Multiple Regressionsanalyse - Kurzabriss

Ziele:

- Schätzung eines Kriteriums aus einer Linearkombination von Prädiktoren
- Meist zu 'Screening-Untersuchungen', um den Einfluß von vermuteten Ursachenvariablen abzuschätzen
- In der Regel mit konkreten Variablen (Messgrößen)

Allgemeine Schätzgleichung:

$$\widehat{X}_{0i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} \qquad (1)$$

	$\widehat{X}_{0i}$	Schätzung des Wertes der i-ten Person auf der Kriteriumsvariable X_0
hierin:	β_j	Regressionsgewicht des j-ten Prädiktors
	X_{ji}	Wert der i-ten Person auf der j-ten Prädiktorvariable

Beispiel:

ABINote $= \beta_0 + \beta_1 \cdot$Rechentestwert$+\beta_2 \cdot$Sprachtestwert$+\beta_3 \cdot$Fleisstestwert$+\beta_4 \cdot$Ausdruckstestwert

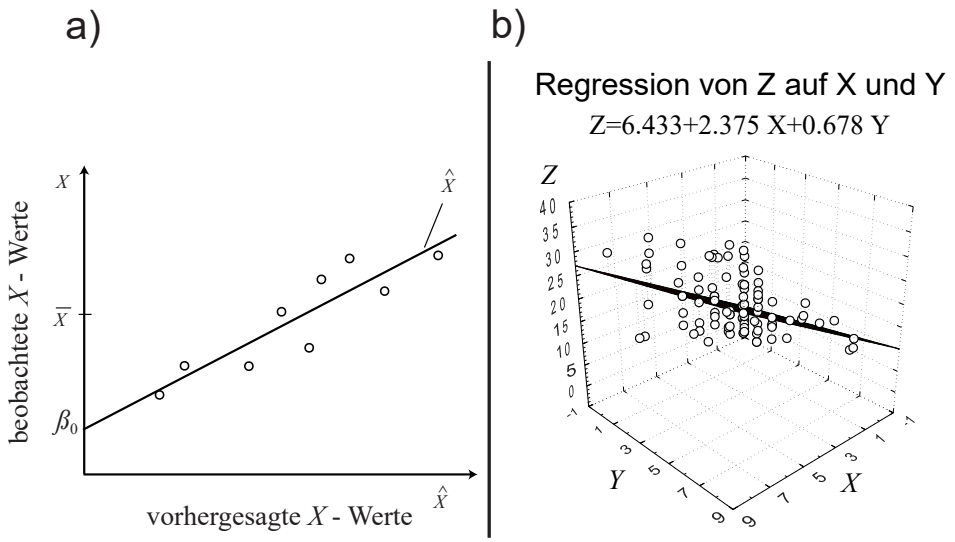

Abbildung 1: Veranschaulichung multiple Regression. a) Plot der durch die Batterie vorhergesagten Werte gegene die beobachteten Werte. b) Z wird aus X und Y vorhergesagt, Darstellung in der Regressionsebene.

0.1 Modellformulierung

Definition einer Fehlerfunktion (Optimierungsfunktion)

Schätzfehler:

$$e_i = X_{0i} - \widehat{X}_{0i} \tag{2}$$

Optimierungskriterium:
Wähle die Regressionsgewichte so, daß die Summe der Schätzfehlerquadrate minimal ist:

$$\begin{aligned} \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left(X_{0i} - \widehat{X}_{0i} \right)^2 \longrightarrow \min \\ &= \sum_{i=1}^{n} \left(X_{0i} - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}) \right)^2 \longrightarrow \min \end{aligned} \tag{3}$$

Man schreibt die multiple Regression i.d. Regel standardisiert:

$$\widehat{z}_{0i} = b_1 z_{1i} + b_2 z_{2i} + \ldots + b_k z_{ki} \tag{4}$$

Für den Schnittpunkt β_0 gilt:

$$\beta_0 = \overline{X}_0 - (\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \tag{5}$$

Weiterhin gibt

$$b_j = \beta_j \frac{s\,(x_j)}{s\,(x_0)} \tag{6}$$

die Beziehung zwischen den Regressionskoeffizienten der standardisierten und der unstandardisierten Variablen an. Beim Übergang zu z- standardisierten Variablen fällt die additive Konstante weg.

