Asymptotics of an empirical bridge of regression on induced order statistics

Artyom Kovalevskii*

Abstract

We propose a class of tests for linear regression on concomitants (induced order statistics). These tests are based on sequential sums of regression residuals. We self-center and self-normalize these sums. The resulting process is called an empirical bridge. We prove weak convergence of the empirical bridge in uniform metrics to a centered Gaussian process. The proposed tests are of chi-square type.

Keywords: concomitants, weak convergence, regression residuals, empirical bridge.

1 Introduction

There are few approaches to regression models testing. An empirical fluctuation process function is implemented in R package by Zeileis at al. [40]. The process uses recursive regression residuals proposed by Brown, Durbin, Ewans [11].

MacNeill [31] studied linear regression for time series. He obtained limit processes for sequences of partial sums of regression residuals. Later Bischoff [10] showed that the MacNeill’s theorem holds in a more general setting. Aue et al. [1] introduced a new test for polynomial regression functions which is analogous to the classical likelihood test. This approach is developed in [2], [3].

Stute [37] proposed a class of tests for one-parametric case.

* pandorra@ngs.ru, Novosibirsk State Technical University, Novosibirsk State University. The research was supported by RFBR grant 17-01-00683
Our approach is adopted specifically for regression on induced order statistics (concomitants). These models arise in applications [28]. Partial results are proposed in [29], [30]. The proofs use the theory of induced order statistics.

David [12] and Bhattacharya [7] have introduced induced order statistics (concomitants) simultaneously. The asymptotic theory was developed in in [4], [5], [6], [8], [9], [13], [14], [15], [16], [17], [18], [19], [21], [22], [24], [38], [39]. Strong convergence to a corresponding Gaussian process can be proved by methods of [20], [27], [32], [35].

New contributions to the theory deal with extremal order statistics [23], [34], [36].

2 Main result and corollaries

Let \((\delta_i, \xi_i, \eta_i) = (\delta_i, \xi_{i1}, \ldots, \xi_{im}, \eta_i)\) be independent and identically distributed random vector rows, \(\delta_i\) has uniform distribution on \([0, 1]\), \(i = 1, \ldots, n\). Random variables \(\delta_1, \ldots, \delta_n\) are not observed. They will be used for ordering.

We assume a linear regression hypothesis \(\eta_i = \xi_i \theta + e_i\). Here

\[ \xi_i \theta \overset{def}{=} \sum_{j=1}^{m} \xi_{ij} \theta_j, \]

\(\{e_i\}_{i=1}^{n}\) and \(\{\delta_i, \xi_i\}_{i=1}^{n}\) are independent, \(\{e_i\}_{i=1}^{n}\) are i.i.d., \(\mathbb{E} e_1 = 0, \text{Var} e_1 = \sigma^2 > 0\).

Vector \(\theta = (\theta_1, \ldots, \theta_m)\) and constant \(\sigma^2\) are unknown.

We order rows by the first component, that is, we change rows \(j\) and \(k\) while \(j < k\) and \(\delta_j > \delta_k\). The result is matrix \((U, X, Y)\) with rows \((U_i, X_i, Y_i) = (U_i, X_{i1}, \ldots, X_{im}, Y_i)\). So \(U_1 < \ldots < U_n\) a.s. are order statistics from uniform distribution on \([0, 1]\). Elements of matrix \((X, Y)\) are concomitants. Let \(\varepsilon_i = Y_i - X_i \theta\). Note that \(\varepsilon_i\) are i.i.d. and have the same distribution as \(e_1\). Sequences \(\{\varepsilon_i\}_{i=1}^{n}\) and \(\{(U_i, X_i)\}_{i=1}^{n}\) are independent.

Let \(\theta\) be LSE:

\[ \hat{\theta} = (X^T X)^{-1} X^T Y. \]

It does not depend on the order of rows.

