Il Metodo dei Minimi Quadrati: Alcuni Esempi Svolti

Alessandro Zaccagnini
alessandro.zaccagnini@unipr.it

14 ottobre 2005

Capitolo 1

Modelli lineari

1.1 Definizioni

Ricordiamo le definizioni: sono date n coppie di numeri reali (x_i, y_i) e si vuole determinare la *retta di regressione* per questi dati. Definiamo

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \qquad \overline{(xy)} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2$$

Ricordiamo le relazioni

$$\mathrm{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \overline{(xy)} - \overline{x}\cdot\overline{y}$$

$$\mathrm{Var}(x) = \mathrm{Cov}(x,x) = \sigma_x^2 = \overline{x^2} - (\overline{x})^2$$

Vogliamo trovare un modello lineare fra i valori x ed i valori y del tipo $\widehat{y} = ax + b$, con $a, b \in \mathbb{R}$, e quindi consideriamo la funzione $\Phi\colon \mathbb{R}^2 \to \mathbb{R}$ definita da

$$\Phi(a,b) = \frac{1}{m}\sum_{i=1}^{n} (y_i - ax_i - b)^2,$$

che rappresenta lo scarto quadratico medio dei valori reali (x_i, y_i) da quelli "previsti" dal modello lineare. Supponiamo che almeno due degli x_i siano differenti fra loro: in altre parole, supponiamo che $\sigma_x \neq 0$. Lo sviluppo della funzione Φ rivela che si tratta di un polinomio di

secondo grado nelle variabili a e b, e non è difficile dimostrare che la superficie $z = \Phi(a,b)$ è quella di un paraboloide ellittico. Per trovare l'unico punto di minimo determiniamo $\nabla\Phi$:

$$\begin{cases} \dfrac{\partial\Phi}{\partial a} = 2a\overline{x^2} + 2b\overline{x} - 2\overline{(xy)} \\ \dfrac{\partial\Phi}{\partial b} = 2a\overline{x} + 2b - 2\overline{y}. \end{cases}$$

Dobbiamo quindi risolvere il sistema

$$\begin{cases} b + \overline{x}a = \overline{y} \\ \overline{x}b + \overline{x^2}a = \overline{(xy)} \end{cases} \qquad \text{oppure} \qquad \begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \overline{y} \\ \overline{(xy)} \end{bmatrix}$$

Chiamiamo M la matrice qui sopra, e sia $D = \det(M) = \overline{x^2} - (\overline{x})^2 = \sigma_x^2$, ricordando che $D \neq 0$ se $n \geq 2$ e gli x_i non sono tutti uguali. Abbiamo dunque

$$M^{-1} = \begin{bmatrix} \overline{x^2}/D & -\overline{x}/D \\ -\overline{x}/D & 1/D \end{bmatrix} \quad \text{e} \quad \begin{bmatrix} b \\ a \end{bmatrix} = M^{-1} \begin{bmatrix} \overline{y} \\ \overline{(xy)} \end{bmatrix} = \begin{bmatrix} \left(\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{(xy)}\right)/\sigma_x^2 \\ \left(\overline{xy} - \overline{x} \cdot \overline{y}\right)/\sigma_x^2 \end{bmatrix}$$

In alternativa, usando le relazioni qui sopra, si trova la soluzione

$$\begin{cases} a_0 = \mathrm{Cov}(x,y)\sigma_x^{-2} \\ b_0 = \overline{y} - a_0\overline{x} \end{cases} \qquad \text{ed} \qquad r = a_0 \cdot \frac{\sigma_x}{\sigma_y} = \frac{\mathrm{Cov}(x,y)}{\sigma_x\sigma_y} = \frac{\overline{(xy)} - \overline{x}\cdot\overline{y}}{\sigma_x\sigma_y}.$$

Il valore minimo di Φ si ottiene sostituendo i valori di a_0 e b_0 e con qualche calcolo si trova

$$\Phi(a_0,b_0) = \frac{\sigma_x^2\sigma_y^2 - \mathrm{Cov}(x,y)^2}{\sigma_x^2}.$$

Ricordiamo che, per la disuguaglianza di Cauchy, si ha $\Phi(a_0,b_0) \geq 0$ ed $r \in [-1,1]$.

