UNIVERSITÀ DEGLI STUDI DI PADOVA
FACOLTÀ DI SCIENZE STATISTICHE
CORSO DI LAUREA TRIENNALE IN
STATISTICA E TECNOLOGIE INFORMATICHE

MODELLI DI REGRESSIONE CON VARIABILE RISPOSTA DISTRIBUITA NELL'INTERVALLO (0,1)

RELATORE: Prof.ssa Alessandra Salvan

LAUREANDO: *Carlo Pianezzola, 553095*

ANNO ACCADEMICO 2008-2009

Lista dei simboli

Simboli matematici

- Y: Vettore di variabili casuali che descrivono la risposta.
- n: Numerosità campionaria.
- i: Variabile naturale contatore che assume valori fino ad n, relativa alle osservazioni.
- Y_i: Variabile casuale che descrive la risposta dell'i-esima unità, $i = 1, \ldots, n$.
- k: Numero dei coefficienti di regressione.
- p: Funzione di densità.
- ℓ: Funzione di log-verosimiglianza.
- ℓ_*: Derivata prima della log-verosimiglianza.
- y: Realizzazione di Y.
- r: Variabile naturale che assume valori fino ad k.
- s: Variabile naturale che assume valori fino ad k, utilizzata per le derivate seconde.
- β: Vettore k-dimensionale che indica i coefficienti di regressione di un modello.
- ε: Vettore n-dimensionale degli errori del modello.
- X: Matrice nxk delle variabili esplicative.

- x_i: Vettore k-dimensionale delle variabili esplicative per l'i-esima osservazione.
- x_{ir}: Valore dell'r-esima variabile esplicativa per l'i-esima unità.
- $E[\cdot]$: Valore atteso.
- $Var[\cdot]$: Varianza.
- $V[\cdot]$: Funzione di varianza.
- μ: Vettore dei valori attesi.
- σ^2: Varianza del modello.
- σ: Scarto quadratico medio del modello.
- $.\hat{}.$: Stima del parametro.
- $g(\cdot)$: Funzione legame.
- $h(\cdot)$: Funzione per trasformare la variabile risposta, nel modello logistico.
- η: Previsore lineare.
- $\mathcal{F}_{en}^{p}$: Famiglia esponenziale naturale di ordine p.
- $\mathcal{F}_{de}^{p}$: Famiglia di dispersione esponenziale di ordine p.
- M: Funzione generatrice dei momenti.
- K: Funzione generatrice dei cumulanti.
- j: Matrice di informazione osservata.
- i: Matrice di informazione attesa.
- ∇: Gradiente, vettore delle derivate prime.
- $B(p,q)$: Funzione beta.
- $\Gamma(\alpha)$: Funzione gamma.
- $\mathbb{R}$: Insieme dei numeri reali.
- $N(0,1)$: Distribuzione normale standard.
- $z_{1-\alpha/2}$: Quantile $1-\alpha/2$ della normale standard.

- χ^2_m: Distribuzione χ^2 con m gradi di libertà.

Prospetto analisi della varianza.

Nel Capitolo 2 l'analisi della varianza, del modello lineare, sarà presentata come segue.

- $\mathcal{F}_k$: Modello completo, con k coefficienti di regressione.
- $\mathcal{F}_{k_0}$ Modello ridotto, con k_0 coefficienti di regressione.
- k_0: Numero dei coefficienti di regressione del modello ridotto, vale la relazione $1 < k_0 < k$.
- $SQT = \sum_{i=1}^{n} (y_i)^2$: Somma totale dei quadrati.
- $SQS_{k_0} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$: Somma dei quadrati degli scarti dalla media dei valori predetti, detta somma dei quadrati spiegata dal modello $\mathcal{F}_{k_0}$.
- $SQS_k = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$: Somma dei quadrati degli scarti dalla media dei valori predetti, detta somma dei quadrati spiegata dal modello $\mathcal{F}_k$.
- $SQE_{k_0} = \sum_{i=1}^{n} (\hat{y}_i - y)^2$: Somma dei quadrati dei residui del modello $\mathcal{F}_{k_0}$.
- $SQE_k = \sum_{i=1}^{n} (\hat{y}_i - y)^2$: Somma dei quadrati dei residui del modello $\mathcal{F}_k$. .
- $SQT_{corr} = SQS_k + SQE_k$: Si intente SQT corretta per il livello medio complessivo delle osservazioni.
- F_{x_0,x_1}: Distribuzione F di Fisher, con x_0 e x_1 gradi di libertà.

Fonte di variabilità	Gradi di libertà	Somma dei quadrati	Test su miglioramento distribuzione nulla
Totale	n	SQT	
Costante	1	ny^2	
Totale corretta	$n-1$	SQT_{corr}	
Miglioramento con, $\mathcal{F}_{k_0}$ rispetto a $\mathcal{F}_1$	k_0-1	SQS_{k_0}	$\frac{SQS_{k_0}/(k_0-1)}{SQE_{k_0}/(n-k_0)}$ $F_{k_0-1,n-k_0}$
Miglioramento con, $\mathcal{F}_k$ rispetto a $\mathcal{F}_{k_0}$	$k-k_0$	$SQS_k - SQS_{k_0}$	$\frac{(SQS_k-SQS_{k_0})/(k-k_0)}{SQE_k/(n-k)}$ $F_{k-k_0,n-k}$
Residui di $\mathcal{F}_k$	$n-k$	SQE_k	

Tabella 1: Prospetto analisi della varianza.

Alfabeto greco Nella Tabella 2 si riporta l'alfabeto greco.

minus	MAIUS	Trascrizione		minus	MAIUS	Trascrizione
α	A	Alfa		ν	N	Ni
β	B	Beta		ξ	Ξ	Xi
γ	Γ	Gamma		o	O	Omicron
δ	Δ	Delta		π	Π	Pi, Pgreco
ε	E	Epsilon		ρ	P	Rho
ζ	Z	Zeta		σ	Σ	Sigma
η	H	Eta		τ	T	Tau
θ	Θ	Theta		υ	Υ	Upsilon
ι	I	Iota		ϕ , φ	Φ	Phi
κ	K	Kappa		χ	X	Chi
λ	Λ	Lambda		ψ	Ψ	Psi
μ	M	Mi, Mu		ω	Ω	Omega

Tabella 2: Alfabeto greco

Indice

Introduzione

La maggior parte dei modelli statistici per analizzare la dipendenza di una risposta continua da variabili esplicative assumono, per la risposta, una distribuzione con supporto non limitato. Si pensi ai modelli di regressione normali, gamma, etc.

Ci si imbatte tuttavia in alcuni fenomeni le cui misurazioni hanno limiti ben definiti. Questa rigidità a volte è dettata da questioni fisiche o anche da condizioni pilotate sotto le quali si svolge l'esperimento, tipicamente in ambito di produzione aziendale, ma il più delle volte è l'analisi di dati proporzionali che origina questa situazione.

Molti studiosi hanno cercato metodi che meglio si adattassero a descrivere dati confinati in un intervallo. Poiché qualsiasi intervallo può essere ricondotto all'intervallo (0,1) con una trasformazione di scala e posizione, gli studi sono stati indirizzati a trovare funzioni che si potessero usare come modelli di regressione per dati proporzionali. Da queste considerazioni sono nati i modelli di regressione parametrica basati sulla distribuzione beta e sulla distribuzione del simplesso, che saranno oggetto di questo lavoro.

L'obbiettivo principale della tesi è la rassegna dei modelli di regressione parametrici per variabili risposte in (0,1), proposti in letteratura. Nel primo capitolo vengono esposti i modelli. Nel secondo due insiemi di dati, con variabile risposta in (0,1), vengono analizzati con i modelli proposti.

La ricerca di metodi inferenziali per lo studio di fenomeni con risposte definite in un intervallo è tuttora molto attiva. La sua applicazione può essere trovata in svariati ambiti: scientifico, industriale, sociale, medico, ambientale. Gli esempi proposti nel Capitolo 2 sono inerenti l'ambito ambientale e industriale.

Capitolo 1

Modelli di regressione per risposta in (0,1)

In questo capitolo, dopo aver richiamato alcune nozioni relative ai modelli lineari normali e ai modelli lineari generalizzati, vengono presentati i principali modelli di regressione per risposta in (0,1) proposti in letteratura. Si richiamano inoltre alcune nozioni base dell'inferenza statistica, per fissare la notazione.

1.1 Richiami sul modello lineare normale

Il modello lineare normale (MLN) è spesso utilizzato per la semplicità della sua specificazione e perché comporta un notevole risparmio computazionale rispetto a modelli più complessi. Di contro, presenta una forte rigidità rispetto alle molteplici esigenze della modellazione statistica.

Il modello lineare spiega la variabile risposta continua tramite una funzione lineare nei parametri, delle variabili esplicative, a cui si aggiunge il termine di errore casuale con distribuzione normale.

Il modello lineare classico con errori normali è specificato dalle seguenti ipotesi:

1. $Y_1, ..., Y_n$ variabili casuali univariate indipendenti,

2. $E[Y_i] = \mu_i = \beta^T x_i = \beta_1 x_{i1} + ... + \beta_k x_{ik}$,

3. $Y_i \sim N(\mu_i, \sigma^2)$,

dove Y_i è la variabile casuale che descrive la risposta per l'i-esima osservazione, $\beta = (\beta_1, ..., \beta_k)^T$ è il vettore dei coefficienti di regressione, $x_i = (x_{i1}, ..., x_{ik})$ è l'i-esima riga della matrice delle variabili esplicative, infine μ_i è il previsore lineare.

La relazione fra la variabile risposta e k variabili esplicative è

$$Y_i = \mu_i + \varepsilon_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \ ,$$

nella quale $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ è il vettore contenente la parte stocastica di errore. La formulazione degli errori è $\varepsilon_i = Y_i - \mu_i$.

Equivalentemente, si può esplicitare Y in forma matriciale

$$Y = X\beta + \varepsilon \ .$$

Si distinguono dunque due parti che formano i valori osservati y della variabile risposta Y: la prima $X\beta$, detta componente sistematica, o anche parte deterministica del modello di regressione, che dipende dai valori assunti dalle variabili esplicative $(x_1, ..., x_k)$, la seconda è la componente aleatoria o erratica, detta anche parte stocastica del modello di regressione.

La funzione del modello dunque è

$$p_Y(y; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 \right] \ .$$

Si ottiene allora la funzione di logverosimiglianza

$$\ell(\mu, \sigma) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - \mu_i]^2 \ . \tag{1.1}$$

La stima di massima verosimiglianza del vettore β si ottiene dalle equazioni normali, si veda Pace e Salvan (2001, § 9.4),

$$X^T X\hat{\beta} = X^T y \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y \ . \tag{1.2}$$

Dalla precedente relazione (1.2), si ha

$$\hat{\mu} = X\hat{\beta} = X(X^T X)^{-1} X^T y \ ,$$

dove $P = X(X^T X)^{-1} X^T$ è detta matrice di proiezione. Ottenuta la stima $\hat{\mu}_i$ di μ_i, gli errori sono stimabili tramite i residui $\hat{\varepsilon}_i = y_i - \hat{\mu}_i$.

La stima di massima verosimiglianza di σ^2 è

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

Mentre $\hat{\mu}$ è uno stimatore non distorto con varianza minima (teorema di Gauss-Markov), così non è per $\hat{\sigma}^2$. Per correggere $\hat{\sigma}^2$ si usa

$$S^2 = \frac{n}{n-k} \hat{\sigma}^2 \ .$$

Sotto ipotesi di normalità si ha,

$$\hat{\beta} \sim N_k(\beta, \sigma^2 (X^T X)^{-1}) \quad \text{e} \quad \frac{(\mathrm{n-k})\mathrm{S}^2}{\sigma^2} \sim \chi^2_{\mathrm{n-k}} \ ,$$

con $\hat{\beta}$ e S^2 stocasticamente indipendenti. Questo è un risultato molto importante per l'inferenza (stima intervallare e verifica di ipotesi) sui parametri del modello di regressione.

Il modello completamente specificato stima tutti i coefficienti β. Alcune variabili esplicative possono non essere influenti per la variabile risposta; in questi casi, per semplificare il modello, si procede all'eliminazione del regressore. Uno strumento utile per la valutazione dei coefficienti di regressione ininfluenti è il test del rapporto di verosimiglianza, interpretabile in termini di scomposizione della devianza. Si veda Pace e Salvan (2001, § 9.4).

Viste le assunzioni molto forti a cui sono soggetti i modelli di regressione lineari il loro utilizzo a volte non risulta possibile, infatti:

- può succedere che la relazione che lega la variabile risposta alle variabili esplicative non sia di tipo lineare nei parametri.

- L'insieme dei valori possibili per la realizzazione di Y_i può essere diverso dalla retta reale.

- La varianza del termine di errore, e quindi anche della Y , può non essere costante (eteroschedasticità).

