Discussion Paper No. 586

MONOTONICITY OF CHI-SQUARE TEST STATISTICS

Keunkwan Ryu

June 2003

The Institute of Social and Economic Research
Osaka University
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Monotonicity of Chi-square Test Statistics

by

Keunkwan Ryu*
Department of Economics
Seoul National University
and
ISER, Osaka University

Abstract

This paper establishes monotonicity of the chi-square test statistic. As the more efficient parameter estimator is plugged into the test statistic, the degrees of freedom of the resulting chi-square test statistic monotonically increase.

Keywords: cell, chi-square test, Fisher information, monotonicity, parametric distribution

JEL Classification: C12, C52

* Keunkwan Ryu is Associate Professor, Department of Economics, Seoul National University, Korea, and Visiting Foreign Scholar, ISER, Osaka University, Japan. His mailing address is Department of Economics, Seoul National University, Seoul 151-742, Korea, phone number is (O) 82-2-880-6397, fax number is (Fax) 82-2-886-4231, and email address is (Email) ryu@snu.ac.kr.

1. Introduction

To test whether a random variable follows a specific parametric distribution, chi-square specification tests have been widely used since Pearson's (1900) pioneering work. For recent econometric extensions, see Andrews (1988a,b). The main idea of the chi-square test is to measure the "distance" between the empirical cell frequencies and their model-based counterparts. If a parametric model is true, then the distance measure will be small; otherwise, it will be large. By suitably defining the distance measure, one can construct a test statistic which asymptotically follows a chi-square distribution with known degrees of freedom. To define the test statistic, one has to estimate the unknown parameters.

This paper proves that the chi-square test statistic monotonically increases in a stochastic sense as the parameter estimator being plugged into the test statistic becomes more and more efficient. When the cell structure used for the parameter estimation is the same as the cell structure employed to design the test statistic, it is well known that the resulting test statistic will asymptotically follow a chi-square distribution with $k-1-r$ degrees of freedom, where k is the number of cells and r is the number of estimated parameters. When the cell structures for estimation and testing are the same, the resulting test statistic loses as many degrees of freedom as there are estimated parameters in its asymptotic distribution. When the cell structure is finer in the case of estimation than in the case of testing, the degrees of freedom loss is not as severe as in the same structure case. In fact, there is a monotonic relationship between the fineness in the estimation cell structure (equivalently, the efficiency of the estimator) and the degrees of freedom loss. As a finer cell structure is used to estimate the parameter and thus a more efficient estimator is plugged into, the lesser will be the loss in the degrees of freedom of the resulting chi-square test statistic.

Chernoff and Lehmann (1954) showed that the test statistic using the estimator obtained from the continuous data is stochastically larger than the one using the estimator from the test cell structure, and that the degrees of freedom of the resulting chi-square test statistic are larger for the continuous data case than for the test cell case. The monotonicity result in this paper extends Chernoff and Lehmann's result to bridge the gap between these two extreme cases.

The rest of the paper is organized as follows. Section 2 introduces the framework. Section 3 contains the main results. Section 4 concludes the paper. Proofs are deferred to the Appendix.

2. Framework

Let X follow a parametric distribution: $X \sim f(x, \theta)$, $x \in \chi$. Let $C_1, \cdots, C_k$ be k cells which partition the support χ of x. Define $p_1, \cdots, p_k$ to be the corresponding cell probabilities, $p_j = \int_{C_j} f(x, \theta)\, dx$. Suppose we have n independent observations, $x_1, \cdots, x_n$, drawn from $f(x, \theta)$. Define an indicator variables d_{ij} such that $d_{ij} = 1$ if $x_i \in C_j$, $d_{ij} = 0$ otherwise. Let $n_1, \cdots, n_k$ denote the number of observations falling into $C_1, \cdots, C_k$, respectively, $n_j = \sum_{i=1}^{n} d_{ij}$. Of course, $\sum_{j=1}^{n} n_j = n$. Empirical cell probabilities, $\hat{p}_1, \cdots, \hat{p}_k$, are obtained as relative frequencies, $\hat{p}_j = n_j/n$.

