# Nonparametric Specification Testing with SpeTestNP

Hippolyte Boucher

2022-09-22

## Contents



# 1 Introduction

In applied work in order to evaluate the effect of a set of exogenous variables on an outcome it is very common to estimate a parametric model such as the linear model with ordinary least squares (OLS). But such parametric specifications may not capture the true relationship between outcome and exogenous variables. In fact if the chosen parametric model is a bad approximation of the true model then counterfactual analysis will be flawed. For this reason in the past forty years a literature on specification tests has developed in order to know if a parametric specification is right or wrong. `SpeTestNP` is a package which implements heteroskedasticity-robust specification tests of parametric models from Bierens (1982), Zheng (1996), Escanciano (2006), Lavergne and Patilea (2008), and Lavergne and Patilea (2012).

The author, creator and maintainer of `SpeTestNP` is Hipppolyte Boucher and Pascal Lavergne is a contributor. This vignette describes the principle behind each test available in `SpeTestNP`, then how to use `SpeTestNP` to test a parametric specification in practice with an application to the wage returns on education and age.

# 2 Testing for a parametric specification

In order to present the specification tests available in `SpeTestNP` we first describe the model being considered and define the null and alternative hypothesis, second we highlight the principle behind each test, third we derive the test statistics and their rejection rules, and fourth we briefly discuss and compare the tests size and power performances.

## 2.1 Model

Consider a sample $(y_j, x_j')_{j=1}^n$ of independent observations with $y_j$ the scalar outcome and $x_j$ a $k \times 1$ vector of exogenous explanatory variables. Then as long as $\mathbb{E}(|y_j|) < +\infty$ there exists some Borel-measurable regression function $g(\cdot)$ such that $g(x_j) = \mathbb{E}(y_j|x_j)$ $a.s$. That is the true model linking $y_j$ and $x_j$ writes

$$y_j = g(x_j) + \varepsilon_j, \qquad \mathbb{E}(\varepsilon_j|x_j) = 0 \quad a.s$$

for $j = 1, 2, \ldots, n$ and where $\varepsilon_j$ denotes the part of $y_j$ which is unexplained by $x_j$ in terms of the mean. But instead in practice some parametric model characterized by a parametric family of functions $\mathcal{F} = \{f(\cdot, \tilde{\theta}) : \tilde{\theta} \in \Theta \subset \mathbb{R}^k\}$ is considered

$$y_j = f(x_j, \theta) + u_j$$

where $\theta = \underset{\tilde{\theta}\in\Theta}{Argmin}\ \mathbb{E}((y_j - f(x_j, \tilde{\theta}))^2)$ is the parameter which yields the best mean square error fit for this parametric model, and where $u_j$ is the error induced by this parametric model. A typical estimator of $\theta$ is the non-linear least squares (NLS) estimator denoted by $\hat{\theta}$, thus when $\mathcal{F}$ is the family of linear functions then $\hat{\theta}$ is the OLS estimator. Next notice that if $g(\cdot) \in \mathcal{F}$ then $\mathbb{E}(u_j|x_j) = 0$ $a.s$ or equivalently $\mathbb{E}(y_j|x_j) = f(x_j, \theta)$. Indeed if $g(\cdot) \in \mathcal{F}$ then by properties of projections

$$g(\cdot) = \underset{\tilde{g}}{Argmin}\ \mathbb{E}((y_j - \tilde{g}(x_j))^2) = \underset{\tilde{g}\in\mathcal{F}}{Argmin}\ \mathbb{E}((y_j - \tilde{g}(x_j))^2) = \underset{\tilde{\theta}\in\Theta}{Argmin}\ \mathbb{E}((y_j - f(x_j, \tilde{\theta}))^2) = f(\cdot, \theta)$$

Consequently when modelling the true relationship between $y$ and $x$ with a parametric model, the implicit null hypothesis is

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s$$

And the alternative hypothesis is

$$H_1 : \mathbb{P}(\mathbb{E}(u_j|x_j) = 0) < 1$$

Equivalently the null and alternative hypothesis write

$$H_0 : g(x_j) = f(x_j, \theta) \quad a.s, \qquad H_1 : \mathbb{P}(g(x_j) = f(x_j, \theta)) < 1$$

## 2.2 Tests principle

Next to construct specification tests the null hypothesis is reformulated into moments conditions from which statistics can be derived. The five reformulations of the null hypothesis are in order.

### 2.2.1 Bierens (1982)

Bierens (1982) proves that the conditional moment condition of the null hypothesis is equivalent to an infinite number of moment conditions which is equivalent to an integrated conditional moment condition

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \ \Leftrightarrow \mathbb{E}(u_j exp(i\beta' x_j)) = 0 \ \ \forall \beta \in \mathbb{R}^k \Leftrightarrow \int_{\mathbb{R}^k} |\mathbb{E}(u_j exp(i\beta' x_j))|^2 \, d\mu(\beta) = 0$$

where $\mu(\cdot)$ is any positive almost everywhere measure, $|\cdot|$ denotes the modulus, and $i$ is the imaginary unit.

