Statistical methods for microarray data

Goeman, Jelle Jurjen

CHAPTER 4

A goodness-of-fit test for multinomial logistic regression

Abstract

This paper presents a score test to check the fit of a logistic regression model with two or more outcome categories. The null hypothesis that the model fits well is tested against the alternative that residuals of samples close to each other in covariate space tend to deviate from the model in the same direction. We propose a test statistic that is a sum of squared smoothed residuals, and show that it can be interpreted as a score test in a random effects model. By specifying the distance metric in covariate space, users can choose the alternative against which the test is directed, making it either an omnibus goodness-of-fit test or a test for lack of fit of specific model variables or outcome categories.

4.1 Introduction

The multinomial logistic regression model is a generalization of logistic regression to outcomes with more than two levels. The model is also known as polytomous or polychotomous logistic regression in the health sciences and as the discrete choice model in econometrics (Hosmer and Lemeshow, 2000). Two variants exist: one for nominal and one for ordinal scale outcomes. This paper considers only the nominal scale version.

When fitting a model it is important to have tools to test for lack of fit. This is especially important for the multinomial logistic model, whose fit is notoriously difficult to visualize. The modelling toolbox should include general tests for the fit of the whole model, but also more specific tests for lack of fit in specific

This chapter will appear as: J. J. Goeman and S. le Cessie (2006). A goodness-of-fit test for multinomial logistic regression. *Biometrics* **62**, in press. The definitive version will be available at `http://www.blackwell-synergy.com`.

covariates or outcome categories. Such tools are remarkably scarce in multinomial logistic regression. Hosmer and Lemeshow (2000) suggested looking at the multinomial model as if it were a set of independent ordinary logistic models of each outcome against the reference outcome, and testing the fit of each of these separately. Lesaffre and Albert (1989) give diagnostics for detecting influential, leverage and outlying samples in multinomial logistic regression, but provided no explicit goodness-of-fit test. The only actual test for the fit of the multinomial logistic regression model is given by Pigeon and Heyse (1999). It is an extension of the test of Hosmer and Lemeshow (2000) for binary regression, which is well known to have low power for detecting quadratic effects (Le Cessie and Van Houwelingen, 1991).

In this paper we present an alternative and flexible goodness-of-fit test for the multinomial logistic regression model. It can be directed against the general alternative that the model does not fit or against more specific alternatives. The test extends the goodness-of-fit test of Le Cessie and Van Houwelingen (1991) for ordinary logistic regression to the multinomial case. The approach is to smooth the regression residuals and to test whether these smoothed residuals have more variance than expected under the null hypothesis, which occurs when residuals which are close together in the covariate space are correlated. This type of test was shown by Le Cessie and van Houwelingen (1995) to be equivalent to a score test in a random effects model, which tests for the presence of a pre-specified correlation structure between the residuals. Their approach to goodness-of-fit testing is quite generally applicable, and has already been extended to generalized linear models (Le Cessie and van Houwelingen, 1995) and to the Cox proportional hazards model (Verweij et al., 1998). This paper extends the methodology to multinomial logistic regression.

The properties of the resulting test are verified using simulated data and illustrated on a liver enzyme data set (Albert and Harris, 1987). Software in *R* for fitting and testing the fit of the model is available on request from the authors.

4.2 The multinomial logistic regression model

Suppose the multinomial outcome variable Y takes values in the unordered set $\{1, \ldots, g\}$. The multinomial logistic regression model assumes that the probability for observation i to have outcome s depends on i's covariates $x_{i1}, \ldots, x_{ip}$ as

$$P(Y_i = s) = \frac{e^{\eta_{is}}}{\sum_{t=1}^{g} e^{\eta_{it}}} \tag{4.1}$$

where $\eta_{is} = \sum_{k=1}^{p} x_{ik}\beta_{ks}$ is a linear predictor. In this formulation of the model we have a regression coefficient β_{ks} for each combination of covariate k and outcome category s, and a separate linear predictor η_{is} for each outcome category (for a more detailed description of the model, see Hosmer and Lemeshow, 2000).

The model as defined in (4.1) is overparametrized. Replacing $(\beta_{k1}, \ldots, \beta_{kg})$ with $(\beta_{k1} + c, \ldots, \beta_{kg} + c)$, for any $c \in \mathbb{R}$ and $k \in \{1, \ldots, p\}$, leads to exactly the same probabilities. The most common way to solve this overparametrization is to designate one outcome category, say outcome 1, as the "reference" category, setting all regression coefficients $\beta_{11}, \ldots, \beta_{p1}$ to zero. A good choice of the reference category will usually facilitate interpretation of the resulting parameter estimates. However, in this paper we are not concerned with estimation but rather with assessment of the fit, which does not depend on the choice of the reference category. We will therefore refrain from choosing a reference category, but instead treat the outcome categories symmetrically, leaving the model overparametrized.

Suppose we have sampled outcomes $Y_1, \ldots, Y_n$ and a corresponding $n \times p$ design matrix X. Then let y_{is} be the indicator of the event $\{Y_i = s\}$, for $i = 1, \ldots, n$, and $s = 1, \ldots, g$, and call the corresponding probabilities $\mu_{is} = \mathrm{P}(Y_i = s)$. Let $\hat{\mu}_{is}$ be the maximum likelihood estimates of μ_{is} for all i and s. The fitted model has $n \times g$ residuals $\hat{r}_{is} = y_{is} - \hat{\mu}_{is}$, one for each individual i and outcome category s. These residuals fulfill $\sum_{s=1}^{g} \hat{r}_{is} = 0$. It will be convenient to gather the residuals together in a long vector $\hat{\mathbf{r}} = \mathbf{y} - \hat{\boldsymbol{\mu}}$, where $\mathbf{y} = (y_{11}, \ldots, y_{n1}, \ldots, y_{1g}, \ldots, y_{ng})'$ and $\boldsymbol{\mu} = (\mu_{11}, \ldots, \mu_{n1}, \ldots, \mu_{1g}, \ldots, \mu_{ng})'$.

