Nonparametric Statistics Exam: January Session (adapted)

Name, family name and ID number

2024-12-12

Instructions

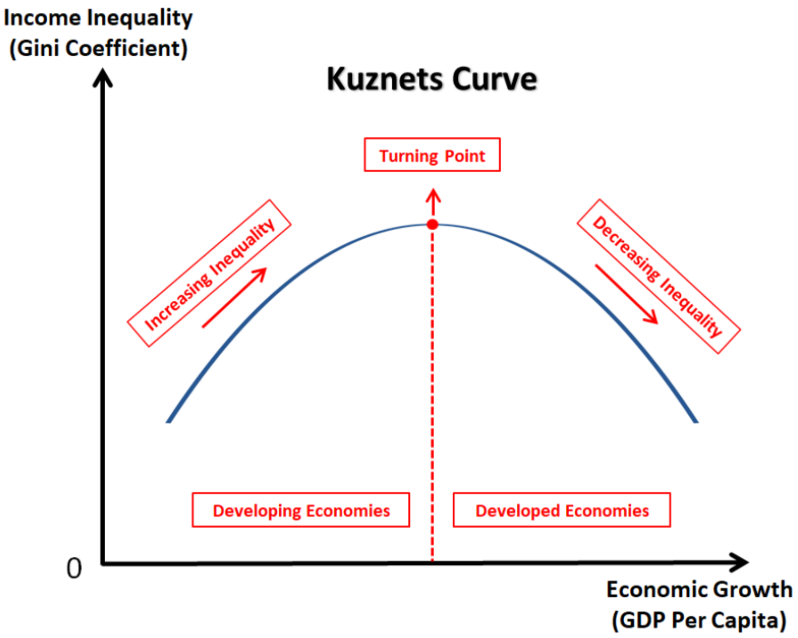
- For all computations based on permutation/bootstrapping, use $B = 1000$ replicates, and seed $= 2024$ every time a permutation/bootstrap procedure is run.
- For Full Conformal prediction intervals, use a regular grid, where, for each dimension, you have $N = 10$ equispaced points with lower bound $min(data) - 0.25 * range(data)$ and upper bound $max(data) + 0.25 * range(data)$. Moreover, do not exclude the test point when calculating the conformity measure. Except for the number of points, these are the default conditions of the `ConformalInference` R package.
- Both for confidence and prediction intervals, as well as tests, if not specified otherwise, set $\alpha = 0.05$.
- When reporting univariate confidence/prediction intervals, always provide upper and lower bounds.
- For solving the exam, you must use one of the templates previously provided and available here. Particularly, for each question you are required to report:
  - Synthetic description of assumptions, methods, and algorithms: which methodological procedure you intend to use to answer the question, succinctly describing the main theoretical characteristics of the chosen approach, and why it is suitable for the analytical task at hand. For *e.g.*, if you perform a statistical test, you should mention its assumptions, null and alternative hypotheses, and the test statistic.
  - Results and brief discussion: the actual result of the procedure applied to the data at hand, including any requested comment, output and plot.

The files with the data are available here

Exercise 1.

After reading Doughnut Economics: Seven Ways To Think Like a 21st Century Economist, Dr. LC Sulla is set out to test empirically whether the Kuznets Curve and its environmental counterpart are verified in reality.

- Classical Kuznets curve: at first, growth increases with inequality and then wealth "lifts everyone up", reducing inequality.



www.economicsonline.co.uk [1]

- Environmental Kuznets Curve (EKC): at first growth increases at the expense of environmental degradation. When people have sufficient income, the economy shifts to a less environment-degrading one. That is, after a turning point "economic growth cleans up".

As a first step to the analysis, Dr. Sulla is performing an exploratory analysis on the gathered data (*cf.* `df1.csv`)[2] for year 2010 with the following variables collected from different countries: `IncomeGroup`, `GDP_per_capita`, `CO2_emissions`, `Gini_pct`, `Population`.

1. Retrieve the country codes with the three highest Tukey depths when taking into account the three dimensions `GDP_per_capita`, `CO2_emissions_per_capita` and the `logit` transformation of the `Gini` coefficient (which is essentially a measure of inequality defined on $(0, 1)$).[3]. Which one of them is the median?

2. Focusing on those same dimensions, draw a bagplot matrix. Report the countries that are inside the bag of the `GDP_per_capita` versus `CO2_emissions` bagplot.[4]. Report which countries are outliers of such bagplot and remove them from the data set for the following exercises.