Let \(h(x) = \mathbb{E}\{\xi_1 | \delta_1 = x\}\) be conditional expectation, \(L(x) = \int_{0}^{x} h(s) \, ds\) be induced theoretical generalised Lorentz curve (see [1]),

\[ b^2(x) = \mathbb{E}\left( (\xi_1 - h(x))^T (\xi_1 - h(x)) \mid \delta_1 = x \right) \]
Let $g_{ij} = E\xi_i\xi_j$, $G = (g_{ij})_{i,j=1}^m$, then $G = \int_0^1 (b^2(x) + h^T(x)h(x))\,dx$.

Let $\bar{\xi}_i = Y_i - X_i\hat{\theta}$, $\hat{\Delta}_k = \sum_{i=1}^k \bar{\xi}_i$, $\hat{\Delta}_0 = 0$.

Let $Z_n = \{Z_n(t), 0 \leq t \leq 1\}$ be a piecewise linear random function with nodes

\[
\left(\frac{k}{n}, \frac{\hat{\Delta}_k}{\sqrt{n}}\right).
\]

We designate weak convergence in $C(0, 1)$ with uniform metrics by $\Rightarrow$.

**Theorem 1** If $G$ exists, $\det G \neq 0$, then $Z_n \Rightarrow Z$. Here $Z$ is a centered Gaussian process with covariation function

\[
K(s, t) = \min(s, t) - L(s)G^{-1}L^T(t), \quad s, t \in [0, 1].
\]

Let $Z_n^0$ be an empirical bridge (see [28], [29], [30]):

\[
Z_n^0(t) = \frac{\sigma}{\hat{\sigma}}(Z_n(t) - tZ_n(1)), \quad 0 \leq t \leq 1,
\]

with $\hat{\sigma}^2 = \sum_{i=1}^n \bar{\xi}_i^2/n$. Let $L^0(t) = L(t) - tL(1)$.

Let $L_{n,j}$ be an empirical induced generalised Lorentz curve:

\[
L_{n,j}(t) = \frac{1}{n} \sum_{i=1}^{[nt]} X_{ij},
\]

$L_n = (L_{n,1}, \ldots, L_{n,m})$, $L_n^0(t) = L_n(t) - tL_n(1)$.

**Corollary 1** Let assumptions of Theorem 1 be held.

1) Then $Z_n^0 \Rightarrow Z^0$, a centered Gaussian process with covariation function

\[
K^0(s, t) = \min\{s, t\} - st - L^0(s)G^{-1}(L^0(t))^T, \quad s, t \in [0, 1].
\]

2) Let $d \geq 1$ be integer,

\[
q = (Z_n^0(1/(d + 1)), \ldots, Z_n^0(d/(d + 1))),
\]

$g_{ij} = \frac{1}{n} \sum_{k=1}^n X_{ki}X_{kj}$, $\tilde{G} = (g_{ij})_{i,j=1}^m$.

\[
\tilde{K}^0(s, t) = \min(s, t) - st - L^0_n(s))^T\tilde{G}^{-1}(L^0_n(t))^T,
\]

$Q = (\tilde{K}^0(i/(d + 1), j/(d + 1)))_{i,j=1}^d$. Then $qQ^{-1}q^T$ converges weakly to a chi-squared distribution with $d$ degrees of freedom.
Note that ordering by $\xi_{i1}, i = 1, \ldots, n$, can be viewed as ordering by $\delta_i$ with $h_1(x) = F^{-1}_{\xi_{i1}}(x)$ (quantile function). In this case $L_1(t) = \int_0^t F^{-1}_{\xi_{i1}}(x) \, dx$.

The next corollary is proved in [29].

**Corollary 2** Let $Y_i = \theta_1 X_{i1} + \varepsilon_i, i = 1, \ldots, n, \theta_1 \in \mathbb{R}, (X_{11}, \ldots, X_{n1})$ are order statistics of i.i.d. $(\xi_{11}, \ldots, \xi_{n1})$, random variables $(\varepsilon_1, \ldots, \varepsilon_n)$ are i.i.d. and independent of them, $0 < E \xi_{11}^2 < \infty$, $E \varepsilon_1 = 0$, $0 < \text{Var} \varepsilon_1 = \sigma^2 < \infty$. Then $Z_n \Rightarrow Z$, a centered Gaussian process with covariance function

$$\min(s, t) - L_1(s)L_1(t)/E \xi_{11}^2.$$ 

The next corollary is a partial case of Theorem 1 in [30].

