Multiple lineare Regression

In diesem Skript sollen die theoretischen Grundlagen zur multiplen linearen Regression sowie deren Umsetzung in MS-Excel dargestellt werden.

Für die Unterstützung bei der Erstellung dieses Skripts gilt Herrn Prof. Dr. W. Kristof, Ph.D., mein besonderer Dank.

1. Vorüberlegungen

Ein lineares Gleichungssystem $\underline{A}\underline{x} = \underline{b}$ heißt homogen, wenn $\underline{b} = \underline{0}$. Ist $\underline{b} \neq \underline{0}$, dann heißt ein solches Gleichungssystem inhomogen.

Ein homogenes Gleichungssystem hat nur dann eine Lösung $\underline{x} \neq \underline{0}$, wenn $\underline{A}$ nicht vollen Rang hat. Dies kann man sich verdeutlichen, wenn man sich die Definition der linearen Unabhängigkeit anschaut[1].

Ein inhomogenes Gleichungssystem ist genau dann lösbar, wenn $\underline{b}$ im Raum liegt, der durch die Spaltenvektoren von $\underline{A}$ aufgespannt wird. Dann ist Rang $\underline{A}$ = Rang $(\underline{A}|\underline{b})$.

Ist $\underline{A}$ quadratisch und hat vollen Rang, dann hat $\underline{A}\underline{x} = \underline{b}$ die eindeutige Lösung

(1) $\underline{x} = \underline{A}^{-1}\underline{b}$.

Im Fall der multiplen Regression werden inhomogene Gleichungssysteme betrachtet, bei denen $\underline{A}$ vom Format (n, m) ist, $n \geq m$ und Rang $\underline{A}$ = m. In der Regel liegt $\underline{b}$ nicht im Raum, der von den Spaltenvektoren von $\underline{X}$ aufgespannt wird. Ein solches System ist dann inkonsistent, denn es existiert keine exakte Lösung.

[1] m Vektoren $\underline{a}_1, \ldots, \underline{a}_m$ sind linear unabhängig, wenn $q_1\underline{a}_1 + q_2\underline{a}_2 + \ldots + q_m\underline{a}_m = \underline{0}$ nur durch $q_j = 0$ für alle q_j erfüllt werden kann.

Wir gehen jedoch zunächst davon aus, daß ein solches Gleichungssystem $\underline{A}x = \underline{b}$ (mit $\underline{A}$ vom Format (n, m), $n \geq m$ und Rang $\underline{A} = m$) eine exakte Lösung hat.

Multiplikation der Gleichung $\underline{A}x = \underline{b}$ von links mit $\underline{A}'$ liefert

(2) $\underline{A}'\underline{A}x = \underline{A}'\underline{b}$.

$\underline{A}'\underline{A}$ ist quadratisch und vom Format (m, m). Aus der linearen Algebra ist bekannt, daß Rang $\underline{A}$ = Rang $\underline{A}'$ = Rang $\underline{A}'\underline{A}$. Folglich ist Rang $\underline{A}'\underline{A} = m$ und damit $\underline{A}'\underline{A}$ invertierbar. Gleichung (2) von links mit $(\underline{A}'\underline{A})^{-1}$ multipliziert liefert

$(\underline{A}'\underline{A})^{-1}\underline{A}'\underline{A}x = (\underline{A}'\underline{A})^{-1}\underline{A}'\underline{b}$. Demnach ist

(3) $\underline{x} = (\underline{A}'\underline{A})^{-1}\underline{A}'\underline{b}$.

Dies ist die (exakte) Lösung für das Gleichungssystem, sofern eine Lösung existiert[2].

Im folgenden soll die „Lösung“ eines inkonsistenten und inhomogenen Gleichungssystems gesucht werden. In diesem Fall liegt $\underline{b}$ nicht im Raum, der in den Spaltenvektoren von $\underline{A}$ aufgespannt wird. Gesucht ist die „beste Lösung im Sinn der kleinsten Quadrate“ – dies wird durch das Symbol „$\doteq$“ ausgedrückt - für die x_j in folgendem Gleichungssystem:

$$\begin{array}{rcccccc} a_{11}x_1 & + & \cdots & + & a_{1m}x_m & \doteq & b_1 \\ \vdots & & & & & & \\ a_{n1}x_1 & + & \cdots & + & a_{nm}x_m & \doteq & b_n \end{array} \quad (4)$$

Man verwendet

$$\underline{A} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}, \underline{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \text{ mit Rang } \underline{A} = m \text{ und } n \geq m. \quad (5)$$

[2] Dies ist gleichzeitig die allgemeine Lösung für alle konsistenten Gleichungssysteme. Falls $\underline{A}$ quadratisch, ist $\underline{x} = (\underline{A}'\underline{A})^{-1}\underline{A}'\underline{b} = \underline{A}^{-1}\underline{A}'^{-1}\underline{A}'\underline{b} = \underline{A}^{-1}\underline{b}$, wie bereits in Gleichung (1) dargestellt.

Folglich ist das Gleichungssystem (4) in Matrizenschreibweise

(6) $\underline{Ax} \doteq \underline{b}$.

Ein inkonsistentes Gleichungssystem hat keine exakte Lösung, also fordert man für

(7)
$$\begin{matrix} a_{11}x_1 & + & \cdots & + & a_{1m}x_m & - & b_1 & = & d_1 \\ \vdots & & & & & & & & \\ a_{n1}x_1 & + & \cdots & + & a_{nm}x_m & - & b_n & = & d_n \end{matrix},$$

daß $d_1^2 + ... + d_n^2$ durch die Wahl der x_j minimal werde. Man steht somit vor dem Problem, die Funktion aller x_j,

(8) $f(x_1, ..., x_m) = (a_{11}x_1 + ... + a_{1m}x_m - b_1)^2 + ... + (a_{n1}x_1 + ... + a_{nm}x_m - b_n)^2$

zu minimieren. Minimierung von $f(x_1, ..., x_m)$ bedeutet

$$\frac{\partial f(x_1,...,x_m)}{\partial x_j} = 0, \qquad j = 1, ..., m.$$

$$\frac{\partial f}{\partial x_j} = 2(a_{11}x_1 + ... + a_{1m}x_m - b_1)a_{1j} + ... + 2(a_{n1}x_1 + ... + a_{nm}x_m - b_n)a_{nj} = 0$$

(9) $$\Leftrightarrow (a_{11}a_{1j} + ... + a_{n1}a_{nj})x_1 + ... + (a_{1m}a_{1j} + ... + a_{nm}a_{nj})x_m = b_1a_{1j} + ... + b_na_{nj} = 0.$$

Dies ergibt, j = 1, ..., m, eingesetzt, folgende m Gleichungen:

