Data Mining / Intelligente Datenanalyse WS 2017/2018
PD Dr.-Ing. habil. Christian Borgelt

Lösung des 3. Übungsblattes

Aufgabe 14 Konfidenzintervalle

Der Punktschätzwert ist einfach die relative Häufigkeit einer Jungengeburt (konsistenter, erwartungstreuer, effizientester und suffizienter Schätzer), also

$$\hat{\theta} = \frac{45195}{87827} \approx 0.5146.$$

Im Skript zur Vorlesung ist folgende Näherungsformel für ein $(1-\alpha)\cdot 100\%$-Konfidenzintervall für den unbekannten Parameter θ einer Binomialverteilung angegeben:

$$A/B \;=\; \frac{1}{n + (A^*)^2}\left(X + \frac{(A^*)^2}{2} \mp A^*\sqrt{\frac{X(n-X)}{n} + \frac{(A^*)^2}{4}}\right),$$

mit $\Phi(A^*) = 1 - \frac{\alpha}{2}$. Da im vorliegenden Fall $n\theta(1-\theta)$ deutlich größer als 9 ist, ist die Annäherung der Binomialverteilung über die Normalverteilung gerechtfertigt. Aus einer Tabelle der Normalverteilung liest man ab (oder berechnet mit dem Program `ndqtl.c` von der WWW-Seite zur Vorlesung):

a) $\alpha = 0.01: \quad \Phi(A^*) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad \Phi(A^*) = 0.995 \quad \Rightarrow \quad A^* \approx 2.576,$
b) $\alpha = 0.001: \quad \Phi(A^*) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad \Phi(A^*) = 0.9995 \quad \Rightarrow \quad A^* \approx 3.291.$

Man erhält so

a) $\alpha = 0.01: \quad [0.5102 \le \theta \le 0.5189],$
b) $\alpha = 0.001: \quad [0.5090 \le \theta \le 0.5201],$

für die gesuchten Konfidenzintervalle.

Aufgabe 15 Konfidenzintervalle

Wie in den Beispielen in der Vorlesung setzen wir an:

$$P(W \le B^*) \;=\; \int_0^{B^*} \frac{1}{\theta} e^{-\frac{w}{\theta}} \mathrm{d}w \;=\; \frac{\alpha}{2},$$
$$P(W \ge A^*) \;=\; \int_{A^*}^{\infty} \frac{1}{\theta} e^{-\frac{w}{\theta}} \mathrm{d}w \;=\; \frac{\alpha}{2}.$$

Aus der ersten Beziehung erhalten wir

$$\int_0^{B^*} \frac{1}{\theta} e^{-\frac{w}{\theta}} \mathrm{d}w = \left[-e^{\frac{w}{\theta}}\right]_0^{B^*} = 1 - e^{-\frac{B^*}{\theta}} = \frac{\alpha}{2},$$

also

$$B^* = -\theta \ln\left(1 - \frac{\alpha}{2}\right).$$

Aus der zweiten Beziehung folgt

$$\int_{A^*}^{\infty} \frac{1}{\theta} e^{-\frac{w}{\theta}} \mathrm{d}w = \left[-e^{\frac{w}{\theta}}\right]_{A^*}^{\infty} = 0 + e^{-\frac{A^*}{\theta}} = \frac{\alpha}{2},$$

also

$$A^* = -\theta \ln\left(\frac{\alpha}{2}\right).$$

Damit haben wir

$$P(B^* < W < A^*) = P\left(-\theta \ln\left(1 - \frac{\alpha}{2}\right) < W < -\theta \ln\left(\frac{\alpha}{2}\right)\right).$$

Durch Bilden der Umkehrfunktionen erhalten wir schließlich

$$P(A < \theta < B) = P\left(-\frac{W}{\ln\left(\frac{\alpha}{2}\right)} < \theta < -\frac{W}{\ln\left(1 - \frac{\alpha}{2}\right)}\right).$$

