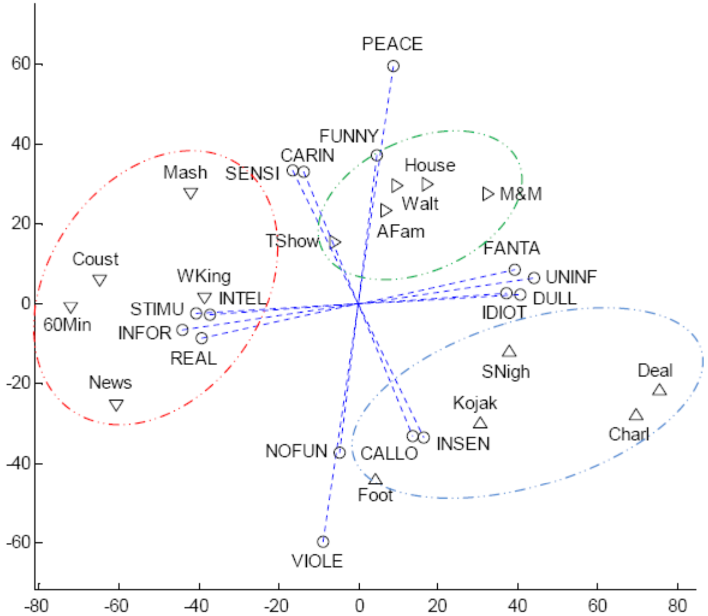


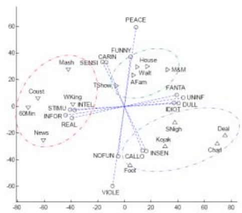
Metodi Statistici per il Management

Modelli statistici di dipendenza

Simone Borra - Roberto Rocci

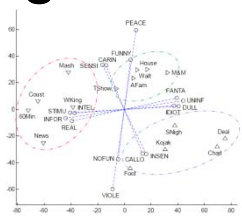
Modelli statistici di dipendenza

- **Modelli**: strutture formali che hanno l’obiettivo di descrivere, spiegare e comprendere (semplificando) fenomeni complessi.
- Si parla di **modelli statistici** quando la formalizzazione si basa sull’utilizzo di metodi e strumenti tipici della scienza matematica.
- Nei modelli statistici si accetta a priori che possa esistere un certo grado di **imprecisione** e di **incertezza.**
- In un modello statistico si cerca di stimare il grado di incertezza e di descriverlo mediante opportune strutture matematiche: le ***variabili casuali***.



Modelli statistici di dipendenza

- L'obiettivo di un modello statistico di analisi della dipendenza è quello di studiare come varia una determinata variabile Y (detta dipendente o risposta), in funzione del variare di alcune variabili $X_1, \ldots, X_k$ (dette indipendenti o esplicative).
- Esempio: Y = prezzo di vendita di un appartamento, X = superficie dell'appartamento.
- Esempio: Y = mi piace/non mi piace fare acquisti nel supermercato ABC, X = ritengo/non ritengo che il supermercato ABC abbia una buona varietà di prodotti.
- Esempio: Y = l'individuo è assunto/non è assunto dopo lo stage, X = punteggio al test di ammissione allo stage.



Modelli statistici di dipendenza

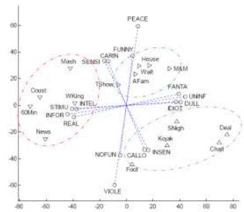
- Un modello statistico di dipendenza è rappresentato dalla funzione di densità/probabilità condizionata:

$$f(y|x_1, \ldots, x_k)$$

dove:

  - y = variabile dipendente o risposta;
  - $x_1, \ldots, x_k$ = variabili indipendenti o esplicative.

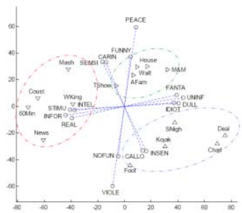
- Condizionatamente ai valori delle variabili esplicative, la variabile risposta rimane comunque una variabile aleatoria $\Rightarrow$ le X non determinano completamente la Y.



