Journal of the Korean Statistical Society (2004), 33: 3, pp 313–321

GOODNESS-OF-FIT TEST USING LOCAL MAXIMUM LIKELIHOOD POLYNOMIAL ESTIMATOR FOR SPARSE MULTINOMIAL DATA†

JANGSUN BAEK[1]

ABSTRACT

We consider the problem of testing cell probabilities in sparse multinomial data. Aerts *et al.* (2000) presented $T = \sum_{i=1}^{k}\{p_i^* - E(p_i^*)\}^2$ as a test statistic with the local least square polynomial estimator p_i^*, and derived its asymptotic distribution. The local least square estimator may produce negative estimates for cell probabilities. The local maximum likelihood polynomial estimator $\hat{p}_i$, however, guarantees positive estimates for cell probabilities and has the same asymptotic performance as the local least square estimator (Baek and Park, 2003). When there are cell probabilities with relatively much different sizes, the same contribution of the difference between the estimator and the hypothetical probability at each cell in their test statistic would not be proper to measure the total goodness-of-fit. We consider a Pearson type of goodness-of-fit test statistic, $T_1 = \sum_{i=1}^{k}\{\hat{p}_i - E(\hat{p}_i)\}^2/p_i$ instead, and show it follows an asymptotic normal distribution. Also we investigate the asymptotic normality of $T_2 = \sum_{i=1}^{k}\{\hat{p}_i - E(\hat{p}_i)\}^2$ where the minimum expected cell frequency is very small.

AMS 2000 subject classifications. Primary 62G10; Secondary 62G05.

Keywords. Goodness of fit, local maximum likelihood, local polynomial estimator, sparse multinomial data.

1. INTRODUCTION

Suppose we observe the cell frequency N_i from the multinomial distribution with the cell probability $p_i, i = 1, \ldots, k$. Then $\sum_{i=1}^{n} N_i = n$ is the total number of observations. When the total number of observations n is relatively small

Recieved December 2003; accepted May 2004.

†This work was supported by grant No. R05-2001-000-00065-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

[1]Department of Statistics, Chonnam National University, Gwangju 500-757, Korea (e-mail : jbaek@chonnam.ac.kr)

comparing to the number of cell k, the multinomial data is called to be sparse. In order to estimate $\mathbf{p} = (p_1, \ldots, p_k)^T$, we may consider the frequency estimator $\bar{\mathbf{p}} = (\bar{p}_1, \ldots, \bar{p}_k)^T$, $\bar{p}_i = N_i/n$, $i = 1, \ldots, k$. The frequency estimator is not consistent under sparse asymptotics where n/k remains small constant as both n and k become infinite. For an ordinal categorical variable it has been proposed to smooth the roughness of the frequency estimators away by borrowing information from neighboring cells to estimate the cell probabilities. Simonoff (1983) considered an estimator based on a maximum penalized likelihood criterion. Burman (1987a), Hall and Titterington (1987) proposed kernel estimators. Aerts *et al.* (1997), Baek (1998) studied the properties of the local polynomial estimator based on the local least square criterion. A drawback to the local least square estimator is that the probability estimate can be negative. Baek and Park (2003) investigated the asymptotic properties of the local maximum likelihood polynomial estimator, which guarantees the positive probability estimates.

If we are interested in testing $H_0 : \mathbf{p} = \mathbf{p}_0$, where $\mathbf{p}_0 = (p_{10}, \ldots, p_{k0})^T$, we usually use the Pearson chi-square statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_{i0})^2}{np_{i0}},$$

or the likelihood ratio statistic

$$G^2 = 2\sum_{i=1}^{k} N_i \log\left(\frac{N_i}{np_{i0}}\right).$$

Under H_0, these statistics are asymptotically distributed as χ^2 with $(k-1)$ degrees of freedom if $\min_i\{np_i\} \to \infty$ as $n \to \infty$. For sparse multinomial data, however, p_i gets small since $k \to \infty$ as $n \to \infty$. Therefore the condition $\min_i np_i \to \infty$ of the $\chi^2_{(k-1)}$ approximation can not be satisfied for sparse multinomial data. Simonoff (1985) proposed a test statistic of the standardized form with his maximum penalized likelihood estimators, and obtained the critical value by the simulation of the statistic under the null hypothesis. Aerts *et al.* (2000) suggested $T = \sum_{i=1}^{k}\{p_i^* - E(p_i^*)\}^2$ with their local least square polynomial estimator p_i^*, and obtained asymptotic normality of the test statistic under the null hypothesis.