(10) $$(a_{11}a_{11} + ... + a_{n1}a_{n1})x_1 + ... + (a_{1m}a_{11} + ... + a_{nm}a_{n1})x_m = b_1a_{11} + ... + b_na_{n1}$$

$$\vdots$$

$$(a_{11}a_{1m} + ... + a_{n1}a_{nm})x_1 + ... + (a_{1m}a_{1m} + ... + a_{nm}a_{nm})x_m = b_1a_{1m} + ... + b_na_{nm}$$

Dies ist

$$(11)\quad \underbrace{\begin{pmatrix} a_{11}^2+\ldots+a_{n1}^2 & \cdots & a_{1m}a_{11}+\ldots+a_{nm}a_{n1} \\ \vdots & & \vdots \\ a_{11}a_{1m}+\ldots+a_{n1}a_{nm} & \cdots & a_{1m}^2a_{nm}^2 \end{pmatrix}}_{\underline{A}'\underline{A}} \underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}}_{\underline{x}} = \underbrace{\begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & & \vdots \\ a_{1m} & \cdots & a_{nm} \end{pmatrix}}_{\underline{A}'} \underbrace{\begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}}_{\underline{b}}, \text{ also}$$

$$(12)\quad \underline{A}'\underline{A}\underline{x} = \underline{A}'\underline{b}, \quad \text{daher}$$

$$(13)\quad \underline{x} = (\underline{A}'\underline{A})^{-1}\,\underline{A}'\underline{b}, \qquad \text{mit Format } (\underline{A}'\underline{A}) = (m, m), \text{ Rang } (\underline{A}'\underline{A}) = m$$
$$\text{und } (\underline{A}'\underline{A})^{-1} \text{ existent nach Voraussetzung.}$$

Gleichung (13) ist also identisch mit der Gleichung (3). Dies ist ein allgemeines Ergebnis. Es ist die (exakte) Lösung für $\underline{x}$, sofern das Gleichungssystem lösbar ist, andernfalls die „beste Lösung“ für $\underline{x}$ im Sinn der kleinsten Quadrate.

2. Multiple Regression

Es ist zu untersuchen, wie der Zusammenhang einer Variablen Y mit den Variablen X_1, ..., X_m aussieht.

Die Variable Y wird in diesem Fall als Kriterium oder *Regressand* bezeichnet, die Variablen X_i als Prädiktoren oder *Regressoren*.

Kann ein solcher Zusammenhang beschrieben werden, so können folgende Aufgaben durchgeführt werden:

- Erkennen eines statistischen Zusammenhangs
- Schätzung der Parameter eines statistischen Zusammenhangs
- Trendanalysen
- Prognose von Y aus X_1, ..., X_m

Gegeben sind die Variablen Y (Regressand) sowie die $X_1, ..., X_m$ (Regressoren).

Der Regressand hat metrisches Skalenniveau, die Prädiktoren haben in der Regel ebenfalls metrisches Skalenniveau, können aber auch dichotome Variablen sein. Werden zusätzlich statistische Testverfahren angewendet, müssen bestimmte Normalverteilungsannahmen erfüllt sein, die jedoch von Fall zu Fall verschieden sind. Diese Verteilungsannahmen sind jedoch keine Voraussetzung für die Durchführung der multiplen Regression.

In vielen Fällen wird zu den Variablen eine Konstante b_0 hinzugenommen. Die Aufgabenstellung ist analog zu (4) die Annäherung von Y im Sinn der kleinsten Quadrate durch Linearkombination $b_o + b_1X_1 + ... + b_mX_m$, also für tatsächliche Meßwerte

$$(14) \quad \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \stackrel{\circ}{=} \begin{pmatrix} b_0 + b_1x_{11} & + \cdots + & b_mx_{1m} \\ \vdots & & \vdots \\ b_0 + b_1x_{n1} & + \cdots + & b_mx_{nm} \end{pmatrix},$$

wir betrachten also n Objekte (Personen) und m Variablen (Regressoren). Die Bezeichung der Variablen entspricht der allgemein üblichen Bezeichnung im Falle der multiplen Regression - im Gegensatz zur Variablenbezeichnung in den Vorüberlegungen sind also hier nicht die x_j unbekannt bzw. gesucht, sondern die b_j!

Gleichung (14) aufgelöst nach b_0 ergibt

$$(15) \quad \begin{pmatrix} b_0 \\ \vdots \\ b_0 \end{pmatrix} \stackrel{\circ}{=} \begin{pmatrix} y_1 - b_1x_{11} & - \cdots - & b_mx_{1m} \\ \vdots & & \vdots \\ y_n - b_1x_{n1} & - \cdots - & b_mx_{nm} \end{pmatrix}.$$

Aus der Statistik 1 ist bekannt, daß die Zahl, die die kleinste Abweichung im Sinn der kleinsten Quadrate von einer Menge von Werten darstellt, das arithmetische Mittel dieser Werte ist. Also ist

$$(16) \quad b_0 = \bar{y} - b_1\bar{x}_1 - ... - b_m\bar{x}_m$$

Aus (14), (15) und (16) folgt, daß

$$(17)\quad \begin{pmatrix} y_1-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix} \overset{\circ}{=} \begin{pmatrix} b_1(x_{11}-\bar{x}_1) & + \cdots + & b_m(x_{1m}-\bar{x}_m) \\ \vdots & & \vdots \\ b_1(x_{n1}-\bar{x}_1) & + \cdots + & b_m(x_{nm}-\bar{x}_m) \end{pmatrix}.$$

Wendet man die Lösung (13) auf das Gleichungssystem (17) an, so erhält man – unter Beachtung der unterschiedlichen Bezeichnungsweise – die Lösung

$$(18)\quad \underline{b} = \left[\begin{pmatrix} x_{11}-\bar{x}_1 & \cdots & x_{n1}-\bar{x}_1 \\ \vdots & & \vdots \\ x_{1m}-\bar{x}_m & \cdots & x_{nm}-\bar{x}_m \end{pmatrix}\begin{pmatrix} x_{11}-\bar{x}_1 & \cdots & x_{1m}-\bar{x}_m \\ \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & \cdots & x_{nm}-\bar{x}_m \end{pmatrix}\right]^{-1}\begin{pmatrix} x_{11}-\bar{x}_1 & \cdots & x_{1m}-\bar{x}_m \\ \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & \cdots & x_{nm}-\bar{x}_m \end{pmatrix}\begin{pmatrix} y_1-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix}$$

Man erkennt, daß es sich bei dem Ausdruck

$$\left[\begin{pmatrix} x_{11}-\bar{x}_1 & \cdots & x_{n1}-\bar{x}_1 \\ \vdots & & \vdots \\ x_{1m}-\bar{x}_m & \cdots & x_{nm}-\bar{x}_m \end{pmatrix}\begin{pmatrix} x_{11}-\bar{x}_1 & \cdots & x_{1m}-\bar{x}_m \\ \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & \cdots & x_{nm}-\bar{x}_m \end{pmatrix}\right]$$

um das n-fache der Kovarianzmatrix der Regressoren handelt und bei dem Ausdruck

$$\begin{pmatrix} x_{11}-\bar{x}_1 & \cdots & x_{1m}-\bar{x}_m \\ \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & \cdots & x_{nm}-\bar{x}_m \end{pmatrix}\begin{pmatrix} y_1-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix}$$ um das n-fache des Kovarianzvektors der n Regressanden

mit dem Regressor.