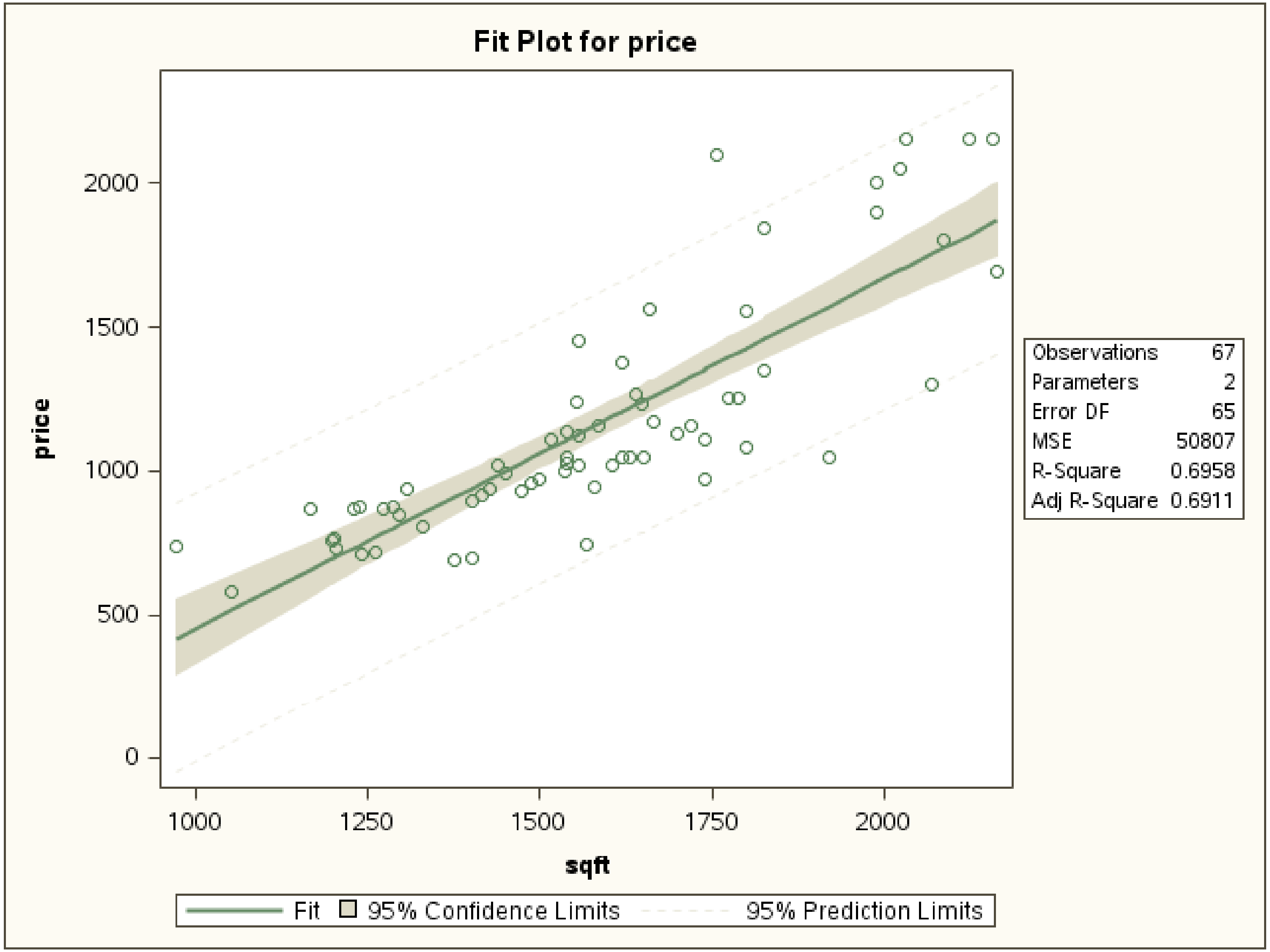
Modelli statistici di dipendenza

Il processo di formulazione di un modello può essere scomposto nei seguenti passaggi logici:

1. Posizione del problema
2. Scelta della classe di modelli
3. Scelta delle variabili e rilevazione dei dati
4. Selezione del modello
5. Stima dei parametri incogniti
6. Analisi della “bontà” del modello.
7. Utilizzo operativo:
    1. Analisi del fenomeno
    2. Previsioni
    3. Scenari



