Modelli per dati ordinali

Scale ordinali

- Le scali ordinali sono molto diffuse nel marketing e in genere nelle scienze sociali allo scopo di misurare attitudini e opinioni
- Per esempio, la soddisfazione per un prodotto o un servizio può essere misurata con una scala Likert del tipo "molto in disaccordo, in disaccordo, indeciso, d'accordo, molto d'accordo"
- Nelle scale ordinali esiste un *chiaro ordinamento dei livelli*, ma le distanze assolute tra questi non sono note (non c'è una misura per esprimere la differenza tra i vari livelli)
- Una variabile ordinale è qualitativa
- Ma può essere anche vista come quantitativa nel senso che ogni livello esprime una dimensione maggiore o minore di una certa caratteristica rispetto a un altro livello

Modellare variabili ordinali

- Siamo interessati a introdurre modelli per descrivere variabili ordinali
- I tipici modelli per variabili quantitative possono rivelarsi non adeguati

Regressione logistica con logit cumulati

- La regressione logistica con logit cumulati è un'estensione della regressione logistica standard, nella quale la variabile risposta Y è misurata su una scala ordinale.
- È importante quindi definire i logit in una maniera tale da riconoscere l'ordinamento all'interno di Y.

Logit cumulati

Per c categorie con probabilità $\pi_1, \cdots, \pi_c$ i logit cumulati sono definiti come

$$\text{logit}[P(Y \leq j)] = \log \frac{P(Y \leq j)}{1 - P(Y \leq j)}$$

$$\log \frac{P(Y \leq j)}{1 - P(Y \leq j)} = \log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c}, \quad j = 1, \cdots, c-1$$

Il logit cumulato usa tutte le categorie c.
Da notare che se le categorie vengono ridotte a due risultati $Y \leq j$ e $Y > j$ si ha il logit ordinario.

Modelli con logit cumulati

Un modello con logit cumulati e variabili esplicative è del tipo

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

per $j = 1, \ldots, c-1$.
Il modello usa simultaneamente tutti i $c-1$ logit cumulati.

Da notare che l'intercetta dipende dall'indice j, α_j.
Le α_j crescono in j perchè $P(Y \leq j)$ cresce in j per *ogni valore fissato* di **x**, e il logit è una funzione crescente di questa probabilità.

Modelli con logit cumulati

Il corrispondente modello per le probabilità cumulate è

$$P(Y \leq j) = \frac{\exp(\alpha_j + \beta' \mathbf{x})}{1 + \exp(\alpha_j + \beta' \mathbf{x})}, \quad j = 1, \ldots, c-1$$

