KYBERNETIKA — VOLUME *51* (2015), NUMBER 6, PAGES 909–922

RANK TESTS IN REGRESSION MODEL BASED ON MINIMUM DISTANCE ESTIMATES

Radim Navrátil

In this paper a new rank test in a linear regression model is introduced. The test statistic is based on a certain minimum distance estimator, however, unlike classical rank tests in regression it is not a simple linear rank statistic. Its exact distribution under the null hypothesis is derived, and further, the asymptotic distribution both under the null hypothesis and the local alternative is investigated. It is shown that the proposed test is applicable in measurement error models. Finally, a simulation study is conducted to show a good performance of the test. It has, in some situations, a greater power than the widely used Wilcoxon rank test.

Keywords: measurement errors, minimum distance estimates, rank tests

Classification: 62J05, 62G10

1. INTRODUCTION

Consider the following model of a regression line

$$Y_i = \beta_0 + x_i\beta + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where β_0 and β are unknown parameters, $x_1, \ldots, x_n$ are fixed or stochastic regressors, the model errors $e_1, \ldots, e_n$ are assumed to be i.i.d. with an unknown distribution function F and a uniformly continuous density f independent from $x_1, \ldots, x_n$ (if they are random). Our aim is to test the hypothesis

$$\mathbf{H}_0 : \ \beta = 0 \quad \text{against} \quad \mathbf{K}_0 : \ \beta \neq 0.$$

Since F is unknown, we should use nonparametric tests. Among them, rank tests play an important role. They use, instead of the original response variables Y_i's, their ranks. Rank tests form a class of statistical procedures which have the advantage of simplicity combined with a surprising power.

Modern development of rank tests began in the 1930's, see e. g. [8] and [12]. In 1945 Wilcoxon [24] introduced a popular Wilcoxon test for comparing two treatments. At first, it was believed that a high price in loss of efficiency when using rank tests has to be paid. However, it turned out that the efficiency of the rank tests behaves quite

DOI: 10.14736/kyb-2015-6-0909

well under the classical normality assumption. In addition, these tests remain valid and have a high efficiency when the normality assumption is not satisfied. These facts were first described by Pitman [19]. At the moment, rank tests remain still very popular and widely used, see [7] and [15].

Let us briefly show the classical approach based on a linear rank statistic (see e. g. [11]). Denote

$$Q_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad \text{with} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Let R_i be the rank of Y_i among $Y_1, \ldots, Y_n$ and let us define a simple linear rank statistic

$$S_n = n^{-1/2}\sum_{i=1}^{n}(x_i - \overline{x})\varphi\left(\frac{R_i}{n+1}\right)$$

for a nondecreasing, nonconstant, square integrable score function $\varphi : (0,1) \mapsto \mathbb{R}$. The test criterion for $\mathbf{H}_0$ is then

$$T_n^2 = \frac{S_n^2}{A^2(\varphi)Q_n}, \tag{2}$$

where

$$A^2(\varphi) = \int_0^1 (\varphi(t) - \overline{\varphi})^2 \,\mathrm{d}t, \quad \overline{\varphi} = \int_0^1 \varphi(t)\,\mathrm{d}t.$$

Under $\mathbf{H}_0$, T_n^2 has asymptotically (under very mild conditions) χ^2 distribution with 1 degree of freedom.

The regression model (1) assumes that the regressors x_i are observed accurately, but in practise this is often not satisfied. We will, therefore, consider a measurement error model

$$\begin{aligned} Y_i &= \beta_0 + \beta x_i + e_i, \\ w_i &= x_i + v_i, \quad i = 1, \ldots, n, \end{aligned} \tag{3}$$

where we observe w_i instead of x_i that are affected by additive measurement errors v_i which are i.i.d. independent from e_i and x_i.

The influence of measurement errors on parameter estimates was first considered by Adcock at the end of the nineteenth century. Adcock [1] showed that in the regression line model with measurement errors the least squares estimate of the slope is downward in magnitude. Since then a lot of methods for dealing with measurement errors have been developed, e. g. the method of moments (see [6, 20]), the maximum likelihood method (see [17]), the total least squares method ([10]). There are even several books devoted entirely to measurement error models, see e. g. [3, 4, 5] and [9].

Most of the methods use parametric approach with its restrictive normality assumptions or with some additional information about the error distribution. However, this is not our case, we will introduce a class of new rank tests applicable in the measurement error model (3) without any further information about the errors.

Jurečková [14] was the first who introduced rank tests into measurement error models. She showed that the test (2) in the model (3) remains valid even if measurement errors are present, they only cause a decrease in the test power. This result was further extended by Navrátil [18] and Jurečková [13] for various other measurement error models.

2. TEST IN SIMPLE LINEAR REGRESSION

Koul [16] considered a class of estimates in a linear regression model based on the minimization of a certain type of distances. He also showed that such estimates might have, in some situations, a greater efficiency than the corresponding R-estimates. This fact motivated us to introduce a class of test statistics based on Cramér–von Mises type of distance involving various weighted empirical processes and investigate their power. We will also pay attention to their robust properties.