Esempio: Prezzo su superficie



Classificazione Modelli

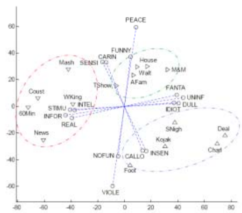
Modelli per tipo di rilevazione del campione

Il campione può essere rilevato in più modi:

1. le variabili sono rilevate su *n* unità distinte, in un certo istante temporale
2. le variabili sono rilevate in *T* diversi istanti temporali, ma riferite sempre alle medesime grandezze
3. le variabili sono rilevate in *T* diversi istanti temporali sulle stesse *n* unità

Abbiamo quindi modelli:

1. sezionali o cross section
2. per serie storiche o time series
3. per osservazioni ripetute o panel data



Regressione lineare (reg)

- Il modello di regressione lineare semplice è così formalizzato:

$$y_i = b_0 + b_1 x_i + w_i$$

  dove:

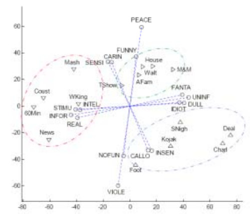
  y_i = variabile dipendente o risposta;

  x_i = variabile indipendente o esplicativa;

  b_0, b_1 = parametri incogniti (intercetta e coefficiente di regressione);

  w_i = termine di errore (o disturbo).

- Il termine $\mathbf{w}_i$ deve presentare i seguenti requisiti:
  – distribuito in modo normale;
  – a media zero;
  – omoschedastico (varianza costante indipendente dalla variabile x);
  – osservazioni indipendenti.

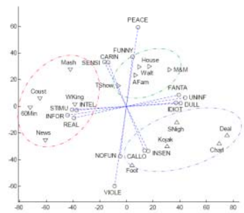


reg: interpretazione

- $\mathbf{E(y_{\mathit{i}} \,|\, x_{\mathit{i}})} = \mu_i = \boldsymbol{b_0} + \boldsymbol{b_1}\mathbf{x}_{\mathit{i}}$, la relazione lineare è vera in media
- $\boldsymbol{b_0} = \mathbf{E(y_{\mathit{i}} \,|\, x_{\mathit{i}} = 0)}$, rappresenta l'influenza di variabili omesse che non variano con ***i***
- $\boldsymbol{b_1}$ incremento di μ_i corrispondente ad un aumento di una unità di $\mathbf{x}_{\mathit{i}}$
- $\mathbf{w}_{\mathit{i}}$ incorpora variabili omesse ed imperfezioni della relazione lineare che intercorre tra $\mathbf{y}_{\mathit{i}}$ e $\mathbf{x}_{\mathit{i}}$
- Il modello può anche essere formulato come

Equazione	*Assunzioni*
$E(y_i \,|x_i) = \mu_i = b_0 + b_1 x_i$	$y_i \,|x_i \sim N(\mu_i, \sigma_w^2)$, indip.

Quindi $f(y_i \,|x_i)$ è una normale con media dipendente da x_i



reg: stima puntuale

- (metodo dei minimi quadrati) le stime di $\boldsymbol{b_0}$ e $\boldsymbol{b_1}$ sono calcolate minimizzando

$$\Sigma_i (y_i - b_0 - b_1 x_i)^2$$

in questo modo otteniamo

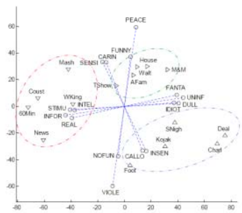
$$\hat{b}_1 = \frac{\sigma_{XY}}{\sigma_X^2}; \hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X}$$

- la varianza dell'errore è stimata come

$$\hat{\sigma}_w^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- sono stimatori corretti ed efficienti nella classe degli stimatori lineari (nei dati) e corretti

$$E(\hat{b}_0) = b_0, E(\hat{b}_1) = b_1, E(\hat{\sigma}_W^2) = \sigma_W^2$$

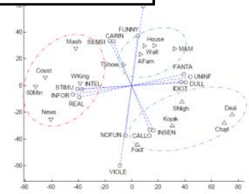


reg: esempio – prezzo su superficie (sqft)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7551970	7551970	148.64	<.0001
Error	65	3302439	50807		
Corrected Total	66	10854409			

Root MSE	225.40352	R-Square	0.6958
Dependent Mean	1161.4627	Adj R-Sq	0.6911
Coeff Var	19.40687		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-758.78383	159.89171	-4.75	<.0001	-1078.10963	-439.458
sqft	1	1.21372	0.09955	12.19	<.0001	1.0149	1.41254