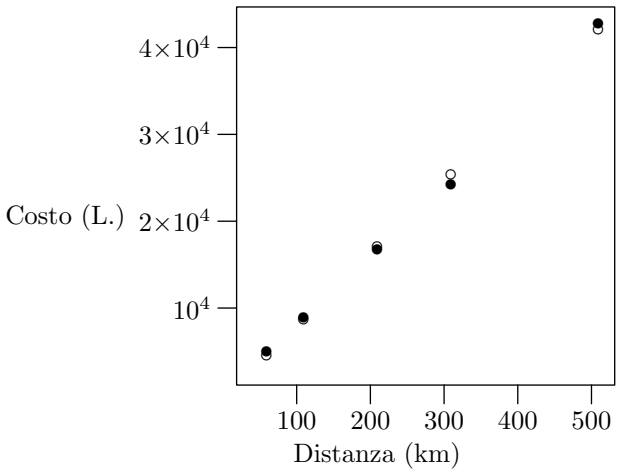
Qui di seguito daremo alcuni esempi svolti di applicazione del metodo dei minimi quadrati con la stessa notazione usata sopra. In particolare, $\widehat{y_i} = ax_i + b$ è il valore di y "previsto" dal modello lineare, mentre y_i è il valore sperimentale. Nei grafici, i valori sperimentali sono indicati da $\bullet$, mentre i valori "previsti" sono indicati da $\circ$.

1.2 Tariffe ferroviarie

Le tariffe ferroviarie per la seconda classe in vigore nell'estate 1999 sono parzialmente riportate nella tabella che segue:

Distanza (km)	50	100	200	300	500
Costo (L.)	4300	8200	16000	23500	42000

Usando opportune unità di misura, determinare un modello ragionevole ed il relativo coefficiente di correlazione per questi dati. Calcolare il costo previsto dal modello per un viaggio di 1000 km.



In questo caso è ragionevole cercare un modello lineare.

$\overline{x} = 230$ $\quad \overline{y} = 18800$

$\overline{x^2} = 78500$ $\quad \overline{y^2} = 531596000$

$\overline{xy} = 6457000$

$\sigma_x = 160$ $\quad \sigma_y = 13347.50913$

$a = 83.3203125000$

$b = -363.6718750000$

$r = 0.9987818603$

i	x_i	y_i	$\widehat{y_i}$
1	50	4300	3802.34375
2	100	8200	7968.35938
3	200	16000	16300.39062
4	300	23500	24632.42188
5	500	42000	41296.48438

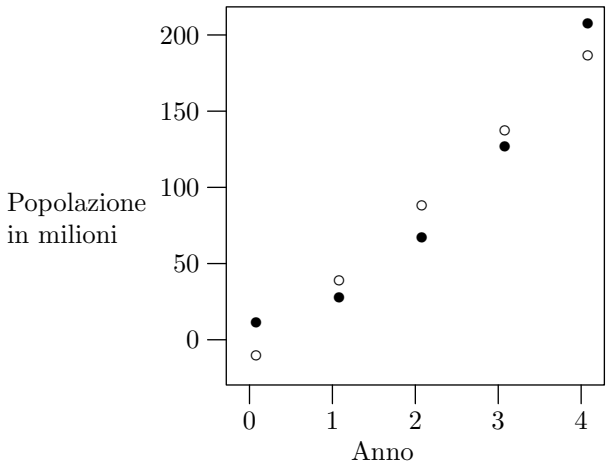
Il costo previsto di un viaggio di 1000 km è dunque 82956.64 L.

1.3 Popolazione degli Stati Uniti, I

La tabella che segue dà la popolazione degli Stati Uniti (in milioni) come risulta da censimenti effettuati negli anni indicati.

Anno	1810	1850	1890	1930	1970
Pop.	7.24	23.2	62.9	122.8	203.2

Usando opportune unità di misura, determinare la retta di regressione ed il coefficiente di correlazione per questi dati, e, mediante un cambiamento di variabili adatto, si trovi un modello non lineare. In entrambi i casi, calcolare i valori previsti dai modelli utilizzati e il numero di abitanti previsto per l'anno 2010. Dire in quale anno gli abitanti supereranno i 300 milioni, utilizzando entrambi i modelli proposti.