Aufgabe 16 Methode der kleinsten Quadrate/Regression

a) Laut Vorlesungsfolien ist die Ausgleichsgerade

$$y = \frac{s_{xy}}{s_x^2}(x - \bar{x}) + \bar{y}.$$

In Aufgabe 10 hatten wir bereits $s_x^2 = 4$ und $s_{xy} = \frac{10}{3}$ berechnet. Die Mittelwerte $\bar{x}$ und $\bar{y}$ erhält man aus den ebenfalls in Aufgabe 10 bereits berechneten Summen

$$\begin{aligned}
\sum_{i=1}^{10} x_i &= 0+1+1+2+3+3+4+5+5+6 = 30, \qquad \text{also} \quad \bar{x} = \tfrac{30}{10} = 3,\\
\sum_{i=1}^{10} y_i &= 0+1+2+3+2+3+4+4+6+5 = 30, \qquad \text{also} \quad \bar{y} = \tfrac{30}{10} = 3.
\end{aligned}$$

Damit ist die gesuchte Ausgleichsgerade

$$y = \frac{10}{3 \cdot 4}(x - 3) + 3 = \frac{1}{2} + \frac{5}{6}x.$$

b) Um das zu lösende System der Normalgleichungen aufzustellen, benutzen wir die bereits in Aufgabe 10 berechneten Summen über die x_i und die y_i (siehe Teilaufgabe a) sowie

$$\begin{aligned}
\sum_{i=1}^{10} x_i^2 &= 0+1+1+4+9+9+16+25+25+36 = 126,\\
\sum_{i=1}^{10} x_i y_i &= 0+1+2+6+6+9+16+20+30+30 = 120.
\end{aligned}$$

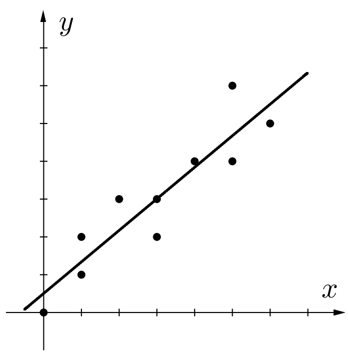
Damit erhalten wir das Gleichungssystem (Normalgleichungen)

$$\begin{aligned}
10a + 30b &= 30,\\
30a + 126b &= 120,
\end{aligned}$$

das die Lösung $a = \frac{1}{2}$ und $b = \frac{5}{6}$ besitzt. Die gesuchte Ausgleichsgerade ist also

$$y = \frac{1}{2} + \frac{5}{6}x.$$

Diese Gerade ist zusammen mit den Datenpunkten in dem rechts gezeigten Diagramm dargestellt.



Aufgabe 17 Methode der kleinsten Quadrate/Regression

Um das zu lösende System der Normalgleichungen aufzustellen, berechnen wir

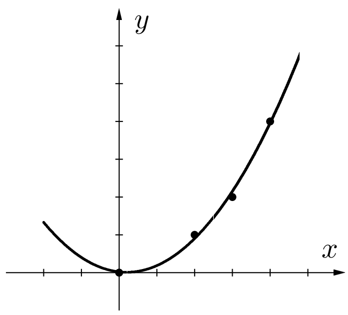
$$\begin{aligned}
\sum_{i=1}^{4} x_i &= 0+2+3+4 &&= 9, & \sum_{i=1}^{4} y_i &= 0+1+2+4 &&= 7,\\
\sum_{i=1}^{4} x_i^2 &= 0+4+9+16 &&= 29, & \sum_{i=1}^{4} x_i y_i &= 0+2+6+16 &&= 24,\\
\sum_{i=1}^{4} x_i^3 &= 0+8+27+64 &&= 99, & \sum_{i=1}^{4} x_i^2 y_i &= 0+4+18+64 &&= 86,\\
\sum_{i=1}^{4} x_i^4 &= 0+16+81+256 &&= 353.
\end{aligned}$$

Damit erhalten wir das Gleichungssystem

$$\begin{aligned}
4a + 9b + 29c &= 7,\\
9a + 29b + 99c &= 24,\\
29a + 99b + 353c &= 86,
\end{aligned}$$

das die Lösung $a = \frac{1}{55}$, $b = -\frac{6}{55}$ und $c = \frac{3}{11}$ besitzt. Die gesuchte Ausgleichsparabel ist folglich

$$y = \frac{1}{55} - \frac{6}{55}x + \frac{3}{11}x^2.$$



Diese Parabel ist zusammen mit den Datenpunkten in dem rechts gezeigten Diagramm dargestellt.