Let us define

$$T_{g,n}(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(x_i) \mathbb{I}\{R_i \le ns\}, \quad 0 \le s \le 1, \tag{4}$$

$$K^*_{g,n} = \int_0^1 T^2_{g,n}(s) \, \mathrm{d}L(s), \tag{5}$$

where R_i is the rank of Y_i among $Y_1, \ldots, Y_n$, L a distribution function on $[0,1]$ and g a real (weight) function, such that $\sum_{i=1}^n g(x_i) = 0$.

Let us discuss some computation aspects of (5). First, let us look at the formula (5) for $K^*_{g,n}$. Inserting (4) into (5) we obtain

$$\begin{aligned} K^*_{g,n} &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} g(x_i) g(x_j) \int_0^1 \mathbb{I}\{R_i \le ns\} \mathbb{I}\{R_j \le ns\} \, \mathrm{d}L(s) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} g(x_i) g(x_j) \int_{\max\left\{\frac{R_i}{n}, \frac{R_j}{n}\right\}}^1 1 \, \mathrm{d}L(s). \end{aligned}$$

Since L is a distribution function, then $L(\max\{a,b\}) = \max\{L(a), L(b)\}$. This also remains true for the limits from the left:

$$K^*_{g,n} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} g(x_i) g(x_j) \left(1 - \max\left\{L\left(\frac{R_i}{n}-\right), L\left(\frac{R_j}{n}-\right)\right\}\right).$$

Since $\sum_{i=1}^n g(x_i) = 0$, we get

$$K^*_{g,n} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} g(x_i) g(x_j) \max\left\{L\left(\frac{R_i}{n}-\right), L\left(\frac{R_j}{n}-\right)\right\}.$$

Using the fact that

$$2\max\{a,b\} = a + b + |a-b|, \quad \forall\, a, b \in \mathbb{R}$$

and $\sum_{i=1}^n g(x_i) = 0$ we have

$$K^*_{g,n} = -\frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} g(x_i) g(x_j) \left| L\left(\frac{R_i}{n}-\right) - L\left(\frac{R_j}{n}-\right) \right|,$$

which is much more convenient for practical computations.

Under $\mathbf{H}_0$ ($\beta = 0$), the model (1) is reduced to

$$Y_i = \beta_0 + e_i, \quad i = 1, \ldots, n. \tag{6}$$

Since the distribution of model errors e_i is absolutely continuous, there cannot be any ties in ranks with probability 1. Thanks to the invariance of ranks with respect to the location, the distribution of $R_1, \ldots, R_n$ under the null hypothesis is uniform over all $n!$ permutations of numbers $\{1, \ldots, n\}$. Therefore, the distribution of $K^*_{g,n}$ given $x_1, \ldots, x_n$, under $\mathbf{H}_0$, is distribution-free and may be even computed directly. To do so, we have to compute all values of the test statistic $K^*_{g,n}$ for each of $n!$ permutations of numbers $\{1, \ldots, n\}$ and order these values in the increasing magnitude. The critical region is then formed by $M = \lfloor \alpha n! \rfloor$ largest values. The combination which leads to the $(M+1)$-st largest value can be possibly randomized.

However, the computation of exact (conditional) distribution may be time consuming for large sample size n. We will, therefore, investigate the asymptotic distribution of $K^*_{g,n}$. We have to distinguish two cases: random or fixed regressors x_i. We will present only the first one because it is often overlooked and because the assumptions and proofs of asymptotic distributions are analogous for fixed regressors.

For $s \in [0, 1]$, let us define empirical processes

$$V_{g,n}(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(x_i) \mathbb{I}\{e_i \le F_n^{-1}(s)\},$$

$$\widehat{V}_{g,n}(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(x_i) \mathbb{I}\{e_i \le F^{-1}(s)\},$$

where $F_n(s) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{e_i \le s)\}$ is an empirical distribution function.

Now, let us state the assumptions needed for proving the asymptotic properties of $K^*_{g,n}$:

$$\sum_{i=1}^{n} (x_i - \bar{x}) > 0 \text{ a.s.} \quad \forall\, n > 1, \tag{7}$$

$$\max_{i=1,\ldots,n} \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \xrightarrow{p} 0, \tag{8}$$

$$g(x_i) \neq 0 \text{ a.s. for some} \quad i = 1, \ldots, n, \tag{9}$$

$$0 < |\mathbb{E} g(X_1)(X_1 - \mathbb{E} X_1)| < \infty, \tag{10}$$

$$x_i g(x_i) \ge 0 \text{ a.s.} \quad \forall i = 1, \ldots, n \text{ or } x_i g(x_i) \le 0 \text{ a.s.} \quad \forall i = 1, \ldots, n, \tag{11}$$

$$\max_{i=1,\ldots,n} g^2(x_i) \xrightarrow{p} 0, \tag{12}$$

$$\sup_{n \in \mathbb{N}} \max_{i=1,\ldots,n} |g(x_i)| \le c \text{ a.s.} \quad \text{for some} \quad 0 < c < \infty, \tag{13}$$

$$0 < \gamma = \sqrt{\mathbb{E} g^2(X_1)} < \infty. \tag{14}$$

Lemma 2.1. Let us assume that (7) – (10) hold, then under $\mathbf{H}_0$

$$\Big|K^*_{g,n} - \int V^2_{g,n}(s)\,\mathrm{d}L(s)\Big| = o_p(1), \quad \text{as} \quad n \to \infty.$$