Praktisch kann man die Konstante jedoch relativ einfach berücksichtigen, indem man eine Variable x_j, die nur aus Einsen besteht, in den Satz der Variablen aufnimmt. Die Gleichungen werden dann rechnerisch so behandelt, als wäre keine Konstante vorhanden (siehe Abschnitt 6.2).

Im Falle von standardisierten Variablen ist die Kovarianzmatrix gleich der Korrelationsmatrix. In diesem Fall gilt

$$(19)\quad \underline{b} = \underline{R}^{-1}\underline{r} \qquad \text{bzw.} \quad \underline{r} = \underline{R}\underline{b}.$$

3. Geometrischer Ansatz

Einem Ansatz von Kristof folgend, läßt sich die Aufgabe auch geometrisch lösen. Der Vorteil der geometrischen Lösung besteht darin, daß man allein mit Mitteln der linearen Algebra auskommt.

Im Falle eines inkonsistenten Systems ist die Lösung im Sinn der kleinsten Quadrate für $\underline{b}$ in der Gleichung

(20) $\underline{Xb} \stackrel{\circ}{=} \underline{y}$,

gesucht. Die Abweichungen der Werte der Variablen Y von den Werten, die durch $\underline{Xb}$ geschätzt werden, wird angegeben durch den Vektor $\underline{\varepsilon}$

(21) $\underline{\varepsilon} = \underline{Xb} - \underline{y}$.

Die nichtnegative skalare Funktion $f(\underline{b}) = \underline{\varepsilon}'\underline{\varepsilon}$ soll so klein wie möglich werden; der Vektor $\underline{b}$ sei die Lösung. Es sei

(22) $\underline{y} = \underline{g} + \underline{h}$,

wobei $\underline{g}$ im durch die Spaltenvektoren von $\underline{X}$ aufgespannten Raum liege und $\underline{h}$ senkrecht dazu. Also ist

(23) $\underline{X}'\underline{h} = \underline{0}$ und $\underline{g}'\underline{h} = \underline{0}$.

Nun wird

$$
\begin{aligned}
f(\underline{b}) = \underline{\varepsilon}'\underline{\varepsilon} &= (\underline{Xb} - \underline{g}-\underline{h})'(\underline{Xb} - \underline{g}-\underline{h}) \\
&= ((\underline{Xb} - \underline{g})'-\underline{h}')(\underline{Xb} - \underline{g}-\underline{h}) \\
&= (\underline{Xb}-\underline{g})'(\underline{Xb}-\underline{g}) - (\underline{Xb}-\underline{g})'\underline{h} - \underline{h}'(\underline{Xb}-\underline{g}) + \underline{h}'\underline{h} \\
&= (\underline{Xb}-\underline{g})'(\underline{Xb}-\underline{g}) - 0 - 0 + \underline{h}'\underline{h}, \text{ also}
\end{aligned}
$$

(24) $f(\underline{b}) = (\underline{Xb}-\underline{g})'(\underline{Xb}-\underline{g}) + \underline{h}'\underline{h}.$

Folglich wird $f(\underline{b})$ minimal, wenn $(\underline{Xb}-\underline{g})$ Null gesetzt wird. Dies wird erreicht, wenn $\underline{Xb} = \underline{g}$. Der Ausdruck ist wiederum die exakte Lösung nach (3). Dies ist möglich, da $\underline{g}$ im durch die Spaltenvektoren von $\underline{X}$ aufgespannten Raum liegt. Hierfür ist

(25) $\underline{b} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{g}.$

Da nach (23) $\underline{X}'\underline{h} = \underline{0}$, gilt auch

(26) $\underline{0} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{h}.$

Addiert man die Gleichungen (25) und (26), so erhält man

$\underline{b} + \underline{0} = (\underline{X}'\underline{X})^{-1}\underline{X}'(\underline{g} + \underline{h})$, bzw.

(27) $\underline{b} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}$

Sofern eine exakte Lösung existiert, ist dieses $\underline{b}$ die Lösung von $\underline{Xb} = \underline{y}$, andernfalls ist es die Lösung von $\underline{Xb} \hat{=} \underline{y}$, also die Lösung im Sinn der kleinsten Quadrate (vgl. (13)).

Unter (22) wurde der Vektor $\underline{y}$ zerlegt in die Vektoren $\underline{g} + \underline{h}$, wobei $\underline{g}$ im durch die Spaltenvektoren von $\underline{X}$ aufgespannten Raum lag und $\underline{h}$ senkrecht dazu. Man kann nun $\underline{g}$ verwenden, um ein Maß zu bilden, das angibt, welcher Anteil des Vektors $\underline{y}$ durch die Lösung $\underline{Xb} \hat{=} \underline{y}$ erfaßt wird. Der Quotient

(28) $$\frac{\underline{g}'\underline{g}}{\underline{y}'\underline{y}}$$

gibt hierbei den erfaßten Anteil von $\underline{y}$ an, $\frac{\underline{h}'\underline{h}}{\underline{y}'\underline{y}}$ entsprechend den nicht erfaßten Anteil.

4. Modellgüte

Gesucht ist ein Maß, das angibt, wie gut das Modell die gegebenen Kriterien Y schätzt.

Hierfür wird die Produkt-Moment-Korrelation zwischen dem Kriterium Y und den $b_0 + b_1X_1 + \ldots + b_mX_m$ betrachtet, also

(29) $$R = \frac{S_{Y, b1X1 + ... + bmXm}}{S_y S_{b1X1 + ... + bmXm}}$$ bzw.

(30) $$R^2 = \frac{S^2_{Y, b1X1 + ... + bmXm}}{S^2_y S^2_{b1X1 + ... + bmXm}}.$$ [3]

Man spricht hier vom multiplen R bzw. R^2. R^2 ist der Anteil der gemeinsamen Varianz zwischen der Kriteriumsvariablen und den Prädiktorvariablen. R^2 ist eine Schätzung des Teils der Varianz der Kriteriumsvariablen, der durch die Prädiktorvariablen vorhergesagt werden kann. Das multiple R^2 entspricht dem Ausdruck (28) (ohne Beweis).

Möchte man eine Variable mittels der multiplen Regression in der Praxis vorherzusagen, sollte mindestens ein R^2 von 0,7 erreicht werden.