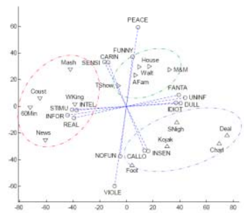
reg: bontà di adattamento

- **SSModel =** $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
- **SSResidual =** $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- **SSTotal =** $\sum_{i=1}^{n} (y_i - \bar{y})^2$

SST = SSM+SSR

Il coefficiente di determinazione è una misura della bontà di adattamento del modello ai dati ed è costruito come **R^2 = SSM/SST = 1 – SSR/SST**.

Ha sempre un valore compreso tra 0 e 1.



reg: test t

- Ipotesi nulla H_0: $b_1=0$
- Statistica test

$$t_{oss} = \text{Coef./(Std. Err.)} = \hat{b}_1 / se(\hat{b}_1),$$

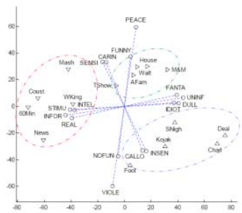
$$t_{oss} \sim T_{n\text{-}p\text{-}1} \text{ se } H_0 \text{ vera}$$

- regola di rifiuto basata sulla statistica test

$$| t_{oss} | > t_{\alpha/2}$$

- p-value

$$\text{p-value} = 2\Pr\{T_{n\text{-}p\text{-}1} > | t_{oss} |\}$$

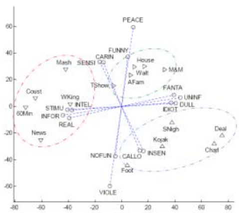


reg: test t

- Se il modello è vero, l'ipotesi nulla H_0: b_1=0 equivale ad affermare che non vi è legame tra X e Y.

- regola di rifiuto basata sul p-value (sempre valida)