P r o o f. For the convenience, we will drop off index g in $K^*_{g,n}$ and $V_{g,n}$. Adding and subtracting $V_n(s)$ in the first integrand and using the Cauchy–Schwarz inequality we get

$$\begin{aligned}
&\Big|\int T_n^2(s)\,\mathrm{d}L(s) - \int V_n^2(s)\,\mathrm{d}L(s)\Big| \\
&= \Big|\int [T_n(s) - V_n(s)]^2\ \mathrm{d}L(s) + 2\int V_n(s)(T_n(s) - V_n(s))\,\mathrm{d}L(s)\Big| \\
&\le \sup_{0\le s\le 1} |T_n(s) - V_n(s)|^2 + 2\sqrt{\int V_n^2(s)\,\mathrm{d}L(s)\int (T_n(s) - V_n(s))^2\,\mathrm{d}L(s)}.
\end{aligned}$$

The fact that

$$\sup_{0\le s\le 1} |T_n(s) - V_n(s)| \le 2 \max_{i=1,\dots,n} |g(x_i)| = o_p(1)$$

together with $\int V_n^2(s)\,\mathrm{d}L(s) = O_p(1)$ proves the Lemma. □

Lemma 2.2. Let us assume that (7) – (10) hold, then under $\mathbf{H}_0$

$$\Big|K^*_{g,n} - \int \widehat{V}^2_{g,n}(s)\,\mathrm{d}L(s)\Big| = o_p(1), \quad \text{as} \quad n \to \infty.$$

P r o o f.

$$\begin{aligned}
&\Big|\int T_n^2(s)\,\mathrm{d}L(s) - \int \widehat{V}_n^2(s)\,\mathrm{d}L(s)\Big| \\
&= \Big|\int \Big[T_n(s) - \widehat{V}_n(s)\Big]^2\ \mathrm{d}L(s) + 2\int \widehat{V}_n(s)(T_n(s) - \widehat{V}_n(s))\,\mathrm{d}L(s)\Big|. \qquad (15)
\end{aligned}$$

Using the Minkowski inequality

$$\begin{aligned}
&\int [T_n(s) - \widehat{V}_n(s)]^2\,\mathrm{d}L(s) = \int [T_n(s) - V_n(s) + V_n(s) - \widehat{V}_n(s)]^2\,\mathrm{d}L(s) \\
&\le 2\int [T_n(s) - V_n(s)]^2\,\mathrm{d}L(s) + 2\int [V_n(s) - \widehat{V}_n(s)]^2\,\mathrm{d}L(s). \qquad (16)
\end{aligned}$$

By the Cauchy–Schwarz inequality

$$\begin{aligned}
\Big|\int \widehat{V}_n(s)(T_n(s) - \widehat{V}_n(s))\,\mathrm{d}L(s)\Big| &\le \sqrt{\int \widehat{V}_n^2(s)\,\mathrm{d}L(s)\int [T_n(s) - \widehat{V}_n(s)]^2\,\mathrm{d}L(s)} \\
&= o_p(1), \qquad (17)
\end{aligned}$$

because $\int \widehat{V}_n^2(s)\,\mathrm{d}L(s) = O_p(1)$ and $\int [T_n(s) - \widehat{V}_n(s)]^2\,\mathrm{d}L(s) = o_p(1)$.

Observe that

$$\widehat{V}_n(FF_n^{-1}(s)) = \sum_{i=1}^{n} g(x_i)\mathbb{I}\{e_i \le F^{-1}FF_n^{-1}(s)\} = \sum_{i=1}^{n} g(x_i)\mathbb{I}\{e_i \le F_n^{-1}(s)\} = V_n(s).$$

Therefore

$$\sup_{0\le s\le 1} |V_n(s) - \widehat{V}_n(s)| = \sup_{0\le s\le 1} |\widehat{V}_n(FF_n^{-1}(s)) - \widehat{V}_n(s)| = o_p(1),$$

because

$$\sup_{0\le s\le 1} |FF_n^{-1}(s) - s| = \sup_{0\le s\le 1} |FF^{-1}(s) - F_nF_n^{-1}(s) + F_nF_n^{-1}(s) - s|$$

$$\le \sup_{x\in\mathbb{R}} |F(x) - F_n(x)| + \sup_{0\le s\le 1} |F_nF_n^{-1}(s) - s| = o_p(1).$$

Now, combining the previous result, Lemma 2.1 and (15), (16) and (17) we have proven the Lemma. □

Remark 2.3. The previous lemma states that the asymptotic distribution of $K_{g,n}^*$ is the same as of $\int \widehat{V}_{g,n}^2(s)\,\mathrm{d}L(s)$, which is easier to investigate. Now, we are able to state the theorem about the asymptotic null distribution of $K_{g,n}^*$.