Conviene utilizzare come unità di misura $x = (A - 1810)/40$, dove A è l'anno dato nella tabella.

$\overline{x} = 2.0$ $\quad$ $\overline{y} = 83.86800$ $\quad$ $a = 49.152000$

$\overline{x^2} = 6.0$ $\quad$ $\overline{y^2} = 12183.42952$ $\quad$ $b = -14.436000$

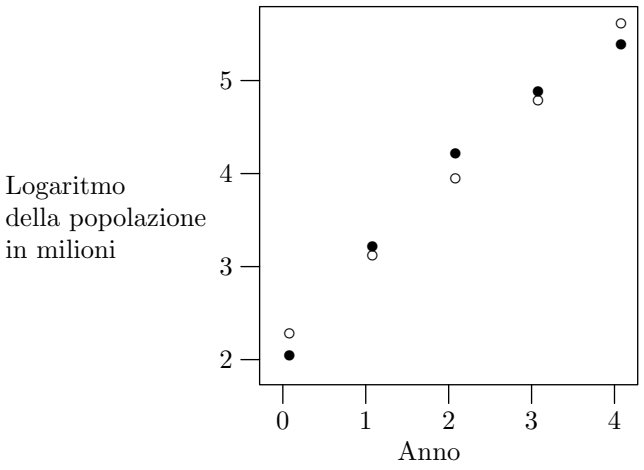
$\overline{xy} = 266.04000$ $\quad$ $r = 0.9686568316$

$\sigma_x = 1.41421$ $\quad$ $\sigma_y = 71.76063$

i	x_i	y_i	$\widehat{y_i}$
1	0.0	7.24000	−14.43600
2	1.0	23.2	34.71600
3	2.0	62.9	83.86800
4	3.0	122.8	133.02000
5	4.0	203.2	182.17200

1.4 Popolazione degli Stati Uniti, II

Si prenda un modello del tipo $y = e^{ax+b}$.



$\overline{x} = 2.0$ $\quad \overline{y} = 3.87801$

$\overline{x^2} = 6.0$ $\quad \overline{y^2} = 16.46782$ $\quad a = 0.8335543756$

$\overline{xy} = 9.42314$ $\quad b = 2.2109047249$

$\sigma_x = 1.41421$ $\quad \sigma_y = 1.19534$ $\quad r = 0.9861864576$

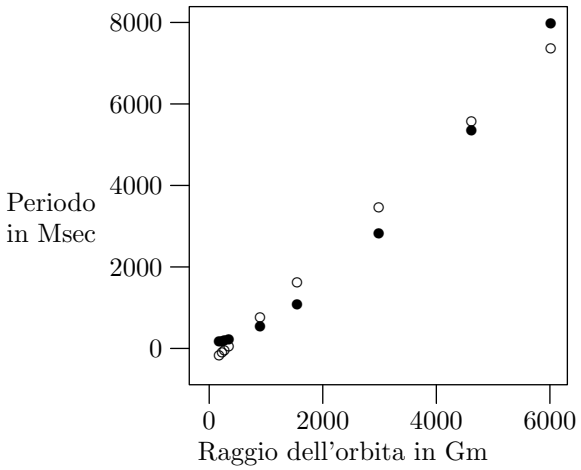
i	x_i	y_i	$\widehat{y}_i$
1	0.0	1.97962	2.21090
2	1.0	3.14415	3.04446
3	2.0	4.14155	3.87801
4	3.0	4.81056	4.71157
5	4.0	5.31419	5.54512

1.5 Terza Legge di Keplero, I

La tabella che segue riporta il raggio medio dell'orbita R ed il periodo di rivoluzione T dei pianeti del sistema solare (espressi rispettivamente in milioni di chilometri ed in milioni di secondi).

