Martin Schindler

Kolmogorov-Smirnov two-sample test based on regression rank scores

Applications of Mathematics, Vol. 53 (2008), No. 4, 297–304

Persistent URL: `http://dml.cz/dmlcz/140323`

KOLMOGOROV-SMIRNOV TWO-SAMPLE TEST BASED ON REGRESSION RANK SCORES*

Martin Schindler, Praha

(Supplement to the special issue of Appl. Math. *53* (2008), No. 3)

Abstract. We derive the two-sample Kolmogorov-Smirnov type test when a nuisance linear regression is present. The test is based on regression rank scores and provides a natural extension of the classical Kolmogorov-Smirnov test. Its asymptotic distributions under the hypothesis and the local alternatives coincide with those of the classical test.

Keywords: regression rank scores, Kolmogorov-Smirnov test, two sample problem, Cramér-von Mises test

MSC 2010: 62G08, 62G10, 62J05

1. Introduction

In [3] Hájek extended the Kolmogorov-Smirnov test of the hypothesis of randomness to tests against alternatives of simple linear regression. He expressed the test criterion (see equation (4)) as a functional of a special rank score process (Hájek's rank scores) for which he proved convergence to Brownian bridge. We mention this fact in Subsection 2.1. Similarly he extended the Cramér-von Mises and the Rényi tests. If, instead of Hájek's rank scores, we consider the process of regression rank scores (see e.g. [1]), we can extend the (two-sample) Kolmogorov-Smirnov test also to a nuisance regression.

So here we deal with the tests of Kolmogorov-Smirnov type on one component of the regression parameter $\boldsymbol{\beta}$ in the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. These tests, based on regression rank scores, were introduced in Jurečková [5]. We derive the two-sample variant of the test and show that this test represents a straightforward extension of the classical Kolmogorov-Smirnov test, more specifically the variant of the classical Kolmogorov-Smirnov test that is the most sensitive to difference in location.

* This work was supported by the Czech Science Foundation under Grant No. 201/05/H007 and by Research Project LC06024.

Note that already in Gutenbrunner and Jurečková [1] the regression rank score process was studied. Further, in Gutenbrunner et al. [2] a broader class of tests of hypothesis in linear regression model based on regression rank scores was derived. This class represents a generalization of simple linear rank tests.

Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{X}^{(1)}\boldsymbol{\beta}^{(1)} + \mathbf{x}^{(p)}\beta_p + \mathbf{e}, \tag{1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ is a vector of observations, $\mathbf{X} = \mathbf{X}_{N\times p}$ is a known design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' = (\boldsymbol{\beta}^{(1)'}, \beta_p)$ are unknown parameters, $\mathbf{e} = (e_1, \ldots, e_N)'$ is the vector of i.i.d. errors, the matrix $\mathbf{X}^{(1)}$ consisting of the first $p-1$ columns of the matrix $\mathbf{X}$ represents the nuisance regression and $\mathbf{x}^{(p)}$ is the pth column of $\mathbf{X}$. Here we do not specify the vector $\mathbf{x}^{(p)}$ but later, in Section 2, we will set $\mathbf{x}^{(p)} = (1, \ldots, 1, 0, \ldots, 0)'$ to derive and describe the two-sample Kolmogorov-Smirnov test.

We want to test the hypothesis

$$H_0\colon\ \beta_p = 0, \quad \boldsymbol{\beta}^{(1)} \ \text{unspecified}.$$

This problem will be tested by a test of Kolmogorov-Smirnov (K-S) type. In the presence of nuisance regression, regression rank scores (RRS) are employed. RRS (see e.g. [2]) in the submodel of (1) given by H_0 are defined as the vector of solutions $\hat{\mathbf{a}}_N(\alpha) = (\hat{a}_{N1}(\alpha), \ldots, \hat{a}_{NN}(\alpha))'$, $0 \leqslant \alpha \leqslant 1$ of the linear programming problem ($\mathbf{1}_N$ denotes the $(N\times 1)$ vector of ones):

$$\max \mathbf{Y}'\hat{\mathbf{a}}_N(\alpha)$$

subject to

$$\mathbf{X}^{(1)'}\hat{\mathbf{a}}_N(\alpha) = (1-\alpha)\mathbf{X}^{(1)'}\mathbf{1}_N, \qquad \hat{\mathbf{a}}_N(\alpha) \in [0,1]^N. \tag{2}$$

1.1. Assumptions

We will impose the following conditions on the regression matrix $\mathbf{X}$ and on the underlying distribution function F.