Aufgabe 18 Multilineare Regression

Um das zu lösende System der Normalgleichungen aufzustellen, berechnen wir

$$\begin{aligned}
\sum_{i=1}^{6} x_i &= 0+0+2+3+2+4 &&= 11, & \sum_{i=1}^{6} y_i &= 1+4+0+1+3+4 &&= 13,\\
\sum_{i=1}^{6} z_i &= 0+2+1+2+3+4 &&= 12,\\
\sum_{i=1}^{6} x_i^2 &= 0+0+4+9+4+16 &&= 33, & \sum_{i=1}^{6} y_i^2 &= 1+16+0+1+9+16 &&= 43,\\
\sum_{i=1}^{6} x_i y_i &= 0+0+0+3+6+16 &&= 25, & \sum_{i=1}^{6} x_i z_i &= 0+0+2+6+6+16 &&= 30,\\
\sum_{i=1}^{6} y_i z_i &= 0+8+0+2+9+16 &&= 35.
\end{aligned}$$

Damit erhalten wir das Gleichungssystem

$$\begin{aligned} 6a &+ 11b + 13c = 12, \\ 11a &+ 33b + 25c = 30, \\ 13a &+ 25b + 43c = 35, \end{aligned}$$

das die Lösung $a = -\frac{151}{567}$, $b = \frac{649}{1134}$ und $c = \frac{91}{162}$ besitzt. Die gesuchte Ausgleichsebene ist folglich

$$z = -\frac{151}{567} + \frac{649}{1134}x + \frac{91}{162}y.$$

Aufgabe 19 Logistische Regression

Bei der logistischen Regression müssen wir die Transformation

$$z = \ln\left(\frac{Y-y}{y}\right)$$

auf die Daten anwenden und für die transformierten Daten eine Ausgleichsgerade bestimmen. Für die vorliegende Aufgabe ist $Y = 1060$. Wir erhalten also die transformierten Daten:

x	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
z	4.059	2.762	0.958	0.405	−1.306	−1.422	−1.757	−5.169	−5.169	−5.169

Wir berechnen nun

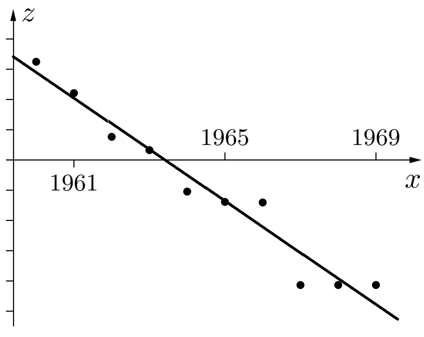
$$\sum_{i=1}^{10} x_i = 19645, \qquad \sum_{i=1}^{10} x_i^2 = 38592685, \qquad \sum_{i=1}^{10} z_i \approx -11.81, \qquad \sum_{i=1}^{10} x_i z_i \approx -23283.$$

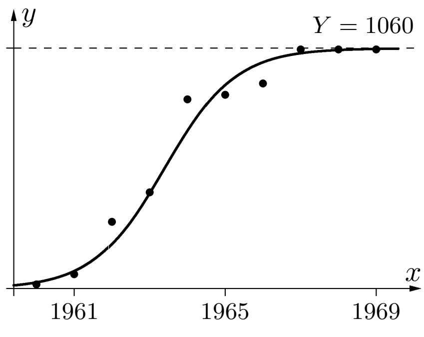
Damit erhalten wir das Gleichungssystem (Normalgleichungen)

$$\begin{aligned} 10a &+ 19645b = -11.81, \\ 19645a &+ 385926856 = -23283, \end{aligned}$$

das die Lösung $a \approx 2091.8$ und $b \approx -1.0654$ besitzt.

