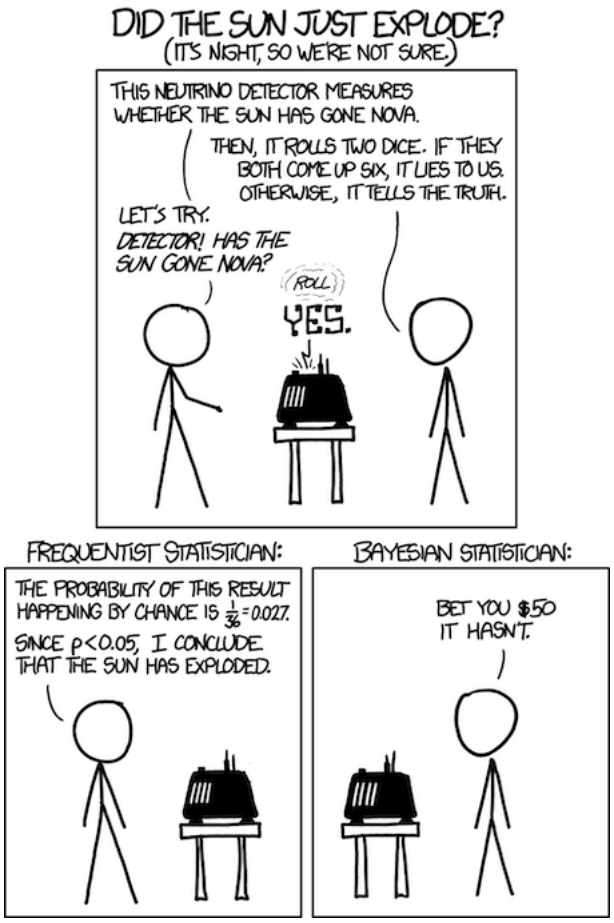
Corrado Caudek

Data Science per psicologi

Psicometria – AA 2021/2022



Indice

Elenco delle figure

Elenco delle tabelle

Prefazione

Data Science per psicologi contiene il materiale delle lezioni dell'insegnamento di *Psicometria B000286* (A.A. 2021/2022) rivolto agli studenti del primo anno del Corso di Laurea in Scienze e Tecniche Psicologiche dell'Università degli Studi di Firenze. *Psicometria* si propone di fornire agli studenti un'introduzione all'analisi dei dati in psicologia. Le conoscenze/competenze che verranno sviluppate in questo insegnamento sono quelle della Data science, ovvero un insieme di conoscenze/competenze che si pongono all'intersezione tra statistica (ovvero, richiedono la capacità di comprendere teoremi statistici) e informatica (ovvero, richiedono la capacità di sapere utilizzare un software).

La psicologia e la Data science

Sembra sensato spendere due parole su un tema che è importante per gli studenti: quello indicato dal titolo di questo Capitolo. È ovvio che agli studenti di psicologia la statistica non piace. Se piacesse, forse studierebbero Data science e non psicologia; ma non lo fanno. Di conseguenza, gli studenti di psicologia si chiedono: "perché dobbiamo perdere tanto tempo a studiare queste cose quando in realtà quello che ci interessa è tutt'altro?" Questa è una bella domanda.

C'è una ragione molto semplice che dovrebbe farci capire perché la Data science è così importante per la psicologia. Infatti, a ben pensarci, la psicologia è una disciplina intrinsecamente statistica, se per statistica intendiamo quella disciplina che studia la variazione delle caratteristiche degli individui nella popolazione. La psicologia studia *gli individui* ed è proprio la variabilità inter- e intra-individuale ciò che vogliamo descrivere e, in certi casi, predire. In questo senso, la psicologia è molto diversa dall'ingegneria, per esempio. Le proprietà di un determinato ponte sotto certe condizioni, ad esempio, sono molto simili a quelle di un altro pon-

te, sotto le medesime condizioni. Quindi, per un ingegnere la statistica è poco importante: le proprietà dei materiali sono unicamente dipendenti dalla loro composizione e restano costanti. Ma lo stesso non può dirsi degli individui: ogni individuo è unico e cambia nel tempo. E le variazioni tra gli individui, e di un individuo nel tempo, sono l'oggetto di studio proprio della psicologia: è dunque chiaro che i problemi che la psicologia si pone sono molto diversi da quelli affrontati, per esempio, dagli ingegneri. Questa è la ragione per cui abbiamo tanto bisogno della Data science in psicologia: perché la Data science ci consente di descrivere la variazione e il cambiamento. E queste sono appunto le caratteristiche di base dei fenomeni psicologici.

Sono sicuro che, leggendo queste righe, a molti studenti sarà venuta in mente la seguente domanda: perché non chiediamo a qualche esperto di fare il "lavoro sporco" (ovvero le analisi statistiche) per noi, mentre noi (gli psicologi) ci occupiamo solo di ciò che ci interessa, ovvero dei problemi psicologici slegati dai dettagli "tecnici" della Data science? La risposta a questa domanda è che non è possibile progettare uno studio psicologico sensato senza avere almeno una comprensione rudimentale della Data science. Le tematiche della Data science non possono essere ignorate né dai ricercatori in psicologia né da coloro che svolgono la professione di psicologo al di fuori dell'Università. Infatti, anche i professionisti al di fuori dall'università non possono fare a meno di leggere la letteratura psicologica più recente: il continuo aggiornamento delle conoscenze è infatti richiesto dalla deontologia della professione. Ma per potere fare questo è necessario conoscere un bel po' di Data science! Basta aprire a caso una rivista specialistica di psicologia per rendersi conto di quanto ciò sia vero: gli articoli che riportano i risultati delle ricerche psicologiche sono zeppi di analisi statistiche e di modelli formali. E la comprensione della letteratura psicologica rappresenta un requisito minimo nel bagaglio professionale dello psicologo.