Let P be the $k \times 1$ column vector composed of $p_1, \cdots, p_k$. Define $\hat{P}$ in a similar manner. Let us make some regularity assumptions:

Assumption 1: The true parameter θ is an interior point of the parameter space.
Assumption 2: The density $f(x, \theta)$ is positive for almost all $x \in \chi$.
Assumption 3: The Lebesgue measure of each cell C_j is positive.
Assumption 4: Each cell probability $p_j(\theta)$ has continuous first-order partial derivatives in a neighborhood of the true parameter θ.
Assumption 5: The Jacobian matrix $\partial P/\partial \theta'$ has full column rank at true θ.

These conditions ensure that $p_j(\theta)$ is positive, locally smooth, and one-to-one at the true parameter θ, and that a Taylor series expansion exists in a neighborhood around the true parameter θ. Assumption 5 implies that the number, say r, of parameters in θ cannot exceed $k-1$, the number of maximum free cells. This is because the row sum of the Jacobian matrix $\partial P/\partial \theta'$ is a zero vector, which implies that the number of independent rows is at most $k-1$.

In the following section, we construct chi-square test statistics to test $\mathrm{H}_0 : X \sim f(x, \theta)$. Depending on how the parameter θ is estimated, the resulting chi-square test statistic will have different degrees of freedom. We consider two different estimates of θ, one based on the test cell structure and the other based on a finer cell structure.

3. Estimation of θ and construction of χ^2 test statistics

We first have to estimate θ, and then based on these estimates, $\hat{\theta}$, compute $p_j(\hat{\theta})$, estimates of the model-implied cell probabilities. By comparing the empirical cell probabilities $\hat{p}_j$ with the model-based estimates $p_j(\hat{\theta})$, we can design a model specification test. Since we have $k-1$ maximum free cells, the number r of free parameters in θ should be less than or equal to $k-1$. In fact, the model $\mathrm{H}_0 : X \sim f(x, \theta)$ contains $\max(k-1-r, 0)$ restrictions relative to $k-1$ free cells. The case of known θ corresponds to $r = 0$, yielding $k-1$ restrictions.

If $r > k-1$, then the model is over-parameterized (under-identified) relative to $k-1$ free cells, which is excluded by Assumption 5.

The test statistic will be based on $\sqrt{n}(\hat{P} - P(\hat{\theta}))$. By using a Taylor expansion, we have

$$\begin{aligned}\sqrt{n}(\hat{P} - P(\hat{\theta})) &\approx \sqrt{n}(\hat{P} - P(\theta)) - J\sqrt{n}(\hat{\theta} - \theta) \\ &\approx \sqrt{n}(\hat{P} - P(\theta)) - JH^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(d_i - P(\theta) - JH^{-1}s_i),\end{aligned} \tag{1}$$

where $J = \partial P/\partial\theta'$, H is the information matrix of a single cell observation, s_i is the ith individual score function, $d_i = (d_{i1}, \cdots, d_{ik})'$, and $\approx$ indicates that both sides of the $\approx$ have the same asymptotic distribution.

Given a data set, one often aggregates the data into a finite set of intervals for the purpose of testing. Of course, if the original data is given as an interval or a categorical data, one does not have to aggregate the data for the testing purpose. Therefore, it is natural to assume that the test cell structure is coarser than (aggregation case) or at best the same as the estimation cell structure (no aggregation).

Alternatively speaking, to compute $\hat{\theta}$, one may use either a frequency data from the test cell structure or a frequency data from a finer cell structure. In section 3, we consider two cases separately, one case where the estimation cell structure and the test cell structure are the same and the other case where the estimation cell structure is a sub-partition of the test cell structure.