Die Überprüfung des ermittelten R^2 auf statistische Signifikanz gegenüber der Hypothese der Unabhängigkeit von Kriterium und Prädiktoren erfolgt – ohne Herleitung - durch einen F-Test:

(31) $$F = \frac{R^2(n-k-1)}{(1-R^2)m}$$, $df_{Zähler} = m$, $df_{Nenner} = n - m - 1$.

[3] Die Konstante b_0 entfällt, da die Produkt-Moment-Korrelation invariant gegenüber einer linearen Transformation ist.

5. Schrittweise Regression

Um zu entscheiden, welche Prädiktoren einen nennenswerten Beitrag für die Gesamtgüte leisten, wird zunächst noch einmal der multiple Korrelationskoeffizient R ($0 \leq R \leq 1$) betrachtet.

Wird die Kriteriumsvariable Y durch nur zwei Prädiktorvariablen (1,2) vorhergesagt, so ist

(32) $R_{Y,12} = \sqrt{b_1 r_{1Y} + b_2 r_{2Y}}$, bzw. allgemein bei m Prädiktorvariablen

(33) $R_{Y,12...m} = \sqrt{\sum_j b_j r_{jY}}$, dieser Ausdruck entspricht dem multiplen R,

wobei die b_j die *standardisierten* Koeffizienten bezeichnen.

Die Matrix $\underline{R}$ ist die Matrix, die die Produkt-Moment-Korrelationen zwischen den Prädiktorvariablen enthält, $\underline{r}$ sei der Vektor der Korrelationen der Prädiktoren mit dem Kriterium. Nach (19) gilt für standardisierte Koeffizienten $\underline{b}$:

$\underline{r} = \underline{R}\underline{b}$ bzw.

(34) $\underline{b} = \underline{R}^{-1}\underline{r}$,

Man kann die Gleichung (34) verwenden, um standardisierte Koeffizienten aus nicht standardisierten zu gewinnen.

Die Prüfung der Regressionskoeffizienten auf Signifikanz erfolgt unter Verwendung standardisierter Koeffizienten $\underline{b}$ mittels der Prüfgröße t, hier (ohne Herleitung)

(35) $t = \frac{b_j}{\sqrt{\frac{r_{jj}(1-R^2)}{n-m-1}}}$, $df = n - m - 1$,

wobei r^{jj} das Element (j, j) der invertierten Korrelationsmatrix $\underline{R}^{-1}$ ist.

Nachdem man die standardisierten Regressionskoeffizienten auf Signifikanz geprüft hat, bildet man (für signifikante Korrelationen) die Produkte aus den standardisierten Regressionskoeffizienten und den Korrelationen der entsprechenden Prädiktoren mit dem Kriterium

(36) $b_j r_j$.

Man kann nun zunächst diejenige Variable X_j mit signifikantem b_j aufnehmen, die den größten Beitrag $b_i r_i$ leistet. Anschließend nimmt man weitere Variablen mit signifikantem b_j in das Regressionsmodell auf, bis der Zugewinn an Güte (R^2) gegenüber dem vorherigen Modell einen zuvor festgelegten Wert (z.B. 0,05) nicht mehr übersteigt. Umgekehrt kann man zunächst von einem Modell ausgehen, daß alle m Prädiktorvariablen enthält und dann die Variablen mit nicht-signifikanten bzw. zu geringen Beiträgen ausschließen. Man spricht hier von einer „Rückwärtstechnik“, im ersten Fall von einer „Vorwärtstechnik“.

Es ist jedoch zu beachten, daß dieses Verfahren in der Praxis zu Problemen führen kann. Für Modelle, die nur einen Teil der m Prädiktorvariablen enthalten, ergeben sich in der Regel andere Gewichte $b_j r_j$ als bei Einbeziehung aller m Variablen! Die Abweichung ist um so stärker, je stärker die Variablen X_j voneinander abhängig sind (Problem der Multikollinearität).

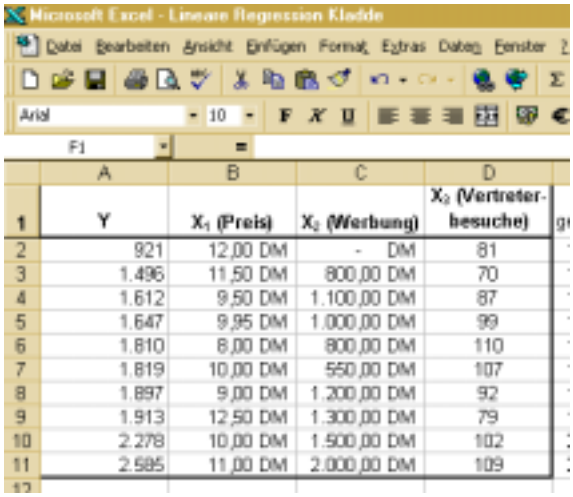
Die Verfahren der schrittweisen Regression sind daher keine exakten Verfahren, sondern müssen als explorative Verfahren verstanden werden.

6. Beispiel

Die Berechnung des Beispiels erfolgt auf Basis von MS-Excel. Die Durchführung erfolgt „manuell“ im Arbeitsblatt – die Erstellung von Makros bzw. VBA-Algorithmen wäre vorteilhaft, soll aber an dieser Stelle nicht explizit vorgestellt werden.

6.1 Ausgangsdaten

Gegeben sind folgende Ausgangsdaten:



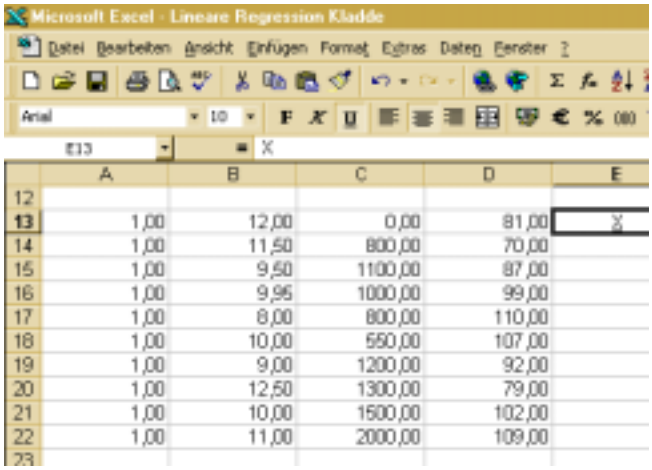
	A	B	C	D
1	Y	X_1 (Preis)	X_2 (Werbung)	X_2 (Vertreter-besuche)
2	921	12,00 DM	- DM	81
3	1.496	11,50 DM	800,00 DM	70
4	1.612	9,50 DM	1.100,00 DM	87
5	1.647	9,95 DM	1.000,00 DM	99
6	1.810	8,00 DM	800,00 DM	110
7	1.819	10,00 DM	550,00 DM	107
8	1.897	9,00 DM	1.200,00 DM	92
9	1.913	12,50 DM	1.300,00 DM	79
10	2.278	10,00 DM	1.500,00 DM	102
11	2.585	11,00 DM	2.000,00 DM	109

Zu bestimmen ist eine Gleichung, die den Umsatz vorhersagt. (Die Tatsache, daß es sich bei n = 10 Daten in der Praxis um eine recht geringe Menge handelt, sei außer Acht gelassen).