#### 2.2.2 Zheng (1996)

Instead Zheng (1996) finds an equivalence between the conditional moment condition and an unconditional one

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \;\Leftrightarrow \mathbb{E}(u_j\mathbb{E}(u_j|x_j)f(x_j)) = 0$$

where $f(\cdot)$ denotes the probability density function of $x_j$.

#### 2.2.3 Escanciano (2006)

Escanciano (2006) proves the equivalence between the null hypothesis, an infinite number of moment conditions which differ from Bierens (1982), and an integrated moment condition

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \;\Leftrightarrow \mathbb{E}(u_j 1\{\beta' x_j \leqslant l\}) = 0 \;\; \forall (t,l) \in \mathbb{S}^k \times \mathbb{R} \Leftrightarrow \int_{\mathbb{S}^k \times \mathbb{R}} \mathbb{E}^2(u_j 1\{\beta' x_j \leqslant l\}) f_\beta(u) d\beta dl = 0$$

where $1\{\cdot\}$ denotes the indicator function, $\mathbb{S}^k = \{t \in \mathbb{R}^k : |t| = 1\}$ denotes the unit sphere, and $f_\beta(\cdot)$ denotes the probability density function of $\beta' x_j$.

#### 2.2.4 Lavergne and Patilea (2008)

Lavergne and Patilea (2008) show that the null hypothesis is equivalent to an infinite number of unconditional moment conditions

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \;\Leftrightarrow \max_{||\beta||=1} \mathbb{E}(u_j\mathbb{E}(u_j|\beta' x_j)\omega(\beta' x_j)) = 0$$

for any $\omega(\cdot)$ such that $\forall \beta \in \mathbb{R}^k$, $\omega(\beta' x_j) > 0$ on the support of $\mathbb{E}(u_j|\beta' x_j)$. This condition resembles that of Zheng (1996) with $\beta' x_j$ replacing $x_j$ in an effort to remove the curse of dimensionality.

#### 2.2.5 Lavergne and Patilea (2012)

Finally Lavergne and Patilea (2012) prove the equivalence between the null and an integrated moment condition

$$H_0 : \mathbb{E}(u_j|x_j) = 0 \quad a.s \;\Leftrightarrow \int_B \mathbb{E}(\mathbb{E}^2(u_j|\beta' x_j) f_\beta(\beta' x_j)) d\beta = 0$$

where $B \subseteq \mathbb{S}^k$ and $f_\beta(\cdot)$ denotes the density of $\beta' x_j$. This moment condition combines the integrated moments approaches of Bierens (1982) and Escanciano (2006) and the dimension reduction devise used in Lavergne and Patilea (2008).

### 2.3 Test statistics

Each test relies on reformulating the null hypothesis into a moment condition for which an empirical counterpart exist. Thus the test statistics are sample analogs of the moments defining the null hypothesis, possibly multiplied by the sample size in order to obtain variation at the limit, and possibly standardized. Denote by $\hat{\theta}$ a consistent estimator of $\theta$ and let $\hat{u}_j = y_j - f(x_j, \hat{\theta})$ denote the residual for individual $j$. The five test statistics are derived in order.

### 2.3.1 Bierens (1982)

An empirical counterpart of the integrated conditional moment $\int_{\mathbb{R}^k} |\mathbb{E}(u_j exp(i\beta' x_j))|^2 \, d\mu(\beta)$ of Bierens (1982) is

$$T_{icm} = \int_{\mathbb{R}^k} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \hat{u}_j exp(i\beta' x_j) \right|^2 d\mu(\beta)$$

with some positive almost everywhere measure $\mu(\cdot)$. Using properties of the modulus and of the Fourier transform it can then be shown that

$$T_{icm} = \frac{1}{n} \sum_{j,j'} \hat{u}_j \hat{u}_{j'} K(x_j - x_{j'}) = \frac{1}{n} \hat{u}' W_{icm} \hat{u}$$

where $K(\cdot)$ is the Fourier transform of $\mu(\cdot)$, $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_n)'$ is the $n \times 1$ vector of stacked residuals, and $W_{icm}$ is the matrix with entries $K(x_j - x_{j'})$ for any row $j$ and column $j'$. Although this statistic can be used as is, $\mu(\cdot)$ is typically assumed to be a symmetric probability measure which is strictly positive almost everywhere. This simplifies the asymptotic theory and the derivation of the test statistic in practice. Indeed as a consequence the Fourier transform of $\mu(\cdot)$ denoted as $K(\cdot)$ is a symmetric bounded density. Hence candidates for $K(\cdot)$ include logistic, triangular, normal, student, or Cauchy densities, see Johnson, Kotz and Balakrishnan (1995, section 23.3) and Dreier and Kotz (2002). Furthermore to control for scale, we impose that either the integral of $K(\cdot)$ to the square equals one or that the distribution associated to $K(\cdot)$ has variance one.