- A volte si ha a che fare con distribuzioni di Y lontane dalla normale, inoltre possono essere discrete.

In particolare il modello lineare normale non sembra poter essere efficacemente utilizzato, per l'analisi di interesse in questo lavoro.

1.2 Richiami sui modelli lineari generalizzati

I modelli di regressione lineari generalizzati (MLG), sono stati introdotti per superare alcune delle carenze dei modelli lineari normali. In particolare questo approccio è conveniente se il supporto della variabile risposta non coincide con $\mathbb{R}$.

Nella definizione del modello lineare generalizzato, la prima ipotesi del modello lineare classico resta invariata e si generalizzano la seconda e la terza ipotesi, con le rispettive seguenti, dove $g(\cdot)$ è una funzione invertibile nota, detta funzione di legame:

- $g(E[Y_i]) = g(\mu_i) = \beta^T x_i$,

- $Y_i \sim DE_1(\mu_i, \sigma^2 V(\mu_i))$.

Il simbolo $DE_1(\mu, \sigma^2 V(\mu))$ indica una famiglia di dispersione esponenziale di ordine 1, con valore atteso μ e varianza $\sigma^2 V(\mu)$, dove $V(\mu)$ è la funzione di varianza. Per alcuni richiami su famiglie esponenziali e famiglie di dispersione esponenziali si rinvia all'Appendice A.1.

Gli aspetti di generalizzazione riguardano dunque la distribuzione e la possibilità

che la relazione lineare nei parametri con le variabili concomitanti non coinvolga direttamente la media della risposta, ma una funzione della media. I più comuni MLG prevedono per la risposta distribuzioni quali la binomiale, Poisson, gamma, normale e altre.

Le famiglie di dispersione esponenziale costituiscono un ampliamento delle famiglie esponenziali ottenuto con l'introduzione di un parametro addizionale. Mentre la funzione del valore atteso resta invariata dalla rispetto alla famiglia esponenziale corrispondente. Il vantaggio di questo ampliamento è nella matrice di covarianza che viene modificata da un fattore di scala o parametro di precisione, così da superare un aspetto di rigidità modellistica delle famiglie esponenziali, Pace e Salvan (1996, Capitolo 6).

La varianza di Y_i, supposto che si distribuisca secondo una $DE_1(\mu_i, \sigma^2 V(\mu_i))$, risulta

$$Var(Y_i) = \sigma^2 V(\mu_i) = \sigma^2 V(g^{-1}(\beta^T x_i)) \ .$$

Quindi, nell'ambito dei modelli lineari generalizzati, vengono permesse certe forme di etroschedasticità. Non viene invece indebolita l'ipotesi di indipendenza delle osservazioni. Resta comunque il fatto che la varianza di Y_i dipende dalla combinazione lineare scelta per il valore atteso $E[Y_i]$.

Per ciascuna specificazione del modello di dispersione esponenziale, tra le possibili scelte della funzione legame $g(\cdot)$, è privilegiata

$$g(\mu) = \theta(\mu) \ ,$$

secondo la quale il parametro canonico θ, si veda Appendice A.1, risulta espresso come combinazione lineare delle variabili esplicative X con i coefficienti β, $\theta_i = \beta^T x_i$. Questa funzione legame è detta funzione di legame canonica.

In generale se $g(\mu_i) = \beta^T x_i$, allora $\mu_i = g^{-1}(\beta^T x_i)$ e pertanto l'i-esima componente di θ è esprimibile con $\theta_i = \theta(\mu_i) = \theta(g^{-1}(\beta^T x_i))$. Visto che, $\theta'(\mu) = 1/V(\mu)$ e definito

$\eta_i = \beta^T x_i$, detto previsore lineare, allora

$$\frac{\partial \theta_i}{\partial \beta_r} = \frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_r} = \frac{1}{V(\mu_i)}\frac{1}{g'(\mu_i)}x_{ir} \ . \tag{1.3}$$

Le quantità riportate sono calcolate in $\mu_i = g^{-1}(\beta^T x_i)$. Se $g(\cdot)$ è la funzione legame canonica si ha $g(\mu) = \theta(\mu)$, di conseguenza $g'(\mu) = 1/V(\mu)$ da cui si ottiene

$$\frac{\partial \theta_i}{\partial \beta_r} = x_{ir} \ .$$

Quest'ultimo risultato è intuibile anche dal fatto che se $g(\cdot)$ è la funzione legame canonica allora $\theta_i = \beta^T x_i$.

Funzione di verosimiglianza

Siano $Y_1, \ldots, Y_n$ variabili casuali secondo le assunzioni dei modelli lineari generalizzati, allora $Y = [Y_1, \ldots, Y_n]$ ha densità

$$p_Y(y; \beta, \sigma^2) = exp\left\{ \frac{1}{\sigma^2} \sum_{i=1}^{n} [\theta(\mu_i) y_i - K(\theta(\mu_i))] \right\} \prod_{i=1}^{n} a\left(\frac{1}{\sigma^2}, y_i \right) \ , \tag{1.4}$$

con $\theta(\mu_i) = \theta(g^{-1}(\beta^T x_i))$ e $K(\cdot)$ funzione generatore dei cumulanti, si veda l'Appendice B.1. Se $g(\cdot)$ è la funzione legame canonica allora la (1.4) si semplifica nella forma:

$$p_Y(y; \beta, \sigma^2) = exp\left\{ \frac{1}{\sigma^2} \left[\beta^T \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} K(\beta^T x_i) \right] \right\} \prod_{i=1}^{n} a\left(\frac{1}{\sigma^2}, y_i \right) \ .$$

Dalla (1.4), espressa in forma generale, si ottiene la log-verosimiglianza

$$\ell(\beta, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[\theta(\mu_i) y_i - K(\theta(\mu_i)) \right] + \sum_{i=1}^{n} \log a\left(\frac{1}{\sigma^2}, y_i \right) \ . \tag{1.5}$$

Per semplificare la notazione d'ora in poi verrà posto: $\theta_i = \theta(\mu_i) = \theta(g^{-1}(\beta^T x_i))$. Il vettore *score* delle derivate rispetto al vettore β e al parametro scalare σ^2 della (1.5)

ha componenti:

$$\ell_r = \frac{\partial \ell(\beta, \sigma^2)}{\partial \beta_r} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[\frac{\partial \theta i}{\partial \beta_r} y_i - \frac{\partial K(\theta_i)}{\partial \beta_r} \right], \quad r = 1, \ldots, k \ , \tag{1.6}$$

$$\ell_{\sigma^2} = \frac{\partial \ell(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{1}{\sigma^4} \sum_{i=1}^{n} \left[\theta_i y_i - K(\theta_i) \right] - \frac{1}{\sigma^4} \sum_{i=1}^{n} \frac{a'(\frac{1}{\sigma^2}, y_i)}{a(\frac{1}{\sigma^2}, y_i)} \ , \tag{1.7}$$

Posto $\lambda = \frac{1}{\sigma^2}$, si ha

$$a'(\lambda, y_i) = \left. \frac{\partial a(\lambda, y_i)}{\partial \lambda} \right|_{\lambda = \frac{1}{\sigma^2}} .$$

Osservando che

$$\frac{\partial K(\theta_i)}{\partial \beta_r} = K'(\theta_i) \frac{\partial \theta_i}{\partial \beta_r} = \mu_i \frac{\partial \theta_i}{\partial \beta_r} \ ,$$

allora la (1.6) può essere riscritta nella forma

$$\ell_r = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \theta i}{\partial \beta_r} \ . \tag{1.8}$$

Per la struttura della (1.8), la stima di massima verosimiglianza per β per un valore fissato di σ^2, $\hat{\beta}_{\sigma^2}$, non dipende da σ^2 e coincide con la massima verosimiglianza non vincolata $\hat{\beta}$. Sfruttando quindi la (1.3), le equazioni di verosimiglianza per β sono:

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i) g'(\mu_i)} x_{ir} = 0 \ .$$

Se ci si riferisce al legame canonico si semplifica molto essendo $g'(\mu_i) = 1/V(\mu_i)$ risulterebbe: $\sum_{i=1}^{n} (y_i - \mu_i) x_{ir} = 0$.

Informazione osservata ed informazione attesa

Si riportano ora le espressioni per le matrici di informazioni osservata ed attesa per β. Derivando rispetto a β_s, dalla (1.8) si ottiene

$$j_{rs} = -\frac{\partial^2 \ell(\beta, \sigma^2)}{\partial \beta_r \partial \beta_s} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left[\frac{\partial \mu_i}{\partial \beta_s} \frac{\partial \theta_i}{\partial \beta_r} - (y_i - \mu_i) \frac{\partial^2 \theta_i}{\partial \beta_r \partial \beta_s} \right] \ . \tag{1.9}$$

Ricordando l'espressione (1.3), l'espressione (1.9) si può riscrivere come

$$j_{rs} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\frac{x_{is}x_{ir}}{(g'(\mu_i))^2 V(\mu_i)}\left(1-\frac{y_i-\mu_i}{V(\mu_i)}\right)\right] \ .$$

Mentre l'informazione attesa risulta

$$i_{rs} = E[j_{rs}] = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\frac{x_{is}x_{ir}}{(g'(\mu_i))^2 V(\mu_i)}\right] \ . \tag{1.10}$$

Indicata con X la matrice delle variabili esplicative, avente riga i-esima x_i e posto $W = diag(w_i)$, con $w_i = 1/[(g'(\mu_i))^2 V(\mu_i)]$, la (1.10) si può riscrivere in forma matriciale

$$i_{\beta\beta} = \frac{1}{\sigma^2}X^T W X \ .$$

Per n sufficientemente elevato, il risultato generale di normalità asintotica dello stimatore di massima verosimiglianza fornisce l'approssimazione

$$\hat{\beta} \dot{\sim} N_k\left(\beta, \sigma^2(X^T W X)^{-1}\right) \ . \tag{1.11}$$

Una stima consistente per la matrice di covarianza di β con σ^2 noto è $\sigma^2(X^T\hat{W}X)^{-1}$, dove $\hat{W}$ indica la matrice W calcolata per $\beta = \hat{\beta}$. Se σ^2 è ignoto sarà utilizzata la sua stima di massima verosimiglianza.

Funzioni legame

Per la situazione in esame in questo lavoro è molto interessante porre l'attenzione su particolari funzioni legame. Ne vengono qui proposte alcune:

- **logistica**: $g(E[Y_i]) = g(\mu_i) = log(\frac{\mu_i}{1-\mu_i})$ con inversa $\hat{\mu}_i = \frac{exp(\beta^T x_i)}{1+exp(\beta^T x_i)}$,
- **log-log complementare**: $g(\mu_i) = \log(-\log(1-\mu_i))$ con inversa $\hat{\mu}_i = 1 - exp(-exp(\beta^T x_i))$,
- **log-log**: $g(\mu_i) = -\log(-\log(\mu_i))$ con inversa $\hat{\mu}_i = exp(-exp(-\beta^T X))$.

Riguardo la funzione di legame detta *logit* o logistica, questa risulta molto conveniente perché per $0 < \mu_i < 1$ si ha che $-\infty < logit(\mu_i) < +\infty$,

$$\lim_{\mu_i \to 0} \left[log(\frac{\mu_i}{1-\mu_i}) \right] = -\infty \;\; e \;\; \lim_{\mu_i \to 1} \left[log(\frac{\mu_i}{1-\mu_i}) \right] = +\infty \;\; ,$$

il codominio di $logit(\mu_i)$ è dunque $\mathbb{R}$ e non più l'intervallo (0,1), infatti $g : [0,1] \to \mathbb{R}$, è funzione monotona crescente.

Riguardo alle altre funzioni legame, con $0 < \mu_i < 1$, tutte hanno come codominio $\mathbb{R}$ e sono monotone crescenti.

Il caso della distribuzione del simplesso, che è un esempio di come i MLG possano essere applicati con successo a variabili casuali con risposte in (0,1), verrà esposto nel § 1.5.

1.3 Modello normale logistico

Una pratica frequentemente utilizzata in ambiti di studio in cui la variabile dipendente ha valori ristretti nell'intervallo (0,1) è di trasformare, con un'opportuna funzione $h(\cdot)$, la stessa variabile risposta, in modo da ottenere una nuova risposta trasformata per la quale si assume un modello lineare normale. Il modello assunto per Y è detto normale logistico. La regressione logistica ha una lunga storia in economia e argomenti ad essa correlati.

Usando questo approccio si assume che

$$h(Y_i) = \log\left(\frac{Y_i}{1-Y_i}\right) = \beta^T x_i + \varepsilon_i \;\; .$$

Dove $h(\cdot) = \log[(Y)/(1-Y)] = Z$ è la trasformazione logistica della variabile risposta. Se Y_i segue la distribuzione normale logistica allora ε_i avrà distribuzione $N(0,\sigma^2)$ e $Z_i = \log[(Y_i)/(1-Y_i)]$ distribuzione $N(\mu,\sigma^2)$. L'approccio standard prevede di confermare che la variabile risposta sia distribuita in modo logistico verificando che la sua trasformazione logistica sia effettivamente una variabile casuale normale e che gli errori sono anch'essi normalmente distribuiti.