Le considerazioni precedenti cercano di chiarire il seguente punto: la Data science non è qualcosa da studiare a malincuore, in un singolo insegnamento universitario, per poi poterla tranquillamente dimenticare. Nel bene e nel male, gli psicologi usano gli strumenti della Data science in tantissimi ambiti della loro attività professionale: in particolare quando costruiscono, somministrano e interpretano i test psicometrici. È dunque chiaro che possedere delle solide basi di Data science è un tassello imprescindibile del bagaglio professionale dello psicologo. In questo insegnamento verranno trattati i temi base della Data science e verrà

adottato un punto di vista bayesiano, che corrisponde all'approccio più recente e sempre più diffuso in psicologia.

Come studiare

Il giusto metodo di studio per prepararsi all'esame di Psicometria è quello di seguire attivamente le lezioni, assimilare i concetti via via che essi vengono presentati e verificare in autonomia le procedure presentate a lezione. Incoraggio gli studenti a farmi domande per chiarire ciò che non è stato capito appieno. Incoraggio gli studenti a utilizzare i forum attivi su Moodle e, soprattutto, a svolgere gli esercizi proposti su Moodle. I problemi forniti su Moodle rappresentano il livello di difficoltà richiesto per superare l'esame e consentono allo studente di comprendere se le competenze sviluppate fino a quel punto sono sufficienti rispetto alle richieste dell'esame.

La prima fase dello studio, che è sicuramente individuale, è quella in cui è necessario acquisire le conoscenze teoriche relative ai problemi che saranno presentati all'esame. La seconda fase di studio, che può essere facilitata da scambi con altri e da incontri di gruppo, porta ad acquisire la capacità di applicare le conoscenze: è necessario capire come usare un software (R) per applicare i concetti statistici alla specifica situazione del problema che si vuole risolvere. Le due fasi non sono però separate: il saper fare molto spesso ci aiuta a capire meglio.

Sviluppare un metodo di studio efficace

Avendo insegnato molte volte in passato un corso introduttivo di analisi dei dati ho notato nel corso degli anni che gli studenti con l'atteggiamento mentale che descriverò qui sotto generalmente ottengono ottimi risultati. Alcuni studenti sviluppano naturalmente questo approccio allo studio, ma altri hanno bisogno di fare uno sforzo per maturarlo. Fornisco qui sotto una breve descrizione del "metodo di studio" che, nella mia esperienza, è il più efficace per affrontare le richieste di questo insegnamento.

- Dedicate un tempo sufficiente al materiale di base, apparentemente facile; assicuratevi di averlo capito bene. Cercate le lacune nella vostra comprensione. Leggere presentazioni diverse dello stesso materiale (in libri o articoli diversi) può fornire nuove intuizioni.
- Gli errori che facciamo sono i nostri migliori maestri. Istintivamente cerchiamo di dimenticare subito i nostri errori. Ma il miglior modo di imparare è apprendere dagli errori che commettiamo. In questo senso, una soluzione corretta è meno utile di una soluzione sbagliata. Quando commettiamo un errore questo ci fornisce un'informazione importante: ci fa capire qual è il materiale di studio sul quale dobbiamo ritornare e che dobbiamo capire meglio.
- C'è ovviamente un aspetto "psicologico" nello studio. Quando un esercizio o problema ci sembra incomprensibile, la cosa migliore da fare è dire: "mi arrendo", "non ho idea di cosa fare!". Questo ci rilassa: ci siamo già arresi, quindi non abbiamo niente da perdere, non dobbiamo più preoccuparci. Ma non dobbiamo fermarci qui. Le cose "migliori" che faccio (se ci sono) le faccio quando non ho voglia di lavorare. Alle volte, quando c'è qualcosa che non so fare e non ho idea di come affontare, mi dico: "oggi non ho proprio voglia di fare fatica", non ho voglia di mettermi nello stato mentale per cui "in 10 minuti devo risolvere il problema perché dopo devo fare altre cose". Però ho voglia di *divertirmi* con quel problema e allora mi dedico a qualche aspetto "marginale" del problema, che so come affrontare, oppure considero l'aspetto più difficile del problema, quello che non so come risolvere, ma invece di cercare di risolverlo, guardo come altre persone hanno affrontato problemi simili, opppure lo stesso problema in un altro contesto. Non mi pongo l'obiettivo "risolvi il problema in 10 minuti", ma invece quello di farmi un'idea "generale" del problema, o quello di capire un caso più specifico e più semplice del problema. Senza nessuna pressione. Infatti, in quel momento ho deciso di non lavorare (ovvero, di non fare fatica). Va benissimo se "parto per la tangente", ovvero se mi metto a leggere del materiale che sembra avere poco a che fare con il problema centrale (le nostre intuizioni e la nostra curiosità solitamente ci indirizzano sulla strada giusta). Quando faccio così, molto spesso trovo la soluzione del problema che mi ero posto e, paradossalmente, la trovo in un tempo minore di quello che, in precedenza, avevo dedicato a "lavorare" al problema. Allora perché non faccio sempre così? C'è ovviamente l'aspetto dei "10 minuti" che non è sempre facile da dimenticare. Sotto pressione, possiamo solo agire in maniera automatica, ovvero possia-

mo solo applicare qualcosa che già sappiamo fare. Ma se dobbiamo imparare qualcosa di nuovo, la pressione è un impedimento.

