Statistical Modeling - Prova 7

Dario Comanducci, 19 maggio 2024

1 Parte A

1.1 Analisi preliminare

Il dataset `MichelinNY.txt` è composto da 164 ristoranti, presenti o meno nella Guida Michelin (Fig. 1a), per i quali viene riportato il prezzo medio Price dei loro piatti (Fig. 1b), oltre ai voti (su una scala da 1 a 30) per le categorie Food, Decor e Service (Fig. 1c).

Dallo scatterplot a matrice di Fig. 2 osserviamo che Price presenta due outlier (pari a 179 e 201\$) suscettibili di effetto leva, dovuti a due ristoranti nella Guida Michelin, con prezzi molto maggiori rispetto al resto dell'intera collezione; pertanto le corrispondenti voci nel dataset sono state rimosse per il resto della trattazione.[1]

Sempre da Fig. 2 notiamo altresì che sui dati complessivi esiste una forte correlazione[2] tra Food e Service ($\rho = 0.8$), di cui occorrerà tener conto nell'analisi del modello lineare che spiega Price in funzione di Michelin, Food, Decor e Service.

Il nostro dataset contiene chiaramente una variabile categorica binaria (Michelin) ed una variabile continua (Price); per quanto riguarda i punteggi di Food, Decor e Service, dato che sono valori ordinabili (con una scala piuttosto fine di 30 livelli) ed è plausibile ritenere costante la variazione tra un valore ed il successivo, decidiamo di interpretare tali grandezze come continue.

1.2 Modellazione

Poiché siamo interessati ad analizzare la dipendenza di Price in funzione di Michelin, Food, Decor e Service, e al fine di dare un significato all'intercetta delle ordinate da parte del modello prodotto, le variabili esplicative Food, Decor e Service verranno rese a media nulla: attraverso questa trasformazione dei dati, avremo che l'intercetta delle ordinate da parte del modello sarà interpretabile come il prezzo predetto quando un ristorante ha voti nella media.[3]

[1] Per i due outlier la tripletta di voti (Food, Decor, Service) vale rispettivamente $(27, 27, 27)$ e $(28, 27, 28)$.

[2] Come sarà descritto più avanti (Fig. 8), questo fenomeno è ancora più marcato nel caso dei ristoranti Michelin ($\rho = 0.9$); nei ristoranti non presenti nella guida invece l'indice di correlazione cala a $\rho = 0.54$.

[3] Inoltre, i voti partono da 1 per cui un punteggio pari a zero non sarebbe contemplato; se anche lo fosse, si può notare da Fig. 1c come i voti a disposizione siano piuttosto distanti da 0 (sono sempre almeno maggiori di 10), per cui l'estrapolazione all'indietro del modello per voti nulli avrebbe qualche problema di tenuta. In ogni caso è sempre possibile ricondursi al modello con le variabili originali dato che $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(x_1 - \bar{x}_1) + \cdots + \hat{\beta}_p(x_p - \bar{x}_p) = (\hat{\beta}_0 - \hat{\beta}_1\bar{x}_1 - \cdots - \hat{\beta}_p\bar{x}_p) + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

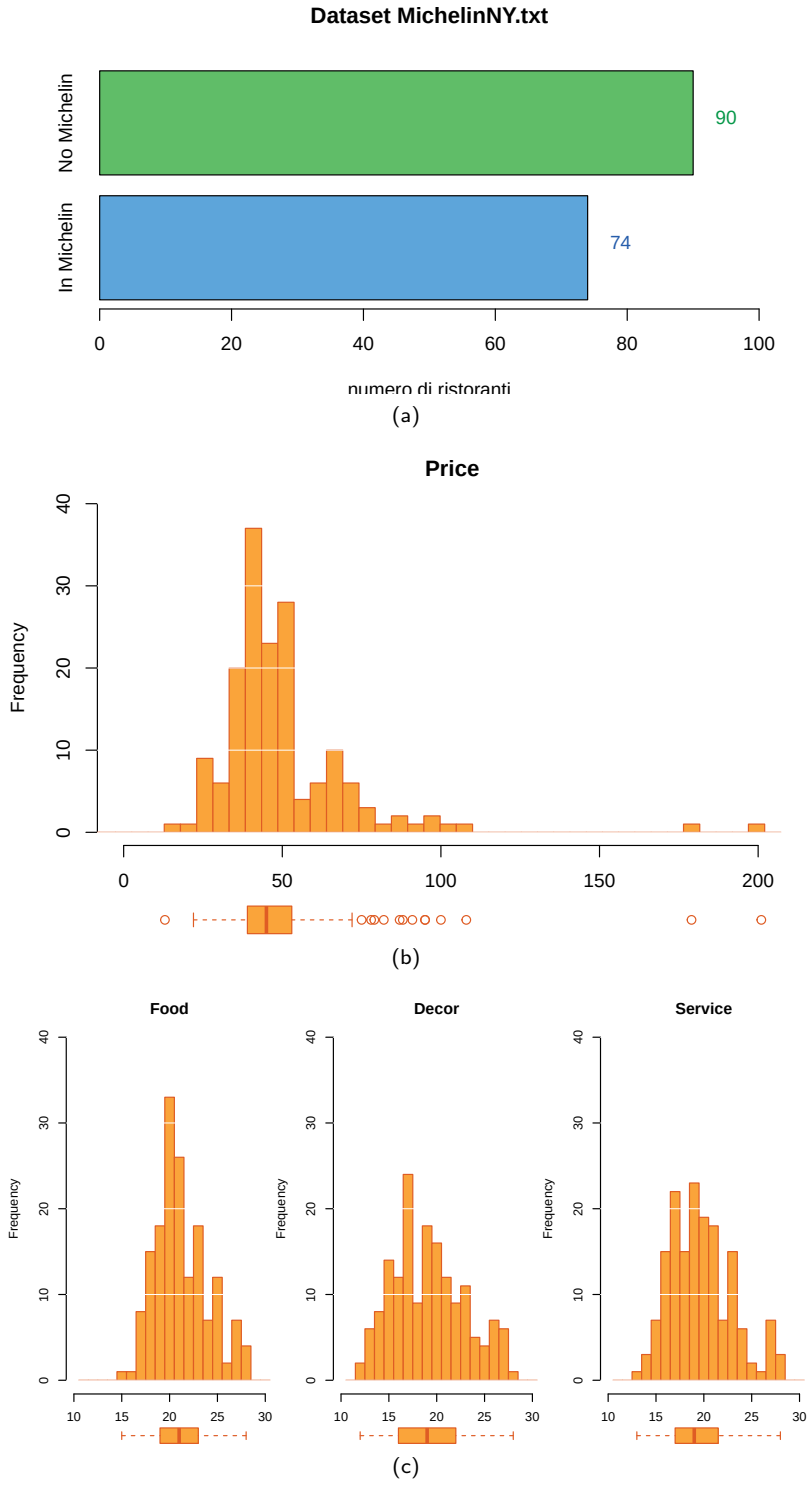


Fig. 1: In (a), la ripartizione del dataset `MichelinNY.txt` rispetto alla variabile Michelin. In (b), le occorrenze della variabile Price ed il relativo boxplot della distribuzione; in (c) quelle dei punteggi Food, Decor e Service.

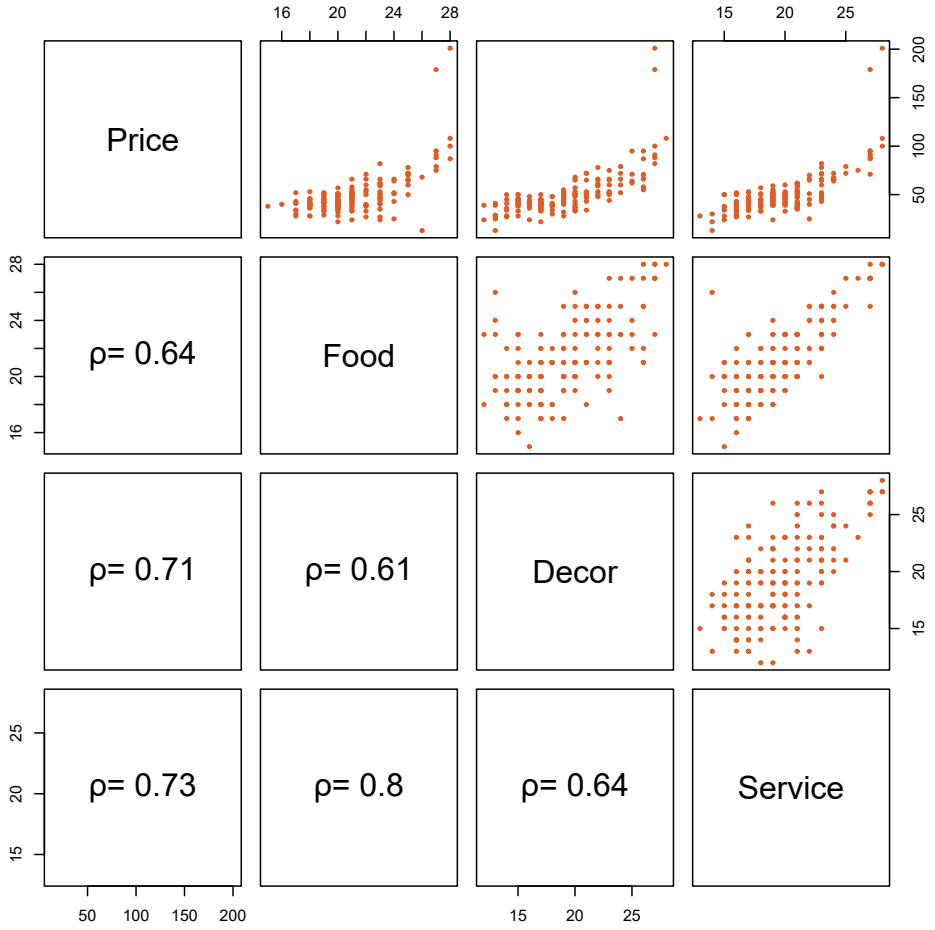


Fig. 2: Scatterplot a matrice per le variabili Price, Food, Decor e Service. Nella parte inferiore della matrice vengono riportati i coefficienti di correlazione per ogni coppia di variabili considerate.

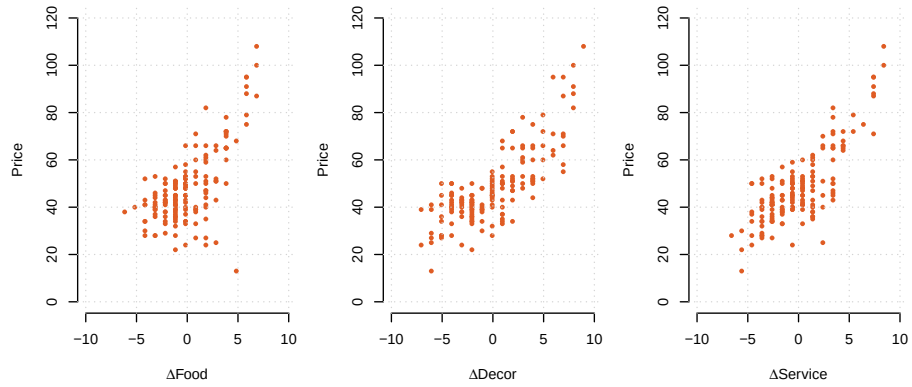


Fig. 3: Grafici di dispersione di Price in funzione di ΔFood, ΔDecor e ΔService. Nel calcolo dei valori medi (Food: 21.167; Decor: 19.068; Service: 19.605) sono stati estromessi i due ristoranti con prezzi outlier.

In virtù delle precedenti considerazioni introduciamo quindi per il nostro dataset, con $N = 164 - 2$ voci, le seguenti variabili

$$\begin{aligned} y &= \mathsf{Price} \in \mathbb{R}^+ \\ x_1 &= \mathsf{Food} \in \{1 \dots 30\} \\ x_2 &= \mathsf{Decor} \in \{1 \dots 30\} \\ x_3 &= \mathsf{Service} \in \{1 \dots 30\} \\ x_4 &= \mathsf{Michelin} \in \{0, 1\} \end{aligned}$$

da cui definiamo per $k = 1 \dots 3$

$$\begin{aligned} \tilde{x}_k &= x_k - \bar{x}_k \\ \bar{x}_k &= \frac{1}{N} \sum_{n=1}^{N} x_{kn} \end{aligned}$$

A tali variabili a media nulla associamo anche i seguenti alias: $\Delta\mathsf{Food}$ per $\tilde{x}_1$, $\Delta\mathsf{Decor}$ per $\tilde{x}_2$, $\Delta\mathsf{Service}$ per $\tilde{x}_3$. Fig. 3 illustra i relativi grafici di dispersione rispetto a Price.

