

Multivariate Regression

Ziel: Man will überprüfen, welche Variablen wieviel Varianz einer Zielvariable erklären und ob die durch die einzelnen Variablen wegerklärte Varianz signifikant von 0 verschieden ist (Anova) oder ob die Koeffizienten der Geradengleichung $y = Ab + r$ signifikant von 0 verschieden sind.

Voraussetzungen: Die Zielvariable ist metrisch skaliert. Die erklärenden Variablen sind metrisch skaliert oder weisen nur zwei Ausprägungen auf. Die Residuen sind normalverteilt. Die Varianzen der Residuen bezüglich der Regressiongerade sind identisch.

Gegeben sei ein Datensatz bezüglich n Objekten für die Variablen $(X_1, \dots, X_k, Y)$. (Ein solcher Datensatz hat die Form einer $n \cdot (k + 1)$ –Matrix!) Es wird ein linearer Zusammenhang zwischen den Variablen $X_1, \dots, X_k$ sowie einer Konstante b_0 einerseits, und der Zielvariable Y andererseits vermutet. Mit anderen Worten gilt laut Vermutung:

$$y_i = \sum_{j=0}^k x_{ij} b_j + r_i$$

für $1 \leq i \leq n$,

$x_{i0} = 1$ (dies wird gesetzt, damit $x_{i0} b_0$ nicht durch Setzung 0 wird).

b_j sind die $k + 1$ Koeffizienten, die wir bestimmen wollen.

r_i ist das Residuum $y_i - \sum_{j=0}^k b_j x_{ij}$ des Datums y_i (Abweichung des Datums von der geschätzten Gerade).

Mit $A = (x_{ij})$ (= die Matrix der Werte der Variablen $X_0, X_1, \dots, X_k$)

$$b = (b_0, \dots, b_k)' \quad (v' \text{ ist der transponierte Vektor von } v; \text{ im Beispiel: } b = \begin{pmatrix} b_0 \\ \vdots \\ b_k \end{pmatrix})$$

$$y = (y_1, \dots, y_n)'$$

$$r = (r_1, \dots, r_n)'$$

kann man $y_i = \sum_{j=0}^k x_{ij} b_j + r_i$ für $0 \leq i \leq n$ auch in Matrizenschreibweise ausdrücken als

$$y = Ab + r$$

Auch hier geht es darum, die Summe der quadrierten Abstände zu minimieren,

d.h. die b_j ($0 \leq j \leq k$) sind so zu bestimmen, dass $\sum_{i=1}^n (y_i - \sum_{j=0}^k b_j x_{ij})^2$ minimal wird.

Man kann beweisen, dass dieser Ausdruck minimal ist genau dann, wenn

$$b = (A'A)^{-1}A'y$$

Dabei ist A' die transponierte Matrix von A . $(A'A)^{-1}$ ist die inverse Matrix von $A'A$ (Die inverse Matrix C einer Matrix B ist definiert als Matrix, für die gilt: $AC = E$, wobei E der Einheitsvektor ist - d.h. für E gilt: alle Zellen ausser der Diagonalen sind null, die Diagonale enthält überall eine 1). Man kann zeigen, dass $(A'A)^{-1}A'y$ ein erwartungstreuer Schätzer des Parametervektors der Verteilung ist, aus der die Daten stammen. Zudem ist dieser Schätzer unter allen linearen Schätzern von minimaler Varianz. Für diese Beweise müssen wir voraussetzen, dass A nicht als Realisation von Zufallsvariablen zu deuten ist, sondern als feste Punkte, unter denen wir die Werte Y erheben.

Berechnungsbeispiel: (es werden wenige Zahlen gewählt, damit die Rechnung mit manuellen Mitteln nachvollziehbar ist). Wir nehmen an, wir hätten die Vermutung, dass es zwischen dem Nettoeinkommen, der Schmuckgeschäftsdichte und dem Geschlecht einerseits (Variablen $X_1, \dots, X_3$) und den Ausgaben für Schmuck andererseits (Variable Y) einen linearen Zusammenhang gibt. Damit gilt für alle i : $y_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + r_i$. Zu schätzen sind die b_j ($0 \leq j \leq 3$).

Wir erheben die folgenden Daten.

