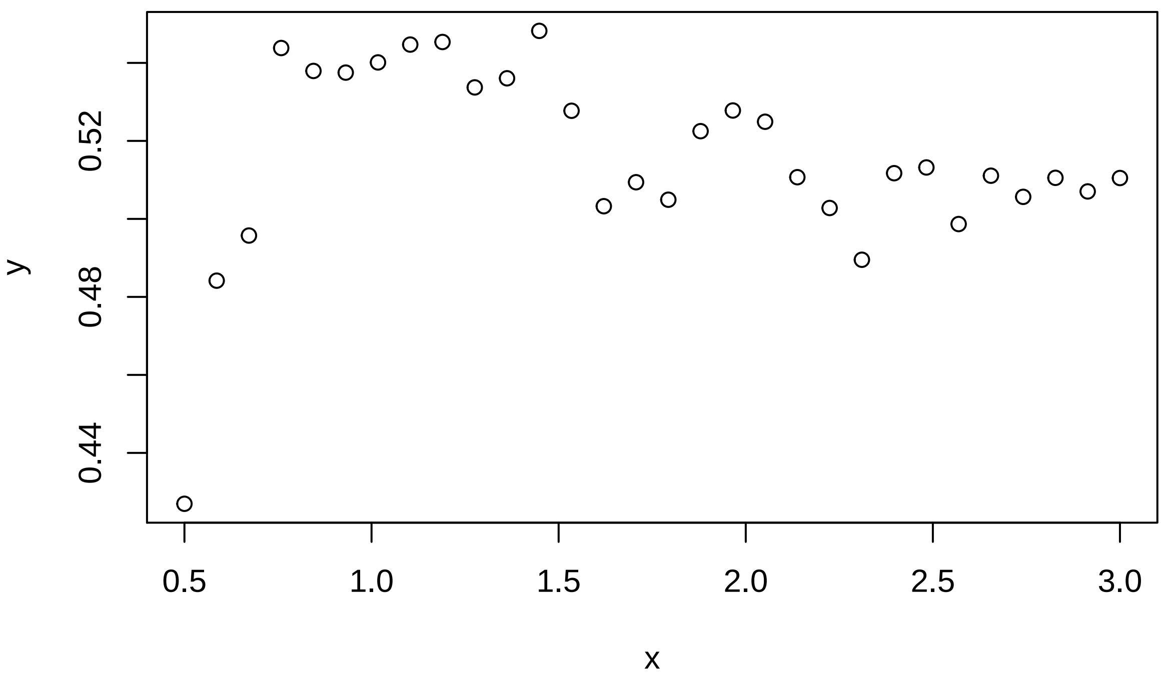
Problema 0

Esercizio tratto dal libro Azzalini e Scarpa (2001), Capitolo 3.

Descrizione del problema

Si consideri il seguente problema illustrativo che ci servirà da propototipo per situazioni più complesse e realistiche.

Supponiamo che ieri abbiamo osservato n ($n = 30$) coppie di dati (x_i, y_i) per $i = 1, \ldots, n$, i dati di addestramento (*training set*), rappresentati nel seguente diagramma di dispersione:



I dati in realtà sono stati generati artificialmente da una legge del tipo

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

dove $\varepsilon_1, \ldots, \varepsilon_n$ sono variabili casuali (v.c.) indipendenti e identicamente distribute (i.i.d.) $N(0, \sigma^2)$ con $\sigma = 10^{-2}$, mentre f è una funzione che lasceremo non specificata, salvo per il fatto che si tratta di una funzione dall'andamento sostanzialmente regolare. Naturalmente per poter generare i dati è stata scelta una funzione specifica (e non è un polinomio).

Si noti che la v.c. viene indicata con Y_i, mentre la sua realizzazione (il valore osservato) con y_i.

Inoltre si assume che $x_1, \ldots, x_n$ sono dei valori costanti (non casuali) fissati dallo sperimentatore.

Si vuole individuare una stima di $f(x)$ che ci consenta di predire i nuovi dati che ci arriveranno domani, i dati di verifica (*test set*), prodotti dallo stesso meccanismo generatore. Per semplicità di ragionamento assumiamo che queste nuove y_i^* siano associate alle stesse ascisse x_i dei dati di ieri. Abbiamo quindi che domani osserveremo n coppie di dati (x_i, y_i^*) per $i = 1, \ldots, n$, i dati di verifica (*test set*) generati come

$$Y_i^* = f(x_i) + \varepsilon_i^*, \quad i = 1, \ldots, n$$

dove $\varepsilon_1^*, \ldots, \varepsilon_n^*$ sono i.i.d. $N(0, \sigma^2)$ con $\sigma = 10^{-2}$.

Le assunzioni fatte corrispondono al cosiddetto *Fixed-X setting*:

- i valori $x_1, \ldots, x_n$ del training set sono fissati (non casuali)
- i valori di x nel test set sono uguali ai valori di x nel training set

A riguardo, si consiglia la lettura Rosset and Tibshirani (2018).

