E C O N O M I C S B U L L E T I N

On distribution approximation: a simple comparative study on procedural variations of the Zheng test

Lawrence Dacuycuy
Kyoto Graduate School of Economics

Abstract

The study analyzes the performance of the Zheng test for functional form in different scenarios concerning the distribution approximation of the test statistic. We apply the test statistic for validating simple wage functions.

Financial and administrative support from the Japan International Cooperation Agency (JICA) and Japan International Cooperation Center (JICE) are gratefully acknowledged. I would also like to thank my supervisor, Prof. Kimio Morimune, Prof. Nishiyama, Prof. Hitomi and Professor Yatchew for allowing me to modify his Splus scripts. Of course, the remaining errors are mine.
Citation: Dacuycuy, Lawrence, (2005) "On distribution approximation: a simple comparative study on procedural variations of the Zheng test." *Economics Bulletin*, Vol. 3, No. 11 pp. 1−10
Submitted: November 11, 2004. **Accepted:** March 1, 2005.
URL: http://www.economicsbulletin.com/2005/volume3/EB−04C40011A.pdf

1 Introduction

In the field of consistent nonparametric tests for functional form, there are competing techniques in approximating the distribution of the test statistic. The issue concerns the adequacy of the usual normal approximation of the distribution of the test statistic via U-statistics theory. This led to the reevaluation of the Zheng (1996) test for functional forms in parametric regression. In resolving the issue of distribution approximation, the most popular method involves the use of wild bootstrap methodology. Li and Wang (1998) noted the superior performance of bootstrap based tests relative to the one that uses asymptotic expansion. They also claimed that the wild bootstrap is applicable, particularly to models with heteroskedastic errors. Gozalo and Linton (2002) noted that by relying on asymptotic theory to derive the distribution, certain orders are left out.

In this simple paper we compare the results from various modifications of the Zheng test with respect to the distribution approximation of the test statistic when dealing with continuous variables only. The study will deal with the comparative performance of the Zheng test when applied to determining the correct functional specification of the wage function using a subsample of male wage earners in the Bicol region during the period 1988–1995. Subjecting wage functions to alternative procedures of the Zheng test is important in light of the observation that specifications might be rendered invalid not because they are but because of the assumed distribution reference used in making the statistical decision. It is also important because much of the studies dealing with wage functions still rely on normal approximation or inconsistent test procedures.

The paper is organized as follows: Section 2 revisits the development of bootstrap–based test procedures for parametric regression. Section 3 introduces the test equations. Section 4 presents and analyzes the results of the tests and finally, section 5 concludes.

2 Asymptotic approximation of the distribution of the Zheng statistic

Consider a regression model of the following form:

$$y_i = \varphi(x_i; \theta) + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

where θ refers to the $k \times 1$ vector of unknown coefficients; ϵ_i refers to the $n \times 1$ vector of unobservable model components and φ is a function that may be linear or nonlinear in coefficients. Obviously, equation 1 is estimable

via nonlinear or linear least squares. We are interested in investigating the adequacy of the assumed functional specification in characterizing the conditional moment $E[y_i|x_i]$. Zheng (1996) used the conditional moment condition $E[\varepsilon_i E[\varepsilon_i|x_i]f(x_i)] = 0$ to derive the test statistic for testing the validity of parametric functional forms. Based on Zheng, $f(x_i)$ is the density function of X. The presence of this component in $E[\varepsilon_i E[\varepsilon_i|x_i]f(x_i)]$ is to address the random denominator problem that arises when $E[\varepsilon_i|x_i]$ is estimated via nonparametric regression. A popular estimator of $E[\varepsilon_i|x_i]$ is known as the Nadaraya–Watson estimator. This is written as

$$E[\epsilon_i|x_i] = \frac{\frac{1}{(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{1}{h^k}K\left(\frac{x_i-x_j}{k}\right)\epsilon_i}{\frac{1}{(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{1}{h^k}K\left(\frac{x_i-x_j}{k}\right)} \tag{2}$$

where $K(\cdot)$ refers to a symmetric and nonnegative kernel function; h is a smoothing parameter and k represents the dimension of the continuous variables. The denominator represents the nonparametric density function. Using the definitional convention for the density function, we have

$$E[\epsilon_i|x_i]f(x_i) = \frac{\frac{1}{(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{1}{h^k}K\left(\frac{x_i-x_j}{k}\right)\epsilon_i}{f(x_i)}f(x_i) \tag{3}$$

It can be recalled that the sample analogue of $E[\varepsilon E[\varepsilon_i|x_i]f(x_i)]$ is written as

$$V_n^{ls} = \frac{1}{n(n-1)h^k}\sum_{i=1}\sum_{j\neq i}K\left(\frac{x_i-x_j}{k}\right)e_i e_j \tag{4}$$

where e_i refers to the residual from a root–N consistent regression.