To decide whether to reject or not the null hypothesis we need to compute quantiles of the distribution under the null of $T_{icm}$ conditional on $x \equiv= (x_1, \ldots, x_n)'$. To do so, $x$ is held fixed (and therefore $W$ is held fixed) and a $n \times 1$ vector of residuals $\hat{u}_b$ is drawn using the wild bootstrap of Wu (1986) or the smooth conditional moment bootstrap of Gozalo (1997) in order to control for potential heteroskedasticity. Using this bootstrapped vector of residuals a bootstrapped statistic can be computed

$$T_{icm,b} = \frac{1}{n} \hat{u}_b' W_{icm} \hat{u}_b$$

By repeating this operation $B$ times we obtain $B$ bootstrapped statistics $(T_{icm,b})_{b=1}^{B}$ which mimic the behavior of $T_{icm}$ under the null hypothesis. Consequently the parametric specification will be rejected at level 5% if $T_{icm} > q_{95\%}$ where $q_{95\%}$ is the 95% quantile of $(T_{icm,b})_{b=1}^{B}$.

### 2.3.2 Zheng (1996)

Zheng (1996) test statistic is the sample analog of $\mathbb{E}(u_j \mathbb{E}(u_j|x_j) f(x_j))$ which is derived by estimating both the density $f(\cdot)$ of $x_j$ and the conditional mean $\mathbb{E}(u_j|x_j = \cdot)$ with Kernels. For any $\tilde{x} \in \mathbb{R}^k$ define

$$\hat{f}(\tilde{x}) = \frac{1}{nh^k} \sum_j K\left(\frac{\tilde{x} - x_j}{h}\right), \qquad \hat{\mathbb{E}}(u_j|x_j = \tilde{x}) = \frac{1}{nh^k} \sum_j \frac{u_j}{\hat{f}(\tilde{x})} K\left(\frac{\tilde{x} - x_j}{h}\right)$$

where $K(\cdot)$ is a Kernel function which is nonnegative, symmetric, bounded, continuous and which integrates to one, and $h$ a bandwidth such that $h \underset{n \to +\infty}{\to} 0$ and $nh^k n \to +\infty \to +\infty$. Then the sample analog $\hat{V}$ of the moment $\mathbb{E}(u_j \mathbb{E}(u_j|x_j) f(x_j))$ and an estimator $\hat{\Sigma}$ of its variance can be defined

$$\hat{V} = \frac{1}{n(n-1)h^k} \sum_{j,j' \neq j} \hat{u}_j \hat{u}_{j'} K\left(\frac{x_j - x_{j'}}{h}\right), \qquad \hat{\Sigma} = \frac{2}{n(n-1)h^k} \sum_{j,j' \neq j} \hat{u}_j^2 \hat{u}_{j'}^2 K^2\left(\frac{x_j - x_{j'}}{h}\right)$$

Then the test statistic is

$$T_{zheng} = \frac{\sqrt{nh^k}\hat{V}}{\hat{\Sigma}} = \frac{\sqrt{nh^k}\hat{u}'W_{zheng}\hat{u}}{\hat{\Sigma}}$$

where $W_{zheng}$ is a matrix whose diagonal elements are equal to zero and its other entries are equal to $K\left(\frac{x_j - x_{j'}}{h}\right)$ for any row $j$ any column $j'$ such that $j \neq j'$. This statistic is pivotal and converge towards the standard normal distribution under the null as proven in Zheng (1996). Sill we can use the wild bootstrap or smooth conditional moment bootstrap to redraw residuals, obtain bootstrapped statistics, and approximate the behavior under the null of $T_{zheng}. In fact using the bootstrap yields a test with better performances which is why the p-value we implement for the Zheng test is also obtained from the bootstrap.

### 2.3.3 Escanciano (2006)

Escanciano (2006) test statistic is the sample analog of $\int_{\mathbb{S}^k \times \mathbb{R}} \mathbb{E}^2(u_j 1\{\beta' x_j \leqslant l\}) d\beta dl$ times $n$ which is derived by approximating the density $f_\beta(\cdot)$ by a probability mass function. Let $\hat{f}_t(l) = \frac{1}{n}\sum_r 1\{\beta' x_r = l\}$ then the statistic is

$$T_{esca} = \int_{\mathbb{S}^k \times \mathbb{R}} \left( \frac{1}{\sqrt{n}} \sum_j \hat{u}_j 1\{\beta' x_j \leqslant l\} \right)^2 \hat{f}_t(l) d\beta dl$$

It can be proven that it has the same form as the other test statistics

$$T_{esca} = \frac{1}{n}\sum_{j,j'} \hat{u}_j \hat{u}_{j'} \frac{1}{n} \sum_r \int_{\mathbb{S}^k} 1\{\beta' x_j \leqslant \beta' x_r, \beta' x_{j'} \leqslant \beta' x_r\} d\beta = \frac{1}{n}\hat{u}' W_{esca} \hat{u}$$

where $W_{esca}$ has elements $\frac{1}{n}\sum_r W_{esca,j,j',r}$ with $W_{esca,j,j',r} = \int_{\mathbb{S}^k} 1\{\beta' x_j \leqslant \beta' x_r, \beta' x_{j'} \leqslant \beta' x_r\} d\beta$ for any row $j$ and column $j'$. Approximating the integrals in $W_{esca}$ is unecessary because

$$W_{esca,j,j',r} = W^{(0)}_{esca,j,j',r} \frac{\pi^{k/2} - 1}{\Gamma(k/2+1)}, \qquad W^{(0)}_{esca,j,j',r} = \left| \pi - arcos\left( \frac{(x_j - x_r)'(x_{j'} - x_r)}{|x_j - x_r||x_{j'} - x_r|} \right) \right|$$

See appendix B in Escanciano (2006) for more details. Again the wild bootstrap and smooth conditional moment bootstrap can be used to resample the residuals and to obtain a vector of bootstrapped statistics from which the quantiles will determine the cutoff point for $T_{esca}$ for the rejection of the null.

