ASML: quarto laboratorio

Regressione splines e GAM

Leonardo Egidi

Ottobre 2024

Università di Trieste

Laboratorio 4

L'obiettivo del laboratorio è familiarizzare con splines e GAM. Nello specifico, imparare a stimare questi modelli di regressione, visualizzare e interpretare le stime, e simulare dati sintetici.

In classe:

- Simulare $n = 150$ dati da una spline cubica naturale con $K = 3$ nodi (knots) fissati e basi troncate, usando i seguenti valori per i parametri 'veri': $\beta_0 = 1,\ \beta_1 = 0.7,\ \beta_2 = -0.1,\ \beta_3 = 5,\ \beta_4 = -1.5,\ \beta_5 = 0.1,\ \beta_6 = -0.3$. Simulare X da una Uniforme(-1,1), e piazzare i nodi in: $(-0.7, 0, 0.7)$. Quanti sono i gdl in questo caso? Raffigurare il grafico di Y in funzione di X.
- Stimare la spline cubica sui dati del punto precedente attraverso la funzione `bs()` del pacchetto `splines`, dapprima specificando i nodi (argomento `knots`), poi stabilendo invece i gdl desiderati (argomento `df`). Che differenza c'è tra le due splines? Raffigurare le curve stimate.
- Stimare una spline naturale cubica con il comando `ns()` e una spline di lisciamento con il comando `smooth.spline()`, facendo variare i valori di λ (bastano 2,3 valori). Raffigurare i risultati. Come varia la spline di lisciamento al variare di λ?

Laboratorio 4

Per la consegna:

- Ripetere tutti i punti precedenti, ma stavolta usando la cosiddette B-splines (usando i dati $y.2$), ovverossia generando le basi iniziali con il comando: `bs(x, knots = knots, intercept = TRUE)`, anziché usare le basi troncate.
- Stimare e raffigurare un numero a scelta di splines sui dati `splines_data.csv`.
- Considerare i dati `trees` del pacchetto `mgcv`. Usare `Volume` come variabile risposta e stimare un modello GAM usando `Girth` e `Height` come covariate. Fare diverse prove (usando solo uno dei due predittori alla volta, poi entrambi), produrre e interpretare anche graficamente le stime dei coefficienti e comparare modelli diversi con il comando `anova()`. Ci sono effetti non lineari delle covariate sulla variabile risposta?