Journal of the KSQC
Vol.21, No.1, June 1993

1993년 6월 韓國品質管理學會誌 제21권 제1호

Asymptotic Expansion of the Distribution of a Studentized Test Statistic for the Slope Parameter in a Simple Linear Structural Relationship

Kyung Chang *
P. Frederic Dahm **

ABSTRACT

Variables, x and y are said to have a linear relation if $y = \beta_0 + \beta_1 x$, and β_0 and β_1 are constants. The relationship is called a structural relationship if x has positive variance (i. e., x is not fixed) and only error-prone measurements of x and y can be obtained. This paper derives (to order $n^{+1/2}$) an approximate distribution of the Studentized test statistic for testing hypotheses about the slope parameter, β_1 in a simple linear structural model. A simulation study suggests our approximate distribution is more accurate approximation to the exact distributions of the Studentized statistic than is the limiting distribution.

* Dankook University
** Texas A & M University

1. Introduction

Regression methodology is one of the most widely used tools for inferences about underlying structures of quantifiable natural and social phenomena. Most such phenomena are measured with error. Standard regression theory does not take into account these errors of measurement, i.e., the usual assumption is that explanatory variables are measured without error. In various cases explanatory variables are measured with error. For an example, root length density(L ; cm root/cm^3 soil) measurements are important for developing and testing many water extraction equations in grain production. For its explanatory variable counts of roots are used at the wall of horizontally installed minihizotrons which is clear tubes that allows continuous and nondestructive observations of rooting(Bland, W.L. and Dugs, W.A., 1988).

Measurement error models(MEM'S) explicitly take into account errors of measurement. We define a generic simple linear MEM with errors in the equation as :

$$(X_t, Y_t) = (x_t, y_t) + (e_t, u_t), \quad t = 1, \cdots, n, \qquad [1]$$

where $y_t = (1 \ \ x_t)\ B$ and $B^T = (\beta_0 \ \ \beta_1)$. The vector$(X_t \ \ Y_t)$ is the vector of observed random variables, $(x_t \ \ y_t)$ is the vector of true unobserved random variables, and$(e_t \ \ u_t)$ is the vector for measurement errors masking the vector of true random variables. Errors u_t and e_t are independent random variables with mean zero and finite positive variances σ_{uu} and σ_{ee}. We can assume either that the x_t's are fixed(i.e., a 'functional' model) or that the x_t's are independent identically distributed random variables(i.e., a 'structural' model). Since, in functional models the number of parameters increases with sample size, problems of inconsistency of maximum likelihood estimators arise for parameters of functional models. The problems associated with an increasing number of parameters have been discussed by Anderson and Rubin(1949, 1950).

The simple 'functional' model with known measurement error variance ratio was investigated by Tsukuda(1985). Building upon the work of Tsukuda, Ollivier(1986) applied Konishi's(1981) transformation to the Studentized test statistic. Extensive coverage of the MEM literature can be found in Fuller(1987) and references cited therein. This provides comprehensive treatment of estimation and large sample inference for linear MEM's. Although exact inferences can be made in a few specific cases, exact small sample inference procedures generally do not exist for MEM's. An alternative to the seemingly intractable problem of deriving exact distributions of test statistics for parameters of MEM's is to derive higher-order approximations to the distribution(and / or density) functions of test statistics.

This paper derives a higher-order approximation to the distribution function of the

Studentized test statistic for β_1 in a simple linear 'structural' model. The model assumes that the ratio of error variances is known, or equivalently that

$$\sigma_{uu} = \sigma_{ee} = \sigma^2. \qquad [2]$$

2. An asymptotic expansion of the distribution of the studentized test statistic

Random variables x_t and y_t are assumed to be under the model with [1] and [2]. That is, $(X_t\ Y_t)$ and $(e_t\ u_t)$ are, respectively, the t th error prone measurement of $(x_t\ y_t)$ and the t th vector of measurement errors. Additionally, we assume that $(x_t\ u_t\ e_t)$. $t = 1, \cdots, n$, are distributed as independent trivariate normal random vectors with mean $(\mu_x\ 0\ 0)$ and variance Diag $(\sigma_{xx}\ \sigma^2\ \sigma^2)$, where Diag means a matrix with its diagomal elements $\sigma_{xx}\ \sigma^2\ \sigma^2$, and its remaining elements 0. Further, all x_t's, u_t's, and e_t's, are assumed to be mutually independent. For convenience we shall call [1] − [2] with the associated assumptions the structural model with $\sigma_{ee} = \sigma_{uu} = \sigma^2$.