Si consideri un modello di regressione polinomiale di grado d:

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \ldots + \beta_{d+1} x^d$$

E' quindi possibile utilizzare i dati di addestramento (training set) per ottenere le stime $\hat{\beta}_1, \hat{\beta}_2, \ldots$ e quindi

$$\hat{f}(x) = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 x^2 + \ldots + \hat{\beta}_{d+1} x^d$$

per predire le nuove y_i^* che osserveremo domani utilizzando

$$\hat{y}_i^* = \hat{f}(x_i), \quad i = 1, \ldots, n$$

Non avendo informazioni che ci guidino nella scelta del grado del polinomio, dovete considerare tutti i gradi possibili con d tra 0 e $n-1$, quindi con un numero $p = d+1$ di parametri che varia da 1 a n, in aggiunta a σ.

Dati

I dati sono disponibili all'indirizzo web http://azzalini.stat.unipd.it/Libro-DM/. In particolare

- i dati "di ieri e di domani": http://azzalini.stat.unipd.it/Libro-DM/ieri-domani.dat dove (`x`, `y.ieri`) sono i dati di training (x_i, y_i) e (`x`, `y.domani`) sono i dati di test (x_i, y_i^*)
- i valori della vera funzione "f" (`f.vera`) in corrispondenza ai punti specificati (`x`) e il valore vero di σ (`sqm.vero <- 0.01`): http://azzalini.stat.unipd.it/Libro-DM/f_vera.R

Domande

0.

Stimare il modello di regressione polinomiale di grado $d = 3$ e aggiungere al diagramma di dispersione di (x_i, y_i), $i = 1, \ldots, n$ i valori previsti dal modello. Si calcoli l'errore quadratico medio sui dati di training (*Training Mean Squared Error*)

$$\mathrm{MSE}_{\mathrm{Tr}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

Per rispondere a questa domanda, potete utilizzare **solo** il training set, ovvero i dati `x` e `y.ieri`.

1.

Si decida il grado d da utilizzare per prevedere i dati di domani, con l'obiettivo di minimizzare l'errore di previsione, ovvero l'errore quadratico medio sui dati di test (*Test Mean Squared Error*)

$$\mathrm{MSE}_{\mathrm{Te}} = \frac{1}{n} \sum_{i=1}^{n} (y_i^* - \hat{f}(x_i))^2$$

Si noti che MSE_{Te} sarà calcolabile solo domani (ovvero dopo aver fatto le previsioni), a differenza dell'errore quadratico medio sui dati di training MSE_{Tr}, che si può calcolare già oggi avendo a disposizione i dati di ieri.

Si giustifichi il motivo (statistico) della scelta effettuata.

Per rispondere a questa domanda, potete utilizzare **solo** il training set, ovvero i dati `x` e `y.ieri` (non è ammissibile utilizzare i dati di domani `y.domani` o la vera f `f.vera`).

2.

La regressione polinomiale è un caso particolare del modello lineare (in notazione matriciale)

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

dove $\underset{n\times 1}{\mathbf{y}} = (y_1, \ldots, y_n)^\mathsf{T}$ è il vettore risposta di dimensione $n \times 1$, $\underset{p\times 1}{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ è il vettore dei coefficienti di dimensione $p \times 1$ e $\underset{n\times p}{\mathbf{X}}$ è la matrice del disegno di dimensione $n \times p$, ovvero

$$\underset{n\times p}{\mathbf{X}} = \begin{bmatrix} x_1^\mathsf{T} \\ x_2^\mathsf{T} \\ \cdots \\ x_i^\mathsf{T} \\ \cdots \\ x_n^\mathsf{T} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

e infine $\underset{n\times 1}{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T}$ ha distribuzione Normale n-variata $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ dove $\mathbf{I}_n$ indica la matrice identità con n righe.

Ad esempio, la matrice del disegno per il polinomio di grado $d = 2$ è la seguente (prime sei righe)

```
X <- model.matrix(lm( y ~ poly(x, degree=2, raw=T), train ))
head(X)
```

```
  (Intercept) poly(x, degree = 2, raw = T)1 poly(x, degree = 2, raw = T)2
1           1                     0.5000000                     0.2500000
2           1                     0.5862069                     0.3436385
3           1                     0.6724138                     0.4521403
4           1                     0.7586207                     0.5755054
5           1                     0.8448276                     0.7137337
6           1                     0.9310345                     0.8668252
```

La stima del polinomio di grado $d = 2$ si ottiene con i seguenti comandi:

```
fit <- lm( y ~ poly(x, degree=2), train)
yhat <- predict(fit, newdata=test)
# si noti che con poly(x, degree=2, raw=TRUE) si ottengono le stesse yhat. Perchè?
```

tuttavia se provo con $d \geq 24$ ottengo (almeno sul mio computer) il seguente messaggio di errore

```
lm( y ~ poly(x, degree=24), train)
```

> Error in poly(x, degree = 24) : 'degree' must be less than number of unique points

Spiegare qual è il problema. **Suggerire** possibili soluzioni (in termini di codice R) per ottenere le stime dei coefficienti del modello con $d = 28$ (riportare le stime dei coefficienti senza arrotondare i valori).