Zheng established the following distribution for the test statistic under the null using Hall's (1984) central limit theorem for degenerate U-statistics.

$$nh^{k/2}V_n^{ls} \rightarrow N\left(0, 2\int K^2(u)du\int[\sigma^2(x)]^2f^2(x)dx\right) \tag{5}$$

The variance can be estimated via the following estimator:

$$\hat{\sigma} = \frac{2}{n(n-1)h^k}\sum_{i=1}\sum_{j\neq i}K\left(\frac{x_i-x_j}{k}\right)e_i^2e_j^2 \tag{6}$$

Letting $\tau = \sqrt{\frac{n-1}{n}}nh^{k/2}V_n/\hat{\sigma}^{1/2}$, the null is rejected if $\tau > z_\alpha$ wherein the latter is the α^{th} quantile in the standard normal distribution. Li and Wang (1998) noted that since the Zheng test statistic does not have finite sample bias, higher ordered kernel functions are not needed. This also allows a wider

set of bandwidths to be used for testing the conditional mean. However, they pointed out that the asymptotic distribution based test may not fare well in small samples. Thus they prescribe the use of resampling methods like the wild bootstrap. The wild bootstrap method is based on a two–point distribution. Following Li and Wang (1998), let the bootstrap residual be denoted by ϵ_*. The bootstrap residual is based on the fitted residuals from equation 1. Conditional on the data $\{(y_i, x_i)\}_{i=1}^n$, the bootstrap residuals should satisfy $E[\epsilon_{i*}] = 0$, $E[\epsilon_{i*}^2] = \hat{e}_i^2$ and $E[\epsilon_{i*}^3] = \hat{e}_i^3$. Li and Wang (1998) outlined the procedure as follows:

1. Generate the bootstrap residuals that satisfy the conditional moments of ϵ_{i*}.

2. Utilizing ϵ_{i*}, calculate $y_{i*} = \varphi(x_i; \hat{\theta}) + \epsilon_{i*}$. Using observations on y_{i*} and x_i, estimate equation 1 and compute the residuals $e_{i*} = y_{i*} - \varphi(x_i; \hat{\theta}_*)$, where $\hat{\theta}_*$ is the coefficient estimate using the bootstrap sample, y_{i*} and x_i.

3. Use the generated residuals to calculate the test statistic in equation 4.

Instead of relying on the normal approximation, derive the empirical distribution by replicating the procedure M times. Given the test statistics across bootstrapped samples, the distribution will be determined. This implies that the null is rejected if $\tau > z_\alpha^*$, wherein the latter is the α^{th} percentile of the empirical bootstrap distribution.

3 Investigating wage functions: A reconsideration

Due to the widespread applicability of wage functions in dealing with labor issues, it is no surprise that heightened interest in specification analysis is observed. For instance, in early studies, Heckman and Polachek (1974) used Box–Cox functional specifications to decide which dependent variable as well as variable transformations are appropriate. Zheng (2000) investigated popular functional forms using Bierens non-smoothing test for functional form. Miles and Mora (2003) used various test procedures to determine the validity of wage functions in Spain and Uruguay. Using Brazilian data, Dougherty and Jimenez (1991) investigated specifications in the spirit of Heckman and Polachek. Investigating wage functional validity in itself should be an empirical regularity not only because we need to have correct estimates for returns to schooling but more importantly to ensure confidence in results coming

from secondary computations that rely heavily on wage functional forms. Dacuycuy (2004) noted the huge disparities among models that include or exclude interaction terms and such disparities are expectedly reflected by the resulting decomposition outcomes.

In this section, we compare the results from various distributional approaches of the Zheng test.[1] First, we test various wage functional forms using ad hoc bandwidth parameters and the usual asymptotic distribution of the test statistic. We then vary this by sticking to the ad hoc bandwidths but now rely on the wild bootstrap approximation. For both the first and second scenarios, the bandwidth selection rule for the constant component would be similar to that of Yatchew (2003) and Li (1999).[2] In both variants, however, the multivariate representation of the kernel function will be the quartic product kernel which is expressed as $K(u) = K(u_1) \times K(u_2)$, where $u_l = \frac{(x_{il} - x_{jl})}{h_l}$ and $K(u_l) = \frac{15}{16}(1 - u^2)^2$ for $u_l \in [-1, 1]$, $l = 1, 2$.

