# NONLINEAR TREND MODELING IN THE ANALYSIS OF CATEGORICAL DATA

**Jan Kalina**

**Abstract**

This paper studies various approaches to testing trend in the context of categorical data. While the linear trend is far more popular in econometric applications, a nonlinear modeling of the trend allows a more subtle information extraction from real data, especially if the linearity of the trend cannot be expected and verified by hypothesis testing. We exploit the exact unconditional approach to propose alternative versions of some trend tests. One of them is the test of relaxed trend (Liu, 1998), who proposed a generalization of the classical Cochran-Armitage test of linear trend. A numerical example on real data reveals the advantages of the test of relaxed trend compared to the classical test of linear trend. Further, we propose an exact unconditional test also for modeling association between an ordinal response and nominal regressor. Further, we propose a robust estimator of parameters in the logistic regression model, which is based on implicit weighting of individual observations. We assess the breakdown point of the newly proposed robust estimator.

**Key words:** contingency tables, exact unconditional test, log-linear model, logistic regression, robust estimation

**JEL Code:** C12, C40, C44

## 1 Linear and nonlinear trend modeling

Analysis of categorical data is a classical field of mathematical statistics, which still obtains an attention in recent references. This paper is devoted to modern statistical methods applied to the analysis of nonlinear trend in categorical data. Let us consider the contingency table of observed counts in the total number of $J$ groups (Table 1). We assume a multinomial model, while the table of probabilities is shown in Table 2. A binary event is observed in objects in these random samples. The presence of a certain trait is denoted as success and its absence as

failure. The contingency table of observed counts can be described by a product of *J* binomial models.

**Tab. 1: Contingency table of size *2xJ*.**

| | Group 1 | Group 2 | ... | Group *J* | Sum |
|---|---|---|---|---|---|
| Success | $n_{11}$ | $n_{12}$ | ... | $n_{1J}$ | $n_{1\bullet}$ |
| Failure | $n_{\bullet 1} - n_{11}$ | $n_{\bullet 2} - n_{12}$ | ... | $n_{\bullet J} - n_{1J}$ | $n_{2\bullet}$ |
| Sum | $n_{\bullet 1}$ | $n_{\bullet 2}$ | ... | $n_{\bullet J}$ | $n$ |

Source: Agresti (2010)

**Tab. 2: Table of probabilities corresponding to the observed counts of Table 1.**

| | Group 1 | Group 2 | ... | Group *J* |
|---|---|---|---|---|
| Success | $\pi_1$ | $\pi_2$ | ... | $\pi_J$ |
| Failure | $1-\pi_1$ | $1-\pi_2$ | ... | $1-\pi_J$ |
| Sum | *1* | *1* | ... | *1* |

Source: Agresti (2010)

The analysis of a linear or nonliner trend is an important topic in econometrics and belongs to the most crucial tasks in epidemiology and clinical trials (Senn, 2002). A trend test is (any possible) test of the null hypothesis of homogeneity in the form

$$H_0: \pi_1 = \pi_2 = ... = \pi_J .$$

against such alternative hypothesis, which describes (models) the trend in a particular way. Testing trend can be used for modeling an ordinal variable in dependence on other (nominal or ordinal) categorical variables. In numerous real examples it has been shown that a test of trend may have a much larger power than a test against the general alternative

$$H_1: H_0 \text{ is not true.}$$

The Cochran-Armitage test of linear trend is the most common test of trend in current applications. It is a test of $H_0$ against $H_2: \pi_j = \pi + \beta x_j,\ j=1,...,J,\ \beta \neq 0,$ where $x_1, x_2, ..., x_J$ are scores corresponding to individual samples of the contingency table. However, the test is known to be sensitive to the assumption of linearity of the trend.

Section 2 of this paper describes the test of relaxed trend and proposes an exact unconditional version. Section 3 is devoted to a trend test in modeling an ordinal variable as a response of

a nominal variable. Here again an exact unconditional test is proposed. Finally Section 4 studies a robust trend modeling, which is obtained by robust estimation of parameters in logistic regression.