Pertanto per l'n-esimo elemento del dataset avremo che il nostro modello lineare assume la forma

$$\begin{aligned} y_n &= \beta_0 + \beta_1 \tilde{x}_{1n} + \beta_2 \tilde{x}_{2n} + \beta_3 \tilde{x}_{3n} + \beta_4 x_{4n} + \epsilon_n \qquad \text{con} \\ \epsilon_n &= \text{componente casuale (a media nulla)} \in \mathbb{R} \end{aligned}$$

o, in forma più compatta,

$$y_n = \mathbf{x}_n^\top \boldsymbol{\beta} + \epsilon_n \tag{1a}$$

$$\mathbf{x}_n = \begin{bmatrix} 1 & \tilde{x}_{1n} & \tilde{x}_{2n} & \tilde{x}_{3n} & x_{4n} \end{bmatrix}^\top \tag{1b}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{bmatrix}^\top \tag{1c}$$

Infine, l'applicazione di Eq. (1) all'intero dataset di N elementi produce il sistema di equazioni

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{y} \in \mathbb{R}^N} = \underbrace{\begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix}}_{\mathtt{X} \in \mathbb{R}^{N \times 5}} \boldsymbol{\beta} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}}_{\boldsymbol{\epsilon} \in \mathbb{R}^N} \tag{2}$$

1.3 Stima del modello

Il risultato ai minimi quadrati per Eq. (2) è ricavabile attraverso le equazioni normali

$$\hat{\boldsymbol{\beta}} = (\mathtt{X}^\top \mathtt{X})^{-1} \mathtt{X}^\top \mathbf{y} \tag{3}$$

ottenendo i valori riportati in Tab. 1, oltre ai relativi errori standard e p-value. Per valutare gli errori standard $\mathrm{se}(\hat{\beta}_k)$, è stata determinata la matrice

$$\begin{aligned} \mathbb{V}(\hat{\boldsymbol{\beta}} | \mathtt{X}) &= S^2 (\mathtt{X}^\top \mathtt{X})^{-1} \quad \text{con} \\ S^2 &= \frac{(\mathbf{y} - \mathtt{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathtt{X}\hat{\boldsymbol{\beta}})}{N - (p+1)} \qquad (\text{dove } N = 162,\ p = 4) \end{aligned}$$

Tab. 1: Valori per $\hat{\boldsymbol{\beta}}$ determinati attraverso Eq. (3), con relativi errori standard e p-value, a confronto con i risultati forniti dalla funzione `lm()` di R.

Metodo	k	$\hat{\beta}_k$	$\text{se}(\hat{\beta}_k)$	$t_{\hat{\beta}_k}$	p-value
Eq. (3)	0 (intercetta)	46.9386	0.8604	54.556	$5.696719 \cdot 10^{-104}$
	1 (ΔFood)	-0.2266	0.3637	-0.623	$5.342026 \cdot 10^{-01}$
	2 (ΔDecor)	1.8799	0.2142	8.776	$2.700553 \cdot 10^{-15}$
	3 (ΔService)	2.5450	0.3111	8.182	$8.962264 \cdot 10^{-14}$
	4 (Michelin)	3.2214	1.4334	2.247	$2.600529 \cdot 10^{-02}$
`lm(...)`	0 (intercetta)	46.9386	0.8604	54.556	$< 2 \cdot 10^{-16}$
	1 (ΔFood)	-0.2266	0.3637	-0.623	0.534
	2 (ΔDecor)	1.8799	0.2142	8.776	$2.70 \cdot 10^{-15}$
	3 (ΔService)	2.5450	0.3111	8.182	$8.96 \cdot 10^{-14}$
	4 (Michelin)	3.2214	1.4334	2.247	0.026

per poi prendere la radice quadrata degli elementi sulla diagonale di $\mathbb{V}(\hat{\boldsymbol{\beta}}|\mathtt{X})$. Per il calcolo dei p-value, occorre valutare la statistica $t_{\beta_k} = \beta_k/\text{se}(\beta_k) \sim t_{N-p}$, su cui valutare

$$p\text{-value} = 2\mathbb{P}\big[t_{\beta_k} > |t_{\hat{\beta}_k}|\big]$$

Tab. 1 mostra che tutti i valori trovati sono equivalenti a quelli determinati tramite la funzione `lm()` di `R`.

1.4 Discussione del modello

Nonostante la statistica F riporti un p-value $< 2.2 \cdot 10^{-16}$, per il modello ottenuto abbiamo $R^2_{\text{adj}} = 0.7817$ che rappresenta solo un livello discreto come indice di qualità.

Da Tab. 1 osserviamo che a ΔFood corrisponde un p-value pari 0.534: questo valore molto alto indica che il relativo coefficiente regressione $\hat{\beta}_2 = -0.2266045$ non è significativamente diverso da 0. La causa di ciò è imputabile all'elevata correlazione che Food ha con Service.

Sebbene rientri nei limiti di accettabilità, anche il p-value corrispondente alla variabile binaria Michelin (ossia x_4) assume un valore piuttosto elevato (0.026). Volendo dare una spiegazione a tale valore, per come x_4 è linearmente inserita in Eq. (1) si ottengono due diversi modelli a seconda che un ristorante sia sulla Guida Michelin oppure no, che differiscono solo sul valore dell'intercetta; x_4 incide solo in maniera marginale ($\hat{\beta}_0 = 46.9386$ mentre $\hat{\beta}_4 = 3.2214$) nel caso dei ristoranti Michelin:

$$\hat{y} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 & (x_4 = 0) \\ (\hat{\beta}_0 + \hat{\beta}_4) + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 & (x_4 = 1) \end{cases}$$

1.4.1 Correzione al modello lineare e analisi dei residui

Poiché la stima è contaminata dalla correlazione tra Food e Service, analizziamo cosa accade rimuovendo la variabile ΔFood (ossia $\tilde{x}_1$) dal modello:

$$y_n = \; e_n + \beta_0 + \beta_2 \tilde{x}_{2n} + \beta_3 \tilde{x}_{3n} + \beta_4 x_{4n} \tag{4}$$

Tab. 2: Stime $\hat{\boldsymbol{\beta}}$ per Eq. (4), con relativi errori standard e p-value.

k	$\hat{\beta}_k$	se$(\hat{\beta}_k)$	$t_{\hat{\beta}_k}$	p-value
0 (intercetta)	47.0307	0.8459	55.595	$< 2 \cdot 10^{-16}$
2 (ΔDecor)	1.8644	0.2124	8.780	$2.55 \cdot 10^{-15}$
3 (ΔService)	2.4188	0.2355	10.269	$< 2 \cdot 10^{-16}$
4 (Michelin)	3.0143	1.3916	2.166	0.0318

Tab. 2 riporta i nuovi parametri per Eq. (4) impiegando la funzione `lm(...)`, ottenendo $R^2_{\text{adj}} = 0.7825$ (p-value della statistica F sempre $< 2.2 \cdot 10^{-16}$): le stime per i vari β_k sono paragonabili a quelle di Tab. 1, ma adesso l'errore standard per $\hat{\beta}_3$ (relativo a ΔService) è inferiore così come di conseguenza anche il corrispondente p-value; il p-value di Michelin è invece ulteriormente peggiorato ma sempre entro i limiti di accettabilità.

L'analisi grafica dei residui[4]

$$\hat{\epsilon}_n = y_n - \hat{y}_n \qquad (\text{dove } \hat{y}_n = \mathbf{e}_n^\top \mathbf{X}\hat{\boldsymbol{\beta}})$$

può rilevare sia se i presupposti della regressione sono soddisfatti, sia quello di evidenziare eventualmente dei pattern di cui tener conto per migliorare la parametrizzazione del modello. Sebbene l'istogramma[5] di Fig. 4 ed il grafico Q-Q plot[6] di Fig. 5a sembrino essere grossomodo in accordo con una distribuzione normale dei residui, lo scatterplot rispetto ai valori predetti di Fig. 5b suggerisce che sia necessario introdurre nel modello delle dipendenze ulteriori.

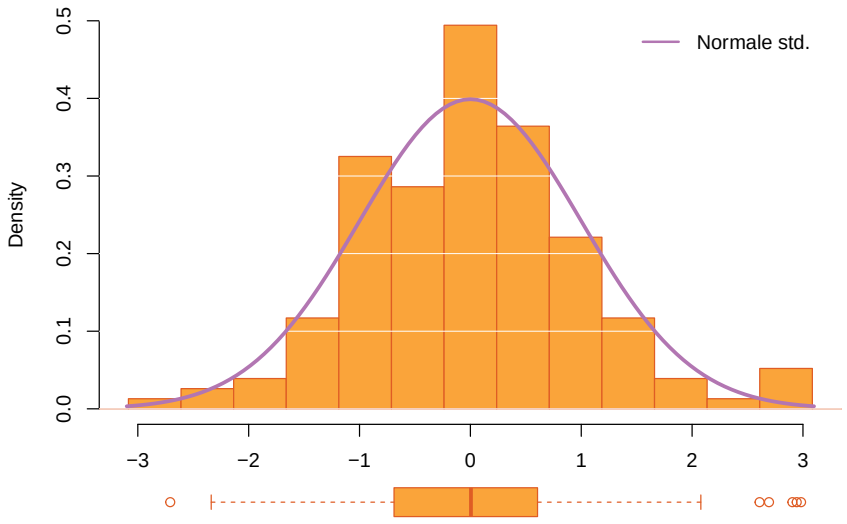
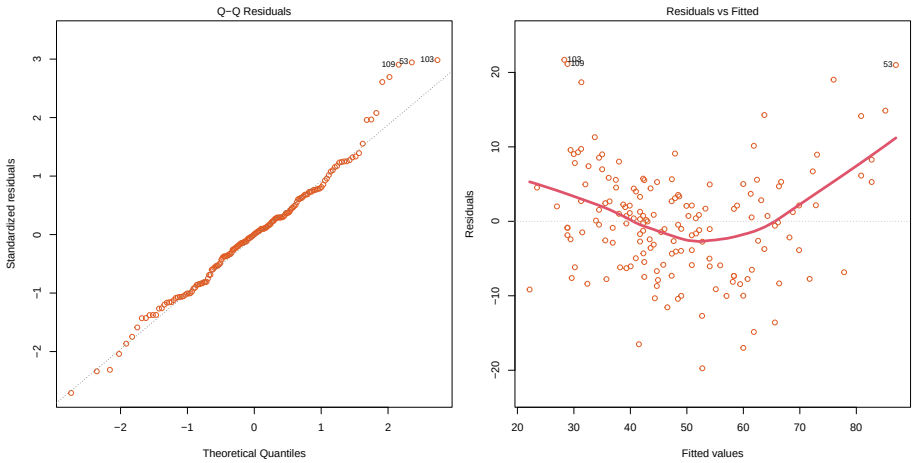


Fig. 4: Residui standardizzati (istogramma e boxplot) per Eq. (4) su Tab. 2.

[4] Il vettore $\mathbf{e}_n$ denota un vettore nullo $\in \mathbb{R}^N$ tranne la n-esima componente pari invece a 1.

[5] I residui standardizzati obbediscono alla normale standard nell'intervallo $(-3, +3)$; inoltre, sotto l'ipotesi di normalità, la mediana è equidistante 1° e 3° quartile, ed il "box" dovrebbe essere centrato tra i due baffi: di fatto si osservano solo alcuni punti un po' esterni ai baffi.

[6] Solo alcuni punti con residui elevati si discostano dalla diagonale del 1° e 3° quadrante.



(a) Diagramma Q-Q plot (b) Residui rispetto ai valori predetti

Fig. 5: Analisi grafica dei residui per il modello definito da Eq. (4) e Tab. 2.

Possiamo approfondire quest'ultimo aspetto visualizzando l'andamento dei residui rispetto alle variabili che compongono il modello (Fig. 6): separando gli scatterplot in base alla variabile Michelin, si osserva che nei ristoranti Michelin appare una chiara dipendenza quadratica tra Price e ΔService.

1.5 Oltre la regressione lineare pura

Possiamo tentare di produrre un modello migliore di Eq. (1) o Eq. (4) osservando separatamente le variabili del dataset all'interno delle due classi di ristoranti definite dalla variabile Michelin (Fig. 7 e Fig. 8).

La separazione dei dati nelle due classi evidenzia che il prezzo di 95$ di un ristorante non Michelin è suscettibile di essere un outlier con effetto leva per quella categoria di ristoranti: a scanso di problemi il ristorante corrispondente viene eliminato dal dataset così come era stato fatto inizialmente per i due ristoranti con prezzi pari a 179 e 201$.[7]

Dai coefficienti di correlazione in Fig. 8 si osserva che, nel caso dei ristoranti Michelin, per Food e Service vale $\rho = 0.9$, mentre nell'altra categoria lo stesso coefficiente cala a $\rho = 0.54$. Notando inoltre che nei ristoranti non Michelin Price è scarsamente correlato con Food, a differenza di Service, oltre a Michelin conviene di nuovo impiegare come variabili esplicative solo Decor e Service (sempre trasformate nella loro versione a media nulla, ΔDecor e ΔService).

Pertanto possiamo tentare di produrre un modello migliore introducendo in Eq. (4) sia degli effetti d'interazione con Michelin $\in \{0, 1\}$ (ossia x_4), sia dei legami quadratici per ΔService ($\tilde{x}_3$):[8]

$$y_n = e_n + \beta_0 + \beta_2 \tilde{x}_{2n} + \beta_3 \tilde{x}_{3n} + \beta_5 \overbrace{(x_{4n}\, \tilde{x}_{3n})}^{x_{5n}} \tag{5}$$

[7] Eliminando fin da subito tale elemento dal dataset non cambia sostanzialmente le prestazioni dei modelli già proposti in Eq. (1) e Eq. (4).