### 2.3.4 Lavergne and Patilea (2008)

Lavergne and Patilea (2008) consider a sample analog of the moment $\mathbb{E}(u_j \mathbb{E}(u_j|x_j)\omega(\beta' x_j))$ with $\omega(\cdot) = f_\beta(\cdot)$ the density of $\beta' x_j$ then its maximum in $t$ in the unit hypersphere. Thus first define for any $t \in \mathbb{S}^k$

$$\mathcal{Q}(\beta) = \frac{1}{n(n-1)h} \sum_{j,j' \neq j} \hat{u}_j \hat{u}_{j'} K\left( \frac{\beta'(x_j - x_{j'})}{h} \right)$$

where $K(\cdot)$ is a bounded symmetric density with bounded variation, $h$ is a bandwidth such that $h \underset{n\to+\infty}{\to} 0$ and $(nh^2)^\alpha / log(n) \underset{n\to+\infty}{\to} +\infty$ for some $\alpha \in (0;1)$. But instead of directly considering $\underset{\beta\in\mathbb{S}^k}{max}\, n\sqrt{h}\mathcal{Q}(\beta)$ as the test statistic, the authors further reduce the dimension problem and normalise. Let

$$v^2(\beta) = \frac{2}{n(n-1)h} \sum_{j,j' \neq j} \hat{\sigma}^2(x_j)\hat{\sigma}^2(x_{j'}) K^2 \left( \frac{\beta'(x_j - x_{j'})}{h} \right)$$

where $\hat{\sigma}(\cdot)$ is a consistent estimator of the conditional variance of $y_j$. We implement the conditional variance estimator of Yin, Geng, Li and Wang (2010) which writes

$$\hat{\sigma}(\tilde{x}) = \frac{\frac{1}{nh_v} \sum_j (y_j - \bar{y}(\tilde{x}))^2 K_v \left( \frac{\tilde{x} - x_j}{h_v} \right)}{\frac{1}{nh_v} \sum_j K_v \left( \frac{\tilde{x} - x_j}{h_v} \right)}, \qquad \bar{y}(\tilde{x}) = \frac{\frac{1}{nh_v} \sum_j y_j K_v \left( \frac{\tilde{x} - x_j}{h_v} \right)}{\frac{1}{nh_v} \sum_j K_v \left( \frac{\tilde{x} - x_j}{h_v} \right)}$$

where $K_v$ is a Kernel function and $h_v$ a bandwidth which can be different from $K$ and $h$. Next define $\hat{\beta} = \underset{\beta \in \mathbb{S}^k}{Argmax} \, |n\sqrt{h}\mathcal{Q}(\beta) - \alpha 1\{\beta \neq \beta^*\}|$ where $\alpha \underset{n \to +\infty}{\to} 0$. $\beta^*$ is a user chosen favorite direction for capturing the information of $x_j$ thus to compute $\hat{\beta}$ candidates which are far from $\beta^*$ are penalized. Consequently the test statistic is

$$T_{pala} = \frac{n\sqrt{h}\mathcal{Q}(\hat{\beta})}{v(\hat{\beta})} = \frac{n\sqrt{h}}{v(\hat{\beta})} \hat{u}' W_{pala} \hat{u}$$

where $W_{pala}$ is a matrix with diagonal elements equal to zero and its other entries equal to $\frac{1}{n(n-1)h} K \left( \frac{\hat{\beta}'(x_j - x_{j'})}{h} \right)$ for any row $j$ and column $j'$ such that $j \neq j'$. Once again the test decision will be based on the statistic boostrapped distribution.

#### 2.3.5 Lavergne and Patilea (2012)

Finally Lavergne and Patilea (2012) use the sample analog of $\int_B \mathbb{E}(\mathbb{E}^2(u_j|\beta' x_j) f_\beta(\beta' x_j)) d\beta = 0$ for some $B \subseteq \mathbb{S}^k$ as a test statistic. To derive it notice that an empirical counterpart of $\mathbb{E}(\mathbb{E}^2(u_j|\beta' x_j) f_\beta(\beta' x_j))$ is $\mathcal{Q}(\beta)$ as defined in the previous subsection. Hence their test statistic which they call smooth integrated conditional moment statistic writes

$$T_{sicm} = \frac{n\sqrt{h} \int_B \mathcal{Q}(\beta) d\beta}{\int_B v(\beta) d\beta} = \frac{n\sqrt{h}}{\int_B v(\beta) d\beta} \hat{u}' W_{sicm} \hat{u}$$

where $v(\beta)$ is defined as in the previous subsection and $W_{sicm}$ has diagonal elements equal to zero and its other elements are equal to $\frac{1}{n(n-1)h} \int_B K \left( \frac{\beta'(x_j - x_{j'})}{h} \right)$ for any row $j$ and any column $j' \neq j$. Clearly $T_{sicm}$ is a smooth version of $T_{icm}$ because of the bandwidth $h$. Furthermore it is also a smooth version of $T_{pala}$ in the sense that instead of being based on the squared error in the worst direction of $\beta' x_j$, it is based on a continuum of directions. In practice to compute the integral a number `nbeta` of points are drawn randomly from the unit (half) hypersphere. Similarly to the other test, the decision to reject or not the null hypothesis is based on the statistic bootstrapped distribution.

