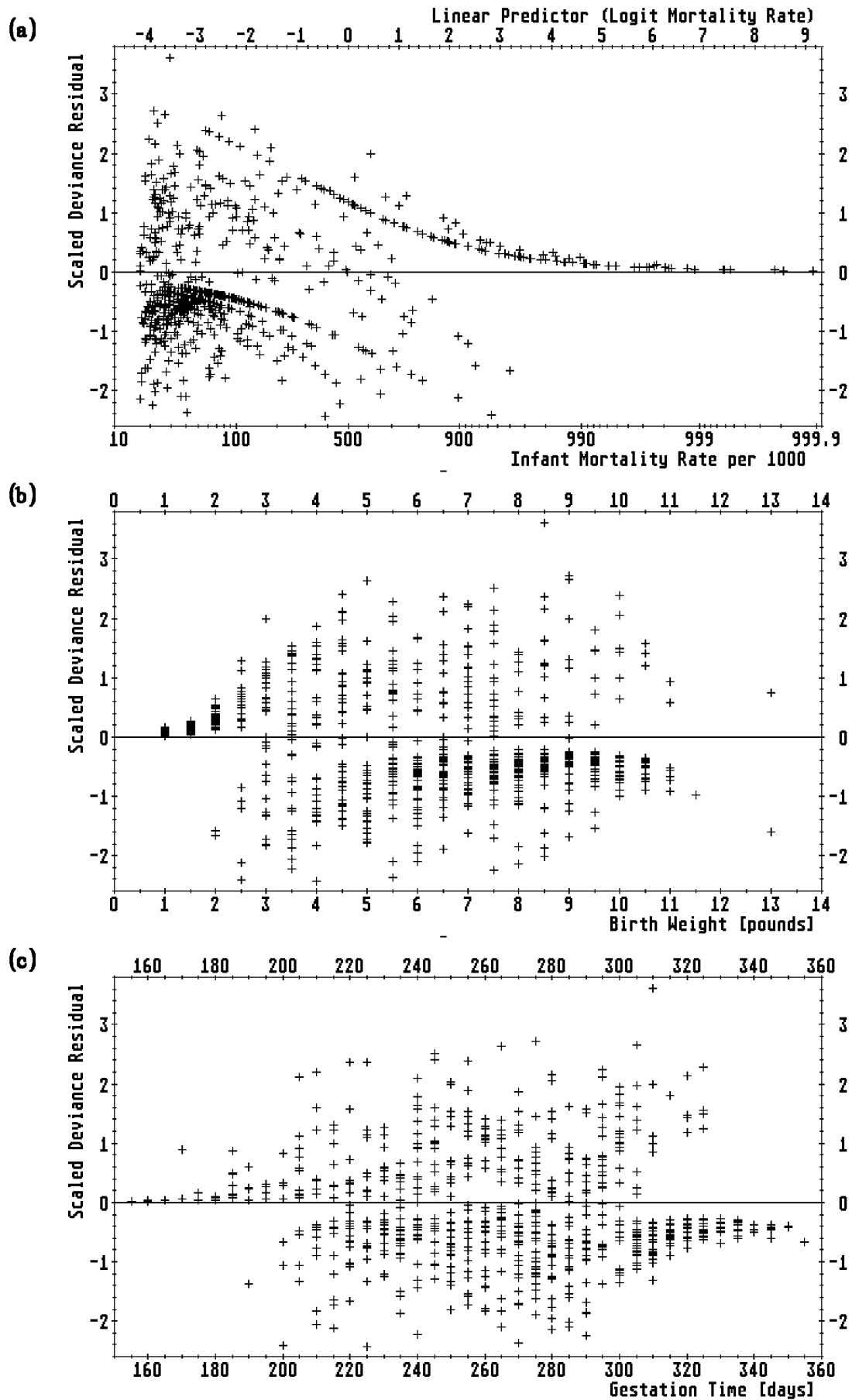


Figure A.3: Plots of the scaled deviance residual against the mortality rate resp. linear predictor for the logit model $1+S+W+W^2+G+G^2$ and different group sizes N_j (see text).

