

Vorlesung Multivariate Verfahren

Prof. Dr. Thomas Staufenbiel
Universität Osnabrück
Evaluation und Forschungsmethodik

Multivariate Verfahren

Multiple Regression & Korrelation

Literatur Multiple Regression

Einführende Literatur:

- Bortz, J. (2005). *Statistik* (6. Auflage). Berlin: Springer. [Kap. 13]
- Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill. [vor allem Kap. 6, 14]
- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum. [Kap. 3]

Weiterführende Literatur:

- Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.
- Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: LEA.

Multiple Regression & Korrelation: Gliederung Teil I

1. Grundlagen

(Ein einführendes Beispiel, Wiederholung
bivariate Regression)

2. Modell der Multiplen Regression

(Modellgleichung, Ziel und Verfahren, ein
rechnerisches Beispiel)

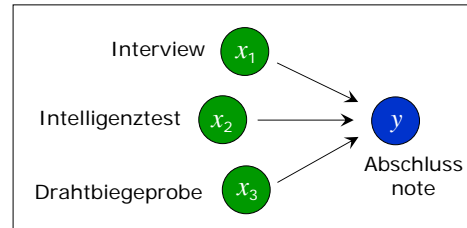
3. Multiple Korrelation

4. Zusammenwirken der Prädiktoren

(Unabhängigkeit, Redundanz, Komplementarität,
Suppression)

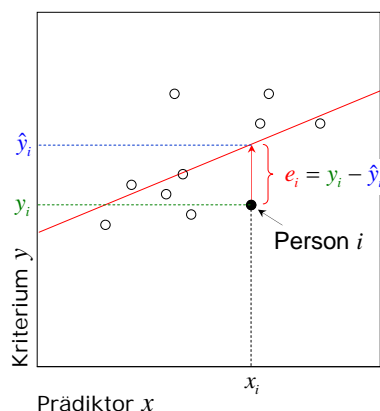
Multiple Regression / Korrelation: Ein Beispiel

- **Beispiel:** Anhand einer Reihe von Verfahren (u.a. Interview, Intelligenztest, Drahtbiegeprobe) sollen Auszubildende für den Beruf des Industriemechanikers ausgewählt werden.



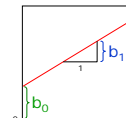
- **Fragestellung:** Wie gut gelingt die Vorhersage der Kriteriumsvariable (=Ausbildungserfolgs, gemessen durch die Berufsschulabschlussnote), wenn man alle Prädiktoren (=Auswahlverfahren) gemeinsam nutzt?
- Weitere Fragestellungen:
 - Wie groß ist der relative Einfluss der einzelnen Prädiktoren?
 - Welche und wie viele Prädiktoren sollten berücksichtigt werden?
 - Wie lassen sich Kriteriumswerte für neue Personen (Bewerber) bestimmen?

Einfache (bivariate) Regression



- Bestimme Regressionsgewichte b_0 und b_1 für $\hat{y}_i = b_0 + b_1 \cdot x_i$ derart, dass

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 \rightarrow \min$$



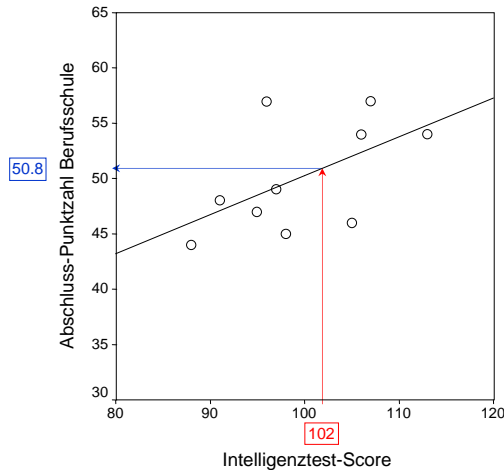
- Durch Differenzierung nach b_1 und b_0 resultiert:

$$b_1 = r \frac{s_y}{s_x} \quad \text{und} \quad b_0 = \bar{y} - b_1 \bar{x}$$

- Standardschätzfehler = Standardabweichung der Fehler e_i :