Theorem 2.4. Let us assume that (7) – (14) hold. Then in the model (1), under $\mathbf{H}_0$,

$$K_{g,n}^* \xrightarrow{d} \gamma^2 \cdot Y_L, \quad \text{with} \quad Y_L = \int_0^1 B^2(s)\,\mathrm{d}L(s),$$

where $B(s)$ is a Brownian bridge in $\mathcal{C}[0,1]$.

P r o o f . Recall that

$$\begin{aligned}\widehat{V}_{g,n}(s) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\mathbb{I}\{e_i \le F^{-1}(s)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\mathbb{I}\{F(e_i) \le s\} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\mathbb{I}\{U_i \le (s)\},\end{aligned}$$

where $U_1, \ldots, U_n$ are i.i.d. random variables with the uniform $\mathcal{U}(0,1)$ distribution.

By [16] we have

$$\widehat{V}_{g,n}(s) \Rightarrow \gamma \cdot B(s) \quad \text{in} \quad \mathcal{D}[0,1]$$

and therefore $\int \widehat{V}_{g,n}^2(s)\,\mathrm{d}L(s) \xrightarrow{d} \gamma^2 \int B^2(s)\,\mathrm{d}L(s)$. That, together with Lemma 2.2, proves the Theorem. □

The distribution of the random variable Y_L for $L(s) = s$ was first investigated by Smirnov [21]. The values of its distribution function may be found for example in [2] or in [22] and [23], some quantiles are listed in Table 1. One has to use simulated values for other choices of the function L.

$(1-\alpha)$	0.90	0.95	0.99	0.999
$(1-\alpha)$-quantile	0.34730	0.46136	0.74346	1.16786

Tab. 1. Quantiles of the distribution of Y_L for $L(s) = s$.

Now, we will investigate behavior of $K^*_{g,n}$ under the local alternative

$$\mathbf{K}_{0,n}: \ \beta = n^{-1/2}\beta^*, \quad 0 \neq \beta^* \in \mathbb{R} \text{ fixed.}$$

For $t \in \mathbb{R}$, let us define

$$K^*_{g,n}(t) = \int_0^1 \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n g(x_i) \mathbb{I}\{R_{i,t} \leq ns\} \right)^2 \mathrm{d}L(s), \tag{18}$$

$$\widehat{K}^*_{g,n}(t)$$
$$= \int_0^1 \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n g(x_i)) \mathbb{I}\{U_i \leq s\} - \frac{t}{\sqrt{n}} \sum_{i=1}^n g(x_i)(x_i - \bar{x}) f(F^{-1}(s)) \right)^2 \mathrm{d}L(s), \tag{19}$$

where $R_{i,t}$ is the rank of $Y_i - x_i t$ among $Y_1 - x_1 t, \ldots, Y_n - x_n t$ and $U_1, \ldots, U_n$ are i.i.d. random variables with uniform $\mathcal{U}(0,1)$ distribution.

Remark 2.5. Koul [16] defined an estimator of β in the model (1) as a minimizer of (18) with respect to $t \in \mathbb{R}$. Hence, the proposed test statistic $K^*_{g,n}$ is $K^*_{g,n}(t)$ computed in the hypothetical value $t = 0$, i. e. $K^*_{g,n} = K^*_{g,n}(0)$ is the test statistic under $\mathbf{H}_0$, while $K^*_{g,n}(n^{-1/2}\beta^*)$ is the test statistic under $\mathbf{K}_{0,n}$.

Lemma 2.6. Let us assume that (7) – (10) hold. Then for every $0 < b < \infty$

$$\sup_{|u| \leq b} |K^*_{g,n}(n^{-1/2}u) - \widehat{K}^*_{g,n}(n^{-1/2}u)| = o_p(1), \quad \text{as} \quad n \to \infty.$$

P r o o f. See [16, Theorem 5.5.5]. □

Remark 2.7. Particularly, if $u = 0$ in Lemma 2.6, we get $K^*_{g,n}(0) = \widehat{K}^*_{g,n}(0) + o_p(1)$, i. e.

$$K^*_{g,n} = \int_0^1 \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n g(x_i)) \mathbb{I}\{U_i \leq s\} \right)^2 \mathrm{d}L(s) + o_p(1).$$

Lemma 2.2 is then a special case of Lemma 2.6.