Let $\mathbf{x}_i'$ denote the ith row of the matrix $\mathbf{X}$, $i = 1, \ldots, N$. We assume that the matrix $\mathbf{X} = \mathbf{X}_N$ satisfies the regularity conditions

(X.1) $x_{i1} = 1$, $i = 1, \ldots, N$,

(X.2) $\max\limits_{\substack{1\leqslant i\leqslant N\\ 1\leqslant j\leqslant p}} |x_{ij}| = \mathcal{O}(N^{(2(b-a)-\delta)/(1+4b)})$
for some a, b, δ, $0 < a \leqslant \frac{1}{4} - \varepsilon$, $0 < b - a \leqslant \frac{1}{2}\varepsilon$, $\varepsilon > 0$, $\delta > 0$,

(X.3) $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^3 = \mathcal{O}(1)$ as $N\to\infty$,

(X.4) $\mathbf{D}_N = N^{-1}\mathbf{X}_N'\mathbf{X}_N \overset{N\to\infty}{\longrightarrow} \mathbf{D}$, where $\mathbf{D}$ is a positively definite matrix.

Further assume that the errors $e_1,\dots,e_N$ in (1) are i.i.d. with an absolutely continuous distribution function F whose tails are assumed to satisfy the following regularity conditions (F.1)–(F.4) (these conditions are satisfied by many common densities f including t-distributions with 5 or more d.f.):

(F.1) F has an absolutely continuous density f, positive for $A<x<B$ and decreasing monotonously when $x\to A+$, $x\to B-$, where $-\infty\leqslant A=\sup\{x\colon F(x)=0\}$ and $+\infty\geqslant B=\inf\{x\colon F(x)=1\}$. The derivative f' of f is bounded a.e.

(F.2) $|F^{-1}(\alpha)|\leqslant c(\alpha(1-\alpha))^{-a}$ (with a from (X.2)) for $0<\alpha\leqslant\alpha_0$, $1-\alpha_0\leqslant\alpha<1$ for some $0<\alpha_0\leqslant\frac{1}{2}$ and for some $c>0$.

(F.3) $1/f(F^{-1}(\alpha))\leqslant c(\alpha(1-\alpha))^{-1-a}$ for $0<\alpha\leqslant\alpha_0$, $1-\alpha_0\leqslant\alpha<1$, $c>0$.

(F.4) $\left|\frac{f'(x)}{f(x)}\right|\leqslant c(|x|+1)$, $x\in\mathbf{R}^1$, $c>0$.

1.2. Statistic of K-S type

Consider the model (1) and define the projection matrix

$$\mathbf{H}^{(1)}=\mathbf{H}_N^{(1)}=\big(h_{ij}^{(1)}\big)_{i=1,\dots,N}^{j=1,\dots,N}=\mathbf{X}_N^{(1)}(\mathbf{X}_N^{(1)'}\mathbf{X}_N^{(1)})^{-1}\mathbf{X}_N^{(1)'}$$

and $\mathbf{x}^*=(x_1^*,\dots,x_N^*)'=\mathbf{H}^{(1)}\mathbf{x}^{(p)}$ the projection of $\mathbf{x}^{(p)}$ into the space spanned by the columns of $\mathbf{X}_N^{(1)}$.