[8] Si ricorda inoltre che $\tilde{x}_2$ corrisponde a ΔDecor. Per brevità è già stata rimossa la dipendenza lineare con Michelin (x_4), a cui corrisponde un p-value pari a 0.0565 se inserita nella stima.

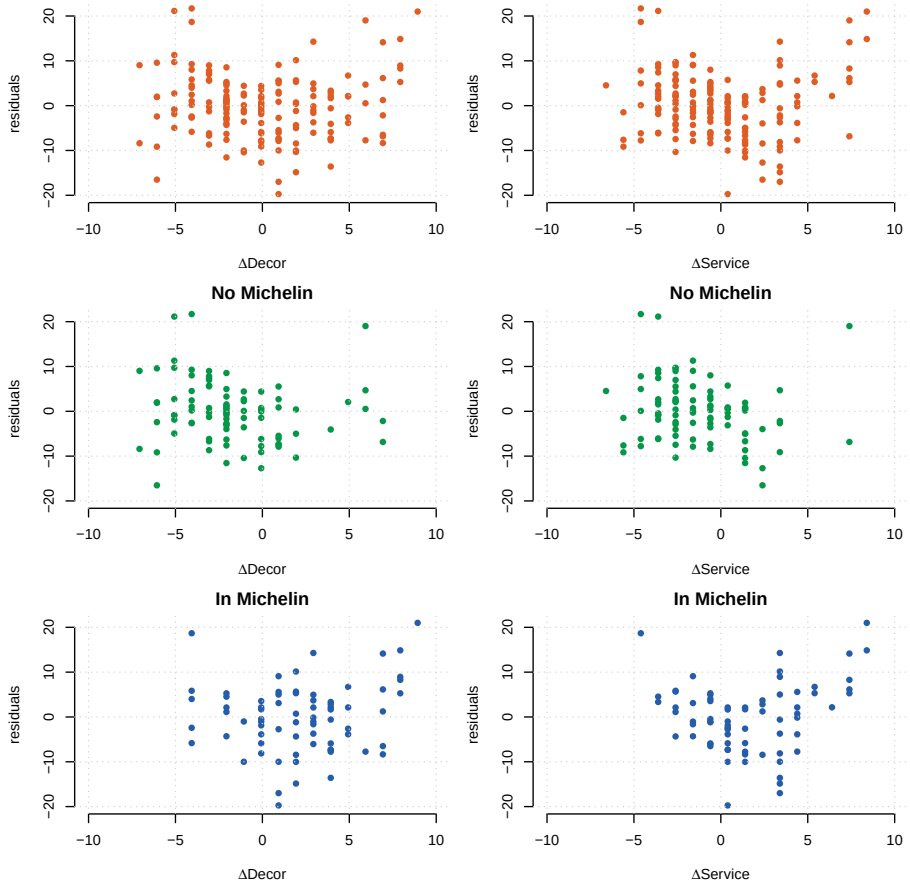


Fig. 6: Residui per il modello in Eq. (4) e Tab. 2 rispetto alle variabili del modello. Nella riga superiore i dati sono presi complessivamente, nelle successive righe sono divisi in base alla presenza o meno sulla Guida Michelin.

I valori ottenuti per Eq. (5) sono riportati in Tab. 3: tutti i $\hat{\beta}_k$ delle variabili considerate sono significativamente diversi da 0. Il modello presenta inoltre $R^2_{\text{adj}} = 0.8322$ (e una statistica F con p-value inferiore a $2.2 \cdot 10^{-16}$): il confronto tra l'analisi grafica dei nuovi residui in Fig. 9 rispetto a Fig. 5 testimonia la maggior qualità del modello trovato, in particolare per l'assenza di trend rilevanti in Fig. 9c. Anche in Fig. 10 il trend parabolico è scomparso; resta tuttavia una curiosa disposizione "verticale" di alcuni residui rispetto a $\Delta\text{service}$ nel caso dei ristoranti Michelin (grossomodo per $\Delta\text{service} \approx 4$), disposizione comunque visibile anche in Fig. 6.

Volendo perfezionare la qualità del modello in Eq. (5) si potrebbe rimuovere dal dataset quei 3 elementi lontani dalla diagonale del QQ-plot di Fig. 9b, con residui di circa 3 deviazioni standard, ed applicare nuovamente la stima dei parametri (omesso per brevità).

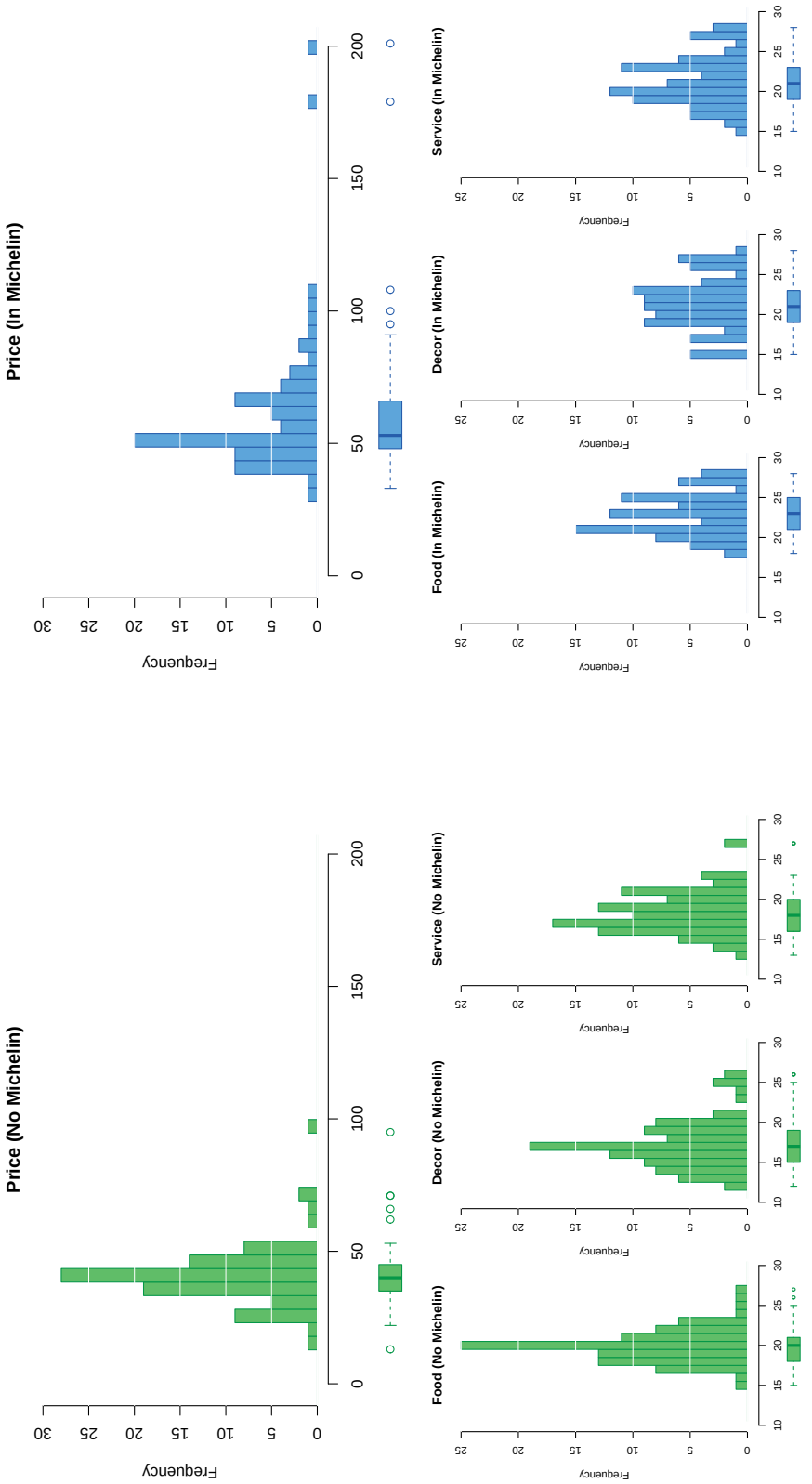


Fig. 7: Occorrenze della variabile Price e dei punteggi per le categorie Food, Decor e Service, divise per tipologia di ristorante.

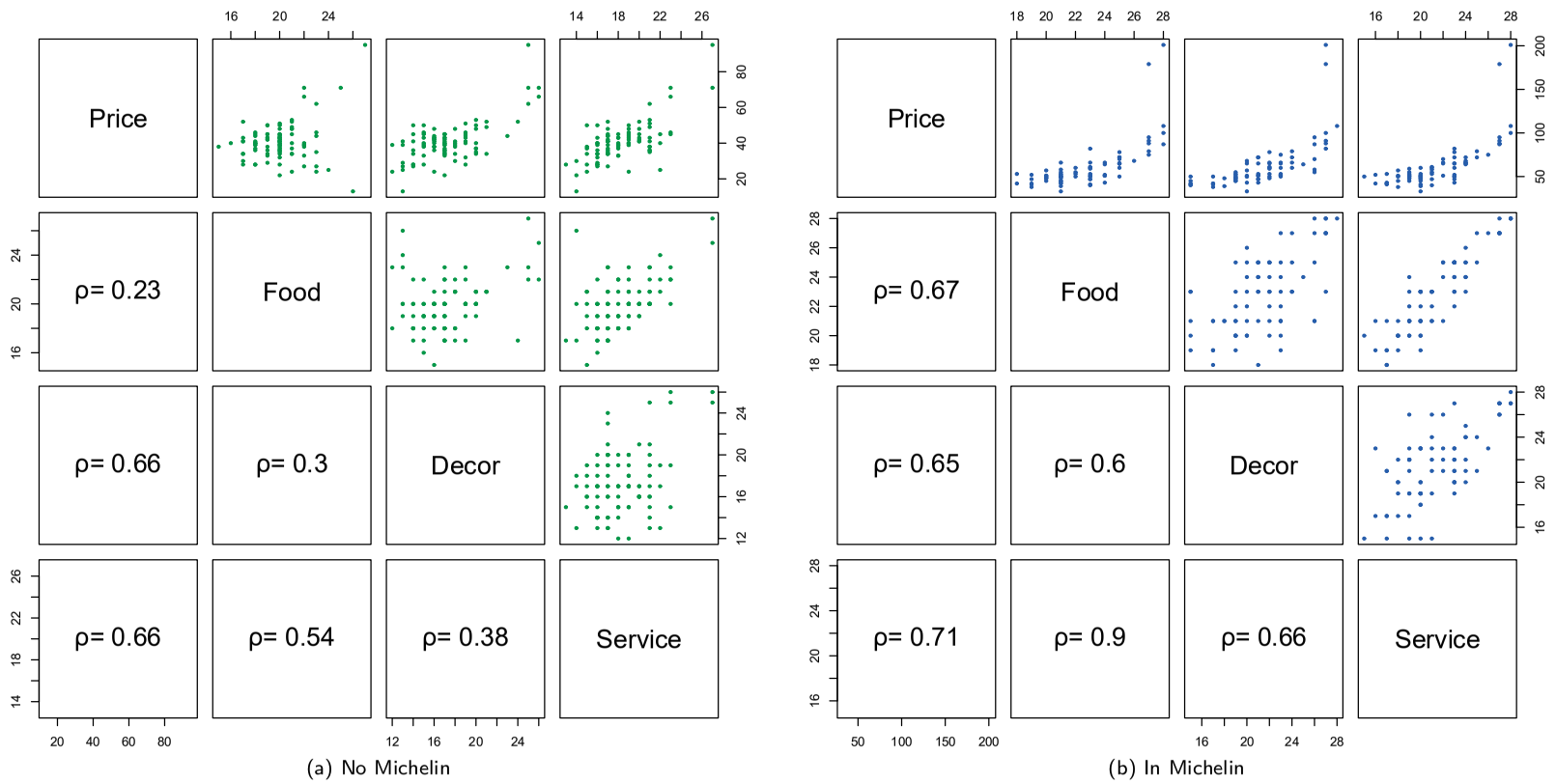
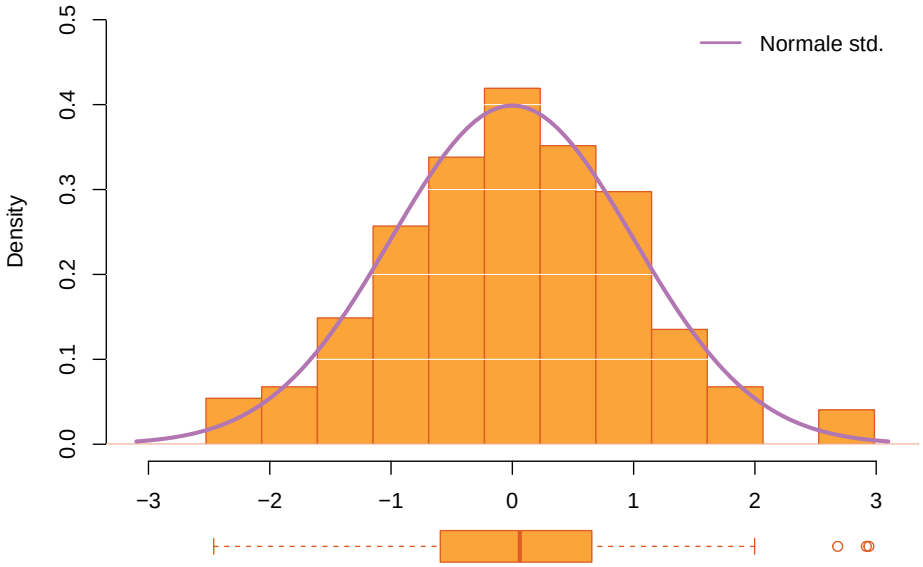
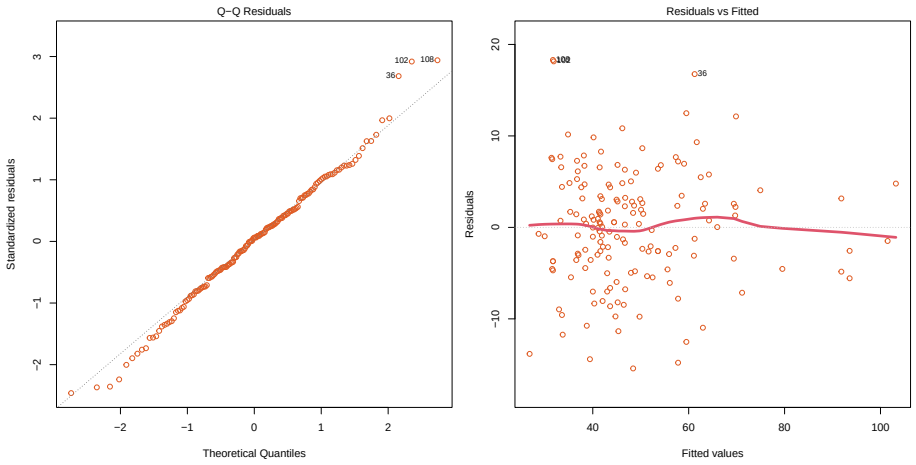