3.1 Estimator from the test cell information

Now the ith log-likelihood function is

$$\log l_i = \sum_{j=1}^{k} d_{ij} \log p_j(\theta). \tag{2}$$

By taking first-derivative,

$$s_i = \frac{\partial \log l_i}{\partial\theta} = \sum_{j=1}^{k} d_{ij}\frac{1}{p_j(\theta)}\frac{\partial p_j(\theta)}{\partial\theta} = J'D^{-1}d_i. \tag{3}$$

Since $J'D^{-1}P(\theta) = J'\iota_k = \frac{\partial}{\partial\theta}(P'\iota_k) = \frac{\partial}{\partial\theta}1 = 0$ (with ι_k being a $k \times 1$ vector of ones), s_i can also be written as

$$s_i = J'D^{-1}(d_i - P(\theta)). \tag{4}$$

By taking expectation of the outer product of the score in (3), we obtain the information matrix

$$\begin{aligned} H &= E[s_i s_i'] = J'D^{-1}E[d_i d_i']D^{-1}J \\ &= J'D^{-1}DD^{-1}J = J'D^{-1}J. \end{aligned} \tag{5}$$

By putting (4) and (5) into (1), we have

$$\begin{aligned} &\sqrt{n}(\hat{P} - P(\hat{\theta})) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [I_k - J(J'D^{-1}J)^{-1}J'D^{-1}](d_i - P(\theta)) \\ &\sim N(0, [I_k - J(J'D^{-1}J)^{-1}J'D^{-1}](D - PP')[I_k - J(J'D^{-1}J)^{-1}J'D^{-1}]'), \end{aligned} \tag{6}$$

where the asymptotic normality is established through the Lindberg-Levy central limit theorem. By using $J'D^{-1}P = 0$, the asymptotic variance-covariance matrix of $\sqrt{n}(\hat{P} - P(\hat{\theta}))$ in (6) can be simplified to yield

$$\sqrt{n}(\hat{P} - P(\hat{\theta})) \sim N(0, D - PP' - J(J'D^{-1}J)^{-1}J'). \tag{7}$$

From (7),

$$D^{-1/2}\sqrt{n}(\hat{P} - P(\hat{\theta})) \sim N(0, I_k - P^{1/2}P^{1/2'} - D^{-1/2}J(J'D^{-1}J)^{-1}J'D^{-1/2}). \tag{8}$$

Note that the variance-covariance matrix is in the form of the projection matrix orthogonal to the space spanned by the columns of $(P^{1/2} : D^{-1/2}J)$, since $P^{1/2}P^{1/2'} = P^{1/2}(P^{1/2'}P^{1/2})^{-1}P^{1/2'}$ and $P^{1/2\prime}D^{-1/2}J = \iota_k'J = 0$. Therefore, the variance-covariance matrix is idempotent and its rank is equal to $k-1-r$.

Lemma 1. *Let Y be distributed multivariate normal with mean 0 and covariance matrix B. A necessary and sufficient condition for $Y'CY$ to have a chi-squared distribution, with degrees of freedom equal to the rank of CB, is $BCBCB = BCB$. (see Rao, 1973, p. 188).*

To apply Lemma 1 to (8), take $Y = D^{-1/2}\sqrt{n}(\hat{P} - P(\hat{\theta}))$, $B = I_k - P^{1/2}P^{1/2'} - D^{-1/2}J(J'D^{-1}J)^{-1}J'D^{-1/2}$, and $C = I_k$. Then, since B is idempotent with rank(B) =trace$(B) = k-1-r$, we have $BCBCB = B^3 = B^2 = BCB$ and rank(CB) =rank$(B) = k-1-r$. Therefore,

$$\begin{aligned} n(\hat{P} - P(\hat{\theta}))'D^{-1}(\hat{P} - P(\hat{\theta})) &= n\sum_{j=1}^{k} \frac{(\hat{p}_j - p_j(\hat{\theta}))^2}{p_j} \\ \approx n\sum_{j=1}^{k} \frac{(\hat{p}_j - p_j(\hat{\theta}))^2}{p_j(\hat{\theta})} &\sim \chi^2(k-1-r). \end{aligned} \tag{9}$$