We define the process $\{S_N(t)\colon 0\leqslant t\leqslant 1\}$ on $C[0,1]$:

$$S_N(t)=\left(\sum_{i=1}^{N}(x_i^{(p)}-x_i^*)^2\right)^{-1/2}\sum_{i=1}^{N}(x_i^{(p)}-x_i^*)\hat{a}_{Ni}(t).$$

It is shown in [5] that under the conditions (X.1)–(X.4) and (F.1)–(F.4) it follows from [2, Theorem 3.2] that

$$\sup_{0\leqslant t\leqslant 1}|S_N(t)-\tilde{S}_{N(t)}|\overset{p}{\to}O\quad\text{as}\quad N\to\infty,\tag{3}$$

where

$$\tilde{S}_N(t)=\left(\sum_{i=1}^{N}(x_i^{(p)}-x_i^*)^2\right)^{-1/2}\sum_{i=1}^{N}(x_i^{(p)}-x_i^*)I[e_i>F^{-1}(t)],\quad 0\leqslant t\leqslant 1$$

and that $S_N(t)$ converges to the Brownian bridge in the uniform topology on $C[0,1]$. In the next section we show how to construct a test based on this fact in the case of a two sample problem.

Consider the model (1) and let $\mathbf{x}^{(p)} = (1, \ldots, 1, 0, \ldots, 0)'$ be the vector with m ones and n zeros, $m + n = N$.

We want to test the hypothesis H_0 of no difference between the samples. This two-sample problem will be tested by a test of Kolmogorov-Smirnov (K-S) type which is a generalization of the classical rank test of K-S type (the variant that is the most sensitive to difference in location) that works in the model (1) without nuisance regression ($\mathbf{X}^{(1)} = \mathbf{1}_N$).

2.1. Classical K-S two-sample test

In the location model (model (1) with $\mathbf{X}^{(1)} = \mathbf{1}_N$) the solution $\hat{\mathbf{a}}_N(\alpha)$ of (2) specializes to Hájek's rank scores $\mathbf{a}_N^*(\alpha) = (a_{N1}^*(\alpha), \ldots, a_{NN}^*(\alpha))$ where

$$a_{Ni}^*(\alpha) = a_N^*(R_i, \alpha) = \begin{cases} 1, & 0 \leqslant \alpha \leqslant (R_i - 1)/N, \\ R_i - \alpha N, & (R_i - 1)/N < \alpha \leqslant R_i/N, \\ 0, & R_i/N < \alpha \leqslant 1, \end{cases}$$

where R_i is the rank of Y_i among $Y_1, \ldots, Y_N$, $i = 1, \ldots, N$. Hájek in [3] or Hájek & Šidák in [4] considered the process $T_N = \{T_N(t)\colon\ 0 \leqslant t \leqslant 1\}$,

$$T_N(t) = \Big(\sum_{i=1}^{N}(c_{Ni} - \overline{c}_N)^2\Big)^{-1/2} \sum_{i=1}^{N}(c_{Ni} - \overline{c}_N) a_N^*(R_i, t), \tag{4}$$

with a triangular array $\mathbf{c}_N = (c_{N1}, \ldots, c_{NN})'$ of constants satisfying

$$\sum_{i=1}^{N}(c_{Ni} - \overline{c}_N)^2 / \max_{1\leqslant i\leqslant N}(c_{Ni} - \overline{c}_N)^2 \overset{N\to\infty}{\longrightarrow} \infty, \quad \overline{c}_N = N^{-1}\sum_{i=1}^{N} c_{Ni}$$

and showed that T_N converges in the uniform topology on $C[0,1]$ to the Brownian bridge. We define the empirical distribution functions of the two samples $\hat{F}_m(x) = m^{-1}\sum\limits_{i=1}^{m} I[Y_i \leqslant x]$ and $\hat{G}_n(x) = n^{-1}\sum\limits_{i=m+1}^{N} I[Y_i \leqslant x]$ and the zero-one quantity V_i, $V_i = 1$ if $Y_{(i)}$ is one of $Y_1, \ldots, Y_m$, $i = 1, \ldots, N$.