Large sample inference concerning β_1 under our model can be made using the Studentized statistic

$$T = (\hat{\beta}_1 - \beta_1) / \{\hat{V}(\hat{\beta}_1)\}^{1/2} \qquad [3]$$

where β_1, the maximum likelihood estimator of $\hat{\beta}_1$ and $\hat{V}(\hat{\beta}_1)$, a consistent estimator of the variance of the limiting distribution of $n^{1/2}\hat{\beta}_1$ are given by Fuller (1987) :

$$\hat{\beta}_1 = [-(M_{XX} - M_{YY}) + \{(M_{XX} - M_{YY})^2 + 4M_{XY}^2\}^{1/2}] / (2M_{XY}) \qquad [4]$$

and $\hat{V}(\hat{\beta}_1) = (\hat{\sigma}^2 / \hat{\sigma}_{xx})^2 + (\hat{\sigma}^2 / \hat{\sigma}_{xx})(1 + \hat{\beta}_1^2) / (n-2) + (\hat{\sigma}^2 / \hat{\sigma}_{xx})^2 \hat{\beta}_1^2 /$
$\{(n-1)(n-2)\}$ [5]

where $\overline{X} = \sum_{t=1}^{n} X_t / n, \quad \overline{Y} = \sum_{t=1}^{n} Y_t / n,$

$$M_{XX} = \sum_{t=1}^{n} (X_t - \overline{X})^2 / (n-1), \quad M_{XY} = \sum_{t=1}^{n} (X_t - \overline{X})(Y_t - \overline{Y}) / (n-1),$$

$$M_{YY} = \sum_{t=1}^{n} (Y_t - \overline{Y})^2 / (n-1), \qquad [6]$$

$\hat{\sigma}_{XX} = M_{XX} - \hat{\sigma}^2$, and $\hat{\sigma}^2 = (M_{YY} - 2\hat{\beta}_1 M_{XY} + \hat{\beta}_1^2 M_{XX}) / (1 + \hat{\beta}_1^2)$.

When we ignore the last term in (5) because it is $0(n^{+2})$ and so it is trivial, a canonical representation of (3), derived by Tsukuda(1985) is :

$$T = (n-2)^{1/2}(1-\beta\hat{\beta}^*)V_{12} / \{\det(V)\}^{1/2} \qquad [7]$$

where $V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = (n-1)\, Q^T M Q$

$$Q = (1+\beta_1^2)^{-1/2} \begin{pmatrix} 1 & -\beta_1 \\ \beta_1 & 1 \end{pmatrix}, \quad M = \begin{pmatrix} M_{XX} & M_{XY} \\ M_{XY} & M_{YY} \end{pmatrix}$$

and $\hat{\beta}^* = [-(V_{11} - V_{22}) + \{(V_{11} - V_{22})^2 + 4V_{12}^2\}^{1/2}] / (2V_{12}) = (\hat{\beta}_1 - \beta_1)/(1+\beta_1\hat{\beta}_1)$. The following steps (i)-(iv) are used to expand T : Steps (i-iii) are obtained using Taylor's expansion(Technical details and derivation of the expressions for N_0, N_1, D_0, D_1, t_0, t_1, and U_{ij} have been relegated to Chang(1990)) :

(i) The expansion of $m^{1/2}(1-\beta\hat{\beta}^*)V_{12}$ is

$$m^{1/2}(1-\beta\hat{\beta}^*)V_{12} = m\{N_0 + m^{-1/2}N_1 + O_p(m^{+1})\},$$

where $N_0 = U_{12}$, $N_1 = -\beta_1 U_{12}^2 / c + \{U_{11}U_{12} / a + U_{12}U_{22} / b\} / 3$, $(U_{11}U_{12}U_{22})^T$ is normally distributed with mean(0, 0, 0) and variance-covariance matrix Diag$(2a^2, ab, 2b^2)$, $a = (1 + \beta_1^2)\sigma_{xx} + \sigma^2$, $b = \sigma^2$, and $c = a - \sigma^2$.