When there are cell probabilities with relatively much different sizes, the same contribution of the difference between the estimator and the hypothetical probability at each cell in their test statistic would not be proper to measure the total goodness-of-fit. Let $\hat{p}_i$ be the local maximum likelihood polynomial

estimator. We consider a Pearson type of goodness-of-fit test statistic, $T_1 = \sum_{i=1}^{k}\{\hat{p}_i - E(\hat{p}_i)\}^2/p_i$, and show it follows an asymptotic normal distribution. Also we investigate the asymptotic normality of $T_2 = \sum_{i=1}^{k}\{\hat{p}_i - E(\hat{p}_i)\}^2$ when the minimum expected cell frequency is very small. Section 2 contains the main results. There we define the local maximum likelihood polynomial estimator of sparse multinomial probability, and derive the limiting distributions of the test statistics, T_1 and T_2.

2. ASYMPTOTIC DISTRIBUTION OF THE TEST STATISTICS

We assume that the p_i's are generated by a latent density function $f(\cdot)$ on $[0,1]$ through the relation

$$p_i = \int_{(i-1)/k}^{i/k} f(u)du, \ i = 1, \ldots, k.$$

When $\sum_{i=1}^{k} N_i = n$ is fixed, the multinomial observed frequency vector $(N_1, \ldots, N_k)^T$ can be viewed as a set of independent Poisson random variables (Kendall and Stuart, 1979, p. 449). Namely, N_i's are independent Poisson random variables, conditional on $\sum_{i=1}^{k} N_i = n$, with the mean $\mu(N_i|X_i) = np_i$, respectively, if we let $X_i = (i - 1/2)/k$. The log-likelihood of k independent Poisson random variables is then

$$l = \sum_{i=1}^{k}\{N_i \log(np_i) - np_i\},$$

where $\sum_{i=1}^{k} N_i = n$ (ignoring constants) (Simonoff, 1996, p. 240).

In order to estimate the probability p at the cell with its center $X = x$, we first get the estimator of $\log(np)$ by local fitting, and then apply the inverse of the link function. This involves centering the data about x and weighting the conditional log-likelihood with $K_h(X_i - x)$, where K is the usual kernel function and h is the bandwidth. The local maximum likelihood polynomial estimator of $\log(np)$ is then given by $\hat{\beta}_0$, where $(\hat{\beta}_0, \ldots, \hat{\beta}_t)^T$ maximizes

$$\sum_{i=1}^{k}\left[N_i\left\{\beta_0 + \cdots + \beta_t(X_i - x)^t\right\} - \exp\left\{\beta_0 + \cdots + \beta_t(X_i - x)^t\right\}\right]K_h(X_i - x).$$

The probability p can then be estimated by applying the inverse function to give

$$\hat{p}\,(x; t, h) = \frac{1}{n}\exp(\hat{\beta}_0).$$

This guarantees that the estimate will be nonnegative. Throughout the paper we assume the following regularity conditions.

(C1) $K(\cdot)$ is a symmetric, continuous kernel with bounded support $[-1, 1]$.

(C2) $f^{(t+1)}(\cdot)$ is continuous on $[0, 1]$.

Baek and Park (2003) showed that the asymptotic bias of $\hat{p}\,(x; t, h)$ equals

$$\int z^{t+1} K_t(z) dz \left\{ \frac{S_{(t+1)}(f, \ldots, f^{(t+1)})}{(t+1)!k} \right\} h^{t+1}$$

where $S_{(t+1)}$ is a function of $f, \ldots, f^{(t+1)}$, the asymptotic variance is

$$(n^2 kh)^{-1} np_i \int K_t(z)^2 dz,$$

and the estimator attains the optimal rate of convergence. For the definition of K_t, let

$$\nu_l = \int z^l K(z) dz,$$

and let $\mathbf{N}_t$ be the $(t+1) \times (t+1)$ matrix having (i, j) entry equal to ν_{i+j-2}, and $\mathbf{M}_t(z)$ be the same as $\mathbf{N}_t$ but with the first column replaced by $(1, z, \ldots, z^t)^T$. For $|\mathbf{N}_t| \neq 0$, $K_t(z)$ is defined by