Fig. 8: Scatterplot a matrice per le variabili Price, Food, Decor e Service, con i dati ripartiti a seconda della presenza nella Guida Michelin (a) oppure no (b).



(a) Istogramma e boxplot dei residui standardizzati



(b) Diagramma Q-Q plot

(c) Residui rispetto ai valori predetti

Fig. 9: Analisi grafica dei residui per il modello dato da Eq. (5) e Tab. 3

Tab. 3: Stime $\hat{\boldsymbol{\beta}}$ per Eq. (5), con relativi errori standard e p-value.

k	$\hat{\beta}_k$	$\text{se}(\hat{\beta}_k)$	$t_{\hat{\beta}_k}$	p-value
0	45.97454	0.56804	80.936	$< 2 \cdot 10^{-16}$
2	1.72526	0.17391	9.921	$< 2 \cdot 10^{-16}$
3	1.57261	0.23050	6.823	$1.83 \cdot 10^{-10}$
5	0.39984	0.05255	7.608	$2.40 \cdot 10^{-12}$

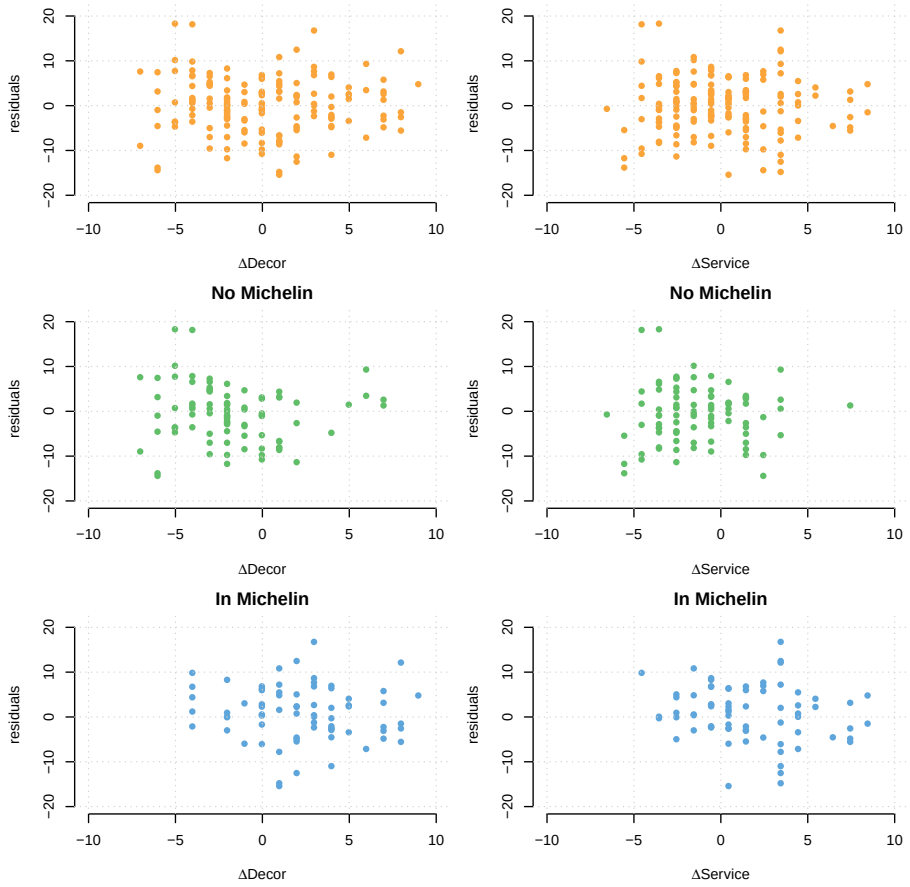


Fig. 10: Residui per il modello in Eq. (5) e Tab. 3 in maniera analoga a Fig. 6.

2 Parte B

2.1 Analisi preliminare dei dati

Il dataset Efige riporta varie informazioni relative ad aziende operanti in diversi settori economici, organizzate in 1115 record. La versione fornita in `efige.txt` presenta un numero di campi superiore a quelli indicati nella legenda di `efige.xls`: le variabili non presenti nella legenda sono state pertanto rimosse dopo il caricamento all'interno di R, rimanendo con 51 variabili così suddivise:[9]

- 2 variabili categoriche, sector (Tab. 4) e region (Tab. 5);[10]
- 31 variabili binarie (Tab. 6);
- 18 variabili quantitative (Tab. 7).

[9] Oltre alle variabili in eccesso, sono state estromessi anche i campi rel_weight e tfp_va.
[10] Le descrizioni dei settori e le codifiche delle regioni sono state recuperate dal foglio dati del file .xls.

Tab. 4: Settori delle aziende (codifica e descrizione).

codice	descrizione
10	prodotti alimentari
11	bevande
13	tessili
14	abbigliamento
15	pelletteria e prodotti affini
16	prodotti in legno/sughero/paglia (no mobili)
17	carta e prodotti di carta
18	stampa e riproduzione di supporti registrati
19	coke e prodotti petroliferi raffinati
20	prodotti chimici
21	prodotti farmaceutici di base e preparati farmaceutici
22	prodotti in gomma e plastica
23	altri prodotti minerali non metallici
24	metalli di base
25	prodotti in metallo fabbricati (no macchinari/attrezzature)
26	prodotti informatici, elettronici e ottici
27	materiale elettrico
28	macchinari e attrezzature n.c.a.
29	autoveicoli, rimorchi e semirimorchi
30	altri mezzi di trasporto
31	mobilia
32	altra manifattura
33	riparazione e installazione di macchinari e attrezzature

Dataset df_0 Indichiamo tale dataset con la sigla df_0; per iterazione verranno definiti in seguito nuovi dataset sempre più circoscritti (si faccia riferimento a Tab. 13 per un quadro complessivo dei dataset impiegati). La variabile da analizzare è la risposta binaria deloc_fdi (ossia "nel 2008 l'impresa ha effettuato investimenti diretti esteri"): tale variabile presenta solo 38 casi positivi sui 1115 record del dataset df_0.[11]

2.1.1 Analisi delle variabili categoriche

Con una risposta così sbilanciata può convenire innanzi tutto analizzare come deloc_fdi si ripartisca rispetto alle due variabili categoriche, individuando i settori e le regioni in cui si manifestano i pochi casi positivi, in modo da apportare una prima scrematura ai dati d'interesse. Tab. 8 mostra che solo un sottoinsieme di region (9 su 20) e sector (15 su 23) comporta deloc_fdi = 1; Fig. 11 ordina per numero di occorrenze le regioni interessate.

Gestione delle Regioni Fig. 11 evidenzia che le aziende con investimenti all'estero sono pressoché solo in regioni del centro-nord (con in testa Lombardia e Veneto), e solo una al sud (Abruzzo): nell'ottica di ridurre le variabili da

[11] Il riferimento temporale al 2008 non pare significativo, per cui verrà inteso che deloc_fdi indichi in generale l'iniziativa di investire all'estero.

Tab. 5: Codifica delle regioni.

cod.	regione	cod.	regione	cod.	regione	cod.	regione
1	Piemonte	6	Friuli	11	Marche	16	Puglia
2	Valle d'Aosta	7	Liguria	12	Lazio	17	Basilicata
3	Lombardia	8	Emilia	13	Abruzzo	18	Calabria
4	Trentino	9	Toscana	14	Molise	19	Sicilia
5	Veneto	10	Umbria	15	Campania	20	Sardegna

analizzare possiamo evitare un dettaglio così spinto nella localizzazione delle aziende limitandoci a raggruppare le regioni a seconda che siano del nord, centro o sud. Di fatto possiamo rimuovere region dal dataset in quanto l'area geografica di appartenenza è già codificata nelle variabili binarie north, centre e south_isl; inoltre, essendo tali variabili mutuamente esclusive, e considerando che le aziende del centro-nord possono essere raggruppate in una sola classe per quanto riguarda gli investimenti all'estero, possiamo fare affidamento sulla sola variabile south_isl:

- quando è a 1 (true), quasi sicuramente deloc_fdi = 0;
- viceversa, è possibile che deloc_fdi = 1.

Gestione dei settori Dalle descrizioni in Tab. 4 non si notano dei possibili raggruppamenti per tipologia dei settori per i quali deloc_fdi = 1; possiamo in ogni caso almeno raggruppare in un solo macro-settore "0" tutti quelli tali che deloc_fdi = 0, riducendo in ogni caso la gamma dei settori da analizzare.

Dataset df_1 La rimozione di region e le modifiche alle categorie di sector di df_0, oltre all'eliminazione delle variabili binarie north e centre portano alla definizione del nuovo dataset df_1 (Tab. 13).

2.1.2 Analisi delle variabili binarie

Avendo come variabile dipendente la risposta binaria deloc_fdi, nell'ottica di scremare il più possibile le variabili da gestire una prima analisi che possiamo compiere è attraverso dei test d'indipendenza di deloc_fdi rispetto alle altre variabili del dataset, osservando le corrispondenti tabelle di contingenza.[12]

La variabile deloc_fdi_china_india viene rimossa a priori nel processo di stima per eccessive analogie con la risposta d'interesse deloc_fdi: infatti

- deloc_fdi_china_india = 1 $\Rightarrow$ deloc_fdi = 1;
- deloc_fdi = 0 $\Rightarrow$ deloc_fdi_china_india = 0.

essendo la matrice di contingenza per la coppia di variabili Y = deloc_fdi e X = deloc_fdi_china_india

[12] Un simile approccio è piuttosto approssimativo in quanto, anche se il fatto che più v.a. siano indipendenti complessivamente implica che tali risultino anche i sottogruppi di variabili prese due a due, tre a tre e così via, non vale la proprietà inversa (cioè se più v.a. sono indipendenti a due a due, non necessariamente risultano del tutto indipendenti nel loro complesso). In questo contesto una tale strategia "golosa" ha solo lo scopo di semplificare il problema da trattare; metodi più rigosori, ad esempio basati su modelli grafici, sarebbero più appropriati.

