

6.2.4 Linearisierende Transformationen:

Sehr häufig wirkt die Variable X nicht „direkt linear“ auf die Variable Y (Streudiagramm anschauen!). Die lineare Regression passt die „optimale Gerade“ an. Was kann man aber tun, wenn selbst diese optimale Gerade nicht passt, da der Zusammenhang eben nicht linear ist.

Viele (nicht alle) der auf den ersten Blick nichtlinearen Modelle lassen sich durch geeignete Variablentransformationen in die lineare Regressionsrechnung einbetten.

Das Modell

$$g(y_i) = \beta_0 + \beta_1 \cdot h(x_i) + \varepsilon_i$$

lässt sich auch völlig analog mit dem KQ-Prinzip behandeln.

Demgegenüber lässt sich etwa das Modell

$$y_i = \beta_0 + x_i^{\beta_1} + \varepsilon_i$$

nicht durch eine Variablentransformation in ein lineares Modell überführen.

6.2.5 Multiple lineare Regression

Verallgemeinerung der linearen Einfachregression: Betrachte mehrere unabhängige metrische Variablen X_1, X_2, \dots, X_p gemeinsam, da typischerweise ja kein monokausaler Zusammenhang vorliegt. Man versucht dann, die Datenpunkte durch eine optimale Hyperebene zu beschreiben.

Modellgleichung:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i.$$

Dabei bezeichnet x_{i1} den für die i -te Beobachtung beobachteten Wert der Variablen X_1 , x_{i2} den Wert der Variablen X_2 , usw.

Interpretation: Die Interpretation von β_0 und β_1, \dots, β_p erfolgt analog zu oben, insbesondere ist β_j die Änderung in Y , wenn X_j um eine Einheit vergrößert wird – und alle anderen Größen gleichbleiben („*ceteris paribus* Effekt“).

KQ-Prinzip: Die Schätzung von $\beta_0, \beta_1, \dots, \beta_p$ erfolgt wieder über das KQ-Prinzip: Bestimme $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ so, dass mit

$$\hat{\varepsilon}_i = y_i - \hat{y}_i := y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_p x_{pi})$$

der Ausdruck

$$\sum_{i=1}^n \hat{\varepsilon}_i^2$$

minimal wird.

Die Schätzungen $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ sind nur mit Matrizenrechnung einfach darzustellen und insbesondere nur noch schwer „von Hand“ zu berechnen. Im Rahmen dieser Veranstaltung brauchen Sie bei der multiplen Regression nicht mehr rechnen, sondern „nur“ typische Outputs korrekt interpretieren können.

Bestimmtheitsmaß: Analog zur linearen Einfachregression lässt sich ein Bestimmtheitsmaß

$$R^2 = \frac{SQE}{SQT}$$

über die Streuungszerlegung definieren. In der multiplen Regression verwendet man allerdings meistens das korrigierte Bestimmtheitsmaß

$$\tilde{R}^2 := 1 - \frac{n-1}{n-p-1}(1 - R^2)$$

das die Anzahl der in das Modell mit einbezogenen Variablen mit berücksichtigt. (Das übliche R^2 würde ja auch durch das Einführen irrelevanter Variablen ansteigen, während bei \tilde{R}^2 sozusagen für jede aufgenommene Variable einen Preis zu bezahlen ist.)

SPSS-Output einer multiplen Regression:

Coefficients^a

Model	Unstandardized Coefficients		t	Sig.	
	B	Std. Error			
1	(Constant)	$\hat{\beta}_0$	$\hat{\sigma}_0$	T_0	p-Wert
	X_1	$\hat{\beta}_1$	$\hat{\sigma}_1$	T_1	"
	X_2	$\hat{\beta}_2$	$\hat{\sigma}_2$	T_2	"
	⋮	⋮	⋮	⋮	"
	X_p	$\hat{\beta}_p$	$\hat{\sigma}_p$	T_p	"

^a Dependent Variable: Y

Stilisierter SPSS-Output einer multiplen Regression

Im Rahmen von Statistik 1 ist nur die Spalte „B“ mit den unstandardisierten Koeffizienten $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ relevant.

Anmerkung: SPSS gibt auch noch –als „ β “ bezeichnete– „standardisierten Koeffizienten“ aus. Das sind nicht etwa die $\hat{\beta}$'s im Sinne der Vorlesung, sondern die Schätzer, wenn man die Variablen vorher standardisiert. Bei der linearen Einfachregression findet man hier den Korrelationskoeffizienten von Bravais Pearson wieder.