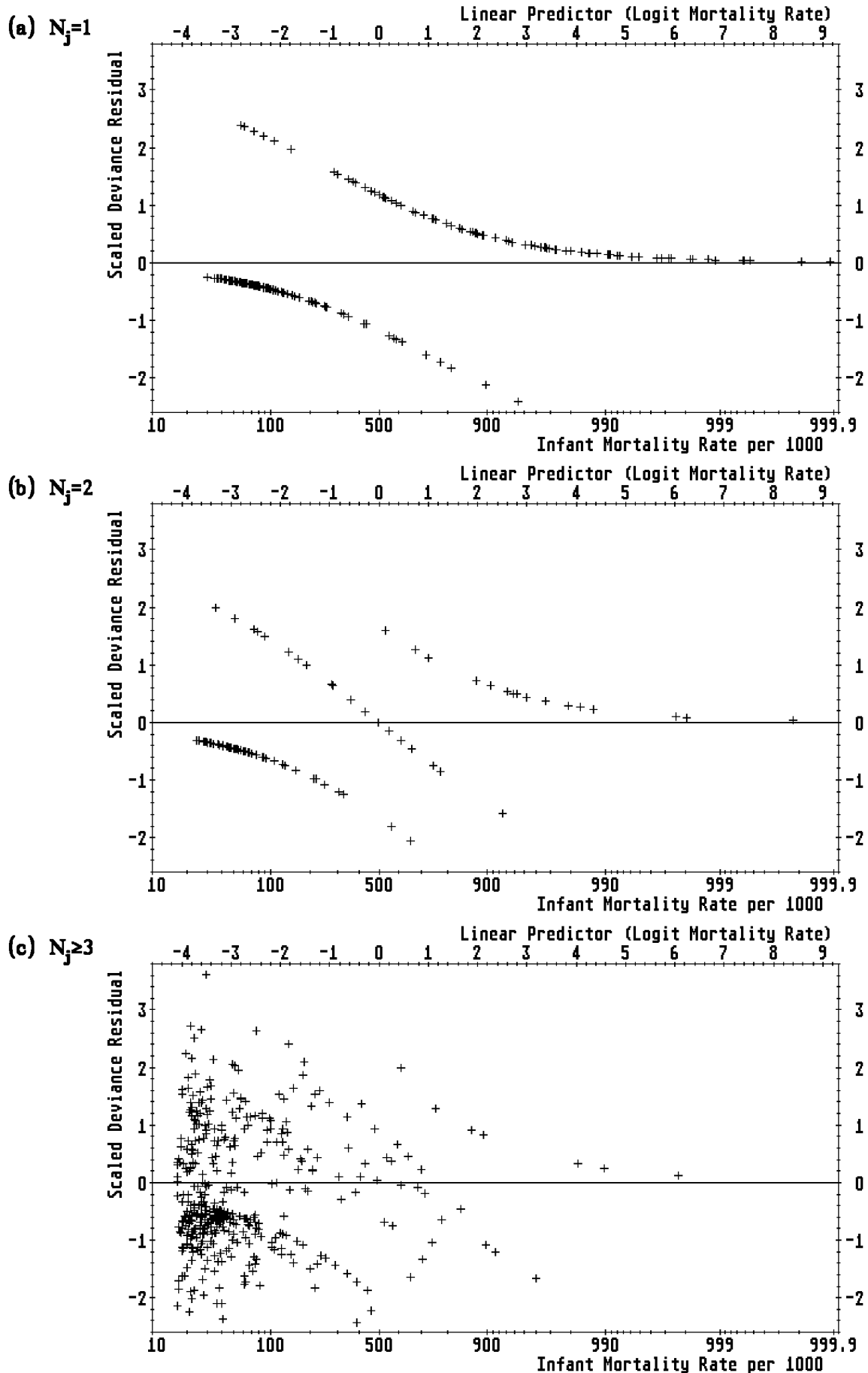


Table A.8: The logit model $1+S+G+G^2$ without variable birth weight

a) Power-divergence statistics for original J=708 groups

Distance	λ	SD	D.F.	χ^2 -P-Level
Likelihood Ratio	0	1127.9039	704	0.0000 %
Cressie-Read	2/3	1472.7266	704	0.0000 %
Pearson	1	2051.2329	704	0.0000 %

b) Normal test (general case) for original J=708 groups

Distance	λ	Statistic T	P-Level 1-sided
Likelihood Ratio	0	17.3916	0.0000 %
Cressie-Read	2/3	19.0161	0.0000 %
Pearson	1	9.7544	0.0000 %

c) Analysis of deviance

Source of variation	Deviance	D.F.	P-Level
Difference due to submodel	469.3932	2	0.0000 %
Model : $1+S+W+W^2+G+G^2$	658.5107	702	
Submodel: $1+S+G+G^2$	1127.9039	704	