Pianeta	*Me*	*Ve*	*Te*	*Ma*	*Gi*	*Sa*	*Ur*	*Ne*	*Pl*
R	57.9	108	150	228	778	1430	2870	4500	5900
T	7.6	19.4	31.6	59.4	374	930	2650	5200	7820

Usando opportune unità di misura, determinare la retta di regressione ed il coefficiente di correlazione per questi dati, e, mediante il cambiamento di variabili adatto, si trovi un modello non lineare del tipo $T = AR^B$. (Si confronti con la terza legge di Keplero).



Conviene usare le unità di misura indicate. Il modello lineare $T = aR + b$ prevede per Mercurio, Venere, Terra e Marte un periodo negativo!

$\overline{x} = 1780.21111$ $\overline{y} = 1899.11111$ $a = 1.2890531142$

$\overline{x^2} = 7337398.26778$ $\overline{y^2} = 10691626.33778$ $b = -395.6755655833$

$\overline{xy} = 8753910.04889$ $r = 0.9887294930$

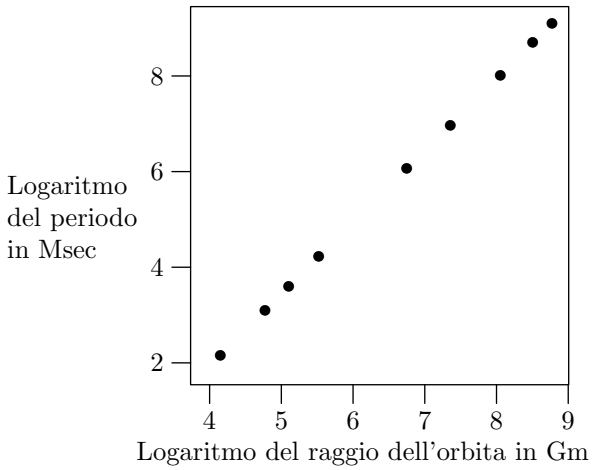
$\sigma_x = 2041.62844$ $\sigma_y = 2661.76696$

i	x_i	y_i	$\widehat{y}_i$
0	57.9	7.6	−321.03939
1	108	19.4	−256.45783
2	150	31.6	−202.31760
3	228	59.4	−101.77146
4	778	374	607.20776
5	1430	930	1447.67039
6	2870	2650	3303.90687
7	4500	5200	5405.06345
9	5900	7820	7209.73781

i	x_i	y_i	$\widehat{y}_i$
1	4.05872	2.02815	2.02798
2	4.68213	2.96527	2.96296
3	5.01064	3.45316	3.45564
4	5.42935	4.08429	4.08360
5	6.65673	5.92426	5.92439
6	7.26543	6.83518	6.83730
7	7.96207	7.88231	7.88209
8	8.41183	8.55641	8.55664
9	8.68271	8.96444	8.96288

1.6 Terza Legge di Keplero, II

Modello non lineare $\log T = a \log R + b$. La tabella relativa è data qui sopra.



$$\overline{x} = 6.46218 \qquad \overline{y} = 5.63261$$
$$\overline{x^2} = 44.40235 \qquad \overline{y^2} = 37.67031$$
$$\overline{xy} = 40.36222$$
$$\sigma_x = 1.62561 \qquad \sigma_y = 2.43804$$

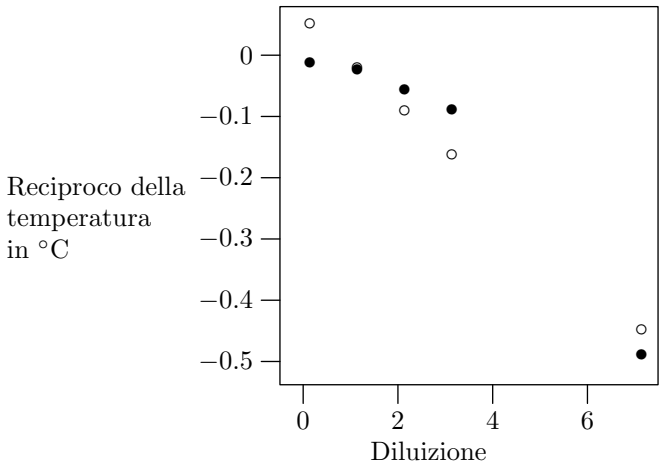
$$a = 1.4997654908$$
$$b = -4.0591410481$$
$$r = 0.9999998222$$

1.7 Anticongelante, I

Sulla confezione di un prodotto anticongelante per auto è riprodotta la tabella qui sotto, il cui significato è che il miscuglio di 1 l di questo prodotto ed x l d'acqua congela alla temperatura T. In altre parole, il prodotto puro congela a -44°C, diluito in parti uguali di acqua congela a -30°C, e cosí via.

