Package 'YatchewTest'

March 29, 2024

Title Yatchew (1997), de Chaisemartin & D'Haultfoeuille (2024) Linearity Test

Version 1.0.0

Maintainer Diego Ciccia <diego.ciccia@sciencespo.fr>

Description Test of linearity originally proposed by Yatchew (1997) and improved by de Chaisemartin and D'Haultfoeuille to be robust under heteroskedasticity.

License MIT + file LICENSE

Imports Rcpp, ggplot2

LinkingTo Rcpp

Author Diego Ciccia [aut, cre],
Felix Knau [aut],
Doulo Sow [aut],
Clément de Chaisemartin [aut],
Xavier D'Haultfoeuille [aut]

Encoding UTF-8

RoxygenNote 7.2.3

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

R topics documented:

`yatchew_test` *Main function*

Description

Heteroskedasticity Robust Test of Linearity (Yatchew, 1997; de Chaisemartin and D'Haultfoeuille, 2024)

Usage

```
yatchew_test(data, ...)
```

Arguments

data	A data object.
...	Undocumented.

Value

Method dispatch depending on the data object class.

---

yatchew_test.data.frame

General yatchew_test method for unclassed dataframes

Description

General yatchew_test method for unclassed dataframes

Usage

```
## S3 method for class 'data.frame'
yatchew_test(data, Y, D, het_robust = FALSE, path_plot = FALSE, ...)
```

Arguments

data	(data.frame) A dataframe.
Y	(char) Dependent variable.
D	(char) Independent variable.
het_robust	(logical) If FALSE, the test is performed under the assumption of homoskedasticity (Yatchew, 1997). If TRUE, the test is performed using the heteroskedasticity-robust test statistic proposed by de Chaisemartin and D'Haultfoeuille (2024).
path_plot	(logical) if TRUE, the assigned object will include a plot of the sequence of (D_{1i}, D_{2i})s that minimizes the euclidean distance between each pair of consecutive observations (see Overview for further details).
...	Undocumented.

Value

A list with test results.

Overview

This program implements the linearity test proposed by Yatchew (1997) and its heteroskedasticity-robust version proposed by de Chaisemartin & D'Haultfoeuille (2024). In this overview, we sketch the intuition behind the two tests, as to motivate the use of the package and its options. Please refer to Yatchew (1997) and Section 3 of de Chaisemartin & D'Haultfoeuille (2024) for further details.

Yatchew (1997) proposes a useful extension of the test with multiple independent variables. The program implements this extension when the D argument has length > 1. It should be noted that the power and consistency of the test in the multivariate case are not backed by proven theoretical results. We implemented this extension to allow for testing and exploratory research. Future theoretical exploration of the multivariate test will depend on the demand and usage of the package.

Univariate Yatchew Test:

Let Y and D be two random variables with continuous and bounded support. The core function of the `yatchew_test()` program is to check that the Y argument is linear in the `D` argument. The null hypothesis of the test is $E[Y|D] = m(D)$, where $m(.)$ is a continuous linear function. The outcome variable can be decomposed as $Y = m(D) + \varepsilon$, with $E[\varepsilon|D] = 0$. If the null holds, $\Delta Y = \Delta\varepsilon$ for $\Delta D \to 0$. In a dataset with N sample realisations of Y and D, one can test this hypothesis as follows:

1. sort the dataset by D;
2. denote the corresponding observations by $(Y_{(i)}, D_{(i)})$, with $g \in \{1, ..., N\}$;
3. compute $\hat{\sigma}^2_{\text{diff}}$, i.e. the variance of the residuals $\varepsilon_{(i)} - \varepsilon_{(i-1)} = Y_{(i)} - Y_{(i-1)}$;
4. compute $\hat{\sigma}^2_{\text{lin}}$, i.e. the variance of the residuals from an OLS regression of Y on D.

