Empirische Forschungsmethoden
Lineare Regression II

Evelyn C. Ferstl

IIG, Abteilung Kognitionswissenschaft

Universität Freiburg

Beim letzten mal…

- Einfache Lineare Regression $\hat{y}_i = a + b \cdot x_i$
- Bestimmung der Regressionskoeffizienten

Steigung der Geraden

$$b_{yx} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Höhenlage

$$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$$

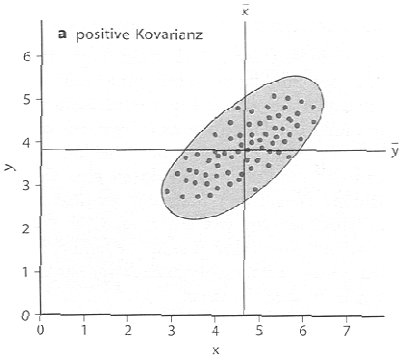
- Bestimmung der Güte der Regression
  - Determinationskoeffizient $r_{xy}^2 = \frac{s_{reg}^2}{s_y^2}$
  - Standardschätzfehler $s_{res} = s_{y.x} = s_y \sqrt{(1 - r_{x.y}^2)}$

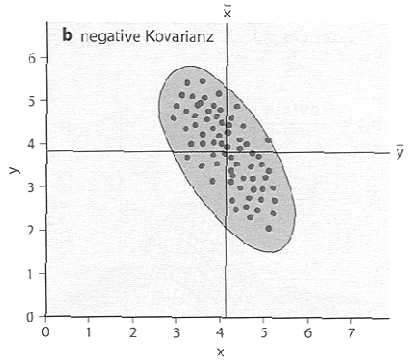
Kovarianz

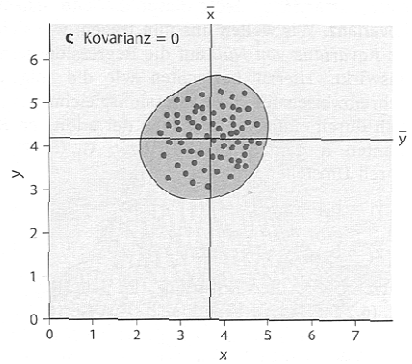
- Kovarianz ist (ähnlich wie die Korrelation) ein Maß für den linearen Zusammenhang zwischen zwei Variablen.
- Sie ist der (nicht- normierte) Mittelwert der Abweichungsprodukte und gibt an wie stark zwei Variablen miteinander variieren (= kovariieren)

$$\mathrm{cov} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n}$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot s_x \cdot s_y} = \frac{\mathrm{cov}}{s_x \cdot s_y}$$

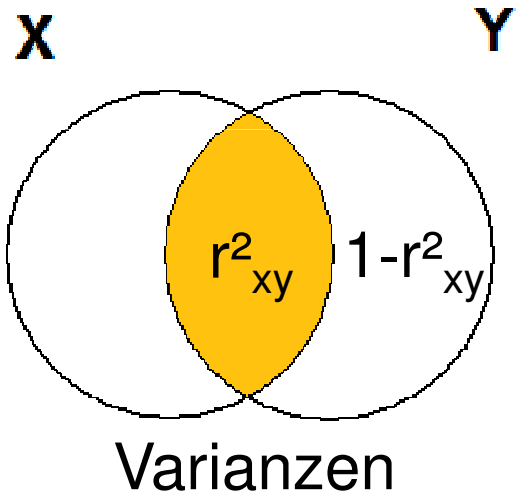






Korrelation

- Korrelationen können nur als *Koinzidenzen* interpretiert werden, nicht als kausaler Wirkmechanismus!
  - Koinzidenz = Zusammenfallen, gemeinsame Auftreten
  - Beispiel: Anzahl der Kinder ~ Anzahl der Störche

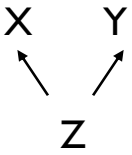


$$s_y^2 = r_{xy}^2 \cdot s_y^2 + (1 - r_{xy}^2) \cdot s_y^2$$

- Determinationskoeffizient r^2_{xy} gibt den Teil der Varianz von Y an, der sich durch X erklären lässt.

Mögliche Interpretation von Korrelationen

- Einseitige Steuerung
  - X bewirkt Y (oder Y bewirkt X) $X \rightarrow Y$ (oder $Y \rightarrow X$)
  - Bsp.: X = Temperament der Mutter, Y = Temperament der Tochter
- Gegenseitige Steuerung
  - X wirkt auf Y, und Y wirkt auf X zurück $X \rightleftarrows Y$
  - Bsp.: X = Sympathie, Y = Kontakthäufigkeit
- Drittseitige Steuerung
  - X und Y hängen von einer dritten Variablen Z ab
  - Bsp.: X = Selbst-, Y = Fremdeinschätzung
    Z = tatsächliche Leistung



- Komplexe Steuerung
  - Das Bedingungsgefüge (A B C … X) bewirkt Y
  - Bsp.: Y = Leistung in einem Beruf
    X = Begabung für den Beruf,
    andere Faktoren (A=Einstellung, B=Erfahrung, C=Umfeld, etc.)