## 2 Test of relaxed trend

This section describes an exact unconditional test of relaxed trend for tables *2xJ*. This is a novel exact unconditional version of the asymptotic test of Liu (1998), which was proposed for epidemiological applications. We consider again the null hypothesis $H_0$, but a more general alternative hypothesis than $H_2$.

We consider the total number of *J* independent random samples with fixed sample sizes $n_1$, $n_2,...,n_J$. We denote $n = n_1 + n_2 + ... + n_J$.

Firstly let us describe the model of relaxed trend (Liu, 1998). For *j=1,...,J,* let us use the notation $B_j$ for the random event, that a randomly selected object out of the total number of *n* objects belongs to one of groups *1,...,j (j < J)*. The complement of $B_j$ is the random event that the same object belongs to one of the remaining groups *j+1,...,J*. We introduce the notation *P(j)* and *Q(j)* for $P(j)=P(success|B_j)$ and $Q(j)=P(failures|B_j)$. Thus *P(j)* is the probability of success for such object, which belongs to one of the first *j* groups and *Q(j)* is the probability of success of such object, which does not belong to any of the first *j* groups. If *Q(j) - P(j) > 0* holds for a certain *j*, it means that an object in the first *j* groups has a smaller probability than an object in the remaining groups.

The null hypothesis $H_0$ of homogeneity can be expressed as

$$H_0 : Q(j)-P(j)=0 \text{ for each } j=1,...,J,$$

while the alternative hypothesis of relaxed trend is formulated as

$$H_3 : Q(j)-P(j)>0 \text{ for each } j=1,...,J.$$

The model $H_2$ may be suitable for applications with a nonlinear trend, which would not be significant for the test of linear trend.

We describe the asymptotic test of $H_0$ against the alternative of relaxed trend. The condition

$$\min\{U_1, U_2,...,U_{J-1}\}>0 \quad (1)$$

gives evidence in favor of $H_3$, where $(U_1, U_2,...,U_{J-1})^T$ is a score statistic of the test of homogeneity. Liu (1998) defined the statistic as

$$T= \chi^2 \cdot 1\{\min\{U_1, U_2,...,U_{K-1}\}>0\},$$

where $\chi^2$ is the test statistic of the test of Pearson's test of $H_0$ against $H_1$ and *1* denotes an indicator function. The asymptotic test is conditional on the value of $n_{1\bullet}$. The asymptotic test also requires to estimate the probability of (1) under $H_0$, denoted by $\alpha_M$.

The asymptotic test is based on computing Rao's score test (Rao, 2002) of homogeneity. We can say that the tedious computations of Liu (1998) were superfluous, because the Rao's score test was derived by Day and Byar (1979) to be precisely equal to Pearson's $\chi^2$ test of homogeneity.

Now we describe the exact unconditional test of relaxed trend. Exact unconditional tests studied by Agresti (2001) or Kalina (2011) are suitable for smaller sample sizes, because their level is guaranteed not to exceed the nominal level 5 %. We consider all possible contingency tables with fixed marginal counts $n_{1\bullet}$, $n_{2\bullet}$, $n_{\bullet 1}$, $n_{\bullet 2}$,..., $n_{\bullet J}$ in the following form.

**Tab. 3: Contingency table of size *2xJ* with fixed marginal counts.**

| | Group 1 | Group 2 | ... | Group *J* | Sum |
|---|---|---|---|---|---|
| Success | $a_1$ | $a_2$ | ... | $a_J$ | $n_{1\bullet}$ |
| Failure | $n_{\bullet 1}$ - $a_1$ | $n_{\bullet 2}$ - $a_2$ | ... | $n_{\bullet J}$ - $a_J$ | $n_{2\bullet}$ |
| Sum | $n_{\bullet 1}$ | $n_{\bullet 2}$ | ... | $n_{\bullet J}$ | $n$ |

Source: Agresti (2001)

The *p*-value is computed as the sum of likelihoods of all such forms of Table 3, which fulfil (1) and have the $\chi^2$ statistic larger or equal to the value of the $\chi^2$ statistic computed in the observed table. The likelihood of each version of Table 3 is evaluated as sum of *J* binomial probabilies under $H_0$. This however depends on an unknown probability of success (across groups). Therefore the value of $\pi$ is considered, which maximizes the likelihood

among all possible values of $\pi$, which are covered by the confidence interval for $\pi$ on a 95 % confidence.