Setting $\mathbf{c}_N = \mathbf{x}^{(p)}$, $\max\limits_{0\leqslant t\leqslant 1} T_N(t)$ coincides with the classical K-S two-sample test statistic T^+. We use the fact that (2) implies $\sum\limits_{i=1}^{N} a_N^*(R_i, j/N) = (1 - j/N)N = N - j$ and that $\max\limits_{0\leqslant t\leqslant 1} T_N(t) = \max\limits_{1\leqslant j\leqslant N} T_N(j/N)$ since the process $T_N(t)$ is linear on every interval $[(j-1)/N, j/N]$, $j = 1, \dots, N$:

$$
\begin{aligned}
T^+ &= \Big(\frac{mn}{N}\Big)^{1/2} \max_{1\leqslant j\leqslant N} [\hat{G}_n(Y_{(j)}) - \hat{F}_m(Y_{(j)})] \\
&= \Big(\frac{mn}{N}\Big)^{1/2} \max_{1\leqslant j\leqslant N} \Big[\frac{1}{n}((1-V_1)+\ldots+(1-V_j)) - \frac{1}{m}(V_1+\ldots+V_j)\Big] \\
&= \Big(\frac{N}{mn}\Big)^{1/2} \max_{1\leqslant j\leqslant N} \Big[\frac{jm}{N} - (V_1+\ldots+V_j)\Big] \\
&= \Big(\frac{N}{mn}\Big)^{1/2} \max_{1\leqslant j\leqslant N} \Big[\frac{jm}{N} - \sum_{i=1}^{m}\Big(1 - a_N^*\Big(R_i, \frac{j}{N}\Big)\Big)\Big] \\
&= \Big(\frac{N}{mn}\Big)^{1/2} \max_{1\leqslant j\leqslant N} \Big[\sum_{i=1}^{m} a_N^*\Big(R_i, \frac{j}{N}\Big) - \frac{m}{N}(N-j)\Big] \\
&= \Big(\frac{N}{mn}\Big)^{1/2} \max_{1\leqslant j\leqslant N} \Big[\Big(1-\frac{m}{N}\Big)\sum_{i=1}^{m} a_N^*\Big(R_i, \frac{j}{N}\Big) - \frac{m}{N}\sum_{i=m+1}^{N} a_N^*\Big(R_i, \frac{j}{N}\Big)\Big] \\
&= \max_{0\leqslant t\leqslant 1} T_N(t).
\end{aligned}
$$

2.2. Main result

Let us first recall that $\mathbf{x}^{(p)} = (1, \dots, 1, 0, \dots, 0)'$ and $\mathbf{x}^* = (x_1^*, \dots, x_N^*)' = \mathbf{H}^{(1)}\mathbf{x}^{(p)}$. The projection matrix $\mathbf{H}^{(1)} = \big(h_{ij}^{(1)}\big)_{i=1,\dots,N}^{j=1,\dots,N}$ corresponding to $\mathbf{X}^{(1)}$ is idempotent, so

$$
\begin{aligned}
\sum_{i=1}^{N}(x_i^{(p)} - x_i^*)^2 &= (\mathbf{x}^{(p)} - \mathbf{x}^*)'(\mathbf{x}^{(p)} - \mathbf{x}^*) = \mathbf{x}^{(p)'}(\mathbf{I}_N - \mathbf{H}^{(1)})(\mathbf{I}_N - \mathbf{H}^{(1)})\mathbf{x}^{(p)} \\
&= \mathbf{x}^{(p)'}(\mathbf{I}_N - \mathbf{H}^{(1)})\mathbf{x}^{(p)} = m - \sum_{i=1}^{m}\sum_{j=1}^{m} h_{ij}^{(1)} > 0.
\end{aligned}
$$

Theorem 1. *Assume that* $\mathbf{X}_N$ *satisfies* (X.1)–(X.4) *and* F *satisfies* (F.1)–(F.4). *Let* $\hat{\mathbf{a}}_N(\alpha) = (\hat{a}_{N1}(\alpha), \dots, \hat{a}_{NN}(\alpha))'$, $0 \leqslant \alpha \leqslant 1$ *be the regression rank scores corresponding to the submodel of the model* (1), *i.e. under* H_0. *Then the process* $\{S_N(t)\colon\ 0 \leqslant t \leqslant 1\}$,