$$\text{p-value} < \alpha$$

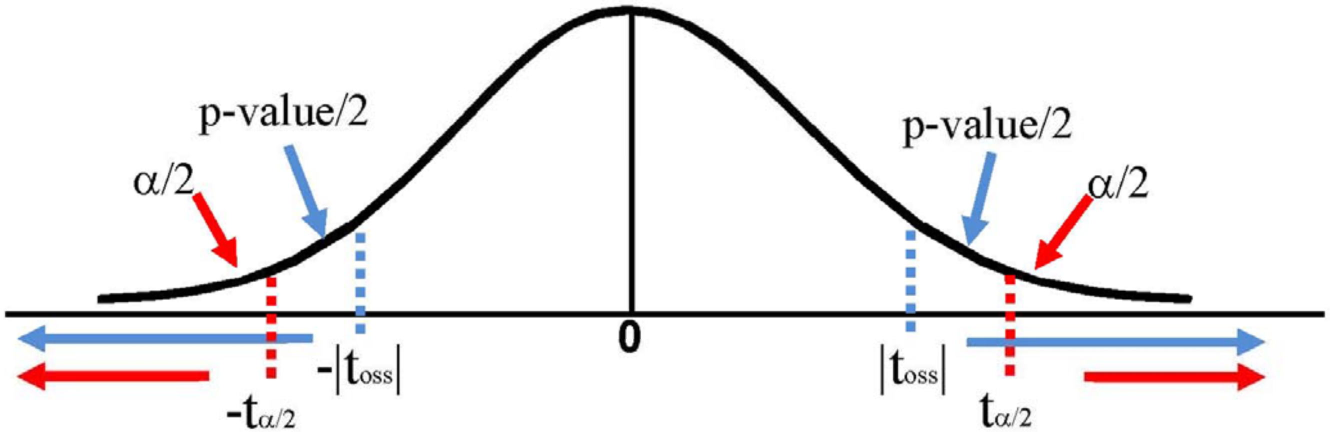


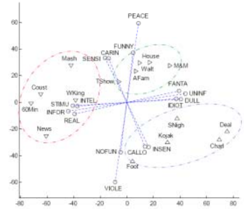
reg: test t

P-value: Accetto H_0 al livello α

$|t_{oss}| < t_{\alpha/2}$

p-value > α



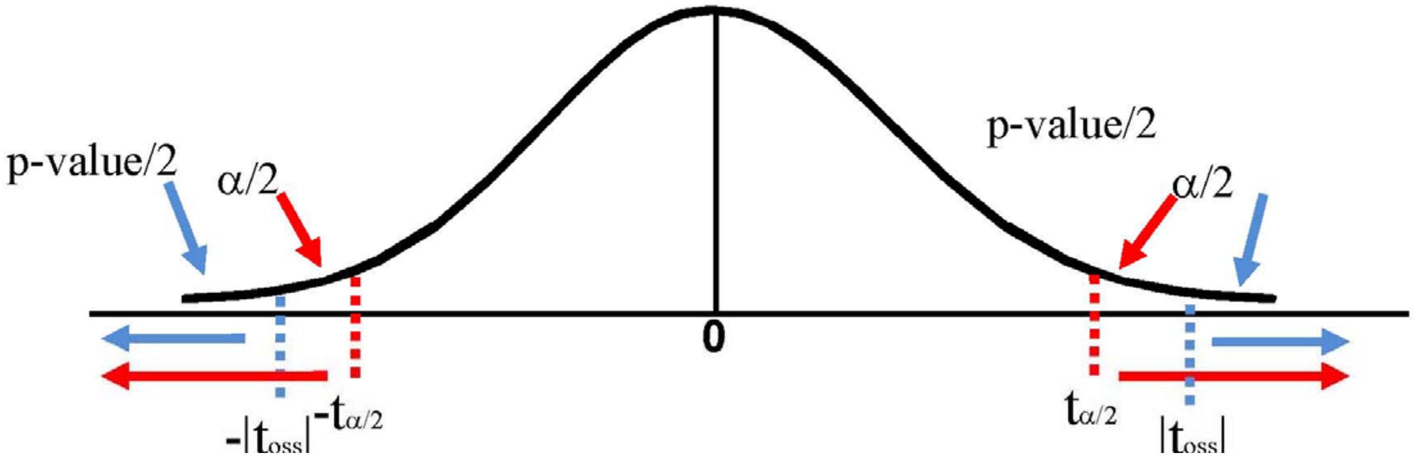


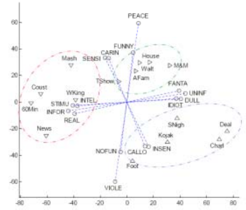
reg: test t

P-value: Rifiuto H_0 al livello α

$|toss| > t\alpha/2$

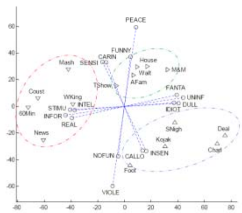
p-value < α





Y binaria: Logit e Probit

- In alcuni casi la variabile dipendente è di natura binaria ed assume solo i due valori 0 o 1. Questi sono generalmente il risultato di una codifica.
- Esempi:
  - pazienti di una patologia potrebbero reagire o meno ad una terapia innovativa;
  - in un test aziendale per una promozione interna alcuni impiegati potrebbero superare la prova e altri no;
  - dopo un periodo di prova alcuni possono essere assunti e altri no.

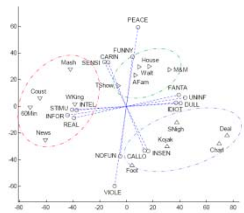


Regressione logit

Ad esempio potremmo essere interessati a capire se per un cliente è importante che in un supermercato ci sia una sufficiente varietà di prodotti.

Has a sufficient choice of brands/types/sizes of products * Shopping there makes me feel good Crosstabulation

			Shopping there makes me feel good		
			No	Yes	Total
Has a sufficient choice of brands/types/sizes of products	No	Count	90	38	128
			70.3%	29.7%	100.0%
	Yes	Count	92	140	232
			39.7%	60.3%	100.0%
Total		Count	182	178	360
			50.6%	49.4%	100.0%



logit: variabile riposta

Variabile aleatoria di Bernoulli

Distribuzione

y	$p(y)$
0	$1-\vartheta$
1	ϑ

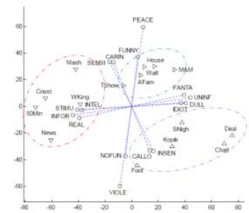
Caratteristiche

$$E(y) = \vartheta$$

$$V(y) = \vartheta(1-\vartheta)$$

Importante osservare che:

- la particolare codifica adottata ci permette di interpretare la probabilità come una media;
- la varianza dipende dalla media.