Einkommen (steuerbares Nettoeinkommen pro Monat in Fr.)	Schmuckgeschäftsdichte (Geschäfte pro 10 000 Einwohner)	Geschlecht (1 weiblich, 0 männlich)	Ausgaben für Schmuck (pro Jahr in Franken)
5	4	1	160
3	3	1	140
2	1	0	130
1	2	0	100
3	2	1	110

Damit erhalten wir die Matrix und den Vektor:

$$A = \begin{pmatrix} 1 & 5 & 4 & 1 \\ 1 & 3 & 3 & 1 \\ 1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 1 & 3 & 2 & 1 \end{pmatrix}; y = \begin{pmatrix} 160 \\ 140 \\ 130 \\ 100 \\ 110 \end{pmatrix}$$

$$A'A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5 & 3 & 2 & 1 & 3 \\ 4 & 3 & 1 & 2 & 2 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 5 & 4 & 1 \\ 1 & 3 & 3 & 1 \\ 1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 1 & 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 14 & 12 & 3 \\ 14 & 48 & 39 & 11 \\ 12 & 39 & 34 & 9 \\ 3 & 11 & 9 & 3 \end{pmatrix}$$

$$(A'A)^{-1} = \begin{pmatrix} 5 & 14 & 12 & 3 \\ 14 & 48 & 39 & 11 \\ 12 & 39 & 34 & 9 \\ 3 & 11 & 9 & 3 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{53}{34} & -\frac{9}{34} & -\frac{15}{34} & \frac{25}{34} \\ -\frac{9}{34} & \frac{15}{34} & -\frac{9}{34} & -\frac{19}{34} \\ -\frac{15}{34} & -\frac{9}{34} & \frac{19}{34} & -\frac{9}{34} \\ \frac{25}{34} & -\frac{19}{34} & -\frac{9}{34} & \frac{83}{34} \end{pmatrix}$$

$$(A'A)^{-1}A' = \begin{pmatrix} \frac{53}{34} & -\frac{9}{34} & -\frac{15}{34} & \frac{25}{34} \\ -\frac{9}{34} & \frac{15}{34} & -\frac{9}{34} & -\frac{19}{34} \\ -\frac{15}{34} & -\frac{9}{34} & \frac{19}{34} & -\frac{9}{34} \\ \frac{25}{34} & -\frac{19}{34} & -\frac{9}{34} & \frac{83}{34} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5 & 3 & 2 & 1 & 3 \\ 4 & 3 & 1 & 2 & 2 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{27}{34} & \frac{3}{17} & \frac{10}{17} & \frac{7}{17} & \frac{21}{34} \\ \frac{11}{34} & -\frac{5}{17} & \frac{6}{17} & -\frac{6}{17} & -\frac{1}{34} \\ \frac{7}{34} & \frac{3}{17} & -\frac{7}{17} & \frac{7}{17} & -\frac{13}{34} \\ -\frac{23}{34} & \frac{12}{17} & -\frac{11}{17} & -\frac{6}{17} & \frac{33}{34} \end{pmatrix}$$

$$(A'A)^{-1}A'y = \begin{pmatrix} -\frac{27}{34} & \frac{3}{17} & \frac{10}{17} & \frac{7}{17} & \frac{21}{34} \\ \frac{11}{34} & -\frac{5}{17} & \frac{6}{17} & -\frac{6}{17} & -\frac{1}{34} \\ \frac{7}{34} & \frac{3}{17} & -\frac{7}{17} & \frac{7}{17} & -\frac{13}{34} \\ -\frac{23}{34} & \frac{12}{17} & -\frac{11}{17} & -\frac{6}{17} & \frac{33}{34} \end{pmatrix} \begin{pmatrix} 160 \\ 140 \\ 130 \\ 100 \\ 110 \end{pmatrix} = \begin{pmatrix} \frac{1415}{17} \\ \frac{305}{17} \\ \frac{55}{17} \\ -\frac{375}{17} \end{pmatrix} \approx \begin{pmatrix} 83.235 \\ 17.941 \\ 3.2353 \\ -22.059 \end{pmatrix} \approx b$$

Damit erhalten wir für alle i jeweils die Gleichung:

$$y_i = \frac{1415}{17} + \frac{305}{17}x_{i1} + \frac{55}{17}x_{i2} - \frac{375}{17}x_{i3} + r_i$$

wobei zur Schätzung von Werten $\hat{y}_i = \frac{1415}{17} + \frac{305}{17}x_{i1} + \frac{55}{17}x_{i2} - \frac{375}{17}x_{i3}$ verwenden. Die Koeffizienten b_0, b_1 und b_2 sind positiv. Für b_0 bedeutet dies, dass b_0 an Schmuck gekauft wird, wenn alle Variablen 0 sind (kein Einkommen, kein Geschäft auf 10'000 Einwohner, männlich; b_0 ist der Ordinatenabschnitt). Ein positives b_1 bedeutet, dass mehr Schmuck gekauft wird, wenn das Einkommen steigt. Dabei gibt b_1 die Steigung des entsprechenden Zusammenhangs an. Analoges gilt für b_2 . b_3 ist negativ. Nimmt x_{i3} den Wert 0 an (= Mann), so wird nichts abgezählt, nimmt x_{i3} den Wert 1 an (= Frau), so wird $\frac{375}{17} = 22.059$ von den zu schätzenden Ausgaben für Schmuck abgezählt.

