Dr. Rudolf Riedi
HTA-FR, 2017-18

Statistik
Serie 1A: Lineare Regression

1. Die Methode der kleinsten Quadrate kann auch verwendet werden, um allgemeinere Gesetzmässigkeiten zwischen Variablen x und y aus Daten zu bestimmen. Wir diskutieren in dieser Aufgabe zwei solcher Gesetzmässigkeiten, nämlich ein Logarithmisches Gesetz von der Form

$$y = \log_b(x) + c$$

mit Basis b und Offset c, sowie ein Potenzgesetz von der Form

$$y = c \cdot x^b$$

mit Exponent b und Faktor c.

(a) Logarithmisches Gesetz
Betrachten wir zuerst ein einfaches Beispiel:

$$y = \log_2(x) - 3$$

Wie Sie schnell bestätigen, liegen die folgenden Punkte auf dieser Kurve (Siehe Figur links):

x	1	2	4	8	16	32	64
y	-3	-2	-1	0	1	2	3

Wird dies in Semi-log-Koordinaten gezeichnet (x in logarithmischer Skala, y normal), so wird der Graph linear. Wird der Logarithmus zur Basis 2 verwendet, so sind die Koordinaten der Punkte nun (Siehe Figur Mitte):

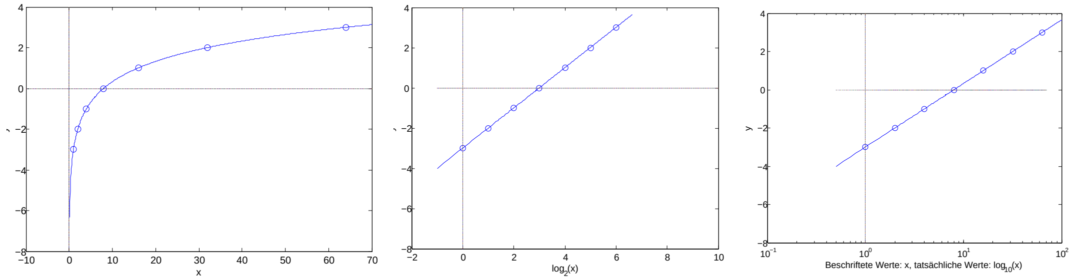
$\log_2(x)$	0	1	2	3	4	5	6
y	-3	-2	-1	0	1	2	3

In der Praxis kann man jedoch nicht für jede Basis entsprechendes Semi-Logarithmisches Papier bereit stellen. Normalerweise wird die Basis 10 gewählt. In dieser Basis sind die Koordinaten der Punkte nun:

$\log_{10}(x)$	0	0.3010	0.6021	0.9031	1.2041	1.5051	1.8062
y	-3	-2	-1	0	1	2	3

Siehe Figur rechts. Man beachte: in horizontaler Richtung ist $\log_{10}(x)$ abgetragen, die Beschriftung gibt jedoch die Werte von x an, nicht von $\log_{10}(x)$.
Man beachte weiter, dass die Steigung geändert hat. Sie beträgt nicht mehr 1 wie bei der Verwendung der Basis 2. Tatsächlich können wir einfach nachrechnen, dass ein Wechsel der Basis einfach die Steigung ändert:

$$y = \log_2(x) - 3 = \frac{1}{\log_d(2)} \cdot \log_d(x) - 3$$



Die Aufgabe besteht nun darin, das ursprüngliche Logarithmische Gesetz aus dem Graphen rechts, also aus der dritten Werte-Tabelle zu bestimmen. Die Punkte liegen nicht ganz perfekt auf einer Geraden, aber praktisch perfekt (der Korrelationskoeffizient beträgt 1.000).
Wir bestimmen die Regressionsgerade durch diese Punkte und erhalten die Steigung $m = 3.322259136212624$ und Offset $q = -2.999999999999999$. Dies bedeutet, dass wir schätzen

$$y \simeq 3.322 \cdot \log_{10}(x) - 3.000$$

Wir sehen dass der Offset $c = -3$ sehr gut geschätzt wurde. Um eine Schätzung der ursprünglichen Basis zu erhalten benützen wir den Basiswechsel, den wir oben schon bemerkt hatten. Wir schreiben b für die noch unbekannte Basis des Logarithmischen Gesetzes; die Basis d des Semi-Log Graphen beträgt $d = 10$:

$$y = \log_b(x) - 3.000 = \frac{1}{\log_{10}(b)} \cdot \log_{10}(x) - 3.000$$

Dementsprechend ist anscheinend

$$m = 3.322 = \frac{1}{\log_{10}(b)}$$

Daraus ergibt sich $b = 10^{1/m} = 10^{1/3.322} = 1.999'97$; also wurde die korrekte Basis $b = 2$ ebenfalls sehr gut geschätzt.

Aufgabe

Gegeben sind die folgenden Punkte (gemessen aus einem Semi-Log Graphen)

x	1	2	5	10	20	50	100
$\log_{10}(x)$	0	0.3010	0.6990	1.0000	1.3010	1.6990	2.0000
y	1.1101	3.1544	5.6524	7.4947	9.5696	12.1816	14.5228

i. Schätzen Sie Basis und Offset eines Logarithmischen Gesetzes.
ii. Wie gut gilt das Gesetz (praktisch perfekt, sehr gut, recht gut, nicht vertrauenswürdig, schlecht)? Beachten Sie: es geht nicht darum, wie gut Sie Basis und Offset schätzen, sondern wie gut die Daten einem Logarithmischen Gesetz folgen.
iii. Wie können Sie die Basis d des Semi-Log-Plots bestimmen, falls sie nicht bekannt ist?

(b) Potenzgesetz

Ein Potenzgesetz ist von der Form

$$y = c \cdot x^b$$

mit Exponent b und Faktor c. Will man ein solches Gesetz für gemessene Daten nachweisen, so trägt man die Daten auf Doppelt-Logarithmisches oder Log-Log Papier auf.
Mit anderen Worten, man betrachtet wie $\log(y)$ von $\log(x)$ abhängt. Durch logarithmieren des Potenzgesetzes erhält man:

$$\log_d(y) = \log_d(c \cdot x^b) = \log_d(c) + b \cdot \log_d(x).$$

Wie man sieht, hängt $\log_d(y)$ linear von $\log_d(x)$ ab, wobei die Steigung gerade b beträgt, egal wie man die Basis d des Logarithmus' wählt. Der Achsenabschnitt oder Offset ist $\log_d(c)$ und hängt davon ab, wie man d wählt.

Aufgabe

Gegeben sind die folgenden Punkte (gemessen aus einem Log-Log Graphen)

x	0.1	0.2	0.5	1	2	5	10
$\log_d(x)$	-1.0000	-0.6990	-0.3010	0	0.3010	0.6990	1.0000
$\log_d(y)$	-2.6825	-1.9303	-1.1206	-0.6265	0.0216	0.6793	1.2831

Es wird ein Potenzgesetz vermutet. Schätzen Sie den Exponenten b dieses Potenzgesetzes.
Wie gut folgen die Daten dem Gesetz (praktisch perfekt, sehr gut, recht gut, nicht vertrauenswürdig, schlecht)?