Questo approccio presenta due inconvenienti: il primo è la scelta fissa della funzione $h(\cdot)$, ce ne potrebbero essere altre migliori. Il secondo è che una tale trasformazione potrebbe non stabilizzare la varianza. La prima preoccupazione è attenuata da alcuni risultati riportati in Cox (1996), sulle diverse funzioni legame per questi dati. La seconda invece resta, anche perché in modelli alternativi (basati sulle distribuzioni beta e simplex) c'è la condizione che una tale trasformazione non stabilizzi la varianza, si veda Kieschnick e McCullough (2003, § 2.1). Nel Capitolo 2 verra utilizzato il software R con la funzione `lm`, che restituisce le stime dei coefficienti di regressione del modello con relativo standard error, valore del test 't' e significatività, in più il coefficiente R^2, che indica la bontà di adattamento del modello.

1.4 Modelli di regressione con risposta beta

La distribuzione beta, avendo supporto [0,1], e grazie alla sua flessibilità, è facilmente adattabile ad un insieme molto grande di fenomeni casuali con risposte in (0,1). Molti autori, ad esempio Ferrari e Cribari-Neto (2004), indicano la distribuzione beta come la migliore scelta per la descrizione di dati proporzionali. La densità della distribuzione beta è data da

$$p(y;p,q) = \frac{1}{B(p,q)} y^{p-1}(1-y)^{q-1} \quad , \quad 0<y<1 \ , \tag{1.12}$$

dove

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx, \quad p>0, \ q>0 \ .$$

Al variare di p e q, la distribuzione beta assume forme differenti: se $p=q$ la distribuzione è simmetrica, se in più $p=q=1$ la densità dà origine ad una distribuzione uniforme nell'intervallo [0,1].

La funzione beta è legata alla funzione gamma, si veda il l'Appendice A.2, tramite la seguente relazione,

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \ .$$

Se Y_i è una variabile aleatoria beta di parametri p e q, allora

$$E[Y_i] = \frac{p}{p+q_i} \ ,$$

$$Var[Y_i] = \frac{pq_i}{(p+q_i)^2(p+q_i+1)} \ .$$

Un approccio standard, Kieschnick e McCullough (2003, § 2.1), è quello di definire il valore atteso come funzione lineare delle variabili esplicative x_i, assumendo

$$E[Y_i] = \frac{p}{p+q_i} = \beta^T x_i \ ,$$

invertendo

$$q_i = \frac{p}{\beta^T x_i} - p \ . \tag{1.13}$$

Si procede quindi alla sostituzione nella (1.12) di (1.13), per ricavare la funzione di densità $p(y_i; p, q_i)$. In seguito si deriva la funzione di log-verosimiglianza per il modello di regressione beta utilizzando questa specifica.

Questo non risulta peraltro il metodo più adatto per utilizzare la regressione beta, in quanto richiede di porre un vincolo su β in modo che $E[Y_i]$ appartenga a (0,1). Un metodo migliore risulta la ricerca e l'applicazione di una funzione legame (link function) per il valore atteso, secondo un procedimento analogo a quello applicato nei modelli lineari generalizzati. In particolare, viene utilizzata la funzione legame *logit*, (cfr. §1.2). Si assume quindi $\log(\frac{\mu_i}{1-\mu_i}) = \beta^T x_i$ con $\mu_i = \frac{p}{p+q_i}$. Riproponendo l'approccio di prima cioè specificando q in base al predittore lineare $\eta_i = \beta^T x_i$ si ottiene

$$q_i = pe^{-\beta^T x_i} \ .$$

Nuovamente, si sostituisce l'espressione di q_i nella (1.12). Per ottenere la stima di β si procede con la massima verosimiglianza. Dato che la distribuzione beta è un membro della famiglia esponenziale, queste stime di massima verosimiglianza hanno tutte le proprietà statistiche stabilite per gli stimatori di questa classe di distribuzioni.

Regressione beta con parametrizzazione alternativa

Una variante è suggerita in Ferrari e Cribari-Neto (2004). Consiste nell'utilizzare una riparametrizzazione della distribuzione beta, in modo tale da ottenere un parametro di media e uno di dispersione:

$$\mu = \frac{p}{p+q} \qquad \text{e} \qquad \phi = p+q \ .$$

Conseguentemente, $p = \mu\phi$ e $q = (1-\mu)\phi$. Con la riparametrizzazione in (1.4), si ha

$$E(Y_i) = \mu_i \qquad \text{e} \qquad Var[Y_i] = \frac{\mu_i(1-\mu_i)}{1+\phi} \ .$$

In questo modo il parametro μ_i viene visto come la media della risposta Y_i, ed il parametro ϕ può essere interpretato come un parametro di precisione nel senso che, per μ fissato, più aumenta ϕ più la varianza della risposta diminuisce.

Con questa nuova parametrizzazione la funzione di densità risulta

$$p(y_i;\mu_i,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma((1-\mu_i)\phi)} y_i^{\mu_i\phi-1}(1-y_i)^{1-\mu_i}\phi - 1 \ . \tag{1.14}$$

Se $\mu = 1/2$ allora la densità risulterà simmetrica e, come anticipato, al crescere di ϕ la densità sarà sempre più concentrata attorno al valore μ; il caso della distribuzione uniforme si ottiene con $\mu = 1/2$ e $\phi = 2$. Si veda Figura 1.1, per ogni grafico si mantiene fisso il parametro ϕ e si disegna la densità per 5 diversi valori di μ.

Siano $Y_1, ..., Y_n$ variabili casuali indipendenti dove ogni Y_i segue la distribuzione in (1.14), con media μ_i e parametro di dispersione ϕ. Il modello è definito assumendo

$$g(\mu_i) = \beta^T x_i = \eta_i \ .$$

Nella precedente espressione $\beta = (\beta_1, ..., \beta_k)$ è il vettore dei parametri ignoti di regressione, $(x_{i1}, ..., x_{ik})$ sono elementi dell'i-esima riga della matrice dei termini noti X. Infine $g(\cdot) : (0,1) \to \mathbb{R}$ è la funzione legame. Si nota che la varianza di Y_i è una

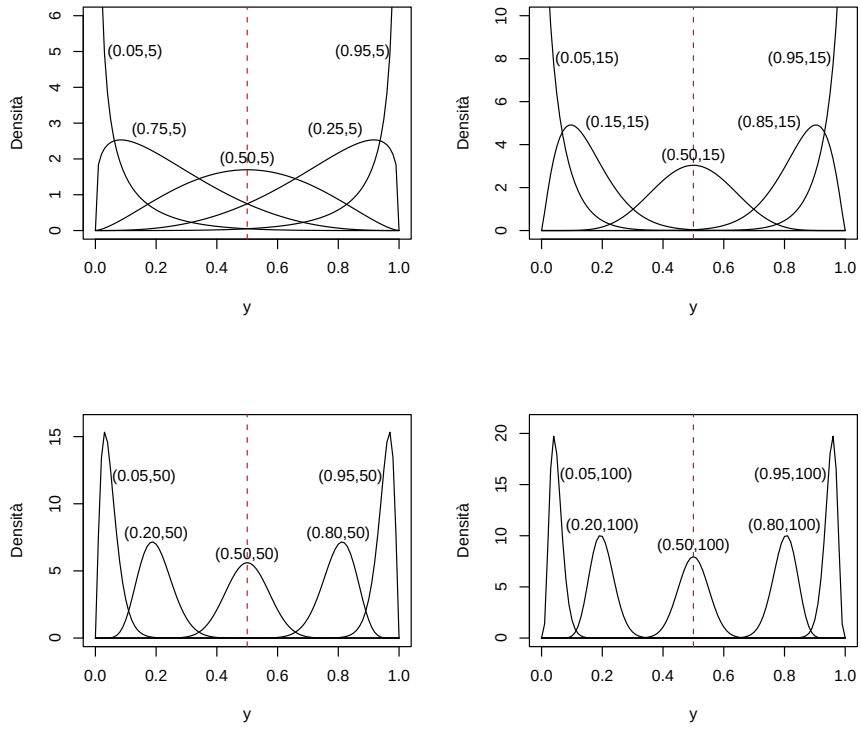

Figura 1.1: Densità beta, con parametri (μ, ϕ). Parametrizzazione in (1.14).

funzione di μ_i.

Ci sono molte scelte di diverse funzioni legame, (cfr. § 1.2). Qui verrà usata la funzione legame *logit* tale che

$$\mu_i = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \ ,$$

La funzione di logverosimiglianza risulta

$$\ell(\beta, \phi) = \sum_{i=1}^{n} \ell_i(\mu_i, \phi) \ , \tag{1.15}$$

dove

$$\ell_i(\mu_i, \phi) = \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) + (\mu_i\phi - 1)\log(y_i) + ((1-\mu_i)\phi - 1)\log(1-y_i) \ . \tag{1.16}$$

Ora si definisce la funzione di punteggio, o funzione score, data dal vettore delle derivare parziali prime della funzione di log-verosimiglianza. Dalla (1.15) si ottiene

$$\frac{\partial \ell(\beta,\phi)}{\partial \beta_r} = \sum_{i=1}^{n} (\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_r}) \ . \tag{1.17}$$

Si noti che $d\eta_i/d\mu_i = g'(\mu_i) \Rightarrow d\mu_i/d\eta_i = 1/g'(\mu_i)$. Viene ora introdotta $\psi(x)$, detta funzione digamma, definita come la derivata del logaritmo di una funzione gamma,

$$\psi(x) = \frac{d \log \Gamma(x)}{dx}, \quad x > 0 \ .$$

Dalla (1.16) si perviene alla forma

$$\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i} = \phi[\log(\frac{y_i}{1-y_i}) - (\psi(\mu_i\phi) - \psi((1-\mu_i)\phi))] \ . \tag{1.18}$$

Definendo come $\mu_i^* = [\psi(\mu_i\phi) - \psi((1-\mu_i)\phi)]$ e $y_i^* = \log(\frac{y_i}{1-y_i})$, si giunge all'espressione finale

$$\ell_r = \frac{\partial \ell(\beta,\phi)}{\partial \beta_r} = \phi \sum_{i=1}^{n} [(y_i^* - \mu_i^*) \frac{1}{g'(\mu_i)} x_{ir}] \ . \tag{1.19}$$

In forma matriciale

$$\ell_\beta = \left[\ell_r\right] = \frac{\partial \ell(\beta,\phi)}{\partial \beta} = \phi X^T T(y^* - \mu^*) \ .$$

Dove X^T è la matrice $k \times n$ dove la colonna i-esima sono le osservazioni rispetto alla variabile y_i, $T = Diag\{1/g'(\mu_1), ..., 1/g'(\mu_n)\}$, $y^* = \{y_1^*, ..., y_n^*\}$ e infine $\mu^* = \{\mu_1^*, ..., \mu_n^*\}$.