- È utile farsi da soli delle domande sugli argomenti trattati, senza limitarsi a cercare di risolvere gli esercizi che vengono assegnati. Quando studio qualcosa mi viene in mente: "se questo è vero, allora deve succedere quest'altra cosa". Allora verifico se questo è vero, di solito con una simulazione. Se i risultati della simulazione sono quelli che mi aspetto, allora vuol dire che ho capito. Se i risultati sono diversi da quelli che mi aspettavo, allora mi rendo conto di non avere capito e ritorno indietro a studiare con più attenzione la teoria che pensavo di avere capito – e ovviamente mi rendo conto che c'era un aspetto che avevo frainteso. Questo tipo di verifica è qualcosa che dobbiamo fare da soli, in prima persona: nessun altro può fare questo al posto nostro.
- Non aspettatevi di capire tutto la prima volta che incontrate un argomento nuovo.[1] È utile farsi una nota mentalmente delle lacune nella vostra comprensione e tornare su di esse in seguito per carcare di colmarle. L'atteggiamento naturale, quando non capiamo i dettagli di qualcosa, è quello di pensare: "non importa, ho capito in maniera approssimativa questo punto, non devo preoccuparmi del resto". Ma in realtà non è vero: se la nostra comprensione è superficiale, quando il problema verrà presentato in una nuova forma, non riusciremo a risolverlo. Per cui i dubbi che ci vengono quando studiamo qualcosa sono il nostro alleato più prezioso: ci dicono esattamente quali sono gli aspetti che dobbiamo approfondire per potere migliorare la nostra preparazione.
- È utile sviluppare una visione d'insieme degli argomenti trattati, capire l'obiettivo generale che si vuole raggiungere e avere chiaro il contributo che i vari pezzi di informazione forniscono al raggiungimento di tale obiettivo. Questa organizzazione mentale del materiale di studio facilita la comprensione. È estremamente utile creare degli schemi di ciò che si sta studiando. Non aspettate che sia io a fornirvi un riepilogo di ciò che dovete imparare: sviluppate da soli tali schemi e tali riassunti.
- Tutti noi dobbiamo imparare l'arte di trovare le informazioni, non solo nel caso di questo insegnamento. Quando vi trovate di fronte a qualcosa che non capite, o ottenete un oscuro messaggio di errore da

[1] Ricordatevi inoltre che gli individui tendono a sottostimare la propria capacità di apprendere (Horn and Loewenstein, 2021).

un software, ricordatevi: “Google is your friend”!

Corrado Caudek
Marzo 2022

Parte I

Il modello lineare

1

Introduzione

Lo scopo della ricerca è trovare le associazioni tra le variabili e fare confronti fra le condizioni sperimentali. Nel caso della psicologia, il ricercatore vuole scoprire le leggi generali che descrivono le relazioni tra i costrutti psicologici e le relazioni che intercorrono tra i fenomeni psicologici e quelli non psicologici (sociali, economici, storici, ...). Abbiamo già visto come la correlazione di Pearson sia uno strumento adatto a questo scopo. Infatti, essa ci informa sulla direzione e sull'intensità della relazione lineare tra due variabili. Tuttavia, la correlazione non è sufficiente, in quanto il ricercatore ha a disposizione solo i dati di un campione, mentre vorrebbe descrivere la relazione tra le variabili nella popolazione. A causa della variabilità campionaria, le proprietà dei campioni sono necessariamente diverse da quelle della popolazione: ciò che si può osservare nella popolazione potrebbe non emergere nel campione e, al contrario, il campione manifesta caratteristiche che non sono necessariamente presenti nella popolazione. È dunque necessario chiarire, dal punto di vista statistico, il legame che intercorre tra le proprietà del campione e le proprietà della popolazione da cui esso è stato estratto. Il modello lineare utilizza la funzione matematica più semplice per descrivere la relazione fra due variabili, ovvero la funzione lineare. In questo Capitolo vedremo come si possa fare inferenza sulla relazione tra due variabili mediante il modello lineare bayesiano. Inizieremo a descrivere le proprietà geometriche della funzione lineare per poi utilizzare questa semplice funzione per costruire un modello statistico secondo un approccio bayesiano.

1.1 La funzione lineare

Iniziamo con un ripasso sulla funzione di lineare. Si chiama *funzione lineare* una funzione del tipo

$$f(x) = a + bx, \tag{1.1}$$

dove a e b sono delle costanti. Il grafico di tale funzione è una retta di cui il parametro b è detto *coefficiente angolare* e il parametro a è detto *intercetta* con l'asse delle y [infatti, la retta interseca l'asse y nel punto $(0, a)$, se $b \neq 0$].

Per assegnare un'interpretazione geometrica alle costanti a e b si consideri la funzione

$$y = bx. \tag{1.2}$$

Tale funzione rappresenta un caso particolare, ovvero quello della *proporzionalità diretta* tra x e y. Il caso generale della linearità

$$y = a + bx \tag{1.3}$$

non fa altro che sommare una costante a a ciascuno dei valori $y = bx$. Nella funzione lineare $y = a + bx$, se b è positivo allora y aumenta al crescere di x; se b è negativo allora y diminuisce al crescere di x; se $b = 0$ la retta è orizzontale, ovvero y non muta al variare di x.