Tab. 6: Variabili binarie del dataset; in rosso, la risposta da modellare.

campo	descrizione
north	sede legale nel Nord Italia
centre	sede legale nel Centro Italia
south_isl	sede legale nel Sud Italia
group	l'azienda appartiene ad un gruppo
individual_first_shr	il principale azionista è un individuo
foreign_first_shr	il principale azionista è straniero
decentr_manag	manager con autonomia decisionale
labour_flex	uso contratti part-time/tempo determinato
female_ceo	il CEO è una donna
fam_ceo	il CEO appartiene alla famiglia proprietaria
prod_inn	innovazioni di prodotto (ultimi 3 anni)
proc_inn	innovazioni di processo (ultimi 3 anni)
patent	brevetti (ultimi 3 anni)
RD_inv	investimenti in R&S (ultimi 3 anni)
direct_export	esportaz. all'estero parte del fatturato (2008)
dir_export_eu	esportaz. in EU parte del fatturato (2008)
dir_export_outside_eu	esportaz. fuori EU parte del fatturato (2008)
import_goods	importaz. beni dall'estero (2008)
import_goods_china_india	importaz. beni dalla Cina o India (2008)
deloc_fdi	investimenti diretti esteri (2008)
deloc_fdi_china_india	investimenti in Cina o India (2008)
qual_cert	certificazioni di qualità
competitors_from_abroad	concorrenti esteri
widened_prod_range	esteso la gamma di prodotti offerti (2008)
increase_margins	incrementato il margine sui costi (2008)
external_financing	richiesta di un nuovo credito
local_bank	relazioni con almeno una banca locale
national_bank	relazioni con almeno una banca nazionale
foreign_bank	relazioni con almeno una banca straniera
credit_requested	richiesta di un nuovo credito
credit_denied	la richiesta di nuovo credito è stata rifiutata

	$X = 0$	$X = 1$
$Y = 0$	1077	0
$Y = 1$	26	12

Scrematura delle variabili binarie Ai fini di valutare l'indipendenza tra deloc_fdi e le altre variabili binarie, è stato inpiegato in prima battuta il test del chi-quadro; tuttavia la funzione di R `chisq.test()` tra deloc_fdi ed alcune variabili (deloc_fdi_china_india, import_goods_china_india, foreign_bank, national_bank, female_ceo, credit_denied, foreign_first_shr, increase_margins) produce il seguente avviso

```
Warning message:
In chisq.test(df1$deloc_fdi, x_k) :
```

Tab. 7: Variabili continue del dataset.

campo	descrizione	#NA
age	Età dell'azienda	2
employees	Numero di dipendenti	
age_ceo	Età del CEO	
RD_empl_share	Quota di dipendenti impiegati in R&S	
grad_empl_share	Quota di dipendenti laureati	
dir_export_share	Vendite all'estero (quota fatturato)	16
import_share_goods	Beni importati dall'estero (quota fatturato)	
deloc_fdi_share	Investimenti diretti esteri (quota fatturato)	
banks_number	Numero di banche con cui l'impresa ha rapport	1
totalassets	Totale attivo (Capitale investito)	
sales	Vendite	
addedvalue	Valore aggiunto	23
ebit	Reddito operativo	5
roa	Risultato Netto / Totale Attivo	
roi	Reddito Operativo / Totale Attivo	5
ros	Reddito Operativo / Vendite	8
roe	Risultato Netto / Capitale Proprio	
leverage	Totale Attivo / Capitale Proprio	

Tab. 8: Ripartizione delle 38 occorrenze per deloc_fdi = 1 secondo sector e region all'interno del dataset df_0.

	region									
sector	1	3	5	6	8	9	11	12	13	Tot.
10	·	1	·	·	·	·	·	·	·	1
13	·	2	2	·	·	·	·	·	·	4
14	·	2	·	·	·	2	·	·	1	5
18	·	1	·	·	·	·	·	·	·	1
20	·	2	·	·	·	·	·	·	·	2
21	·	·	·	·	1	·	·	·	·	1
22	1	·	1	·	·	·	·	·	·	2
23	·	·	1	·	·	·	1	·	·	2
25	1	1	1	·	·	·	·	·	·	3
26	·	1	·	·	·	·	·	1	·	2
27	·	·	2	·	1	·	1	1	·	5
28	1	·	·	2	2	·	·	·	·	5
30	·	2	·	·	·	·	·	·	·	2
31	·	·	1	·	·	·	·	·	·	1
32	·	1	·	1	·	·	·	·	·	2
Tot.	3	13	8	3	4	2	2	2	1	38

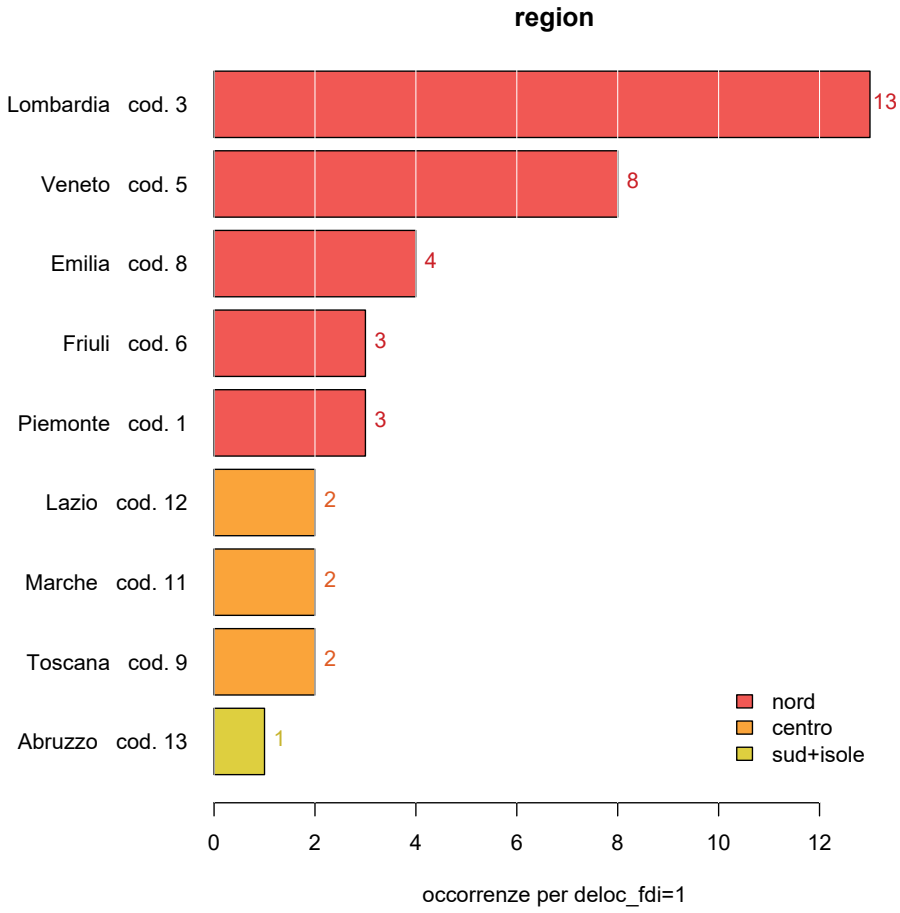


Fig. 11: Regioni e settori per i quali deloc_fdi=1.

```
Chi-squared approximation may be incorrect
```

per cui tale test è stato affiancato da un altro criterio basato sugli odds ratio come dettagliato di seguito. In entrambi i test le due variabili binarie vengono ritenute dipendenti quando il p-value è sufficientemente piccolo (inferiore a 0.05).

Indipendenza tramite odds ratio Date due variabili binarie X e Y, consideriamo i seguenti odds

$$o_{Y|X=1} = \frac{\mathbb{P}[Y=1|X=1]}{\mathbb{P}[Y=0|X=1]}$$
$$o_{Y|X=0} = \frac{\mathbb{P}[Y=1|X=0]}{\mathbb{P}[Y=0|X=0]}$$

Nell'ipotesi che X e Y siano variabili indipendenti abbiamo che, per $x, y \in \{0,1\}$, $\mathbb{P}[Y=y|X=x] = \mathbb{P}[Y=y]$, e quindi per lo odds ratio vale

$$\Theta_{Y|X} = \frac{o_{Y|X=1}}{o_{Y|X=0}} = \frac{\mathbb{P}[Y=1]}{\mathbb{P}[Y=0]} \Big/ \frac{\mathbb{P}[Y=1]}{\mathbb{P}[Y=0]} = 1 \tag{6}$$

Considerando pertando N osservazioni di due v.a. binarie X e Y, data la loro matrice di contingenza (con $N = n_{00} + n_{01} + n_{10} + n_{11}$)

	$X = 0$	$X = 1$
$Y = 0$	n_{00}	n_{01}
$Y = 1$	n_{10}	n_{11}

è possibile ricavare una stima per Eq. (6) come

$$\hat{\Theta}_{Y|X} = \overbrace{\left(\frac{n_{11}}{n_{01}}\right)}^{\hat{o}_{Y|X=1}} \Big/ \overbrace{\left(\frac{n_{10}}{n_{00}}\right)}^{\hat{o}_{Y|X=0}} = \frac{n_{11}\, n_{00}}{n_{01}\, n_{10}} \tag{7}$$

Da un punto di vista implementativo, è stata impiegata la funzione di R `fisher.test()`, che restituisce al posto di Eq. (7) una stima a massima verosimiglianza *condizionale* per $\Theta_{Y|X}$, con relativo p-value sulla verifica che $\hat{\Theta}_{Y|X}$ sia significativamente diverso da 1.

Tab. 9 riporta i p-value ottenuti con `chisquare.test()` e `fisher.test()`, segnalando in rosso le variabili indipendenti rispetto a deloc_fdi (almeno un test fornisce un p-value < 0.05 valido) e pertanto scartate dal processo di stima: i due criteri forniscono risultati concordi in ogni caso.

Complessivamente questa prima ripartizione appare plausibile anche da un punto di vista "semantico" delle variabili: l'unica perplessità può semmai riguardare foreign_first_shr in quanto potrebbe essere plausibile che un azionista di maggioranza straniero sia più incline ad investire all'estero rispetto ad azionisti nazionali. Tuttavia dalla matrice di contingenza di X = foreign_first_shr rispetto a Y = deloc_fdi questo fenomeno non appare

	$X = 0$	$X = 1$
$Y = 0$	1029	48
$Y = 1$	36	2

Per quanto riguarda invece la variabile national_bank la stima dell'odds ratio vale ∞ in quanto $n_{10} = 0$, come si vede dalla relativa matrice di contingenza (con X = national_bank, Y = deloc_fdi)

	$X = 0$	$X = 1$
$Y = 0$	137	940
$Y = 1$	0	38

Notiamo altresì che il 100% delle aziende che investono all'estero hanno rapporti con banche nazionali, e che tra le imprese che non investono all'estero solo una piccola parte (meno del 13%) non ha nemmeno relazioni con banche nazionali: tutto sommato non pare una variabile particolarmente informativa per i nostri scopi (quasi tutte le aziende hanno rapporti con banche nazionali), per cui national_bank viene scartata dal pool di variabili impiegabili nel modello.

Occorre adesso valutare quali tra le restanti variabili binarie portino al modello informazioni paragonabili. Alcune possono essere raggruppate semanticamente:

- direct_export, dir_export_outside_eu e dir_export_eu sono chiaramente collegate (in particolare le ultime due implicano la prima);

Tab. 9: Test d'indipendenza delle variabili binarie X_k rispetto alla risposta d'interesse Y = deloc_fdi basati su test chi-quadro e sugli odds ratio $\Theta_{Y|X_k}$. In rosso le variabili ritenute indipendenti con deloc_fdi; in *italico* le variabili con test del chi-quadro non attendibile.

variabile X_k	p-value(χ^2)	$\widehat{\Theta}_{Y\|X_k}$	p-value($\Theta_{Y\|X_k}$)
national_bank	0.036	∞	0.010
direct_export	$1.285 \cdot 10^{-04}$	18.71	$9.092 \cdot 10^{-06}$
competitors_from_abroad	$1.315 \cdot 10^{-06}$	9.12	$1.834 \cdot 10^{-07}$
import_goods_china_india	$2.875 \cdot 10^{-12}$	8.80	$3.670 \cdot 10^{-08}$
labour_flex	0.018	8.74	0.005
foreign_bank	$5.815 \cdot 10^{-12}$	8.35	$3.169 \cdot 10^{-08}$
RD_inv	$1.157 \cdot 10^{-05}$	7.84	$2.811 \cdot 10^{-06}$
group	$1.450 \cdot 10^{-07}$	5.24	$1.607 \cdot 10^{-06}$
import_goods	$1.155 \cdot 10^{-05}$	4.57	$1.661 \cdot 10^{-05}$
patent	$5.021 \cdot 10^{-06}$	4.49	$3.292 \cdot 10^{-05}$
prod_inn	$2.668 \cdot 10^{-04}$	4.40	$1.056 \cdot 10^{-04}$
qual_cert	0.002	3.65	0.001
dir_export_outside_eu	0.001	3.42	0.001
dir_export_eu	0.007	3.38	0.003
widened_prod_range	0.025	2.34	0.020
fam_ceo	0.426	1.48	0.365
external_financing	0.451	1.38	0.401
proc_inn	0.579	1.27	0.511
decentr_manag	0.845	1.20	0.653
credit_requested	0.827	1.19	0.670
foreign_first_shr	1	1.19	0.686
increase_margins	1	0.81	1
local_bank	0.469	0.74	0.382
female_ceo	0.793	0.72	0.790
credit_denied	0.805	0.67	0.763
individual_first_shr	0.019	0.43	0.022
south_isl	0.085	0.17	0.050

- import_goods_china_india implica import_goods;
- widened_prod_range è legata a prod_inn.

Nell'ottica di avere un modello il meno specifico possibile, vengono mantenute nel pool di variabili solo quelle più generali (direct_export, import_goods e prod_inn).

Applichiamo adesso nuovamente il test d'indipendenza con gli odds ratio su ogni possibile coppia tra le variabili ancora a disposizione, per avere un quadro della situazione: da Tab. 10 si evince che le variabili binarie sono perlopiù associabili le une alle altre, ma in questa fase dell'analisi è difficile capire quali eliminare ulteriormente (ad esempio è plausibile che prod_inn e RD_inv siano legate, ma quale agisce di più su deloc_fdi?).