Multiple Korrelation (Leonhart, 2004)

- Speiseeiskonsum/Woche ~ Anzahl Ertrinkender
- Korreliert vermutlich mit der Lufttemperatur
- „Scheinkorrelationen“ entdecken und bereinigen
- Den gemeinsamen Einfluss mehrerer (Prädiktor-)variablen auf eine Kriteriumsvariable erkennen
- Die relativen Anteile des Einflusses einer oder mehrerer Variablen auf eine Kriteriumsvariable bestimmen
- Bildungsforschung:
  - Erfolg eines Unterrichtskonzepts
  - Motivation der Lernenden
  - Qualität der Unterrichtsmaterialien

Varianzen

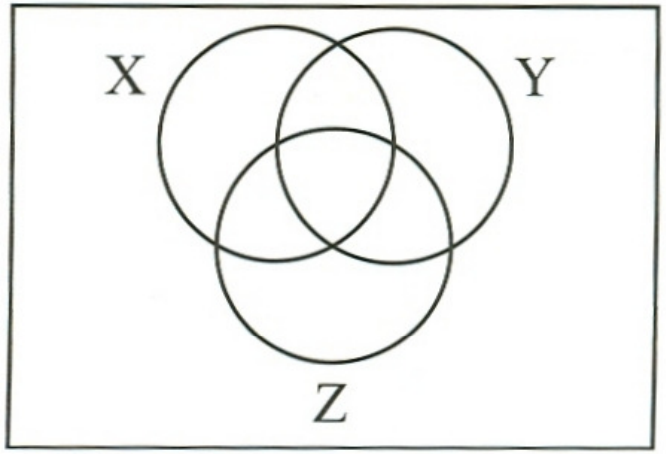


Abbildung 13.1: Ein Venn-Diagramm für den Zusammenhang zwischen drei Variablen

aus: Leonhart, 2004

Partialkorrelation $r_{xy.z}$

Korrelation von (x und y), ohne z

- Die Partialkorrelation $r_{xy.z}$ beschreibt den linearen Zusammenhang von zwei Variablen, aus dem der Einfluss einer dritten Variable eliminiert wurde.
- Sie gibt somit die Korrelation der Variablen x und y an, nachdem das Merkmal z aus x und y *herauspartialisiert* wurde.
- Vorgehen:
  Es werden zwei Regressionen durchgeführt, bei denen jeweils von der Drittvariablen z auf die Merkmale x und y geschlossen wird.
  Anschließend wird die Korrelation der jeweiligen Regressionsresiduen berechnet.
- Erweiterbar auf *Partialkorrelationen höherer Ordnung*.

$$r_{xy.z} = \frac{r_{xy} - r_{yz} \cdot r_{xz}}{\sqrt{(1 - r_{yz}^2) \cdot (1 - r_{xz}^2)}}$$

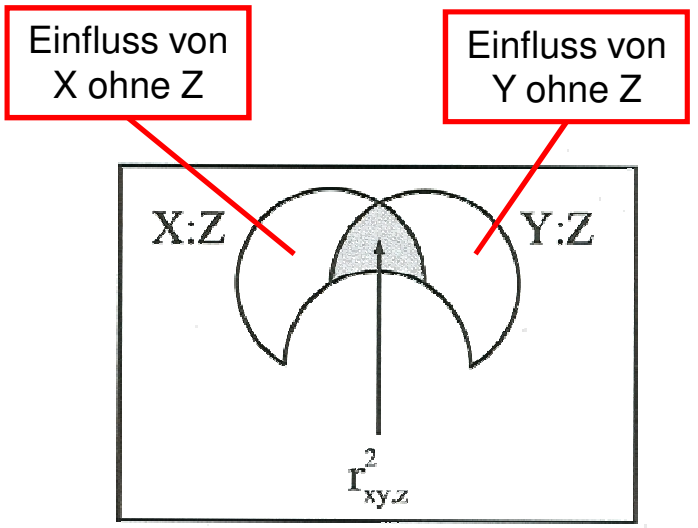


Abbildung 13.2: Partialkorrelation als Venn-Diagramm

aus: Leonhart, 2004

Rechenbeispiel

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot s_x \cdot s_y} = 0.89$$

$$r_{yz} = 0.80 \qquad r_{xz} = 0.77$$

$$r_{xy.z} = \frac{r_{xy} - r_{yz} \cdot r_{xz}}{\sqrt{(1 - r_{yz}^2) \cdot (1 - r_{xz}^2)}}$$

$$r_{xy.z} = \frac{0.89 - 0.80 \cdot 0.77}{\sqrt{(0.36) \cdot (0.41)}} = \frac{0.274}{0.384} = 0.71$$

	x	y	z
VP	Abstraktions-fähigkeit	sensomotorische Koordination	Alter
1	9	8	6
2	11	12	8
3	13	14	9
4	13	13	9
5	14	14	10
6	9	8	7
7	10	9	8
8	11	12	9
9	10	8	8
10	8	9	7
11	13	14	10
12	7	7	6
13	9	10	10
14	13	12	10
15	14	12	9
m	10.9	10.8	8.4
s	2.21	2.40	1.36

Beispiel

In Iowa und Nebraska wurde eine Zufallsstichprobe von 142 älteren Frauen gründlich untersucht (Swanson, Pearl P., Ruth Leverton, Mary R. Gram, Harriet Roberts and Isabel Pesek, Journal of Gerontology 10 (1955) 41 , zitiert von Snedecor, G. W., Statistical Methods, 5. ed., Ames, 1959, p. 430).