Consideriamo ora il coefficiente b. Si consideri un punto x_0 e un incremento arbitrario ε come indicato nella figura 1.1. Le differenze $\Delta x = (x_0 + \varepsilon) - x_0$ e $\Delta y = f(x_0 + \varepsilon) - f(x_0)$ sono detti *incrementi* di x e y. Il coefficiente angolare b è uguale al rapporto

$$b = \frac{\Delta y}{\Delta x} = \frac{f(x_0 + \varepsilon) - f(x_0)}{(x_0 + \varepsilon) - x_0}, \tag{1.4}$$

indipendentemente dalla grandezza degli incrementi Δx e Δy. Il modo più semplice per assegnare un'interpretazione geometrica al coefficiente angolare (o pendenza) della retta è dunque quello di porre $\Delta x = 1$. In tali circostanze infatti $b = \Delta y$.

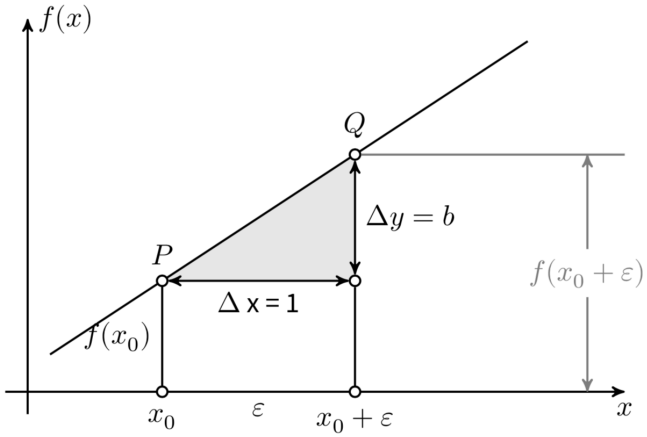


Figura 1.1: *La funzione lineare* $y = a + bx$.

1.2 Una media per ciascuna osservazione

In precedenza abbiamo visto come sia possibile stimare i parametri di un modello bayesiano nel quale le osservazioni sono indipendenti e identicamente distribuite secondo una densità gaussiana,

$$Y_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma), \quad i = 1, \dots, n. \tag{1.5}$$

Il modello (1.5) assume che ogni Y_i sia la realizzazione di una v.c. descritta da una $\mathcal{N}(\mu, \sigma^2)$. Da un punto di vista bayesiano, si assegnano distribuzioni a priori ai parametri μ e σ, si genera la verosimiglianza in base ai dati osservati e, con queste informazioni, si generano le distribuzione a posteriori dei parametri (Gelman et al., 2020):

$$\begin{aligned} Y_i \mid \mu, \sigma &\overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \\ \mu &\sim \mathcal{N}(\mu_0, \tau^2) \\ \sigma &\sim \text{Cauchy}(x_0, \gamma) \end{aligned}$$

È comune però che vengano registrate altre variabili x_i che possono essere associate alla risposta di interesse y_i. La variabile x_i viene chiamata *predittore* (o variabile indipendente) in quanto il ricercatore è tipicamente

interessato a predire il valore y_i a partire da x_i. Come si può estende il modello (1.5) per lo studio della possibile relazione tra y_i e x_i?

Il modello (1.5) assume una media μ comune per ciascuna osservazione Y_i. Dal momento che desideriamo introdurre una nuova variabile x_i che assume un diverso valore per ciascuna osservazione y_i, il modello (1.5) può essere modificato in modo che la media comune μ venga sostituita da una media μ_i specifica a ciascuna osservazione i-esima:

$$Y_i \mid \mu_i, \sigma \overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma), \quad i = 1, \dots, n. \tag{1.6}$$

Si noti che le osservazioni $Y_1, \dots, Y_n$ non sono più identicamente distribuite poiché hanno medie diverse, ma sono ancora indipendenti come indicato dalla notazione `ind` posta sopra il simbolo $\sim$ nella (1.6).

1.2.1 Relazione lineare tra la media $y \mid x$ e il predittore

L'approccio che consente di mettere in relazione un predittore x_i con la risposta Y_i è quello di assumere che la media di ciascuna Y_i, ovvero μ_i, sia una funzione lineare del predittore x_i. Una tale relazione lineare è scritta come

$$\mu_i = \beta_0 + \beta_1 x_i, \quad i = 1, \dots, n. \tag{1.7}$$

Nella (1.7), ciascuna x_i è una costante nota (ecco perché viene usata una lettera minuscola per la x) e β_0 e β_1 sono parametri incogniti. Questi parametri rappresentano l'intercetta e la pendenza della retta di regressione e sono delle variabili casuali.[1] L'inferenza bayesiana procede assegnando una distribuzione a priori a β_0 e a β_1 e si esegue l'inferenza riassumendo la distribuzione a posteriori di questi parametri.