6.2 Bestimmung der Matrix $\underline{X}$

Anders als in Abschnitt 2 dargestellt, ist die erste Spalte der Matrix $\underline{X}$ ein Vektor, der nur aus Einsen besteht. Hierdurch erreicht man, daß die Zahl b_0 direkt im entsprechenden Lösungsvektor $\underline{b}$ steht und nicht über Gleichung (16) bestimmt werden muß.

Man kopiere die Inhalte der Zellen (b2:d11) in den Bereich (b13:d22) und füge Einsen in die Zellen (a13:a22) ein:

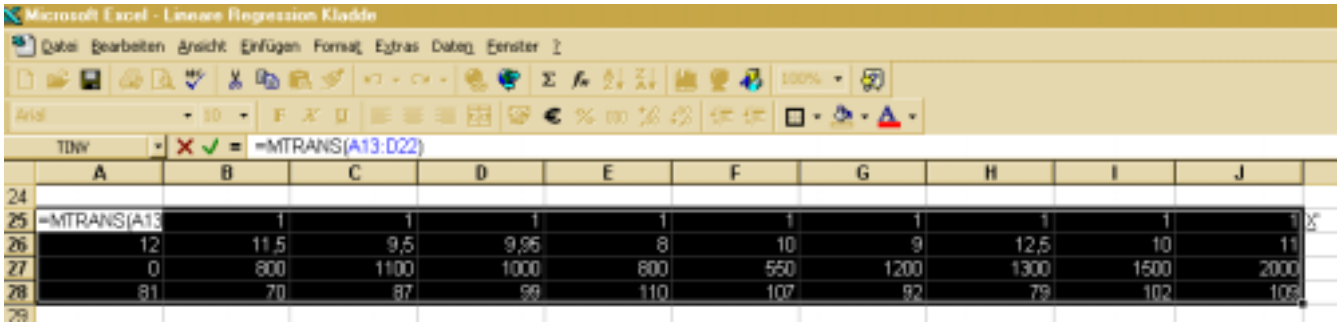


	A	B	C	D	E
12					
13	1,00	12,00	0,00	81,00	X
14	1,00	11,50	800,00	70,00	
15	1,00	9,50	1100,00	87,00	
16	1,00	9,95	1000,00	99,00	
17	1,00	8,00	800,00	110,00	
18	1,00	10,00	550,00	107,00	
19	1,00	9,00	1200,00	92,00	
20	1,00	12,50	1300,00	79,00	
21	1,00	10,00	1500,00	102,00	
22	1,00	11,00	2000,00	109,00	
23					

Dieser Bereich ist die Marix X; der Bereich (a2:a11) entspricht dem Vektor y.

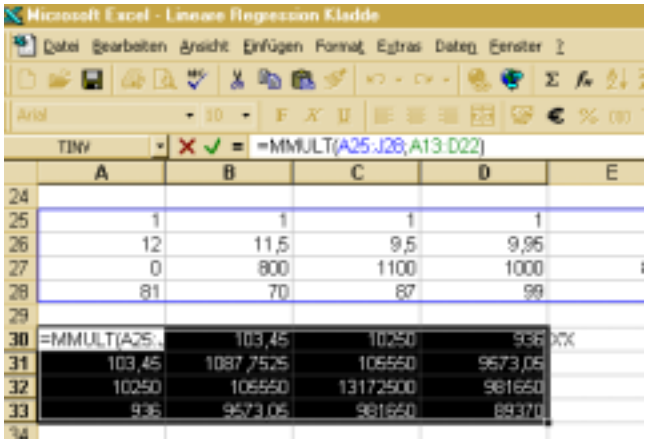
6.3 Transponieren der Matrix X, Bildung der Matrix X´X

Man Markiere die Zellen (a25:j28). Nun ist die Formel „=MTRANS(a13:d22)“ einzugeben. Damit die Formel für den gesamten Bereich der Matrix eingegeben wird, muß die Eingabe mit der Tastenkombination Strg-Shift-Enter erfolgen:



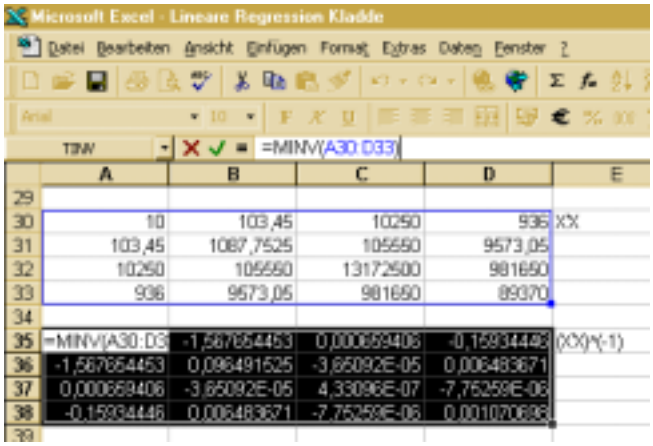
	A	B	C	D	E	F	G	H	I	J
24										
25	=MTRANS(A13	1	1	1	1	1	1	1	1	1
26	12	11,5	9,5	9,95	8	10	9	12,5	10	11
27	0	800	1100	1000	800	550	1200	1300	1500	2000
28	81	70	87	99	110	107	92	79	102	109
29										

Um die Matrix X´X zu bilden, muß der Bereich (a30:d33) markiert werden und hier das Produkt der beiden Matrizen X´ und X hineingeschrieben werden. Der Befehl lautet „=mmult(a25:j28;a13:d22)“. Die Eingabe – wie alle Eingaben von Matrix-Formeln – mit Strg-Shift-Enter bestätigen!