3.2 Estimator from a finer cell information

Let a finer cell structure be obtained by sub-partitioning each interval in the previous k cell structure. For notational convenience, let us assume that each cell C_j is partitioned into m sub-cells $C_{j1}, \cdots, C_{jm}$, $j = 1, \cdots, k$. The fact that m is constant across the original k cells is only for convenience. Define p_{jl} to be the sub-cell probability corresponding to C_{jl}: $p_{jl} = \int_{C_{jl}} f(x, \theta)\, dx$, $j = 1, \cdots, k$, and $l = 1, \cdots, m$. Define indicator variables d_{ijl} such that $d_{ijl} = 1$ if $x_i \in C_{jl}$, $d_{ijl} = 0$ otherwise, $i = 1, \cdots, n$, $j = 1, \cdots, k$, and $l = 1, \cdots, m$. Let n_{jl} denote the number of observations falling into C_{jl}: $n_{jl} = \sum_{i=1}^{n} d_{ijl}$. Of course, $\sum_{l=1}^{m} n_{jl} = n_j$ and $\sum_{j=1}^{k} n_j = n$. The empirical sub-cell probability $\hat{p}_{jl}$ is obtained as the corresponding relative frequency: $\hat{p}_{jl} = n_{jl}/n$. Let P^* be the $mk \times 1$ column vector composed of p_{jl}: $P^* = (p_{11}, \cdots, p_{1m} : \cdots : p_{k1}, \cdots, p_{km})'$. Define $\hat{P}^*$, d^*, and D^* in a similar manner.

Similar to (1), we derive

$$\sqrt{n}(\hat{P} - P(\hat{\theta}^*)) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (d_i - P(\theta) - JH^{*-1} s_i^*), \tag{1'}$$

where $\hat{\theta}^*$ is the maximum likelihood estimator calculated from the sub-partitioned data, $J = \partial P / \partial \theta'$, H^* is the Fisher information matrix of a single observation, s_i^* is the ith score function, and $\approx$ indicates that both sides of the $\approx$ have the same asymptotic distribution.

Now the ith log-likelihood function is

$$\log l_i^* = \sum_{j=1}^{k} \sum_{l=1}^{m} d_{ijl} \log p_{jl}(\theta). \tag{2'}$$

By taking first-derivative,

$$s_i^* = \frac{\partial \log l_i^*}{\partial \theta} = \sum_{j=1}^{k} \sum_{l=1}^{m} d_{ijl} \frac{1}{p_{jl}(\theta)} \frac{\partial p_{jl}(\theta)}{\partial \theta} = J^{*\prime} D^{*-1} d_i^*, \tag{3'}$$

where $J^* = \partial P^* / \partial \theta'$. Since $J^{*'} D^{*-1} P^*(\theta) = J^{*\prime} \iota_{mk} = \frac{\partial}{\partial \theta}(P^{*'} \iota_{mk}) = \frac{\partial}{\partial \theta} 1 = 0$, s_i^* can also be written as

$$s_i^* = J^{*\prime} D^{*-1} (d_i^* - P^*(\theta)). \tag{4'}$$

Using (4'), we obtain the information matrix

$$\begin{aligned} H^* &= E[s_i^* s_i^{*\prime}] = J^{*\prime} D^{*-1} E[d_i^* d_i^{*\prime}] D^{*-1} J^* \\ &= J^{*\prime} D^{*-1} D^* D^{*-1} J^* = J^{*\prime} D^{*-1} J^*. \end{aligned} \tag{5'}$$

By putting (4′) and (5′) into (1'), we have

$$\begin{aligned}\sqrt{n}(\hat{P}-P(\hat{\theta}^*)) \approx & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[T-J(J^{*\prime}D^{*-1}J^*)^{-1}J^{*'}D^{*-1}](d_i^*-P^*(\theta)) \\ \sim & N(0, A(D^*-P^*P^{*'})A'),\end{aligned} \tag{6′}$$

where $A = T - J(J^{*\prime}D^{*-1}J^*)^{-1}J^{*'}D^{*-1}$ and T is a $k \times mk$ matrix $T = I_k \otimes \iota_m$ with $\otimes$ being the Kronecker product operator (see Amemiya 1985, p. 462). The matrix T plays the role of linking d_i, $P(\theta)$, and J to d_i^*, $P^*(\theta)$, and J^*: $d_i = Td_i^*$, $P(\theta) = TP^*(\theta)$, and $J = TJ^*$. By using (i) $T(D^* - P^*P^{*'})T = D - PP'$, (ii) $J = TJ^*$, and (iii) $J^{*\prime}D^{*-1}P^* = J^{*'}\iota_{mk} = 0$, the asymptotic variance-covariance matrix in (6′) can be simplified to yield

$$\sqrt{n}(\hat{P}-P(\hat{\theta}^*)) \sim N(0, D-PP'-J(J^{*'}D^{*-1}J^*)^{-1}J'). \tag{7′}$$