Nel modello (1.7), la funzione lineare $\beta_0 + \beta_1 x_i$ è interpretata come il valore atteso della Y_i per ciascun valore x_i, mentre l'intercetta β_0 rappresenta il valore atteso della Y_i quando $x_i = 0$. Il parametro β_1 (pendenza) rappresenta invece l'aumento medio della Y_i quando x_i aumenta di un'unità. È importante notare che la relazione lineare (1.6) di parametri β_0 e β_1 descrive l'associazione tra *la media* μ_i e il predittore x_i. In altri termini, tale relazione lineare ci fornisce una predizione sul valore medio μ_i, non sul valore *effettivo* Y_i.

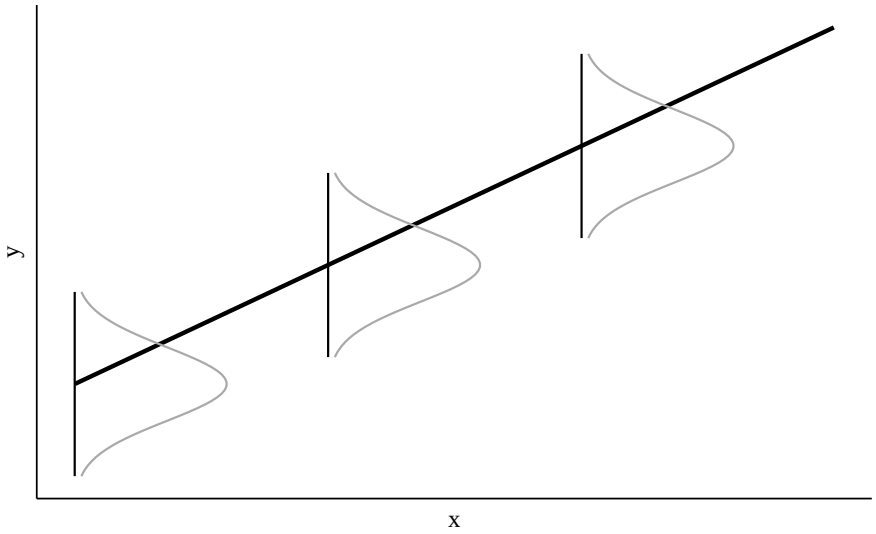
[1] Una notazione alternativa per tali parametri è α, β, anziché β_0, β_1.

1.2.2 Il modello lineare

Sostituendo la (1.7) nella (1.6) otteniamo il modello lineare:

$$Y_i \mid \beta_0, \beta_1, \sigma \overset{ind}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma), \quad i = 1, \dots, n. \tag{1.8}$$

Questo è un caso speciale del modello di campionamento Normale, dove le Y_i seguono indipendentemente una densità Normale con una media $(\beta_0 + \beta_1 x_i)$ specifica per ciascuna osservazione e con una deviazione standard (σ) comune a tutte le osservazioni. Poiché include un solo predittore (x), questo modello è comunemente chiamato *modello di regressione lineare semplice*.



Commenti e considerazioni finali

Il modello lineare semplice viene usato per descrivere la relazione tra due variabili e per determinare il segno e l'intensità di tale relazione. Inoltre, il modello lineare ci consente di prevedere il valore della variabile dipendente in base ai valori della variabile indipendente.

2

Modello lineare in Stan

Mostreremo qui come sia possibile usare il linguaggio probabilistico Stan per la stima dei parametri del modello di regressione. Vedremo anche come interpretare i risultati dell'analisi bayesiana (si vedano anche le Appendici **??** e **??**).

2.1 Una distribuzione a priori debolmente informativa

Per implementare l'approccio bayesiano è necessario assegnare una distribuzione a priori ai parametri. Nel contesto del modello di regressione è desiderabile scegliere distribuzioni a priori che abbiano uno scarso impatto sulla distribuzione a posteriori.

Supponiamo che le nostre credenza a priori sui parametri del modello, β_0, β_1 e σ siano tra loro indipendenti. Allora possiamo scrivere la distribuzione congiunta dei parametri nel modo seguente:

$$p(\beta_0, \beta_1, \sigma) = p(\beta_0)p(\beta_1)p(\sigma).$$

Possiamo dunque assumere $\beta_0 \sim \mathcal{N}(\mu_0, s_0)$ e $\beta_1 \sim \mathcal{N}(\mu_1, s_1)$. Per σ possiamo assumere $\sigma \sim \text{Cauchy}(a, b)$. Moltiplicando la verosimiglianza

$$\prod_{i=1}^{n} p(y_i \mid x_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$$

per le distribuzioni a priori dei parametri, si ottiene la distribuzione a posteriori. Tuttavia, tale distribuzione non è risolvibile per via analitica. È dunque necessario utilizzare un algoritmo MCMC per ottenere una sequenza di campioni casuali dalla distribuzione a posteriori.

2.2 Linguaggio Stan

Leggiamo in R il dataset `kidiq`:

```
library("rio")
df <- rio::import(here::here("data", "kidiq.dta"))
head(df)
#>   kid_score mom_hs mom_iq mom_work mom_age
#> 1        65      1 121.12        4      27
#> 2        98      1  89.36        4      25
#> 3        85      1 115.44        4      27
#> 4        83      1  99.45        3      25
#> 5       115      1  92.75        4      27
#> 6        98      0 107.90        1      18
```