4.3 Testing goodness-of-fit by smoothing

A goodness-of-fit test tests a model against the alternative that the model 'does not fit'. This is an extremely broad class of alternatives: lack of fit comes in many different shapes and sizes. A linear model, for example, can display lack of fit when the distribution of the residuals is skewed or heavy-tailed, or when there are non-linear relationships which fit the data better. Typically, there is no single goodness-of-fit test which has good power against all kinds of lack of fit. For better interpretation, a goodness-of-fit test should therefore be specific about the type of lack of fit is directed against.

The goodness-of-fit test of this paper is directed against the alternative that any non-linearities or interaction effects have been missed. Such neglected effects can be detected by looking for patterns in the residuals: observations close to each other in covariate space which deviate from the model in the same direction. One looks for this same kind of behaviour when making a scatterplot

of the residuals against a covariate. The test can also detect different kinds of lack of fit which show up as patterns of correlation in the residuals, such as over-dispersion.

One can formally test for patterns in the residuals by smoothing the residuals: the smoothed residuals are a weighted average of the residual itself and the other residuals which are close to it in covariate space. If residuals close to each other are strongly correlated, the smoothing will not affect the magnitude of the residuals much, while if they are not correlated smoothing will shrink the residuals toward zero. The sum of squares of the smoothed residuals is therefore a good measure of the correlations of residuals close to each other in covariate space (Le Cessie and Van Houwelingen, 1991).

Based on these arguments we propose to reject for large values of the test statistic

$$Q = \sum_{s=1}^{g} \sum_{i=1}^{n} \Big[\sum_{j=1}^{n} u_{ij}(y_{js} - \hat{\mu}_{js}) \Big]^2 \tag{4.2}$$

where $u_{ij} \geq 0$ is the i,j-th entry of a smoothing matrix U, fulfilling $\sum_{j=1}^{n} u_{ij} = 1$ for all i. The statistic Q is a sum of squared smoothed residuals, as each $\tilde{r}_{is} = \sum_{j=1}^{n} u_{ij}(y_{js} - \hat{\mu}_{js})$ is a smoothed version of the residual $\hat{r}_{is}$. Note that smoothing of the residual $\hat{r}_{is}$ only involves residuals of the same outcome category s, as the residuals corresponding to different categories are not expected to be positively correlated.

There are various possibilities for the choice of the smoothing matrix U. This choice has two aspects: the choice of a distance measure and the choice of a smoothing method. Of these two, the choice of distance measure deserves most consideration. To test globally for lack of fit one could take euclidian distance using all covariates. As euclidian distance is sensitive to the scaling of the variables, it is wise to rescale the variables to unit variance to prevent one covariate dominating the distance measure. If, on the other hand, the interest is in testing lack of fit for a specific subset of the covariates, one should only use that subset for constructing the distance measure. The choice of a smoothing method is less of an issue. Let d_{ij} be the chosen distance between observations i and j. Following Le Cessie and van Houwelingen (1995) one could choose a kernel smoother based on this distance. A convenient choice is the uniform kernel which has $K(t) = 1$ if $-1 \leq t \leq 1$, and $K(t) = 0$ otherwise. The resulting smoothing matrix U will have entries

$$u_{ij} = \frac{K(d_{ij}/h)}{\sum_{k=1}^{n} K(d_{ik}/h)}.$$

Here h is the bandwidth, which should be chosen carefully: taking h too large results in oversmoothing, while taking h too small results in undersmoothing.

Both will lead to low power. The choice of h can be related to the distribution of the distances d_{ij}, $i \neq j$: our experience is that taking h as the 25-th percentile of this distribution is a often good choice. Using a kernel smoother, the smoothed residual $\tilde{r}_{is}$ will be the average of all residuals $\hat{r}_{js}$ with $d_{ij} \leq h$.

4.4 Distribution of the test statistic

To be able to use the test statistic Q for testing we must calculate or approximate its distribution function.

Write $\mathbf{U} = I_g \otimes U$, where $\otimes$ denotes the Kronecker product and I_g the $g \times g$ identity matrix, and write $\mathbf{R} = \mathbf{U}\mathbf{U}'$. Then we can write $\tilde{\mathbf{r}} = \mathbf{U}'\hat{\mathbf{r}}$ and

$$Q \;=\; \|\tilde{\mathbf{r}}\|^2 \;=\; (\mathbf{y} - \hat{\boldsymbol{\mu}})'\mathbf{R}(\mathbf{y} - \hat{\boldsymbol{\mu}}),$$

which is a non-negative quadratic form.

There is no exact expression for the null distribution function of Q, but there are various approaches for finding an approximation. The most promising approach follows asymptotic arguments. Assuming that as n grows new observations are added which have the same covariate patterns as those already present, it can be shown that Q converges in distribution to a linear combination of chi-squared variables with one degree of freedom. There is no simple explicit expression for the distribution function of a such a distribution, but it is known that it can be well approximated by a general scaled chi-squared (or gamma) distribution. This is often used as an approximate distribution for quadratic forms (Cox and Hinkley, 1974, p. 462–463), although more accurate approximations exist (Solomon and Stephens, 1978). The gamma approximation was also used for the test of Le Cessie and van Houwelingen (1995) which this paper extends. It should be calibrated to have the same mean and variance as Q as well as to the fact that $Q \geq 0$, resulting in a gamma distribution with parameters $\alpha = (\mathrm{E}Q)^2/\mathrm{Var}(Q)$ and $\lambda = \mathrm{E}Q/\mathrm{Var}(Q)$. The accuracy of this approximation will be checked with a simulation example in section 4.7.