Si desidera ora ottenere la funzione score per ϕ. Dalla (1.15) otteniamo

$$\frac{\partial \ell(\beta,\phi)}{\partial \phi} = \sum_{i=1}^{n} (\frac{\partial \ell_i(\mu_i,\phi)}{\partial \phi}) \ .$$

Allora, nuovamente dall'espressione (1.16), si ottiene

$$\frac{\partial \ell_i(\mu_i,\phi)}{\partial \phi} = \mu_i[\log(\frac{y_i}{1-y_i}) - \psi(\mu_i\phi) + \psi((1-\mu_i)\phi)] + \log(1-y_i) - \psi((1-\mu_i)\phi) + \psi(\phi) \ .$$

Ricordando le notazioni per y^* e μ^* ,

$$\ell_\phi = \sum_{i=1}^{n} (\frac{\partial \ell_i(\mu_i, \phi)}{\partial \phi}) = \sum_{i=1}^{n} [\mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi((1-\mu_i)\phi) + \psi(\phi)] \ .$$

Ora si può scrivere il vettore delle $k+1$ derivate prime:

$$\nabla \ell(\beta, \phi) = \{\ell_{\beta_1}(\beta, \phi), ..., \ell_{\beta_k}(\beta, \phi), \ell_\phi(\beta, \phi)\} \ ,$$

nel quale

$$\ell_\beta(\beta, \phi) = \frac{\partial \ell(\beta, \phi)}{\partial \beta} = \phi X^T T(y^* - \mu^*) \ , \tag{1.20}$$

$$\ell_\phi(\beta, \phi) = \sum_{i=1}^{n} [\mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi((1-\mu_i)\phi) + \psi(\phi)] \ . \tag{1.21}$$

Il prossimo passo consiste nel calcolare l'informazione osservata definita come la matrice delle derivate seconde cambiate di segno. Quest'ultima ha la struttura:

$$\begin{bmatrix} \begin{pmatrix} \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_1 \partial \beta_k} \\ \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_2 \partial \beta_2} & \cdots & \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_2 \partial \beta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_k \partial \beta_1} & \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_k \partial \beta_2} & \cdots & \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_k \partial \beta_k} \end{pmatrix} & \begin{pmatrix} \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_1 \partial \phi} \\ \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_2 \partial \phi} \\ \vdots \\ \frac{\partial^2 \ell(\beta,\phi)}{\partial \beta_k \partial \phi} \end{pmatrix} \\ \begin{pmatrix} \frac{\partial^2 \ell(\beta,\phi)}{\partial \phi \partial \beta_1} & \frac{\partial^2 \ell(\beta,\phi)}{\partial \phi \partial \beta_2} & \cdots & \frac{\partial^2 \ell(\beta,\phi)}{\partial \phi \partial \beta_k} \end{pmatrix} & \begin{pmatrix} \frac{\partial^2 \ell(\beta,\phi)}{\partial \phi \partial \phi} \end{pmatrix} \end{bmatrix}$$

Dalla (1.17) otteniamo la seguente scrittura per la derivata in $\partial\beta_r \partial\beta_s$

$$\frac{\partial^2 \ell(\beta, \phi)}{\partial \beta_r \partial \beta_s} = \sum_{i=1}^{n} \left[\frac{\partial^2 \ell_i(\mu_i, \phi)}{\partial \mu_i^2} \frac{1}{g'(\mu_i)^2} x_{ir} x_{is} \right] \ . \tag{1.22}$$

Quindi, vista la (1.22), occorre derivare ulteriormente rispetto a μ_i la (1.18), ottenendo

$$\frac{\partial^2 \ell_i(\mu_i, \phi)}{\partial \mu_i^2} = -\phi[(\psi'(\mu_i \phi))\phi + \psi'((1-\mu_i)\phi)\phi] \ ,$$

posto $W^* = diag\{w_1^*, ..., w_n^*\}$ matrice diagonale con

$$w_i^* = \phi[(\psi'(\mu_i\phi)) + \psi'((1-\mu_i)\phi)]\frac{1}{g'(\mu_i)^2} \ ,$$

si giunge alla forma

$$j_{\beta\beta}(\beta,\phi) = \left[-\frac{\partial^2 \ell(\beta,\phi)}{\partial\beta_r \partial\beta_s} \right] = \phi \left[\sum_{i=1}^{n} w_i x_{ir} x_{is} \right] \ .$$

Questo è il blocco relativo a β dell'informazione osservata, dove con $[a_{rs}]$ si indica la matrice con generico elemento a_{rs}. In questo caso risulta essere uguale al blocco dell'informazione attesa, visto che si tratta di una quantità non stocastica. Viene qui espressa in forma matriciale,

$$i_{\beta\beta}(\beta,\phi) = E(-\frac{\partial \ell(\beta,\phi)}{\partial\beta_r \partial\beta_s}) = j_{\beta\beta}(\beta,\phi) = \phi X^T W X \ .$$

Dalla (1.19), si ottengono le derivate seconde miste

$$\frac{\partial^2 \ell(\beta,\phi)}{\partial\beta\partial\phi} = \sum_{i=1}^{n} \left[\left((y_i^* - \mu_i^*) - \phi \frac{(\partial\mu_i^*)}{\partial\phi} \right) \frac{1}{g'(\mu_i)} x_{ir} \right] \ .$$

Posto $c_i = \phi\frac{(\partial\mu_i^*)}{\partial\phi} = \phi[\psi(\mu_i\phi)\mu_i - \psi((1-\mu_i)\phi)(1-\mu_i)]$, l'elemento dell'informazione osservata risulta

$$j_{\beta_r\phi}(\beta,\phi) = \sum_{i=1}^{n} [(y_i^* - \mu_i^*) - c_i] \frac{1}{g'(\mu_i)} x_{ir} \ .$$

In questo caso, differentemente da prima, la quantità risulta essere stocastica. Siccome $E[y^*] = \mu^*$ allora l'elemento corrispondente dell'informazione attesa è

$$i_{\beta_r\phi}(\beta,\phi) = -\sum_{i=1}^{n} c_i \frac{1}{g'(\mu_i)} x_{ir} \ ,$$

in forma matriciale $i_{\beta\phi}(\beta,\phi) = -X^T T c$.

Si ricorda che le derivate seconde miste sono simmeriche perciò, $i_{\beta_r\phi} = i_{\phi\beta_r}$.

Infine, si valuta la derivata seconda pura in ϕ, derivando la (1.21)

$$\frac{\partial^2 \ell(\beta,\phi)}{\partial\phi\partial\phi} = \sum_{i=1}^{n} -[\psi'(\mu_i\phi)\mu_i^2 + (1-\mu_i)^2\psi'((1-\mu_i)\phi) - \psi'(\phi)] \ ,$$

posto ora $D = diag\{d_1,...,d_n\}$ con $d_i = \psi'(\mu_i\phi)\mu_i^2 + (1-\mu_i)^2\psi'((1-\mu_i)\phi) - \psi'(\phi)$, risulta

$$i_{\phi\phi}(\beta,\phi) = E(\sum_{i=1}^{n} d_i) = tr(D) \ . \tag{1.23}$$

Si ottiene allora la matrice completa di informazione attesa:

$$i_{\beta\phi} = \begin{pmatrix} \phi X^T W X & X^T T c \\ \\ (X^T T c)^T & tr(D) \end{pmatrix} \ .$$

Si noti che i parametri β e ϕ non sono ortogonali, questo è in contrasto con quanto risulta per i modelli lineari generalizzati.

L'informazione attesa è importante per esprimere la distribuzione asintotica dello stimatore di massima verosimiglianza perché la sua inversa, i^{-1}, fornisce una stima della matrice di covarianza di $(\hat\beta,\hat\phi)$ sotto (β,ϕ). Risulta noto che, sotto condizioni di regolarità, quando il campione è abbastanza numeroso,

$$\begin{pmatrix} \hat\beta \\ \hat\phi \end{pmatrix} \sim N_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, i^{-1}_{\beta\phi}\right) \ ,$$

nella quale $\hat\beta$ e $\hat\phi$ sono gli stimatori di massima verosimiglianza.

Da qui si possono ottenere gli intervalli di confidenza per i parametri, come mostrato in Raydonal, Cribari-Neto e Vasconcellos (2006, Capitolo 2). Posto $0 < \alpha < 1/2$, $z_{1-\alpha/2}$ quantile $1-\frac{\alpha}{2}$ di una distribuzione $N(0,1)$ e $i^{rr}(\hat\theta)$ come l'r-esimo elemento

della diagonale della matrice i^{-1}, $r = (1,\ldots,k+1)$.

$$\left[\hat{\beta}_r - z_{1-\alpha/2}(i(\hat{\theta})^{rr})^{1/2}, \hat{\beta}_r + z_{1-\alpha/2}(i(\hat{\theta})^{rr})^{1/2}\right] \quad \text{e}$$

$$\left[\hat{\phi} - z_{1-\alpha/2}(i(\hat{\theta})^{(k+1)(k+1)})^{1/2}, \hat{\phi} + z_{1-\alpha/2}(i(\hat{\theta})^{(k+1)(k+1)})^{1/2}\right]$$

sono gli intervalli di confidenza asintotici per β_r e ϕ, con copertura nominale $(1-\alpha)$.

L'espressione di i^{-1} si ottiene con la formula per le matrici partizionate (cfr. Appendice A.3),

$$i^{-1} = i^{-1}\begin{pmatrix}\beta\\ \phi\end{pmatrix} = \begin{pmatrix} i^{\beta\beta} & i^{\beta\phi} \\ \\ i^{\phi\beta} & i^{\phi\phi}\end{pmatrix} \ ,$$

dove

$$i^{\beta\beta} = \tfrac{1}{\phi}(X^T W X)^{-1}\left(I_k + \frac{X^T T c c^T T^T X (X^T W X)^{-1}}{\gamma\phi}\right) \ ,$$

$$i^{\beta\phi} = (i^{\phi\beta})^T = -\tfrac{1}{\gamma\phi}(X^T W X)^{-1} X^T T c \ ,$$

$$i^{\phi\phi} = \gamma^{-1} \ ,$$

con I_k matrice identità (kxk) e

$$\gamma = tr(D) - \phi^{-1} c^T T^T X (X^T W X)^{-1} X^T T c \ .$$

Le stime di β, ϕ, si ottengono dalle equazioni di verosimiglianza. Non essendo risolvibili esplicitamente, si deve ricorrere a metodi numerici, quali il metodo di Newton-Raphson. Nel Capitolo 2 verrà utilizzato il software R con la funzione `betareg` , che restituisce le stime dei coefficienti di regressione del modello con relativo standard error, significatività, in più un coefficiente R^2_p, detto *pseudo* R^2. La funzione è contenuta nella libreria `betareg` sviluppata da Raydonal, Cribari-Neto e Vasconcellos (2006) richiede come argomenti la variabile risposta, le variabili esplicative, la funzione legame e il dataset da cui ottenere i dati.

Misure di diagnostica

Dopo la specificazione del modello è importante ottenere delle misure che permettano di provare la bontà del modello.

Una misura generale dell'adattamento del modello si ottiene con il coefficiente R_p^2, si veda §1.4, che ha la stessa interpretazione di R^2 nei modelli di regressione lineare.

Un altro modo per verificare il modello è calcolare i residui standardizzati:

$$res_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{var}(y_i)}} \ ,$$

dove $\hat{\mu}_i = g^{-1}(\hat{\beta}^T x_i)$ e $\hat{var} = (\hat{\mu}_i(1-\hat{\mu}_i))/(1+\hat{\theta})$. Il diagramma di dispersione (i, res_i) con $i = \{1,\ldots,n\}$, non deve evidenziare andamenti visibilmente sistematici. Allo stesso modo nel diagramma di dispersione $(\hat{\eta}_i, res_i)$, $\hat{\eta}_i = \beta^T x_i$, trend evidenti potrebbero suggerire l'errata scelta della funzione legame.

Altri residui che possono essere utili allo scopo sono i residui di devianza, definiti da:

$$r_i^d = sign(y_i - \hat{\mu}_i)\sqrt{\{2(\ell_i(\tilde{\mu}_i, \hat{\theta}) - \ell(\hat{\mu}_i, \hat{\theta}))\}} \ ,$$

dove $\tilde{\mu}_i$ è il valore di μ_i che risolve $\partial \ell_i / \partial \mu_i = 0$. Anche questi residui vengono analizzati con strumenti grafici tramite il diagramma di dispersione, ci si aspetta, come nel caso dei residui standardizzati, una 'nuvola' casuale di punti.

Come ultima misura di diagnostica viene proposta la distanza di Cook. La distanza di Cook misura l'influenza di un singolo caso sulla stima dei coefficienti di regressione, quando il singolo caso viene rimosso dal processo di stima. Un valore della distanza di Cook > 1 indica che il punto è influente. Tale distanza è definita come

$$Cook_i = \frac{1}{k}\left(\hat{\beta} - \hat{\beta}_{\bar{i}}\right) X^T W X \left(\hat{\beta} - \hat{\beta}_{\bar{i}}\right) \ .$$

Per ovviare al fatto di dover calcolare n stime in più per il vettore β, in quanto $\hat{\beta}_{\bar{i}}$ in questo caso indica la stima del vettore dei coefficienti di regressione tolta l'i-esima

osservazione, si utilizza la seguente approssimazione:

$$C_i = \frac{h_{ii} res_i^2}{k(1-h_{ii})^2} \ ,$$

dove h_{ii} è l'i-esimo elemento diagonale della matrice $I_n - X(X^TX)^{-1}X^T$. Per eseguire la verifica si osserva il diagramma di dispersione (i, C_i). Per approfondimenti si veda Ferrari e Cribari-Neto (2004).

1.5 Distribuzione sul simplesso

In questo paragrafo verrà descritto il modello di regressione parametrico basato sulla distribuzione *simplex* (Kieschnick e McCullough, 2003), sviluppata da Barndorff-Nielsen e Jørgensen negli anni '90 appositamente per dati misurati sull'intervallo (0,1). Viene reso noto che, consultando Song (2006, § 2.6), l'articolo Kieschnick e McCullough (2003, § 2.1.6) presenta una dimenticanza nella formulazione della densità della distribuzione *simplex*. Questo modello è particolarmente utile perché, a differenza del modello beta, la distribuzione sul simplesso risulta una famiglia di dispersione esponenziale. Di conseguenza, tutti i risultati riguardo la teoria dei modelli generalizzati possono essere applicati ad un modello di regressione basato su tale distribuzione.