The feasibility of step (3) derives from the fact that $Y_{(i)} - Y_{(i-1)} = m(D_{(i)}) - m(D_{(i-1)}) + \varepsilon_{(i)} - \varepsilon_{(i-1)}$ and, under the null, the first difference term is close to zero for $D_{(i)} \approx D_{(i-1)}$. Sorting at step (1) ensures that consecutive $D_{(i)}$s are as close as possible. Lastly, Yatchew (1997) shows that under homoskedasticity and regularity conditions

$$T := \sqrt{G}\left(\frac{\hat{\sigma}^2_{\text{lin}}}{\hat{\sigma}^2_{\text{diff}}} - 1\right) \xrightarrow{d} \mathcal{N}(0, 1)$$

To this end, one can reject the linearity of $m(.)$ with significance level α if $T > \Phi(1 - \alpha)$.

If the homoskedasticity assumption fails, this test leads to overrejection. De Chaisemartin & D'Haultfoeuille (2024) propose a heteroskedasticity-robust version of the test statistic above. This version of the Yatchew (1997) test can be implemented by running the command with the option `het_robust = TRUE`.

Multivariate Yatchew Test:

Let $\mathbf{D}$ is a $N \times M$ random matrix and denote with $||.,.||$ the Euclidean distance between two vectors. The new null hypothesis of the test is $E[Y|\mathbf{D}] = g(\mathbf{D})$, where $g : \mathbb{R}^M \to \mathbb{R}$ is a continuous linear function. Following the same logic as the univariate case, we want to reduce ΔY to $\Delta\varepsilon$ by valuing $g(.)$ between consecutive observations. However, multivariate sorting does not ensure that $||\mathbf{D}_{(i)}, \mathbf{D}_{(i-1)}|| \to 0$. To this end, the program runs a nearest neighbor algorithm to find the sequence of observations such that the Euclidean distance between consecutive positions is minimized. The algorithm has been programmed in C++ and it has been integrated in R thanks to the `Rcpp` library. The program follows a very simple nearest neighbor approach:

1. collect all the Euclidean distances between all the possible unique pairs of rows in $\mathbf{D}$ in the matrix M, where $M_{n,m} = ||\mathbf{D}_n, \mathbf{D}_m||$ with $n, m \in \{1, ..., N\}$ row indices of $\mathbf{D}$;
2. setup the queue to $Q = \{1, ..., N\}$, the (empty) path vector $I = \{\}$ and the starting index $i = 1$;
3. remove i from Q and find the column index j of M such that $M_{i,j} = \min_{c \in Q} M_{i,c}$;
4. append j to I and start again from step 3 with $i = j$ until Q is empty.

To improve efficiency, the program collects only the $N(N-1)/2$ Euclidean distances corresponding to the lower triangle of matrix M and chooses j as $\min_{c \in Q} 1\{c < i\}M_{i,c} + 1\{c > i\}M_{c,i}$. The output of the algorithm is a sequence of row numbers such that the distance between the corresponding rows $\mathbf{D}_i$s is minimized. The program also uses two refinements suggested in Appendix A of Yatchew (1997):

- The entries in $\mathbf{D}$ are normalized in $[0, 1]$;
- The algorithm above is applied to sub-cubes, i.e. partitions of the $[0, 1]^M$ space, and the full path is obtained by joining the extrema of the subpaths.

By convention, the program computes $(2\lceil \log_{10} N \rceil)^M$ subcubes, where each univariate partition is defined by grouping observations in $2\lceil \log_{10} N \rceil$ quantile bins. If $M = 2$, the user can visualize in a ggplot graph the exact path across the normalized $\mathbf{D}_i$s by running the command with the option `path_plot = TRUE`.

The vector I at the end of the algorithm is finally used to sort the dataset and the program resumes from step (2) of the univariate case.

References

de Chaisemartin, C., d'Haultfoeuille, X., Gurgand, M. (2024). Two-way Fixed Effects and Difference-in-Difference Estimators in Heterogeneous Adoption Designs.

Yatchew, A. (1997). An elementary estimator of the partial linear model.

Examples

```
df <- as.data.frame(matrix(NA, nrow = 1E3, ncol = 0))
df$x <- rnorm(1E3)
df$b <- runif(1E3)
df$y <- 2 + df$b * df$x
yatchew_test(data = df, Y = "y", D = "x")
```

Index