Diluizione	0	1	2	3	7
Temperatura ($^\circ$C)	-44	-30	-15	-10	-2

Usando opportune unità di misura, determinare un modello ragionevole ed il relativo coefficiente di correlazione per questi dati. Si osservi che in questo caso un modello lineare $T = ax + b$ è completamente fuori luogo (perché?). Si studino il modello $T^{-1} = ax + b$ (spiegando perché anche questo non va bene) ed il modello $\log(-T) = ax + b$, determinandone i relativi coefficienti di correlazione.



Il modello lineare $T = ax + b$ prevede che per $x \to +\infty$ la temperatura di congelamento tenda a $+\infty$, mentre deve necessariamente tendere a 0 restando negativa. Il modello $T^{-1} = ax + b$ non va bene (a posteriori) perché "prevede" una temperatura di congelamento positiva per il prodotto puro. Questo fatto è dovuto alla presenza di un asintoto verticale.

$$\overline{x} = 2.60000 \qquad \overline{y} = -0.14455$$
$$\overline{x^2} = 12.60000 \qquad \overline{y^2} = 0.05321$$
$$\overline{xy} = -0.79333$$
$$\sigma_x = 2.41661 \qquad \sigma_y = 0.17978$$

$$a = -0.0714923205$$
$$b = 0.0413345787$$
$$r = -0.9609996536$$

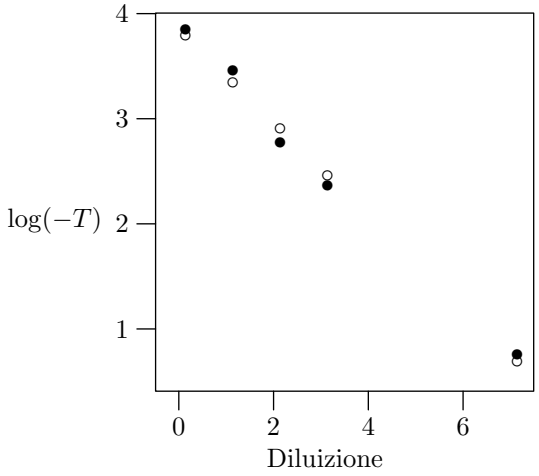
i	x_i	y_i	$\widehat{y}_i$
1	0	−0.02273	0.04133
2	1	−0.03333	−0.03016
3	2	−0.06667	−0.10165
4	3	−0.10000	−0.17314
5	7	−0.50000	−0.45911

1.8 Anticongelante, II

Modello $\log(-T) = ax + b$.

$$\overline{x} = 2.60000 \qquad \overline{y} = 2.57783$$
$$\overline{x^2} = 12.60000 \qquad \overline{y^2} = 7.80082$$
$$\overline{xy} = 4.11542$$
$$\sigma_x = 2.41661 \qquad \sigma_y = 1.07499$$

$$a = -0.4429711420$$
$$b = 3.7295588674$$
$$r = -0.9958150835$$



i	x_i	y_i	$\widehat{y}_i$
1	0	3.78419	3.72956
2	1	3.40120	3.28659
3	2	2.70805	2.84362
4	3	2.30259	2.40065
5	7	0.69315	0.62876