0.2 Gang der Lösung

Für den multivariaten Fall lautet der Modellansatz bei vorheriger z- Standardisierung aller k Variablen:

$$\widehat{z}_{0i} = b_1 z_{1i} + b_2 z_{2i} + \ldots + b_k z_{ki} \tag{7}$$

Schreibt man die Gleichung nur für die Variablen, lautet sie:

$$\widehat{z}_0 = b_1 z_1 + b_2 z_2 + \ldots + b_k z_k \tag{8}$$

Die Regressionskoeffizienten b_j erhält man aus der Bedingung

$$\sum_{i=1}^{n} (z_{0i} - \widehat{z}_{0i})^2 \longrightarrow \min$$

durch Einsetzen von (7), partielles Ableiten nach jedem Regressionskoeffizienten und anschließendes Nullsetzen, denn dann erhält man ein System von Normalgleichungen. Dieses erhält man ebenso, wenn man (8) nacheinander mit jedem Prädiktor multipliziert und über Personen i summiert sowie durch die Anzahl n teilt. Es ergeben sich nacheinander für $j = 1, \ldots, k$ folgende Gleichungen:

$$\begin{array}{lcllcllcl}
b_1 \frac{1}{n}\sum z_1 z_1 & + & b_2 \frac{1}{n}\sum z_1 z_2 & + & \cdots & + & b_k \frac{1}{n}\sum z_1 z_k & = & \frac{1}{n}\sum z_1 z_0 \\
b_1 \frac{1}{n}\sum z_2 z_1 & + & b_2 \frac{1}{n}\sum z_2 z_2 & + & \cdots & + & b_0 \frac{1}{n}\sum z_2 z_k & = & \frac{1}{n}\sum z_2 z_0 \\
\vdots & & \vdots \quad \vdots & & \vdots \quad \vdots & & \vdots \quad \vdots & & \vdots \quad \vdots \\
b_1 \frac{1}{n}\sum z_0 z_1 & + & b_2 \frac{1}{n}\sum z_0 z_2 & + & \cdots & + & b_k \frac{1}{n}\sum z_0 z_0 & = & \frac{1}{n}\sum z_0 z_0
\end{array}$$

die offenbar einfach auf das folgende Gleichungssystem hinauslaufen:

$$\begin{array}{lcllcllcl}
b_1 & + & b_2 r_{12} & + & \cdots & + & b_k r_{1k} & = & r_{10} \\
b_1 r_{21} & + & b_2 & + & \cdots & + & b_k r_{2k} & = & r_{20} \\
b_1 r_{31} & + & b_2 r_{32} & + & \cdots & + & b_k r_{3k} & = & r_{3k} \\
\vdots & & \vdots \quad \vdots & & \vdots \quad \vdots & & \vdots \quad \vdots & & \vdots \quad \vdots \\
b_1 r_{k1} & + & b_2 r_{k2} & + & \cdots & + & b_k & = & r_{k0}
\end{array} \tag{9}$$

Rechts stehen hier die Kriteriumskorrelationen und links die Prädiktorinterkorrelationen. Auf der Diagonalen des Gleichungssystems sthen die gesuchten Koeffizienten. Das System hat genausoviel Zeilen wie Unbekannte (die Regressionsgewichte) $\rightarrow$ es ist lösbar. Man kann (9) als eine Matrixgleichung schreiben.

Wir schreiben

$$\mathbf{Rb} = \mathbf{r}_{k0} \tag{10}$$

wobei boldface Symbole für Vektoren stehen. Insbesondere ist:

$\mathbf{R}$	$k \times k$ Matrix der Prädiktorinterkorrelationen
$\mathbf{b}$	gesuchter $k \times 1$ Vektor der Regressionskoeffizienten
$\mathbf{r}_{k0}$	$k \times 1$ Vektor der Kriteriumskorrelationen

Vormultiplizieren mit der Inversen der Prädiktor-Interkorrelationsmatrix gibt:

$$\begin{aligned}
\mathbf{R}^{-1}\mathbf{Rb} &= \mathbf{R}^{-1}\mathbf{r}_{k0} \\
\mathbf{b} &= \mathbf{R}^{-1}\mathbf{r}_{k0}
\end{aligned}$$

Die Lösung zeigt uns, daß nur die Inverse der Interkorrelationsmatrix der Prädiktoren gefunden werden muß, um das Problem der Bestimmung der Beta-Gewichte zu lösen.