To use this approximation we have to calculate expectation and variance of Q. This involves the distribution of the estimated residuals $\mathbf{y} - \hat{\boldsymbol{\mu}}$, which can be related to the easier distribution of the true residuals $\mathbf{y} - \boldsymbol{\mu}$ through its first order approximation, using standard theory from generalized linear models (McCullagh and Nelder, 1989). If n is not too small,

$$\mathbf{y} - \hat{\boldsymbol{\mu}} \approx (I - \mathbf{H})(\mathbf{y} - \boldsymbol{\mu}) \tag{4.3}$$

where $\mathbf{H}$ is the asymmetric form of the hat matrix for the multinomial logistic regression model. It is defined as $\mathbf{H} = \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\mathbf{X}'$, where $\mathbf{X} = I_g \otimes X$,

superscript minus denotes a generalized inverse, and $\mathbf{W}$ is given by

$$\mathbf{W} = \begin{pmatrix} W^{11} & W^{12} & \cdots & W^{1g} \\ W^{21} & W^{22} & & \vdots \\ \vdots & & \ddots & \\ W^{g1} & \cdots & & W^{gg} \end{pmatrix}, \tag{4.4}$$

where each W^{ij} is an $n \times n$ diagonal matrix with

$$\text{diag}(W^{st}) = \text{diag}(W^{ts}) = \begin{cases} (-\mu_{1s}\mu_{1t}, \ldots, -\mu_{ns}\mu_{nt})' & \text{if } s \neq t \\ (\mu_{1s}(1-\mu_{1s}), \ldots, \mu_{ns}(1-\mu_{ns}))' & \text{if } s = t \end{cases}$$

The hat matrix $\mathbf{H}$ also plays an important role in the paper of Lesaffre and Albert (1989), where it is used to detect influential observations. From the approximation (4.3) it follows that if n is not too small, the distribution of Q is approximately the same as the distribution of

$$\tilde{Q} = (\mathbf{y} - \boldsymbol{\mu})'\tilde{\mathbf{R}}(\mathbf{y} - \boldsymbol{\mu}),$$

where $\tilde{\mathbf{R}} = (I - \mathbf{H})'\mathbf{R}(I - \mathbf{H})$.

Under the null hypothesis, $\text{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'] = \mathbf{W}$, so that

$$\text{E}\tilde{Q} = \text{trace}(\tilde{\mathbf{R}}\mathbf{W}).$$

The variance under H_0 of Q is calculated in Section 4.10. It is given by

$$\text{Var}(\tilde{Q}) = 2\text{trace}(\tilde{\mathbf{R}}\mathbf{W}\tilde{\mathbf{R}}\mathbf{W}) + \sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{u=1}^{g}\sum_{v=1}^{g}\sum_{i=1}^{n} \tilde{R}_{ii}^{st}\tilde{R}_{ii}^{uv}\kappa_i^{stuv} \tag{4.5}$$

In this expression, $\tilde{R}_{ij}^{st}$ is the i,j-th element of the submatrix $\tilde{R}^{st}$ of $\tilde{\mathbf{R}}$, which is similarly decomposed as $\mathbf{W}$ in (4.4). The value of κ_i^{stuv} does not depend on the order of s, t, u and v: it can be calculated with

$$\begin{aligned} \kappa_i^{ssss} &= \mu_{is} - 7\mu_{is}^2 + 12\mu_{is}^3 - 6\mu_{is}^4 \\ \kappa_i^{ssst} &= -\mu_{it}\mu_{is} + 6\mu_{it}\mu_{is}^2 - 6\mu_{it}\mu_{is}^3 \\ \kappa_i^{sstt} &= -\mu_{is}\mu_{it} + 2\mu_{is}\mu_{it}^2 + 2\mu_{is}^2\mu_{it} - 6\mu_{is}^2\mu_{it}^2 \\ \kappa_i^{sstu} &= 2\mu_{is}\mu_{it}\mu_{iu} - 6\mu_{is}^2\mu_{it}\mu_{iu} \\ \kappa_i^{stuv} &= -6\mu_{is}\mu_{it}\mu_{iu}\mu_{iv}, \end{aligned} \tag{4.6}$$

after recoding s, t, u and v to denote unique outcomes.

The mean and variance of Q involve the unknown vector $\boldsymbol{\mu}$, which should be estimated by its maximum likelihood estimate $\hat{\boldsymbol{\mu}}$ in applications.

4.5 Testing for the presence of a random effect

The test proposed in Section 4.3 was motivated by heuristic arguments. These arguments give a good impression of the type of alternative the test can be expected to have good power against, but the alternative was not yet precisely specified. In this section we present a fully specified alternative model from which the goodness-of-fit test proposed in Section 4.3 can be derived as a score test. This model explicitly lets observations which are close to each other in covariate space have correlated residuals.