Inoltre

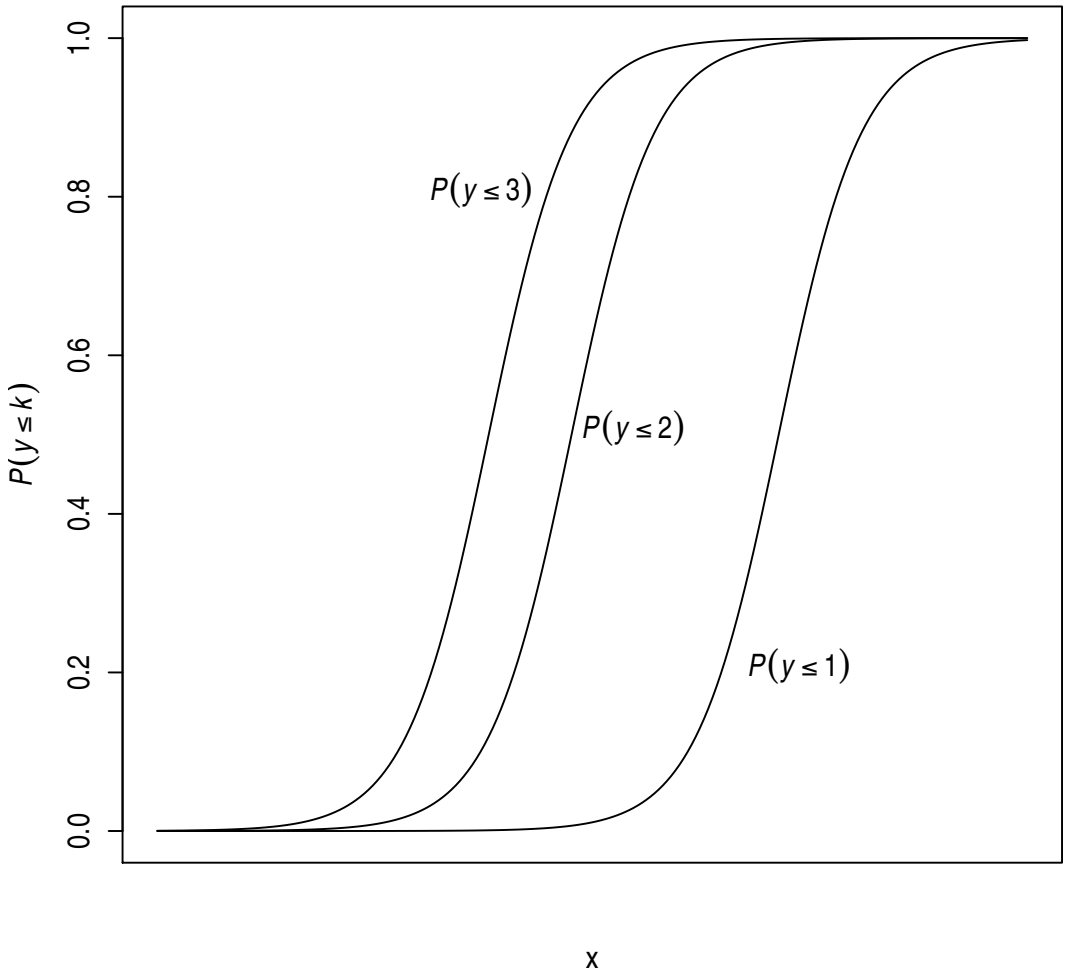
$$P(Y = j) = \frac{\exp(\alpha_j + \beta' \mathbf{x})}{1 + \exp(\alpha_j + \beta' \mathbf{x})} - \frac{\exp(\alpha_{j-1} + \beta' \mathbf{x})}{1 + \exp(\alpha_{j-1} + \beta' \mathbf{x})}$$

Modelli con logit cumulati: caso con un predittore

Consideriamo il caso di un singolo predittore x.
Il modello sarà

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta x, \quad j = 1, \ldots, c-1.$$

Modelli con logit cumulati: caso con un predittore



Per ogni fissato j le curve sono delle logistiche con risposta binaria $Y \leq j$ e $Y > j$. Effetto comune di β per $j = 1, 2, 3$.