$$s_e = s_y \sqrt{1 - r^2}$$

Einfache (bivariate) Regression



$$N = 10$$

$$\bar{x} = 99.60, \quad s_x = 7.86$$

$$\bar{y} = 50.10, \quad s_y = 4.95$$

$$r_{xy} = 0.56 \quad (p = .09)$$

$$b_0 = 15.12 \quad b_1 = 0.35$$

$$\hat{y}_i = 15.12 + 0.35 \cdot x_i$$

$$x = 102 \Rightarrow$$

$$\hat{y}_i = 15.12 + 0.35 \cdot 102 = 50.8$$

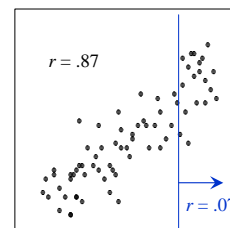
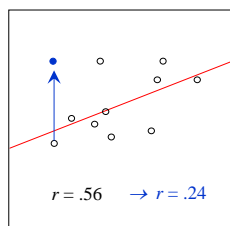
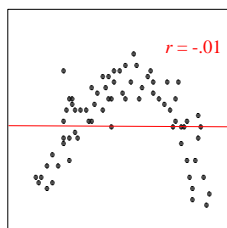
Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Bivariate Regression / Korrelation

➤ Zu beachten bei der linearen Regression/Korrelation ist allg.:

- quantifiziert wird der lineare Zusammenhang (d.h. es gibt auch andere Formen, für die Produkt-Moment-Korrelation nicht geeignet ist)
- Korrelationen können stark durch Ausreißer ("outlier") beeinflusst werden
- Korrelationen können durch Einschränkung der Wertebereiche der Variablen ("range-restriction") beträchtlich verändert – d.h. meist reduziert – werden (was man entsprechend statistisch korrigieren kann).
- aus einer Korrelation kann nicht auf die Kausalrichtung geschlossen werden



Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Multiple Regression & Korrelation: Gliederung Teil I

1. Grundlagen

(Ein einführendes Beispiel, Wiederholung bivariate Regression)

2. Modell der Multiplen Regression

(Modellgleichung, Ziel und Verfahren, ein rechnerisches Beispiel)

3. Multiple Korrelation

4. Zusammenwirken der Prädiktoren

(Unabhängigkeit, Redundanz, Komplementarität, Suppression)

Multiple Regression: Modellgleichung

➤ Modellgleichung der multiplen Regressionsanalyse:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_q x_{iq} + \dots + b_m x_{im} = b_0 + \sum_{j=1}^m b_j x_{ij}$$

$$\Leftrightarrow y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_q x_{iq} + \dots + b_m x_{im} + e_i$$

mit

x_{iq} beobachteter Wert der Person i im Prädiktor q ($q=1..m$ Prädiktoren)

y_i beobachteter Wert der Person i im Kriterium ($i=1..n$ Personen)

\hat{y}_i vorhergesagter Wert der Person i im Kriterium

e_i Fehlerwert (=Residuum) der Person i (mit $e_i = y_i - \hat{y}_i$)

b_q (unstandardisiertes Regressionsgewicht) des Prädiktors q ($q=1..m$)

Multiple Regression: Modellgleichung

- Modell in Matrixdarstellung:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + \dots + b_m \begin{pmatrix} x_{1m} \\ x_{2m} \\ \vdots \\ x_{nm} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$\mathbf{y} = b_0 \mathbf{u} + b_1 \mathbf{x}_1 + \dots + b_m \mathbf{x}_m + \mathbf{e}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X} \mathbf{b} + \mathbf{e}$$

$${}_n \mathbf{y}_1 = {}_n \mathbf{X}_m \quad {}_m \mathbf{b}_1 + {}_n \mathbf{e}_1$$

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Multiple Regression: Modellgleichung

- Standardisiert man alle Variablen vor der Berechnung, so erhält man

$$\hat{y}_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \dots + \beta_q^* x_{iq}^* + \dots + \beta_m^* x_{im}^*$$

der * an allen Prädiktoren und dem Kriterium zeigt an, dass diese Variablen alle z-standardisiert vorliegen.