We propose to add an extra random effect z_{is} to the linear predictor η_{is} for each combination of observation i and outcome category s. Given the random effect, the distribution of Y becomes

$$\mathrm{P}(Y_i = s \mid \mathbf{z}) = \frac{e^{\eta_{is}+z_{is}}}{\sum_{t=1}^{g} e^{\eta_{it}+z_{it}}} \tag{4.7}$$

where $\mathbf{z} = (z_{11}, \ldots, z_{n1}, \ldots, z_{1g}, \ldots, z_{ng})'$ is the vector of all random effects. We do not specify a distributional form for $\mathbf{z}$, but we specify its first and second moments as $\mathrm{E}(\mathbf{z}) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{z}) = \tau^2\mathbf{R}$, where τ^2 is an unknown parameter. The matrix $\mathbf{R} = \mathbf{U}\mathbf{U}'$ here is the same matrix as defined in section 4.4. It can be written $\mathbf{R} = I_g \otimes R$ where $R = UU'$. Let R_{ij}^{st} be the element of $\mathbf{R}$ corresponding to the covariance of the random effects z_{is} and z_{jt}. If U is a smoothing matrix, R_{ij}^{st} is positive when $s = t$ and the distance d_{ij} is small, and zero otherwise. For example, when using a uniform kernel with bandwidth h, $R_{ij}^{st} > 0$ if $s = t$ and there is a k such that $d_{ki} \leq h$ and $d_{kj} \leq h$; R_{ij}^{st} is zero otherwise. If $\tau^2 > 0$, the presence of the random effect causes extra variation in the regression residuals with a covariance structure similar to $\mathbf{R}$: correlated random effects cause correlated residuals. Therefore, if $\tau^2 > 0$ observations which are close to each other tend to have correlated residuals.

The null hypothesis that the multinomial logistic regression model fits well can be phrased in terms of the above random effects model (4.7) as

$$H_0 : \ \tau^2 = 0,$$

which implies $\mathbf{z} = \mathbf{0}$, against the one-sided alternative

$$H_A : \ \tau^2 > 0.$$

We test H_0 with a score test. An advantage of score testing is that it only requires fitting the model under the null hypothesis, not under the alternative hypothesis. This is an important advantage for our H_A, because the random effects model (4.7) is difficult to fit. Furthermore, the score test is by definition

a one-sided test, so problems due to a null hypothesis on the boundary of the parameter space do not arise.

The score test statistic is the derivative of the loglikelihood $\ell(\tau^2)$ with respect to τ^2 at $\tau^2 = 0$. If nuisance parameters are present, as in this case the model parameters $\boldsymbol{\beta}$, the loglikelihood is replaced by the profile loglikelihood $\hat{\ell}(\tau^2) = \ell(\tau^2, \hat{\boldsymbol{\beta}}(\tau^2))$. We have

$$\frac{\partial \hat{\ell}}{\partial \tau^2} = \frac{\partial \ell}{\partial \tau^2} + \frac{\partial \ell}{\partial \boldsymbol{\beta}} \cdot \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \tau^2}.$$

As $\partial \ell / \partial \boldsymbol{\beta}$ is zero if $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, the score test statistic of the profile likelihood is simply the score test statistic of the ordinary likelihood with maximum likelihood estimates of the nuisance parameters under the null plugged in.

The loglikelihood of the general model (4.7) is given by

$$\ell(\tau^2) = \log \Big[\mathrm{E}_{\mathbf{z}} \Big\{ \exp \Big(\sum_{i=1}^{n} \sum_{s=1}^{g} y_{is} \log\{\nu_{is}(\mathbf{z})\} \Big) \Big\} \Big], \tag{4.8}$$

where $\nu_{is}(\mathbf{z}) = \mathrm{P}(Y_i = s \mid \mathbf{z})$ and $\mathrm{E}_{\mathbf{z}}$ denotes taking the expectation over $\mathbf{z}$. In Section 4.11 we calculate the derivative of this likelihood with respect to τ^2 at $\tau^2 = 0$, in the spirit of Le Cessie and van Houwelingen (1995), using a Taylor approximation of $\nu_{is}(\mathbf{z})$ with respect to $\mathbf{z}$ at $\mathbf{z} = \mathbf{0}$. This results in the score test statistic

$$T = \frac{\partial \hat{\ell}(0)}{\partial \tau^2} = \tfrac{1}{2}(\mathbf{y} - \hat{\boldsymbol{\mu}})' \mathbf{R} (\mathbf{y} - \hat{\boldsymbol{\mu}}) - \tfrac{1}{2} \mathrm{trace}(\mathbf{R}\hat{\mathbf{W}}). \tag{4.9}$$

We see that the score test statistic in this model is equivalent to the test statistic proposed in (4.2), as, ignoring the constants, T is simply Q minus the estimated expectation of Q.

This alternative construction of Q as a score test statistic gives interesting insights in the power properties of the test. A score test is a locally most powerful test (Cox and Hinkley, 1974) in the sense that it optimizes the slope of the power function at the test value of $\tau^2 = 0$. It is therefore the optimal test to use against the alternative model (4.7) when the value of τ^2 is small. These alternatives tend to have small, but non-zero values of the random effect $\mathbf{z}$. The goodness-of-fit test proposed in this paper is therefore the optimal test for detecting a small, but consistent deviation from the model.

The random effects model of this section is interesting in its own right as a general test for the existence of a random effect with a specified covariance structure $\mathbf{R}$, which may be any positive semi-definite matrix. This type of test has many applications outside the context of goodness-of-fit testing, for example in variance components analysis in genetics (Houwing-Duistermaat et al., 1995) and in high-dimensional data analysis in genomics (Goeman et al., 2004).

4.6 Connection to binary logistic regression

Here we show that for $g = 2$, when multinomial logistic regression becomes binary logistic regression, the test in this paper is exactly the same as the goodness-of-fit test of Le Cessie and van Houwelingen (1995), so that it is a generalization of that test.