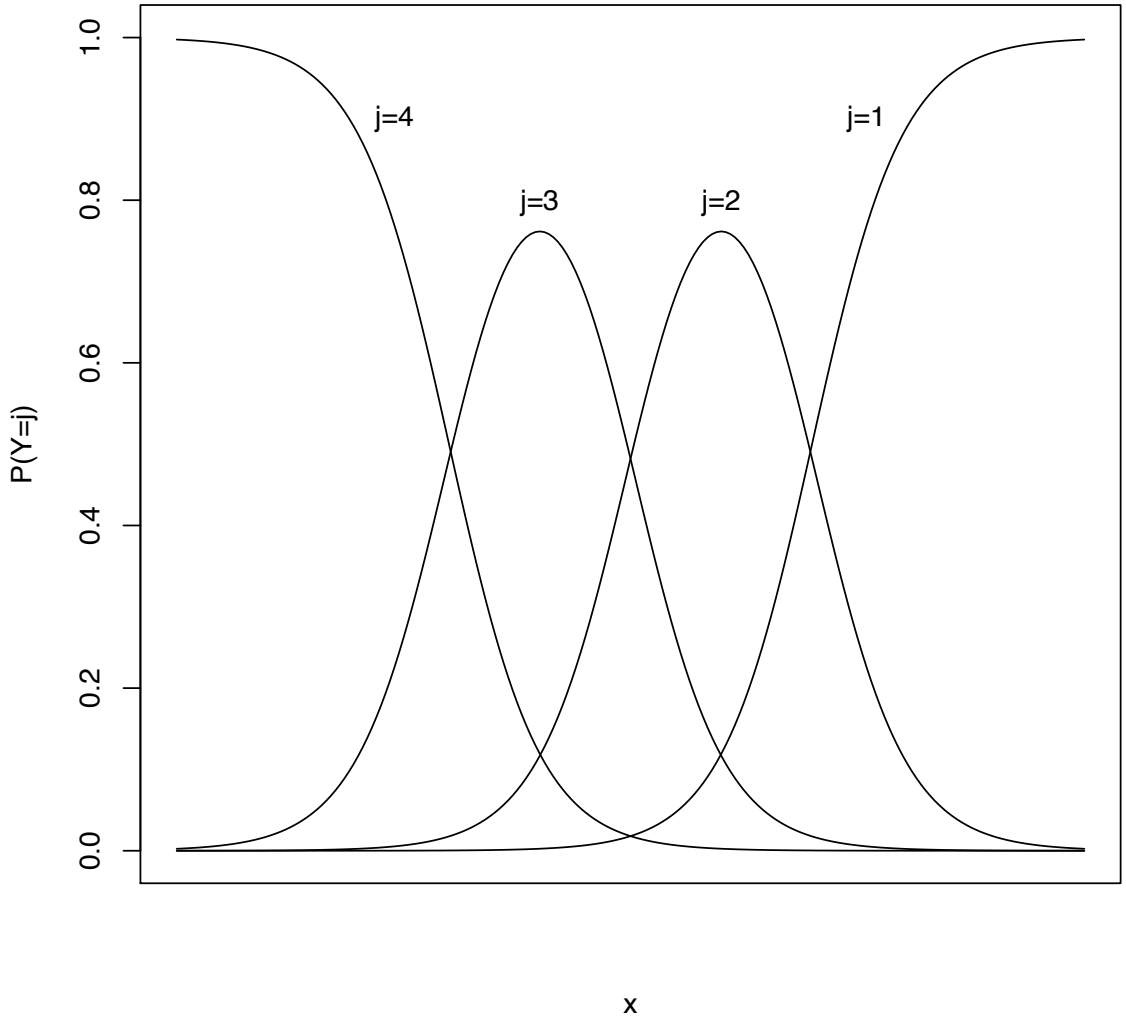
Modelli con logit cumulati: caso con un predittore

- Dal momento che curve per differenti probabilità cumulate hanno la stessa forma, qualunque curva è identica a qualsiasi altra, solo traslata a destra o a sinistra.
- Per $j < k$, la curva per $P(Y \leq k)$ è la curva per $P(Y \leq j)$ traslata di $(\alpha_k - \alpha_j)/\beta$ unità nella direzione di x,

$$P[Y \leq k|X = x] = P\left[Y \leq j|X = x + \frac{\alpha_k - \alpha_j}{\beta}\right].$$

- Le intercette α_j sono necessarie per determinare le probabilità cumulate, ma il parametro di interesse è β, in quanto descrive l'effetto di x.

Modelli con logit cumulati: caso con un predittore



$$P(Y = j) = P(Y \leq j) - P(Y \leq j - 1) \text{ con } \beta > 0.$$

Modelli con logit cumulati: caso con un predittore

- Se $\beta > 0$, ogni logit cumulato cresce al crescere di x, ovvero $P(Y \leq j)$ cresce al crescere di x.
- Ciò implica che le distribuzioni condizionate di Y sono stocasticamente inferiori per livelli più elevati di x.
- Questo può creare problemi di interpretazione.