**Corollary 3** Let $Y_i = \theta_1 X_{i1} + \theta_2 + \varepsilon_i, i = 1, \ldots, n, \theta_1, \theta_2 \in \mathbb{R}, (X_{11}, \ldots, X_{n1})$ are order statistics of i.i.d. $(\xi_{11}, \ldots, \xi_{n1}),$ random variables $(\varepsilon_1, \ldots, \varepsilon_n)$ are i.i.d. and independent of them, $0 < \text{Var} \xi_{11} < \infty$, $E \varepsilon_1 = 0$, $0 < \text{Var} \varepsilon_1 = \sigma^2 < \infty$. Then $Z_n \Rightarrow Z$, a centered Gaussian process with covariance function

$$\min(s, t) - st - L_0(s)L_0(t)/\text{Var} \xi_{11}.$$ 

### 3 Proof of Theorem 1

Note that

$$\hat{\Delta}_k = \sum_{i=1}^k (Y_i - X_i \hat{\theta}) = \sum_{i=1}^k (X_i(\theta - \hat{\theta}) + \varepsilon_i)$$

$$= \sum_{i=1}^k (X_i(\theta - (X^T X)^{-1}X^T Y) + \varepsilon_i)$$

$$= \sum_{i=1}^k (X_i(\theta - (X^T X)^{-1}X^T (X \theta + \varepsilon)) + \varepsilon_i)$$

$$= \sum_{i=1}^k (\varepsilon_i - X_i(X^T X)^{-1}X^T \varepsilon).$$

Note that $X_{[nt]}/n \rightarrow L(t)$ a.s. uniformly on compact sets, and $X^T X/n \rightarrow G$ a.s.

So we study process

$$\left\{ \sum_{i=1}^{[nt]} (\varepsilon_i - L(t)G^{-1}X^T \varepsilon), \quad t \in [0, 1] \right\}.$$
This process is a bounded linear functional of \((m+1)\)-dimensional process
\[
\left\{ \sum_{i=1}^{[nt]} (X_i \varepsilon_i, \varepsilon_i), \ t \in [0,1] \right\}.
\]

We use the functional central limit theorem for induced order statistics by Davydov and Egorov [17].
We assume that \(\eta_i = \xi_i \theta + \epsilon_i\), \(\{\epsilon_i\}_{i=1}^n\) and \(\{\xi_i\}_{i=1}^n\) are independent, \(\{\epsilon_i\}_{i=1}^n\) are i.i.d., \(\text{E} \epsilon_1 = 0\), \(\text{Var} \epsilon_1 = \sigma^2 > 0\).

Let see rows \((\delta_i, \xi_i \epsilon_i, \epsilon_i) = (\delta_i, \xi_i \epsilon_i, \ldots, \xi_{im} \epsilon_i, \epsilon_i)\). We have
\[
\text{E}(\xi_1 \epsilon_1 | \delta_1 = x) = 0, \ \text{E}(\epsilon_1 | \delta_1 = x) = 0, \ x \in [0,1].
\]

The conditional covariance matrix of the vector \((\xi_1 \epsilon_1, \epsilon_1)\) is
\[
\tilde{b}^2(x) = \text{E} \left( (\xi_1 \epsilon_1, \epsilon_1)^T (\xi_1 \epsilon_1, \epsilon_1) | \delta_1 = x \right) = \sigma^2 \begin{pmatrix} b^2(x) + h^T(x)h(x) & h^T(x) \\ h(x) & 1 \end{pmatrix}.
\]

Let \(\tilde{b}(x)\) be an upper triangular matrix such that \(\tilde{b}(x)\tilde{b}(x)^T = \tilde{b}^2(x)\). Then
\[
\tilde{b}(x) = \sigma \begin{pmatrix} b(x) & h^T(x) \\ 0 & 1 \end{pmatrix}.
\]

Here \(b(x)\) is an upper triangular matrix such that \(b(x)b(x)^T = b^2(x)\). By Theorem 1 of Davydov and Egorov [17] the process
\[
\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} X_i \varepsilon_i, \sum_{i=1}^{[nt]} \varepsilon_i \right\}^T, \ t \in [0,1]
\]
converges weakly in the uniform metrics to the Gaussian process
\[
\left\{ \int_0^t \tilde{b}(x) \ dW_{m+1}(x), \ t \in [0,1] \right\}.
\]

Here \(W_{m+1} = (W_1, \ldots, W_{m+1})^T\) is an \((m+1)\)-dimensional standard Wiener process.