All of these will be applied in ascertaining the correct functional form for simple wage functions. These specifications are as follows:

$$
\begin{aligned}
log\ W_i &= \beta_0 + \beta_1 AGE_i + \beta_2 AGE_i^2 + \beta_3 SCH_i + \varepsilon_{1i} \\
log\ W_i &= \beta_0 + \beta_1 AGE_i + \beta_2 AGE_i^2 + \beta_3 AGE_i^3 + \beta_4 AGE_i^4 + \beta_5 SCH_i + \varepsilon_{2i} \\
log\ W_i &= \beta_0 + \beta_1 AGE_i + \beta_2 AGE_i^2 + \beta_3 SCH_i + \beta_4 AGE \times SCH_i + \varepsilon_{3i} \\
log\ W_i &= \beta_0 + \beta_1 AGE_i + \beta_2 AGE_i^2 + \beta_3 SCH_i + \beta_4 SCH_i^2 + \varepsilon_{4i} \\
log\ W_i &= \beta_0 + \beta_1 AGE_i + \beta_2 AGE_i^2 + \beta_3 AGE_i^3 + \beta_4 AGE_i^4 + \beta_5 SCH_i \\
&+ \beta_6 SCH_i^2 + \varepsilon_{5i}
\end{aligned}
$$

where $log\ W_i$ refers to the real wage, SCH, years of schooling and AGE, age of individual i. The real wage is computed by dividing third quarter real earnings by the total number of hours. Consumer price indices for the Bicol region are used to deflate earnings. Descriptive statistics are reported in the appendix.

4 Results

Results in tables 1 and 2 demonstrate the risks that are associated with ad hoc bandwidth selection methods. They are called ad hoc bandwidth procedures because the constant, c, in $cn^{-1/5}$ for each regressor is pegged at an arbitrary value of 1, instead of allowing grid search using cross validation

[1] Miles and Mora (2003) for instance, still relied on ad hoc bandwidth procedures and employed the bootstrap method for Stute's and Bierens tests.

[2] We modified a similar program written by Yatchew (2003) in Splus to handle the computation of the test statistics for wage function analysis.

methods. Both tests rely on the asymptotic distribution of the test statistic. Based on the results, there are instances wherein models are accepted in one test but rejected in another, indicating that the resulting bandwidth choice may not fall within the allowable band for admissible bandwidth parameters.

Table 1: P–values based on asymptotic method: 1988–1995

Specification	1988	1989	1990	1991	1992	1993	1994	1995
(1)	0.001	0.022	0.005	0.000	0.396	0.039	0.035	0.531
(2)	0.004	0.066	0.015	0.000	0.380	0.186	0.119	0.573
(3)	0.446	0.061	0.568	0.070	0.742	0.082	0.129	0.696
(4)	0.006	0.341	0.183	0.000	0.400	0.388	0.548	0.921
(5)	0.020	0.598	0.274	0.000	0.384	0.730	0.733	0.933

Note: The bandwidths which follow that of Li (1999) are equal to $C_0\hat{\sigma_x}n^{-1/5}$, where $C_0 = 1$ and $\hat{\sigma_x}$ is the standard deviation of x.

Table 2: P–values based on asymptotic method: 1988–1995

Specification	1988	1989	1990	1991	1992	1993	1994	1995
(1)	0.000	0.012	0.000	0.000	0.478	0.021	0.002	0.164
(2)	0.000	0.035	0.000	0.000	0.463	0.089	0.014	0.190
(3)	0.258	0.063	0.237	0.099	0.906	0.056	0.023	0.373
(4)	0.000	0.454	0.005	0.000	0.510	0.451	0.310	0.781
(5)	0.000	0.695	0.016	0.000	0.495	0.719	0.492	0.807

Note: The bandwidths which follow that of Yatchew (2003) are equal to $C_0\frac{max(x)-min(x)}{2}n^{-1/5}$, where $C_0 = 1$.

A possible way to analyze the relationship between distribution approximation and test outcomes is to resort to bootstrap methods. We implement the Zheng test using two ad hoc bandwidth parameters but use the wild bootstrap method. We will follow the procedure outlined in Li and Wang (1998) and peg bootstrap replications at 100. The original calculation for the test statistic would then be compared with the quantile of interest derived from the empirical bootstrap distribution. Results for bootstrap–based tests are reported in tables 3 and 4.

Table 3: P–values based on bootstrap method: 1988–1995

Specification	1988	1989	1990	1991	1992	1993	1994	1995
(1)	0.000	0.003	0.000	0.000	0.277	0.014	0.046	0.259
(2)	0.000	0.015	0.000	0.000	0.261	0.052	0.014	0.257
(3)	0.219	0.059	0.286	0.009	0.382	0.058	0.009	0.379
(4)	0.016	0.148	0.121	0.000	0.221	0.197	0.255	0.591
(5)	0.015	0.260	0.082	0.000	0.121	0.245	0.315	0.619

Note: The bandwidths which follow that of Li (1999) are equal to $C_0\hat{\sigma_x}n^{-1/5}$, where $C_0 = 1$ and $\hat{\sigma_x}$ is the standard deviation of x.