$$
\begin{aligned}
S_N(t) &= \Big(\sum_{i=1}^{N}(x_i^{(p)} - x_i^*)^2\Big)^{-1/2} \sum_{i=1}^{N}(x_i^{(p)} - x_i^*)\hat{a}_{Ni}(t) \\
&= \Big(m - \sum_{i=1}^{m}\sum_{j=1}^{m} h_{ij}^{(1)}\Big)^{-1/2} \Big[\sum_{i=1}^{m}\Big(1 - \sum_{j=1}^{m} h_{ij}^{(1)}\Big)\hat{a}_{Ni}(t) \\
&\qquad + \sum_{i=m+1}^{N}\Big(-\sum_{j=1}^{m} h_{ij}^{(1)}\Big)\hat{a}_{Ni}(t)\Big],
\end{aligned}
$$

converges to the Brownian bridge in the uniform topology on $C[0,1]$. *Thus, for* $K_N^+ = \max_{0\leqslant t\leqslant 1} S_N(t)$ *and* $K_N = \max_{0\leqslant t\leqslant 1} |S_N(t)|$ *we can write, under* H_0,

$$\lim_{N\to\infty} P(K_N^+ < x) = \begin{cases} 1 - \mathrm{e}^{-2x^2}, & x \geqslant 0,\\ 0, & x < 0,\end{cases}$$

$$\lim_{N\to\infty} P(K_N < x) = \begin{cases} 1 - 2\sum\limits_{z=1}^{\infty}(-1)^{z+1}\mathrm{e}^{-2z^2x^2}, & x \geqslant 0,\\ 0, & x < 0.\end{cases}$$

Proof. It follows from (3) and from the properties of the Brownian bridge. □

The statistics K_N^+ and K_N, similarly to the classical K-S statistics, can be used for testing the two sample problem with nuisance regression against one-sided and two-sided alternatives.

By Theorem 1 the test based on K_N^+ rejects H_0 on the asymptotic significance level α provided $K_N^+ \geqslant (-\frac{1}{2}\log\alpha)^{1/2}$.

The asymptotic power of the test based on K_N^+, against the local alternative

$$H_N\colon\ \beta_p = N^{-1/2}\Delta, \quad \text{with } \beta_1,\dots,\beta_{p-1} \text{ unspecified},$$

can be obtained from the following theorem.

Theorem 2. *Under the conditions of Theorem* 1 *and under* H_N, *the process*

$$S_N(t) - \Biggl[\Bigl(m - \sum_{i=1}^m\sum_{j=1}^m h_{ij}^{(1)}\Bigr)^{1/2}\Delta N^{-1/2} f(F^{-1}(t))\Biggr]$$

converges to the Brownian bridge $\{Z(t)\colon\ 0\leqslant t\leqslant 1\}$ *in the uniform topology on* $C[0,1]$ *from which it follows that*

$$\begin{aligned}&\lim_{N\to\infty} P(K_N^+ \geqslant x|H_N)\\ &\qquad = P\biggl(\max_{0\leqslant t\leqslant 1}\Bigl\{Z(t) + \Bigl(m - \sum_{i=1}^m\sum_{j=1}^m h_{ij}^{(1)}\Bigr)^{1/2}\Delta N^{-1/2} f(F^{-1}(t))\Bigr\} \geqslant x\biggr)\end{aligned}$$

for any $x > 0$. *Additionally,*

$$\begin{aligned}&\lim_{N\to\infty} P\Bigl(K_N^+ \geqslant \Bigl(-\frac{\log\alpha}{2}\Bigr)^{1/2}|H_N\Bigr) - \alpha\\ &\quad = \Biggl[2\Bigl(m - \sum_{i=1}^m\sum_{j=1}^m h_{ij}^{(1)}\Bigr)^{1/2}\Delta N^{-1/2}\alpha\Bigl(-\frac{\log\alpha}{2}\Bigr)^{1/2}\\ &\qquad\times\int_0^1 -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\psi(u,\alpha)\,\mathrm{d}u\Biggr](1+o(1))\end{aligned}$$

holds for $\Big(m - \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} h_{ij}^{(1)}\Big)^{1/2} \Delta N^{-1/2} \to 0$, *where*

$$\psi(u,\alpha) = 2\Phi\Big[\Big(-\frac{\log\alpha}{2}\Big)^{1/2}(2u-1)(u(1-u))^{-1/2}\Big] - 1,$$

$0 < \alpha < 1$ *and* Φ *is the standard normal distribution function.*

Proof. It follows from (3) and from [4, Theorem VI.3.2]. The last assertion follows from [4, Theorem VI.4.5]. □

Remark 1 (2-way ANOVA model). For example, in two-way layout, we can use this two-sample K-S test, similarly to e.g. the Friedman test, for comparing two treatments applied on I blocks. The effects of the blocks would represent the nuisance regression here.