Modelli con logit cumulati: parametrizzazione alternativa

Spesso si ricorre a una scrittura alternativa del modello con logit cumulati

$$\text{logit}[P(Y \leq j)] = \alpha_j - \beta x$$

- In questo caso, con il segno negativo prima di β, si ha l'interpretazione usuale.
- Per esempio, con $\beta > 0$, Y si trova sulla *parte superiore della scala*, al crescere di x.

Modelli con logit cumulati: caso con più predittori

Possiamo generalizzare il caso di un singolo predittore, considerando più predittori

$$\text{logit}[P(Y \leq j)] = \alpha_j - \beta_1 x_1 - \cdots - \beta_p x_p, \qquad j = 1, \ldots, c-1.$$

Ogni logit cumulato ha la propria intercetta α_j, ma l'effetto β_k della k-esima covariata, per $k = 1, \cdots, p$, è lo stesso per tutte le categorie.

Modelli con logit cumulati: variabile latente

- Un'interessante interpretazione del modello con logit cumulati, considera la variabile ordinale Y come una discretizzazione di una variabile continua latente Y^*.
- Si supponga che Y^* vari attorno a un parametro di posizione η, come la media, che dipende da $\mathbf{x}$, ovvero $\eta(\mathbf{x}) = \beta'\mathbf{x}$.
- Si supponga inoltre che

$$P(Y^* \leq y^*|\mathbf{x}) = G(y^* - \eta) = G(y^* - \beta'\mathbf{x})$$

Modelli con logit cumulati: variabile latente

Fissato un valore di $\mathbf{x}$, si ha

$$Y^* = \beta' \mathbf{x} + \epsilon$$

dove ϵ si distribuisce secondo $G(\cdot)$ (funzione di ripartizione) con $\mathbf{E}(\epsilon) = 0$