Dataset df_2 La scrematura delle variabili binarie porta al dataset df_2 (Tab. 13), avendo rimosso deloc_fdi_china_india, fam_ceo, external_financing, proc_inn,

Tab. 10: Nella parte superiore della matrice: odds ratio calcolati con `fisher.test()` per ciascuna coppia di variabili binarie. Nella parte inferiore: sulla base dei p-value forniti da `fisher.test()`, viene indicato quali coppie di variabili siano indipendenti (•, per p-value $\geqslant 0.5$) oppure se esista un'associazione statisticamente significativa (○).

	01	02	03	04	05	06	07	08	09	10	11
01	–	0.03	1.94	1.72	1.73	2.22	1.66	2.13	2.53	1.81	3.57
02	○	–	0.64	0.65	0.60	0.42	0.52	0.49	0.48	0.51	0.31
03	○	○	–	1.46	1.58	1.54	1.79	1.88	1.41	1.45	1.79
04	○	○	○	–	8.46	7.31	3.64	2.59	1.53	2.29	1.47
05	○	○	•	○	–	7.25	6.31	2.55	1.79	2.24	2.01
06	○	○	○	○	○	–	2.94	2.51	2.29	2.35	1.17
07	○	○	○	○	○	○	–	3.56	1.42	4.66	4.92
08	○	○	○	○	○	○	○	–	1.68	2.43	2.67
09	○	○	○	○	○	○	○	○	–	1.46	2.25
10	○	○	○	○	○	○	○	○	○	–	1.98
11	○	○	•	•	○	•	○	○	○	○	–

01: group
02: individual_first_shr
03: labour_flex
04: prod_inn
05: patent
06: RD_inv
07: direct_export
08: import_goods
09: qual_cert
10: competitors_from_abroad
11: foreign_bank

• variabili indipendenti
○ associaz. significativa

decentr_manag, credit_requested, foreign_first_shr, increse_margin, local_bank, female_ceo, credit_denied, south_isl, national_bank, import_goods_china_india, dir_export_eu, dir_export_outside_eu, widened_prod_range.

2.1.3 Analisi delle variabili quantitative

Analisi di banks_number Tra le variabili quantitative, banks_number è l'unica che deve essere trattata come discreta. Fig. 12 illustra come il numero di banche sia ripartito nel dataset, sia nel caso delle sole aziende con investimenti all'estero sia per le restanti imprese. Non notando particolari differenze tra le due distribuzioni (tenendo anche conto delle poche aziende tali che deloc_fdi = 1), un'informazione così dettagliata viene ritenuta inutile; pertanto banks_number viene rimossa dalle variabili a disposizione per stimare il modello.

Analisi delle variabili continue Dall'insieme di variabili continue in Tab. 7 è stata rimossa a priori age_ceo in quanto i suoi valori sembrano indicare gli anni in carica del CEO (range di interi da 2 a 7), per cui appare poco legata alla risposta binaria d'interesse (deloc_fdi).

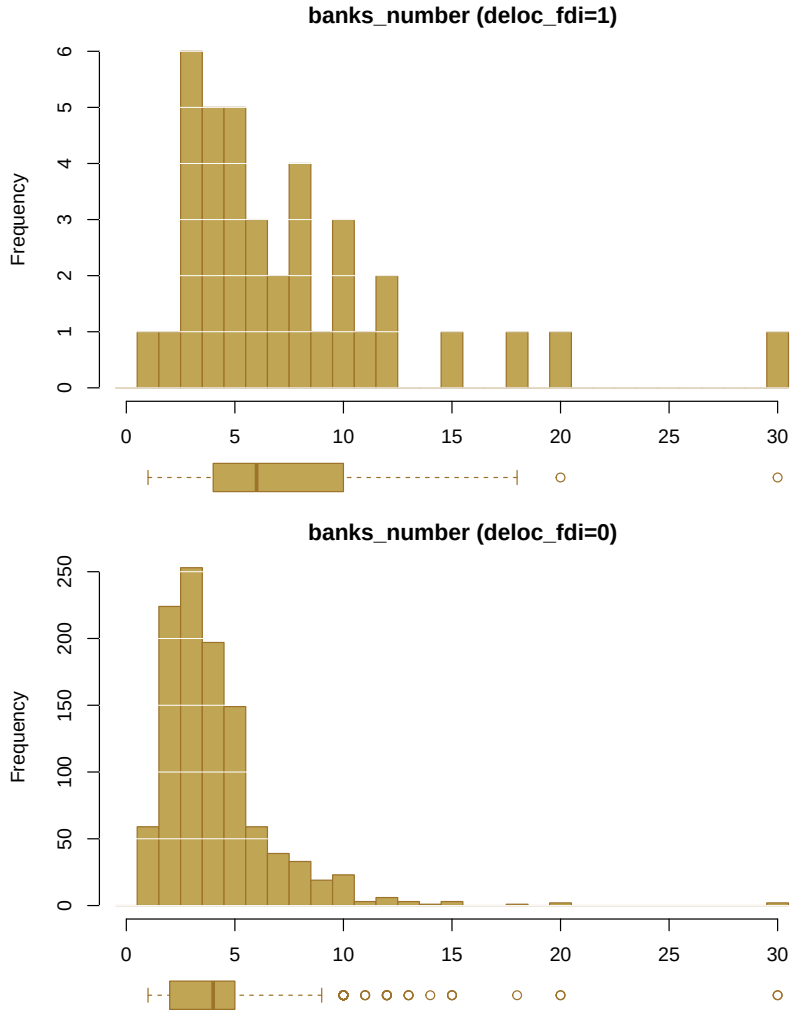


Fig. 12: Ripartizione del numero di banche: solo per le imprese con investimenti esteri (in alto), e su tutte le restanti aziende (in basso).

Inoltre, anche deloc_fdi_share viene rimossa a priori in quanto ridondante: di fatto essa è diversa da 0 solo quando deloc_fdi vale 1; la percentuale di fatturato investita all'estero che tale variabile rappresenta è superflua ai fini della classificazione.

Outliers Prima di calcolare i coefficienti di correlazione, da un primo sguardo alle variabili continue notiamo che molte presentano degli outlier: allo scopo di evitare casi fuori scala, vengono rimossi dal dataset i record per con valori nelle code escludendo quelli inferiori al quantile $q_{0.005}$ e quelli superiori a $q_{0.995}$ per ciascuna variabile continua. Da una successiva ispezione visiva, è stato poi rimosso manualmente qualche dato in più:

- sono stati eliminati 10 record per i quali sale > 150000;
- sono stati rimossi 2 elementi per ebit > 10000;

- infine 1 record è stato cancellato per totalassets > 250000.

Fig. 13, 14 mostrano la ripartizione dei valori rimasti per le variabili più critiche.

Indici di correlazione A questo punto dai coefficienti di correlazione di Tab. 11 notiamo che alcune variabili sono piuttosto correlate tra di loro; in particolare sono in gran parte "raggruppabili"

- employees, totalassets, sales, addedvalue, ebit;
- roa, roi, ros e roe;
- RD_empl_share e grad_empl_share.[13]

Inoltre ebit esibisce una certa correlazione anche con roa, roi, ros ($\rho \approx 0.35$). Prima di decidere quale variabile mantenere nei vari gruppi appena emersi, conviene anche visualizzarne gli scatterplot:

- per il primo gruppo viene tenuto addedvalue come elemento rappresentativo, poiché mostra la maggior correlazione con le altre variabili del gruppo;
- similmente, per il secondo manteniamo invece roi.

Infine Fig. 16a mostra la matrice di scatter plot per employees, RD_empl_share e grad_empl_share, mentre in Fig. 16b, 16c sia RD_empl_share che grad_empl_share sono riportati rispetto a deloc_fdi_share (qui usata come alias continuo per deloc_fdi).[14] Da tali figure si nota che le aziende con investimenti all'estero hanno relativamente pochi dipendenti (molte sotto i 100) e non laureati. Tenendo anche conto che solitamente gli assunti nel reparto R&D sono perlopiù laureati, possiamo escludere RD_empl_share dal pool di variabili per il modello.

Per completezza, inoltre Fig. 17 riporta la matrice di scatter plot per age, dir_export_share, import_share_goods e leverage.

Dataset df_3 L'eliminazione di banks_number, age_ceo, deloc_fdi_share e delle variabili quantitative elencate in Tab. 12 definisce il dataset df_3 come riassunto in Tab. 13; la rimozione dei record con outlier ha portato il numero di esempi dai 1115 iniziali a 1010.

2.1.4 Similarità tra variabili continue e binarie

Osservando la lista delle variabili continue rimaste in Tab. 12, notiamo che alcune di loro sembrano strettamente legate alle variabili discrete ancora a disposizione (Tab. 10). Di tali coppie, nell'idea che le variabili continue veicolino un'informazione più ricca, vengono quindi rimosse le variabili discrete:

- tra import_share_goods (continua) e import_goods (binaria) viene eliminata import_goods;
- tra dir_export_share (continua) e direct_export (binaria) viene rimossa direct_export;

[13] A loro volta con una certa correlazione con employees ($\rho = 0.29$ e $\rho = 0.24$ rispettivamente).
[14] Ai fini di mostrare meglio dove si concentrano i valori per cui deloc_fdi=1, in Fig. 16a i corrispondenti valori delle variabili visualizzate sono stati leggermente sporcati con del rumore tramite la funzione `jitter()`.

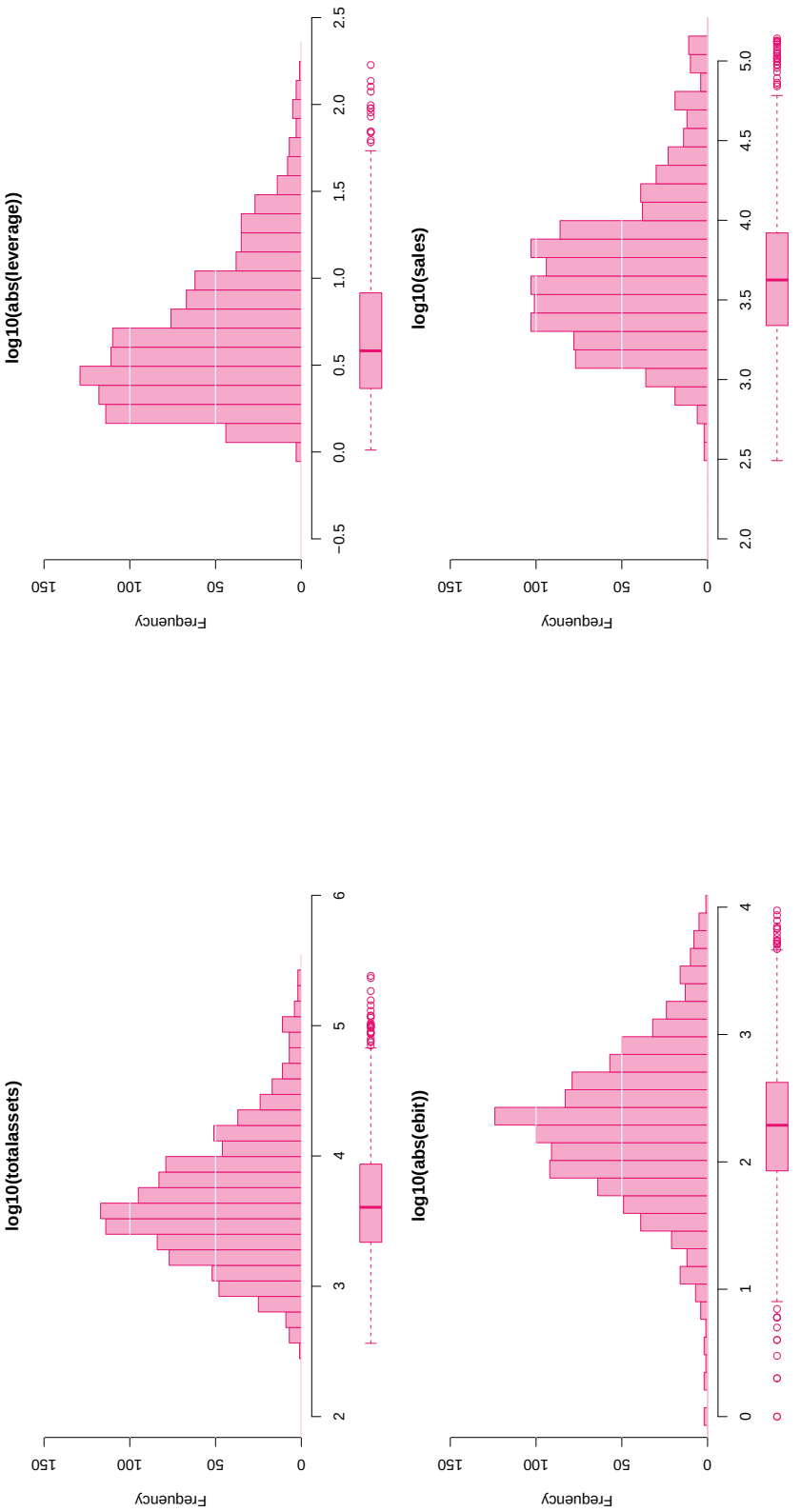


Fig. 13: Occorrenze e boxplot in scala logaritmica per totalassets (a), leverage (b), ebit (c) e sales (d). Per leverage e ebit si è dovuto applicare anche il valore assoluto a causa di elementi tra -848.75 e -0.46 (leverage), e tra -12984 e -1 (ebit).