$$K_t(z) = \frac{|\mathbf{M}_t(z)|}{|\mathbf{N}_t|} K(z).$$

We will use the result of the asymptotic variance of the local maximum likelihood estimator to get the limiting distribution of the test statistics T_1 and T_2. In the case where the expected cell frequency is bounded below, *i.e.* $\min_i\{np_i\} > 0$, T_1 would be preferable since the relative difference between the probability estimate and the hypothetical probability is more proper measure of goodness-of-fit. When the expected cell frequency gets smaller as the sample size increases, we cannot calculate the relative difference and T_2 could be used.

Burman (1987b) derived asymptotic normality for quadratic forms in $N_i - np_i$, which include the Pearson statistic, $\chi^2 = \sum(N_i - np_i)^2/(np_i)$ and the test statistic like $\sum(N_i - np_i)^2$. The results of the limiting distributions of Pearson χ^2 statistic and $\sum(N_i - np_i)^2$ are stated as Lemma 2.1 and Lemma 2.2, respectively.

LEMMA 2.1 (Burman, 1987b, Corollary 3.3 (a) (i)). *Let* $\min_i\{np_i\} = \epsilon_{1n} > 0$ *for all* n. *Then as* $n \to \infty$ *and* $k \to \infty$,

$$\frac{\chi^2 - (k-1)}{\sigma_{1n}} \xrightarrow{d} N(0,1),$$

if $k\epsilon_{1n}^3 \to \infty$, *where* $\sigma_{1n}^2 = 2k + n^{-1}\sum p_i^{-1}$.

LEMMA 2.2 (Burman, 1987b, Corollary 3.4). *Let* $\max_i\{p_i\} = \epsilon_{2n} \to 0$. *Then as* $n \to \infty$ *and* $k \to \infty$,

$$\frac{\sum(N_i - np_i)^2 - n(1 - \sum p_i^2)}{\sigma_{2n}} \xrightarrow{d} N(0,1),$$

if $n\epsilon_{2n}^2 \to 0$, *where* $\sigma_{2n}^2 = 2n^2\sum p_i^2 + n$.

Now we derive the asymptotic distribution of T_1 when $\min_i\{np_i\} > 0$ is satisfied.

THEOREM 2.1. *Assume (C1), (C2), and* $h \to 0$, $nh^3 \to \infty$ *as* $n \to \infty$. *Let* $\min_i\{np_i\} = \epsilon_{1n} > 0$ *for all* n, *and let* $n_* = \min(n, nkh)$. *If* $k\epsilon_{1n}^3 \to \infty$ *and* $\sum_{i=1}^k p_i^{-1}/(n_*\sigma_{1n}) \to 0$ *as* $n_* \to \infty$, *where* $\sigma_{1n}^2 = 2k + n^{-1}\sum_{i=1}^k p_i{}^{-1}$, *then*

$$\left(\frac{n}{\sigma_{1n}}\right)\left[\sum_{i=1}^k \frac{\{\hat{p}_i - E(\hat{p}_i)\}^2}{p_i} - \frac{(k-1)}{n}\right] \xrightarrow{d} N(0,1).$$

PROOF. First we decompose $\sqrt{n}(\hat{p}_i - E(\hat{p}_i))$ into three parts as follows;

$$\sqrt{n}(\hat{p}_i - E(\hat{p}_i)) = \sqrt{n}(\bar{p}_i - p_i) + \sqrt{n}(\hat{p}_i - \bar{p}_i) + \sqrt{n}(p_i - E(\hat{p}_i)) \tag{2.1}$$