Die dann resultierenden Regressionsgewichte β_q^* bezeichnet man als standardisierte Regressionsgewichte (oder auch Beta-Gewichte oder Standardpartialregressionskoeffizienten; der griechische Buchstabe steht also nicht – wie sonst häufig – für einen Populationskennwert)

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Multiple Regression: Ziel und Verfahren

- Ziel: Die Prädiktoren sollen so gewichtet werden, dass ihre gewichtete additive Verknüpfung (=Linearkombination) den Kriteriumswerten "möglichst nahe" kommt.
- "Möglichst nahe" wird z.B. wie bei der bivariaten Regression definiert, d.h. die Regressionsgewichte sollen so bestimmt werden, dass die quadrierten Residuen minimiert werden ["ordinary least squares" (OLS)-Schätzungen]:

$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2 \rightarrow \min$$

- Voraussetzungen für die Bestimmung:
 - Die Prädiktorenmatrix X hat den vollen Rang $(m+1)$
 - notwendig dazu: Zahl der Personen $n >$ Zahl der Variablen m
- Mathematisches Verfahren z.B. modifizierte Cholesky-Zerlegung von $X'X$

Multiple Regression: Beispiel I

Re- gion	Verkaufte Stückzahl (y)	Preis in Euro pro Stück (x_1)	Werbung in Euro (x_2)	Zahl Vertreter- besuche (x_3)
1	2298	12.50	2000	109
2	1814	10.00	550	107
3	1647	9.95	1000	99
4	1496	11.50	800	70
5	969	12.00	0	81
6	1918	10.00	1500	102
7	1810	8.00	800	110
8	1896	9.00	1200	92
9	1715	9.50	1100	87
10	1699	12.50	1300	79

- Vorhersage des Umsatzes ("Verkaufte Stückzahl") aus den drei Prädiktoren Preis, Werbung und Zahl Vertreterbesuche pro Periode (aus: Backhaus et al., 1990).

$$\begin{pmatrix} 2298 \\ 1814 \\ \vdots \\ 1699 \end{pmatrix} = b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} 12.50 \\ 10.00 \\ \vdots \\ 12.50 \end{pmatrix} + b_2 \begin{pmatrix} 2000 \\ 550 \\ \vdots \\ 1300 \end{pmatrix} + b_3 \begin{pmatrix} 109 \\ 107 \\ \vdots \\ 79 \end{pmatrix} + \begin{pmatrix} 648.65 \\ 68.83 \\ \vdots \\ 49.65 \end{pmatrix}$$

$$y = b_0 u + b_1 x_1 + b_2 x_2 + b_3 x_3 + e$$

$$y = 725.5 - 26.3 x_1 + 0.5 x_2 + 8.4 x_3 + e$$

Es resultieren folgende optimale Regressionsgewichte

Multiple Regression: Beispiel I

➤ Aus der Regressionsgleichung

$$\text{"Stückzahl"} = 725.5 - 26.3 \cdot \text{"Preis"} + 0.5 \cdot \text{"Werbung"} + 8.4 \cdot \text{"Zahl Vertreterbesuche"} + \text{Fehler}$$

kann man schließen:

- *Werbung* und *Vertreterbesuche* wirken sich förderlich auf die Zahl der verkauften Einheiten aus, ein hoher *Preis* hinderlich.
- dass z.B. für die folgenden Werte einer weiteren Region vorhergesagt werden kann:
 $\text{"Geschätzte Stückzahl"} = 725.5 - 26.3 \cdot 9.90 + 0.5 \cdot 800 + 8.4 \cdot 95 = 1663.13$
- dass vorhergesagt wird, dass 10 weitere Vertreterbesuche den Verkauf um 84 Stück erhöhen (was bei einem Preis von 10 € dann 840 € brächte).