Additionally, we remark that $\alpha_M$ can be easily evaluated exactly as the sum of likelihoods of all tables (2), which fulfil (1).

## 3 Examples

We illustrate the test of relaxed trend with a numerical example. We consider a dose-response analysis on laboratory mice. Let us consider three samples with fixed sample sizes. A presence of a binary outcome is observed in mice in these samples. A certain chemical is added to nourishment for the mice. Group 1 obtains the smallest amount and group 3 obtains the largest amount of the chemical. There exists a hypothesis that a larger amount of the chemical is associated with the death of the mice, which is the binary outcome of the experiment. We compute the test of linear trend (using values of scores 2, 1 and 0) and the test of relaxed trend.

**Tab. 4: Observed counts in the dose-response experiment.**

| | Group 1 | Group 2 | Group 3 | Sum |
|---|---|---|---|---|
| Death | 3 | 4 | 7 | 14 |
| Survival | 11 | 9 | 2 | 22 |
| Sum | 14 | 13 | 9 | 40 |

Source: own research

**Tab. 5: Results of various tests of trend.**

| Test | Test statistic | *p*-value |
|---|---|---|
| $\chi^2$ test of homogeneity | $\chi^2$ =7.88 | *p*=0.0194 |
| Test of linear trend | $\chi^2_T$ =6.67 | *p*=0.0098 |
| Test of relaxed trend: asymptotic | *T*= 7.88 | *p*=0.0065 |
| Test of relaxed trend: exact unconditional | - | *p*=0.0061 |

Source: own research

In this example, the tests of linear trend and relaxed trend give a more significant conclusion than the test of homogeneity. The test of linear trend depends on the selection of scores. For this reason we consider the test of relaxed trend to be more objective. The test of relaxed trend requires to compute the constant $\alpha_M$, which is equal to $\alpha_M = 0.336$.

Further, we perform a simulation study based on a random generation of samples of binomial distribution, which together form a contingency table. Thus we randomly generate 5000 tables of sizes 2x3 for different situations determined by various values of the probability of success under homogeneity (say $\pi$) and various values of marginal counts $n_{\bullet 1}, n_{\bullet 2}, n_{\bullet 3}$. The following table summarizes estimated probabilities of rejecting $H_0$ by the exact unconditional test of relaxed trend. The test has a nominal 5 % level. The computations are performed in Matlab software.

**Tab. 6: *P*-values of the exact unconditional test of relaxed trend.**

| $\pi$ | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $n_{\bullet 3}$ | *p*-value |
|---|---|---|---|---|
| 0.2 | 5 | 5 | 5 | 0.0890 |
| 0.2 | 10 | 10 | 10 | 0.0738 |
| 0.2 | 5 | 10 | 5 | 0.0608 |
| 0.5 | 5 | 5 | 5 | 0.0576 |
| 0.5 | 10 | 10 | 10 | 0.0506 |
| 0.5 | 5 | 10 | 5 | 0.0562 |
| 0.8 | 5 | 5 | 5 | 0.0852 |
| 0.8 | 10 | 10 | 10 | 0.0730 |
| 0.8 | 5 | 10 | 5 | 0.0550 |

Source: own research

The test has a large computational complexity for larger sample sizes. The results of the simulations allow to conclude that the test of relaxed trend holds the probability of type I error close to the nominal value (5 %) already for small sample sizes, assuming $n_{\bullet 1} = n_{\bullet 2} = n_{\bullet 3}$ and $\pi$=0.5. However, for more extreme values of $\pi$, the asymptotic test exceeds the level above 5 % for small samples.

## 4 Trend modeling in the analysis of ordinal data

This section proposes an exact unconditional test for trend in a contingency table assuming an ordinal response with $I$ outcome values (in rows) in $J$ samples (in columns). We propose a novel exact unconditional version for a log-linear model model, which was described by Agresti (2010).