Modelli con logit cumulati: variabile latente

Definiamo delle soglie sulla scala continua come

$$-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$$

e assumiamo che $Y = j$ se

$$\alpha_{j-1} < Y^* \leq \alpha_j$$

Modelli con logit cumulati: variabile latente

Dato che $Y^* = \beta'\mathbf{x} + \epsilon$, possiamo scrivere

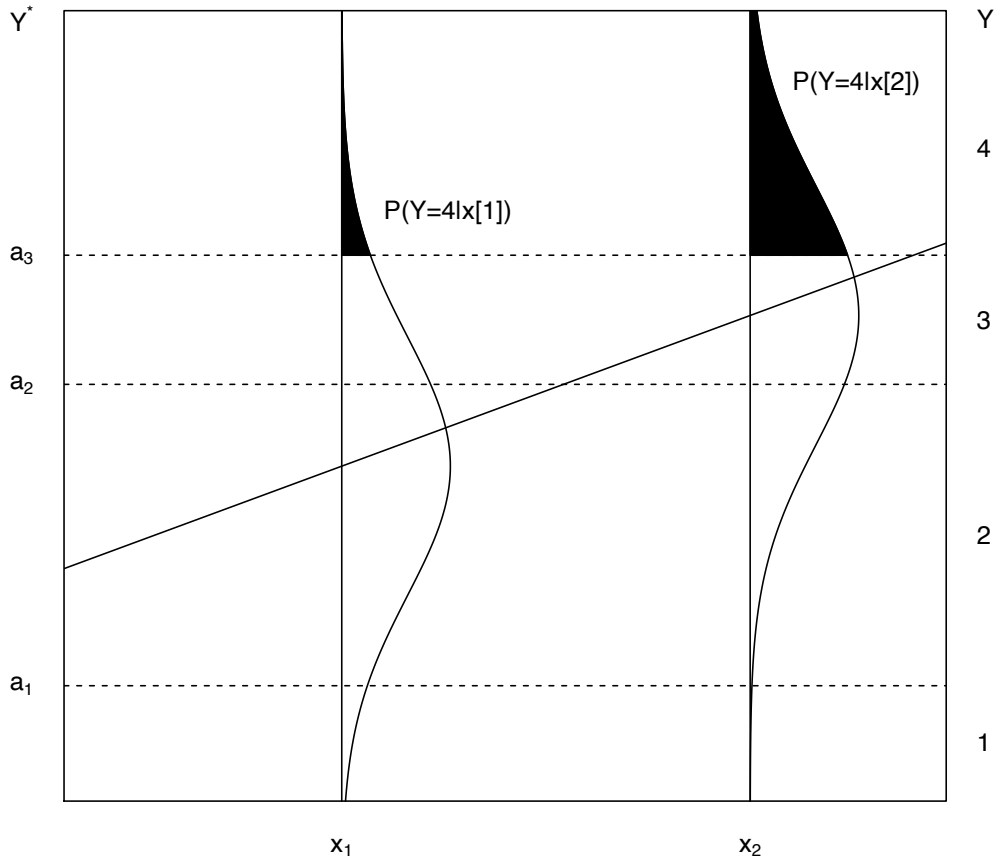
$$\begin{aligned} P(Y \leq j|\mathbf{x}) &= P(Y^* \leq \alpha_j|\mathbf{x}) \\ &= P(\beta'\mathbf{x} + \epsilon \leq \alpha_j|\mathbf{x}) \\ &= P(\epsilon \leq \alpha_j - \beta'\mathbf{x}|\mathbf{x}) \\ &= G(\alpha_j - \beta'\mathbf{x}) \end{aligned}$$

dato che $G(t) = P(\epsilon \leq t)$ e dunque

$$G^{-1}[P(Y \leq j|\mathbf{x})] = \alpha_j - \beta'\mathbf{x}.$$

- Se G è la distribuzione logistica standard, allora G^{-1} è la funzione logit e il modello è il modello a logit cumulati.
- Da notare che ora β è preceduto dal segno negativo.

Modelli con logit cumulati: variabile latente



Modelli con logit cumulati: esempio

Consideriamo il caso di un'indagine di customer satisfaction condotta da una banca.
Per 500 clienti della banca, selezionati casualmente, sono state rilevate, tra le altre, le seguenti informazioni:

- soddisfazione (4 livelli)
- età
- possesso dell'auto

Modelli con logit cumulati: esempio

	Estimate	SE	*t*-value	Wald 95% conf. limits	
Model with age and car possession					
(Intercept 1|2)	−0.5803	0.3569	−1.6256	−1.2798	0.1193
(Intercept 2|3)	0.1778	0.3499	0.5081	−0.5080	0.8636
(Intercept 3|4)	1.5289	0.3560	4.2951	0.8312	2.2265
age	0.0386	0.0068	5.6450	0.0252	0.0519
car possession	−0.4080	0.2259	−1.8060	−0.8508	0.0348
$D = 1182.31$ with 5 d.f.					
Model with age only					
(Intercept 1|2)	−0.2901	0.3187	−0.9105	−0.9147	0.3344
(Intercept 2|3)	0.4678	0.3110	1.5041	−0.1418	1.0774
(Intercept 3|4)	1.8137	0.3198	5.6719	1.1870	2.4405
age	0.0374	0.0068	5.5240	0.0242	0.0573
$D = 1185.64$ with 4 d.f.					

Come interpretiamo questi risultati?