We examine the limiting behavior of the second part $\sqrt{n}(\hat{p}_i - \bar{p}_i)$. Since Theorem 1 of Baek and Park (2003) shows

$$\mathrm{Var}(\hat{p}_i) \sim (n^2kh)^{-1} np_i \int K_t(z)^2 dz,$$

it follows that

$$\mathrm{Var}\left(\sqrt{n}\hat{p}_i\right) \sim (nkh)^{-1} np_i \int K_t(z)^2 dz,$$

which equals $O((nkh)^{-1})$ because $\min_i\{np_i\} = \epsilon_{1n} > 0$. Also we know

$$E\left(\sqrt{n}\bar{p}_i\right) = \sqrt{n}p_i\ ,\ \mathrm{Var}\left(\sqrt{n}\bar{p}_i\right) = \frac{np_i(1-p_i)}{n},$$

which is $O(n^{-1/2})$. So it is easy to see that

$$\sqrt{n}\bar{p}_i = \sqrt{n}p_i + O_p\left(n^{-1/2}\right),$$
$$\sqrt{n}\hat{p}_i = \sqrt{n}E(\hat{p}_i) + O_p\left((nkh)^{-1/2}\right).$$

Therefore

$$\sqrt{n}\left(\hat{p}_i - \bar{p}_i\right) = \sqrt{n}\left(E(\hat{p}_i) - p_i\right) + O_p\left({n_*}^{-1/2}\right),$$

where $n_* = \min(n, nkh)$. Plugging the last result into the equation (2.1) leads to

$$\sqrt{n}(\hat{p}_i - E(\hat{p}_i)) = \sqrt{n}(\bar{p}_i - p_i) + O_p\left({n_*}^{-1/2}\right),\ i = 1, \ldots, k. \tag{2.2}$$

Let

$$\mathbf{W}_n = \sqrt{n}(\hat{p}_1 - E(\hat{p}_1), \ldots, \hat{p}_k - E(\hat{p}_k))^T,$$
$$\mathbf{X}_n = \sqrt{n}(\bar{p}_1 - p_1, \ldots, \bar{p}_k - p_k)^T,$$
$$\mathbf{O}_p\left({n_*}^{-1/2}\right) = \left({n_*}^{-1/2}\right)(O_p(1), \ldots, O_p(1))^T.$$

Then the equation (2.2) can be expressed with the vectors $\mathbf{W}_n$, $\mathbf{X}_n$, $\mathbf{O}_p({n_*}^{-1/2})$ as $\mathbf{W}_n = \mathbf{X}_n + \mathbf{O}_p({n_*}^{-1/2})$. Let $\mathbf{C} = \mathrm{diag}(p_1^{-1}, \ldots, p_k^{-1})$. Then

$$\mathbf{W}_n^T\mathbf{C}\mathbf{W}_n = \mathbf{X}_n^T\mathbf{C}\mathbf{X}_n + 2\mathbf{O}_p({n_*}^{-1/2})^T\mathbf{C}\mathbf{X}_n + \mathbf{O}_p({n_*}^{-1/2})^T\mathbf{C}\mathbf{O}_p({n_*}^{-1/2}).$$

$\mathbf{C}\mathbf{X}_n$ in the second part $\mathbf{O}_p({n_*}^{-1/2})^T\mathbf{C}\mathbf{X}_n$ in the right hand side of the above equation is

$$\left(\sqrt{n}\,\frac{(\bar{p}_1 - p_1)}{p_1}, \ldots, \sqrt{n}\,\frac{(\bar{p}_k - p_k)}{p_k}\right)^T,$$

and this becomes $O_p({n_*}^{-1/2})(p_1^{-1}, \ldots, p_k^{-1})^T$ since

$$\sqrt{n}(\bar{p}_i - p_i) = O_p\left({n_*}^{-1/2}\right),\ i = 1, \ldots, k.$$

So the second part becomes $O_p({n_*}^{-1})\sum_{i=1}^k p_i^{-1}$. The third part $\mathbf{O}_p({n_*}^{-1/2})^T\mathbf{C}\times$ $\mathbf{O}_p({n_*}^{-1/2})$ is obtained similarly to be $O_p({n_*}^{-1})\sum_{i=1}^k p_i^{-1}$, which is the same as the second part asymptotically. Thus we can rewrite $\mathbf{W}_n^T\mathbf{C}\mathbf{W}_n$ as

$$\mathbf{W}_n^T\mathbf{C}\mathbf{W}_n = \mathbf{X}_n^T\mathbf{C}\mathbf{X}_n + O_p({n_*}^{-1})\sum_{i=1}^k {p_i}^{-1}. \tag{2.3}$$