Seguendo Jørgensen (1997), si definisce la distribuzione *simplex* come

$$p(y;\mu,\sigma^2) = \left[\frac{1}{\sqrt{2\pi\sigma^2(y(1-y))^3}}\right] exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\} \ ,$$

per $0 < y < 1$, dove

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}$$

è la devianza unitaria, $0 < \mu < 1$. La funzione legame $logit$, come per la distribuzione beta, è quella che più si adatta alla situazione.

La derivata prima della log-verosimiglianza è

$$\frac{\partial \ell(\mu,\sigma^2)}{\partial \beta_r} = \sum_{i=1}^{n}\left[\frac{\partial}{\partial \mu_i}\log(a(y_i;\sigma^2)) - \frac{1}{2\sigma^2}\frac{(y_i-\mu_i)^2}{y_i(1-y_i)\mu_i^2(1-\mu_i)^2}\right]\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_r} \ , \qquad (1.24)$$

con $a(y_i;\sigma) = 1/\sqrt{2\pi\sigma^2(y(1-y))^3}$ funzione dipendente dal solo parametro σ. Tralasciando il segno di sommatoria e di conseguenza anche l'indice i, la derivata necessaria per calcolare la (1.24) risulta,

$$\Rightarrow \frac{\partial}{\partial\mu}\Big[-\frac{1}{2}\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}\Big] = \frac{y-\mu}{\mu(1-\mu)}\Big[d(y;\mu)+\frac{1}{\mu(1-\mu)}\Big] = \xi(y;\mu) \quad . \quad (1.25)$$

Perciò, posta $\xi(y;\mu)$ come la derivata rispetto a μ, si perviene alla forma finale di ℓ_r

$$\ell_r = \frac{1}{\sigma^2}\sum_{i=1}^{n}\xi(y_i;\mu_i)\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_r} = \frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ir}\Big(\frac{1}{g'(\mu_i)}\Big)\xi(y_i;\mu_i) \quad .$$

Come si può notare la forma di questa espressione non è lineare, quindi, per una soluzione dell'equazione $\ell_r(\mu,\sigma^2) = 0$, bisogna procedere con metodi numerici quali l'algoritmo di Newton-Raphson.

Si desidera ora ottenere l'espressione dell'informazione attesa per β. Per l'ortogonalità di β e σ si procede al calcolo della derivata seconda rispetto a β come se σ fosse noto. Derivando quindi rispetto a μ la relazione (1.25) si ottiene

$$\frac{\partial\xi(y,\mu)}{\partial\mu} = \xi'(y;\mu) =$$

$$-\Big\{\frac{1}{\mu(1-\mu)}d(y;\mu)+\frac{1-2\mu}{\mu^2(1-\mu)^2}(y-\mu)d(y;\mu)-\frac{y-\mu}{\mu(1-\mu)}d'(y;\mu)+\frac{1}{\mu^3(1-\mu)^3}-\frac{3-6\mu}{\mu^4(1-\mu)^4}(y-\mu)\Big\} \quad ,$$

quindi

$$j_{\beta\beta} = \left[\frac{\partial\ell(\mu\sigma^2)}{\partial\beta_r\partial\beta_s}\right] = \left[\frac{1}{\sigma^2}\sum_{i=1}^{n}\xi'(y_i;\mu_i)\frac{1}{(g'(\mu_i))^2}x_{ir}x_{is}\right] \quad ,$$

è il blocco dell'informazione osservata relativo a β. Per ottenere l'informazione attesa per β si calcola il valore atteso, in questo caso il risultato sarà diverso dall'informazione osservata visto che $\xi'(y;\mu)$ dipende anche da y. Il valore atteso è

$$E\Big[\xi'(Y;\mu)\Big] = \frac{3\sigma^2}{\mu(1-\mu)}+\frac{1}{\mu^3(1-\mu)^3} \quad ,$$

visto che, si veda Song (2006, § 2.6):

- $E[Y-\mu] = 0$,

- $E[d(Y;\mu)] = \sigma^2$,
- $E[(Y-\mu)d'(Y;\mu)] = -2\sigma^2$,
- $E[(Y-\mu)d(Y;\mu)] = 0$.

Allora l'informazione attesa risulta:

$$i_{\beta\beta} = \left[\frac{1}{\sigma^2}\sum_{i=1}^{n} E\left[\xi'(y_i;\mu_i)\right]\frac{1}{(g'(\mu_i))^2}x_{ir}x_{is}\right] = \left[\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[u_i x_{ir}x_{is}\right]\right] ,$$

con

$$u_i = \frac{3\sigma^2\mu_i^2(1-\mu_i)^2+1}{\mu_i^2(1-\mu_i)^2} ,$$

in forma matriciale

$$i_{\beta\beta} = \frac{1}{\sigma^2}X^T UX .$$

Richiamando la teoria dei modelli lineari generalizzati, (cfr. § 1.2, in particolare (1.11)), si può ottenere la distribuzione dello stimatore di massima verosimiglianza $\hat{\beta}$

$$\hat{\beta} \dot{\sim} N_k\left(\beta, \sigma^2(X^T UX)^{-1}\right) .$$

Attualmente non esistono funzioni R che implementino modelli basati sulla distribuzione del simplesso. In questo lavoro ci si limita limita quindi solo alla parte teorica. Lo sviluppo di una libreria R potrà essere oggetto di un lavoro futuro.

Capitolo 2

Analisi di insiemi di dati in ambiente R

In questo capitolo si prendono in considerazione due insiemi di dati che saranno trattati con le tecniche descritte nel Capitolo 1, escluso il modello basato sulla distribuzione del simplesso, allo scopo di confrontare i metodi disponibili.

2.1 Dati atmosferici rilevati presso stazione meteo

Insieme di dati

Viene qui considerato un dataset di 60 osservazioni tratte dalla banca dati del centro di rilevazione metereologica del rifugio la Guardia nelle Piccole Dolomiti, Arpav (2009). Si tratta di informazioni giornaliere dal 27/03/09 al 25/05/09. Si sono rilevate giornalmente: la temperatura media ($Tmedia$, °C), la temperatura minima ($Tmin$, °C), la temperatura massima ($Tmax$, °C), i millimetri di pioggia (*Pioggia*, mm), la radiazione solare ($Radiaz$, MJ/m^2), il vento ($Vento$, km/g), la direzione prevalente del vento (*Direz*, punti cardinali). L'intenzione è di modellare l'andamento della percentuale di umidità media dell'aria rispetto alle altre variabili rilevate. La variabile risposta è l'umidità (Umedia, %), definita dalla media dell'indice di umidità massimo e quello di umidità minima. Le variabili esplicative sono tutte quantitative tranne la direzione prevalente del vento che assume i quattro valori dei punti cardinali.

Struttura del dataset

Il dataset ha la seguente struttura:

Cont	Umedia	Tmedia	Tmin	Tmax	Pioggia	Radiaz	Vento	Direz
1	0.800	3.0	-0.5	5.5	0.0	13.161	98.9	S
2	0.990	3.2	2.4	4.0	36.4	0.948	136.7	E
3	0.945	3.9	2.4	4.8	64.2	1.418	174.8	N
4	0.870	4.4	2.4	7.5	15.4	6.979	285.4	N
5	0.820	5.6	3.3	9.2	6.4	10.958	108.9	S
6	0.815	6.9	3.2	10.2	9.4	4.688	90.4	O
7	0.915	5.5	4.4	7.4	50.6	1.469	194.5	E
8	0.795	8.0	4.5	11.8	0.0	15.596	120.1	N
9	0.785	8.4	6.3	11.8	2.4	4.563	139.0	N
...	...	...	...	...	...	...	...	...
57	0.645	17.5	13.9	21.5	0.0	27.325	134.9	S
58	0.815	17.6	13.8	21.0	0.0	25.216	130.8	N
59	0.630	19.5	14.9	23.1	0.0	25.606	163.3	N
60	0.590	22.6	17.2	26.8	0.2	28.599	290.5	N

Tabella 2.1: Dataset misure atmosferiche.

Analisi dei dati, confronto tra modelli

Nella Tabella 2.2 vengono riportati i risultati delle analisi tramite modelli di regressione considerati nel Capitolo 1. Tutti i risultati sono conseguiti tramite l'utilizzo del software (freeware) R versione 2.6; vengono riportati per ogni coefficiente di regressione la stima e tra parentesi, il *p-value* di significatività dei coefficienti.

	Normale lineare	Normale logistico	Beta legame logit
Intercetta	1.05728	3.04717	2.80381
	($< 2e^{-16}$)	($5.62e^{-12}$)	(0)
Tmedia	0.02702	0.27561	0.22616
	(0.28923)	(0.143853)	(0.0778)
Tmin	0.01608	0.10455	0.06743
	(0.30146)	(0.360487)	(0.387)
Tmax	-0.04479	-0.39138	-0.30197
	(0.00126)	(0.000173)	(0.000012)
Pioggia	0.00107	0.01627	0.015691
	(0.04184)	($7.78e^{-05}$)	(0.000164)
Radiaz	-0.00217	-0.00072	-0.00894
	(0.34029)	(0.965335)	(0.433)
Vento	-0.000456	-0.00086	-0.00208
	(0.00766)	(0.478590)	(0.0156)
DirezN	0.00217	-0.07941	0.02479
	(0.94038)	(0.710793)	(0.887)
DirezO	0.01271	-0.01368	0.04943
	(0.80862)	(0.971720)	(0.860)
DirezS	0.00621	-0.04958	0.02890
	(0.86514)	(0.853699)	(0.888)
Valore R^2	0.7717	0.8285	0.81619

Tabella 2.2: Risultati per i modelli di regressione proposti.

Si nota che alcune delle variabili prese in esame sono ininfluenti. Per i modelli normale e beta si può provare a togliere la direzione del vento la radiazione al suolo e la temperatura minima. I modelli ridotti danno i seguenti risultati:

	Normale lineare	Beta legame logit
Intercetta	1.069607	2.828119
	$(< 2e^{-16})$	(0)
Tmedia	0.052565	0.325250
	$(8.49e^{-06})$	$(2.43e^{-08})$
Tmax	-0.057803	-0.348095
	$(5.71e^{-07})$	$(9.34e^{-10})$
Pioggia	0.001419	0.017858
	(0.002468)	$(8.37e^{-07})$
Vento	-0.000552	-0.002472
	(0.000417)	$(1.41e^{-03})$
Valore R^2	0.7528	0.8143
Valore AIC		-173.2934

Tabella 2.3: Risultati per i modelli lineare e beta ridotti.

Fonte di variabilità	Gradi di libertà	Somma dei quadrati	Test su miglioramento distribuzione nulla
Totale	60	35.3013	
Costante	1	34.4662	
Totale corretta	59	0.8350	
Miglioramento con, $\mathcal{F}_{k_0}$ rispetto a $\mathcal{F}_1$	4	0.6286	41.8787 $(4.4408e^{-16})$
Miglioramento con, $\mathcal{F}_k$ rispetto a $\mathcal{F}_{k_0}$	5	0.0157	0.8255 (0.5375)
Residui di $\mathcal{F}_k$	50	0.1906	

Tabella 2.4: Analisi della varianza, modello lineare.

Viene eseguito il test di analisi della varianza per confermare la riduzione eseguita, si veda Tabella 2.4. Il test per il modello normale esclude che i dati siano generati da un processo di campionamento casuale semplice. Mentre si accetta l'ipotesi H_0 per il modello ridotto $\mathcal{F}_{k_0}$. Per il modello beta esiste una funzione in R, sviluppata da Ferrari e Cribari-Neto (2004), `anova.betareg` , che esegue il test di analisi della varianza come log-rapporto di verosimiglianza dando come risultato 2.6276 che confrontato con la distribuzione χ^2_5, da come valore del test 0.7571, allora si accetta l'ipotesi nulla del modello ridotto. (Per consultare il prospetto teorico dell'analisi della varianza si veda la Tabella 1, nell'Indice dei Simboli).

Nel modello logistico c'è la possibilità di escludere anche la variabile *Vento*.

	Intercetta	Tmedia	Tmax	Pioggia	R^2	AIC
Normale	2.95165	0.46922	-0.4912	0.01553	0.8196	75.6929
logistico	$(3.39e^{-16})$	$(2.00e^{-08})$	$(1.09e^{-09})$	$(7.04e^{-08})$		

Tabella 2.5: Risultati per il modello logistico ridotto.

Fonte di variabilità	Gradi di libertà	Somma dei quadrati	Test su miglioramento distribuzione nulla
Totale	60	169.2598	
Costante	1	109.0913	
Totale corretta	59	60.1685	
Miglioramento con, $\mathcal{F}_{k_0}$ rispetto a $\mathcal{F}_1$	3	49.3129	84.7955 (0)
Miglioramento con, $\mathcal{F}_k$ rispetto a $\mathcal{F}_{k_0}$	6	0.5389	0.4353 (0.8517)
Residui di $\mathcal{F}_k$	50	10.3167	

Tabella 2.6: Analisi della varianza, modello logistico.