From (7′),

$$\begin{aligned}& D^{-1/2}\sqrt{n}(\hat{P}-P(\hat{\theta}^*)) \\ \sim & N(0, I_k - P^{1/2}P^{1/2'} - D^{-1/2}J(J^{*'}D^{*-1}J^*)^{-1}J'D^{-1/2}).\end{aligned} \tag{8′}$$

Ryu (1993) showed that as the data cell structure gets finer, the resulting Fisher information is monotonically increasing. Using his result, we immediately notice that the variance-covariance matrix in (8′) is larger than the corresponding one in (8) in the matrix sense (meaning that their difference is a positive semi-definite matrix).

Now let us compare these two variance-covariance matrices in (8) and (8′):

$$\begin{aligned}\Sigma_k =& I_k - P^{1/2}P^{1/2'} - D^{-1/2}J(J'D^{-1}J)^{-1}J'D^{-1/2}; \\ \Sigma_{mk} =& I_k - P^{1/2}P^{1/2'} - D^{-1/2}J(J^{*'}D^{*-1}J^*)^{-1}J'D^{-1/2}.\end{aligned}$$

We have $\Sigma_{mk} \geq \Sigma_k$ in the matrix sense.

Lemma 2. *The characteristic roots of Σ_{mk} are: $k-1-r$ ones, 1 zero, and r fractions. Let $\lambda_1, \cdots, \lambda_r$ be these fractions. Then these λ's are the roots of the equation $|J'D^{-1}J - (1-\lambda)J^{*'}D^{*-1}J^*| = 0$. (proof in the Appendix)*

Lemma 3. *Let Y be multivariate normal with mean 0 and covariance matrix Ω. Let $\delta_1, \cdots, \delta_k$ be the characteristic roots of Ω. Then the distribution of $Y'Y$ is asymptotically equivalent to the distribution of $\delta_1 z_1^2 + \cdots + \delta_k z_k^2$, where $z_1, \cdots, z_k$ are i.i.d. standard normal random variates (see Chernoff and Lehmann 1954, p. 584).*

By combining Lemmas 2 and 3, we finally derive the asymptotic distribution of $n(\hat{P}-P(\hat{\theta}^*))'D^{-1}(\hat{P}-P(\hat{\theta}^*)) \approx n\sum_{j=1}^{k}(\hat{p}_j-p_j(\hat{\theta}^*))^2/p_j(\hat{\theta}^*)$. Its asymptotic distribution is equivalent to that of $z_1^2+\cdots+z_{k-1-r}^2+\lambda_1 z_{k-r}^2+\cdots+\lambda_r z_{k-1}^2$, where $z_1,\cdots,z_{k-1}$ are i.i.d. standard normal variates and $\lambda_1,\cdots,\lambda_k$ are as defined as in Lemma 2.

Note that if $m=1$, the results here reproduce the previous results since $J^{*'}D^{*-1}J^* = J'D^{-1}J$ implies $\lambda_1=\cdots=\lambda_r=0$. When the true θ is known, we can regard it as an infinite information, $J^{*'}D^{*-1}J^*=\infty$, implying $\lambda_1=\cdots=\lambda_r=1$. Therefore, in the known parameter case, the resulting test statistic is stochastically the largest, and asymptotically distributed as chi-square with degrees of freedom equal to the number of cells minus one.

4. Concluding Remarks

The results in this paper extend to models with covariances. These models are specified through conditional distribution of y given x.