Subtracting $(k-1)$ and dividing σ_{1n} from the both sides of the equation (2.3) gives us that

$$\frac{\mathbf{W}_n^T\mathbf{C}\mathbf{W}_n-(k-1)}{\sigma_{1n}}=\frac{\mathbf{X}_n^T\mathbf{C}\mathbf{X}_n-(k-1)}{\sigma_{1n}}+\frac{O_p(n_*^{-1})\sum_{i=1}^k p_i^{-1}}{\sigma_{1n}}.$$

From the condition of the theorem,

$$\sum_{i=1}^k \frac{p_i^{-1}}{n_*\sigma_{1n}}\to 0 \quad \text{as} \quad n_*\to\infty,$$

the third part of the last equation is $o_p(1)$. As $\{\mathbf{X}_n^T\mathbf{C}\mathbf{X}_n-(k-1)\}/\sigma_{1n}$ converges in distribution to $N(0,1)$ by Lemma 2.1, the result is immediate by Slutsky theorem. □

When $\min_i\{np_i\}$ is very small or if np_i's are not too different from one another, $T_2=\sum_{i=1}^k\{\hat{p}_i-E(\hat{p}_i)\}^2$ could be used instead of Pearson chi-square type of statistic. The limiting distribution of T_2 is obtained in the following Theorem 2.2.

THEOREM 2.2. *Assume (C1), (C2), and* $h\to 0$, $nh^3\to\infty$ *as* $n\to\infty$. *Let* $\max_i\{p_i\}=\epsilon_{2n}\to 0$, *and let* $k_*=\min(k,k^2h)$. *If* $n\epsilon_{2n}^2\to 0$ *and* $k/(k_*\sigma_{2n})\to 0$ *as* $k_*\to\infty$, *where* $\sigma_{2n}^2=2n^2\sum p_i^2+n$, *then*

$$\left(\frac{n}{\sigma_{2n}}\right)\left[\sum_{i=1}^k\{\hat{p}_i-E(\hat{p}_i)\}^2-\left(1-\sum p_i^2\right)\right]\xrightarrow{d} N(0,1).$$

PROOF. Note that $n\sum_{i=1}^k\{\hat{p}_i-E(\hat{p}_i)\}^2=\sum_{i=1}^k\{\sqrt{n}(\hat{p}_i-E(\hat{p}_i))\}^2$, and we will investigate the limiting behavior of $\sqrt{n}(\hat{p}_i-E(\hat{p}_i))$ first as in the proof of Theorem 2.1. $\mathrm{Var}\,(\sqrt{n}\hat{p}_i)$ is $(nkh)^{-1}np_i\int K_t(z)^2dz$ asymptotically, but the condition $\min_i\{np_i\}>0$ for all n is not guaranteed anymore. We can approximate np_i with $nf(x_i)/k$ since $p_i\sim f(x_i)/k+O(k^{-3})$, and

$$\mathrm{Var}\left(\sqrt{n}\hat{p}_i\right)\sim(k^2h)^{-1}f(x_i)\int K_t(z)^2dz=O\left((k^2h)^{-1}\right).$$

We can approximate $\mathrm{Var}\,(\sqrt{n}\bar{p}_i)$ similarly as

$$\mathrm{Var}\left(\sqrt{n}\bar{p}_i\right)\sim\frac{f(x_i)}{k}\left(1-\frac{f(x_i)}{k}\right)=O\left(k^{-1}\right).$$

Now let $k_* = \min(k, k^2h)$, then

$$\sqrt{n}(\hat{p}_i - \bar{p}_i) = \sqrt{n}(E(\hat{p}_i) - p_i) + O_p\left({k_*}^{-1/2}\right)$$

because

$$\begin{aligned}\sqrt{n}\hat{p}_i &= \sqrt{n}E(\hat{p}_i) + O_p\left((k^2h)^{-1/2}\right),\\ \sqrt{n}\bar{p}_i &= \sqrt{n}p_i + O_p\left(k^{-1/2}\right).\end{aligned}$$

Hence

$$\sqrt{n}(\hat{p}_i - E(\hat{p}_i)) = \sqrt{n}(\bar{p}_i - p_i) + O_p({k_*}^{-1/2}),\ i = 1, \ldots, k. \tag{2.4}$$