0.3 Der multiple Korrelationskoeffizient

Für die multiple Regression läßt sich ebenfalls ein multipler Korrelationskoeffizient berechnen. Er lautet:

$$R_{0\cdot 12\cdots k} = \sqrt{\sum_{j=1}^{k} b_j \cdot r_{j0}} \tag{11}$$

Was bedeutet dieser multiple Korrelationskoeffizient? Er erhält seine Bedeutung im Rahmen der Varianzzerlegung, die man für die multiple, wie auch für die bivariate Regression, folgendermassen anschreiben kann:

$$\begin{array}{ccccc} \text{Gesamt-Streuung} & = & \text{Erklärte Streuung} & + & \text{Nicht Erklärte Streuung} \\ \frac{1}{n}\sum_{i=1}^{n}\left(X_{0i}-\overline{X}_{0i}\right)^2 & = & \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{X}_{0i}-\overline{X}_{0i}\right)^2 & + & \frac{1}{n}\sum_{i=1}^{n}\left(X_{0i}-\widehat{X}_{0i}\right)^2 \end{array}$$

Wegen

$$1 = \frac{\text{Erklärte Streuung}}{\text{Gesamt-Streuung}} + \frac{\text{Nicht Erklärte Streuung}}{\text{Gesamt-Streuung}} \tag{12}$$

hat man

$$R^2 = \frac{\text{Erklärte Streuung}}{\text{Gesamt-Streuung}} \tag{13}$$

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{X}_{0i}-\overline{X}_{0i}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_{0i}-\overline{X}_{0i}\right)^2} \tag{14}$$

$$\Rightarrow R^2 = 1 - \frac{\text{Nicht Erklärte Streuung}}{\text{Gesamt-Streuung}}. \tag{15}$$

Der multiple Korrelationskoeffizient kann genauso interpretiert werden wie der bivariate Produkt - Moment Korrelationskoeffizient.

Aber:

- Je mehr Prädiktoren, desto größer der Anteil an erklärter Streuung
- Je mehr Prädiktoren, desto ungünstiger für die Schätzeigenschaften des Modells (lernen wir später beim Thema Inferenzstatistik kennen)

0.4 Statistische Signifikanz

Inferenzstatistisch interessant sind folgende Fragen

1. Ist der Zusammenhang zwischen dem Kriterium und den Prädiktoren statistisch signifikant? (Läßt sich die gefundene multiple Korrelation mit der Annahme vereinbaren, daß die 'wahre' multiple Korrelation zwischen dem Kriterium und den Prädiktoren gleich Null ist ?)

2. Welche Regressionskoeffizienten leisten einen statistisch bedeutsamen Beitrag zur Gesamtschätzung des Kriteriums (Für welche Regressionskoeffizienten läßt sich die Annhame nicht halten, daß deren wahrer Regressionskoeffizient gleich Null ist?)

Die erster Frage beantwortet ein $F-$ Test:

$$F = \frac{R^2 \left(n - k - 1\right)}{\left(1 - R^2\right) \cdot k} \tag{16}$$

mit k Zählerfreiheitsgraden und $n - k - 1$ Nennerfreiheitsgraden.

Die zweite Frage beantwortet ein $t-$ Test:

$$t = \frac{b_j}{\sqrt{\frac{r^{jj}(1-R^2)}{n-k-1}}} \tag{17}$$

mit $n - k - 1$ Freiheitsgraden. Hier ist r^{jj} das $jj-$ te Element der invertierten Korrelationsmatrix. Voraussetzung der Signifikanzprüfung ist die multivariate Normalverteilung der beteiligten Variablen.

0.5 Interpretation der Regressionskoeffizienten

0.5.1 Unabhängige Prädiktoren

Die Schwierigkeit der Interpretation der Regressionsgewichte besteht in der Beurteilung des Einflusses der *Interkorrelationen* der Prädiktoren. Für unkorrelierte Prädiktoren ist die Deutung einfach, da in diesem Fall $\mathbf{R} = \mathbf{R}^{-1} = \mathbf{I}$ ist. Dann wird der Vektor der Regressionskoeffizienten

$$\mathbf{b} = \mathbf{I}\mathbf{r}_{k0} \tag{18}$$

$$\mathbf{b} = \mathbf{r}_{k0} \tag{19}$$

d.h. die beta- Gewichte sind gleich den Kriteriumskorrelationen. Dies entspricht dem Fall einer Varianzzerlegung mit *rein additiven Komponenten*, d.h. es gibt keine Kovarianzterme. Es ist dann

$$R^2_{0\cdot 12\cdots k} = \sum_{j=1}^{k} r^2_{j0} \tag{20}$$

d.h.

$$\frac{\text{Erklärte Streuung}}{\text{Gesamt-Streuung}} = \frac{\text{Erklärte Streuung (1)}}{\text{Gesamt-Streuung}} + \frac{\text{Erklärte Streuung (2)}}{\text{Gesamt-Streuung}} + \ldots + \frac{\text{Erklärte Streuung } k}{\text{Gesamt-Streuung}}.$$