Drei der Variablen waren Alter, Blutdruck und die Cholesterin-Konzentration im Blut mit den Korrelationskoeffizienten

$$r_{AB}=0{,}3332, \quad r_{AC}=0{,}5029, \quad r_{BC}=0{,}2495.$$

Da ein erhöhter Blutdruck mit einer vermehrten Cholesterineinlagerung in den Wänden der Blutgefäße zusammenhängen könnte, erscheint es uns interessant, dieser Frage näher nachzugehen. Da B und C mit dem Alter zunehmen, ergibt sich die Frage, ob der an sich schwache Zusammenhang lediglich auf das Alter zurückzuführen ist, oder ob auf jeder Altersstufe ein echter Zusammenhang existiert. Der Alterseffekt wird eliminiert durch die Berechnung von $r_{BC.A}$ [vgl. (5.79c)]:

$$r_{BC.A}=\frac{r_{BC}-r_{AB}\cdot r_{AC}}{\sqrt{(1-r_{AB}^2)(1-r_{AC}^2)}}$$

$$r_{BC.A}=\frac{0{,}2495-0{,}3332\cdot 0{,}5029}{\sqrt{(1-0{,}3332^2)(1-0{,}5029^2)}}=0{,}1005.$$

Für $142-3=139\,FG$ läßt sich diese Korrelation auf dem 5%-Niveau nicht sichern.

Nach Sachs, 1984;
http://www.sgipt.org/wisms/statm/kor/partkor.htm

Partialkorrelation vs. Semipartialkorrelation

- Häufig ist man an derjenigen bivariaten Korrelation interessiert, die sich bei Konstanthaltung einer Variablen ergibt. Diese Korrelation heißt **Partialkorrelation**.
- Ein Beispiel: Will man den korrelativen Zusammenhang zwischen den Variablen „Größe“ und „Gewicht“ bei Kindern ermitteln, sollte man die Variable „Alter“ als Ganzes auspartialisieren.
- Möchte man jedoch wissen wieviel Varianz der Variablen Y nur durch X erklärt werden kann, darf man nur die gemeinsame Varianz von Y und Z auspartialisieren. Man spricht dann von einer **Semipartialkorrelation**.
- Ein Beispiel: Will man den korrelativen Zusammenhang zwischen den Variablen „körperliche Gesundheit“ und „akademische Leistung“ ermitteln, wird man den Einfluss der Variable „in sportliche Betätigungen investierte Zeit“ auf die Variable „körperliche Gesundheit“ kontrollieren.

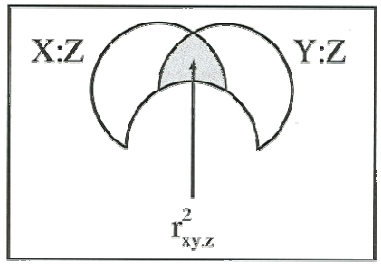


Abbildung 13.2: Partialkorrelation als Venn-Diagramm

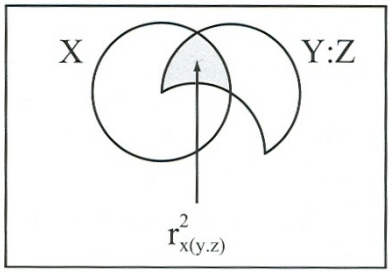


Abbildung 13.3: Semipartialkorrelation als Venn-Diagramm

aus: Leonhart, 2004

Semipartialkorrelation $r_{x(y,z)}$

Korrelation von x und (y ohne z)

- Eine Semipartialkorrelation $r_{xy.z}$ ist die Korrelation eines Residuums einer ursprünglichen Variablen, nachdem z **nur** aus y herauspartialisiert wurde. Mit $r^2_{x(y,z)}$ kann der Varianzanteil bestimmt werden, den eine Variable allein am Kriterium erklärt.
- Über die Semipartialkorrelation kann bestimmt werden, ob es sinnvoll ist, eine weitere Variable z in die Regressionsgleichung aufzunehmen, da über sie zusätzliche Varianz an x erklärt wird (*inkrementelle Validität*). Siehe auch multiple Regressionsanalyse.
- Erweiterbar auf *Semipartialkorrelationen höherer Ordnung*.