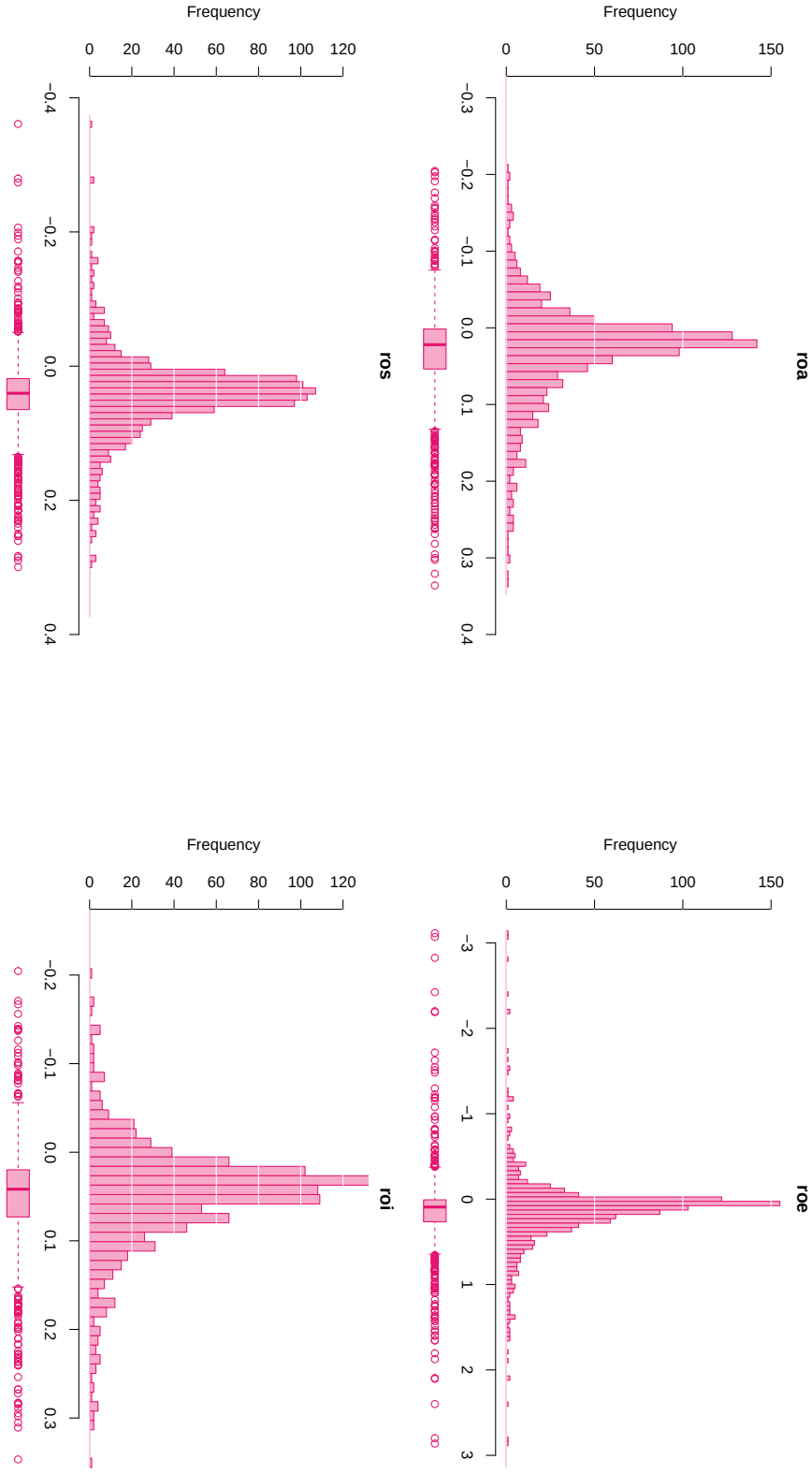


Fig. 14: Occorrenze e boxplot per roa (a), roe (b), ros (c) e roi (d).

Tab. 11: Nella parte superiore della matrice: indici di correlazione per ciascuna coppia di variabili continue. Nella parte inferiore: viene indicato quali coppie di variabili siano altamente correlate (•) e quali no (○). Le due variabili evidenziate in vari colori sono molto correlate le une con le altre.

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
age 01	–	0.17	−0.01	0.02	0.11	−0.02	0.16	0.16	0.17	0.10	−0.03	−0.08	−0.05	−0.13	−0.16
employees 02	○	–	0.14	0.29	0.24	0.10	0.72	0.80	0.86	0.45	−0.03	−0.05	−0.03	−0.03	−0.09
RD_empl_share 03	○	○	–	0.30	0.07	0.01	0.05	0.05	0.07	0.02	−0.02	−0.03	−0.01	−0.05	0.00
grad_empl_share 04	○	○	•	–	0.09	0.06	0.23	0.24	0.29	0.20	0.02	0.01	0.02	0.03	−0.03
dir_export_share 05	○	○	○	○	–	0.07	0.21	0.23	0.24	0.16	−0.03	−0.04	−0.03	−0.02	−0.06
import_share_goods 06	○	○	○	○	○	–	0.11	0.12	0.10	0.10	−0.05	−0.04	−0.02	−0.04	−0.05
totalassets 07	○	•	○	○	○	○	–	0.83	0.86	0.55	−0.03	−0.06	0.00	−0.06	−0.09
sales 08	○	•	○	○	○	○	•	–	0.90	0.66	0.02	0.00	−0.01	−0.02	−0.09
addedvalue 19	○	•	○	○	○	○	•	•	–	0.64	0.07	0.04	0.06	−0.02	−0.10
ebit 10	○	•	○	○	○	○	•	•	•	–	0.36	0.35	0.38	0.13	−0.10
roa 11	○	○	○	○	○	○	○	○	○	•	–	0.96	0.83	0.49	−0.13
roi 12	○	○	○	○	○	○	○	○	○	•	•	–	0.85	0.51	−0.05
ros 13	○	○	○	○	○	○	○	○	○	•	•	•	–	0.38	−0.04
roe 14	○	○	○	○	○	○	○	○	○	○	•	•	•	–	0.21
leverage 15	○	○	○	○	○	○	○	○	○	○	○	○	○	○	–

• alta/media correlazione ($|\rho| \geqslant 0.3$)
○ bassa/leggera correlazione

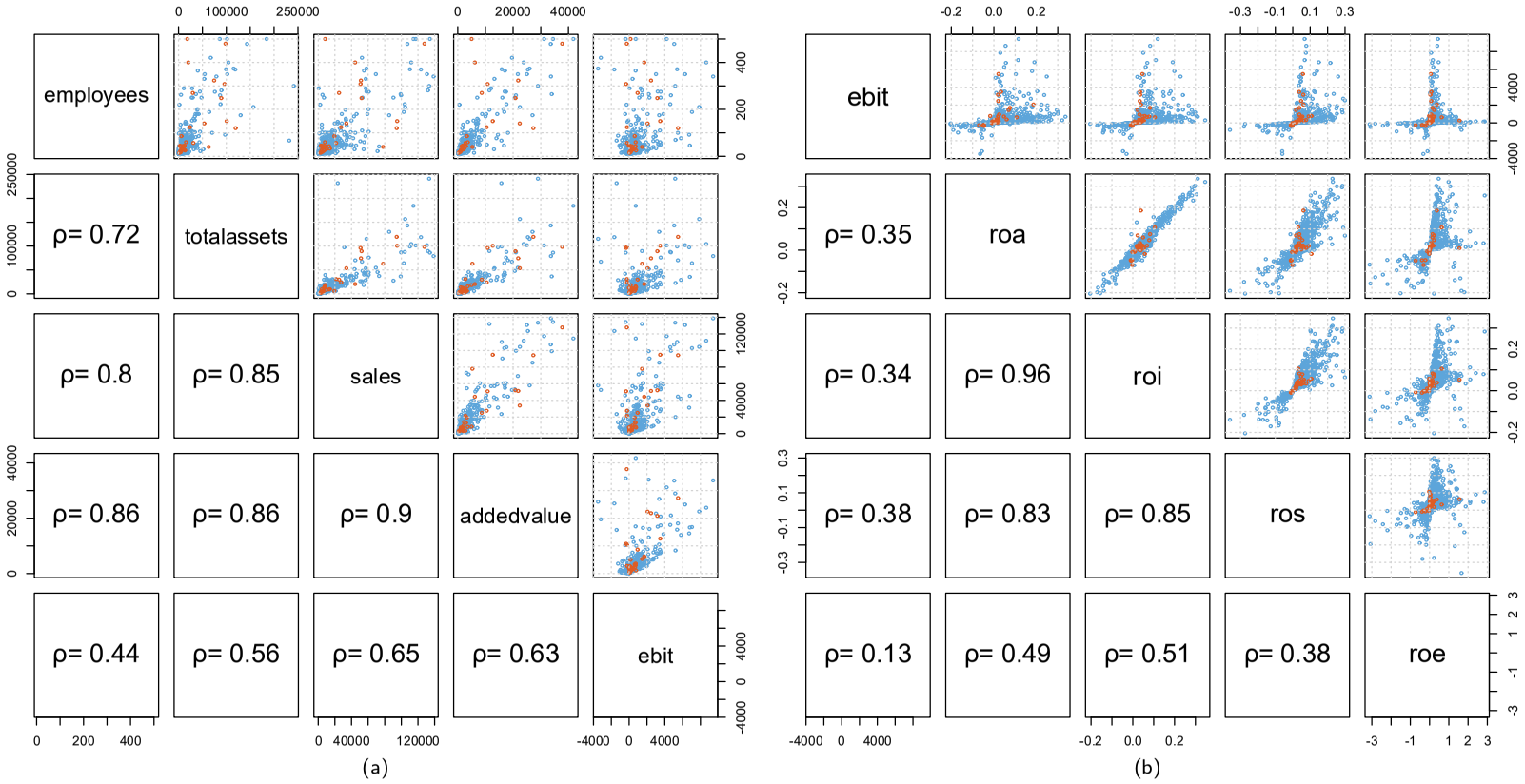


Fig. 15: Matrici di scatter plot per totalassets, sales, addedvalue, ebit (a) e per ebit, textsfroa, roi, textsfros, roe. I valori per cui deloc_fdi vale 1 sono evidenziati con • rispetto a •

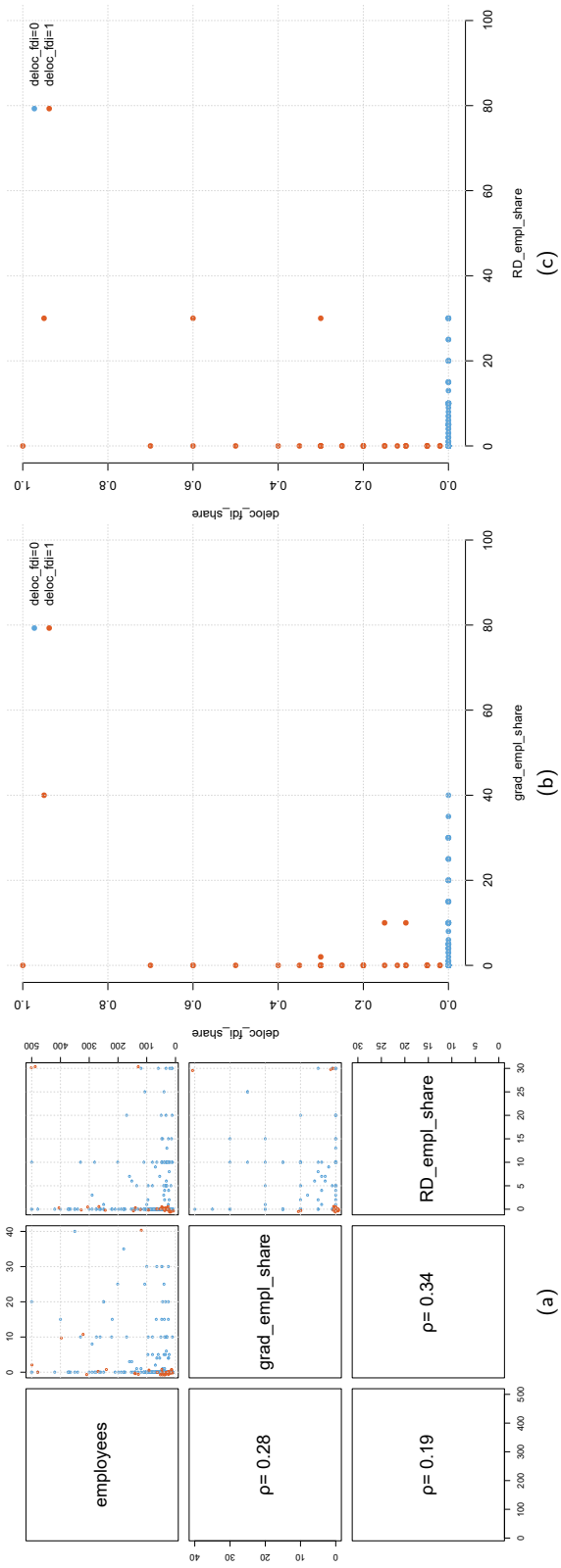


Fig. 16: Ripartizione di employees, RD_empl_share e grad_empl_share in (a), e di deloc_fdi_share rispetto a RD_empl_share (b) e grad_empl_share (c), a seconda che ai dati visualizzati corrisponda deloc_fdi=0 (●) oppure deloc_fdi=1 (●).