Now, let (7) – (14) be satisfied. Rewrite (19) as

$$\widehat{K}^*_{g,n}(t) = \widehat{K}^*_{g,n}(0) - \frac{2t}{n}\sum_{j=1}^{n} g(x_j)(x_j-\bar{x})\sum_{i=1}^{n} g(x_i)\int_0^1 \mathbb{I}\{U_i \le s\} f(F^{-1}(s))\,\mathrm{d}L(s)$$

$$+ \frac{t^2}{n}\left(\sum_{i=1}^{n} g(x_i)(x_i-\bar{x})\right)^2 \int_0^1 f^2(F^{-1}(s))\,\mathrm{d}L(s)$$

$$= \widehat{K}^*_{g,n}(0) + \frac{2t}{n}\sum_{j=1}^{n} g(x_j)(x_j-\bar{x})\sum_{i=1}^{n} g(x_i)\varphi(U_i) + \frac{t^2}{n}\sigma_{f,L}\left(\sum_{i=1}^{n} g(x_i)(x_i-\bar{x})\right)^2,$$

where $\varphi(u) = \int_0^u f(F^{-1}(s))\,\mathrm{d}L(s)$ and $\sigma_{f,L} = \int_0^1 f^2(F^{-1}(s))\,\mathrm{d}L(s)$. From Lemma 2.6, (18) and (19) we finally get

$$\begin{aligned} K^*_{g,n}(n^{-1/2}\beta^*) &= K^*_{g,n}(0) + 2\beta^*\frac{1}{n}\sum_{j=1}^{n} g(x_j)(x_j-\bar{x})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\varphi(U_i) \\ &+ (\beta^*)^2\sigma_{f,L}\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i)(x_i-\bar{x})\right)^2 + o_p(1). \end{aligned} \tag{20}$$

The right-hand side of (20) converges to the convolution of two (dependent) random variables $\gamma^2 \cdot Y_L$ and $Z \sim \mathcal{N}(a,b)$, where

$$\begin{aligned} a &= (\beta^*)^2\sigma_{f,L}(\mathbb{E}\{g(X_1)\cdot(X_1-\mathbb{E}X_1)\})^2, \\ b &= 4(\beta^*)^2\left[\mathbb{E}\{g(X_1)\cdot(X_1-\mathbb{E}X_1)\}\right]^2\mathbb{E}g^2(X_1)\operatorname{var}\varphi(U_i). \end{aligned}$$

Hence, under (7) – (14) the asymptotic distribution of $K^*_{g,n}$ under the local alternative $\mathbf{K}_{0,n}$ is the above convolution of dependent random variables.

As far as practical applications are concerned, there arises a natural question how to choose the functions g and L. The function g is in fact a weight function, so it can downweight outlying observations (regressors) to robustify our test against the extreme values of x_i (if g is bounded for example). The function L has a similar interpretation as the score-function φ in the standard rank test theory. The optimal L could be chosen based on the estimate of unknown model errors. The simplest choice $L(s) = s$ provides very reasonable results (see the simulations).

3. EXTENSIONS OF THE TEST

Now, let us return to the problem with measurement errors in the model (3). We would like to use our test although the original regressors are not observable. We apply our test based on the observed regressors w_i, denote the corresponding test statistic $K^*_{w,n}$ and show that the test works: under $\mathbf{H}_0$ ($\beta = 0$), the measurement error model (3) is reduced to the model (6) – the same model as in the case without measurement errors. Hence, the exact distribution (given $w_1, \ldots, w_n$) of $K^*_{w,n}$, under $\mathbf{H}_0$, might be derived in

the same way as in the model without measurement errors. Using the same arguments, if (7) – (14) hold for w_i, then the asymptotic null distribution of $K^*_{w,n}$ is the same as in the model without measurement errors.

The previous ideas might be summarized in the following theorem.

Theorem 3.1. Let the conditions (7) – (14) for w_i's be satisfied. Then in the measurement error model (3) under $\mathbf{H}_0$

$$K^*_{w,n} \xrightarrow{d} \gamma^2 \cdot Y_L, \quad \text{with} \quad \gamma = \sqrt{\mathbb{E}g^2(W_1)} \quad \text{and} \quad Y_L = \int_0^1 B^2(s)\,\mathrm{d}L(s).$$

Remark 3.2. The presence of measurement errors decreases the power of our test because we do not use the values of the function g in the optimal points $x_1, \ldots, x_n$ but in w_i's.

