Linear regression with NPI

Damjan Škulj

University of Ljubljana, Faculty of Social Sciences, Slovenia

Second Workshop on Principles and Methods of Statistical Inference with Interval Probability
Munich, September 2009

Outline

Assumptions

- *Nonparametric predictive inference (NPI)* model (Coolen and Coolen-Schrijner (2000), Coolen and Van der Laan (2001), Augustin and Coolen (2004)) is based on Hill's assumption $A_{(n)}$ (Hill (1968)):
- n observations $x_1, \ldots, x_n$ are given corresponding to r.v.s $X_1, \ldots, X_n$, enumerated in the increasing order, $x_0 := -\infty$ and $x_{n+1} = \infty$:

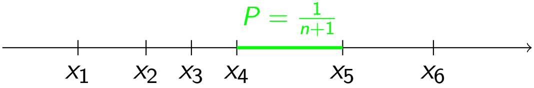
Assumptions

- *Nonparametric predictive inference (NPI)* model (Coolen and Coolen-Schrijner (2000), Coolen and Van der Laan (2001), Augustin and Coolen (2004)) is based on Hill's assumption $A_{(n)}$ (Hill (1968)):
- n observations $x_1, \ldots, x_n$ are given corresponding to r.v.s $X_1, \ldots, X_n$, enumerated in the increasing order, $x_0 := -\infty$ and $x_{n+1} = \infty$:

- What is the probability distribution corresponding to X_{n+1}?
- Basic assumption: $P(X_{n+1} \in [x_i, x_{i+1}]) = \dfrac{1}{n+1}$ for all $i = 0, \ldots, n$.

Assumptions

- *Nonparametric predictive inference (NPI)* model (Coolen and Coolen-Schrijner (2000), Coolen and Van der Laan (2001), Augustin and Coolen (2004)) is based on Hill's assumption $A_{(n)}$ (Hill (1968)):
- n observations $x_1, \ldots, x_n$ are given corresponding to r.v.s $X_1, \ldots, X_n$, enumerated in the increasing order, $x_0 := -\infty$ and $x_{n+1} = \infty$:



- What is the probability distribution corresponding to X_{n+1}?
- Basic assumption: $P(X_{n+1} \in [x_i, x_{i+1}]) = \dfrac{1}{n+1}$ for all $i = 0, \ldots, n$.

Lower and upper expectations

- The probability assignments under the basic assumption can be extended to a lower and an upper probability $\underline{P}$ and $\overline{P}$ on the σ-field of all Borel subsets of $\mathbb{R}$.
- For linear regression we need (conditional) expectations.
- In our case we can calculate the lower and the upper expectations.
- Problem: The lower and the upper expectation under NPI are $\pm\infty$ respectively regardless of the values of $x_1, \dots, x_n$.

Bounded NPI

- The only solution is to bound the possible values of X_{n+1}.
- The assumption added: $\underline{P}(X_{n+1} \in [L, U]) = 1$ for some lower and upper bounds L and U respectively:

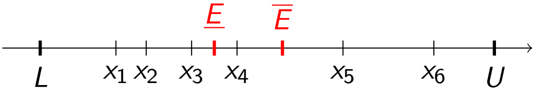
Calculation of the lower and upper expectations

- Let the points $x_1, \dots, x_n$ and $x_0 := L, x_{n+1} := U$ be given.
- Under the above assumptions we have:

$$\underline{E}(X_{n+1} | x_0, \dots, x_{n+1}) = \frac{1}{n+1} \sum_{i=0}^{n} x_i$$

and

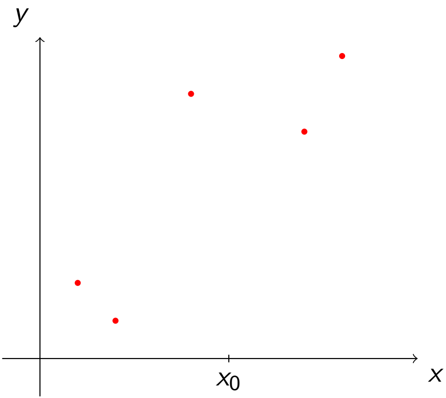
$$\overline{E}(X_{n+1} | x_0, \dots, x_{n+1}) = \frac{1}{n+1} \sum_{i=1}^{n+1} x_i$$



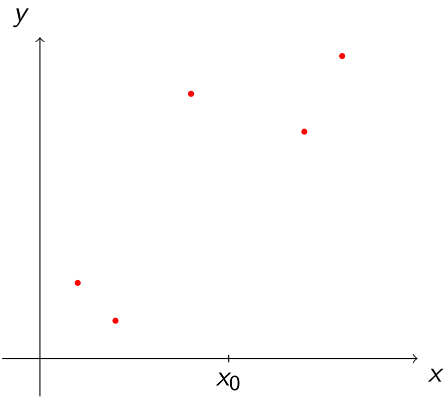
Combining linearity assumption with NPI

- Linear regression is applied when a linear relationship between dependent and independent varible(s) is assumed.
- A regression estimate is then obtained as a conditional expectation calculated on the basis of this assumption, the data and the value(s) of the independent variable(s).
- To apply NPI, a set of real valued points is needed for each possible value (set of values) of independent (variables).