Appendix

Proof of Lemma 2: Given $J'D^{-1}J$ and $J^{*'}D^{*-1}J^*$, there exists an $r\times r$ nonsingular matrix S and an $r\times r$ diagonal matrix M such that $(J'D^{-1}J)^{-1} = SS' \geq (J^{*'}D^{*-1}J^*)^{-1} = SMS'$ where the diagonal elements of M are the roots of $|(J^{*'}D^{*-1}J^*)^{-1}-\mu(J'D^{-1}J)^{-1}|=0$ and hence of $|J'D^{-1}J-\mu J^{*'}D^{*-1}J^*|=0$ (see Rao 1973). Since $0\leq J'D^{-1}J\leq J^{*'}D^{*-1}J^*$, all these roots are between zero and one. Now we can re-write Σ_k and Σ_{mk} as:

$$\begin{aligned}\Sigma_k =& I_k - P^{1/2}P^{1/2'} - D^{-1/2}JSS'J'D^{-1/2}\\ \Sigma_{mk} =& I_k - P^{1/2}P^{1/2'} - D^{-1/2}JSMS'J'D^{-1/2}.\end{aligned}$$

Let those r columns of $D^{-1/2}JS$ be $u_1,\cdots,u_r$, that is, $D^{-1/2}JS=(u_1:\cdots:u_r)$. We can easily show that $(P^{1/2},u_1,\cdots,u_r)$ is a collection of $1+r$ orthogonal unit vectors and that Σ_k is the projection matrix to their orthogonal column space. Let $\eta_1\cdots,\eta_{k-1-r}$ be a complementary set of orthogonal unit

vectors. Now Σ_{mk} can be re-written as:

$$\begin{aligned}\Sigma_{mk} =& I_k - P^{1/2}P^{1/2'} - D^{-1/2}JSMS'J'D^{-1/2} \\ =& I_k - P^{1/2}P^{1/2'} - \sum_{j=1}^{r}\mu_j u_j u_j' \\ =& [I_k - P^{1/2}P^{1/2'} - \sum_{j=1}^{r} u_j u_j'] + \sum_{j=1}^{r}(1-\mu_j)u_j u_j' \\ =& \sum_{j=1}^{k-1-r}\eta_j\eta_j' + \sum_{j=1}^{r}(1-\mu_j)u_j u_j',\end{aligned}$$

where the last equality follows since $P^{1/2}P^{1/2'} + \sum_{j=1}^{r} u_j u_j' + \sum_{j=1}^{k-1-r}\eta_j\eta_j' = I_k$.

From the resulting expression

$$\Sigma_{mk} = \sum_{j=1}^{k-1-r} 1\cdot\eta_j\eta_j' + \sum_{j=1}^{r}(1-\mu_j)\cdot u_j u_j',$$

it is clear that the characteristic roots of Σ_{mk} are: $k-1-r$ ones, 1 zero, and r fractions $1-\mu_j$ with μ_j's being the roots of the equation $|J'D^{-1}J - \mu J^{*'}D^{*-1}J^*| = 0$, $j = 1, \cdots, r$. **QED**

References

Amemiya, T. (1985), *Advanced Econometrics*, Harvard Univ. Press.

Andrews, D. W. K. (1988a), "Chi-square Diagnostic Tests for Econometric Models: Theory," *Econometrica*, 56, 1419–1453.

Andrews, D. W. K. (1988b), "Chi-square Diagnostic Tests for Econometric Models: Introduction and Applications," *Journal of Econometrics*, 37, 135–156.

Chernoff, H. and E. L. Lehmann, "The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit," *Annals of Mathematical Statistics*, 1954, 579–586.

Pearson, K. (1900), "On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arison from Random Sampling," *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 50, 157–175.

Rao, C. R. (1973), *Linear Statistical Inference and its Application*, 2nd ed., Wiley.

Ryu, Keunkwan (1993), "Monotonicity of the Fisher Information and the

Kullback-Leibler Divergence Measure," *Economics Letters*, 42, 121–128.

Ryu, Keunkwan and T. E. MaCurdy (forthcoming), "Equivalence results in chi-square tests," *Economics Letters*.