Take $g = 2$. Call $R = UU'$, $W = W^{11}$, as defined in (4.4), and $H = WX(X'WX)^{-1}X'$. Call $\mathbf{y}_1 = (y_{11}, \ldots, y_{n1})'$, $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{n1})'$, using the notation of Section 4.2. Then the test statistic of Le Cessie and van Houwelingen (1995) is given by

$$Q_1 = (\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)'R(\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)$$

To show that this test statistic is equivalent to the test statistic in this paper for $g = 2$, remark that $\mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{f} \otimes (\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)$ where $\mathbf{f} = (1, -1)'$. Combining this with $\mathbf{R} = I_g \otimes R$, it follows that

$$Q = \mathbf{f}'\mathbf{f} \otimes (\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)'R(\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1) = 2Q_1.$$

The test statistics are therefore equivalent.

To show that also the approximations to the distribution of the test statistic are the same, we must show that also $\tilde{Q} = 2\tilde{Q}_1$, where

$$\tilde{Q}_1 = (\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)'(I - H)'R(I - H)(\mathbf{y}_1 - \hat{\boldsymbol{\mu}}_1)$$

This can be shown by remarking that $\mathbf{W} = F \otimes W$, where

$$F = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Writing $\mathbf{X} = I \otimes X$, remarking that F has generalized inverse $F^- = (1/4)F$ and expanding the Kronecker products, we get $\mathbf{H} = (1/2)F \otimes H$, from which $(I - \mathbf{H})(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{f} \otimes (I - H)(\mathbf{y}_1 - \boldsymbol{\mu}_1)$. Finally, combining this with $\mathbf{R} = I \otimes R$, the result $\tilde{Q} = 2\tilde{Q}_1$ follows. Therefore the two test statistics and the approximations to their distribution are completely equivalent in case of binary logistic regression.

4.7 Simulation results

To check the adequacy of the gamma approximation to the distribution of Q in a concrete case and to give an illustration of the power of the test, we conducted a small simulation experiment (compare Le Cessie and van Houwelingen, 1995, for the case $g = 2$).

We constructed a data set of 108 observations and three covariates x_1, x_2 and x_3, each taking values 1, 0 and -1. The 108 observations were taken as

TABLE 4.1: *Fraction rejected for the goodness-of-fit test of this paper, based on 10,000 simulated data sets under the null hypothesis ($t = 0$) and under alternatives with a quadratic effect ($t > 0$).*

alternative	*nominal test size* α				
	0.10	0.05	0.01	0.005	0.001
$t = 0$	0.125	0.061	0.014	0.007	0.002
$t = 1$	0.243	0.148	0.046	0.026	0.009
$t = 2$	0.618	0.487	0.259	0.189	0.088
$t = 3$	0.882	0.800	0.581	0.485	0.300
$t = 4$	0.979	0.954	0.844	0.781	0.606

four replicates from each of the 27 possible combinations of the three covariate values. We modelled the probabilities of three possible outcomes as in (4.1) with

$$\begin{aligned} \eta_1 &= 2x_1 + tx_1^2 \\ \eta_2 &= 2x_2 \\ \eta_3 &= 2x_3 \end{aligned}$$

By varying the value of t we can generate outcomes both from the null hypothesis that a multinomial logistic regression model in x_1, x_2 and x_3 fits well, and various alternative hypotheses.

We generated 10,000 multinomial outcome vectors Y from the model, taking $t = 0$, 1, 2, 3 and 4. For each realisation of Y we fitted a multinomial logistic regression model in x_1, x_2 and x_3 and calculated the goodness-of-fit test statistic Q, estimated its expectation and variance, and calculated the p-value using the gamma approximation. The smoothing matrix U was constructed using a uniform kernel with a bandwidth at the 25-th percentile of the distance distribution, which meant that each smoothed residual was the average of all residuals at most $\sqrt{2}$ distance away. The results are given in table 4.1, rounded to three decimal places.

Judging from this table, it seems that the gamma approximation to the distribution of the test statistic performs quite well, although it is slightly anti-conservative. The rejection rates for $t = 0$ are close to the nominal α level. It can also be concluded that the goodness-of fit test has good power for detecting deviations from the null hypothesis. It would be interesting to look at the effect of different choices of the bandwidth and to study different alternatives, but we lack space in this paper.

TABLE 4.2: *The p-values of the goodness-of-fit test for the liver enzyme data, with different choices of bandwidth (measured as percentiles of the distribution of distances between observations).*

model	*bandwidth (percentile)*						
	10	20	30	40	50	60	70
non-log-transformed	0.004	0.001	0.000	0.000	0.013	0.022	0.091
log-transformed	0.491	0.576	0.341	0.297	0.579	0.580	0.397

4.8 Application: liver enzyme data

We applied the goodness-of-fit test to a dataset of patients with liver disease (Albert and Harris, 1987). This data set has 218 patients in four disease categories: acute viral hepatitis (57 patients), persistent chronic hepatitis (44), aggressive chronic hepatitis (40) and post-necrotic cirrhosis (77). For each patient the concentrations of three liver enzymes was measured: aspartate aminotransferase (AST), alanine aminotransferase (ALT) and glutamate dehydrogenase (GLDH). All these variables had markedly skewed distributions. The data were analyzed with a multinomial logistic regression model by Albert and Harris (1987), but Lesaffre and Albert (1989) argued for a multinomial logistic regression model with log-transformed covariates.

We tested the fit of the model with AST, GLDH and ALT using the goodness-of-fit test of this paper and kernel smoothing using a uniform kernel with bandwidth equal to the 25-th percentile of the distribution of the distances between observations. We found $Q = 8.41$ with $\mathrm{E}Q = 2.78$ and $\mathrm{sd}(Q) = 1.27$. On a scaled chi-squared distribution with 9.52 degrees of freedom (gamma{4.76, 1.71}), this gave a p-value of 0.001, clearly indicating lack of model fit. Log-transforming the covariates before fitting the model gives a clearly non-significant p-value of 0.37.