Die Gleichung können wir nun verwenden, um zu schätzen, wieviel jemand für Schmuck ausgibt, wenn wir wissen, wieviel er verdient, welches die Schmuckgeschäftsdichte seines Wohnortes ist und welches Geschlecht er hat.

z.B. Jemand verdiene 3000 pro Monat, die Dichte sei 2 und das Geschlecht sei männlich. Dies ergibt in die Gleichung eingesetzt:

$$\hat{y}_i = \frac{1415}{17} + \frac{305}{17} \cdot 3 + \frac{55}{17} \cdot 2 - \frac{375}{17} \cdot 0 = 143.53 \text{ Fr.}$$

Diese Person wird also im Mittel 143.53 Franken für Schmuck ausgeben.

Wäre die Person eine Frau, würde sich ergeben:

$$\hat{y}_i = \frac{1415}{17} + \frac{305}{17} \cdot 3 + \frac{55}{17} \cdot 2 - \frac{375}{17} \cdot 1 = \frac{2065}{17} = 121.47 \text{ Fr.}$$

Berechnung mit Excel

Sie finden das folgende Beispiel unter "multivariate Regression mit Excel.xls" auf der Homepage. Eingabe der Datenmatrix A in Excel:

8	1	5	4	1
9	1	3	3	1
10	1	2	1	0
11	1	1	2	0
12	1	3	2	1

Berechnung der transponierten Matrix A' (die transponierte Matrix ist die Matrix, die entsteht, wenn wir eine Matrix an der Diagonale spiegeln): A kopieren. Auf die Zelle gehen, von der weg nach links-unten die transponierte Matrix eingefügt werden soll (im Beispiel A16). Anklicken. Dann "bearbeiten", "Inhalte einfügen", transponiert, ok. Wir erhalten fürs Beispiel:

16	1	1	1	1	1
17	5	3	2	1	3
18	4	3	1	2	2
19	1	1	0	0	1

Als nächstes berechnen wir $A'A$ mit Hilfe des Excelbefehles "mmult(A';A)": Dazu müssen wir vorgängig eine leeren Teil der Tabelle beleuchten (mit $(k+1) \times (k+1)$ Zellen, im Beispiel mit 4×4 Zellen und A23:D26) und schreiben dann fürs Beispiel

"mmult(A16:E19;A8:D12)" (Sie können auch die Matrizen eingeben, indem Sie die erste mit dem Cursor beleuchten, ";" eingeben, die zweite mit dem Cursor beleuchten) und bestätigen mit ctrl + shift + enter. Es erscheint die Matrix $A'A =$

23	5	14	12	3
24	14	48	39	11
25	12	39	34	9
26	3	11	9	3

Wir berechnen nun die inverse Matrix $(A'A)^{-1}$ dieser Matrix: Dazu muss ein leerer Teil der Exceltabelle beleuchtet werden, der gleichviele Zeilen und Spalten hat wie die obige Datenmatrix (im Beispiel 4×4 ; Zellen A30:E34): Dann wird "minv(A23:D26)" (im Beispiel) geschrieben und bestätigt mit ctrl + shift + enter. Wir erhalten $(A'A)^{-1} =$

30	1.558823529	-0.264705882	-0.44117647	0.73529412
31	-0.264705882	0.441176471	-0.26470588	-0.55882353
32	-0.441176471	-0.264705882	0.55882353	-0.26470588
33	0.735294118	-0.558823529	-0.26470588	2.44117647
34				

Nun berechnen wir die Multiplikation (wie bereits beschrieben) dieser Matrix mit der A' , indem wir in einem leeren Teil der Tabelle $(k+1) \times n$ Zellen beleuchten (im Beispiel A37:E40). Wir erhalten: $(A'A)^{-1}A' =$

37	-0.794117647	0.176470588	0.58823529	0.41176471	0.61764706
38	0.323529412	-0.294117647	0.35294118	-0.35294118	-0.02941176
39	0.205882353	0.176470588	-0.41176471	0.41176471	-0.38235294
40	-0.676470588	0.705882353	-0.64705882	-0.35294118	0.97058824

Schliesslich müssen wir $(A'A)^{-1}A'y$ berechnen, wobei $b =$

45	160
46	140
47	130
48	100
49	110

(Multiplikation von $(A'A)^{-1}A'$ mit y mit dem Befehl mmult(A37:E40;A45:A49) (im Beispiel), bestätigen mit: ctrl, shift, enter.

Wir erhalten:

53	83.23529412
54	17.94117647
55	3.235294118
56	-22.05882353

{Ende der Erläuterungen zur Excelberechnung}