Table A.9: The logit model $1+S+G+G^2$ with $J=76$ *pooled* groups, obtained by collapsing the original groups over birth weight (see also Table A.3)a) Power-divergence statistics for *pooled* $J=76$ groups

Distance	λ	SD	D.F.	χ^2 -P-Level
Likelihood Ratio	0	108.4410	72	0.3566 %
Cressie-Read	2/3	111.9297	72	0.1804 %
Pearson	1	117.1283	72	0.0619 %

b) Normal test (general case) for *pooled* $J=76$ groups

Distance	λ	Statistic T	P-Level 1-sided
Likelihood Ratio	0	3.5484	0.0194 %
Cressie-Read	2/3	2.3581	0.9185 %
Pearson	1	0.4014	34.4052 %

Table A.10: The normal test for the logit model $1+S+G+G^2$ based on all 13 730 *individual* groups.

Distance	Lambda	Statistic T	P-Level 2-sided
Likelihood Ratio	0	not defined	
Cressie-Read	2/3	-2.9772	0.2909 %
Pearson	1	-2.0852	3.7049 %

Table A.11: The logit model $1+S+W+W^2$ without variable gestation time

a) Power-divergence statistics for original J=708 groups

Distance	λ	SD	D.F.	χ^2 -P-Level
Likelihood Ratio	0	697.7130	704	55.9724 %
Cressie-Read	2/3	795.8668	704	0.8972 %
Pearson	1	1016.4541	704	0.0000 %

b) Normal test (general case) for original J=708 groups

Distance	λ	Statistic T	P-Level 1-sided
Likelihood Ratio	0	14.3599	0.0007 %
Cressie-Read	2/3	14.3068	0.0008 %
Pearson	1	3.1109	0.0933 %

c) Analysis of deviance

Source of variation	Deviance	D.F.	P-Level
Difference due to submodel	39.2023	2	0.0000 %
Model : $1+S+W+W^2+G+G^2$	658.5107	702	
Submodel: $1+S+W+W^2$	697.7130	704	

Table A.12: The logit model $1+S+W+W^2$ with $J=45$ *pooled* groups, obtained by collapsing the original groups over gestation time (see also Table A.3).a) Power-divergence statistics for *pooled* $J=45$ groups

Distance	λ	SD	D.F.	χ^2 -P-Level
Likelihood Ratio	0	47.8086	41	21.5753 %
Cressie-Read	2/3	48.6018	41	19.3464 %
Pearson	1	50.2625	41	15.2198 %

b) Normal test (general case) for *pooled* $J=45$ groups

Distance	λ	Statistic T	P-Level 1-sided
Likelihood Ratio	0	0.3766	35.3233 %
Cressie-Read	2/3	0.5248	29.9878 %
Pearson	1	0.3909	34.7932 %

Table A.13: The normal test for the logit model $1+S+W+W^2$ based on all 13 730 *individual* groups.

Distance	Lambda	Statistic T	P-Level 2-sided
Likelihood Ratio	0	not defined	
Cressie-Read	2/3	-0.1147	90.8652 %
Pearson	1	-0.0007	99.9447 %

Table A.14: The logit model $1+W+W^2+G+G^2$ without variable sex

a) Power-divergence statistics for original J=708 groups

Distance	λ	SD	D.F.	χ^2 -P-Level
Likelihood Ratio	0	662.8639	703	85.8499 %
Cressie-Read	2/3	686.4157	703	66.5674 %
Pearson	1	797.3371	703	0.7570 %

b) Normal test (general case) for original J=708 groups

Distance	λ	Statistic T	P-Level 1-sided
Likelihood Ratio	0	1.9972	2.2901 %
Cressie-Read	2/3	1.7310	4.1727 %
Pearson	1	0.4974	30.9462 %

c) Analysis of deviance

Source of variation	Deviance	D.F.	P-Level
Difference due to submodel	4.3532	1	3.6939 %
Model : $1+S+W+W^2+G+G^2$	658.5107	702	
Submodel: $1+W+W^2+G+G^2$	662.8639	703	

Table A.15: The logit model $1+W+W^2+G+G^2$ with $J=423$ *pooled* groups, obtained by collapsing the original groups over sex (see also Table A.3).a) Power-divergence statistics for *pooled* $J=423$ groups

Distance	λ	SD	D.F.	χ^2 -P-Level
Likelihood Ratio	0	392.4233	418	81.0604 %
Cressie-Read	2/3	411.7469	418	57.7017 %
Pearson	1	473.8105	418	3.0515 %

b) Normal test (general case) for *pooled* $J=423$ groups

Distance	λ	Statistic T	P-Level 1-sided
Likelihood Ratio	0	0.8496	19.7778 %
Cressie-Read	2/3	1.1558	12.3882 %
Pearson	1	0.2980	38.2834 %

Table A.16: The normal test for the logit model $1+W+W^2+G+G^2$ based on all 13 730 *individual* groups.