Vogliamo descrivere l'associazione tra il QI dei figli e il QI delle madri mediante un modello lineare. Per farci un'idea del valore dei parametri, adattiamo il modello lineare ai dati mediante la procedura di massima verosimiglianza:

```
summary(lm(kid_score ~ mom_iq, data = df))
#>
#> Call:
#> lm(formula = kid_score ~ mom_iq, data = df)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -56.75 -12.07   2.22  11.71  47.69
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  25.7998     5.9174    4.36  1.6e-05 ***
#> mom_iq        0.6100     0.0585   10.42  < 2e-16 ***
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
```

```
#> Residual standard error: 18.3 on 432 degrees of freedom
#> Multiple R-squared:  0.201,  Adjusted R-squared:  0.199
#> F-statistic:  109 on 1 and 432 DF,  p-value: <2e-16
```

Sulla base delle informazioni precedenti, giungiamo alla seguente formulazione bayesiana del modello lineare:

$$
\begin{aligned}
y_i &\sim \mathcal{N}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 x_i \\
\beta_0 &\sim \mathcal{N}(25, 10) \\
\beta_1 &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \text{Cauchy}(18, 5)
\end{aligned}
$$

La prima riga definisce la funzione di verosimiglianza e le righe successive definiscono le distribuzioni a priori dei parametri. Il segno $\sim$ (tilde) si può leggere "si distribuisce come". La prima riga ci dice che ciascuna osservazione y_i è una variabile casuale che segue la distribuzione gaussiana di parametri μ_i e σ. La seconda riga specifica, in maniera deterministica, che ciascun μ_i è una funzione lineare di x_i, con parametri β_0 e β_1. Le due righe successive specificano le distribuzioni a priori per β_0 e β_1. La distribuzione a priori di β_0 è una distribuzione gaussiana di parametri $\mu_\alpha = 25$ e deviazione standard $\sigma_\alpha = 10$; la distribuzione a priori di β_1 è una distribuzione gaussiana standardizzata. L'ultima riga definisce la distribuzione a priori di σ, ovvero una Cauchy di parametri 18 e 5.

Dobbiamo ora specificare il modello bayesiano descritto sopra in linguaggio Stan[1]. Il codice Stan viene eseguito più velocemente se l'input è standardizzato così da avere una media pari a zero e una varianza unitaria.[3]

[1] Nella discussione che segue ripeto pari pari ciò che è riportato nel manuale del linguaggio Stan[2].

[3] Si noti un punto importante. Il fatto di standardizzare i dati fa in modo che le distribuzioni a priori sui parametri vadano espresse sulla scala delle v.c. normali standardizzate. Se centriamo sullo 0 tali distribuzioni a priori, con una deviazione standard dell'ordine di grandezza dell'unità, i discorsi sull'arbitrarietà delle distribuzioni a priori perdono di significato: nel caso di dati standardizzati le distribuzioni a priori formulate come indicato sopra sono distribuzioni debolmente informative il cui unico scopo è la regolarizzazione dei dati, ovvero di mantenere le inferenze in una gamma ragionevole di valori. L'uso di distribuzioni a priori debolmente informative contribuisce nel contempo a limitare l'influenza eccessiva delle osservazioni estreme (valori anomali). Il punto importante qui è che tali distribuzioni a priori

Ponendo $y = (y_1, \dots, y_n)$ e $x = (x_1, \dots, x_n)$, il modello lineare può essere scritto come

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

dove

$$\varepsilon_i \sim \mathcal{N}(0, \sigma).$$

Seguendo la notazione del manuale Stan, i parametri del modello lineare sono qui denotati da α e β. Per eseguire la standardizzazione dei dati, è necessario centrare i dati, sottraendo da essi la media campionaria, per poi scalarli dividendo per la deviazione standard campionaria. Una singola osservazione u viene standardizzata dalla funzione z definita da

$$z_y(u) = \frac{u - \bar{y}}{\texttt{sd}(y)}$$

dove la media $\bar{y}$ è

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

e la deviazione standard è

$$\texttt{sd} = \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{-\frac{1}{2}}.$$

La trasformata inversa è definita invertendo i due passaggi precedenti: la deviazione standard è usata per scalare i valori u e la media campionaria è usata per traslare la distribuzione dei valori u scalati:

$$z_y^{-1}(u) = \texttt{sd}(y)u + \bar{y}.$$

Consideriamo il seguente modello iniziale in linguaggio Stan:

non introducono alcuna distorsione sistematica nella stima a posteriori.

```
modelString <- "
data {
  int<lower=0> N;
  vector[N] y;
  vector[N] x;
}
parameters {
  real alpha;
  real beta;
  real<lower=0> sigma;
}
model {
  // priors
  alpha ~ normal(25, 10);
  beta ~ normal(0, 1);
  sigma ~ cauchy(18, 5);
  // likelihood
  for (n in 1:N)
    y[n] ~ normal(alpha + beta * x[n], sigma);
}
"
writeLines(modelString, con = "code/simpleregkidiq.stan")
```

La funzione `modelString()` registra una stringa di testo mentre `writeLines()` crea un file nell'indirizzo specificato. Tale file deve avere l'estensione `.stan`.