### 2.4 Validity, consistency and power properties

Each test can be proven to be valid, as in under the null hypothesis the probability to reject the null converges to nominal level, and to be consistent, as in under any fixed alternative the probability to reject the null converges to one.

But these five tests differ significantly in terms of power in practice. The test of Bierens (1982) seem to be the least powerful test in practice, under the alternative it diverges at a slow rate and has difficulty rejecting the null when the number $k$ of exogenous variables is large. By construction the test of Zheng (1996) diverges at a faster rate than Bierens (1982) under the alternative but it has trouble rejecting the null when $k$ is

large. The test of Escanciano (2006) does not depend on a choice of weighting function and does not require numerical integration however to derive its statistic it requires $n^3$ operations making it very slow and hard to apply in practice. In addition its power however largely depends on the true alternative and is low when $k$ is large. The tests of Lavergne and Patilea (2008), and Lavergne and Patilea (2012) are more powerful than the other two when $k$ is large because of their use of a continuum of single index $\beta' x_j$ to summarize the correlation between $u_j$ and $x_j$. At the same time when $k$ is small the two tests are at least as powerful as the others. As mentioned the power of Lavergne and Patilea (2008) test comes from the "worst" single-index alternative whereas the power of Lavergne and Patilea (2012) test comes from a continuum of single-index alternatives. Thus in practice under the alternative the nature of the correlation between $u_j$ and $x_j$ will determine which of these two tests is more powerful.

See the references for more details.

# 3 Using `SpeTestNP`

Previously we have described the principle behind the five nonparametric specification tests, how to derive the test statistics and the rejection rules, and discussed their properties. Next we show how to use `SpeTestNP` to test parametric models in practice, with first the installation, second a description of how to use the test, third a thorough description of the arguments of the package main function `SpeTest`, and fourth an application to determine the true shape of expected wages conditional on years of education and age.

## 3.1 Installation

`SpeTestNP` can be downloaded and installed using the package `devtools` by running the following lines of code:

```
install.packages("devtools")

library("devtools")

install_github("HippolyteBoucher/SpeTestNP")
```

To choose where and how the package is installed check `help(install_github)` and `help(install.packages)`. Alternatively users can download the package and directly install it with the CMD. `SpeTestNP` will soon be available on CRAN. `SpeTestNP` requires the packages `stats` (already installed and loaded by default in Rstudio), `foreach`, `parallel` and `doParallel` (if parallel computing is used to generate the vector) to be installed.

## 3.2 Testing with `SpeTestNP`

Recall the true model and the model induced by the parametric specification characterized by $\mathcal{F} = \{f(\cdot, \tilde{\theta} : \theta \in \Theta \subset \mathbb{R}^k\}$

$$y_j = g(x_j) + \varepsilon_j, \qquad y_j = f(x_j, \theta) + u_j$$

where $\mathbb{E}(y_j|x_j) = g(x_j)$ $a.s$ and $\theta = \underset{\tilde{\theta}\in\Theta}{Argmin}\ \mathbb{E}((y_j - f(x_j, \tilde{\theta}))^2)$.

Then to test the parametric specification or equivalently to test $H_0 : \mathbb{E}(u_j|x_j) = 0$ $a.s$ the function `SpeTest` of the package `SpeTestNP` can be directly used by filling the first argument `eq` with a fitted model of class `lm` or `nls`. In case the parametric specification is linear or can be rewritten in a linear form `eq` should be an object of class `lm`. In case of non-linear models `eq` should be an object of class `nls` which stands for non-linear least squares (from the package `stats`). Note that in order to perform the specification test by feeding `SpeTest` with an `nls` model then the arguments in `nls` must be given in the right order. Then by running the following command the parametric specification characterized by $\mathcal{F}$ is tested

```
SpeTest(eq)
```

The function returns an object of class `STNP` which when printed with `print` or `print.STNP` returns the test statistic and its p-value. An object of type `STNP` is a list which not only contains the test statistic `stat` and its p-value `pval` but also the type of the test `type`, the rejection rule `rejection`, the test statistic normalization `norma`, the Kernel function denoted as $K(\cdot)$ used to compute the test statistic central matrix `ker`, the standardization method of test the statistic central matrix `knorm`, the type of bootstrap used to compute the p-value `boot`, the number of bootstrap samples used to compute the p-value `nboot`, the bandwidths `cch` and `hv`, etc... To obtain a summary of the test and its options the method `summary` or `summary.STNP` can be used on objects of class `STNP`.