So process
\[
\left\{ \frac{1}{\sigma \sqrt{n}} \sum_{i=1}^{[nt]} (\varepsilon_i - L(t)G^{-1}X^T \varepsilon)^T, \ t \in [0,1] \right\}
\]
converges weakly in the uniform metrics to the Gaussian process $Z = \{Z(t), \ t \in [0,1]\}$,

$$Z(t) = W_{m+1}(t) - L(t)G^{-1} \left( \int_0^1 b(x) \ dW_m(x) + \int_0^1 h^T(x) \ dW_{m+1}(x) \right)^T,$$

$W_m = (W_1, \ldots, W_m)^T$.

By the noted convergencies $X_{[nt]}/n \to L(t)$ a.s. uniformly on compact sets, $X^T X/n \to G$ a.s., the process $Z_n$ has the same weak limit $Z$.

The covariance function of the limiting Gaussian process $Z$ is

$$K(s, t) = \mathbf{E}Z(s)Z(t)$$

$$\quad = \min(s,t) - L(s)G^{-1} \int_0^t h^T(x) \ dx - L(t)G^{-1} \int_0^s h^T(x) \ dx$$

$$\quad \quad + L(s)G^{-1} \int_0^1 (b^2(x) + h^T(x)h(x)) \ dx G^{-1}L^T(t)$$

$$\quad = \min(s,t) - L(s)G^{-1}L^T(t).$$

The proof is complete.

**Acknowledgement**

The research was supported by RFBR grant 17-01-00683.

**References**

[1] Aue A., Horvath L., Huskova M., Kokoszka P., 2008. Testing for change in polynomial regression. Bernoulli 14, 637–660.

[2] Aue A., Horvath L., 2013. Structural breaks in time series. Journal of Time Series Analysis 34:1, 1–16.

[3] Aue A., Rice G., Sonmez O., 2018. Detecting and dating structural breaks in functional data without dimension reduction. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80:3, 509–529.

[4] Balakrishnan N. and Cohen A.C., 1991. Order Statistics and Inference: Estimation Methods, Academic Press, Boston, MA.
[5] Barnett V., 1976. The ordering of multivariate data, J. Roy. Statist. Soc. Ser. A 139, 318–354.

[6] Barnett V., Green P.J. and Robinson A., 1976. Concomitants and correlation estimates, Biometrika 63, 323–328.

[7] Bhattacharya P.K., 1974. Convergence of sample paths of normalized sums of induced order statistics, The Annals of Statist. 2, 1034–1039.

[8] Bhattacharya P.K., 1976. An invariance principle in regression analysis, The Annals of Statist. 4, 621–624.

[9] Bhattacharya P.K., 1984. Induced order statistics: Theory and applications, in: P.R. Krishnaiah and P.K. Sen, eds., Handbook of Statistics, Vol. 4 (North-Holland, Amsterdam), 383–403.

[10] Bischoff W., 1998. A functional central limit theorem for regression models. Ann. Stat. 26, 1398–1410.

[11] Brown R. L., Durbin J., Evans J. M., 1975. Techniques for testing the constancy of regression relationships over time. J. R. Statist. Soc. 37, 149–192.

[12] David H. A., 1973. Concomitants of order statistics, Bull. Internat. Statist. Inst. 45, 295–300.

[13] David H. A., 1991. Concomitants of order statistics: Review and recent developments, Symposium in Honor of C.W. Dunnett, Hamilton, Ont., Canada.

[14] David H. A. and Galambos J., 1974. The asymptotic theory of concomitants of order statistics, J. Appl. Probab. 11, 762–770.

[15] David H. A., Nagaraja H. N., 2003. Order Statistics. New Jersey: John Wiley & Sons.