Modificando il codice precedente otteniamo il modello Stan per dati standardizzati. Il blocco `data` è identico a quello del caso precedente. I predittori e la risposta standardizzati sono definiti nel blocco `transformed data`. Per semplificare la notazione (e velocizzare l'esecuzione), nel blocco `model` l'istruzione di campionamento è espressa in forma vettorializzata: `y_std ~ normal(alpha_std + beta_std * x_std, sigma_std);`.

```
modelString <- "
data {
  int<lower=0> N;
  vector[N] y;
```

```
  vector[N] x;
}
transformed data {
  vector[N] x_std;
  vector[N] y_std;
  x_std = (x - mean(x)) / sd(x);
  y_std = (y - mean(y)) / sd(y);
}
parameters {
  real alpha_std;
  real beta_std;
  real<lower=0> sigma_std;
}
transformed parameters {
  vector[N] mu_std = alpha_std + beta_std * x_std;
}
model {
  alpha_std ~ normal(0, 1);
  beta_std ~ normal(0, 1);
  sigma_std ~ normal(0, 1);
  y_std ~ normal(mu_std, sigma_std);
}
generated quantities {
  // transform to the original data scale
  real alpha;
  real beta;
  real<lower=0> sigma;
  alpha = sd(y) * (alpha_std - beta_std * mean(x) / sd(x)) + mean(y);
  beta = beta_std * sd(y) / sd(x);
  sigma = sd(y) * sigma_std;
}
"
writeLines(modelString, con = "code/simpleregstd.stan")
```

Si noti che i parametri vengono rinominati per indicare che non sono i parametri "naturali", ma per il resto il modello è identico. Sono qui utilizzate distribuzioni a priori debolmente informative per i parametri `alpha` e `beta`.

I valori dei parametri sulla scala originale dei dati vengono calcolati nel blocco `generated quantities` e possono essere recuperati con un po' di algebra.

$$\begin{aligned}
y_n &= \mathrm{z}_y^{-1}(\mathrm{z}_y(y_n)) \\
&= \mathrm{z}_y^{-1}\left(\alpha' + \beta' \mathrm{z}_x(x_n) + \epsilon'_n\right) \\
&= \mathrm{z}_y^{-1}\left(\alpha' + \beta'\left(\frac{x_n - \bar{x}}{\mathtt{sd}(x)}\right) + \epsilon'_n\right) \\
&= \mathtt{sd}(y)\left(\alpha' + \beta'\left(\frac{x_n - \bar{x}}{\mathtt{sd}(x)}\right) + \epsilon'_n\right) + \bar{y} \\
&= \left(\mathtt{sd}(y)\left(\alpha' - \beta'\frac{\bar{x}}{\mathtt{sd}(x)}\right) + \bar{y}\right) + \left(\beta'\frac{\mathtt{sd}(y)}{\mathtt{sd}(x)}\right) x_n + \mathtt{sd}(y)\epsilon'_n, \quad (2.1)
\end{aligned}$$

da cui

$$\alpha = \mathtt{sd}(y)\left(\alpha' - \beta'\frac{\bar{x}}{\mathtt{sd}(x)}\right) + \bar{y}; \qquad \beta = \beta'\frac{\mathtt{sd}(y)}{\mathtt{sd}(x)}; \qquad \sigma = \mathtt{sd}(y)\sigma'.$$

Per svolgere l'analisi bayesiana sistemiamo i dati nel formato appropriato per Stan:

```
data_list <- list(
  N = length(df$kid_score),
  y = df$kid_score,
  x = df$mom_iq
)
```

La funzione `file.path()` ritorna l'indirizzo del file con il codice Stan:

```
file <- file.path("code", "simpleregstd.stan")
```

Il codice Stan può essere stampato usando il metodo `$print()`:

```
mod$print()
```

Prendendo come input un file contenente un programma Stan, la funzione `cmdstan_model()` ritorna un oggetto di classe `CmdStanModel`. In pra-

tica, `CmdStan` traduce un programma Stan in C++ e crea un eseguibile compilato.

```
mod <- cmdstan_model(file)
```

L'indirizzo dell'eseguibile compilato viene ritornato da `$exe_file()`:

```
mod$exe_file()
```

Applicando il metodo `$sample()` ad un oggetto `CmdStanModel` eseguiamo il campionamento MCMC:

```
fit <- mod$sample(
  data = data_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  chains = 4L,
  parallel_chains = 2L,
  refresh = 0,
  thin = 1
)
```

Al metodo `$sample()` possono essere passati molti argomenti. La pagina di documentazione è disponibile al seguente link[4].

Un sommario della distribuzione a posteriori per i parametri stimati si ottiene con il metodo `$summary()`, il quale chiama la funzione `summarise_draws()` del pacchetto `posterior`:

```
fit$summary(c("alpha", "beta", "sigma"))
#> # A tibble: 3 x 10
#>   variable   mean median     sd    mad     q5    q95
#>   <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
#> 1 alpha    25.9   25.8   6.02   6.02   16.0   35.8
#> 2 beta      0.609  0.609 0.0596 0.0603  0.511  0.707
#> 3 sigma    18.3   18.3   0.634  0.644  17.3   19.4
```

[4] https://mc-stan.org/cmdstanr/reference/model-method-sample.html

```
#> # ... with 3 more variables: rhat <dbl>,
#> #   ess_bulk <dbl>, ess_tail <dbl>
```

Da questo output possiamo valutare rapidamente la convergenza del modello osservando i valori di Rhat per ciascun parametro. Quando questi sono pari o vicini a 1, le catene hanno realizzato la convergenza. Ci sono molti altri test diagnostici, ma questo test è importante per Stan. Oppure è possibile usare:

```
fit$cmdstan_summary()
```

Le statistiche diagnostiche sono fornite dal metodo `$cmdstan_diagnose()`:

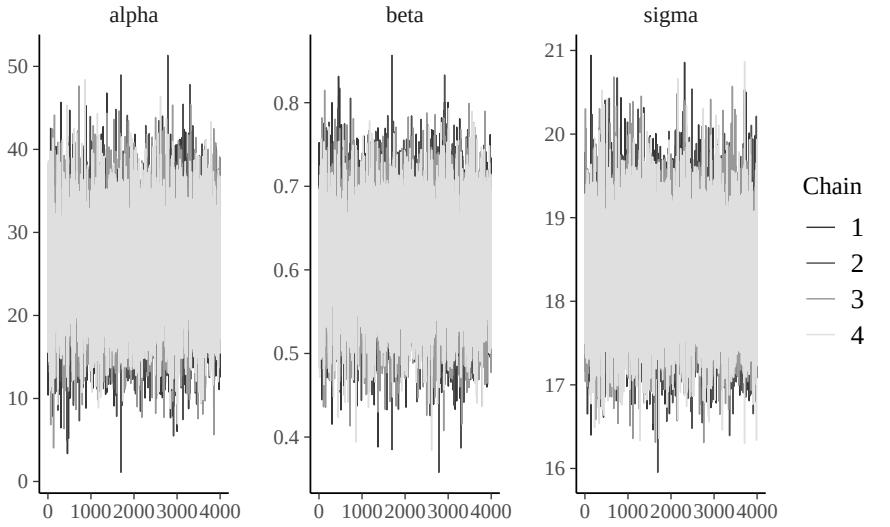
```
fit$cmdstan_diagnose()
```

È possibile creare un oggetto di classe `stanfit`

```
stanfit <- rstan::read_stan_csv(fit$output_files())
```

per poi utilizzare le funzioni del pacchetto `bayesplot`. Ad esempio:

```
stanfit %>%
  mcmc_trace(pars = c("alpha", "beta", "sigma"))
```



Infine, eseguendo la funzione `launch_shinystan(fit)`, è possibile analizzare oggetti di classe `stanfit` mediante le funzionalità del pacchetto `ShinyStan`.

2.3 Interpretazione dei parametri

Assegnamo ai parametri la seguente interpretazione.

- L'intercetta pari a 25.9 indica il QI medio dei bamini la cui madre ha un QI = 0. Ovviamente questo non ha alcun significato. Vedremo nel modello successivo come trasformare il modello in modo da potere assegnare all'intercetta un'interpretazione sensata.
- La pendenza di 0.61 indica che, all'aumentare di un punto del QI delle madri, il QI medio dei loro bambini aumenta di 0.61 unità. Se consideriamo la gamma di variazione del QI delle madri nel campione, il QI medio dei bambini cambia di 41 punti. Questo indica un sostanziale effetto del QI delle madri sul QI dei loro bambini: $(138.89 - 71.04) * 0.61 = 41.39$.
- Il parametro $\sigma = 18.3$ fornisce una stima della dispersione delle osservazioni attorno al valore predetto dal modello lineare, ovvero forni-

sce una stima della deviazione standard dei residui attorno al valore atteso del modello lineare.

2.3.1 Centrare i predittori

Per migliorare l'interpretazione dell'intercetta possiamo "centrare" la x, ovvero esprimere la x nei termini degli scarti dalla media: $x - \bar{x}$. In tali circostanze, la pendenza della retta specificata dal modello lineare resta immutata, ma l'intercetta corrisponde a $\mathbb{E}(y \mid x = \bar{x})$. Per ottenere questo risultato, modifichiamo i dati da passare a Stan:

```
data2_list <- list(
  N = length(df$kid_score),
  y = df$kid_score,
  x = df$mom_iq - mean(df$mom_iq)
)
```

Adattiamo il modello:

```
fit2 <- mod$sample(
  data = data2_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  chains = 4L,
  parallel_chains = 2L,
  refresh = 0,
  thin = 1
)
```

Trasformiamo l'oggetto `fit` in un oggetto di classe `stanfit`:

```
stanfit <- rstan::read_stan_csv(fit2$output_files())
```

Le stime a posteriori dei parametri si ottengono con

```
fit2$summary(c("alpha", "beta", "sigma"))
#> # A tibble: 3 x 10
```

```
#>   variable   mean median     sd    mad     q5    q95
#>   <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
#> 1 alpha    86.8   86.8   0.876  0.871  85.4   88.2
#> 2 beta      0.609  0.609 0.0591 0.0589  0.513  0.707
#> 3 sigma    18.3   18.3   0.630  0.624  17.3   19.4
#> # ... with 3 more variables: rhat <dbl>,
#> #   ess_bulk <dbl>, ess_tail <dbl>
```

Si noti la nuova intercetta, ovvero 86.8. Questo valore indica il QI medio dei bambini le cui madri hanno un QI pari alla media del campione. Centrare i dati consente dunque di assegnare all'intercetta un'interpretazione utile.

Commenti e considerazioni finali

La presente discussione suggerisce che è conveniente standardizzare i dati prima di procedere con l'analisi. Ciò può essere fatto all'interno del codice Stan (come negli esempi di questo Capitolo), oppure prima di passare i dati a Stan. Se vengono usati dati standardizzati diventa poi facile utilizzare distribuzioni a priori debolmente informative per i parametri. Tali distribuzioni a priori hanno, come unico scopo, quello di regolarizzare i dati e di facilitare la stima dei parametri mediante MCMC.

Bibliografia

Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Horn, S. and Loewenstein, G. (2021). Underestimating learning by doing. *Available at SSRN 3941441*.