Capitolo 2

Modelli Quadratici

Vogliamo cercare un modello del tipo $y = ax^2 + bx + c$ per le n coppie di numeri reali (x_i, y_i). Per $j \in \mathbb{N}$ definiamo

$$\Sigma_j = \sum_{i=1}^{n} x_i^j \qquad \Sigma'_j = \sum_{i=1}^{n} x_i^j y_i$$

Ricordiamo che si deve risolvere il sistema

$$\begin{cases} \Sigma_0 c + \Sigma_1 b + \Sigma_2 a = \Sigma'_0 \\ \Sigma_1 c + \Sigma_2 b + \Sigma_3 a = \Sigma'_1 \\ \Sigma_2 c + \Sigma_3 b + \Sigma_4 a = \Sigma'_2 \end{cases} \quad \text{cioè} \quad \begin{bmatrix} \Sigma_0 & \Sigma_1 & \Sigma_2 \\ \Sigma_1 & \Sigma_2 & \Sigma_3 \\ \Sigma_2 & \Sigma_3 & \Sigma_4 \end{bmatrix} \begin{bmatrix} c \\ b \\ a \end{bmatrix} = \begin{bmatrix} \Sigma'_0 \\ \Sigma'_1 \\ \Sigma'_2 \end{bmatrix}$$

Il sistema può essere risolto determinando la matrice inversa, oppure ricavando una variabile da una delle equazioni e sostituendo nelle altre, fino alla determinazione di tutti i valori. Applicheremo questa teoria ad un solo caso, che studieremo nei dettagli.

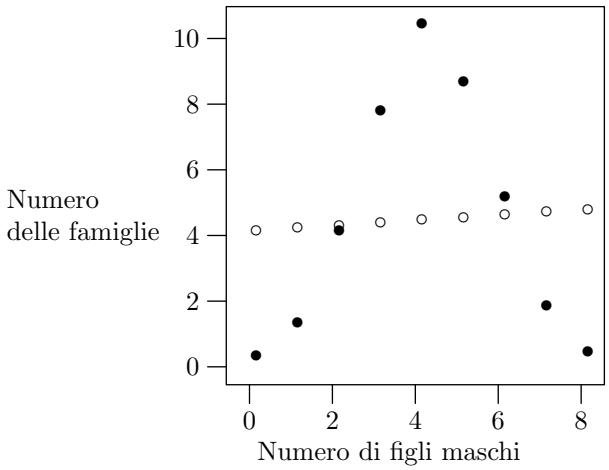
2.1 Statistica sul sesso dei figli nelle famiglie numerose

Un'indagine statistica sulle famiglie con esattamente 8 figli ha dato il risultato riassunto nella tabella che segue:

n. maschi	0	1	2	3	4	5	6	7	8
n. famiglie	161	1152	3951	7603	10263	8498	4984	1655	264

2.1.1 Il modello lineare

In questo caso è evidente a priori che un modello lineare è del tutto fuori luogo: infatti si ottengono i seguenti risultati.



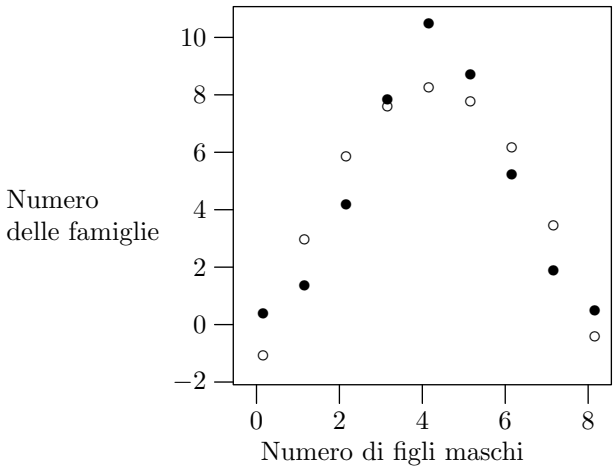
$$\overline{x} = 4.00000 \qquad \overline{y} = 4.28122$$
$$\overline{x^2} = 22.66667 \qquad \overline{y^2} = 31.10702$$
$$\overline{xy} = 17.66733$$
$$\sigma_x = 2.58199 \qquad \sigma_y = 3.57465$$

$$a = 0.0813666667$$
$$b = 3.9557555556$$
$$r = 0.0587715017$$

i	x_i	y_i	$\widehat{y_i}$
1	0.0	0.16100	3.95576
2	1.0	1.15200	4.03712
3	2.0	3.95100	4.11849
4	3.0	7.60300	4.19986
5	4.0	10.26300	4.28122
6	5.0	8.49800	4.36259
7	6.0	4.98400	4.44396
8	7.0	1.65500	4.52532
9	8.0	0.26400	4.60669

2.1.2 Il modello quadratico

Cerchiamo dunque un modello del tipo $\widehat{y} = ax^2 + bx + c$: il risultato è indicato nel grafico seguente.