0.5.2 Abhängige Prädiktoren

Abhängigkeiten zwischen den Prädiktoren können prinzipiell in zweierlei Hinsicht untersucht werden:

1. Bedeutet die Abhängigkeit Redundanz, d.h. messen die vielen Variablen Aspekte gemeinsam, so daß man prinzipiell weniger (latente) Variablen benötigt? ($\rightarrow$ unerwünschter Aspekt)

2. Erfassen die Abhängigkeiten Teile der ($\rightarrow$ Kontamination) der Variablen und wirken so optimierend auf die gesamte Schätzgleichung (Suppressionseffekt, erwünscht)?

Zusammenhang mit der Partialkorrelation. Partialkorrelationen geben die Korrelation zweier Variablen an, die vom Effekt anderer (spezifizierter) Variablen bereinigt wurden. So gibt $r_{12\cdot 3}$ die Korrelation von X_1 und X_2 bereinigt vom Effekt der Variablen X_3. Das Prinzip dabei ist, aus den Variablen X_1 und X_2 den Vorhersageanteil zulasten von X_3 herauszuziehen (die Schätzwerte von den beobachteten Werten abzuziehen) und nunmehr die Korrelation der Residuen zu betrachten, dieses ist die "reine" Korrelation der Variablen X_1 und X_2. Ähnliches kann durch eine geeignete Gewichtung von Variablen, die keine hohe Kriteriumskorrelation haben, in der Schätzgleichung erreicht werden (Suppressionseffkt). Für den Fall von nur 3 Variablen in der Schätzgleichung gibt es Ungleichungen, anhand deren man Suppressorvariablen identifizieren kann. Für mehr Variablen ist dies aber nicht so einfach. Als Faustregel für die Identifikation der Suppressorwirkung gilt

$$U_j > r^2_{j0}, \tag{21}$$

wobei U_j die sog. *Nützlichkeit* der Variablen X_j ist.

$$U_j = +\Delta R^2(j)$$

meint den Betrag, um den die multiple Korrelation zunimmt, wenn Variable X_j zusätzlich in die Gleichung aufgenommen wird. Anhaltspunkte für die Rolle der Variablen in der Schätzgleichung bieten die Step-Methode, iterative Methoden und die Betrachtung der Partialkorrelationen. Man kann *Strukturkoeffizienten* bilden, dessen Quadrat angibt, welchen Anteil eine Prädiktorvariable an der *vorhergesagten* Kriteriumsvarianz hat

$$c_j = \frac{r_{j0}}{R}.$$

Weiter läßt sich zeigen:

$$R^2 = r_{01}^2 + r_{0(2\cdot1)}^2 + r_{0(3\cdot21)}^2 + r_{0(4\cdot321)}^2 + \ldots \tag{22}$$

d.h. die aufgeklärte Varianz ist darstellbar über eine Reihe von Semipartialkorrelationen (nur der Prädiktor ist vom linearen Anteil der anderen Prädiktoren bereinigt), wobei jede neu hinzukommende Prädiktorvariable vom Effekt aller bisher berücksichtigten Variablen bereinigt ist. Die Semipartialkorrelationen sind für die Interpretation wichtig, denn es gilt

$$r_{0(j\cdot12\cdots k)}^2 = \frac{Varianz(X_{j\cdot12\cdots k})}{Gesamtvarianz}$$

Die quadrierten Semipartialkorrelationen geben also den Anteil der Varianz eines von den anderen Prädiktoren residualisierten Prädiktors an der gesamten Kriteriumsvarianz (seinen dann eigenständigen Varianzbeitrag) an. Für 3 Variablen berechnet sich die Semipartialkorrelation nach (X_0 Kriterium, X_2 von X_1 befreit)

$$r_{0(2.1)} = \frac{r_{20} - r_{10}r_{12}}{\sqrt{1 - r_{12}^2}} \tag{23}$$

Aus (22) folgt, dass die Nützlichkeit (21) eines Prädiktors j das Quadrat seiner Semipartialkorrelation ist, denn die Differenz

$$r_{0(j\cdot12\cdots k)}^2 = R_{0\cdot(12\cdots k)}^2 - R_{0\cdot(12\cdots k|-j)}^2 \tag{24}$$

definiert ja die Nützlichkeit der Variablen X_j nach (21) (Näheres hierzu s. Bortz, 2005, Kap. 13).