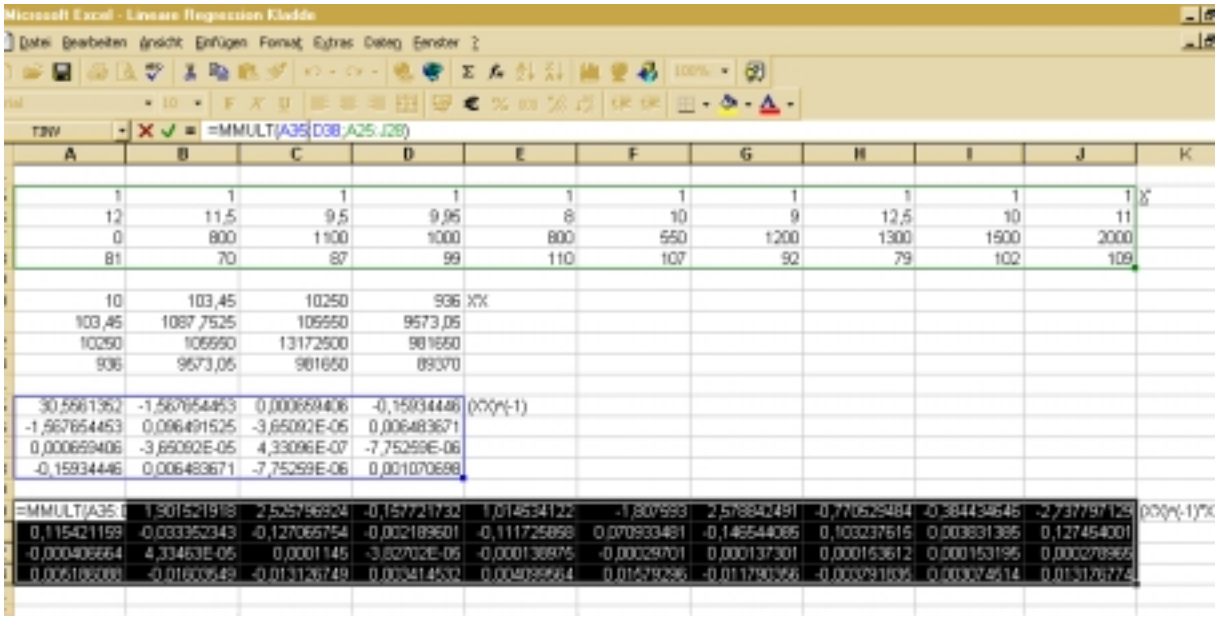
6.4 Bestimmung der Inversen von X´X

Markierung des Bereichs (a:35;d38). Die Formel zur Berechnung der Inversen lautet „=MINV(a30;d33)“.

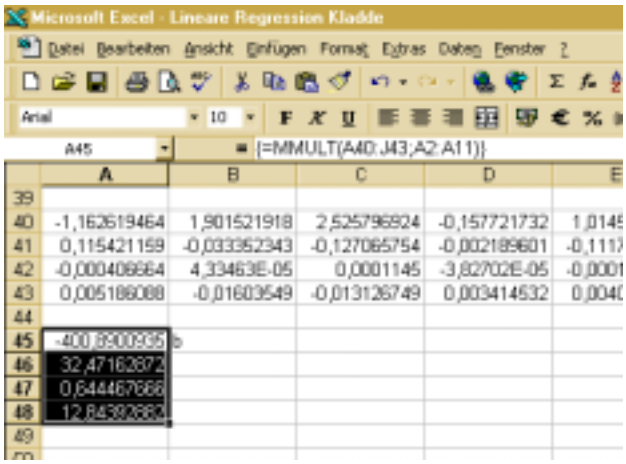


6.5 Bestimmung des Lösungsvektors b

Die Berechnung des Lösungsvektors b muß in MS-Excel in zwei Schritten erfolgen, da sich nur jeweils zwei Matrizen pro Operation multiplizieren lassen. Zunächst bilde man die Matrix $(X´X)^{-1}X´$, indem man in den Bereich (a40:j43) die folgende Formel einträgt („=mmult(a35:d38;a25;j28)“):



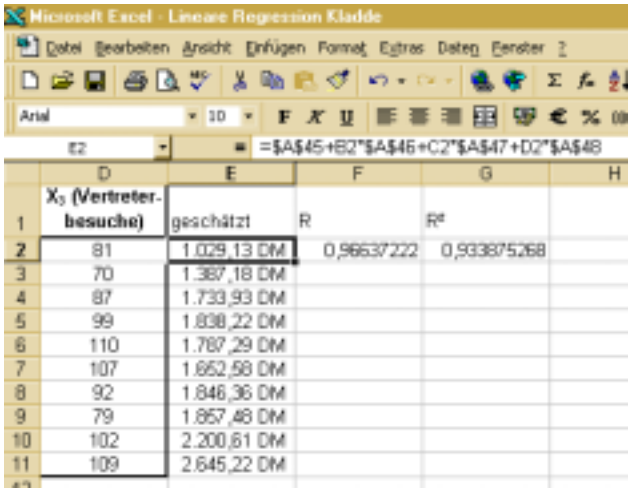
Anschließend wird dieses Matrizenprodukt mit dem Vektor y multipliziert („=mmult(a40:j43;a2:a11)“):



6.6 Bestimmung der Güte

Man kann nun die Schätzwerte der Variablen Y bestimmen, indem man in die Zellen (e2:e11) die entsprechende Formel („=a45+b2*a46+c2*a47+d2*a48“) einträgt. Wird die folgende Formel in die erste Zelle eingetragen, so kann man sie einfach in die folgenden Zellen kopieren. Die Eingabe von „$“ bewirkt, daß es sich für die entsprechenden Zellen um absolute Bezüge handelt.

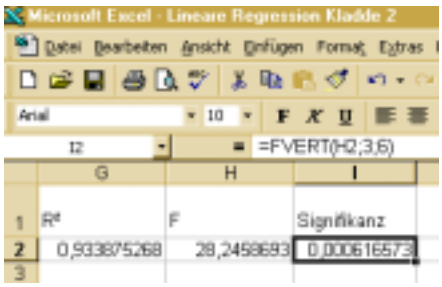
Man kann das R bzw. R² (hier in den Zellen f2 bzw. g2) bestimmen, indem man die Produkt-Moment-Korrelation der Reihe der Schätzwerte (e2:e11) und der Ausgangswerte (a2:a11) berechnen läßt. Der entsprechende Excel-Befehl lautet („=korrel(e2:e11;a2:a11)“). R² erhält man durch Quadrieren der Zelle f2 („=f2^2“):



6.7 Überprüfung des R² auf Signifikanz

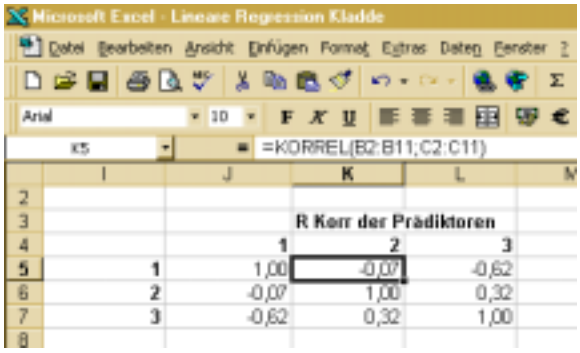
Die Überprüfung des R² auf Signifikanz erfolgt nach (31). In Excel wird der F-Wert mit der entsprechenden Formel in Zelle h2 berechnet:

Die Überprüfung von F auf Signifikanz erfolgt in Excel durch die Funktion „Fvert(F; $df_{Zähler}$; df_{Nenner})“ (Zelle i2). Die Prüfgröße ist mit 0,00062 hoch signifikant.

