Rank score test for independence with nuisance nonlinear regression

Martin Schindler

Technical University of Liberec & Charles University in Prague, Czech Republic

email: martin.schindler@karlin.mff.cuni.cz

SUMMARY: **We construct a class of tests for independence with nuisance nonlinear regression. The test statistic is based on the nonlinear regression rank scores. We illustrate the proposed test on the "Scottish hill races data - 2000" that gives record-winning times (male and female) in year 2000, distance and climb for 77 Scottish long distance races. We test the hypothesis of independence of the record-winning times and the record-winning times for females taking into account the influence of distance and climb on the record times (the form of the influence is given by nonlinear models).**

Introduction

We construct a class of tests for independence with nuisance nonlinear regression. This class is a generalization of a class of nonparametric tests for independence with nuisance linear regression introduced by Picek in [1].

Definition of nonlinear regression rank scores

Consider the following nonlinear regression model

$$Y_i = g(\mathbf{x}_i, \boldsymbol{\theta}) + e_i = \theta_0 + g^*(\mathbf{x}_i, \boldsymbol{\theta}^*) + e_i, \qquad i = 1, \dots, n \tag{1}$$

$\mathbf{x}_i \in \mathbf{R}_+^q \subset \mathcal{X}$ are given vectors. $e_1, \dots, e_n$ are i.i.d. errors with a positive but generally unknown density f and $\mathbf{Y} = (Y_1, \dots, Y_n)'$ is a vector of observations.

We define the nonlinear regression rank scores as a vector $\hat{\mathbf{a}}_n(\alpha) = \hat{\mathbf{a}}(\alpha) = (\hat{a}_{n1}(\alpha), \dots, \hat{a}_{nn}(\alpha))'$ solving

$$\mathbf{Y}'\hat{\mathbf{a}}_n(\alpha) := \max \tag{2}$$

$$\mathbf{V}_n'(\hat{\boldsymbol{\theta}}_{n\alpha})\hat{\mathbf{a}}_n(\alpha) = (1-\alpha)\mathbf{V}_n'(\hat{\boldsymbol{\theta}}_{n\alpha})\mathbf{1}_n \tag{3}$$

$$\hat{\mathbf{a}}_n(\alpha) \in [0,1]^n, \quad 0 < \alpha < 1. \tag{4}$$

$$\hat{a}_{ni}(\alpha) = \begin{cases} 1 & \text{if } Y_i > g(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{n\alpha}), \\ \\ 0 & \text{if } Y_i < g(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{n\alpha}), \end{cases} \tag{5}$$

where

$$\mathbf{V}_n(\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\theta}) = \left[\left[\frac{\partial g(\mathbf{x}_i, \boldsymbol{\theta} + \boldsymbol{\delta})}{\partial \delta_j}\right]_{\boldsymbol{\delta}=\mathbf{0}}\right]_{i=1,\dots,n}^{j=0,\dots,p}, \tag{6}$$

$$\hat{\boldsymbol{\theta}}_{n\alpha} = arg \min_{t \in \mathbf{R}^{p+1}} \sum_{i=1}^{n} \rho_\alpha(Y_i - g(\mathbf{x}_i, t)) \tag{7}$$

is the nonlinear α-th regression quantile and

$$\rho_\alpha(u) = |u|\{(1-\alpha)I[u<0] + \alpha I[u>0]\}, \quad u \in \mathbf{R},\ \alpha \in (0,1).$$

Construction of the test statistic

Suppose we have the pairs of observations $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ that follow the nonlinear regression models

$$X_i = g(\mathbf{c}_i, \boldsymbol{\theta}) + e_i, \qquad i = 1, \dots, n \tag{8}$$

and

$$Y_i = \tilde{g}(\mathbf{d}_i, \boldsymbol{\vartheta}) + \delta_i, \qquad i = 1, \dots, n, \tag{9}$$

where $\mathbf{X} = (X_1, \dots, X_n)'$, $\mathbf{Y} = (Y_1, \dots, Y_n)'$ are vectors of observations, g and $\tilde{g}$ are of known shape and $\mathbf{c}_i \in \mathbf{R}_+^q$, $\mathbf{d}_i \in \mathbf{R}_+^{\tilde{q}}$, $i = 1, \dots, n$ are given vectors. Further, $\boldsymbol{\theta} = (\theta_0, \dots, \theta_p)'$, $\boldsymbol{\vartheta} = (\vartheta_0, \dots, \vartheta_{\tilde{p}})'$ are unknown parameters. Finally, $e_1, \dots, e_n$ are i.i.d. errors with a positive but generally unknown density f and $\delta_1, \dots, \delta_n$ are i.i.d. errors with a positive but generally unknown density $\tilde{f}$.