Anche nel modello logistico, si veda Figura 2.6, l'analisi della varianza suggerisce di rifiutare l'ipotesi che i dati siano generati da campionamento casuale semplice, mentre accetta, l'ipotesi H_0 corrispondente al modello ridotto $\mathcal{F}_{k_0}$.

Per operare una scelta tra i modelli logistico e quello basato sulla regressione beta si può utilizzare il criterio di Akaike. Questo criterio permette di confrontare due modelli anche non annidati, si sceglie il modello con indice minore.

Generalmente è espresso tramite la formula

$$AIC = -2\left[\log(\ell(\hat{\theta})) - k\right] \quad , \tag{2.1}$$

dove $\log(\ell(\hat{\theta}))$ è il logaritmo naturale del valore massimo della funzione di verosimiglianza del modello stimato, k è il numero dei coefficienti di regressione del modello.

Analisi dei residui Si valutano ora la normalità residui dei modelli normale e logistico, tramite strumenti grafici, si veda 2.1.

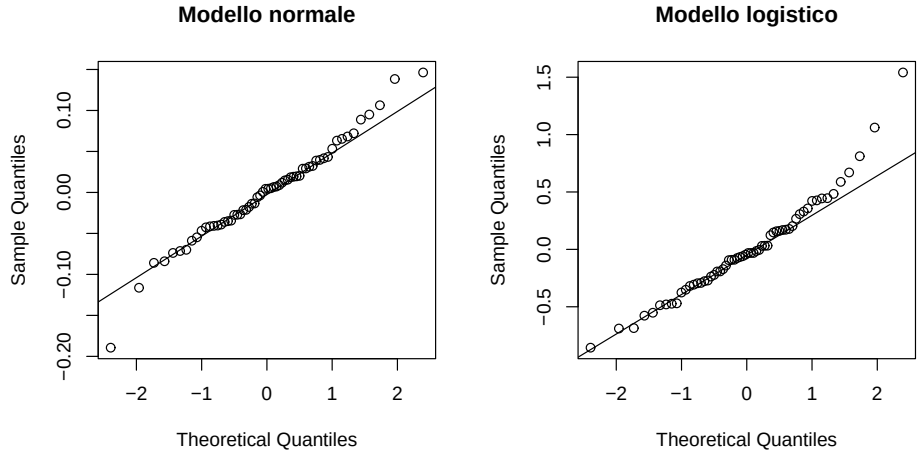

Figura 2.1: Grafici dei residui, per i modelli normale e logistico. Dati atmosferici.

I residui presentano leggeri scostamenti dalla normalità sulle due code per il modello lineare, mentre per il logistico c'è solo uno scostamento sulla coda desta ma più consistente.

La Figura 2.2 presenta le quattro misure di diagnostica descritte nel §1.4 per il modello basato sulla regressione beta. Anche l'analisi grafica dei residui del modello beta non presenta evidenti anomalie: non vengono visualizzati particolari trend o va-

lori *outlier*. Unico accorgimento è sulla distanza di Cook l'osservazione n. 28 sembra essere particolarmente influente.

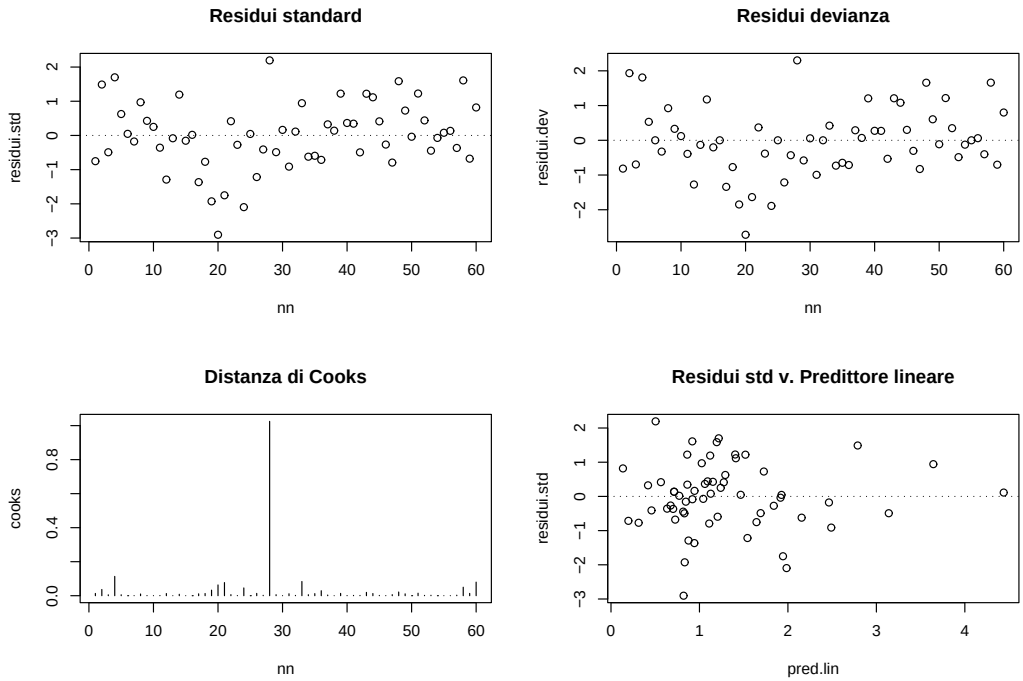

Figura 2.2: Grafici dei residui, per il modello beta. Dati atmosferici.

Confrontando i due modelli, logistico e regessione beta, con l'AIC, si vedano le Tabelle 2.5 e 2.3, si conviene che il modello basato sulla regressione beta sia migliore del logistico anche se quest'ultimo ha il vantaggio di una minore complessità computazionale.

2.2 Dati Prater, oli combustibili

Insieme di dati

Questo paragrafo propone un esempio tratto da Ferrari e Cribari-Neto (2004), si veda la Tabella C.2, in cui la risposta riguarda la percentuale di gasolio risultante dopo la raffinazione del greggio. Le possibili variabili esplicative sono: peso specifico del greggio (SG, grado API), la pressione del vapore del greggio (VP, lbf/in^2), i punti 10% ASTIM del greggio ($V10$, la temperatura in cui il 10% del greggio evapora), la temperatura di evaporazione del greggio (EP, °F). Il dataset contiene 32 osservazioni senza dati mancanti. Le tre variabili SG, VP e $V10$ sono variabili di esplorazione e corrispondono a dieci differenti tipi di greggio e sono soggette a condizioni sperimentali di raffinazione controllate. Questo database è stato analizzato da Atkinson (1985), usando un modello di regressione lineare normale.

Preparazione del dataset:

Le tre variabili SG, VP e $V10$ sono sintetizzate nel fattore No con 10 livelli, quindi la tabella verrà semplificata mantenendo solo la variabile No. Verrà modificato anche il supporto della variabile risposta Y portandolo all'intervallo (0,1). Ecco come si presenta il dataset pronto per l'analisi:

	No	EP	Y
1	A	205	0.122
2	A	275	0.223
...	...	...	...
32	J	428	0.180

Tabella 2.7: Dataset misurato da Prater (1956).

	Normale lineare	Normale logistico	Beta legame logit
Intercetta	-0.2016	-4.58017	-4.4318
	($1.62e^{-09}$)	(0)	(0)
TipoB	-0.08275	-0.44392	-0.4051
	($1.10e^{-05}$)	(0.00474)	($2.36e^{-05}$)
TipoC	-0.04767	-0.22593	-0.1554
	(0.00346)	(0.12373)	($9.40e^{-02}$)
TipoD	-0.1440	-0.73167	-0.6680
	($2.81e^{-07}$)	($1.35e^{-05}$)	($1.11e^{-15}$)
TipoE	-0.1103	-0.69103	-0.5939
	($2.07e^{-08}$)	($8.38e^{-05}$)	($4.60e^{-12}$)
TipoF	-0.1211	-0.75748	-0.6875
	($4.03e^{-07}$)	($2.51e^{-05}$)	($8.66e^{-15}$)
TipoG	-0.1751	-1.35296	-1.1840
	($1.30e^{-11}$)	($8.96e^{-10}$)	(0)
TipoH	-0.1949	-1.39659	-1.2318
	($1.15e^{-11}$)	($2.99e^{-09}$)	(0)
TipoI	-0.2274	-1.42191	-1.3419
	($8.49e^{-12}$)	($2.31e^{-08}$)	(0)
TipoJ	-0.2811	-1.83613	-1.7277
	($1.58e^{-14}$)	($3.49e^{-11}$)	(0)
temp	-0.001580	0.011602	0.01096
	(0)	($1.63e^{-15}$)	(0)
Valore R^2	0.9792	0.9638	0.9617
Valore AIC		-9.4026	-147.5951

Tabella 2.8: Risultati per i modelli di regressione proposti.

Analisi dei dati, confronto tra modelli

Nella tabella 2.8 vengono riportati i risultati ottenuti applicando i metodi di regressione proposti nel Capitolo a relativamente al dataset in Tabella 2.7. Tutti i risultati sono conseguiti tramite R versione 2.6.

I dati sembrano suggerire che il tipo di lavorazione sia in ogni caso molto importante per determinare la percentuale di gasolio. Anche la temperatura di evaporazione gioca un ruolo molto importante per spiegare il fenomeno.

Visto il valore molto alto di R^2, sembra che i modelli spieghino tutti in modo ottimale la variabile risposta.

Analisi dei residui Si valutano ora la normalità residui dei modelli, tramite strumenti grafici. Si veda 2.3, per il modello normale e per il modello logistico.

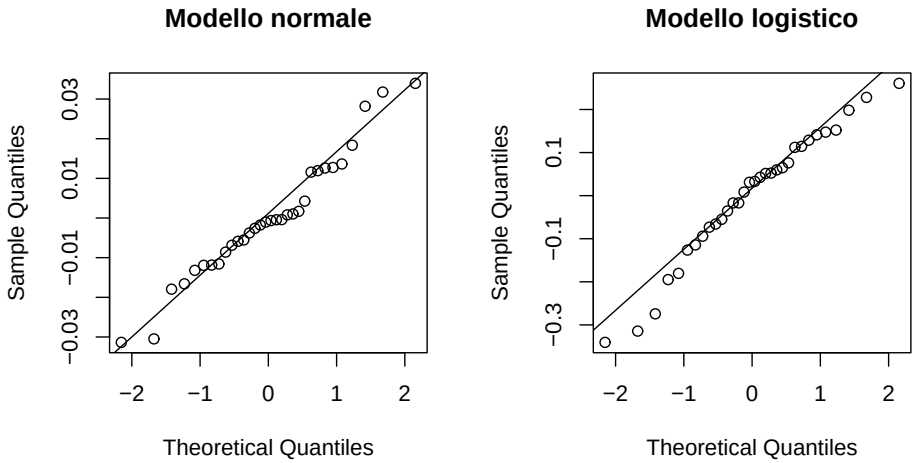

Figura 2.3: Grafici dei residui, per i modelli normale e logistico. Oli combustibili.

I grafici mostrano che i residui di entrambi i modelli non seguono esattamente la distribuzione normale, nel caso del modello normale la parte centrale è un po' scostata dalla retta, mentre per il modello logistico sono le code a generare qualche dubbio.

In Figura 2.4, si riportano i quattro grafici esposti nel § 1.4 per il modello basato sulla regressione beta. Anche l'analisi grafica dei residui del modello beta non presenta evidenti anomalie: non vengono visualizzati particolari trend o valori outlier. La distanza di Cook, questa volta, non evidenzia nessun valore particolarmente influente, infatti tutti i valori sono inferiori ad 1.

Anche con questo dataset, per il criterio di Akaike, si veda 2.1, il modello basato sulla regressione beta sembra preferibile, tenendo comunque presente la maggiore semplicità computazionale del modello logistico.