A possible approach

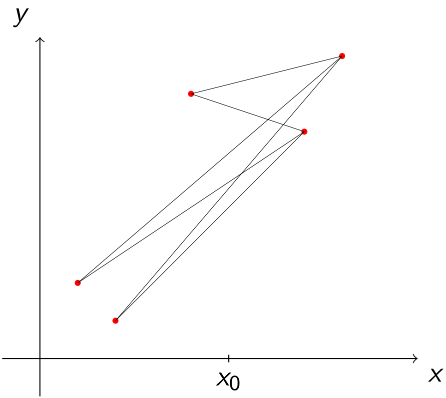


A possible approach



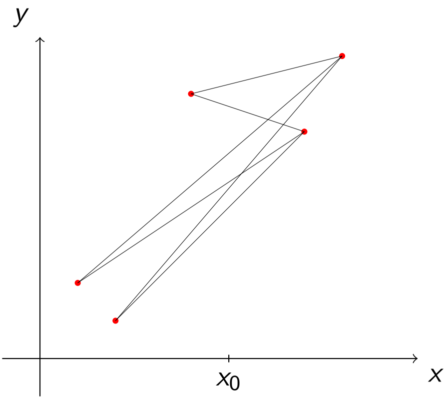
- Connect data points with lines (linearity assumption).

A possible approach



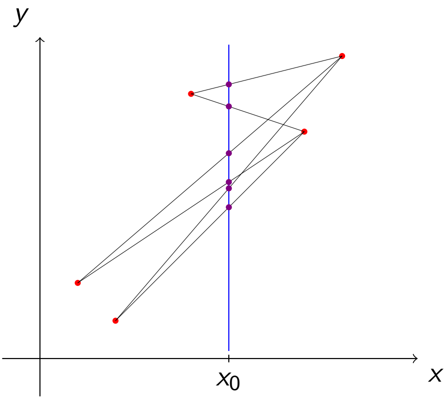
- Connect data points with lines (linearity assumption).

A possible approach



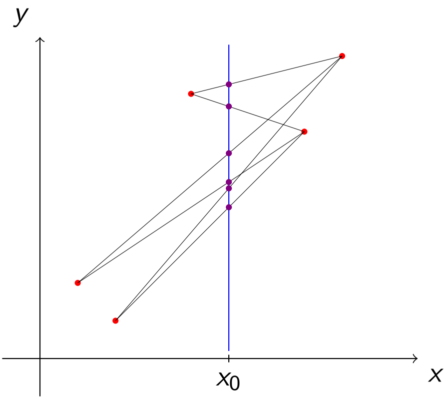
- Connect data points with lines (linearity assumption).
- Some of the lines intersect the vertical line $x = x_0$.

A possible approach



- Connect data points with lines (linearity assumption).
- Some of the lines intersect the vertical line $x = x_0$.

A possible approach



- Connect data points with lines (linearity assumption).
- Some of the lines intersect the vertical line $x = x_0$.
- The intersection points can be used for NPI estimation.

Example

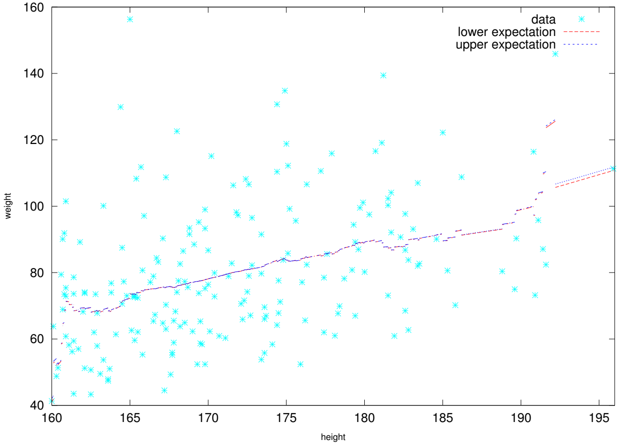


Figure: A demonstrantion of the NPI linear regression on real data: Hable (2009)

Multiple regression

- The method can be generalised to the case with multiple independent variables.
- Take a point $\mathbf{x}_0$.
- Instead of lines, we take n-dimensional hyperplanes that are defined by $n+1$ data points.
- We only consider those hyperplanes where $\mathbf{x}_0$ is a convex combination of the set of independent parts of the data points.
- Data points used for NPI are the intersections between the hyperplanes and the line $\mathbf{x} = \mathbf{x}_0$.