2.3. Cramér-von Mises type test

Similarly to the Kolmogorov-Smirnov test, we can generalize also the Cramér-von Mises type two-sample test for nuisance regression.

We first look at the location model. With the same notation as in Subsection 2.1, we put again $\mathbf{c}_N = \mathbf{x}^{(p)} = (1,\ldots,1,0,\ldots,0)'$, and for $T_N(t)$ from (4) we have that the classical Cramér-von Mises two-sample test statistic M equals (see [4, III.1.3.11 and V.3.8])

$$M = \frac{1}{mn}\sum_{j=1}^{N-1}\Big[\frac{jm}{N} - (V_1 + \ldots + V_j)\Big]^2 = \int_0^1 T_N^2(t)\,\mathrm{d}t + \frac{1}{6N}.$$

In model (1) (for nuisance linear regression) the test criterion of the Cramér-von Mises type two-sample test is then $\int_0^1 S_N^2(t)\,\mathrm{d}t$, where $S_N(t)$ is the process from Theorem 1, and it can be seen from the form of the test statistic that this test with a critical region $\{\int_0^1 S_N^2(t)\,\mathrm{d}t \geqslant C\}$ is suitable only for two-sided alternatives (similarly to the statistic K_N). The critical values can be obtained from the following theorem.

Theorem 3. *Under the conditions of Theorem* 1 *we have*

$$\lim_{N\to\infty} P\Big(\int_0^1 S_N^2(t)\,\mathrm{d}t < x\Big) = P\Big(\sum_{j=1}^{\infty}\frac{X_j^2}{j^2\pi^2} < x\Big),$$

where $X_1, X_2, \ldots$ *are independent standardized normal random variables.*

Proof. It follows from Theorem 1 and from the property of the Brownian bridge stated in [4, Theorem V.3.3.c]. □

All the tests proposed in this paper are based on the regression rank scores and their construction is inspired by the structure of the classical Kolmogorov-Smirnov (Cramér-von Mises) test. Therefore, they do not require a preliminary estimation of the nuisance parameter and their asymptotic distributions coincide with the classical tests.

Acknowledgement. The author would like to thank his supervisor Prof. RNDr. Jana Jurečková, DrSc. for helpful discussion.

References

[1] *C. Gutenbrunner, J. Jurečková*: Regression rank scores and regression quantiles. Ann. Stat. *20* (1992), 305–330.
[2] *C. Gutenbrunner, J. Jurečková, R. Koenker, S. Portnoy*: Tests of linear hypotheses based on regression rank scores. J. Nonparametric Stat. *2* (1993), 307–331.
[3] *J. Hájek*: Extension of the Kolmogorov-Smirnov test to regression alternatives. Proc. Bernoulli-Bayes-Laplace Seminar (L. LeCam, ed.). Univ. of California Press, 1965, pp. 45–60.
[4] *J. Hájek, Z. Šidák*: Theory of Rank Tests. Academia, Praha, 1967.
[5] *J. Jurečková*: Tests of Kolmogorov-Smirnov type based on regression rank scores. Information Theory, Stastistical Decision Functions, Random Processes. Trans. 11th Prague Conf., Prague 1990, Vol. B (J. Á. Víšek, ed.). Academia & Kluwer, Praha & Dordrecht, 1992, pp. 41–49.
[6] *R. Koenker, G. Bassett (1978)*: Regression quantiles. Econometrica *46* (1978), 33–50.

Author's address: *M. Schindler*, Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Statistics, Sokolovská 83, 186 75 Prague 8, Czech Republic, e-mail: `schindle@karlin.mff.cuni.cz`, and Technical University in Liberec, Studentská 2, 461 17 Liberec 1, Czech Republic.