$$r_{x(y,z)} = \frac{r_{xy} - r_{yz} \cdot r_{xz}}{\sqrt{1 - r_{yz}^2}}$$

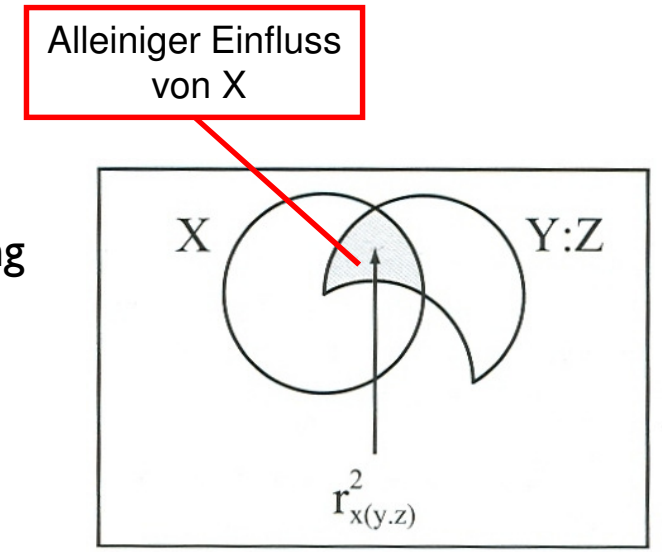


Abbildung 13.3: Semipartialkorrelation als Venn-Diagramm

aus: Leonhart, 2004

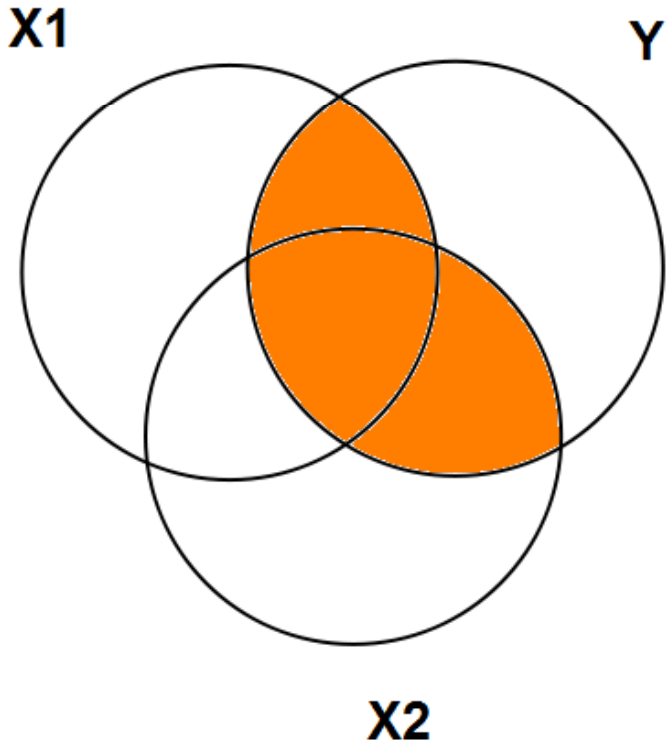
Multiple Korrelation $R_{y.x1x2}$

- Bisher: Zusammenhang zwischen zwei Variablen und einer herauspartialisierten dritten.
- Der **multiple Korrelationskoeffizient** erfasst den Zusammenhang zwischen **mehreren** Prädiktorvariablen x_i und einem Kriterium y.
- Er entspricht damit der Produkt-Moment-Korrelation zwischen dem von mehreren Prädiktorvariablen gemeinsam vorhergesagten Wert des Kriteriums und dem tatsächlichen Kriteriumswert (Korrelation von y mit $\hat{y}$).
- Beispiel für 2 Prädiktorvariablen:

$$R_{y.x1x2} = \sqrt{\frac{r_{x1y}^2 + r_{x2y}^2 - 2 \cdot r_{x1x2} \cdot r_{x1y} \cdot r_{x2y}}{1 - r_{x1x2}^2}}$$

Multiple Korrelation $R_{y.x1x2}$

$$R_{y.x1x2} = \sqrt{\frac{r_{x1y}^2 + r_{x2y}^2 - 2 \cdot r_{x1x2} \cdot r_{x1y} \cdot r_{x2y}}{1 - r_{x1x2}^2}}$$



Multiple Korrelation $R_{y.x1x2}$

- Speziallfall: x1 und x2 sind unkorreliert
- $r_{x1x2} = 0$

$$R_{y.x1x2} = \sqrt{\frac{r_{x1y}^2 + r_{x2y}^2 - 2 \cdot r_{x1x2} \cdot r_{x1y} \cdot r_{x2y}}{1 - r_{x1x2}^2}} =$$

$$= \sqrt{\frac{r_{x1y}^2 + r_{x2y}^2}{1}}$$

Multiple Lineare Regression

- Definition: Die **multiple Regression** ist eine lineare Regression mit mehreren Prädiktoren. Sie ist somit eine **Erweiterung der einfachen linearen Regression:**

Einfache Lineare Regression

$$\hat{y}_i = b \cdot x_i + a$$

- Wie dort wird mit der Methode der kleinsten Quadrate die bestmögliche Vorhersage mit einem möglichst geringen Vorhersagefehler angestrebt.