We will propose a test of the hypothesis H_0 of independence between $\mathbf{X}$ and $\mathbf{Y}$. The test will be based on the nonlinear regression rank scores.

Let $\hat{\mathbf{a}}_n^X(\alpha) = (\hat{a}_{n1}^X(\alpha), \dots, \hat{a}_{nn}^X(\alpha))'$ denote the regression rank scores (2)–(5) in the nonlinear model (8). Similarly, $\hat{\mathbf{a}}_n^Y(\alpha) = (\hat{a}_{n1}^Y(\alpha), \dots, \hat{a}_{nn}^Y(\alpha))'$ denote the regression rank scores (2)–(5) in the nonlinear model (9), for $0 \le \alpha \le 1$.

Further, we choose a nondecreasing, bounded score function $\varphi(u)$ such that

$$\varphi(1-u) = -\varphi(u), \qquad 0 < u < 1. \tag{10}$$

For fixed $\varepsilon \in (0, \frac{1}{2})$ we define $\varphi_\varepsilon(u)$ to be the $\varphi(u)$ truncated on both ends of the interval $(0,1)$ to facilitate the asymptotic distribution of S_n in (13) and assume that

$$0 < A^2(\varphi_\varepsilon) = \int_0^1 (\varphi_\varepsilon(u) - \bar{\varphi}_\varepsilon)^2\, du < \infty, \qquad \bar{\varphi}_\varepsilon = \int_0^1 \varphi_\varepsilon(u)\, du. \tag{11}$$

We use the Wilcoxon scores $\varphi(u) = u - 1/2$ and calculate the scores generated by φ_ε

$$\hat{b}_{ni}^X = \int_0^1 \hat{a}_{ni}^X(\alpha)\, d\varphi_\varepsilon(\alpha) \quad \text{and} \quad \hat{b}_{ni}^Y = \int_0^1 \hat{a}_{ni}^Y(\alpha)\, d\varphi_\varepsilon(\alpha), \qquad i = 1, \dots, n \tag{12}$$

in the model (8) and the model (9).

We will use the test criterion

$$S_n = \frac{1}{\sqrt{n}}(A(\varphi_\varepsilon))^{-2} \sum_{i=1}^{n} \hat{b}_{ni}^X \hat{b}_{ni}^Y \tag{13}$$

which is, under the hypothesis of independence

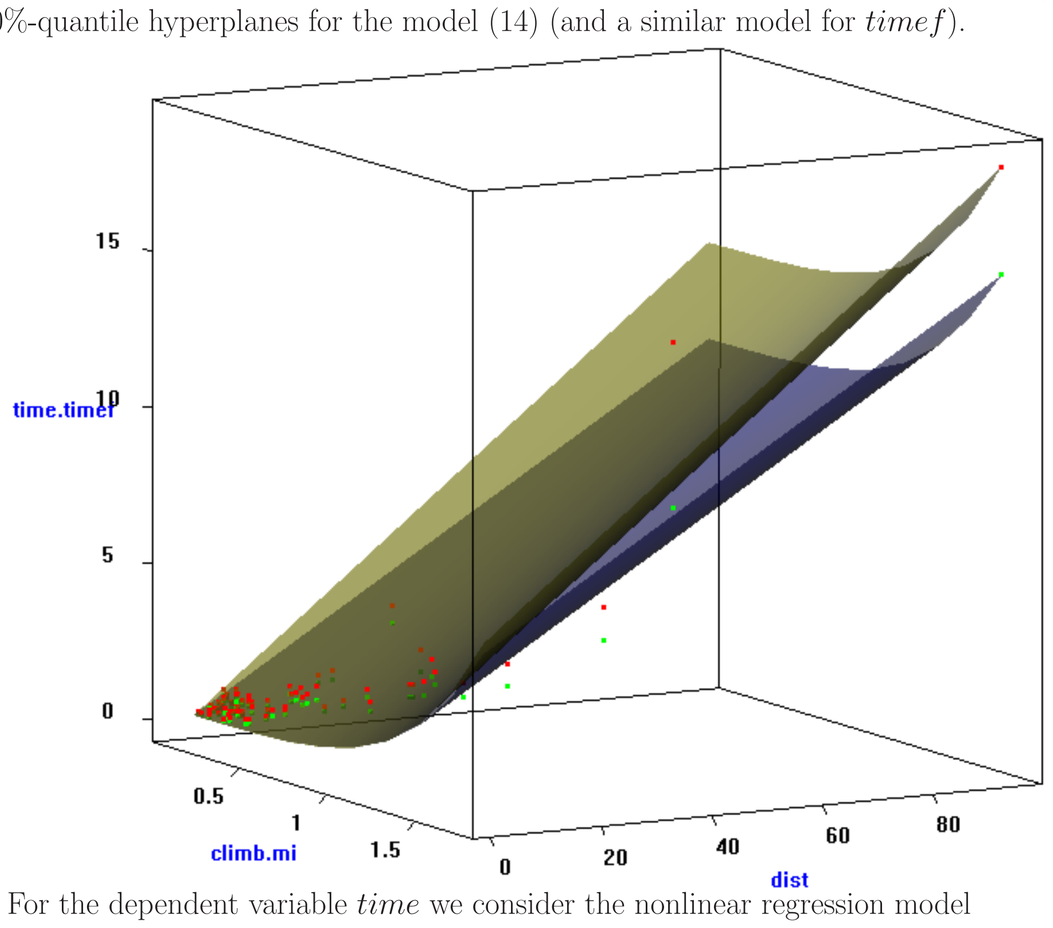
$$H_0 : \mathbf{X}, \mathbf{Y} \quad \text{independent},$$