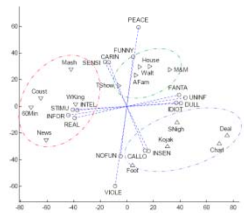
logit

Possiamo utilizzare il modello lineare

$$E(y_i \mid x_i) = \mu_i = b_0 + b_1 x_i \ ?$$

No perchè:

- la *y* ha distribuzione di Bernoulli e non Normale;
- l'ipotesi di omoschedasticità non è sicuramente verificata;
- i valori stimati di $\mathbf{E(y_{\mathit{i}} \mid x_{\mathit{i}})}$ non necessariamente ricadono nell'intervallo [0,1];
- in molte situazioni reali si è visto che la probabilità di un evento varia in funzione di una o più variabili esplicative in modo non lineare.



logit: modello

Equazione

$$\vartheta_i = \frac{\exp(b_0 + b_1 x_i)}{1 + \exp(b_0 + b_1 x_i)}$$

Assunzioni

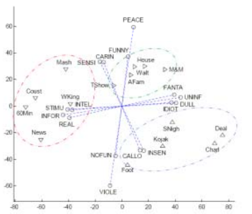
$y_i | x_i \sim \text{Be}(\vartheta_i)$, indip.

o equivalentemente

$$\text{logit}(\vartheta_i) = b_0 + b_1 x_i$$

$y_i | x_i \sim \text{Be}(\vartheta_i)$, indip.

dove $\text{logit}(\vartheta_i) = \log\left(\frac{\vartheta_i}{1-\vartheta_i}\right)$ è il logaritmo dell’ “odds”.



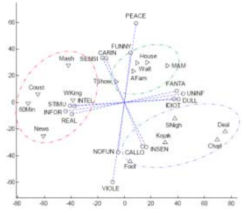
logit: odds

Odds = numero di chance favorevoli all’accadimento contro una sfavorevole.

Le odds sono un diverso modo di misurare il grado di fiducia nell’accadimento di un evento.

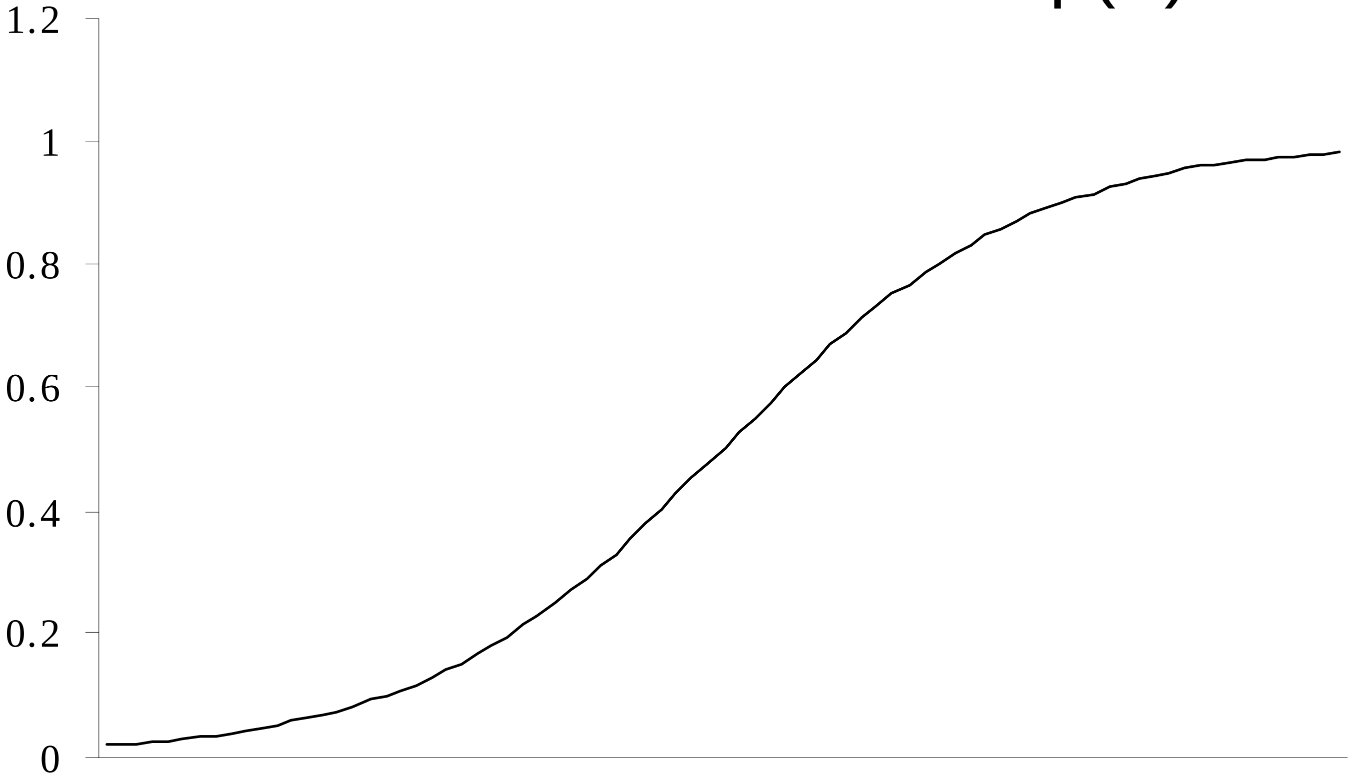
Sono così legate con le probabilità:

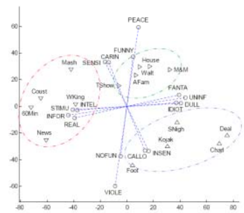
$$\text{probabilità dell'evento} = \frac{\text{odds}}{\text{odds} + 1}; \qquad \text{odds} = \frac{\text{probabilità}}{1 - \text{probabilità}}$$



logit: logistica

Grafico della funzione logistica $\vartheta = \dfrac{\exp(x)}{1+\exp(x)}$





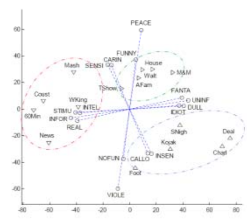
logit: interpretazione

b_0 valore di logit(ϑ) quando $\mathbf{x} = \mathbf{0}$, b_1 incremento di logit(ϑ) se x_j aumenta di una unità.

Importante osservare che se x_j aumenta di una unità allora l'incremento di ϑ dipende dal valore di x_j.

Esempio Per studiare la relazione tra *feel* e *choice* consideriamo il modello $\text{logit}(\vartheta_i) = b_0 + b_1 x_i$

$$\text{logit}(\Pr(feel=1)) = \begin{cases} b_0 & \text{scelta insufficiente} \\ b_0 + b_1 & \text{scelta sufficiente} \end{cases}$$



logit: interpretazione

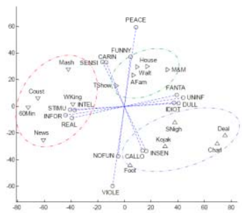
risulta

$$b_1 = \text{logit}(\vartheta_{Ssi}) - \text{logit}(\vartheta_{Sno}) = \log\left(\frac{\vartheta_{Ssi}/(1-\vartheta_{Ssi})}{\vartheta_{Sno}/(1-\vartheta_{Sno})}\right)$$

quindi

$$\exp(b_1) = \frac{\vartheta_{Ssi}/(1-\vartheta_{Ssi})}{\vartheta_{Sno}/(1-\vartheta_{Sno})} \Rightarrow \exp(b_1)\frac{\vartheta_{Sno}}{1-\vartheta_{Sno}} = \frac{\vartheta_{Ssi}}{1-\vartheta_{Ssi}}$$

In generale possiamo dedurre che se il regressore aumenta di una unità, allora l'odds aumenta $\exp(b_1)$ volte



logit: stima ML

La stima dei parametri avviene con il metodo della massima verosimiglianza. L'idea è quella di considerare il campione osservato come quello che aveva la maggiore probabilità di essere estratto.