Distance	Lambda	Statistic T	P-Level 2-sided
Likelihood Ratio	0	not defined	
Cressie-Read	2/3	-1.4958	13.4702 %
Pearson	1	-1.0703	28.4484 %

Appendix B: Tables for Example 2 (Ile-et-Villaine Study)

Logit Model: 1+AgeGrp+AlcGrp+TobGrp

Table B.1: Power divergence statistics and P-Levels (with resp. to χ^2)

Distance	λ	SD	D.F.	χ^2 -P-Level
Likelihood Ratio	0	82.3369	76	28.9754 %
Cressie-Read	2/3	79.4093	76	37.2057 %
Pearson	1	86.5574	76	19.1302 %

Table B.2: Distribution and characteristics of group sizes N_j for all J=88 groups

Size N_j	Frequency	Number of Groups:	88
1	13.64%	Minimum of Sizes:	1
2	7.95%	Maximum of Sizes:	60
3	9.09%	Harmonic Mean HM	3.5
4	10.23%	Arithmetic Mean AM	11.1
5	3.41%	Quadratic Mean QM	16.8
6 - 9	18.18%	Variance	160.0
10 - 19	21.59%		
20 - 39	9.09%		
40 -	6.82%		

Table B.3: Different normalizations of SD (according to special cases 1-3) with 1-sided normal P-level

Distance	λ	T	$T^{(3)}$	$T^{(2)}$	$T^{(1)}$
<i>Normalized static</i>					
Likelihood Ratio	0	-0.1579	-0.1498	-0.1368	-0.4269
Cressie-Read	2/3	-0.0634	-0.0513	-0.0618	-0.6475
Pearson	1	-0.0669	-0.0531	-0.1087	-0.1087
<i>Normal P-level</i>					
Likelihood Ratio	0	56.2725 %	55.9556 %	55.4411 %	66.5265 %
Cressie-Read	2/3	52.5256 %	52.0463 %	52.4652 %	74.1361 %
Pearson	1	52.6685 %	52.1181 %	54.3295 %	54.3295 %

Table B.4: Expectation μ , variance v^2 and quadratic form Q for the statistics SD

Distance	λ	SD	$\mu(\hat{\theta})$	$v^2(\hat{\theta})$	$Q(\hat{\theta})$
Likelihood Ratio	0	82.3369	84.1519	146.7202	14.5594
Cressie-Read	2/3	79.4093	80.2296	255.5301	87.8649
Pearson	1	86.5574	88.0000	737.5930	273.1526

References

- Breslow, N.E. & Day, N.E.(1980): *Statistical Methods in Cancer Research, Volume 1, The Analysis of Case-Control Studies*. Lyon, IARC.
- Dale, J.R.(1986): *Asymptotic normality of goodness-of-fit statistics for sparse product multinomials*. J.Royal Statist.Soc. B, 48, 48-59.
- Karn, M.N. & Penrose, L.S. (1951-52): *Birth Weight and Gestation Time in Relation to Maternal Age, Parity and Infant Survival*. Ann. Eugenics 16, 147-164.
- Köhler, K.J (1986): *Goodness-of-fit tests for log-linear models in sparse contingency tables*. J.Amer.Statist.Soc. 81, 483-493.
- McCullagh, P.(1985a): *On the asymptotic distribution of Pearson's statistic in linear exponential family models*. Int. Statist.Review 53, 61-67
- McCullagh, P.(1985b): *Sparse data and conditional tests*. Bull.Int.Statist.Inst., Proc. 45th Session of ISI (Amsterdam), Invited Paper 28.3,1-10.
- McCullagh, P.(1986): *The conditional distribution of goodness-of-fit statistics for discrete data*. J.Amer.Statist.Ass. 81, 104-107.
- McCullagh, P. & Nelder, J.A. (1989): *Generalized Linear Models* (2nd. Ed.). London (Chapman & Hall).
- Osius, G.(1985): *Goodness-of-fit tests for binary data with (possible) small expectation but large degrees of freedom*. Statistics&Decision, Suppl.No. 2, 213-224
- Osius, G. (1986): *Anpassungstest: Neuer Beweis-Fahrplan*. Unpublished manuscript.
- Read, T.R.C. & Cressie, N.A.C.(1984): *Multinomial goodness-of-fit tests*. J.Royal Statist.Soc. B, 46, 440-464.
- Read, T.R.C. & Cressie, N.A.C.(1988): *Goodness-of-fit statistics for discrete multivariate data*. New York (Springer)
- Rojek, D.(1989): *Asymptotik für Anpassungstests in Produkt-Multinomial-Modellen bei wachsendem Freiheitsgrad*. Ph.D. Thesis, Universität Bremen, FB3.
- Date: 19-Dec-1989 (printed edition)
10-Jan-2001 (PDF-File, with minor corrections)