To investigate the sensitivity of this result to the choice of the smoothing method, we calculated the p-value for different choices of the bandwidth parameters (table 4.2). Bandwidth values are given as percentiles of the distribution of distances between the observations. From table 4.2 it can be seen that the test is quite robust to the choice of bandwidth.

There are various ways of making use of the flexibility of the goodness-of-fit test of this paper for looking more closely into a more significant test result. One is to break down the omnibus test for all variables to see which variables are responsible for the lack of fit. This can be done by using subsets of the original covariates AST, ALT and GLDH for constructing the distance measure for use by the test, testing whether the relationship between that subset of the covari-

TABLE 4.3: *Results of goodness-of-fit test of the liver enzyme data in which the distance measure between observations depends on different subsets of the covariates. The table gives raw p-values and multiplicity adjusted p-values from a closed testing procedure.*

Distance based on	*p-value*	*adjusted p-value*
AST, ALT and GLDH	0.001	0.001
AST and GLDH	0.003	0.003
AST and ALT	0.001	0.001
GLDH and ALT	0.000	0.001
AST	0.000	0.003
GLDH	0.314	0.314
ALT	0.001	0.001

ates and the outcome has been adequately modelled. Taking all $2^3 - 1$ subsets, we can set up a closed testing procedure (Marcus et al., 1976) to control for multiple testing. In this procedure each subset of covariates is only tested when all its supersets are significant (for example the subset {ALT} is only tested when tests based on the subsets {AST, ALT}, {GLDH, ALT} and {GLDH, ALT, AST} are all significant). In that case all tests can be performed at level α, while still keeping the family-wise error rate at α (Marcus et al., 1976). The multiplicity adjusted p-values (Dudoit et al., 2003) for this procedure are the maximum of the p-values of the test itself and all supersets. These multiplicity-adjusted p-values are never smaller than the p-value for the test for global lack of fit. We performed these tests using kernel smoothing with a bandwidth at the 25-th percentile of the distance distribution as above. The raw and multiplicity adjusted p-values are given in table 4.3. The lack of fit is most clear in ALT and AST, while there is no evidence for lack of fit in GLDH. This is in line with the analysis of Lesaffre and Albert (1989), who concluded that there was no real need to log-transform GLDH.

Just as the test result can be split up in its component variables, it can be split into its component outcome categories. The test statistic can be written as

$$Q = \sum_{s=1}^{g} Q_s$$

where Q_s is the sum of the squared smoothed residuals $\tilde{r}_{1s}, \ldots, \tilde{r}_{ns}$, corresponding to outcome category s. We plotted the Q_s, $s = 1, \ldots, 4$ in figure 4.1, standardized to z-scores. From the plot we can see that the lack of fit is clear in the residuals of categories 1, 2 and 3 (acute viral, persistent chronic and aggressive chronic hepatitis), but that there is no clear evidence for lack of fit in the residuals of category 4 (post-necrotic cirrhosis).

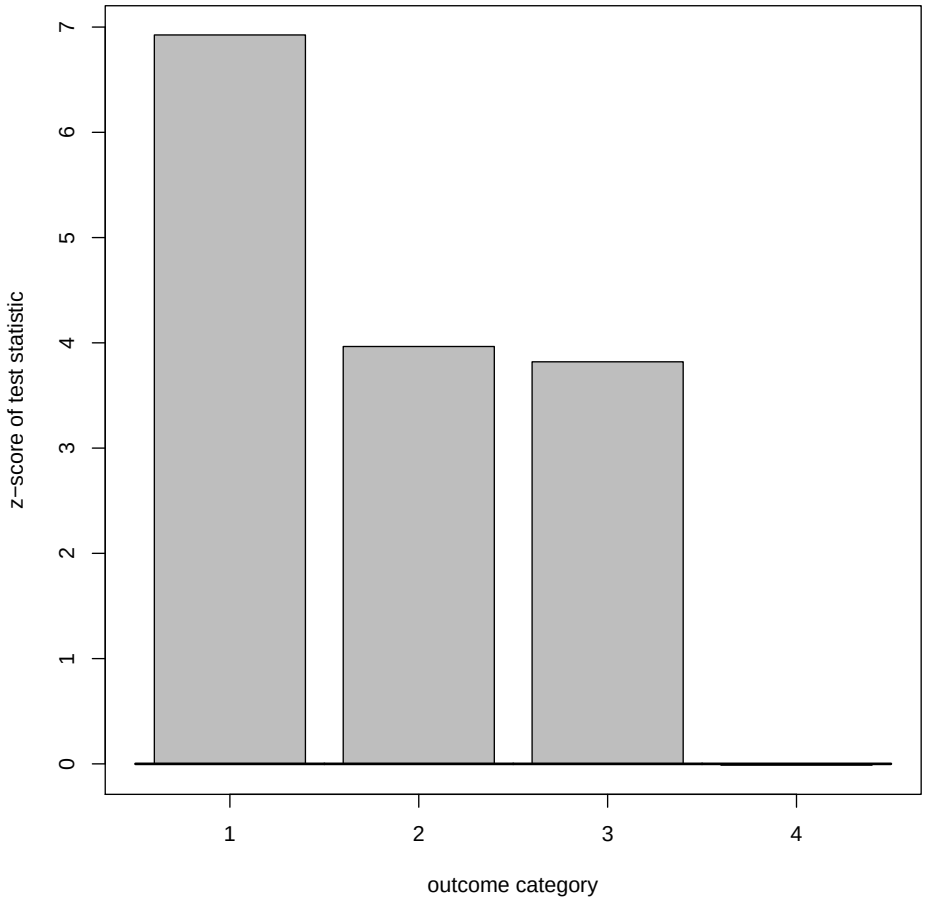

FIGURE 4.1: *Influence of the four outcome categories on the goodness-of-fit test result. Depicted are the sum of squared smoothed residuals Q_s of each outcome category s, standardized to z-scores. The total goodness-of-fit test statistic is the sum of the unstandardized Q_s-scores.*