(ii) The expansion of $\{\det(V)\}^{1/2}$ is, using Nel's(1978) formula,

$$\{\det(V)\}^{1/2} = \{V_{11}V_{22} - V_{12}^2\}^{1/2} = m\{D_0 + m^{-1/2}D_1 + O_p(m^{+1})\},$$

where $D_0 = \{a\ \sigma^2\}^{1/2}$, and $D_1 = \{aU_{22} + \sigma^2 U_{11}\} / (2D_0)$.

(iii) The expansion of T is

$$T = (n-2)^{1/2}, (1-\beta\hat{\beta}^*)V_{12} / \{\det(V)\}^{1/2}$$

$$= \frac{m\{N_0 + m^{-1/2}N_1 + O_p(m^{+1})\}}{m\{D_0 + m^{-1/2}D_1 + O_p(m^{+1})\}} = t_0 + m^{-1/2}t_1 + O_p(m^{+1}),$$

where $t_0 = U_{12} / \{a\ \sigma^2\}^{1/2}$,

$$t_1 = \{U_{11}U_{12} / (3a) + U_{12}U_{22} / (3b) - U_{12}^2\beta_1 / c\} / \{a\sigma^2\}^{1/2} - U_{12}(aU_{22} + \sigma^2 U_1) / \{2(a\ \sigma^2)^{3/2}\},$$

and the term t_0 is a standard normal random variable.

(iv) The conditional expectation of t_1 given t_0 is used to express the approximate characteristic function of T in terms of t_0. This conditional expectation is $E(t_1 \mid t_0) = -\{a \; \sigma^2\}^{1/2} \beta_1 t_0^2 / C$. The characteristic function of T is formally expanded as

$$C_T(\theta) = E\{\exp(i\theta^T)\} = E[\exp\{i\theta(t_0 + m^{-1/2}t_1 + O_p(m^{+1}))\}]$$
$$= \exp(-\theta^2) - K^{**}(-i\theta)E\{t_0^2\exp(i\theta t_0)\}/m^{-1/2} + O(m^{+1})$$

where $K^{**} = -\{a \quad \sigma^2\}^{1/2}\beta_1 / c$, and $a = (1+\beta_2)\sigma_{xx} + \sigma^2$.

(v) The approximate cumulative distribution function of T is obtained by applying the inverse Fourier transformation of $C_T(\theta)$ of step(iv) :

$$Pr[T \le \xi] = \int_{-\infty}^{\xi} \frac{1}{2\pi}\int_{-\infty}^{+\infty} \exp(-i\theta x)C_T(\theta)\,d\theta\,dx$$
$$= \Phi(\xi) - K^{**}\xi^2\phi(\xi)m^{-1/2} + O(m^{+1}) \qquad [8]$$

where Φ and ϕ are the cumulative distribution function and the probability density function, respectively, of a standard normal random variable, (we need Tsukuda's(1985) formulas for (iv – v)).

3. Simulation Results and Discussion

Monte Carlo experiments were performed to evaluate empirically quantiles of the distribution function of the test statistic T in [3]. Let

$$M_{ZZ} = \begin{pmatrix} M_{XX} & M_{XY} \\ M_{XY} & M_{YY} \end{pmatrix}$$

To generate T, we need M_{XX}, M_{XY}, and M_{YY} of expressions [3] – [5]. The matrix $(n-1)$ M_{ZZ} is distributed as a Wishart distribution with $n-1$ degrees of freedom, variance-covariance matrix Σ, where

$$\Sigma = \begin{pmatrix} \sigma_{XX} + \sigma^2 & \beta_1\sigma_{XX} \\ \beta_1\sigma_{XX} & \beta_1^2\sigma_{XX} + \sigma^2 \end{pmatrix}. \qquad [9]$$