Vertrieb der Hefte 4, 14, 23, 26 durch Universitätsbuchhandlung, Bibliothekstr. 3, D-28359 Bremen. Vertrieb der übrigen Hefte (soweit nicht vergriffen) durch die Autoren oder FB 3 Mathematik/Informatik Universität Bremen, Postfach 330440, D-28334 Bremen.

1. Ulrich Krause (1976): Strukturen in unendlichdimensionalen konvexen Mengen, 74 S.
2. Fritz Colonius, Diederich Hinrichsen (1976): Optimal control of hereditary differential systems. Part I, 66 S.
3. Günter Matthiessen (1976): Theorie der heterogenen Algebren, 88 S.
4. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer (1976): Skript zur Analysis, Band 1 (11. Auflage 2000), 286 S.
5. Wolfgang Schröder (1977): Operator-algebraische Ergodentheorie für Quantensysteme, 59 S.
6. Rolf Röhrig, Michael Unterstein (1977): Analyse multivariabler Systeme mit Hilfe komplexer Matrixfunktionen, 216 S.
7. Horst Herrlich, Hans-Eberhard Porst, Rudolf-Eberhard Hoffmann, Manfred Bernd Wischnewsky (1976): Nordwestdeutsches Kategorienseminar, 193 S.
8. Fritz Colonius, Diederich Hinrichsen (1977): Optimal Control of Hereditary Differential Systems. Part II: Differential State Space Description, 36 S.
9. Ludwig Arnold (1977): Differentialgleichungen und Regelungstheorie, 185 S.
10. Rudolf Lorenz (1977): Iterative Verfahren zur Lösung großer, dünnbesetzter symmetrischer Eigenwertprobleme, 104 S.
11. Konrad Behnen, Hans-Peter Kinder, Gerhard Osius, Rüdiger Schäfer, Jürgen Timm (1977): Dose-Response-Analysis, 206 S.
12. Hans-Friedrich Münzner, Dieter Prätzel-Wolters (1978): Minimalbasen polynomialer Moduln, Strukturindizes und BRUNOVSKY-Transformationen, 53 S.
13. Konrad Behnen (1978): Vorzeichen-Rangtests mit Nullen und Bindungen, 53 S.
14. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer, Eberhard Oeljeklaus (1978): Skript zur Linearen Algebra, Band 1 (13. Auflage 2000), 249 S.
15. Günter Ludyk (1978): Abtastregelung zeitvarianter Einfach- und Mehrfachsysteme, 54 S.
16. Momme Johs Thomsen (1977): Zur Theorie der Fastalgebren, 146 S.
17. Klaus Horneffer, Horst Diehl (1978): Modellrechnungen zur anaeroben Reduktionskinetik des Cytochroms P-450, 34 S.
18. Horst Herrlich, Rudolf-Eberhard Hoffmann, Hans-Eberhard Porst, Manfred Bernd Wischnewsky (1979): Structure of Topological Categories, 252 S.
19. Hans-Friedrich Münzner, Dieter Prätzel-Wolters (1979): Geometric and moduletheoretic approach to linear systems. Part I: Basic categories and functors, 28 S.
20. Hans-Friedrich Münzner, Dieter Prätzel-Wolters (1979): Geometric and moduletheoretic approach to linear systems. Part II: Moduletheoretic characterization and reachability, 28 S.
21. Eckart Beutler, Hans Kaiser, Günter Matthiessen, Jürgen Timm (1979): Biduale Algebren, 165 S.
22. Horst Diehl, Detlef Harbach, Jürgen Timm (1980): Planung und Auswertung von Atomabsorptions-Spektrometrie-Untersuchungen mit der Additionsmethode, 44 S.
23. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer (1981): Skript zur Analysis, Band 2 (6. Auflage 1996), 299 S.
24. Horst Herrlich (1981): Categorical Topology 1971-1981, 105 S.
25. Horst Herrlich, Rudolf-Eberhard Hoffmann, Hans-Eberhard Porst, Manfred Bernd Wischnewsky (1981): Special Topics in Topology and Category Theory, 108 S.

26. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer (1984): Skript zur Linearen Algebra, Band 2 (7. Auflage 1999), 257 S.
27. Rudolf-Eberhard Hoffmann (1982): Continuous Lattices and Related Topics, 314 S.
28. Horst Herrlich, Rudolf-Eberhard Hoffmann, Hans-Eberhard Porst (1987): Workshop on Category Theory, 169 S.
29. Harald Boehme (1987): Zur Berufspraxis des Diplommathematikers, 16 S.
30. Jürgen Timm (1986): Mathematische Modelle der Dosis-Wirkungsanalyse bei den experimentellen Untersuchungen der Arbeitsgruppe zur karzinogenen Belastung des Menschen durch Luftverunreinigung, 65 S.
31. Dieter Denneberg (1988): Mathematik für Wirtschaftswissenschaftler. I. Lineare Algebra, 97 S.
32. Peter E. Crouch, Diederich Hinrichsen, Anthony J. Pritchard, Dietmar Salamon (1988, previous edition University of Warwick 1981): Introduction to Mathematical Systems Theory, 244 S.
33. Gerhard Osius (1989): Some Results on Convergence of Moments and Convergence in Distribution with Applications in Statistics, 27 S.
34. Dieter Denneberg (1989): Verzerrte Wahrscheinlichkeiten in der Versicherungsmathematik, Quantilsabhängige Prämienprinzipien, 24 S.
35. Eberhard Oeljeklaus (1989): Birational splitting of homogeneous Albanese bundles, 30 S.
36. Gerhard Osius, Dieter Rojek (1989): Normal Goodness-of-Fit Tests for Parametric Multinomial Models with Large Degrees of Freedom, 38 S.
37. Dieter Denneberg (1990): Mathematik zur Wirtschaftswissenschaft. II. Analysis, 59 S.
38. Ulrich Krause, Cornelia Zahlten (1990): Arithmetik in Krull monoids and the cross number of divisor class groups, 29 S.
39. Dieter Denneberg (1990): Subadditive Measure and Integral, 39 S.
40. Ulrich Krause, Peter Ranft (1991): A limit set trichotomy for monotone nonlinear dynamical systems, 31 S.
41. Angelika van der Linde (1992): Statistical analyses with splines: are they well defined? 22 S.
42. Dieter Denneberg (1992): Lectures on non-additive measure and integral (new edition: Non-additive measure and integral. TDLB 27, Kluwer Academic, Dordrecht (1994)), 114 S.
43. Gerhard Osius (1993): Separating Agreement from Association in Log-linear Models for Square Contingency Tables With Applications, 23 S.
44. Hans-Peter Kinder, Friedrich Liese (1995): Bremen-Rostock Statistik Seminar, 5. - 7. März 1992, 110 S.
45. Dieter Denneberg (1995): Extension of a measurable space and linear representation of the Choquet Integral, 30 S.
46. Dieter Denneberg, Michael Grabisch (1996): Shapley value and interaction index, 20 S.
47. Angelika Bunse-Gerstner, Heike Faßbender (1996): A Jacobi-like method for solving algebraic Riccati equations on parallel computers, 24 S.
48. Hans-Eberhard Porst editor (1997): Categorical methods in algebra and topology - a collection of papers in honour of Horst Herrlich, 498 S.
49. Angelika van der Linde, Gerhard Osius (1997): Estimation of nonparametric risk functions In matched case-control studies, 28 S.
50. Angelika van der Linde (1997): Estimating the smoothing parameter in generalized spline-based regression, 46 S.
51. Ursula Müller, Gerhard Osius (1998): Asymptotic normality of goodness-of-fit statistics for sparse Poisson data, 15 S.
52. Ursula Müller (1999): Nonparametric regression for threshold data, 18 S.
53. Gerhard Osius (2000): The association between two random elements – A complete characterization in terms of odds ratios, 32 S.