4.9 Discussion

Formal goodness-of-fit testing is important in model-building of the multinomial logistic regression model, because the fitted model is very difficult to visualize. So far, however, only one goodness-of-fit test was available for this model (Pigeon and Heyse, 1999), which stands in the tradition of the goodness-of-fit test of Hosmer and Lemeshow (2000) for binary logistic regression. In this paper we have presented a very different goodness-of-fit test based on a sum of squared smoothed residuals, extending a test of Le Cessie and Van Houwelin-

gen (1991). It has power against consistent patterns of non-linearity: observations close to each other in covariate space which deviate in the same direction.

To illustrate the power properties of the test, we have constructed a random effects model for which the proposed test is optimal. Such a precise specification of the alternative hypothesis against which the test is optimal clarifies the type of lack of fit the test is directed against and therefore gives some insight into its power properties. Tools were also provided to look more closely into a significant test result.

Just like the test of Le Cessie and van Houwelingen (1995), the test proposed in this paper has potential applications outside the goodness-of-fit testing context, for example in genetics (Houwing-Duistermaat et al., 1995) and in high-dimensional data analysis (Goeman et al., 2004). This paper allows these applications to be generalized to the case of multinomial outcome variables.

4.10 Variance of the test statistic

We calculate the variance of $\tilde{Q}$ as given in (4.5). Write $\tilde{Q} = \sum_{s=1}^{g}\sum_{t=1}^{g} Q_{st}$, where $Q_{st} = \sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{R}_{ij}^{st}(y_{is}-\mu_{is})(y_{jt}-\mu_{jt})$ for $s,t = 1,\ldots g$. We will calculate the g^4 covariances of all Q_{st} terms and sum them to find the variance of $\tilde{Q}$.

Define $S_{ij}^{st} = (y_{is}-\mu_{is})(y_{jt}-\mu_{jt})$. Then $\mathrm{Cov}(S_{ij}^{st}, S_{kl}^{uv}) = 0$ unless $i=k$ and $j=l$ or $i=l$ and $j=k$, due to the independence of the samples under the null hypothesis. Therefore

$$\begin{aligned}\mathrm{Cov}(Q_{st}, Q_{uv}) &= \sum_i \tilde{R}_{ii}^{st}\tilde{R}_{ii}^{uv}\mathrm{Cov}(S_{ii}^{st}, S_{ii}^{uv}) + \sum_i\sum_{j\neq i}\tilde{R}_{ij}^{st}\tilde{R}_{ij}^{uv}\mathrm{Cov}(S_{ij}^{st}, S_{ij}^{uv}) \\ &\quad + \sum_i\sum_{j\neq i}\tilde{R}_{ij}^{st}\tilde{R}_{ji}^{uv}\mathrm{Cov}(S_{ij}^{st}, S_{ji}^{uv}).\end{aligned}$$

If $i \neq j$, $\mathrm{E}(S_{ij}^{st}) = 0$, for all s and t, so that

$$\begin{aligned}\mathrm{Cov}(S_{ij}^{st}, S_{ij}^{uv}) &= \mathrm{E}[(y_{is}-\mu_{is})(y_{iu}-\mu_{iu})]\cdot\mathrm{E}[(y_{jt}-\mu_{jt})(y_{jv}-\mu_{jv})] \\ &= W_{ii}^{su}W_{jj}^{tv},\end{aligned}$$

while if $i=j$,

$$\mathrm{Cov}(S_{ii}^{st}, S_{ii}^{uv}) = \mathrm{E}[S_{ii}^{st}S_{ii}^{uv}] - \mathrm{E}[S_{ii}^{st}]\mathrm{E}[S_{ii}^{uv}] = \mathrm{E}[S_{ii}^{st}S_{ii}^{uv}] - W_{ii}^{st}W_{ii}^{uv}.$$

Using these expressions,

$$\begin{aligned}\mathrm{Cov}(Q_{st}, Q_{uv}) &= \sum_{i=1}^{n}\tilde{R}_{ii}^{st}\tilde{R}_{ii}^{uv}\kappa_i^{stuv} + \sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{R}_{ij}^{st}\tilde{R}_{ji}^{vu}W_{ii}^{su}W_{jj}^{tv} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{R}_{ij}^{st}\tilde{R}_{ji}^{uv}W_{ii}^{sv}W_{jj}^{tu}\end{aligned}$$

where $\kappa_i^{stuv} = \mathrm{E}(S_{ii}^{st} S_{ii}^{uv}) - W_{ii}^{st} W_{ii}^{uv} - W_{ii}^{su} W_{ii}^{tv} - W_{ii}^{sv} W_{ii}^{tu}$. It is easy to check that the value of κ_i^{stuv} does not depend on the order of s, t, u and v. Calculation of the values of κ_i^{stuv} as given in (4.6) is straightforward but tedious.