By default the test of Bierens (1982) with the standard normal density as the central matrix function is applied and the test p-value is obtained using 50 wild bootstrap samples with a naive estimator of the conditional variance of the errors. Among many options, by changing the argument `rejection` from `bootstrap` (the default) to `asymptotics` if `type = "zheng"` or `type = "pala"` or`type = "sicm"` the test p-value is then based on the asymptotic normality of these normalized test statistics under the null. In addition by default the test statistic is not normalized as in by default the denominator in $T_{zheng}$, $T_{pala}$ and $T_{sicm}$ is set to one. This can be changed by setting `norma = "naive"` to normalize the statistic using a naive estimator of the errors conditional variance as in $T_{zheng}$, or by setting `norma = "np"` to normalize the statistic using a nonparametric estimator of the errors conditional variance as in $T_{pala}$ and $T_{sicm}$. If `rejection = "bootstrap"` setting `para` to `T` greatly speeds up the computation of the p-value by deriving bootstrapped statistics in parallel. For more details refer to the next section or `help(SpeTest)`.

Note that the functions `SpeTest_Stat` and `SpeTest_Dist` are also available. Both functions take similar arguments to `SpeTest`. `SpeTest_Stat` computes the specification test statistic, while `SpeTest_Dist` generates a vector of size `nboot` from the specification test statistic distribution under the null hypothesis using the bootstrap. The argument `para` is also available to `SpeTest_Dist`. `SpeTest_Stat` and `SpeTest_Dist` allow to easily perform simulation exercises.

## 3.3 Arguments description and additional features

To be more specific about the arguments of the function `SpeTest`:

- Argument `eq` should be the fitted parametric model of class `lm` or `nls`of the parametric specification of interest $\mathcal{F}$
- Argument `type` refers to the type of the test

  If `type = "icm"` the test of Bierens (1982) is performed (default)

  If `type = "zheng"` the test of Zheng (1996) is performed

  If `type = "esca"` the test of Escanciano (2006) is performed, significantly increases computing time

  If `type = "pala"` the test of Lavergne and Patilea (2008) is performed

  If `type = "sicm"` the test of Lavergne and Patilea (2012) is performed
- Argument `rejection` refers to the rejection rule

  If `rejection = "bootstrap"` the p-value of the test is based on the bootstrap (default)

  If `rejection = "asymptotics"` and `type = "zheng"` or `type = "esca"` or `type = "sicm"` the p-value of the test is based on asymptotic normality of the normalized version of one of these test statistic under the null hypothesis

  If `type = "icm"` or `type = "esca"` the argument `rejection` is ignored and the p-value is based on the bootstrap
- Argument `norma` refers to the normalization of the test statistic

If `norma = "no"` the test statistic is not normalized (default)

If `norma = "naive"` the test statistic is normalized with a naive estimator of the errors variance

If `norma = "np"` the test statistic is normalized with a nonparametric estimator of the errors variance

- Argument `boot` refers to the bootstrap method used to compute the test p-value when `rejection = "bootstrap"`

  If `boot = "wild"` the wild bootstrap of Wu (1986) is used (default)

  If `boot = "smooth"` the smooth conditional moments bootstrap of Gozalo (1997) is used

- Argument `nboot` is the number of bootstraps used to compute the test p-value, by default 'nboot = 50}

- Argument `para` determines if parallel computing is used or not when `rejection = "bootstrap"`

  If `para = F` parallel computing is not used to generate the bootstrap samples to compute the test p-value (default)

  If `para = T` parallel computing is used to generate the bootstrap samples to compute the test p-value, significantly decreases computing time, makes use of all CPU cores except one

- Argument `ker` refers to the Kernel function used in the central matrix and for the nonparametric covariance estimator if there is any

  If `ker = "normal"` the central matrix Kernel function is the normal p.d.f (default)

  If `ker = "triangle"` the central matrix Kernel function is the triangular p.d.f

  If `ker = "logistic"` the central matrix Kernel function is the logistic p.d.f

  If `ker = "sinc"` the central matrix Kernel function is the sine cardinal function

- Argument `knorm` refers to the normalization of the Kernel function

  If `knorm = "sd"` then the standard deviation using the Kernel function equals 1 (default)

  If `knorm ="sq"` then the integral of the squared Kernel function equals 1

- Argument `cch` is the central matrix Kernel bandwidth

  If `type = "icm"` or `type = "esca"` then `cch` always equals `1`

  If `type = "zheng"` the `"default"` bandwidth is the scaled rule of thumb: `cch = 1.06*n^(-1/5)`

  If `type = "sicm"` and `type = "pala"` the `"default"` bandwidth is the scaled rule of thumb: `cch = 1.06*n^(-1/(4+k))` where `k` is the number of regressors

  The user may change the bandwidth when `type = "zheng"`, `type = "sicm"` or `type = "pala"`.