➤ Hingegen kann man nicht auf die relative Wichtigkeit der Prädiktoren schließen: die *b*-Gewichte hängen vom Maßstab der Prädiktoren ab! Beispiel: Eine Angabe des Prädiktors "Preis" in Cent statt Euro führt zu einem Gewicht von $b_1/100$.

➤ Um dies zu können, muss man die Variablen vorher standardisieren und erhält dann die β -Gewichte: $\beta(\text{Preis}) = -.12$, $\beta(\text{Werbung}) = .77$, $\beta(\text{Vertreterbes.}) = .35$

Multiple Regression & Korrelation: Gliederung Teil I

1. Grundlagen

(Ein einführendes Beispiel, Wiederholung bivariate Regression)

2. Modell der Multiplen Regression

(Modellgleichung, Ziel und Verfahren, ein rechnerisches Beispiel)

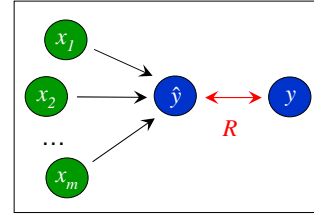
3. Multiple Korrelation

4. Zusammenwirken der Prädiktoren

(Unabhängigkeit, Redundanz, Komplementarität, Suppression)

Multipler Korrelationskoeffizient

- Die multiple Korrelation R bezeichnet analog zum bivariaten Fall die Korrelation zwischen den empirisch beobachteten und den durch die Regressionsgleichung vorhergesagten Kriteriumswerten: $R = r_{y\hat{y}}$
- R ist damit ein Maß für die Güte der Vorhersage.



Eigenschaften:

- R kann Werte von 0 bis 1 annehmen: $0 \leq R \leq 1$
- R kann nie kleiner werden als die größte aller bivariaten Korrelationen zwischen den Prädiktoren q und dem Kriterium: $R \geq |r_{qy}| \quad \forall q=1..m$
- R lässt sich bei bekannten β_j und Korrelationen zwischen den Prädiktoren und dem Kriterium bestimmen via: $R = \sqrt{\beta_1 r_{1y} + \beta_2 r_{2y} + \dots + \beta_m r_{my}}$
- R hängt von den Korrelationen der Prädiktoren mit dem Kriterium und den Korrelationen der Prädiktoren untereinander ab:

$$R = \sqrt{\frac{r_{1y}^2 + r_{2y}^2 - 2r_{1y}r_{2y}r_{12}}{1 - r_{12}^2}} \quad (\text{für } m = 2 \text{ Prädiktoren})$$

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Multipler Determinationskoeffizient

- Der multiple Determinationskoeffizient R^2 ist die quadrierte multiple Korrelation.

Eigenschaften:

- R^2 kann ebenfalls Werte von 0 bis 1 annehmen: $0 \leq R^2 \leq 1$
- R^2 ist immer kleiner gleich R (gleich nur bei $R=0$ und $R=1$): $R^2 \leq R$
- R^2 ist interpretierbar als der Anteil der Varianz im Kriterium, der durch die Vorhersage der Regressionsgleichung erklärt wird:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

← Erklärte Varianz
← Gesamtvarianz

Dabei gilt: $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Fehlervarianz

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Multiple Regression & Korrelation: Gliederung Teil I

1. Grundlagen

(Ein einführendes Beispiel, Wiederholung bivariate Regression)

2. Modell der Multiplen Regression

(Modellgleichung, Ziel und Verfahren, ein rechnerisches Beispiel)

3. Multiple Korrelation

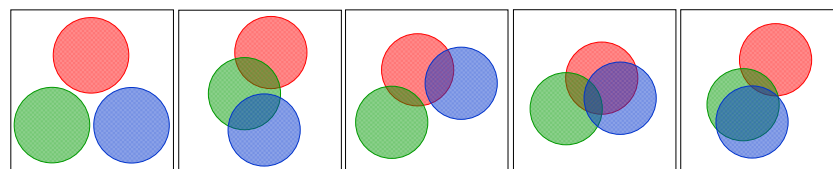
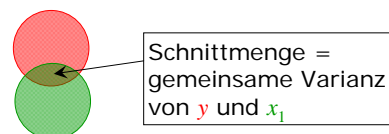
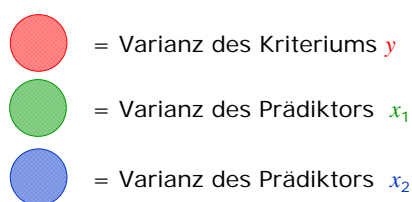
4. Zusammenwirken der Prädiktoren