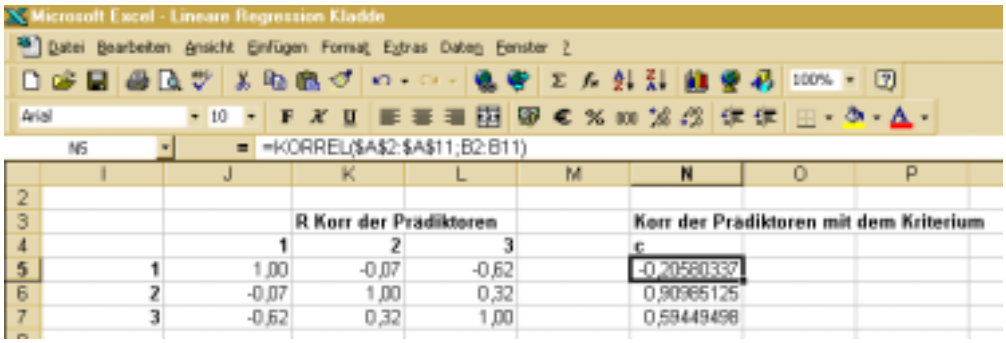


6.8 Bestimmung der standardisierten Koeffizienten

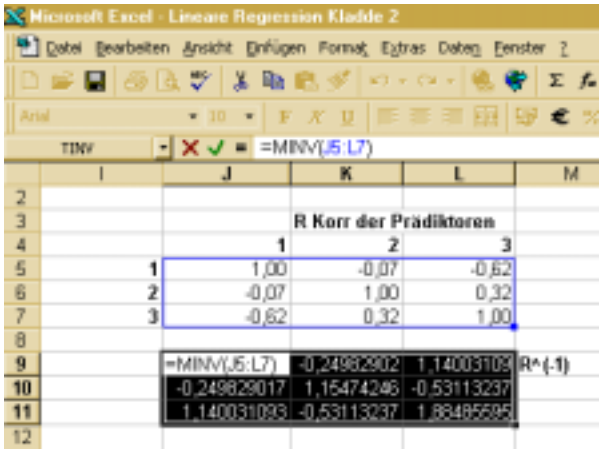
Um eine schrittweise Korrelation durchzuführen, muß man die standardisierten Koeffizienten b_i bestimmen. Hierzu benötigt man zunächst die Korrelationsmatrix der Prädiktoren. Hierfür werden die Korrelationen zwischen den Variablen X_1 ... X_3 paarweise mit dem Befehl „korrel“ (s.o.) berechnet. Die Korrelationsmatrix ist hier in den Bereich (j5:l7) eingetragen:



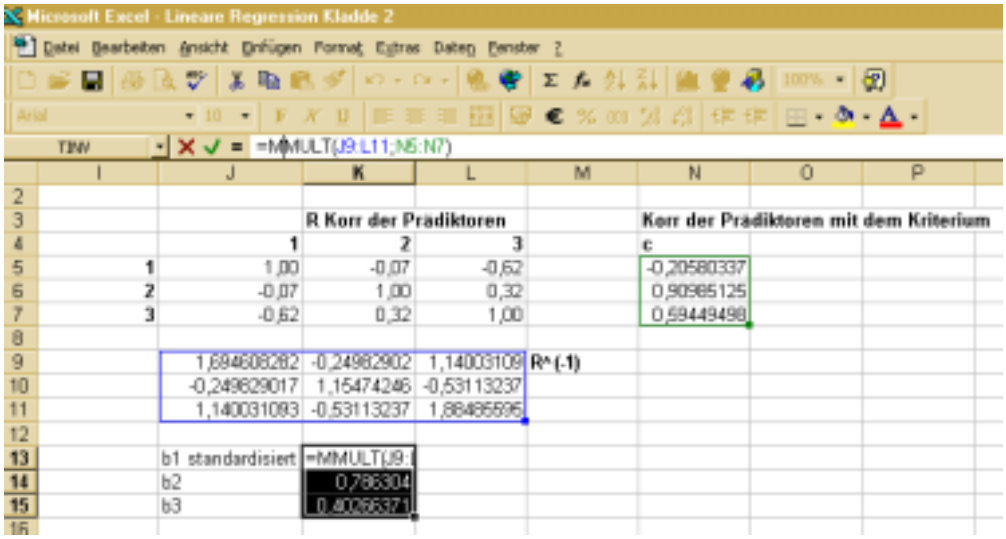
Anschließend berechnet man die Korrelationen zwischen den Prädiktoren und dem Kriterium, die hier in den Bereich (n5:n7) eingetragen werden:



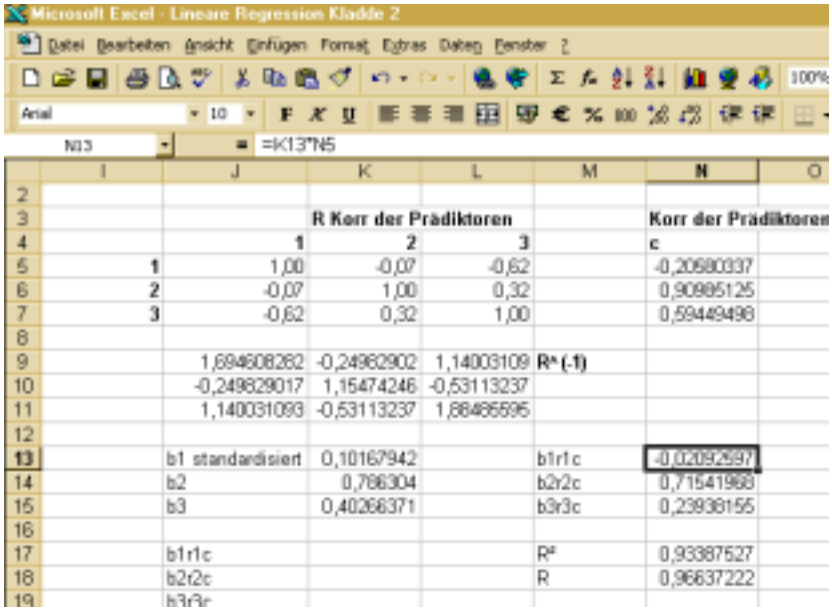
Nun ist die Inverse der Korrelationsmatrix der Prädiktoren zu bestimmen:



Die standardisierten Regressionskoeffizienten (hier im Bereich k13:k15) erhält man durch Multiplikation von $\underline{R}^{-1}$ mit $\underline{r}$.