Case Study

Studio di caso: soddisfazione del cliente

- In molti ambiti commerciali e aziendali, l'analisi della soddisfazione del cliente, customer satisfaction, rappresenta un indicatore chiave del successo dell'azienda
- Le indagini di customer satisfaction vengono usualmente condotte per mezzo di questionari
- La soddisfazione complessiva si misura tipicamente tramite risposta ordinale ad una specifica domanda
- Un questionario contiene varie domande riguardanti l'opinione e le aspettative del cliente rispetto all'azienda, al prodotto, ai servizi offerti

Studio di caso: soddisfazione del cliente

- Si analizza il caso di un'azienda nel settore ICT, che produce software e servizi di consulenza collegati
- I clienti di questa azienda sono banche che adottano software, applicazioni e servizi correlati
- Campione casuale di 4000 'clienti' (dipendenti delle banche, utilizzatori dei prodotti e servizi dell'azienda)
- Le opinioni sono raccolte, chiedendo ai clienti di dare una valutazione sui vari aspetti che caratterizzano la relazione tra azienda e cliente

Studio di caso: soddisfazione del cliente

La soddisfazione complessiva viene misurata attraverso una sola domanda, alla fine del questionario:

"Richiamando tutti gli aspetti analizzati nel questionario, quanto si ritiene soddisfatto dall'azienda nel complesso?"

Studio di caso: soddisfazione del cliente

La risposta è stata codificata in 6 livelli

Livello	*Descrizione*
1	estremamente soddisfatto
2	molto soddisfatto
3	abbastanza soddisfatto
4	abbastanza insoddisfatto
5	molto insoddisfatto
6	estremamente insoddisfatto

La "soddisfazione complessiva" è una variabile categoriale ordinale

Studio di caso: soddisfazione del cliente

Variabili presenti nel questionario

- prodotti/servizi utilizzati: *quali prodotti/servizi dell'azienda utilizza?*

Variabile	*Prodotto/servizio*
v1	1
v2	2
v4	3
v5	4
v6	5
v7	6
v8	7
v9	altro

Studio di caso: soddisfazione del cliente

Variabili presenti nel questionario
Soddisfazione per staff e prodotti

- 10 livelli (1: totalmente in disaccordo, 10: totalmente d'accordo)
- v11 (1: no, 2: sì, una volta, 3: sì, qualche volta, 4: sì, spesso)

Variabile	*Domanda*
v11	contatti con personale
v24	prodotti facili da usare
v25	prodotti facili da adattare a esigenze
v26	prodotti sono quel che mi serve
v27	prodotti sono affidabili
v28	prodotti facili da integrare

Studio di caso: soddisfazione del cliente

Variabili presenti nel questionario

- *Importanza* di singoli aspetti in una azienda IT
  (1: per nulla importante, 10: molto importante)

Variabile	*Domanda*
v29	esperienza del personale
v30	servizio di consulenza efficiente
v31	problem solving
v32	affidabilità di prodotti e servizi
v33	flessibilità di prodotti e servizi
v34	efficienza di prodotti e servizi
v35	velocità
v36	utilità del personale
v37	efficienza del personale
v38	predisposizione verso le esigenze del cliente
v39	capacità di rispondere alle esigenze
v40	flessibilità
v41	innovatività

Studio di caso: soddisfazione del cliente

Variabili presenti nel questionario

- *Soddisfazione* verso singoli aspetti in una azienda IT
  (1: per nulla soddisfatto, 10: molto soddisfatto)

Variabile	*Domanda*
v42	esperienza del personale
v43	servizio di consulenza efficiente
v44	problem solving
v45	affidabilità di prodotti e servizi
v46	flessibilità di prodotti e servizi
v47	efficienza di prodotti e servizi
v48	velocità
v49	utilità del personale
v50	efficienza del personale
v51	predisposizione verso le esigenze del cliente
v52	capacità di rispondere alle esigenze
v53	flessibilità
v54	innovatività