Now, let us show the extension of our test into a multiple regression model

$$Y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n, \tag{21}$$

where $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are unknown parameters, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are fixed or stochastic vectors of regressors, the model errors $e_1, \ldots, e_n$ are assumed to be i.i.d. with an unknown distribution function F and a uniformly continuous density f independent from $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (if they are random). Now, we test the hypothesis

$$\mathbf{H}_0 : \boldsymbol{\beta} = 0 \quad \text{against} \quad \mathbf{K}_0 : \boldsymbol{\beta} \neq 0.$$

We will introduce the test statistic $K^*_{g,n}$ into the multiple regression model (21). Let us define

$$\begin{aligned} T^j_{g,n}(s) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_j(x_{i,j}) \mathbb{I}\{R_i \leq ns\}, \quad 0 \leq s \leq 1, \\ \mathbf{T}_{g,n}(s) &= (T^1_{g,n}(s), \ldots, T^p_{g,n}(s))^\top, \\ K^*_{g,n} &= \int_0^1 \mathbf{T}^\top_{g,n}(s) \mathbf{T}_{g,n}(s)\,\mathrm{d}L(s), \end{aligned} \tag{22}$$

where R_i is the rank of Y_i among $Y_1, \ldots, Y_n$, L a distribution function on $[0,1]$ and $\mathbf{g} = (g_1, \ldots, g_p) : \mathbb{R}^p \mapsto \mathbb{R}$ a (weight) function, such that $\sum_{i=1}^n g_j(x_{i,j}) = 0$ and $\sum_{i=1}^n g_j^2(x_{i,j}) = 1$ for all $j = 1, \ldots p$.

Remark 3.3. Similarly as in Section 2, the formula (22) might be simplified for practical computations. Based on the permutation principle its exact null distribution might be derived. A detailed analysis of $K^*_{g,n}$ in the multiple regression model (21) will be part of our future study.

4. SIMULATIONS

To support the previous theoretical results, we conducted a large simulation study. Let us present several interesting results.

Let us start with the model (1) without measurement errors for a moderate sample size $n = 30$. We have compared the empirical power of our test based on the test statistic $K^*_{g,n}$ with $g(x_i) = x_i - \bar{x}$ and $L(s) = s$ (call it *the minimum distance test*) with the Wilcoxon test for regression (based on (2) with $\varphi(u) = u$) and the standard t-test for regression.

The regressors $x_1, \ldots, x_{30}$ were generated from the uniform $\mathcal{U}(-2, 10)$ distribution, the model errors e_i were generated from the normal, logistic, Laplace and t-distribution with 6 degrees of freedom, respectively, always with 0 mean and variance $3/2$. The empirical powers of the tests were computed as a percentage of rejections of $\mathbf{H}_0$ among 10 000 replications, at the significance level $\alpha = 0.05$. The results are summarized in Table 2.

$\beta \setminus e_i$	$\mathcal{N}\left(0, \frac{3}{2}\right)$			$Log\left(0, \frac{\sqrt{2}\pi}{3}\right)$			$Lap\left(0, \frac{\sqrt{3}}{2}\right)$			$t(6)$		
test	MD	W	t	MD	W	t	MD	W	t	MD	W	t
0	4.98	4.42	5.00	5.06	4.55	5.00	5.00	4.55	5.04	5.00	4.32	4.93
0.1	28.7	28.3	31.5	32.7	31.4	32.0	42.4	39.0	33.5	34.6	33.1	32.9
−0.1	28.3	28.2	30.9	32.7	31.2	32.2	42.5	39.0	33.7	33.3	32.1	31.9
0.2	78.2	78.8	82.3	82.5	81.8	81.9	88.3	86.6	82.0	84.5	83.9	82.6
−0.2	78.3	78.7	82.9	83.3	82.7	82.9	89.2	87.5	83.1	84.0	83.4	82.5

Tab. 2. The percentage of rejections of the hypothesis $\mathbf{H}_0 : \beta = 0$ of the minimum distance test (MD), the Wilcoxon test for regression (W) and the t-test for regression (t); $n = 30$.

The t-test achieves (not surprisingly) the largest power for normal model errors, but the differences among the three tests are not very distinct. For the distributions with heavier tails than normal, our test has the largest power, even for the logistic distribution (for which the Wilcoxon test is locally most powerful rank test). It is caused by the slow convergence of the Wilcoxon test statistic to its asymptotic distribution.

Now, let us compare the three previous tests in the measurement error model (3) – under the same simulation design as before. The empirical errors of the first kind for various measurement errors are summarized in Table 3. The empirical powers for various measurement errors are summarized in Table 4 (with the true value of parameter $\beta = 0.2$).

According to Table 3, the minimum distance test preserves the error of the first kind at the prescribed α even if measurement errors are present. The presence of measurement errors decreases the power of all tests – the larger variance of measurement errors, the smaller power.