Multiple Lineare Regression

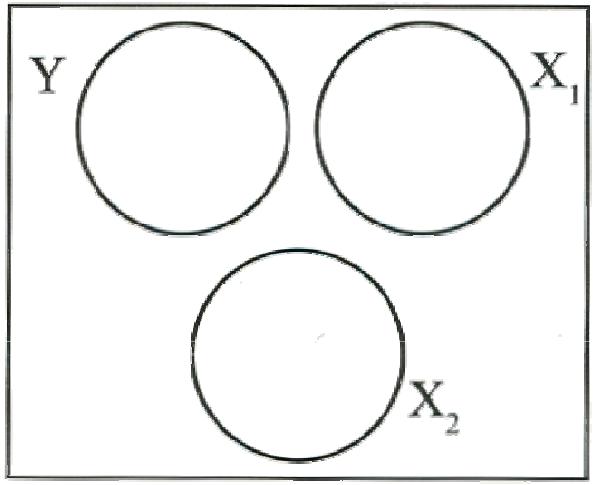
$$\hat{y}_i = b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \ldots + b_k \cdot x_{ik} + a_{1.23\ldots k}$$

Multiple Korrelation: Beispielhafte Zusammenhänge

- Um einen ersten Eindruck zu bekommen welche Prädiktoren geeignet sind um die Werte der Kriteriumsvariablen vorherzusagen, erstellt man eine **Korrelationsmatrix**.
- Null-Korrelation

Tabelle 13.1: Null-Korrelation

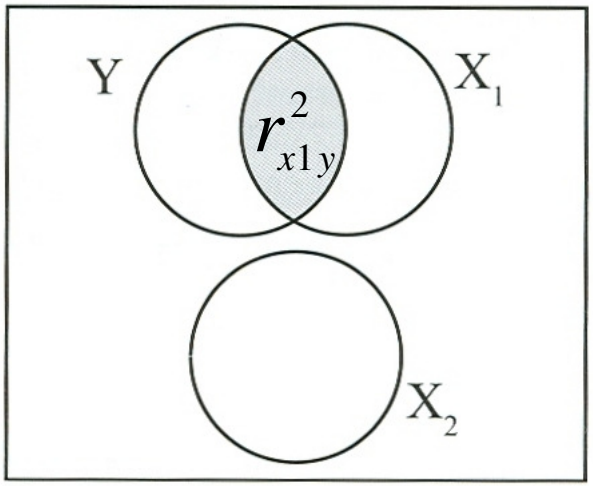
	y	x_1	x_2
y	1,00	,00	,00
x_1		1,00	,00
x_2			1,00



- Nur ein sinnvoller Prädiktor

Tabelle 13.2: Ein sinnvoller Prädiktor

	y	x_1	x_2
y	1,00	,60	,00
x_1		1,00	,00
x_2			1,00



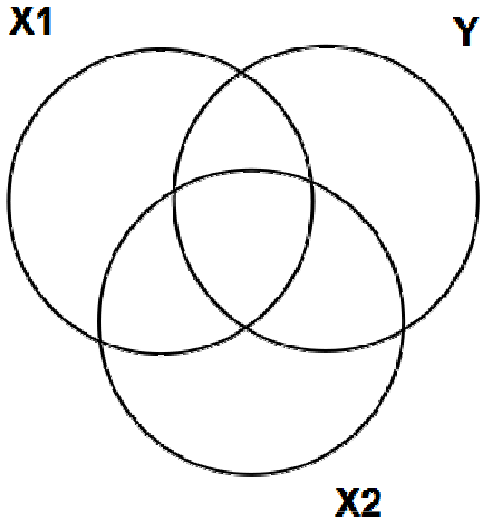
$$r^2_{x1y} = 0.6^2 = 0.36$$

$$\hat{y} = b \cdot x_1 + a$$

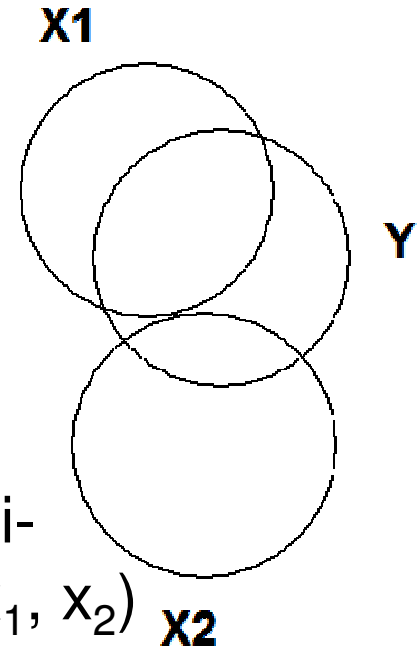
Multikollinearität

- Wenn die Prädiktorvariablen miteinander korrelieren, spricht man von **Multikollinearität**.
- Das heißt die Prädiktorvaiablen sind **nicht** unabhängig von einander.
- Da fast immer eine (wenn auch kleine) Korrelation vorliegt stellt sich meistens die Frage nach der **Stärke** der vorliegenden Multikollinearität. (nicht danach ob sie überhaupt vorliegt)
- Beispiele
  - Alter und Einkommen
  - Gelaufener Weg und benötigte Zeit

Multikollinearität (x_1, x_2)



keine Multi-kollinearität (x_1, x_2)



Inkrementelle Validität

(inkrementell = „in kleinen Stufen hinzukommend“)

- Eine Variable besitzt inkrementelle Validität, wenn ihre Aufnahme als **zusätzlicher Prädiktor** in einer Regression mit mehreren Prädiktoren den **Anteil der aufgeklärten Varianz** im Kriterium signifikant **erhöht**.