(Unabhängigkeit, Redundanz, Komplementarität, Suppression)

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Multiple Regression: Vorhersage im Venn-Diagramm

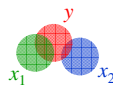
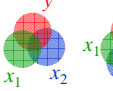
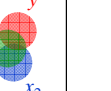


Prädiktoren abh.?	nein	ja	nein	ja	ja
Vorhersage möglich durch?	weder x_1 noch x_2	nur x_1	x_1 und x_2 unabhängig	x_1 und x_2 partiell redundant	beide, aber x_2 vollständig redundant zu x_1

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Zusammenwirken der Prädiktoren (bei m=2)

	Unabhängigkeit	Redundanz	Komplementarität	Suppression
Venn-Diagramm:		<div style="display: flex; justify-content: space-around;"> <div>  <p>partiell</p> </div> <div>  <p>vollständig</p> </div> </div>	nicht darstellbar	
Kennzeichen:	Die Prädiktoren sind unkorreliert.	Die Prädiktoren weisen einen gemeinsamen Anteil an der Varianz des Kriteriums auf.		
R^2 :	$R^2 = r_{1y}^2 + r_{2y}^2$	$R^2 < r_{1y}^2 + r_{2y}^2$		

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Multiple Regression: Komplementarität

- Komplementarität erkennt man daran, dass die Prädiktoren beide positiv (oder beide negativ) mit dem Kriterium korrelieren und negativ untereinander korrelieren. In diesem Fall kann R^2 erheblich größer als im Falle der Unabhängigkeit werden.
- Dies erkennt man an folgendem Beispiel mit zwei vollständig komplementären Prädiktoren, bei der die Vorhersage sogar perfekt ist:

$$\begin{pmatrix} 7 \\ 8 \\ 6 \\ 7 \\ 6 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix} + \begin{pmatrix} 6 \\ 6 \\ 3 \\ 3 \\ 1 \\ 3 \end{pmatrix}$$

$$y = x_1 + x_2$$

$$r_{1y} = .23$$

$$r_{2y} = .38$$

$$r_{12} = -.82$$

$$R^2 = 1$$

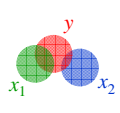
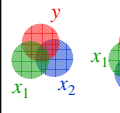
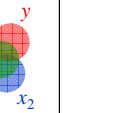
$$r_{1y}^2 + r_{2y}^2 = 0.20$$

- Inhaltliches Beispiel: y = Leistung eines Richters, x_1 = fachliche Kompetenz, x_2 = Verhandlungsgeschick. r_{1y} und r_{2y} könnten positiv sein und r_{12} negativ.

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Zusammenwirken der Prädiktoren (bei m=2)

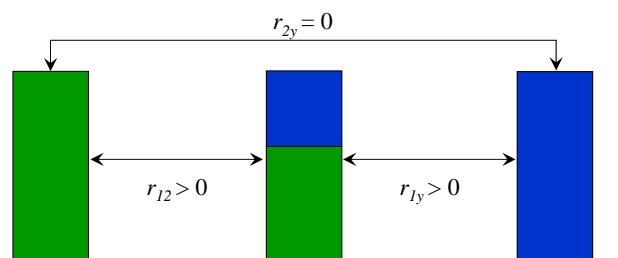
	Unabhängigkeit	Redundanz	Komplementarität	Suppression
Venn-Diagramm:		<div> <div>partiell</div>  </div> <div> <div>vollständig</div>  </div>	nicht darstellbar	
Kennzeichen:	Die Prädiktoren sind unkorreliert.	Die Prädiktoren weisen einen gemeinsamen Anteil an der Varianz des Kriteriums auf	Die Prädiktoren korrelieren beide positiv (oder beide negativ) mit dem Kriterium und negativ untereinander.	
R^2 :	$R^2 = r_{1y}^2 + r_{2y}^2$	$R^2 < r_{1y}^2 + r_{2y}^2$	$R^2 > r_{1y}^2 + r_{2y}^2$	