The equation (2.4) can be expressed with the vectors $\mathbf{W}_n, \mathbf{X}_n, \mathbf{O}_p({k_*}^{-1/2})$ defined in the proof of Theorem 2.1, as $\mathbf{W}_n = \mathbf{X}_n + \mathbf{O}_p({k_*}^{-1/2})$. Then $n\sum_{i=1}^{k}\{\hat{p}_i - E(\hat{p}_i)\}^2$ is $\mathbf{W}_n^T\mathbf{W}_n$, and

$$\mathbf{W}_n^T\mathbf{W}_n = \mathbf{X}_n^T\mathbf{X}_n + 2\mathbf{O}_p\left({k_*}^{-1/2}\right)^T\mathbf{X}_n + \mathbf{O}_p\left({k_*}^{-1/2}\right)^T\mathbf{O}_p\left({k_*}^{-1/2}\right).$$

It is easy to see that $\mathbf{O}_p({k_*}^{-1/2})^T\mathbf{X}_n = k\,O_p(k_*^{-1})$ because

$$\sqrt{n}(\bar{p}_i - p_i) = O_p({k_*}^{-1/2}),$$

and $\mathbf{O}_p({k_*}^{-1/2})^T\mathbf{O}_p({k_*}^{-1/2})$ is also $k\,O_p(k_*^{-1})$. Therefore

$$\mathbf{W}_n^T\mathbf{W}_n = \mathbf{X}_n^T\mathbf{X}_n + kO_p(k_*^{-1}). \tag{2.5}$$

By subtracting $n(1 - \sum p_i^2)$ and dividing σ_{2n} from the both sides of the equation (2.5), we get

$$\frac{\mathbf{W}_n^T\mathbf{W}_n - n(1 - \sum p_i^2)}{\sigma_{2n}} = \frac{\mathbf{X}_n^T\mathbf{X}_n - n(1 - \sum p_i^2)}{\sigma_{2n}} + \frac{kO_p(k_*^{-1})}{\sigma_{2n}}.$$

The third part of the last equation is $o_p(1)$ under the condition of $k/(k_*\sigma_{2n}) \to 0$ as $k_* \to \infty$. Since

$$\frac{\mathbf{X}_n^T\mathbf{X}_n - n(1 - \sum p_i^2)}{\sigma_{2n}} \xrightarrow{d} N(0,1)$$

by Lemma 2.2, we get the result by Slutsky theorem. □

ACKNOWLEDGEMENTS

The author is grateful for the helpful comments of two referees. Especially the conditions of the main theorems are refined by one of the referees.

REFERENCES

AERTS, M. A., AUGUSTYNS, I. AND JANSSEN P. (1997). "Smoothing sparse multinomial data using local polynomial fitting", *Nonparameteric Statistics*, **8**, 127–147.

AERTS, M. A., AUGUSTYNS, I. AND JANSSEN P. (2000). "Central limit theorem for the total squared error of local polynomial estimators of cell probabilities", *Journal of Statistical Planning and Inference*, **91**, 181–193.

BAEK, J. (1998). "A local linear kernel estimator for sparse multinomial data", *Journal of the Korean Statistical Society*, **27**, 515–529.

BAEK, J. AND PARK, J. (2003). "On the asymptotic properties of a local maximum likelihood polynomial estimator for sparse multinomial probabilities", submitted.

BURMAN, P. (1987a). "Smoothing sparse contingency tables", *Sankhyā*, **A49**, 24–36.

BURMAN, P. (1987b). "Central limit theorem for quadratic forms for sparse tables", *Jonrnal of Multivariate Analysis*, **22**, 258–277.

HALL, P. AND TITTERINGTON, D. M. (1987). "On smoothing sparse multinomial data", *The Australian Journal of Statistics*, **29**, 19–37.

KENDALL, M. AND STUART, A. (1979). *The Advanced Theory of Statistics 2*, 4th ed., Charles Griffin & Company, London.

SIMONOFF, J. S. (1983). "A penalty function approach to smoothing large sparse contingency tables", *The Annals of Statistics*, **11**, 208–218.

SIMONOFF, J. S. (1985). "An improved goodness-of-fit statistic for sparse multinomials", *Journal of the American Statistical Association*, **80**, 671–677.

SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*, Springer, New York.