$v_i \setminus e_i$	$\mathcal{N}\left(0, \frac{3}{2}\right)$			$Log\left(0, \frac{\sqrt{2}\pi}{3}\right)$			$Lap\left(0, \frac{\sqrt{3}}{2}\right)$			$t(6)$		
	MD	W	t	MD	W	t	MD	W	t	MD	W	t
0	4.98	4.42	5.00	5.06	4.55	5.00	5.00	4.55	5.04	5.00	4.32	4.93
$\mathcal{N}(0,4)$	4.45	3.94	4.39	4.90	4.29	5.03	4.90	4.29	5.04	4.99	4.66	4.92
$\mathcal{N}(0,6)$	4.53	3.97	4.44	4.81	4.41	5.05	4.81	4.41	5.06	4.77	4.59	4.95
$\mathcal{U}(-\sqrt{18}, \sqrt{18})$	5.49	4.78	5.36	5.13	4.53	4.97	5.13	4.53	4.81	4.51	3.85	4.34
$2t(6)$	5.09	4.63	5.04	5.11	4.59	4.94	5.11	4.59	4.96	5.17	4.51	4.81
$\mathcal{U}(-6,6)$	5.50	4.73	5.42	5.18	4.62	5.12	5.18	4.62	4.85	4.87	4.19	4.55

Tab. 3. The percentage of rejections of the hypothesis $\mathbf{H}_0: \beta = 0$ of the minimum distance test (MD), the Wilcoxon test for regression (W) and the t-test for regression (t); true $\beta = 0$, $n = 30$.

$v_i \setminus e_i$	$\mathcal{N}\left(0, \frac{3}{2}\right)$			$Log\left(0, \frac{\sqrt{2}\pi}{3}\right)$			$Lap\left(0, \frac{\sqrt{3}}{2}\right)$			$t(6)$		
	MD	W	t	MD	W	t	MD	W	t	MD	W	t
0	78.2	78.8	82.3	82.5	81.8	81.9	88.3	86.6	82.0	84.5	83.9	82.6
$\mathcal{N}(0,4)$	64.1	63.8	68.2	69.1	67.9	68.2	76.4	74.0	68.8	71.5	70.4	69.6
$\mathcal{N}(0,6)$	58.1	57.8	61.7	63.0	62.2	62.6	71.0	68.0	63.5	65.9	64.7	63.5
$\mathcal{U}(-\sqrt{18}, \sqrt{18})$	58.4	58.4	62.5	62.7	61.9	62.8	70.7	67.8	63.8	65.8	64.3	63.6
$2t(6)$	59.0	59.0	62.5	63.8	62.5	63.0	71.4	68.6	64.1	66.8	65.8	64.6
$\mathcal{U}(-6,6)$	45.2	44.7	48.0	49.8	48.4	49.4	57.0	54.1	50.3	51.6	50.2	49.6

Tab. 4. The percentage of rejections of the hypothesis $\mathbf{H}_0: \beta = 0$ of the minimum distance test (MD), the Wilcoxon test for regression (W) and the t-test for regression (t); true $\beta = 0.2$, $n = 30$.

We performed more simulations for various regressors x_i (both random and fixed), sample sizes n and the model errors e_i. We also compared the tests according to the choice of the functions L and g. However, the corresponding results are very similar to those presented in Tables 2–4.

Remark 4.1. Koul [16] derived a formula for the asymptotic variance of the minimum distance estimator from Remark 2.5 with $g(x_i) = x_i - \bar{x}$ and $L(s) = s$. It is given by

$$\sigma_0^2 = \frac{\int\int [F(x \wedge y) - F(x)F(y)] f^2(x) f^2(y)\, \mathrm{d}x\mathrm{d}y}{\left(\int f^3(x)\, \mathrm{d}x\right)^2}.$$

We may compare the minimum distance estimator with the Wilcoxon and the least squares estimate via the relative asymptotic efficiency (ARE) for various distributions of the model errors, see Table 5.

Pitman [19] generalized asymptotic relative efficiencies for tests; we tried to arrive at similar formulae for the minimum distance tests. However, in Pitman's definition

F	$ARE(MD, W)$	$ARE(MD, LSE)$
Laplace	1.667	1.309
Logistic	0.988	1.034
Normal	0.957	0.914
Cauchy	1.278	∞

Tab. 5. The asymptotic relative efficiencies of the minimum distance estimate (MD), the Wilcoxon estimate (W) and the least squares estimate (LSE) for various model errors.

the asymptotic distribution of the test statistic needs to be χ^2, which is not our case. In addition, we did not obtain a closed formula for the asymptotic distribution of our test statistic under Pitman's alternative.

However, we compared the empirical powers of the three tests. The simulation results are in accordance with Table 5 which compares variances of the corresponding estimates. The minimum distance test has not only a slightly greater power than the Wilcoxon test or the t-test for some model errors, but it also converges faster to its asymptotic distribution than the Wilcoxon test (see Table 4).

CONCLUSION

In this paper we proposed a new rank test for hypothesis testing in a simple linear regression model. The test statistic is based on minimum distance estimates (Cramér–von Mises distance) and unlike the classical rank tests (such as Wilcoxon, or van der Waerden) it is not a linear function of the ranks.

We derived its exact null distribution and asymptotic distribution under both the null and alternative hypotheses. In the simulation study we showed a good performance of the test. It achieves a greater power than the Wilcoxon test for the distribution of model errors with heavy tails. Moreover, it converges faster to its asymptotic distribution than the Wilcoxon test.

Our test is neither sensitive to leverage observations, nor outliers and has robust properties. It might be also used in measurement error models, the errors cause only a decrease of its power. The extension into a multiple regression model is straightforward, a detailed analysis will be included in our further study.