**Tab. 7: Contingency table of size *IxJ*.**

| | Sample 1 | Sample 2 | ... | Sample J |
|---|---|---|---|---|
| Outcome 1 | $n_{11}$ | $n_{12}$ | ... | $n_{1J}$ |
| Outcome 2 | $n_{21}$ | $n_{22}$ | ... | $n_{2J}$ |
| ... | ... | ... | ... | ... |
| Outcome $I$ | $n_{I1}$ | $n_{I2}$ | ... | $n_{IJ}$ |

Source: Agresti (2010)

Each of the observed counts $n_{ij}$ represents an observation of a random variable $N_{ij}$, which is modeled by Poisson distribution $Po(m_{ij})$. We assume a log-linear model

$$log\ m_{ij} = u + u_1(i) + u_2(j) + \beta_j(x_i - \bar{x}), \quad i=1,...I, j=1,...,J, \tag{2}$$

where parameters $u$, $u_1(i)$ and $u_2(j)$ are standard parameters of a log-linear model (fulfilling a certain set of parametrization constraints), $x_1,...,x_I$ are scores assigned to individual rows of the table and $\beta_1,...,\beta_J$ are parameters for individual columns of the table.

Now we describe an exact unconditional test of homogeneity against the alternative hypothesis, which is described by (2). The null hypothesis $H_0$ can be expressed as

$$H_0: \beta_1 = \beta_2 = ... = \beta_J .$$

The exact unconditional test considers all possible forms of Table 3. The $p$ value is computed as the sum of likelihoods of all such tables, which have a larger residual deviance than the observed table. Here residual deviance is a test statistic comparing the model (2) with a saturated model, which can be computed by software allowing an analysis of generalized linear models. In an analogous way as in Section 2, the likelihood of each form of Table 3 under $H_0$ is evaluated as sum of $J$ binomial probabilities, which depend on an unknown probability of success $\pi$ (across groups).

## 5 Robust logistic regression

The logistic regression is a basic tool for modeling trend of an ordinal variable depending on one or several regressors (continuous or categorical). At the same time it is the most common method among generalized linear models. The maximum likelihood estimation of parameters in the logistic regression is known to be too vulnerable to the presence of outliers. This section proposes a novel robust estimator for parameters of logistic regression.

Some robust estimation procedures have been proposed as an alternative to the classical maximum likelihood method. However, most of them do not possess a high breakdown point, which is a measure of sensitivity of estimators against noise or outliers in the data and a crucial concept in robust statistics. Christmann (1994) explains outliers in the logistic regression mainly by typing errors. Buonaccorsi (2010) warned that outliers appear in real data more commonly as measurement errors. There is a connection between robust statistics and statistical theory of measurement error models, which has obtained an intensive attention in econometrics, which however goes beyond the scope of this paper. We can say briefly that practical situations with errors in measurements in both response and regressors require to use robust statistical methods, which take the measurement errors into account (Saleh et al., 2012). Another example of econometric methods for data contaminated by measurement errors is the (robust) instrumental variables estimator (Kalina, 2012).

Christmann (1994) proposed the least median of squares method for estimating parameters in the logistic regression model. He proved the estimator to possess the maximal possible breakdown point. Nevertheless, the least median of squares is known to possess a very low efficiency (Hekimoglu et al., 2009). In this section, we propose a robust estimator of logistic regression parameters based on the least weighted squares estimator (Víšek, 2002; Čížek, 2011; Kalina, 2012) and derive its breakdown point.

We recall the least weighted squares (LWS) regression, which is a highly robust estimator in linear regression proposed by Víšek (2002). There must be nonnegative weights $w_1, w_2, ..., w_n$ specified before the computation of the estimator. These are assigned to the data after a permutation, which is determined automatically only during the computation based on the residuals. It is reasonable to choose such weights so that the sequence $w_1, w_2, ..., w_n$ is

decreasing (non-increasing), so that the most reliable observations obtain the largest weights, while outliers with large values of the residuals get small (or zero) weights.