Studio di caso: soddisfazione del cliente

Variabili presenti nel questionario

- *Soddisfazione complessiva e caratteristiche dei clienti*

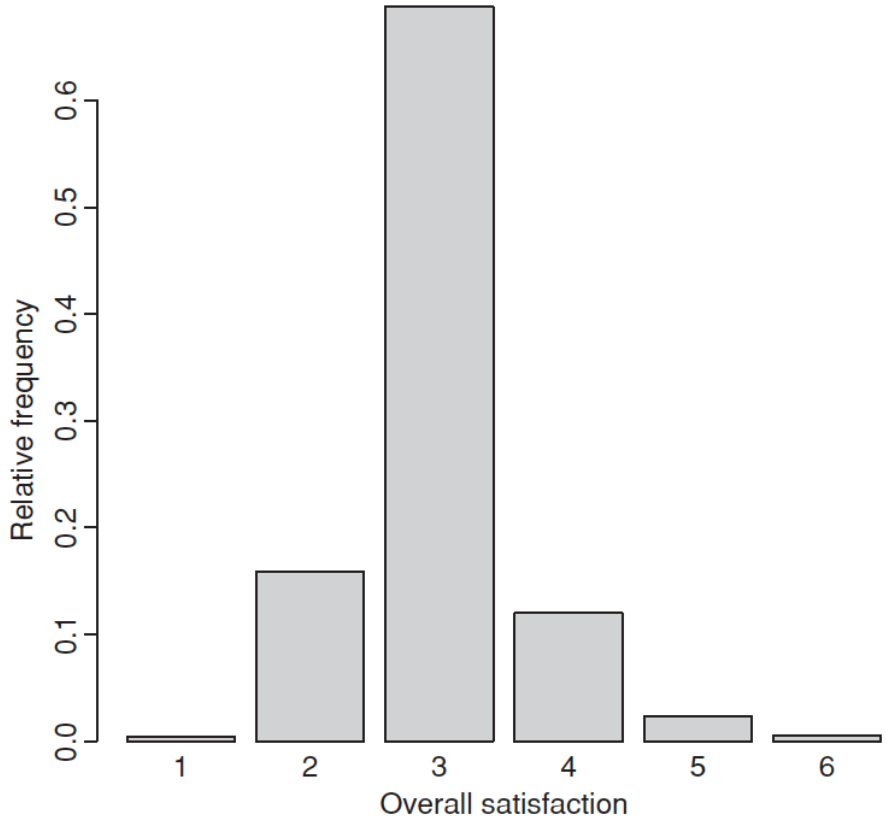
Variabile	*Domanda*
v56	soddisfazione complessiva - variabile risposta
v58	categoria professionale
v59	status professionale
v60	età
v61	esperienza lavorativa
v62	livello di istruzione
v63	genere

Studio di caso: soddisfazione del cliente

Ovviamente i marketing manager sono interessati a identificare gli aspetti specifici più strettamente connessi con la soddisfazione complessiva.
Per far questo è possibile prevedere la variabile di interesse, seguendo tre diverse strategie:

- considerare la variabile risposta come categoriale ordinale
- considerare la variabile risposta come categoriale, ignorando l'ordinalità
- considerare la variabile risposta come una quantitativa discreta

Studio di caso: soddisfazione del cliente



Soddisfazione complessiva nel training set (3000 clienti).
Circa il 69% si dichiara abbastanza soddisfatto.

Studio di caso: soddisfazione del cliente

Variabile risposta come categoriale ordinale

Il primo modello che viene stimato è il modello con logit cumulati, per sfruttare l'ordinamento nella variabile risposta.
Il miglior modello viene selezionato con una procedura stepwise basata sull'AIC.

Studio di caso: soddisfazione del cliente

- Uno degli aspetti interessanti del modello con logit cumulati è la sua *interpretabilità*
- Da notare che le categorie della variabile risposta sono in ordine invertito, rispetto a quello usuale
  1: estremamente soddisfatto, 6: estremamente insoddisfatto
- Dunque se un coefficiente stimato ha segno positivo, al crescere di quella variabile, aumenta la probabilità di appartenere alle categorie più alte . . .
  . . . quindi di essere più insoddisfatto