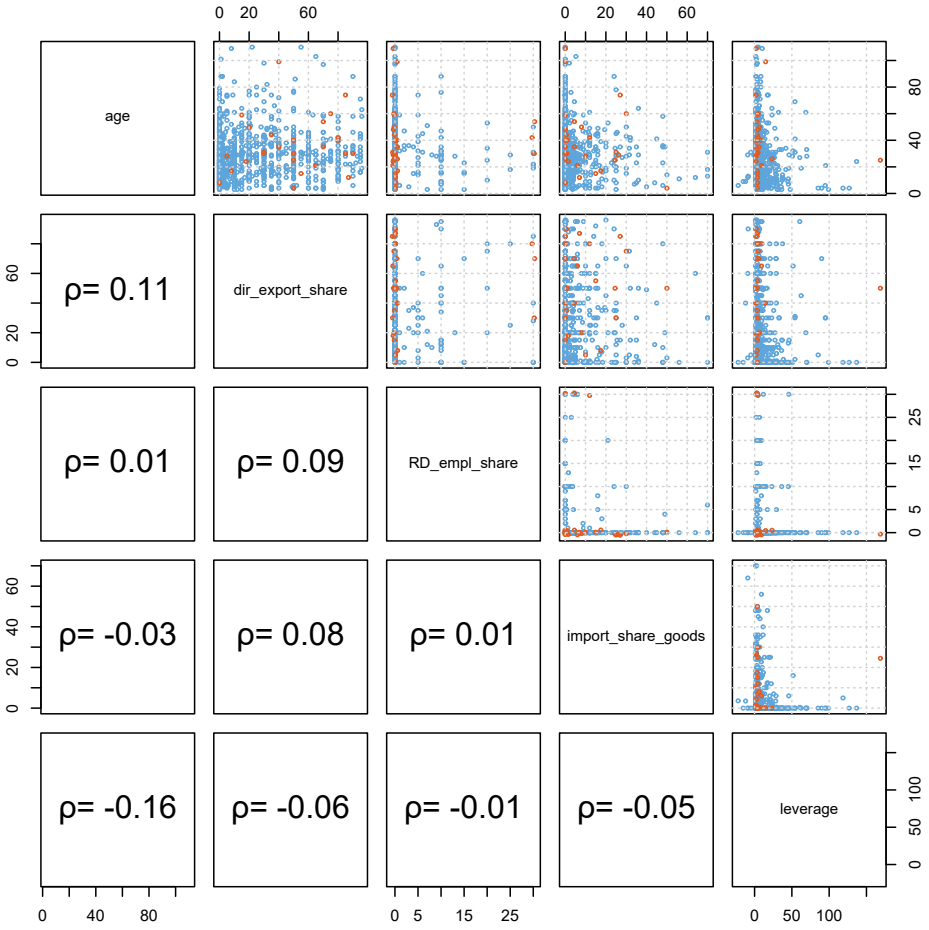


Fig. 17: Matrice di scatterplot per le restanti variabili continue (• deloc_fdi=0; • deloc_fdi=1).

- inoltre, ricordando che grad_empl_share era stata preferita a RD_empl_share perché in genere i dipendenti in Ricerca e Sviluppo sono in genere laureati, viene rimossa anche RD_inv.

Dataset df_4 Queste ultime eliminazioni, seguite dalla rimozione di tutti i record con dati mancanti, portano a definire il dataset df_4 (Tab. 13), composto da 18 variabili (deloc_fdi compresa) e 977 record rispetto alle 51 variabili iniziali con 1115 elementi ciascuna; i record per cui deloc_fdi=1 sono scesi da 38 a 27. Tab. 14 riepiloga le variabili selezionate e quelle eliminate al termine dell'intera fase di scrematura.

2.2 Determinazione del modello

Per individuare un modello logit capace di spiegare deloc_fdi rispetto ad un sottoinsieme delle variabili selezionate di Tab. 14 si è ricorso alla tecnica di backward selection tramite la funzione `stepAIC()` a partire da un modello full

Tab. 12: Scrematura delle variabili continue in base alla loro correlazione.

mantenute	rimosse
age	employees
grad_empl_share	totalassets
dir_export_share	sales
import_share_goods	ebit
addedvalue	roa
roi	ros
ros	roe
leverage	RD_empl_share

Tab. 13: Riduzione delle variabili d'interesse.

dataset	caratteristiche
df_0	dataset iniziale
df_1	esclusione di region, north e centre uso di south_isl per l'area geografica raggruppamento in sector = 0 dei settori per i quali deloc_fdi = 0
df_2	esclusione di deloc_fdi_china_india e national_bank rimoz. di variabili binarie indipendenti da deloc_fdi (Tab. 9) rimoz. di variabili binarie semanticamente simili
df_3	rimozione di banks_number esclusione a priori di age_ceo e deloc_fdi_share rimozione outliers rimoz. di variabili continue su base correlazione (Tab. 12)
df_4	eliminazione di import_goods, dir_export e RD_inv (v.a. discrete) esclusione dei record con dati mancanti

Tab. 14: Riepilogo delle variabili selezionate per df_4.

v.a. categoriche	v.a. binarie	v.a. continue
sector (con $sector_0$)	group	age
	individual_first_shr	grad_empl_share
	labour_flex	dir_export_share
	prod_inn	import_share_goods
	patent	banks_number
	RD_inv	addedvalue
	direct_export	roi
	import_goods	leverage
	deloc_fdi	
	qual_cert	
	competitors_from_abroad	
	foreign_bank	

Tab. 15: Primo modello logit (al termine di `stepAIC()`), con relativi errori standard e p-value.

k	variabili	$\hat{\beta}_k$	se$(\hat{\beta}_k)$	z-value	p-value
0	(intercetta)	-7.512	1.215	-6.182	$6.34 \cdot 10^{-10}$
1	labour_flex	1.245	1.044	1.193	0.233
2	patent	$8.5 \cdot 10^{-1}$	$4.7 \cdot 10^{-1}$	1.798	0.072
3	dir_export_share	$1.4 \cdot 10^{-2}$	$7.9 \cdot 10^{-3}$	1.737	0.082
4	competitors_from_abroad	2.022	$7.7 \cdot 10^{-1}$	2.621	0.009
5	foreign_bank	2.031	$4.4 \cdot 10^{-1}$	4.582	$4.60 \cdot 10^{-6}$
6	addedvalue	$3.7 \cdot 10^{-5}$	$2.4 \cdot 10^{-5}$	1.554	0.120

ricavato con `glm()` su tutte le variabili disponibili in Tab. 14: il modello così ottenuto presenta diverse variabili con p-value eccessivi (Tab. 15), per cui non accettabile.

Allo scopo di determinare un modello valido statisticamente, è stato quindi deciso di rimuovere dalle variabili quella con il p-value più elevato (ossia labour_flex) e di confrontare la stima sul modello annidato con quello in Tab. 15 tramite test della devianza:

Modello	Resid. Dev	ΔDev.	p-value
no labour_flex	182.64		
Tab. 15	180.61	2.0286	0.1544

Avendo avuto esito positivo (ossia il modello annidato non si discosta significativamente dal precedente, poiché il p-value per l'incremento della devianza supera l'usuale soglia di 0.05), la procedura viene iterata (rimozione della variabile peggiore, e test della devianza del modello annidato rispetto al precedente) fino ad ottenere dei p-value soddisfacenti per tutte le variabili del modello: Tab. 16 riassume i test di devianza condotti; il procedimento è terminato in due iterazioni, quando le variabili impiegate hanno mostrato tutte p-value inferiore a 0.05 (inoltre, il tentativo di iterare ulteriormente non è stato più supportato dal test della devianza).

Il modello finale è riportato in Tab. 17; per scrupolo viene testato anche rispetto al modello iniziale in Tab. 15, con conferma che il modello ridotto è paragonabile ad esso (ma preferibile secondo il rasoio di Occam):

Modello	Resid. Dev	ΔDev.	p-value
Tab. 17	185.28		
Tab. 15	180.61	4.6638	0.09711

2.2.1 Interpretazione del modello

L'impiego degli odds ratio, tramite

$$\Theta_i = \exp(\hat{\beta}_i)$$

ci consente d'interpretare le variabili nel modello di Tab. 17 come segue.

Tab. 16: Riduzione del modello a partire da quello in Tab. 15, eliminando ad ogni iterazione la variabile con p-value (fit) peggiore, con test della devianza rispetto al modello precedente.

it.	v.a. rimossa	p-value (fit)	ΔDev.	p-value (Dev.)
1	labour_flex	0.23281	2.0286	0.1544
2	addedvalue	0.08586	2.6352	0.1045

Tab. 17: Modello logit finale, con relativi errori standard e p-value.

k	variabili	$\hat{\beta}_k$	se($\hat{\beta}_k$)	z-value	p-value
0	(intercetta)	-6.488563	0.778314	-8.337	$< 2 \cdot 10^{-16}$
1	patent	0.966237	0.450825	2.143	0.03209
2	dir_export_share	0.015820	0.007637	2.071	0.03831
3	competitors_from_abroad	2.172436	0.764398	2.842	0.00448
4	foreign_bank	2.150410	0.434596	4.948	$7.5 \cdot 10^{-7}$

Intercetta ($\Theta_0 \approx 1.52 \cdot 10^{-3}$) La propensione che un'azienda effettui investimenti diretti esteri, quando non è presente nessuno degli altri fattori (ovvero nessun brevetto, nessuna vendita all'estero, nessun concorrente straniero, nessuna banca straniera), è molto bassa in quanto $\hat{\beta}_0$ è negativo ed altamente significativo (p-value $< 2 \cdot 10^{-16}$).

patent ($\Theta_1 \approx 2.63$) In questo caso, un'azienda con brevetti recenti è circa 2,63 volte più propensa ad effettuare investimenti esteri rispetto a un'azienda senza depositi di brevetto, mantenendo costanti tutte le altre variabili.

dir_export_share ($\Theta_2 \approx 1.02$) Essendo dir_export_share una variabile continua occorre ragionare in termini di variazioni unitarie, in questo caso relativi alla percentuale del fatturato dovuto alle esportazioni: ad un incremento unitario è associato un aumento molto leggero della propensione a realizzare investimenti esteri; nondimeno, tale risultato suggerisce che le aziende con una quota maggiore di esportazioni tendano di più ad investire all'estero.

competitors_from_abroad ($\Theta_3 \approx 8.78$) La presenza di concorrenti stranieri è associabile ad un aumento ampio (quasi 9 volte) della possibilità da parte di un'azienda di effettuare investimenti esteri, cioè le aziende che affrontano la concorrenza estera siano più inclini ad investire esse stesse all'estero (possibilmente per espandere la propria portata sul mercato o contrastare la concorrenza straniera).

foreign_bank ($\Theta_4 \approx 8.59$) Avere rapporti con almeno una banca estera è associato a un aumento pari a oltre 8 volte delle probabilità di effettuare investimenti all'estero rispetto a non farli; verosimilmente, l'accesso ai finanziamenti esteri potrebbe svolgere un ruolo nel facilitare gli investimenti (magari nello stesso paese della banca).

Conclusioni Nel complesso, il modello trovato appare abbastanza sensato:

- esso suggerisce che le aziende con brevetti, concorrenti stranieri o rapporti con banche estere siano tutte più propense ad effettuare investimenti all'estero rispetto a quelle che invece non hanno tali caratteristiche;
- l'effetto di avere un brevetto è il più debole tra i tre, mentre avere concorrenti stranieri o rapporti bancari esteri ha un'associazione positiva molto più forte con gli investimenti esteri;
- inoltre, esportare all'estero mostra un'associazione leggermente positiva con tali investimenti, forse per ridurre i costi verso i paesi d'esportazione.

Riflessioni sull'inclusione di patent L'unica perplessità riguarda semmai l'inclusione di patent tra le variabili impiegate (seppur la meno influente rispetto a competitors_from_abroad e foreign_bank): ad intuito i brevetti sono riconducibili ad aziende che si avvalgono di personale con alto profilo (laureati), mentre gli scatter plot in Fig. 16, condotti durante l'analisi delle variabili continue, mostravano una concentrazione delle aziende con investimenti esteri tra quelle con pochi laureati. Possibili spiegazioni di tale contraddizione (fatta valida la congettura sui laureati necessari per produrre brevetti) sono:

- la selezione delle variabili da mettere a disposizione del modello è da rivedere in qualche elemento (una conoscenza specifica delle dinamiche aziendali in merito agli investimenti esteri probabilmente guiderebbe in modo più pertinente il processo di scrematura del dataset);
- la determinazione del modello tramite backward selection basata sull'indice AIC non è particolarmente performante, e metodi più evoluti (LASSO e affini) potrebbero essere impiegati al suo posto;
- la risposta deloc_fdi è particolarmente difficile da modellare data l'esiguità degli esempi positivi nel dataset a disposizione (solo il 3% circa).

.