[16] David H. A., O’Connell M. J. and Yang S. S., 1977. Distribution and expected value of the rank of a concomitant of an order statistic, The Annals of Statist. 5, 216–223.

[17] Davydov Y., Egorov V., 2000. Functional limit theorems for induced order statistics. Mathematical Methods of Statistics 9(3), 297–313.
[18] Davydov Y., Zitikis R., 2004. Convex rearrangements of random elements. Fields Institute Communications, Vol. 44, 141–171.

[19] Domma F., Giordano S., 2016. Concomitants of m-generalized order statistics from generalized Farlie-Gumbel-Morgenstern distribution family, Journal of Computational and Applied Mathematics, V. 294, 413–435.

[20] Einmah J., Mason D., 1988. Strong limit theorems for weighted quantile processes, Annals. Probab., V. 16, 4, 1623–1643.

[21] Egorov V. A., Nevzorov V. B., 1982. Some theorems on induced order statistics, Theory Prob. Appl. V. 27, N.3, 633–639.

[22] Egorov V. A., Nevzorov V. B., 1984. Convergence to the normal law of sums of induced order statistics. J. of Soviet Math. V. 25, N.3, 1139–1146.

[23] Eryilmaz S. and Bairamov I. G., 2003. On a new sample rank of an order statistics and its concomitant, Statistics & Probability Letters, vol. 63, issue 2, 123–131.

[24] Galambos J., 1987. The Asymptotic Theory of Extreme Order Statistics, (Krieger, Malabar, FL, 2nd. ed.).

[25] Gastwirth J. L., 1971. A general definition of the Lorenz curve. Econometrica 39, 1037–1039.

[26] Goldie C. M., 1977. Convergence theorems for empirical Lorenz curves and their inverses. Adv. Appl. Prob. 9, 765–791.

[27] Koul H. L. 2002. Weighted Empirical Processes in Dynamic Nonlinear Models, SpringerVerlag, New York.

[28] Kovalevskii A., 2013. A regression model for prices of second-hand cars. Applied methods of statistical analysis. Applications in survival analysis, reliability and quality control, 124–128.

[29] Kovalevskii A. P., Shatalin E. V., 2015. Asymptotics of Sums of Residuals of One-Parameter Linear Regression on Order Statistics. Theory of probability and its applications. Vol. 59, iss. 3, 375–387.
[30] Kovalevskii A., Shatalin E., 2016. A limit process for a sequence of partial sums of residuals of a simple regression on order statistics. Probability and Mathematical Statistics, Vol. 36, Fasc. 1, 113–120.

[31] MacNeill I. B., 1978. Limit processes for sequences of partial sums of regression residuals. Ann. Prob. 6, 695–698.

[32] Sakhanenko A. I., Sukhovershina O. A., 2015. On accuracy of approximation in Koul’s theorem for weighted empirical processes, Sib. Elektron. Mat. Izv., 12, 784–794.

[33] Sen P.K., 1976. A note on invariance principles for induced order statistics, The Annals of Probab. 4, 474–479.

[34] Shahbaz M. Q., Shabbaz S., Mohsin M., Rafiq A., 2010. On distribution of bivariate concomitants of records, Applied Mathematics Letters, Volume 23, Issue 5, 567–570.

[35] Shorack G., Wellner J., 1986. Empirical processes with applications to statistics, Wiley N. Y.

[36] Stepanov A., Berred A., Nevzorov V. B., 2016. Concomitants of records: Limit results, generation techniques, correlation, Statistics & Probability Letters, V. 109, 184–188.

[37] Stute W., 1997. Nonparametric model checks for regression. Ann. Statist. 25, 613–641.

[38] Yang S.S., 1977. General distribution theory of the concomitants of order statistics, The Annals of Statist. 5, 996–1002.

[39] Zamanzade E., Vock M., 2015. Variance estimation in ranked set sampling using a concomitant variable, Statistics & Probability Letters, Volume 105, 1–5.

[40] Zeileis A., Leisch F., Hornik K., Kleiber Ch., 2002. Strucchange: An R Package for Testing for Structural Change in Linear Regression Models. Journal of Statistical Software, 7(2), 1–38.