Tabelle 13.3: Inkrementelle Validität

	y	x_1	x_2
y	1,00	,60	,45
x_1		1,00	,30
x_2			1,00

$$r^2_{x1\,y} = 0.6^2 = 0.36$$

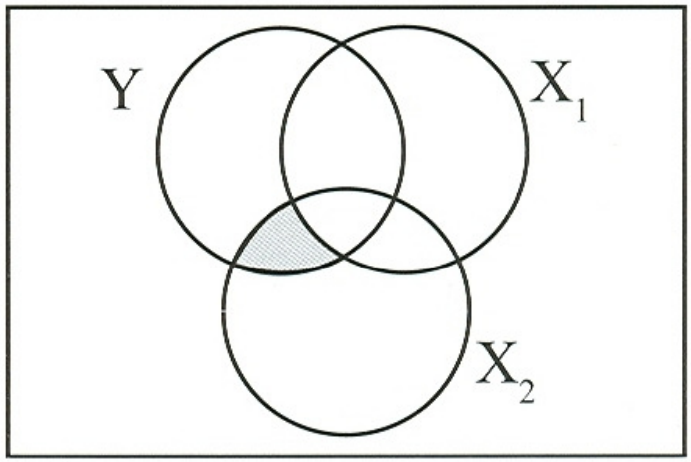


Abbildung 13.6: Zusätzlich erklärte Varianz

- x_1 hat mit einer Varianzaufklärung von 36 % die beste **alleinige** Vorhersagekraft.
- Durch Hinzunahme von x_2 wird die Vorhersage verbessert. x_2 verfügt also über inkrementelle Validität

Inkrementelle Validität - Rechenbeispiel

Tabelle 13.3: Inkrementelle Validität

	y	x_1	x_2
y	1,00	,60	,45
x_1		1,00	,30
x_2			1,00

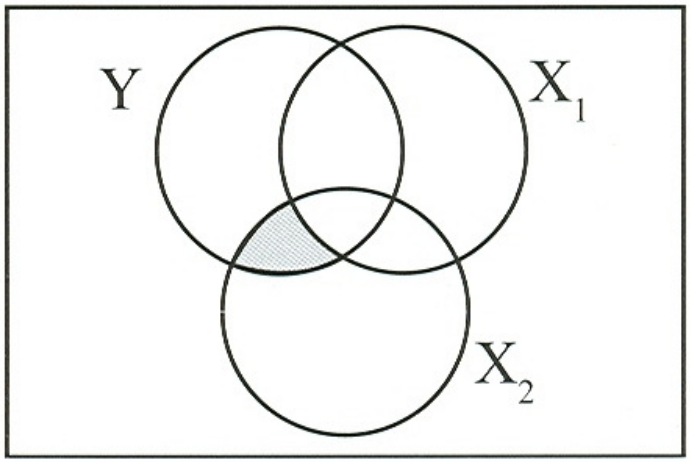


Abbildung 13.6: Zusätzlich erklärte Varianz

$$r^2_{x1y} = 0.6^2 = 0.36$$

$$R_{y.x1x2} = \sqrt{\frac{r^2_{x1y} + r^2_{x2y} - 2 \cdot r_{x1x2} \cdot r_{x1y} \cdot r_{x2y}}{1 - r^2_{x1x2}}}$$

$$R_{y.x1x2} = \sqrt{\frac{.60^2 + .45^2 - 2 \cdot .30 \cdot .60 \cdot .45}{1 - .30^2}}$$

$$R_{y.x1x2} = \sqrt{\frac{.36 + .2025 - .162}{.91}} = .6634$$

$$R^2_{y.x1x2} = .6634^2 = .4401$$

- x_1 hat mit einer Varianzaufklärung von r^2=36 % die beste alleinige Vorhersagekraft.
- Durch Hinzunahme von x_2 wird die Vorhersage deutlich verbessert auf 44%. x_2 verfügt also über inkrementelle Validität.

Signifikanztest (Bortz, 1999, S. 447)

- Wie groß muss die zusätzlich erklärte Varianz sein, damit es sich lohnt sie zusätzlich ins Modell mit aufzunehmen?
- Beispiel: eine Variable ist bereits ins Modell aufgenommen (k), eine zweite soll zusätzlich aufgenommen werden (p).
- H_0 : zusätzlich aufgenommene Variable bringt keine Erhöhung der erklärten Varianz.

$$F = \frac{\left(R^2_{y.x1x2} - r^2_{xy}\right)(n-k-p-1)}{\left(1-R^2_{y.x1x2}\right)p}$$

$n =$ Anzahl der Fälle

$k =$ Anzahl der bereits aufgenommenen Variablen (hier = 1)

$p =$ Anzahl der zusätzlichen aufgenommenen Variablen (hier = 1)

$$df_1 = p$$

$$df_2 = n-k-p-1$$

Rechenbeispiel

$$F = \frac{\left(R^2_{y.x1x2} - r^2_{xy}\right)(n-k-p-1)}{\left(1-R^2_{y.x1x2}\right)p} = \frac{(.44-.36)(40-1-1-1)}{(1-.44)} = 5.29$$

$n =$ Anzahl der Fälle (**angenommen = 40**)

$k =$ Anzahl der bereits aufgenommenen Variablen (hier = 1)