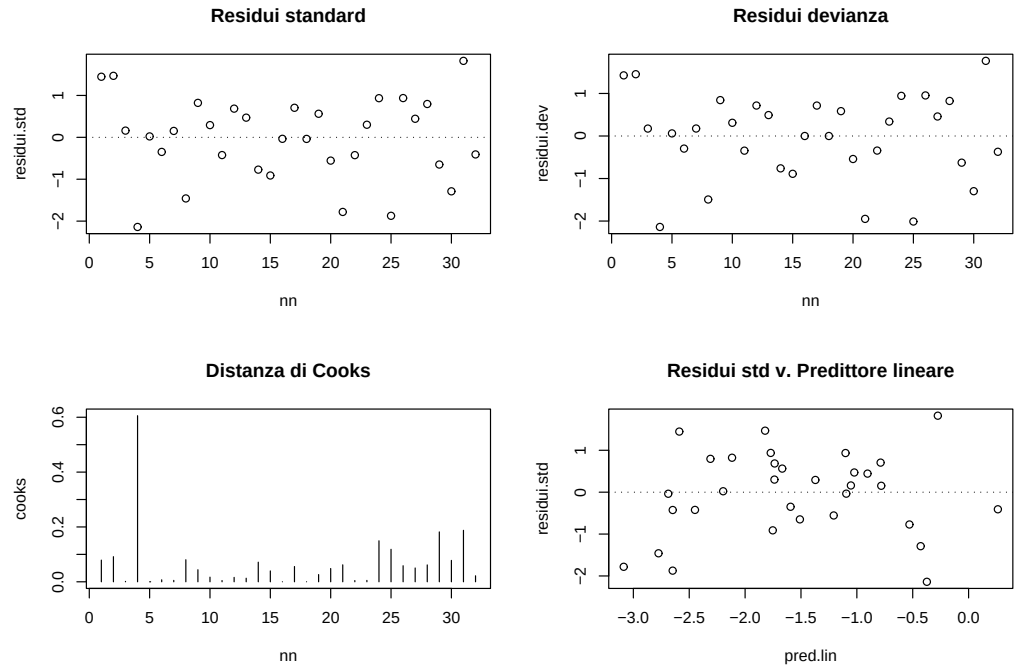

Figura 2.4: Grafici dei residui, per il modello beta. Oli combustibili.

2.3 Conclusioni

In questo capitolo sono state considerate delle applicazioni a due casi dei modelli esposti, nei paragrafi 1.1, 1.3 e 1.4 del precedente capitolo. Ne è emersa una inadattabilità del modello lineare semplice normale, come anticipato anche nella parte di teoria, mentre risultano quasi a 'pari merito' i modelli logistico e quello basato sulla distribuzione beta. Quest'ultimo è preferibile in entrambi i casi secondo criterio di Akaike.

Appendice A

Appendice A

A.1 Famiglie esponenziali e di dispersione esponenziali

Verranno brevemente riassunte delle caratteristiche essenziali delle famiglie esponenziali e di dispersione esponenziale introdotte da Jørgensen, per approfondimenti si veda Pace e Salvan (1996, Capitolo 5).

A.1.1 Famiglie esponenziali

Data $p_0(y)$, funzione di densità rispetto ad una misura dominante μ, per la variabile casuale Y con supporto $\mathcal{Y} \subseteq \mathbb{R}$, tramite l'ampliamento esponenziale si perviene ad una famiglia parametrica che includa $p_0(y)$ come caso particolare e i cui elementi abbiano il medesimo supporto. Le densità della famiglia esponenziale sono proporzionali a $\exp(\theta y)p_0(y)$. Allora la funzione generatrice dei momenti, è

$$M_0(\theta) = \int_{\mathcal{Y}} e^{\theta y} p_0(y) d\mu \ ,$$

se l'integrale esiste finito. Poiché $M_0(0) = 1$, l'insieme $\{\Theta = \theta \in \mathbb{R} : M_0(\theta) < +\infty\}$ non è vuoto.

Viste le assunzioni, allora

$$p(y,\theta) = \frac{e^{\theta y} p_0(y)}{M_0(\theta)} = exp\{\theta y - K(\theta)\} p_0(y) \ ,$$

è una funzione di densità con $K(\theta) = \log(M_0(\theta))$ funzione generatrice dei cumulanti, si noti che $p(y,0) = p_0(y)$.

Si definisce famiglia esponenziale di ordine 1 generata da $p_0(y)$ l'insieme con densità

$$\mathcal{F}_{en}^1 = \{p(y;\theta) = exp\{\theta y - K(\theta)\}p_0(y), \quad y \in \mathcal{Y}, \quad ,\theta \in \Theta\} \ . \tag{A.1}$$

Sono di famiglia $\mathcal{F}_{en}^1$ le distribuzioni: normale, Poisson, binomiale, esponenziale, gamma. Risulta molto usata in quanto:

- costituisce un serbatoio di modelli utili per le applicazioni;
- le procedure inferenziali basate sulla verosimiglianza sono generali, semplici e accurate;
- l'algoritmo numerico di stima è lo stesso per tutti i modelli della classe.

Si può dire quindi che una varabile Y è della famiglia esponenziale se la sua densità può essere scritta nella forma

$$exp(\theta y - K(\theta))p_0(y) \ .$$

Sia Y appartenente alla $\mathcal{F}_{en}^1$ definita in (A.1). Si dimostra che le funzioni generatrici dei momenti e dei cumulanti risultano

$$M_Y(t;\theta) \quad = exp\{K(\theta + t) - K(\theta)\} \ ,$$

$$K_Y(t;\theta) \quad = K(\theta + t) - K(\theta) \ .$$

Esistono momenti di ogni ordine che sono dati dall'espressione $\kappa_r = \frac{\partial^r}{\partial \theta^r} K(\theta)$. In particolare il valore atteso $E_\theta[Y]$ e la varianza $Var_\theta[Y]$ verranno indicati come: $E_\theta[Y] = \kappa_1(Y) = K'(\theta)$ e $Var_\theta[Y] = \kappa_2(Y) = K''(\theta)$.

Finora si è discusso il caso monoparametrico. Risulta comunque agevole l'estensione al caso multiparametrico. Si definisce $\mathcal{F}_{en}^{p}$ come famiglia esponenziale naturale di ordine p generata da $p_0(y)$, densità di Y, con supporto $\mathcal{Y} \subseteq \mathbb{R}^p$,

$$\mathcal{F}_{en}^{p} = \left\{ p(y;\theta) = exp\{\theta \cdot y - K(\theta)\} p_0(y), \quad y \in \mathcal{Y} \subseteq \mathbb{R}^p, \quad \theta \in \Theta \right\} \ , \qquad \text{(A.2)}$$

dove $\theta \cdot y$ è il prodotto scalare tra θ e y. Si possono estendere le $\mathcal{F}_{en}^{p}$ in modo tale da non vincolare le dimensioni dello spazio parametrico e dello spazio campionario a p, con queste estensioni si definiscono le famiglie esponenziali di ordine p, $\mathcal{F}_{e}^{p}$.

Per quanto riguarda la log-verosimiglianza si ottiene, a partire da una $\mathcal{F}_{en}^{p}$,

$$\ell(\theta, y) = \log(p_Y(y;\theta)) = (\theta \cdot y - K(\theta)) + \log(p_0(y)) = \theta \cdot y - K(\theta) + c(y) \ .$$

Il vettore score risulta,

$$\ell_*(\theta) = \frac{\partial \ell(\theta, y)}{\partial \theta} = y - \kappa_1(y) = y - E_\theta[Y] \ ,$$

ossia coincide con il vettore degli scarti di y dal proprio valore atteso. Infine L'informazione osservata $j(\theta)$:

$$j(\theta) = -\frac{\partial \ell_*(\theta)}{\partial \theta} = \kappa_2(y) = Var_\theta[Y] \ .$$

Osservando che l'informazione osservata $j(\theta)$ non dipende da y, allora si conclude che essa coincide con l'informazione attesa di Fisher $i(\theta)$.

La stima di massima verosimiglianza per θ, $\hat{\theta}$, se esiste è data dalla soluzione dell'equazione di verosimiglianza

$$y - E_\theta[Y] = 0 \ ,$$

tale soluzione è unica visto che $-j(\theta)$ è matrice definita positiva.

A.1.2 Famiglie di dispersione esponenziali

Una generalizzazione della classe delle famiglie esponenziali è data dalle famiglie di dispersione esponenziale. Esse prevedono oltre al parametro p-dimensionale θ anche un parametro scalare λ. Da questa classe ampliata nascono i modelli lineari generalizzati, usati nel § 1.2. Come avviene per le famiglie esponenziali anche la generazione astratta, per le famiglie di dispersione, avviene in modo univoco a partire da una funzione generatrice dei momenti assegnata.

Si presenta ora la densità di una variabile Y appartenente alla famiglia di dispersione esponenziale:

$$\mathcal{F}_{de}^{p} = \{p(y;\theta,\lambda) = a(\lambda,y)e^{\lambda(\theta\cdot y - K(\theta))}\} \ . \tag{A.3}$$

Con riferimento alla $\mathcal{F}_{de}^{p}$ in (A.3) si nomina: θ parametro naturale o canonico, $K(\cdot)$ generatore dei cumulanti, λ parametro di precisione.

Se Y appartiene alla famiglia $\mathcal{F}_{de}^{p}$ espressa dalla (A.3), allora la variabile casuale Y ha funzione generatrice dei momenti

$$M_Y(t;\theta,\lambda) = e^{\lambda[K(\theta+t/\lambda)-K(\theta)]} \ .$$

Segue che la funzione generatrice dei cumulanti è

$$K_Y(t;\theta,\lambda) = \lambda[K(\theta+t/\lambda)-K(\theta)] \ .$$

Un generico cumulante di ordine r risulta: $\kappa^r = \lambda^{1-r}\frac{\partial^r K(\theta)}{\partial\theta^r}$. Con $p=1$ il vettore dei valori attesi e la funzione di varianza sono definiti da

$$\mu(\theta) = \frac{\partial K(\theta)}{\partial\theta} = \left(\frac{\partial K(\theta)}{\partial\theta_1},\ldots,\frac{\partial K(\theta)}{\partial\theta_p}\right)^T \ ,$$

$$V(\mu) = \frac{\partial^2 K(\theta)}{\partial\theta}\partial\theta^T\Big|_{\theta=\theta(\mu)} \ .$$

Si ottiene poi la relazione $Var(Y) = \frac{1}{\lambda}V(\mu)$.

Può risultare talvolta utile ricorrere alla riparametrizzazione di una $\mathcal{F}_{de}^{p}$ con (μ,σ^2),

dove $\mu = \mu(\theta)$ e $\sigma^2 = 1/\lambda$ il parametro σ^2 è detto parametro di dispersione. Con riferimento a questa riparametrizzazione si utilizzerà la notazione

$$Y \sim DE_p(\mu, \sigma^2 V(\mu)) \ ,$$

per indicare che la variabile casuale Y ha densità $p(y;\theta,\lambda)$ di tipo $\mathcal{F}_{de}^{p}$, con $\theta = \theta(\mu)$, $\lambda = 1/\sigma^2$ e funzione di varianza $V(\mu)$, $\theta(\mu)$ è l'inversa di $\mu(\theta)$. Questa riparametrizzazione risulta utile per la definizione dei modelli lineari generalizzati per i quali l'assunzione è che si abbiano n osservazioni indipendenti, y_i, determinazioni di una variabile casuale Y_i con distribuzione $DE_1(\mu_i, \sigma^2 V(\mu_i))$, con μ_i funzione di una combinazione lineare dei valori di k variabili esplicative.

A.2 Distribuzione gamma

La distribuzione gamma con parametri (α,λ), $\alpha > 0$ e $\lambda > 0$, ha densità

$$p(y;\alpha,\lambda) = \frac{\lambda e^{-\lambda y}(\lambda y)^{\alpha-1}}{\Gamma(\alpha)} \ , \ y > 0 \ ,$$

dove $\Gamma(\alpha)$, è la funzione gamma definita da

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1} dy \ .$$

Vale la relazione ricorsiva $\Gamma(\alpha) = (\alpha-1)!$ per valori interi di α.

La distribuzione gamma con $\alpha = 1$ coincide con la distribuzione esponenziale, e se $\lambda = \frac{1}{2}$ e $\alpha = \frac{n}{2}$, allora la distribuzione coincide con la χ^2_n, con n gradi di libertà.