Funzione di probabilità del campione

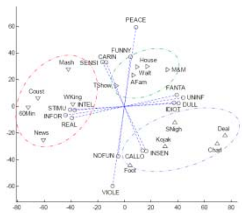
$$p(\mathbf{y};\vartheta) = \vartheta_1^{y_1}(1-\vartheta_1)^{1-y_1}\cdot\ldots\cdot\vartheta_n^{y_n}(1-\vartheta_n)^{1-y_n}$$

Funzione di verosimiglianza (Likelihood function)

$$L(b_0,b_1;\mathbf{y}) = \vartheta_1^{y_1}(1-\vartheta_1)^{1-y_1}\cdot\ldots\cdot\vartheta_n^{y_n}(1-\vartheta_n)^{1-y_n};\mathrm{logit}(\vartheta_i) = b_0 + b_1 x_i$$

Stimatore di massima verosimiglianza

$$(\hat{b}_0,\hat{b}_1) = \arg\max L(b_0,b_1;\mathbf{y})$$



logit: stima ML

Esempio: Y = 1 non assunto dopo stage

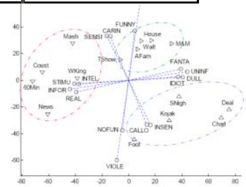
dex = punteggio al test di entrata

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	795.912	748.014
SC	800.438	757.067
-2 Log L	793.912	744.014

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	49.8977	1	<.0001
Score	47.3598	1	<.0001
Wald	44.1038	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.9863	0.5948	25.2094	<.0001
dex	1	-0.0909	0.0137	44.1038	<.0001

Parameter Estimates and Wald Confidence Intervals			
Parameter	Estimate	95% Confidence Limits	
Intercept	2.9863	1.821	4.152
dex	-0.0909	-0.12	-0.0641



logit: test Z

- Ipotesi nulla H_0: $b_1=0$
- Statistica test

$$z_{oss} = \text{Coef./(Std. Err.)} = \hat{b}_1 / se(\hat{b}_1)$$

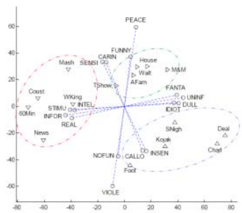
$$z_{oss} \sim N(0,1) \text{ se } H_0 \text{ vera e } n \text{ suff. elevato}$$

- Regola di rifiuto

$$| z_{oss} | > z_{\alpha/2}$$

- P-value

$$2\Pr\{Z > | z_{oss}|\}$$



logit: test di Wald

- Ipotesi nulla H_0: $b_1=0$
- Statistica test

$$w_{oss} = (z_{oss})^2$$

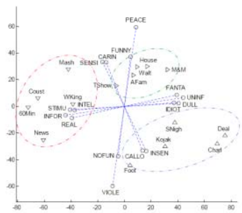
$$w_{oss} \sim \chi^2(1) \text{ se } H_0 \text{ vera e } n \text{ suff. elevato}$$

- Regola di rifiuto

$$W_{oss} > \chi^2_{\alpha}$$

- P-value

$$\Pr\{\chi^2(1) > w_{oss}\}$$



Regressione probit: modello

Modello alternativo al logit. Differisce da questo perché usa la funzione probit (Φ^{-1}) al posto della logit.

Equazione

$$\vartheta_i = \Phi\left(b_0 + b_1 x_i\right)$$

o equivalentemente

$$\Phi^{-1}(\vartheta_i) = b_0 + b_1 x_i$$

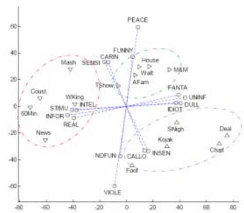
Assunzioni

$y_i|x_i \sim \mathrm{Be}(\vartheta_i)$, indip.

$y_i|x_i \sim \mathrm{Be}(\vartheta_i)$, indip.

dove Φ e Φ^{-1} sono la funzione di ripartizione della variabile aleatoria normale standard e la sua inversa.

Tale funzione è monotona crescente, quindi se b_1 è positivo ϑ cresce al crescere di $\boldsymbol{x}$.



probit: stima ML

Esempio: Y = 1 non assunto dopo stage
dex = punteggio al test di entrata

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	795.912	749.437
SC	800.438	758.49
-2 Log L	793.912	745.437

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	48.4746	1	<.0001
Score	47.3598	1	<.0001
Wald	43.9215	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.6619	0.3445	23.2714	<.0001
dex	1	-0.0516	0.00779	43.9215	<.0001

Parameter Estimates and Wald Confidence Intervals			
Parameter	Estimate	Limits	
Intercept	1.6619	0.9867	2.3371
dex	-0.0516	-0.067	-0.0364