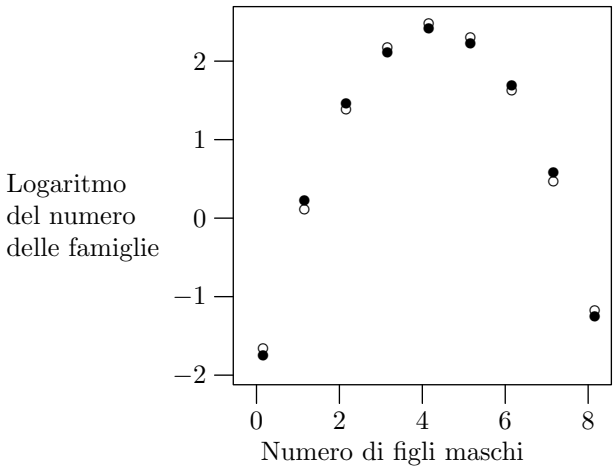
$$a = -0.5615887446$$
$$b = 4.5740766234$$
$$c = -1.2857393939$$

i	x_i	y_i	$\widehat{y}_i$
1	0.0	0.16100	−1.28574
2	1.0	1.15200	2.72675
3	2.0	3.95100	5.61606
4	3.0	7.60300	7.38219
5	4.0	10.26300	8.02515
6	5.0	8.49800	7.54493
7	6.0	4.98400	5.94153
8	7.0	1.65500	3.21495
9	8.0	0.26400	−0.63481

Anche questo modello risulta insoddisfacente, poiché prevede risultati negativi per $x = 0$ e per $x = 8$.

2.1.3 Il modello esponenziale-quadratico

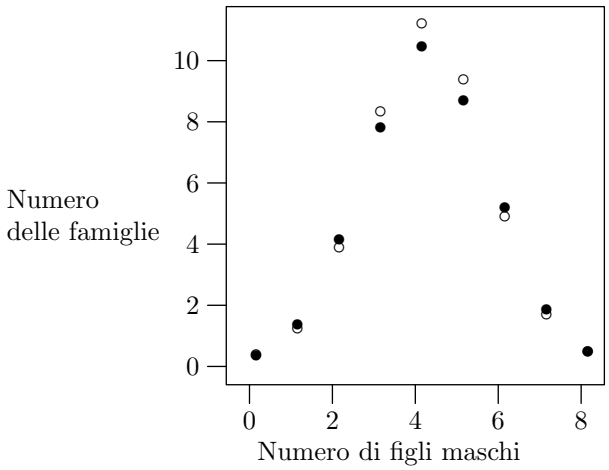
Cerchiamo infine un modello del tipo $\widehat{y} = \exp(ax^2 + bx + c)$: in altre parole, cerchiamo un modello quadratico fra x e $\log y$. Il risultato è indicato nel grafico seguente.



$$a = -0.2437079698$$
$$b = 2.0103454143$$
$$c = -1.7435272405$$

i	x_i	y_i	$\widehat{y}_i$
1	0.0	−1.82635	−1.74353
2	1.0	0.14150	0.02311
3	2.0	1.37397	1.30233
4	3.0	2.02854	2.09414
5	4.0	2.32855	2.39853
6	5.0	2.13983	2.21550
7	6.0	1.60623	1.54506
8	7.0	0.50380	0.38720
9	8.0	−1.33181	−1.25807

L'ultimo grafico riporta gli stessi risultati del precedente, ma con y al posto di $\log y$.



$a = -0.2437079698$

$b = 2.0103454143$

$c = -1.7435272405$

i	x_i	y_i	$\widehat{y}_i$
1	0.0	0.16100	0.17490
2	1.0	1.15200	1.02338
3	2.0	3.95100	3.67786
4	3.0	7.60300	8.11843
5	4.0	10.26300	11.00695
6	5.0	8.49800	9.16600
7	6.0	4.98400	4.68825
8	7.0	1.65500	1.47285
9	8.0	0.26400	0.28420