Die folgenden Diagramme zeigen die Ausgleichsgerade für die transformierten Daten und die zugehörige Ausgleichskurve $y \approx \frac{1060}{1+e^{2091.8-1.0654x}} \approx \frac{1060}{1+e^{1.0654(1963.4-x)}}$ für die Originaldaten.





Aufgabe 20 Exponentielle Regression

Die Gleichung $n = n_0(1-e^{a+bx})$ läßt sich wie folgt leicht in eine Geradengleichung umwandeln (vgl. auch die für die logistische Regression verwendete Transformation):

$$\begin{aligned} n &= n_0(1-e^{a+bt}) &\Leftrightarrow\quad \frac{n}{n_0} &= 1-e^{a+bt} \\ &&\Leftrightarrow\quad 1-\frac{n}{n_0} &= e^{a+bt} \\ &&\Leftrightarrow\quad \ln\frac{n_0-n}{n_0} &= a+bt \end{aligned}$$

Durch die Substitutionen $x = t$ und $y = \ln\frac{n_0-n}{n_0}$ erhalten wir die „Normalform“ $y = a+bx$, auf die sich die Methode der kleinsten Quadrate direkt anwenden läßt. Wir berechnen zunächst die transformierten Koordinaten der Meßpunkte:

x	0	30	60	90	120	150	180	210	240
y	0.000	−0.365	−0.803	−1.064	−1.461	−1.898	−2.313	−2.513	−3.124

Für diese Datenpunkte müssen wir eine Ausgleichsgerade bestimmen. Wir berechnen dazu

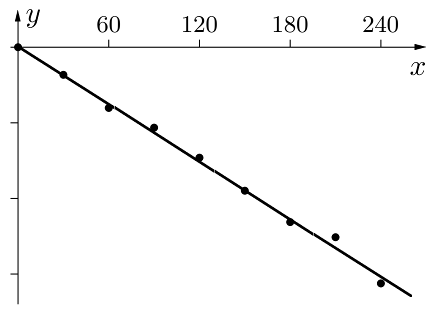
$$\sum_{i=1}^{9} x_i = 1080, \qquad \sum_{i=1}^{9} x_i^2 = 183600, \qquad \sum_{i=1}^{9} y_i \approx -13.63, \qquad \sum_{i=1}^{9} x_i y_i \approx -2322.13.$$

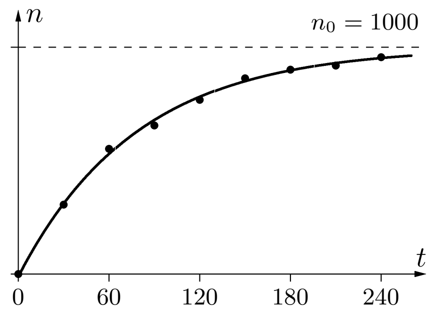
Damit erhalten wir das Gleichungssystem (Normalgleichungen)

$$\begin{aligned} 9a \;+\; 1080b &= -13.63, \\ 1080a \;+\; 183600b &= -2322.13, \end{aligned}$$

das die Lösung $a \approx 0.0109$ und $b \approx -0.0127$ besitzt.

Die folgenden Diagramme zeigen die Ausgleichsgerade für die transformierten Datenpunkte und die zugehörige Ausgleichskurve $n \approx 1000(1 - e^{0.0109-0.0127t})$ für die Originaldaten.





Die Halbwertszeit T der radioaktiven Substanz berechnet sich wie folgt: Bekanntlich ist die Halbwertszeit T definiert als die Zeit, nach der (im Durchschnitt) die Hälfte der anfänglich vorhandenen Teilchen zerfallen sind. Also gilt

$$\frac{1}{2}N_0 = N_0 e^{-\lambda T} = N_0 e^{bT} \qquad \Leftrightarrow \qquad T = \frac{\ln\frac{1}{2}}{b} \approx 54.5s.$$

Bei der radioaktiven Substanz handelt es sich daher wahrscheinlich um Radon Rn_{86}^{220}, dessen Halbwertzeit man in physikalischen Tabellenwerken mit $T = 55s$ angegeben findet.