$p =$ Anzahl der zusätzlichen aufgenommenen Variablen (hier = 1)

$$df_1 = p$$

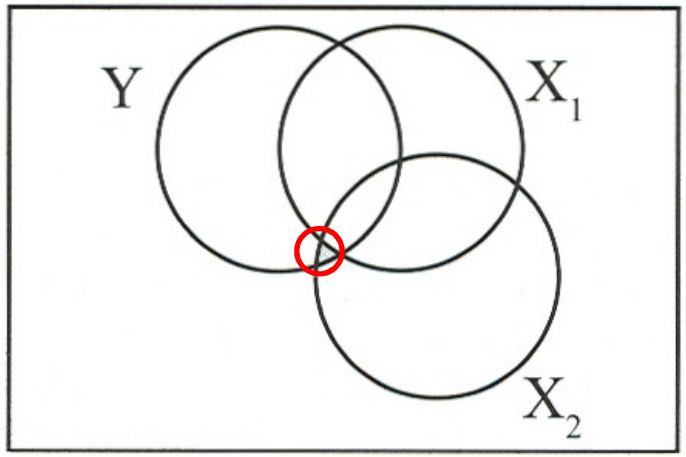
$$df_2 = n - k - p - 1$$

- Kritischer F-Wert$_{\alpha=5\%}$ für $df_1=1$, $df_2=37$ ist < 5.29
- Wir lehnen die H_0, dass die zusätzliche Variable keine zusätzliche varianz erklärt ab und nehmen die Variable in das Modell auf.

Keine inkrementelle Validität (Beispiel)

Tabelle 13.4: Keine inkrementelle Validität

	y	x_1	x_2
y	1.00	,50	,30
x_1		1,00	,50
x_2			1,00



$$R_{y.x1x2} = \sqrt{\frac{r_{x1y}^2 + r_{x2y}^2 - 2 \cdot r_{x1x2} \cdot r_{x1y} \cdot r_{x2y}}{1 - r_{x1x2}^2}}$$

$$R_{y.x1x2} = \sqrt{\frac{.50^2 + .30^2 - 2 \cdot .50 \cdot .50 \cdot .30}{1 - .50^2}}$$

$$R_{y.x1x2} = \sqrt{\frac{.25 + .09 - .15}{.75}} = .5033$$

$$R_{y.x1x2}^2 = .503^2 = .2533$$

- x_1 hat mit einer Varianzaufklärung von r^2_{x1y}=25,0 % die beste **alleinige** Vorhersagekraft.
- Durch Hinzunahme von x_2 wird die Vorhersage verbessert auf 25,3 %. Dies stellt keine signifikante Vergrößerung der erklärten Varianz dar. Ergo: **Keine inkrementelle Validität** von x_2

Suppressor-Effekt

- Ein **Suppressoreffekt** ist gegeben, wenn die Hinzunahme einer Variablen x_2 die Vorhersage verbessert, obwohl dieser Prädiktor selbst nicht mit dem Kriterium y korreliert.
- X_2 unterdrückt die für die Vorhersage von Y irrelevante Varianz
- Verhältnis der gemeinsamen Varianz von X_1 und Y zur Gesamtvarianz von X_1 wird größer.

Tabelle 13.5: Suppressor-Effekt

	y	x_1	x_2
y	1,00	,55	,00
x_1		1,00	,55
x_2			1,00

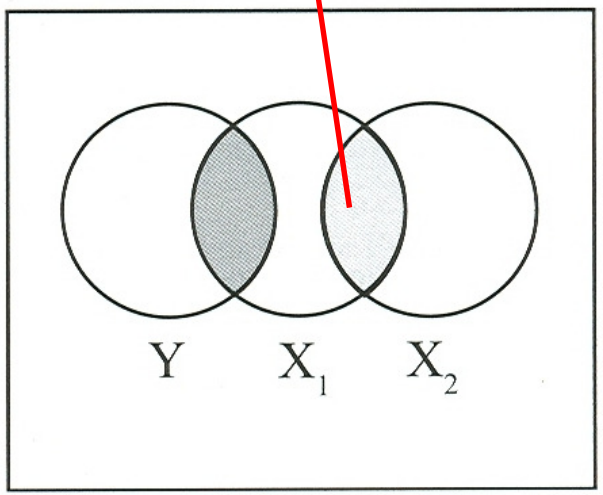


Abbildung 13.8: Suppressoreffekt

$$r^2_{x1y} = 0.55^2 = 0.30$$

$$R_{y.x1x2} = \sqrt{\frac{r^2_{x1y} + r^2_{x2y} - 2 \cdot r_{x1x2} \cdot r_{x1y} \cdot r_{x2y}}{1 - r^2_{x1x2}}}$$

$$R_{y.x1x2} = \sqrt{\frac{.55^2 + .0^2 - 2 \cdot .55 \cdot .0 \cdot .55}{1 - .55^2}} = .6586$$

$$R^2_{y.x1x2} = .43$$

Multiple Lineare Regression

Vorhersage eines Werts $\hat{y}_i$

$$\hat{y}_i = b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \ldots + b_k \cdot x_{ik} + a_{1.23\ldots k}$$

Zusammensetzung von

$$y_i = b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \ldots + b_k \cdot x_{ik} + a_{1.23\ldots k} + e_i$$

Standardisierte Form

$$\hat{z}_{yi} = \beta_1 \cdot z_{i1} + \beta_2 \cdot z_{i2} + \ldots + \beta_k \cdot z_{ik}$$

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

- Die b-Gewichte geben die Einflußstärke und –richtung der einzelnen Variablen an. Da sie jedoch unterschiedlich skaliert sind, kann man nur die standardisierten β -Gewichte miteinander in Beziehung setzen.