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Multiple Regression: Suppression

- Eine Suppressor-Variable ist eine Variable, die den Vorhersagebeitrag einer (oder mehrerer) anderer Variablen erhöht, indem sie irrelevante Varianz in dem (den) anderen Prädiktor(en) unterdrückt.



Prädiktor 2
= Suppressor

Prädiktor 1

Kriterium y

Beispiel I: Prüfungsangst

Examensnote

beruflicher Erfolg

Beispiel II: Leseschwindigkeit

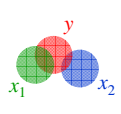
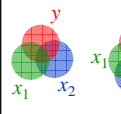
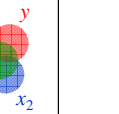
Speed-Test für
Geschichtswissen

Geschichts-
Wissen

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Zusammenwirken der Prädiktoren (bei m=2)

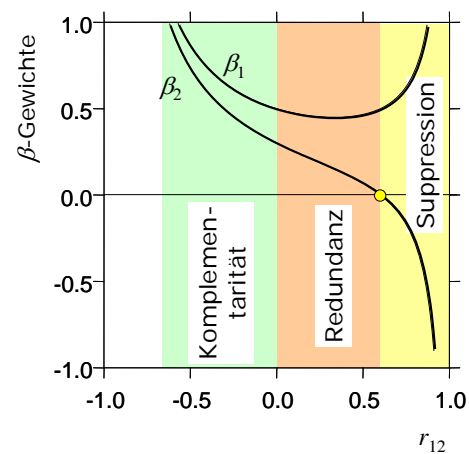
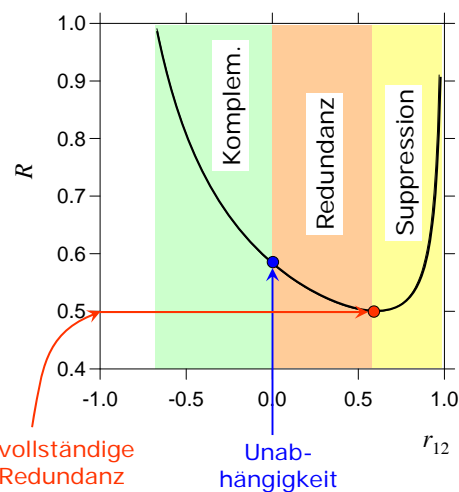
	Unabhängigkeit	Redundanz	Komplementarität	Suppression
Venn-Diagramm:		partiell  vollständig 	nicht darstellbar	
Kennzeichen:	Die Prädiktoren sind unkorreliert.	Die Prädiktoren weisen einen gemeinsamen Anteil an der Varianz des Kriteriums auf	Die Prädiktoren korrelieren beide positiv (oder beide negativ) mit dem Kriterium und negativ untereinander.	Der Suppressor korreliert positiv oder zu 0 mit dem Kriterium und erhält ein negatives Beta (oder umgekehrt).
R^2 :	$R^2 = r_{1y}^2 + r_{2y}^2$	$R^2 < r_{1y}^2 + r_{2y}^2$	$R^2 > r_{1y}^2 + r_{2y}^2$	R^2 kann größer als $r_{1y}^2 + r_{2y}^2$ werden (muss aber nicht)

Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel



Zusammenwirken der Prädiktoren (bei m=2)

Abhängigkeit der multiplen Korrelation und der β -Gewichte von der Interkorrelation r_{12} (im Wertebereich $[-.676, .976]$) zweier Prädiktoren mit $r_{y1}=0.50$ und $r_{y2}=0.30$ (aus Darlington, 1990, S. 156ff)



Vorlesung Multivariate Verfahren (WS 2008/2009)
Prof. Dr. Thomas Staufenbiel