3. Without the aforementioned outliers in the dataset, implement through a permutational a MANOVA test for the equality of the multivariate mean (using `GDP_per_capita`, `CO2_emissions` and `logit(Gini)`) of the three groups of countries according to `IncomeGroup`, utilising the Wilks statistic[5]. Provide the histogram of the Monte Carlo approximation of the permutational distribution of the test statistic, the resulting p-value of the test and your conclusion.

4. Perform a Mann-Whitney U test comparing to test the median `CO2_emissions` of *Middle income* countries and that of *High income* countries.

[1] Image taken from https://www.economicsonline.co.uk/definitions/the-kuznets-curve.html/

[2] The data for these exam are real and were processed from the original data present here.

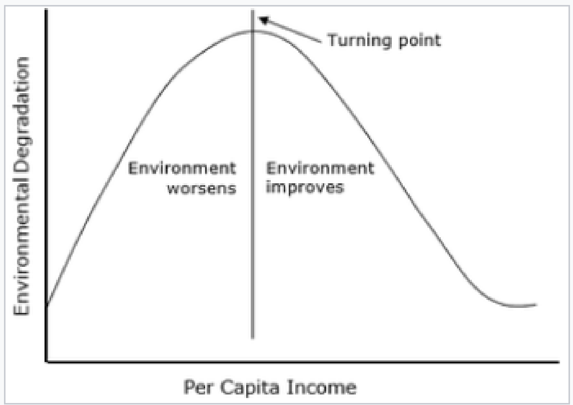
[3] Recall that $logit(x) = \ln \frac{x}{1-x}, x \in (0, 1)$

[4] Hint: The deepest 50% of the data are contained in the bag. Note that the bagplot compares only two dimensions at a time.

[5] Remember that a smaller value of Wilks's Lambda is associated to a higher significance in the test

Exercise 2

After some other touches to the data set (`df2.csv`), Dr. Sulla is ready to test for the existence of the Kuznets and the Environmental Kuznets curve.



Hypothetical environmental Kuznets curve: a translation of the Kuznets curve to the use of natural resources. [6]

1. Implement a quadratic fit utilising raw polynomials for the classical Kuznets curve (`CO2_emissions_per_capita` vs `GDP_per_capita`) and one for the Environmental Kuznets Curve (EKC) (`logit(Gini)` vs `GDP_per_capita`). Provide a plot of the regression lines with parametric pointwise parametric prediction intervals and a table summarising the coefficients. Comment the results: are they coherent with the economic theory?

2. For the EKC model, compute the prediction bands for the regression using a full conformal approach and setting $\alpha = 0.1$ as the miscoverage level. Show it in a plot.

3. For the classical Kuznets curve, fit the following model for the 50th quantile of `logit(Gini)`:

$$logit(Gini) = \sum_{g \in \mathcal{G}} \beta_g + f(GDP) + \epsilon$$

   where $\mathcal{G} = \{\text{Low income, Medium income, High income}\}$, $f(GDP)$ is a function generated by the quadratic B-Spline basis with one interior knot at `GDP_per_capita`$= 30000$. Provide a table summarising the coefficients of the regression, including a comment on statistical significance.

4. After refitting the model with a suitable family of quantiles, utilise an (adaptive) nonparametric estimator to obtain a density of the prediction of the model at `GDP_per_capita`$= 20000$, `g`=`High income`.

Exercise 3

Before having a fallout with his (former) friend Dr. Marius[7], Dr. Sulla received from him a data set of smoothed functions for the observations `CO2_emissions` along time from 1990 to 2010, in an equally spaced grid of 100 elements and only for `High income` and `Low income` countries. `ex3.rds` contains a list with the data frame and the time grid.

[6] Image source: Wikipedia

[7] Yes, we have named the characters after Roman dictator Lucius Cornelius Sulla and Roman consul Gaius Marius, who at first were friends and then enemies!

1. By suitably defining a functional data object, plot the resulting curves for the $N = 57$ countries. Build a functional boxplot without adjusting the F value and report which countries are magnitude outliers.

Provide the (point-wise) $\alpha = 0.1$-trimmed mean of each group [8]. Add them the last plot. Calculate the pointwise p-value of the permutation test using as test statistic the difference of the $\alpha = 0.1$-trimmed means [9] and interpret your results.

3. Following the logic of the EKC, high and low income countries should produce different `C02_emissions`. To see whether this is verified in the `C02_emissions` measured along time, perform a global permutation test for the equality of both functional α-trimmed means of both gorups, the $L2$ distance between them. You can approximate it with the euclidean distance of the discretely sampled curves.

[8] Hint: this reduces to the α - trimmed mean at every point of the grid for each income group.

[9] So at every time point there is a p-value of the local test