ACKNOWLEDGEMENT

This paper is based on a part of the author's dissertation thesis at Charles University in Prague, the paper originated during the author's visit at Michigan State University, East Lansing, MI, USA with a collaboration with Professor Hira L. Koul. The author's research was supported by Student Project Grant at MU (specific research, rector's programme) MUNI/A/1441/2014.

(Received January 12, 2015)

REFERENCES

[1] R. J. Adcock: Note on the method of least squares. The Analyst *4* (1877), 183–184. DOI:10.2307/2635777

[2] T. Anderson and D. Darling: Asymptotic theory of certain goodness of fit criteria based on stochastic processes. Ann. Math. Statist. *23* (1952), 193–212. DOI:10.1214/aoms/1177729437

[3] J. P. Buonaccorsi: Measurement Error Models, Methods and Applications. Chapman and Hall/CRC, Boca Raton 2010. DOI:10.1201/9781420066586

[4] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu: Measurement Error in Nonlinear Models: A Modern Perspective. Chapman and Hall/CRC, Boca Raton 2006. DOI:10.1201/9781420010138

[5] C. L. Cheng and J. W. van Ness: Statistical Regression with Measurement Error. Kendalls Library of Statistics 6. Arnold, London 1999. DOI:10.1002/1097-0258(20000815)19:15¡2077::aid-sim500¿3.0.co;2-7

[6] E. F. Drion: Estimation of the parameters of a straight line and of the variances of the variables, if they are both subject to error. Indagationes Math. *13* (1951), 256–260. DOI:10.1016/s1385-7258(51)50036-7

[7] L. Feng, C. Zou, and Z. Wang: Rank-based inference for the single-index model. Statist. Probab. Lett. *82* (2012), 535–541. DOI:10.1016/j.spl.2011.11.025

[8] M. Friedman: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Amer. Statist. Assoc. *32* (1937), 675–701. DOI:10.1080/01621459.1937.10503522

[9] W. A. Fuller: Measurement Error Models. John Wiley and Sons, New York 1987. DOI:10.1002/jae.3950030407

[10] G. H. Golub and C. F. van Loan: An analysis of the total least squares problem. SIAM J. Numer. Anal. *17* (1980), 883–893. DOI:10.1137/0717073

[11] J. Hájek, Z. Šidák, and P. K. Sen: Theory of Rank Tests. Second Edition. Academic Press, New York 1999. DOI:10.1016/b978-012642350-1/50020-5

[12] H. Hotteling and M. R. Pabst: Rank correlation and tests of significance involving no assumptions of normality. Ann. Math. Statist. *7* (1936), 29–43. DOI:10.1214/aoms/1177732543

[13] J. Jurečková, H. L. Koul, R. Navrátil, and J. Picek: Behavior of R-estimators under measurement errors. To apper in Bernoulli.

[14] J. Jurečková, J. Picek, and A. K. Md. E. Saleh: Rank tests and regression rank score tests in measurement error models. Comput. Statist. Data Anal. *54* (2010), 3108–3120. DOI:10.1016/j.csda.2009.08.020

[15] J. Jurečková, P. K. Sen, and J. Picek: Metodological Tools in Robust and Nonparametric Statistics. Chapman and Hall/CRC Press, Boca Raton, London 2013.

[16] H. L. Koul: Weighted Empirical Processes in Dynamic Nonlinear Models. Springer, New York 2002. DOI:10.1007/978-1-4613-0055-7

[17] D. V. Lindley: Regression lines and the linear functional relationship. Suppl. J. Roy. Statist. Soc. *9* (1947), 218–244. DOI:10.2307/2984115

[18] R. Navrátil and A. K. Md. E. Saleh: Rank tests of symmetry and R-estimation of location parameter under measurement errors. Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica *50* (2011), 95–102.

[19] E. J. G. Pitman: Lecture Notes on Nonparametric Statistics. Columbia University, New York 1948.

[20] E. L. Scott: Note on consistent estimates of the linear structural relation between two variables. Anal. Math. Stat. *21* (1950), 284–288. DOI:10.2307/2984115

[21] N. V. Smirnov: Sur la distribution de ω^2 (criterium de m. r. v. mises). C. R. Akad. Sci. Paris *202* (1936), 449–452.

[22] L. Tolmatz: On the distribution of the square integral of the brownian bridge. The Annals of Probab. *30* (2002), 253–269. DOI:10.1214/aop/1020107767

[23] L. Tolmatz: Addenda: On the distribution of the square integral of the brownian bridge. The Annals of Probab. *31* (2003), 530–532.

[24] F. Wilcoxon: Individual comparisons by ranking methods. Biometrics *1* (1945), 80–83. DOI:10.2307/3001968

Radim Navrátil, Department of Mathematics and Statistics, Masaryk University, Kotlářská 2, Brno. Czech Republic.

e-mail: navratil@math.muni.cz