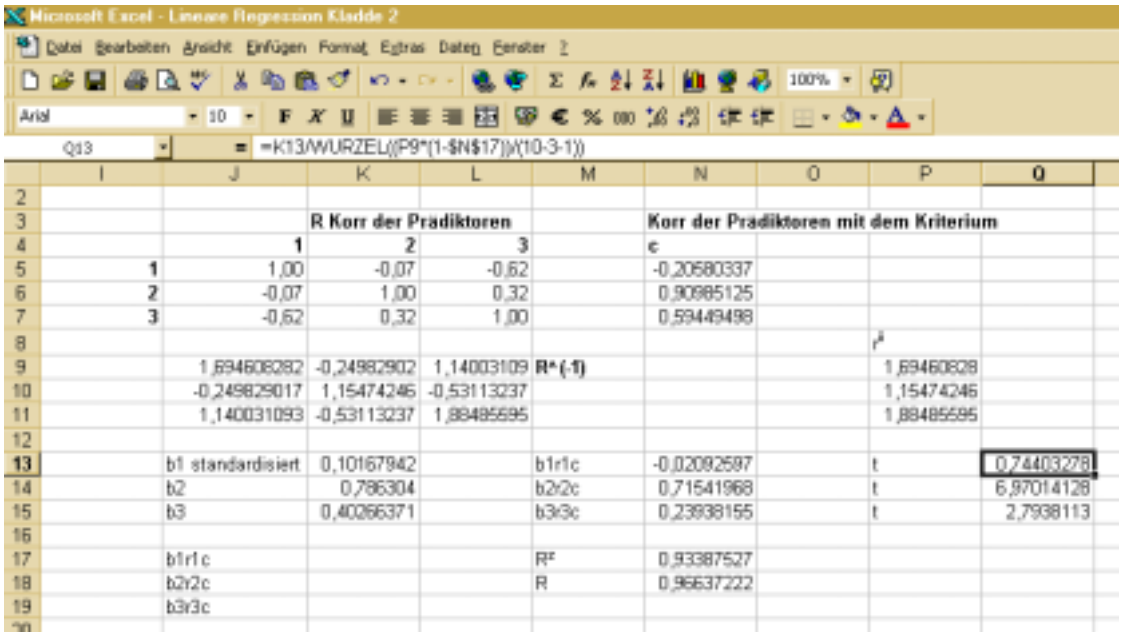
Die standardisierten Koeffizienten b_i werden mit den Korrelationen der Prädiktoren mit dem Kriterium r_{jc} multipliziert, hier im Bereich (n13:n15). Die Summe dieser Produkte ergibt R^2, die Wurzel aus dieser Summe R:



	I	J	K	L	M	N	O
2							
3			R Korr der Prädiktoren			Korr der Prädiktoren	
4		1	2	3		c	
5	1	1,00	-0,07	-0,62		-0,20580337	
6	2	-0,07	1,00	0,32		0,90985125	
7	3	-0,62	0,32	1,00		0,59449498	
8							
9		1,694608282	-0,24962902	1,14003109	R*(-1)		
10		-0,249629017	1,15474246	-0,53113237			
11		1,140031093	-0,53113237	1,88485595			
12							
13		b1 standardisiert	0,10167942		b1r1c	-0,02092597	
14		b2	0,786304		b2r2c	0,71541968	
15		b3	0,40266371		b3r3c	0,23938155	
16							
17		b1r1c			R^2	0,93387527	
18		b2r2c			R	0,96637222	
19		b3r3c					

6.9 Prüfung der standardisierten Regressionskoeffizienten auf Signifikanz

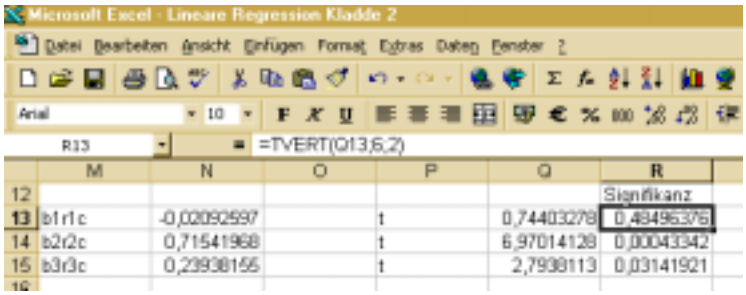
Die Prüfung der standardisierten Regressionskoeffizienten auf Signifikanz erfolgt nach (35), wir unterstellen, daß die Variablen normalverteilt seien. In Excel erfolgt die Berechnung der t-Werte entsprechend in den Zellen im Bereich (q13:q15); die Werte (r^{ii}) wurden zur besseren Berechnung in den Bereich (p9:p11) übertragen:



	I	J	K	L	M	N	O	P	Q
2									
3			R Korr der Prädiktoren			Korr der Prädiktoren mit dem Kriterium			
4		1	2	3		c			
5	1	1,00	-0,07	-0,62		-0,20580337			
6	2	-0,07	1,00	0,32		0,90985125			
7	3	-0,62	0,32	1,00		0,59449498			
8								r^{ii}	
9		1,694608282	-0,24962902	1,14003109	R*(-1)			1,69460828	
10		-0,249629017	1,15474246	-0,53113237				1,15474246	
11		1,140031093	-0,53113237	1,88485595				1,88485595	
12									
13		b1 standardisiert	0,10167942		b1r1c	-0,02092597		t	0,74403278
14		b2	0,786304		b2r2c	0,71541968		t	6,97014128
15		b3	0,40266371		b3r3c	0,23938155		t	2,7938113
16									
17		b1r1c			R^2	0,93387527			
18		b2r2c			R	0,96637222			
19		b3r3c							

Die Funktion „tvert(t-Wert;Freiheitsgrade;Seiten)“ prüft die t-Werte auf Signifikanz. Hier wurde die Funktion in die Zellen des Bereichs (r13:r15) eingetragen. Die Regressionskoef-

fizienten b_2 und b_3 sind also auf einem Niveau von $\alpha = 0{,}05$ (zweiseitig) signifikant; die Variablen „Werbeausgaben“ und „Vertreterbesuche“ leisten positive Beiträge von 0,72 bzw. 0,24, der Preis ist nicht signifikant und leistet einen negativen Beitrag zur Gesamtmodellgüte.



	M	N	O	P	Q	R
12						Signifikanz
13	b1r1c	-0,02092597		t	0,74403279	0,48496376
14	b2r2c	0,71541968		t	6,97014128	0,00043342
15	b3r3c	0,23938155		t	2,7938113	0,03141921

Man würde nun ein neues Modell erstellen, das nur die Variablen „Werbeausgaben“ und „Vertreterbesuche“ enthält.

6.10 Aufgabe

Erstellen Sie ein neues Modell, das nur die Prädiktoren „Werbeausgaben“ und „Vertreterbesuche“ enthält. Errechnen Sie die Modellgüte und die standardisierten Regressionskoeffizienten. Prüfen Sie die Regressionskoeffizienten auf Signifikanz und bestimmen Sie die Beiträge der beiden Variablen für das Modell!