Il valore atteso e la varianza della distribuzione gamma sono:

$$E[Y] = \frac{\alpha}{\lambda} \ ,$$

$$Var[Y] = \frac{\alpha}{\lambda^2} \ .$$

A.3 Inversa matrice a blocchi

Sia A matrice quadrata, con A_{11} e A_{22} sottomatrici quadrate non singolari. Verrà utilizzata la scrittura

$$\mathbf{A^{-1}} = \begin{pmatrix} A_{11} & A_{12} \\ \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ \\ A^{21} & A^{22} \end{pmatrix} ,$$

Per il calcolo dell'inversa è noto che:

$$\begin{cases} A^{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} \\ A^{12} = -A^{11}A_{12}A_{22}^{-1} \\ A^{21} = -A_{22}^{-1}A_{21}A^{11} \\ A^{22} = A_{22}^{-1} + A_{22}^{-1}A_{21}A^{11}A_{12}A_{22}^{-1} \end{cases} .$$

Poiché $(A^T)^{-1} = (A^{-1})^T$, si può riscrivere come

$$\begin{cases} A^{11} = A_{11}^{-1} + A_{11}^{-1}A_{12}A^{22}A_{21}A_{11}^{-1} = A_{11}^{-1}[I + A_{12}A^{22}A_{21}A_{11}^{-1}] \\ A^{12} = -A_{11}^{-1}A_{12}A^{22} \\ A^{21} = -A^{22}A_{21}A_{11}^{-1} \\ A^{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} \end{cases} .$$

Appendice B

Appendice B

B.1 Funzioni generatrici

Funzioni generatrici dei momenti

La funzione generatrice dei momenti $M_Y(t)$ di una variabile aleatoria Y è definita per t reale da: $M_Y(t) = E[e^{tY}]$ che significa,

$$M_Y(t) = \sum_t e^{ty} p(y)$$

$$M_Y(t) = \int_{-\infty}^{+\infty} e^{ty} p(y) dy \ ,$$

nel primo caso se Y è discreta con densità $p(y)$, nel secondo se Y è continua con densità $p(y)$. Se la funzione generatrice dei momenti determina la distribuzione di Y in modo unico; cioè se $M_Y(t)$ esiste ed è finita in un intorno di $t = 0$ allora la distribuzione di Y è univocamente determinata. Inoltre, la funzione $M_Y(t)$ è chiamata generatrice dei momenti in quanto tutti i momenti di Y si possono ottenere derivando rispetto a t successivamente $M_Y(t)$ e valutandone il risultato in $t = 0$.

Per approfondimenti vedi Pace e Salvan (1996, § 3.3.1).

Funzione generatrice dei cumulanti

Sia $M_Y(t)$ funzione generatrice dei momenti di una variabile aleatoria Y, allora si definisce funzione generatrice dei cumulanti

$$K_Y(t) = \log\Big(M_Y(t)\Big) \ ,$$

il cumulante di ordine r è dato da

$$\kappa_r(Y) = \frac{\partial^r}{\partial t^r} K_Y(t)\Big|_{t=0} \ .$$

Si osservi che la funzione generatrice dei cumulanti della variabile casuale somma risulta

$$K_{Sn}(t) = \log\left(M_y(t)\right)^n = nK_Y(t) \ ,$$

per cui vale la relazione,

$$\kappa_r(Sn) = nk_r(Y) \ .$$

Per approfondimenti vedi Pace e Salvan (1996, § 3.3.5).

Appendice C

Appendice C

C.1 Insiemi di dati

C.1.1 Dati stazione metereologica, Arpav

Viene qui riportato il dataset originale tratto da Arpav (2009).

Data	Tmedia	Tmin	Tmax	Pioggia	Umin	Umax	Radiaz	Vento	Direz
27/03/09	3.0	-0.5	5.5	0.0	60	100	13.161	98.9	S
28/03/09	3.2	2.4	4.0	36.4	100	100	0.948	136.7	E
29/03/09	3.9	2.4	4.8	64.2	89	100	1.418	174.8	N
30/03/09	4.4	2.4	7.5	15.4	74	100	6.979	285.4	N
31/03/09	5.6	3.3	9.2	6.4	64	100	10.958	108.9	S
01/04/09	6.9	3.2	10.2	9.4	63	100	4.688	90.4	O
02/04/09	5.5	4.4	7.4	50.6	83	100	1.469	194.5	E
03/04/09	8.0	4.5	11.8	0.0	60	99	15.596	120.1	N
04/04/09	8.4	6.3	11.8	2.4	63	94	4.563	139.0	N
05/04/09	9.0	6.5	12.0	0.8	58	100	6.841	142.5	N
06/04/09	10.3	6.3	14.7	0.0	47	79	13.989	172.6	N
07/04/09	9.9	6.5	13.9	1.2	45	80	12.533	143.5	N
08/04/09	9.2	6.2	13.5	5.6	49	93	12.282	121.6	N
09/04/09	9.2	5.8	12.5	0.0	68	97	16.154	140.9	N

10/04/09	9.3	5.6	13.2	0.0	49	89	22.662	166.6	N
11/04/09	8.7	5.4	12.8	0.0	45	92	20.240	174.2	N
12/04/09	9.3	5.9	13.0	0.0	44	83	19.618	155.9	N
13/04/09	11.7	7.1	16.4	0.4	25	80	21.125	250.6	N
14/04/09	11.4	7.7	14.9	0.4	40	75	18.809	211.6	N
15/04/09	11.7	7.5	15.7	0.0	36	66	17.409	140.7	N
16/04/09	7.0	4.9	10.2	38.8	59	100	1.927	122.5	E
17/04/09	7.2	4.5	11.3	0.0	49	84	16.813	271.9	O
18/04/09	5.7	3.8	7.8	7.2	70	100	8.868	103.2	N
19/04/09	5.9	4.7	8.1	26.4	57	100	1.745	167.6	E
20/04/09	6.1	4.3	8.6	32.0	75	100	1.167	187.0	E
21/04/09	8.9	6.7	11.3	4.0	52	100	2.843	128.6	N
22/04/09	11.6	6.9	16.0	0.0	39	78	25.901	232.0	N
23/04/09	6.1	2.2	12.1	17.8	54	100	2.592	167.4	N
24/04/09	4.6	1.9	6.6	0.0	72	92	10.305	135.6	S
25/04/09	7.5	3.0	11.4	0.0	56	90	20.070	143.6	N
26/04/09	5.8	4.2	7.5	36.8	78	100	2.391	109.7	N
27/04/09	6.4	5.4	7.5	168.6	100	100	0.144	352.1	E
28/04/09	6.0	5.6	6.4	119.6	99	100	1.029	422.1	E
29/04/09	4.7	2.8	7.2	38.6	74	100	6.828	155.6	N
30/04/09	7.2	3.7	9.8	3.6	50	97	12.583	249.7	N
01/05/09	11.9	8.3	16.4	0.0	33	67	20.999	320.8	N
02/05/09	12.9	7.5	16.9	0.0	25	100	26.547	292.2	N
03/05/09	10.7	6.5	14.6	0.0	39	97	24.294	207.7	N
04/05/09	9.4	6.3	13.1	4.6	56	100	9.539	220.7	N
05/05/09	7.8	4.0	11.4	0.0	53	100	17.290	135.5	E
06/05/09	10.3	6.3	14.5	0.0	52	93	23.208	108.2	S
07/05/09	12.5	8.7	16.4	0.0	39	94	27.363	143.6	S
08/05/09	11.7	9.5	13.9	0.0	80	97	10.907	111.4	N
09/05/09	12.3	10.2	14.9	0.6	74	99	14.742	97.7	S

10/05/09	13.1	10.2	15.7	0.0	62	99	17.880	141.2	E
11/05/09	14.7	11.2	18.7	0.0	41	88	24.306	172.3	N
12/05/09	14.6	11.5	17.7	0.0	48	93	24.558	123.4	S
13/05/09	12.9	10.9	16.0	0.0	72	100	17.102	104.8	N
14/05/09	12.7	11.1	14.4	0.0	80	97	10.682	88.4	S
15/05/09	10.3	8.7	11.9	3.4	74	100	5.397	72.9	N
16/05/09	11.7	8.1	14.3	0.0	74	100	12.207	101.9	N
17/05/09	14.6	10.4	17.8	0.0	55	100	19.273	116.8	N
18/05/09	16.3	12.5	20.0	0.0	50	83	23.051	142.4	N
19/05/09	15.9	13.3	19.0	0.0	57	90	20.867	138.8	N
20/05/09	16.4	13.4	19.3	0.0	62	90	21.344	128.1	N
21/05/09	17.3	13.2	21.3	0.0	45	91	23.422	132.0	N
22/05/09	17.5	13.9	21.5	0.0	43	86	27.325	134.9	S
23/05/09	17.6	13.8	21.0	0.0	71	92	25.216	130.8	N
24/05/09	19.5	14.9	23.1	0.0	26	100	25.606	163.3	N
25/05/09	22.6	17.2	26.8	0.2	20	98	28.599	290.5	N

Tabella C.1: Dataset misure atmosferiche, originale.

C.1.2 Dati oli combustibili, Prater

Viene qui riportato il dataset originale tratto da Ferrari e Cribari-Neto (2004).

	No	SG	VP	V10	EP	Y
1	A	50.8	8.6	190	205	12.2
2	A	50.8	8.6	190	275	22.3
3	A	50.8	8.6	190	345	34.7
4	A	50.8	8.6	190	407	45.7
5	B	40.8	3.5	210	218	8.0
6	B	40.8	3.5	210	273	13.1
7	B	40.8	3.5	210	347	26.6

8	C	40.0	6.1	217	212	7.4
9	C	40.0	6.1	217	272	18.2
10	C	40.0	6.1	217	340	30.4
11	D	38.4	6.1	220	235	6.9
12	D	38.4	6.1	220	300	15.2
13	D	38.4	6.1	220	365	26.0
14	D	38.4	6.1	220	410	33.6
15	E	40.3	4.8	231	307	14.4
16	E	40.3	4.8	231	367	26.8
17	E	40.3	4.8	231	395	34.9
18	F	32.2	5.2	236	267	10.0
19	F	32.2	5.2	236	360	24.8
20	F	32.2	5.2	236	402	31.7
21	G	41.3	1.8	267	235	2.8
22	G	41.3	1.8	267	275	6.4
23	G	41.3	1.8	267	358	16.1
24	G	41.3	1.8	267	416	27.8
25	H	38.1	1.2	274	285	5.0
26	H	38.1	1.2	274	365	17.6
27	H	38.1	1.2	274	444	32.1
28	I	32.2	2.4	284	351	14.0
29	I	32.2	2.4	284	424	23.2
30	J	31.8	0.2	316	365	8.5
31	J	31.8	0.2	316	379	14.7
32	J	31.8	0.2	316	428	18.0

Tabella C.2: Dataset misurato da Prater (1956), originale.

Bibliografia

[1] ARPAV (2009), SERVIZIO METEO REGIONALE VENETO. http://www.arpa.veneto.it/datirete.htm, *Arpav*.

[2] ATKINSON, A. C. (1985). *Plots, Transformations and Regression: an Introduction to Graphical Methods of Diagnostic Regression Analysis*, New York: Oxford University Press.

[3] COX, C. (1996). Nonlinear quasi-likelihood models: Application to continuous proportion, *Computational statistic & Data analisys* , **21**, 449-61.

[4] FERRARI, S.L.P. , CRIBARI-NETO, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics*, **31**, 799815.

[5] GREGORIO, E. , SALCE, L. (2005). *Algebra Lineare*, Libreria Progetto, Padova.

[6] JØRGENSEN B (1997). *The Theory of Dispersion Models*, New York: Chapman & Hall.

[7] KIESCHNICK, R. , MCCULLOUGH, B.D. (2003). Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions, *Statistical Modelling*, **3**, 193213.

[8] PACE, L. , SALVAN, A. (1996). *Teoria della Statistica*, Cedam , Padova.

[9] PACE, L. , SALVAN, A. (2001). *Introduzione alla Statistica II*, Cedam , Padova.

[10] PRATER, N. H. (1956). Estimate gasoline yields from crudes, *Petroleum Refiner*, **35**, pp 236-238.

[11] RAYDONAL, O. , CRIBARI-NETO, F. , VASCONCELLOS, K.L.P. (2006). Improved point and interval estimation for a beta regression model, *Computational Statistics & Data Analysis*, **51**, 960981.

[12] ROSS, S.M. (2004). *Calcolo delle Probabilità*, Apogeo, Milano. Edizione Italiana a cura di: FERRANTE, M. E MARICONDA, C.

[13] SONG, P. X.-K. (2006). *Correlated Data Analysis: Modeling, Analytics and Applications*, Springer.

Elenco delle tabelle

Elenco delle figure

Ringraziamenti

Desidero innanzitutto ringraziare la Prof. Alessandra Salvan, per la disponibilità, competenza e serietà dimostrata nel seguirmi lungo la stesura di questo lavoro, ho appreso molte cose che mi potranno essere utili per il prosieguo dei miei studi.

Ringrazio di cuore i miei genitori che nell'arco di questo tempo in cui spesso mi sono assentato, dal consueto svolgersi della vita familiare, mi hanno sostenuto ed apprezzato per gli sforzi compiuti. Anche mia sorella merita una menzione perché anche se piccola ha saputo starmi vicino e darmi motivazioni.

Un grazie anche agli amici di sempre con cui ho trascorso momenti felici e spensierati in cui condividevamo esperienze universitarie. Mi hanno appoggiato ed aiutato, il merito è anche loro per questo traguardo.

Infine, ringrazio le cantine Soldà che mi hanno offerto possibilità di lavorare e quindi di mantenermi, almeno in parte, gli studi. Sono stati pazienti e gentili ad adattarsi alle mie esigenze di studio dimostrandosi affidabili e comprensivi.

Resta da dire che grandi sollecitazioni mi sono arrivate da una mia recente amicizia, enorme è stata la sua pazienza a starmi vicino in molte situazioni, probabilmente è una persona che crede nella provvidenza.