we denote the distribution of $(n-1)$ M_{ZZ} by $W(n-1, \Sigma)$. Let U be the Cholesky decomposition of Σ such that $\Sigma = U^T U$. Then, since $(n-1)(U^T)^{-1} M_{ZZ} U^{-1}$ is $W(n-1, I)$, $(n-1)(U^T)^{-1} M_{ZZ} U^{-1}$ can be represented as

$$V^* = \begin{pmatrix} V_{11}^* & V_{12}^* \\ V_{21}^* & V_{22}^* \end{pmatrix}$$

which can be generated by the relations, $a_{11} = V_{11}^*$, $a_{12} = V_{12}^*(V_{11}^*)^{-1/2}$, $a_{22} = V_{22}^* - a_{12}^2$, where a_{11}, a_{12}, and a_{22} are independently distributed as $\chi^2(n-1)$, $n(0, 1)$, and $\chi^2(n-2)$ random variables, respectively (Kshirsagar, 1972, p. 55), and $N(0, 1)$ and $\chi^2(df)$ denote a standard normal distribution and a chi-square distribution with df degrees of freedom, respectively. Hence, M_{ZZ} can be expressed as : $M_{ZZ} = U^T V^* U / (n-1)$. To generate X, x, a_{11}, a_{12}, and a_{22}, we use the Box-Müller method to generate $N(0, 1)$ random variables and the Cornish-Fisher approximation to generate chi-square random variables, respectively

Since expressions [7] and [9] show that the distribution of T depends soley on the parameter values n, β_1, and the ratio σ_{XX} / σ^2 and we can assume with no loss in generality that $\sigma^2 = 1$, the set of parameters to be specifid in the Monte Carlo experiment is (n, β, σ_{XX}). For 12 sets of parameters ⟨Table 1⟩ we obtained the empirical cumulative distribution functions of the test statistic T, generating 10000 T's. Using these sets, we examine the effects of n, β, and σ_{XX} on the distribution of the Studentized test statistic T

⟨ Table 1. Sets of Parameter Values Simulation Study ⟩

Case	n	β_1	σXX	Case	n	β_1	σXX	Case	n	β_1	σXX
1	20	0	4	5	20	1	2	9	40	1	4
2	20	0.5	4	6	20	1	4	10	60	1	1
3	20	2	4	7	40	1	1	11	60	1	2
4	20	1	1	8	40	1	2	12	60	1	4

There is no completely accurate way to check whether the empirical distribution is a suitable proxy for the unknown exact distribution of T. However, when $\beta_1 = 0$, T has an exact Student's t distribution with $n-2$ degrees of freedom. We can check the error of the empirical distribution for the case $\beta_1 = 0$. Tables 2 and 3 show the errors between the empirical distribution and Student's t distribution with $n-2$ degrees of freedom, denoted by $t(n-2)$. Columns (2), (3), and (4) in Table 2, respectively, give the empirical distribution values, the exact distribution values ($t(18)$), the errors of empirical distribution values (column (2) − column (3)). (5) and (6) can be explained as : Letting $x = 10000\,\hat{p}$, x is distributed as binomial distribution with parameters 10000 and p with 10000 generated T's where p = Prob $[t \leq \xi]$ for Student's $t(18)$ random variable t. Then, the level−0.95 confidence interval for $\hat{p}$ is given by $[p - 1.96\{p(1-p)/10000\}^{1/2},\ p + 1.96\{p(1-p)/10000\}^{1/2}]$. Thus, confidence intervals for $\hat{p}$ are in (6) when ξ and p have specific values. In Table 2 the maximum error is −0.0045, and we can see the probabilities in colum $n(2)$ are located in each corresponding interval in (6), respectively.