Taking all covariances of the Q_{st} terms together, we have

$$\mathrm{Var}(\tilde{Q}) = \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{u=1}^{g} \sum_{v=1}^{g} \mathrm{Cov}(Q_{st}, Q_{uv}).$$

The result (4.5) follows by rewriting

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{R}_{ij}^{st} \tilde{R}_{ji}^{vu} W_{ii}^{su} W_{jj}^{tv} = \mathrm{trace}(\tilde{R}^{st} W^{tv} \tilde{R}^{vu} W^{su})$$

and

$$\sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{u=1}^{g} \sum_{v=1}^{g} \mathrm{trace}(\tilde{R}^{st} W^{tu} \tilde{R}^{uv} W^{sv}) = \mathrm{trace}(\tilde{\mathbf{R}}\mathbf{W}\tilde{\mathbf{R}}\mathbf{W}).$$

4.11 Derivation of the test statistic

We derive the expression (4.9) for the score test statistic from the random effects model (4.7). The likelihood $L(\tau^2) = \exp\{\ell(\tau^2)\}$ can be written

$$L(\tau^2) = \mathrm{E}_{\mathbf{z}}\big[\prod_{i=1}^{n} f_i(\mathbf{z})\big],$$

where $f_i(\mathbf{r}) = \exp\{l_i(\mathbf{z})\}$ and

$$l_i(\mathbf{z}) = \sum_{s=1}^{g} y_{is} \log\{\nu_{is}(\mathbf{z})\}.$$

Compare (4.8). Note that $f_i(\mathbf{z})$ only depends on $(z_{i1}, \ldots, z_{ig})$. Therefore, Taylor expanding $L(\tau^2)$ with respect to $\mathbf{z}$ at $\mathbf{z} = \mathbf{0}$ gives

$$\begin{aligned}
L(\tau^2) &= \mathrm{E}_{\mathbf{z}}\Bigg[\prod_{i=1}^{n} f_i(\mathbf{0}) + \sum_{s=1}^{g} \sum_{i=1}^{n} z_{is} \frac{\partial f_i(\mathbf{0})}{\partial z_{is}} \prod_{j \neq i} f_j(\mathbf{0}) \\
&\quad + \frac{1}{2} \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n} z_{is} z_{it} \frac{\partial^2 f_i(\mathbf{0})}{\partial z_{is} \partial z_{it}} \prod_{j \neq i} f_j(\mathbf{0}) \\
&\quad + \frac{1}{2} \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n} \sum_{j \neq i} z_{is} z_{jt} \frac{\partial f_i(\mathbf{0})}{\partial z_{is}} \frac{\partial f_j(\mathbf{0})}{\partial z_{jt}} \prod_{k \neq i,j} f_k(\mathbf{0}) + o(\mathbf{z}\mathbf{z}')\Bigg] \\
&= \prod_{i=1}^{n} f_i(\mathbf{0}) + \frac{1}{2}\tau^2 \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n} R_{ii}^{st} \frac{\partial^2 f_i(\mathbf{0})}{\partial z_{is} \partial z_{it}} \prod_{j \neq i} f_j(\mathbf{0}) \\
&\quad + \frac{1}{2}\tau^2 \sum_{s=1}^{g} \sum_{t=1}^{g} \sum_{i=1}^{n} \sum_{j \neq i} R_{ij}^{st} \frac{\partial f_i(\mathbf{0})}{\partial z_{is}} \frac{\partial f_j(\mathbf{0})}{\partial z_{jt}} \prod_{k \neq i,j} f_k(\mathbf{0}) + o(\tau^2).
\end{aligned}$$

Using $\frac{\partial f_i(\mathbf{z})}{\partial z_{is}} = f_i(\mathbf{z})\frac{\partial l_i(\mathbf{z})}{\partial z_{is}}$ and $\frac{\partial^2 f_i(\mathbf{z})}{\partial z_{is}\partial z_{it}} = f_i(\mathbf{z})\left[\frac{\partial^2 l_i(\mathbf{z})}{\partial z_{is}\partial z_{it}} + \frac{\partial l_i(\mathbf{z})}{\partial z_{is}}\frac{\partial l_i(\mathbf{z})}{\partial z_{it}}\right]$, this expression can be rewritten to

$$\begin{aligned} L(\tau^2) &= \prod_{i=1}^{n} f_i(\mathbf{0})\Bigg[1 + \frac{1}{2}\tau^2 \sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n} R_{ii}^{st}\frac{\partial l_i(\mathbf{0})}{\partial z_{is}\partial z_{it}} \\ &\quad + \frac{1}{2}\tau^2 \sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij}^{st}\frac{\partial l_i(\mathbf{0})}{\partial z_{is}}\frac{\partial l_j(\mathbf{0})}{\partial z_{jt}}\Bigg] + o(\tau^2) \end{aligned}$$

Because $\frac{\partial \ell(0)}{\partial \tau^2} = \frac{1}{L(0)}\frac{\partial L(0)}{\partial \tau^2}$, the score function at $\tau^2 = 0$ is

$$\frac{\partial \ell(0)}{\partial \tau^2} = \frac{1}{2}\left[\sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n} R_{ii}^{st}\frac{\partial^2 l_i(\mathbf{0})}{\partial z_{is}\partial z_{it}} + \sum_{s=1}^{g}\sum_{t=1}^{g}\sum_{i=1}^{n}\sum_{j=1}^{n} R_{ij}^{st}\frac{\partial l_i(\mathbf{0})}{\partial z_{is}}\frac{\partial l_j(\mathbf{0})}{\partial z_{jt}}\right]$$

The result (4.9) follows from $\frac{\partial l_i(\mathbf{0})}{\partial z_{is}} = y_{is} - \mu_{is}$ and $\frac{\partial^2 l_i(\mathbf{0})}{\partial z_{is} z_{it}} = -W_{it}^{st}$.