asymptotically standard normal under some other regularity assumptions (see [2] for the details).

Application to data

Now we will deal with the data set called *races2000* from package *DAAG* in software *R*. It contains record-winning times in year 2000 for 77 Scottish long distance races. For every race it gives its name, its distance in miles (*dist*), the climb (*climb*) (total height gained during the route) in feet, the record time in hours (*time*) and the record time in hours for females (*timef*). We will, however, use the climb in miles ($climb.mi = climb/5280$) instead of *climb* in our analysis.

Now we will try to use the test for independence with nuisance nonlinear regression. We apply it to the record-winning times (*time*) and record-winning times for females (*timef*) and test H_0, the hypothesis of independence of *time* and *timef*. However, we have to exclude two races from the future analysis, since the record times for females are not available for them, so now the number of observations is $n = 75$. On the figure below a scatterplot of both *time* and *timef* with *dist* and *climb.mi* is shown together with the nonlinear regression 50%-quantile hyperplanes for the model (14) (and a similar model for *timef*).



For the dependent variable *time* we consider the nonlinear regression model

$$time = a + \beta \cdot dist + \gamma \cdot (climb.mi)^\delta + \varepsilon. \tag{14}$$

Considering the Wilcoxon score function φ, we compute the scores $\hat{b}_{ni}^M$ defined in (12) for this model.

Similarly, we will assume that the same model (with the same regressors but different parameters) holds for the variable *timef*. We denote the scores from this model by $\hat{b}_{ni}^F$.

Then we calculate the test statistic (13) which can be simplified to

$$S_n \doteq \frac{12}{\sqrt{n}} \sum_{i=1}^{n} \hat{b}_{ni}^M \hat{b}_{ni}^F$$

and is asymptotically standard normal under H_0.

The approximate values of the test statistic for H_0 are:

$$S_n^{1000} = 7.613 \qquad S_n^{100} = 7.612 \qquad S_n^{10} = 7.409,$$

where the superscripts stand for the number of subintervals we cut the interval $(0,1)$ to numerically approximate the scores (integrals) in (12). This statistic strongly rejects the hypothesis of independence of the record times and the record times for females. This test takes into account the influence (the form of the influence is given by the nonlinear models (14) for *time* and *timef*) of the variables *dist* and *climb.mi* on the record times. The dependence (confirmed by the test) can be caused by an influence of one of the dependent variables on the other one (or a mutual influence) or the dependence is a result of a common unknown (that is not included in the regression model) factor that affects both these variables. Not only distance and climb, but also elevation difference between start and finish, terrain, the time of the year, how long the race is run or some other factors, can affect the record times.

Thus, there are probably many other factors that can influence the record times, but are not included in the regression model. That is probably the reason why the test statistic is so large.

Acknowledgement: The author would like to thank his supervisor Prof. RNDr Jana Jurečková, DrSc. The present work was supported by Research Project LC06024.

References.

[**1**] Picek J.(1999): Nonparametric tests of independence with nuisance regression. *Proceedings of the 14th International Workshop on Statistical Modelling*, (H. Friedl e.a., ed.), pp. 616-619, Graz, Austria.

[**2**] Schindler M.(2008): Inference based on regression rank scores. *Doctoral thesis*, Charles University, Prague. (to appear)