Columns (2), (3), and (4) in Table 3, respectively, represent empirical quantile values, exact quantile values ($t(18)$), errors of empirical quantile values(column (2)−column (3)). (5) is obtained as : After generating 10000 Student's $t(18)$ random variable R's we construct level−0.95 confidence intervals for quantile values for $p = 0.01$, $p = 0.05$, $p = 0.95$, and $p = 0.99$, using a formula [lower bound, upper bound] = $[R_{(L)}, R_{(U)}]$, where $R_{(i)}$ is the i th order statistic of n R's randomly generated, U and L are the integers obtained by rounding U^* and L^* upward to the next higher integers, $U^* = np + 1.96\{np(1-p)\}$ and $L^* = np - 1.96\{np(1-p)\}$, for given p (Conover, 1980). Also, in Table 3 the maximum error is −0.079 and we can see the empirical quantiles in column (2) are located in each corresponding interval (5), respectively. Hence, it is reasonable that we consider our empirical distribution to be a suitable proxy distribution of the exact distribution.

〈 Table 2. Errors of the Empirical Distribution Function of T 〉
($n = 20$, $\beta_1 = 0$ $\sigma_{XX} = 4$)

(1) ξ	(2) Empirical Prob.	(3) $t(18)$ Prob.	(4) Errors (2)−(3)	(5) p	(6) Confidence Interval for $\hat{p}$
−4	0.0004	0.0004	0.0000	0.0004	[0.0000, 0.0008]
−3	0.0044	0.0038	0.0006	0.0038	[0.0026, 0.0050]
−2	0.0334	0.0304	0.0030	0.0304	[0.0270, 0.0338]
−1	0.1672	0.1653	0.0019	0.1653	[0.1580, 0.1726]
0	0.4955	0.5000	−0.0045	0.5000	[0.4902, 0.5098]
1	0.8324	0.8347	−0.0023	0.8347	[0.8274, 0.8420]
2	0.9680	0.9696	−0.0016	0.9696	[0.9662, 0.9730]

For remaining cases in Table 1 except ($n = 20$, $\beta_1 = 0$, $\sigma_{XX} = 4$), empirical quantiles, our approximate quantiles, standard normal quantiles, and Student's $t(n-2)$ quantiles are compared with a format similar with Table 3 (Chang, 1990, pp. 21−25). Almost our all approximate quantiles are shown to have the least error values. Among the results, we present a part of them in Table 4. Since we can see the least absolute error values in parentheses in column (4) compared with other quantiles, we know our approximation is more accurate than any other quantiles using standard normal and Student's $t(n-2)$

approximation with regard to at least our twoelve cases.

〈 Table 3. Errors of the Empirical Quantiles of T 〉

$(n = 20, \beta_1 = 0, \sigma XX = 4)$

(1) ξ	(2) Empirical Quantiles	(3) $t(18)$ Quantiles	(4) Errors (2)−(3)	(5) Confidence Interval for Quantile Values
0.01	−2.631	−2.552	−0.079	$[t_{(81)}, t_{(120)}] = [-2.775, -2.512]$
0.05	−1.756	−1.734	−0.022	$[t_{(458)}, t_{(543)}] = [-1.807, -1.694]$
0.95	1.755	1.734	0.021	$[t_{(9458)}, t_{(9543)}] = [1.686, 1.770]$
0.99	2.558	2.552	0.006	$[t_{(9881)}, t_{(9920)}] = [2.405, 2.591]$

〈 Table 4. Empirical Quantiles of T and Three Approximated Quantiles 〉

(1) Prob.	(2) $(n, \beta_1, \sigma XX)$	(3) Empirical Quantiles	(4) Our Approxi Quantiles	(5) Stan. Normal Quantiles	(6) Stu. $t(n-2)$ Quantile
0.01	(20, 1, 4)	−3.12	−2.80 (0.3)	−2.33 (0.8)	−2.55 (0.6)
0.05	(20, 1, 4)	−2.01	−1.88 (0.1)	−1.65 (0.4)	−1.73 (0.3)
0.95	(20, 1, 4)	1.48	1.41 (0.1)	1.65 (0.2)	1.73 (0.3)
0.99	(20, 1, 4)	2.03	1.85 (0.2)	2.33 (0.3)	2.55 (0.5)
0.01	(20, 2, 4)	−3.16	−2.91 (0.3)	−2.33 (0.8)	−2.55 (0.6)
0.05	(20, 2, 4)	−2.06	−1.94 (0.1)	−1.65 (0.4)	−1.73 (0.3)
0.95	(20, 2, 4)	1.41	1.35 (0.1)	1.65 (0.2)	1.73 (0.3)
0.99	(20, 2, 4)	1.93	1.74 (0.2)	2.33 (0.4)	2.55 (0.6)
0.01	(40, 1, 1)	−3.29	−3.09 (0.2)	−2.33 (1.0)	−2.43 (0.9)
0.05	(40, 1, 1)	−2.12	−2.03 (0.1)	−1.65 (0.5)	−1.69 (0.4)
0.95	(40, 1, 1)	1.25	1.27 (0.0)	1.65 (0.4)	1.69 (0.4)
0.99	(40, 1, 1)	1.72	1.57 (0.2)	2.33 (0.6)	2.43 (0.7)