- Argument `hv` is the bandwidth the nonparametric errors covariance estimator when `norma = "np"` or `rejection = "bootstrap"` and `boot = "smooth"`

  By `"default"` the bandwidth is the scaled rule of thumb `hv = 1.06*n^(-1/(4+k))`

- Argument `nbeta` refers to the number of elements $\beta$ used to represent the unit hypersphere $\mathcal{S}^k$ when `type = "pala"` or `type = "sicm"`

  Computing time increases as `nbeta` gets larger

  By `"default"` it is equal to 20 times the square root of the number of exogenous control variables

- Argument `direct` refers to the default "directions" for the tests of Lavergne and Patilea (2008) and Lavergne and Patilea (2012)

  If `type = "pala"`, `direct` is the favored direction for $\beta$, by `"default"` it is the OLS estimator if `class(eq) = "lm"`

If `type = "sicm"`, `direct` is the initial direction for $\beta$. This direction should be a vector of `0` (for no direction), `1` (for positive direction) and `-1` (for negative direction)

For example, `c(1,-1,0)` indicates that the user thinks that the 1st regressor has a positive effect on the dependent variable, that the 2nd regressor has a negative effect on the dependent variable, and that he has no idea about the effect of the 3rd regressor

By `"default"` no direction is given to the hypersphere

- Argument `alphan` refers to the weight given to the favored direction for $\beta$ when `type = "pala"`

  By `"default"` it is equal to `log(n)*n^(-3/2)`

Before changing the default options of arguments `norma`, `direct` and `alphan` we strongly advise the user to read the tests references.

## 3.4 Application

To finish we use data on 2,000 individuals from the Current Population Survey as in Stock and Watson (2007) to find the true shape of their expected earnings conditional on their years of education and their age using the test of Bierens (1982).

```
library(SpeTestNP)
library(AER)

data( CPSSW8 )

summary( CPSSW8 )
```

```
##     earnings         gender           age              region
##  Min.   : 2.003   male  :34348   Min.   :21.00   Northeast:12371
##  1st Qu.:11.058   female:27047   1st Qu.:33.00   Midwest  :15136
##  Median :16.250                  Median :41.00   South    :18963
##  Mean   :18.435                  Mean   :41.23   West     :14925
##  3rd Qu.:23.558                  3rd Qu.:49.00
##  Max.   :72.115                  Max.   :64.00
##    education
##  Min.   : 6.00
##  1st Qu.:12.00
##  Median :13.00
##  Mean   :13.64
##  3rd Qu.:16.00
##  Max.   :20.00
```

The dependent variable we consider is earnings and the explanatory variables we focus on are education and age.

### 3.4.1 Expected earnings conditional on age and education separately

First we consider the specification of the conditional expectation of earnings given the two explanatory variables separately. Conditional on age only:

```
lm_age <- lm( earnings ~ age, data = CPSSW8[1:2000,] )
summary( lm_age )
```

```
##
## Call:
## lm(formula = earnings ~ age, data = CPSSW8[1:2000, ])
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.931  -7.506  -2.340   5.107  41.055
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.31358    0.93654   13.15  < 2e-16 ***
## age          0.16533    0.02128    7.77 1.24e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.45 on 1998 degrees of freedom
## Multiple R-squared:  0.02933,    Adjusted R-squared:  0.02885
## F-statistic: 60.38 on 1 and 1998 DF,  p-value: 1.244e-14
```

```
    lm_age2 <- lm( earnings ~ age + I(age^2), data = CPSSW8[1:2000,] )
    summary( lm_age2 )
```

```
##
## Call:
## lm(formula = earnings ~ age + I(age^2), data = CPSSW8[1:2000,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.619  -7.171  -2.065   5.022  40.788
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.232863   3.290282  -1.590    0.112
## age          1.056823   0.161746   6.534 8.11e-11 ***
## I(age^2)    -0.010556   0.001899  -5.559 3.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.38 on 1997 degrees of freedom
## Multiple R-squared:  0.04413,    Adjusted R-squared:  0.04317
## F-statistic: 46.09 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```
    SpeTest( lm_age, para = T )
```

```
##
##   Bierens (1982) integrated conditional moment test
##
##   Test statistic :  145.89856
##   Bootstrap p-value :  0
##
```

```
    SpeTest( lm_age2, para = T )
```

```
##
##   Bierens (1982) integrated conditional moment test
##
##   Test statistic :  1.1218
##   Bootstrap p-value :  0.74
##
```

Clearly the quadratic specification of the expectation conditional on age is not rejected, the p-value is above 50%. Conditional on education only:

```
    lm_educ <- lm(earnings ~ education, data = CPSSW8[1:2000,] )
    summary(lm_educ)
```

```
##
## Call:
## lm(formula = earnings ~ education, data = CPSSW8[1:2000, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.142  -6.236  -1.649   4.594  42.268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.08180    1.23459  -6.546 7.48e-11 ***
## education    1.98425    0.08795  22.562  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.472 on 1998 degrees of freedom
## Multiple R-squared:  0.2031, Adjusted R-squared:  0.2027
## F-statistic: 509.1 on 1 and 1998 DF,  p-value: < 2.2e-16
```

```
    lm_educ2 <- lm(earnings ~ education + I(education^2), data = CPSSW8[1:2000,] )
    summary(lm_educ2)
```

```
##
## Call:
## lm(formula = earnings ~ education + I(education^2), data = CPSSW8[1:2000,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.186  -6.125  -1.718   4.609  42.940
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    14.34252    5.46702   2.623  0.00877 **
## education      -1.25377    0.77417  -1.619  0.10550
## I(education^2)  0.11345    0.02695   4.210 2.67e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.433 on 1997 degrees of freedom
## Multiple R-squared:  0.2101, Adjusted R-squared:  0.2093
## F-statistic: 265.5 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```
    SpeTest( lm_educ, para = T )
```

```
##
##   Bierens (1982) integrated conditional moment test
##
##   Test statistic :  41.9029
##   Bootstrap p-value :  0
```

```
##
```

```
    SpeTest( lm_educ2, para = T )
```

```
##
##   Bierens (1982) integrated conditional moment test
##
##   Test statistic :  0.66312
##   Bootstrap p-value :  0.88
##
```

Similarly the quadratic specification of the expectation conditional on education is not rejected.