Let us denote the $i^{\text{th}}$ order value among the squared residuals for a particular value of the estimate $\boldsymbol{b}$ of the parameter $\boldsymbol{\beta}$ by $u_i^2(\mathrm{b})$. The least weighted squares estimator $\boldsymbol{b}_{LWS}$ for the linear regression model is defined as

$$\boldsymbol{b}_{LWS} = \operatorname{argmin} \sum_{i=1}^{h} w_i u_{(i)}^2(b). \tag{3}$$

The least weighted squares estimator combines a high robustness with a high efficiency for normal data. Let us now come to the definition of a robust LWS-based estimator for the logistic regressio model.

We consider a binary variable $(Y_1,...,Y_n)^T$, which is explained in a logistic regression by regressors $X_1,...,X_n$. The regressors are *p*-dimensional variables. The conditional distribution of $Y_i$ assuming fixed values of the regressors is assumed to be binomial $Bi(m_i,\pi_i)$, where $\pi_i$ depends on regression parameters $\beta_1,...,\beta_p$. We introduce the notation

$$v_i = (m_i\pi_i(1-\pi_i))^{1/2},\ \tilde{X}_i = v_i X_i \text{ and } \tilde{Y}_i = v_i \log(\pi_i/(1-\pi_i)).$$

We define the least weighted logistic regression (LWLR) estimator as the least weighted squares estimator (3) computed for the data $(\tilde{X}_1,\tilde{Y}_1)^T,...,(\tilde{X}_n,\tilde{Y}_n)^T$. We recommend to use the data-dependent adaptive weights of Čížek (2011), which yield a high breakdown point. This allows us to derive the breakdown point of the LWLR estimator.

*Theorem 1.*

Under technical assumptions of Christmann (1994), the breakdown point of the least weighted logistic estimator is equal to $[(n+1)/2]-(p+1)$, where $[x]$ denotes the integer part of *x*.

## Acknowledgment

The paper was supported by RVO 67985807.

## References


1. Agresti A.: *Analysis of ordinal categorical data.* Second edition. New York: Wiley, 2010.
2. Agresti A.: Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine* 20, 2709-2722, 2001.
3. Buonaccorsi J.P.: *Measurement error: models, methods, and applications.* Boca Raton: Chapman & Hall/CRC, 2010.
4. Christmann A.: Least median of weighted squares in logistic regression with large strata. *Biometrika* 81 (2), 413-417, 1994.
5. Čížek P.: Semiparametrically weighted robust estimation of regression models. *Computational Statistics and Data Analysis* 55 (1), 774-788, 2011.
6. Day N.E., Byar D.P.: Testing hypotheses in case-control studies. Equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics* 35, 623-630, 1979.
7. Hekimoglu S., Erenoglu R.C., Kalina J.: Outlier detection by means of robust regression estimators for use in engineering science. *Journal of Zhejiang University* 10 (6), 909-921, 2009.
8. Kalina J.: On multivariate methods in robust econometrics. *Prague Economic Papers* 1/2012, 69-82, 2012.
9. Kalina J.: Some tests for evaluation of contingency tables (for biomedical applications). *Journal of Applied Mathematics, Statistics and Informatics* 7 (1), 37-50, 2011.
10. Liu Q.: An order-directed score test for trend in ordered 2×K tables. *Biometrics* 54, 1147-1154, 1998.
11. Rao C.R.: *Linear statistical inference and its applications.* New York: Wiley, 2002.
12. Saleh A.K.Md.E., Picek J., Kalina J.: R-estimation of the parameters of a multiple regression model with measurement errors. *Metrika* 75 (3), 311-328, 2012.
13. Senn S.: *Statistical issues in drug development*. Chichester: Wiley, 2002.

14. Víšek J.Á.: The least weighted squares I. *Bulletin of the Czech Econometric Society* 9 (15), 31-56, 2002.

## Contact

RNDr. Jan Kalina, Ph.D.

Institute of Computer Sciences, Academy of Sciences of the Czech Republic

Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic

kalina@cs.cas.cz