Cosa vi aspettate di ottenere (in termini di valori previsti $\hat{y}_i^*$) se utilizzate il polinomio di grado $n - 1$? Si giustifichi la risposta.

Per rispondere a questa domanda, potete utilizzare **solo** il training set, ovvero i dati `x` e `y.ieri`.

3.

Si supponga di conoscere la vera f. **Si decida il grado** d da utilizzare per prevedere generici dati di domani (non necessariamente il test set `x` e `y.domani`) con generici dati di ieri (non necessariamente il training set `x` e `y.ieri`), con l'obiettivo di minimizzare il valore atteso dell'errore di previsione, ovvero

$$\mathbb{E}[\mathrm{MSE}_{\mathrm{Te}}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(y_i^* - \hat{f}(x_i))^2]$$

dove il valore atteso è rispetto alle v.c. $Y_1, \ldots, Y_n$ e $Y_1^*, \ldots, Y_n^*$.

Si giustifichi la scelta effettuata, commentando il risultato alla luce della risposta fornita alla domanda 1.

Lettura suggerita: Capitolo 7.2 del libro Hastie, Tibshirani, Friedman (2009). The Elements of Statistical Learning. Springer

Per rispondere a questa domanda, potete utilizzare **solo** la vera f e il vero valore di σ, ovvero i dati `f.vera`, `x` e `sqm.vero`.

4.

Si supponga di conoscere la vera f e di aver osservato i dati di ieri. **Si decida il grado** d da utilizzare per prevedere generici dati di domani con i dati effettivamente osservati ieri (il training set `x` e `y.ieri`), con l'obiettivo di minimizzare il valore atteso dell'errore di previsione condizionato ai dati effettivamente osservati ieri, ovvero

$$\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}}|Y_1 = y_1, \ldots, Y_n = y_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Y_i^* - \hat{f}(x_i))^2|Y_1 = y_1, \ldots, Y_n = y_n\right]$$

dove il valore atteso è rispetto alle v.c. $Y_1^*, \ldots, Y_n^*$.

Si giustifichi la scelta effettuata, commentando il risultato alla luce delle risposte fornite alle domande 1. e 3.

Per rispondere a questa domanda, potete utilizzare **solo** il training set, la vera f e il vero valore di σ, ovvero i dati `x` e `y.ieri`, `f.vera` e `sqm.vero`.

Regolamento

1. Bisogna consegnare entro la scadenza prevista (per i gruppi, quella indicata, per i lavori individuali, almeno una settimana prima della della data di esame) un **UNICO** file in formato **.PDF** (non sono ammessi altri tipi di file) contenente le risposte alle domande (e il codice utilizzato). Il file deve essere nominato nel seguente modo: [MATRICOLA]_HW0.pdf (e.g. 2575_HW0.pdf) per i lavori individuali e [NOME DEL GRUPPO]_HW0.pdf per i lavori di gruppo. Il file dovrà essere caricato sulla pagina MOODLE in corrispondenza all'HOMEWORK0. Per i lavori di gruppo, **tutti** i componenti del gruppo devono caricare lo stesso file. Sarà possibile effettuare **una sola** sottomissione finale (vi verrà chiesta conferma, non sono ammesse consegne via e-mail). Per tutti gli studenti, il mancato rispetto della scadenza prevista corrisponde ad un punteggio di 0.

2. Per rispondere alle domande, è ammesso l'utilizzo di qualsiasi linguaggio di programmazione. Tuttavia il codice utilizzato deve essere riportato e deve risultare **RIPRODUCIBILE**. Troverate nella pagina MOODLE un esempio in Rmarkdown con la risposta alla Domanda0.

3. La valutazione si baserà sulla correttezza, chiarezza e precisione delle risposte fornite e del codice utilizzato. Non è previsto un limite di pagine per il file da consegnare, ma verrà premiata la capacità di sintesi, ovvero una struttura argomentativa ben articolata, codice elegante e leggibile, con le conclusioni che rispondono in modo specifico e puntuale alla domanda iniziale. E' possibile utilizzare fonti (libri, Internet, persone e così via) ma è richiesto di citarle nel testo. L'uso di fonti senza citarle si traduce in un voto nullo.

4. Il docente si riserva la possibilità di chiedere a qualunque studente di spiegare le risposte fornite e/o il codice utilizzato. Per i lavori individuali, questa spiegazione (se richiesta) avverrà il giorno della prova scritta. Il punteggio ottenuto scade alla fine dell'Anno Accademico 2020/21. Tutti gli studenti sono tenuti ad aderire ad un codice di condotta, che vieta il plagio, la falsificazione, l'assistenza non autorizzata, imbrogli e altri atti gravi di disonestà accademica. Comportamenti non corretti possono essere soggetti a provvedimenti disciplinari come da Art. 35 e 36 del Regolamento Didattico di Ateneo.