### 3.4.2 Expected earnings conditional on age and education jointly

Second we model the specification of the conditional expectation of earnings given the two explanatory variables jointly. Conditional on age and education we consider a linear and a quadratic specification with age, education, and their square included as control variables:

```
    lm_lin <- lm( earnings ~ age + education,
                  data = CPSSW8[1:2000,] )
    summary(lm_lin)
```

```
##
## Call:
## lm(formula = earnings ~ age + education, data = CPSSW8[1:2000,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.581  -6.210  -1.545   4.627  42.218
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.48238    1.46212 -10.589   <2e-16 ***
## age           0.17059    0.01890   9.024   <2e-16 ***
## education     1.99362    0.08623  23.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.287 on 1997 degrees of freedom
## Multiple R-squared:  0.2343, Adjusted R-squared:  0.2335
## F-statistic: 305.5 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```
    lm_quad <- lm( earnings ~ age + I(age^2) + education + I(education^2),
                  data = CPSSW8[1:2000,] )
    summary(lm_quad)
```

```
##
## Call:
## lm(formula = earnings ~ age + I(age^2) + education + I(education^2),
##     data = CPSSW8[1:2000, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.106  -6.029  -1.634   4.486  41.619
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -12.321907   6.164596  -1.999 0.045763 *
## age              0.969749   0.143304   6.767 1.72e-11 ***
## I(age^2)        -0.009519   0.001682  -5.659 1.74e-08 ***
## education       -0.706899   0.756012  -0.935 0.349882
## I(education^2)   0.094213   0.026318   3.580 0.000352 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.187 on 1995 degrees of freedom
## Multiple R-squared:  0.2514, Adjusted R-squared:  0.2499
## F-statistic: 167.5 on 4 and 1995 DF,  p-value: < 2.2e-16
```

```
    SpeTest( lm_lin, para = T )
```

```
##
##   Bierens (1982) integrated conditional moment test
##
##   Test statistic :  39.41962
##   Bootstrap p-value :  0
##
```

```
    SpeTest( lm_quad, para = T )
```

```
##
##   Bierens (1982) integrated conditional moment test
##
##   Test statistic :  2.62128
##   Bootstrap p-value :  0.04
##
```

We reject these two specifications because the p-value is equal to 0. They do not approximate properly the conditional expectation with respect to both age and education.

Next we test a highly non-linear specification with age, age to the square, education, education to the square, and their products included as controls:

```
    lm_nlin <- lm( earnings ~ age + I(age^2) + education + I(education^2)
                   + I(education*age) + I(education^2*age)
                   + I(education*age^2) + I(education^2*age^2),
                   data= CPSSW8[1:2000,] )

    SpeTest( lm_nlin, para= T )
```

```
##
##   Bierens (1982) integrated conditional moment test
##
##   Test statistic :  0.0184
##   Bootstrap p-value :  0.92
##
```

And we do not reject this specification because the p-value is above 50%. This means that the expectation of earnings conditional on age and education highly depends on the co-movement of age and education.

Indeed, even tough a linear model with two control variables provides better fit than a model with only one control variable it does not manage to properly model the true relationship between the outcome and the controls, ie it does not manage to properly model the conditional expectation. In that sense if we add

more control variables linearly we can obtain better fit but will most likely obtain a worse estimate of the conditional expectation.

# 4 References


H.J. Bierens (1982), "Conistent Model Specification Test", *Journal of Econometrics*, 20 (1), 105-134

J.C. Escanciano (2006), "A Consistent Diagnostic Test for Regression Models Using Projections", *Econometric Theory*, 22 (6), 1030-1051

P.L. Gozalo (1997), "Nonparametric Bootstrap Analysis with Applications to Demographic Effects in Demand Functions", *Journal of Econometrics*, 81 (2), 357-393

P. Lavergne and V. Patilea (2008), "Breaking the Curse of Dimensionality in Nonparametric Testing", *Journal of Econometrics*, 143 (1), 103-122

P. Lavergne and V. Patilea (2012), "One for All and All for One: Regression Checks with Many Regressors", *Journal of Business & Economic Statistics*, 30 (1), 41-52

J.H. Stock and M.W. Watson (2007), "Why Has U.S. Inflation Become Harder to Forecast?", *Journal of Money, Credit and Banking*, 39 (1), 3-33

C.F.J. Wu (1986) "Jacknife, bootstrap and other resampling methods in regression analysis (with discussion)", *Annals of Statistics*, 14 (4), 1261-1350

J. Yin, Z. Geng, R. Li, H. Wang (2010), "Nonparametric covariance model", *Statistica Sinica*, 20 (1), 469-479

J.X. Zheng (1996), "A Consistent Test of Functional Form via Nonparametric Estimation Techniques", *Journal of Econometrics*, 75 (2), 263-289