Notes on Lab Session 4

Giulia Tani
https://tanigiulia.github.io/

TSE - MSc course in Program Evaluation
February 2026

Introduction to non-parametric estimation

- We observe (Y_i, X_i) for $i = 1, \ldots, n$. For a new individual with $X = x$, we want to predict Y. We know that he conditional expectation minimizes the conditional MSE:

$$m(x) = \mathbb{E}[Y \mid X = x] = \arg\min_{a \in \mathbb{R}} \mathbb{E}[(Y - a)^2 \mid X = x].$$

- There are two possibilities:
  - » **Parametric estimation**: assume a functional form, e.g. $m(x) = x^\top \theta$ (linear regression), and estimate θ from the sample.
  - » **Non-parametric estimation**:

$$m(x) = \mathbb{E}[Y \mid X = x] = \int y \, f_{Y|X}(y \mid x) \, dy = \int y \, \frac{f_{Y,X}(y, x)}{f_X(x)} \, dy.$$

Estimate the densities locally:

$$\hat{m}(x) = \int y \, \frac{\widehat{f}_{Y,X}(y, x)}{\widehat{f}_X(x)} \, dy,$$

where $\widehat{f}_{Y,X}(y, x)$ and $\widehat{f}_X(x)$ are kernel density estimators of the joint density of (Y, X) and the marginal density of X.

Kernel density estimation

- Since the empirical CDF is a step function, we cannot differentiate it to get the density.
  We define the smoothed CDF estimator:

  $F_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_i}{h}\right)$, where $H : \mathbb{R} \to [0,1]$ is a smooth increasing function from 0 to 1.

  If x is far below X_i, $H(\cdot) \approx 0$. If x is far above X_i, $H(\cdot) \approx 1$.
  As $h \to 0$, the transition becomes sharper and $H((x - X_i)/h)$ behaves like the empirical CDF.

- Define the **kernel** as the derivative of H:

  $$K(u) = H'(u).$$

  Differentiate $F_{n,h}(x)$ with respect to x (chain rule):

  $$\hat{f}_{n,h}(x) = \frac{d}{dx}F_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

  Each observation X_i contributes a bump centered at X_i. The estimate $\hat{f}_{n,h}(x)$ is the average height of these bumps at the point x. The factor $1/h$ keeps the total area equal to 1.

Kernel regression

- Plugging $\widehat{f}_{Y,X}(y,x)$ and $\widehat{f}_X(x)$ into $\hat{m}(x)$, we obtain the **Nadaraya–Watson estimator**:

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x)\, Y_i, \qquad \text{where } w_i(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}.$$

  This is a *local average* of the Y_i: $w_i(x)$ is large when X_i is close to x and small when X_i is far. Smaller h makes weights go to zero faster.

- Choosing **bandwidth** h means balancing bias and variance.

  Small bandwidth:

  - » Only very nearby points receive non-negligible weight.
  - » The estimate is very local (low bias), but it uses fewer effective observations (high variance).

  Large bandwidth:

  - » Many points receive weight.
  - » The estimate is smoother and more stable (lower variance), but less local (higher bias).

Non-parametric regression: pros and cons

Pros:

- More flexible relationship, lower risk of misspecification
- Can improve predictive performance (higher R^2)

Cons:

- Requires larger samples
- Overfitting: model captures too much noise from the estimating sample, so it performs poorly in other samples

Application: recap

- We work with simulated data: we know the data generating process, we measure how far our estimates are from the true coefficients.
- There is a (fictional) government program aiming at reducing the use of pollutant fertilizer in carrot production by subsidizing the supply of low-pollution fertilizers.
- To investigate whether the program is effective, a pilot is run in 5 (eligible) French regions: a fixed quantity of minerals per hectare is given to the first 200 farms that apply.

– Unit of observation i: the farm.
– Treatment: farm receives the minerals (`d`).
– Outcome: use of pollutant fertilizer after the program (`fertilizer_2020`).
– No randomization: to receive treatment, farms must be quick to apply.
– Selection bias: farms who are quick to apply are systematically different (bigger, located in protected areas, with younger owners...), and their usage of the pollutant fertilizer would have been different even without the treatment.

Application (cont'd)

- We assume that, if we control for farm's characteristics X_i, the treatment is as good as randomly assigned (**selection on observables**):

$$Y_i = \alpha + \beta D_i + X_i'\gamma + \epsilon_i.$$

- If the dimension of X_i is large, instead of conditioning on X_i we can **condition on the propensity score** $\hat{\pi}(X_i)$ (regression method):

$$Y_i = \alpha + \beta D_i + \gamma\hat{\pi}(X_i) + \epsilon_i.$$

  The propensity score can be estimated:
  - » parametrically (e.g. logit);
  - » non-parametrically (e.g. Kernel estimation).

- Alternatively, we can use **matching methods**: compare observations in treatment and control group having similar X_i or $\hat{\pi}(X_i)$.

Matching methods

- Write the treatment effect on the treated as $\mathbb{E}\big(Y_i(1) - Y_i(0) \mid D_i = 1\big)$. Then, under CIA:

  $$ATT = \mathbb{E}\Big(Y_i(1) - \mathbb{E}\big(Y_i(0) \mid X_i, D_i = 1\big) \;\Big|\; D_i = 1\Big) = \mathbb{E}\Big(Y_i(1) - \mathbb{E}\big(Y_i(0) \mid X_i, D_i = 0\big) \;\Big|\; D_i = 1\Big).$$

  Idea: for a treated unit i with observed x_i, find individuals in the control group with (approximately) the same x_i to estimate $\mathbb{E}\big(Y_i(0) \mid X_i = x_i, D_i = 0\big)$.

- Example: **nearest-neighbor matching**.
  1) Take a treated unit with covariates x_i (so $D_i = 1$ and we observe $Y_i(1)$).
  2) Find closest control unit j with $D_j = 0$ based on propensity score.
     - If multiple control units have same distance, one is chosen randomly
     - Control unit can be used with or without replacement.
  3) Use the outcome of the matched control unit to estimate the missing counterfactual untreated outcome $\widehat{Y_i(0)}$.
  4) Compute the treated-minus-matched-control difference $Y_i(1) - \widehat{Y_i(0)}$
  5) Take the average difference.

- We still need common support: $0 < \pi(x) < 1$ for all x in the relevant support (so treated units have comparable controls).

Application: Estimates

- We know the true impact of policy pilot (treatment) on pollutant fertilizer use: -10 kg/Ha.
- Estimation results:
  - » $\hat{\beta} = -10.85$ from regression with controls
  - » $\hat{\beta} = -11.31$ from regression with (parametrically estimated) propensity score as control
  - » $\hat{\beta} = -13.76$ from propensity score matching with parametric propensity score model
  - » $\hat{\beta} = -9.64$ from propensity score matching with nonparametric propensity score model
- Note:
  - » Regression relies on a chosen functional form, while matching is more flexible.
  - » But matching on many covariates runs into the curse of dimensionality.
  - » The propensity score helps by collapsing X into the single index $\pi(X)$, but it still must be estimated: parametric logit uses a small fixed number of parameters, whereas nonparametric estimation can again suffer from sparsity.