Table 4: P–values based on bootstrap method: 1988–1995

Specification	1988	1989	1990	1991	1992	1993	1994	1995
(1)	0.000	0.000	0.000	0.000	0.275	0.014	0.011	0.078
(2)	0.000	0.000	0.000	0.000	0.204	0.008	0.000	0.053
(3)	0.127	0.042	0.071	0.056	0.631	0.043	0.007	0.107
(4)	0.006	0.160	0.010	0.006	0.240	0.249	0.083	0.392
(5)	0.003	0.220	0.000	0.000	0.154	0.258	0.123	0.363

Note: The bandwidths which follow that of Yatchew (2003) are equal to $C_0\frac{max(x)-min(x)}{2}n^{-1/5}$, where $C_0 = 1$.

It is noticeable that the general effect of resorting to the bootstrap method to approximate the distribution of the test statistic is to lower the p-value which means that there are some hypotheses accepted by non-bootstrap test which are rejected when the bootstrap distribution is used.

5 Concluding remarks

The study underscores the importance of distributional assumptions associated with nonparametric consistent based test procedures for parametric functional forms. The results highlighted at times, diverging test outcomes, indicating that there is a need to properly discern which technique is applicable.

The results may also reflect partly the choice of bandwidths. In view of the critical dependence of the nonparametric test procedures on bandwidth selection considerations, more systematic data–driven selection procedures should be adopted to replace ad hoc bandwidth procedures, a critical point addressed by a new generation of consistent nonparametric based test that

rely on cross–validation for bandwidth determination. (Racine and Li (2004), Hsiao, Racine and Li (2004))

References

Bierens, H. (1990). "A consistent conditional moment test of functional form." *Econometrica* **58**, 1443 – 1458.

Dacuycuy, L. (2004) "On Wage Specifications and Male Inequality Decompositions: A Reexamination of Evidence", Manuscript.

Dougherty, C. and E. Jimenez (1991) "The specification of earnings functions: Tests and Implications." *Economics of Education Review* **10**, 85–98.

Fan, Y., Q. Li (1996) "Consistent model specification tests: omitted variables, parametric and semiparametric functional forms." *Econometrica* **64**, 865–890.

Gozalo, P. and O.B. Linton (2001) "Testing additivity in generalized nonparametric regression models with estimated parameters." *Journal of Econometrics* **104**, 1–48.

Hall, P. (1984) "Cental limit theorem for integrated square error of multivariate nonparametric density estimators." *Journal of Multivariate Analysis* **14**, 1 – 16.

Heckman, J. and S. Polachek (1974) "Empirical Evidence on the Functional Form of the Earnings–Schooling Relationship." *Journal of the American Statistical Association* **69**, 350–354.

Hsiao, C., Racine, J. and Li, Q. (2004). A consistent model specification test with mixed categorical and continuos data.

Li, Q. (1999) "Consistent model specification tests for time series econometric models." *Journal of Econometrics* **92**, 101–147

Li, Q. and S. Wang (1998) "A simple consistent bootstrap test for a parametric regression function." *Journal of Econometrics* **87**, 145 - 165.

Miles, D. and J. Mora (2003) "On the performance of nonparametric specifi-

cation tests in regression models." *Computational Statistics and Data Analysis* **42**, 477 – 490.

National Statistics Office. Labor Force Survey, various years.

Powell, J., J.H. Stock and T.M. Stoker (1989) "Semiparametric estimation of index coefficients." *Econometrica* **57**, 1403 – 1430.

Yatchew, A. (2003) Semiparametric Regression for the Applied Econometrician. Cambridge University Press.

Zheng, J.X. (1996) "A consistent test of functional form via nonparametric estimation technique." *Journal of Econometrics* **75**, 263–289.

Zheng, J.X. (2000)"A Nonparametric Analysis of the U.S. Earnings Distribution." in *Advances in Econometrics: Applying Kernel and Nonparametric Estimation to Economic Topics* by Fomby, Thomas B. and R. Carter Hill eds., JAI Press.

Table 5: Descriptive Statistics for Male Workers in the Bicol Region: 1988–1995

Variables	1988	1989	1990	1991	1992	1993	1994	1995
Logarithm of Real Wage	1.51 (0.71)	1.53 (0.67)	1.60 (0.74)	1.59 (0.67)	1.59 (0.72)	1.59 (0.65)	1.65 (0.72)	1.63 (0.74)
Age	33.54 (12.49)	34.04 (12.37)	34.44 (12.64)	33.43 (12.20)	34.91 (12.70)	34.39 (12.30)	35.22 (12.49)	34.54 (12.07)
Schooling	7.92 (3.37)	8.28 (3.45)	8.17 (3.42)	8.01 (3.20)	8.14 (3.33)	7.99 (3.27)	8.16 (3.26)	8.35 (3.19)
Number of Observations	611	578	634	680	646	615	663	722