4. Summary and Future Research

This paper investigates small-sample inference concerning the slope parameter in a simple linear structural model with known error variance ratio. Derivation of the exact distribution of the Studentized statistics appears to be intractable. Application of the large sample results of the limiting distribution for the Studentized statistics to small samples appears to yield inaccurate inferences. As an alternative, using Taylor series, characteristic function and Fourier inverse transformation, this study proposes an expansion of the Studentized statistic. Simulation experiments show that approximations to the exact distribution of the Studentized statistic based on our expansion are more accurate than are their limiting distribution approximations.

Several important extentions of these results will be attempted. These include the following :

(i) We shall next consider the simple linear structural model with replicated observations $X_{tj}, t = 1, \cdots, n; j = 1, \cdots, r_t$ and $Y_{tk}, t = 1, \cdots, n; k = 1, \cdots, s_t$.

(ii) In many experiments it is unrealistic to assume that measurement errors are uncorrelated. Refined inference procedures can be developed for cases where $\sigma_{eu} = Cov(u_{tj}, e_{tj}) \neq 0$.

Finally, because validity of the expansion [8] is moot, a task proving it remains on mathematicians. However, we believe arguments similar to those presented in Morimune and Kunitomo(1980) can be used to prove the validity. The results of this paper and Tskuda (1985) support the validity empirically in simulation experiments.

REFERENCES

1. Anderson, T. W. and Rubin, H. (1949), "*Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations*," The Annals of Mathematical Statistics, 20, pp. 46−63.
2. Anderson, T. W. and Rubin, H. (1950), "*The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations*," The Annals of Mathematical Statistics, 21, pp. 570−582.
3. Bland, W. L. (1988) and Dugas W. A., "*Root length Density from Minirhizotron Observations*," Agronomy Journal, 80, 2, pp. 271−275.
4. Chang, K. (1990), "*Asymptotic Expansions of the Distributions of Studentized Test Statistics for the Slope Parameter in Simple Linear Structural Relationships*," Ph. D. Dissertation Texas A & M University, Department of Statistics.
5. Conover, W. J. (1980), Practical Nonparametric Statistics, New York, John Wiley and Sons.
6. Fuller, W. A (1987), Measurement Error Models, New York, John Wiley and Sons.
7. Konishi, S. (1981), "*Normalizing Transformations of Some Statistics in Multivariate Analysis*," Biometrika, 68, pp. 647−651.
8. Kshirsagar, A. M. (1972), Multivariate Analysis, New York, M. Dekker.
9. Morimune, K. and Kunitomo, N. (1980), "*Improving the Maximum Likelihood Estimate in Linear Functional Relationships for Alternative Parameter Sequences*," Journal of the American Statistical Association, 75, pp. 230−237.
10. Nel, D. G. (1978), "*On the Symmetric Multivariate Normal Distribution and the Asymptotic Expansion of a Wishart Matrix*," South African Statistical Journal, 12, pp. 145−159.
11. Ollivier, V. (1986), "*Improved Test Statistics for the Slope Parameter in a Simple Linear Functional Model*," M. S. Thesis, Texas A & M University, Department of Statistics.
12. Tsukuda, Y. (1985), "*Asymptotic Expansions of the Distributions of Test Statistics for the Slope in a Simple Linear Functional Relationship*," Ph. D. Dissertation, Texas A&M University, Department of Statistics.