Ockhams Rasiermesser

- Prinzip der Sparsamkeit
- „Von mehreren Theorien, die die gleichen Sachverhalte erklären, ist die einfachste allen anderen vorzuziehen.“

$\hat{y} = b_1 \cdot x_1 + a$ ist einfacher als $\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + a$

- Ziel bei der Multiplen Regression ist es immer mit möglichst **wenigen Prädiktorvariablen** möglichst **viel Varianz** der Kriteriumsvariablen aufzuklären.
- Eine Voraussetzung der Multiplen Regression:
  - k < n
  - k = Anzahl der Prädiktorvariablen
  - n = Anzahl der Versuchspersonen

Multiple Lineare Regression

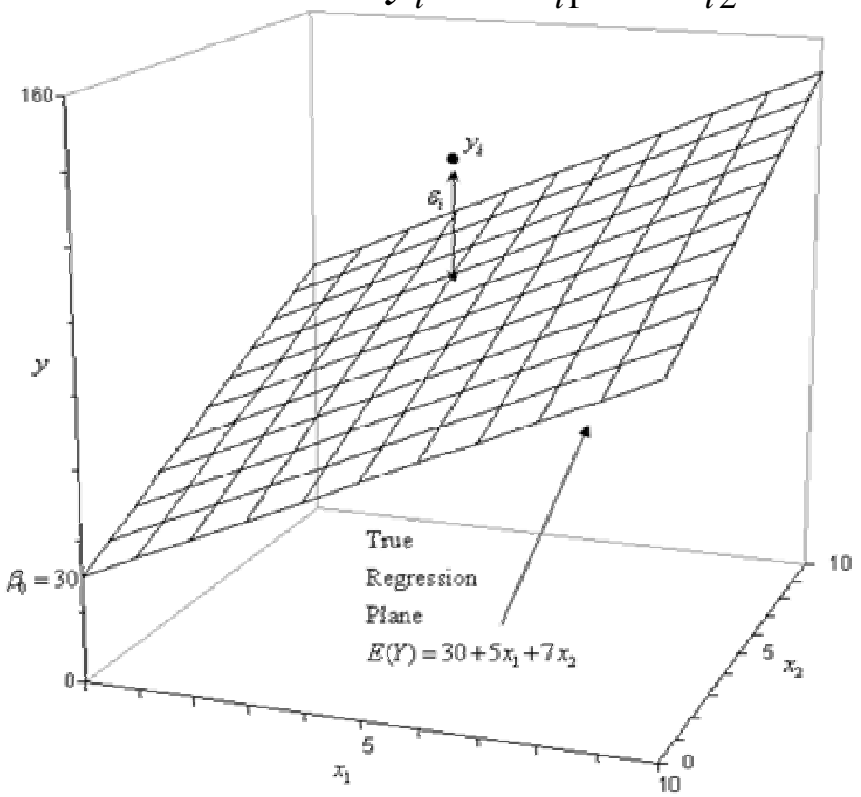
- Bestimmung der b-Gewichte (Partialregressionskoeffizienten)
- Die b-Gewichte sind eine Maß für das relative Gewicht der Prädiktoren für die Vorhersage.

$$\hat{y}_i = 5x_{i1} + 7x_{i2} + 30$$

$$b_{x1} = \frac{s_y}{s_{x1}} \cdot \frac{r_{yx1} - r_{yx2} \cdot r_{x1x2}}{1 - r_{x1x2}^2}$$

$$b_{x2} = \frac{s_y}{s_{x2}} \cdot \frac{r_{yx2} - r_{yx1} \cdot r_{x1x2}}{1 - r_{x1x2}^2}$$

$$a = \bar{y} - (b_{x1} \cdot \bar{x}_1 + b_{x2} \cdot \bar{x}_2)$$



http://www.weibull.com/DOEWeb/multiple_linear_regression_analysis.htm

Strategien der Auswahl von Variablen für multiple Regression

- b-Gewichte sind nicht unabhängig von einander.
  Man will möglichst viel Varianz mit möglichst wenigen Variablen aufklären.
  Ein (oft großer) Teil der Varianz ist aber mehreren Variablen gemeinsam.
  Welche Variablen auswählen?

$$\hat{y}_i = b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \ldots + b_k \cdot x_{ik} + a_{1.23\ldots k}$$

- Die a-priori Auswahl
  Theorie- und evidenzgeleitet werden inhaltlich bedeutsame Prädiktoren in die Regressionsgleichung aufgenommen
- Die a-posteriori Auswahl
  Prädiktorenauswahl iterativ, mehrere Regressionsanalysen
  - Alle möglichen Untermengen
  - Vorwärtsselektion
  - Rückwärtsselektion
  - Schrittweise Regression

Strategien der Auswahl von Variablen für multiple Regression

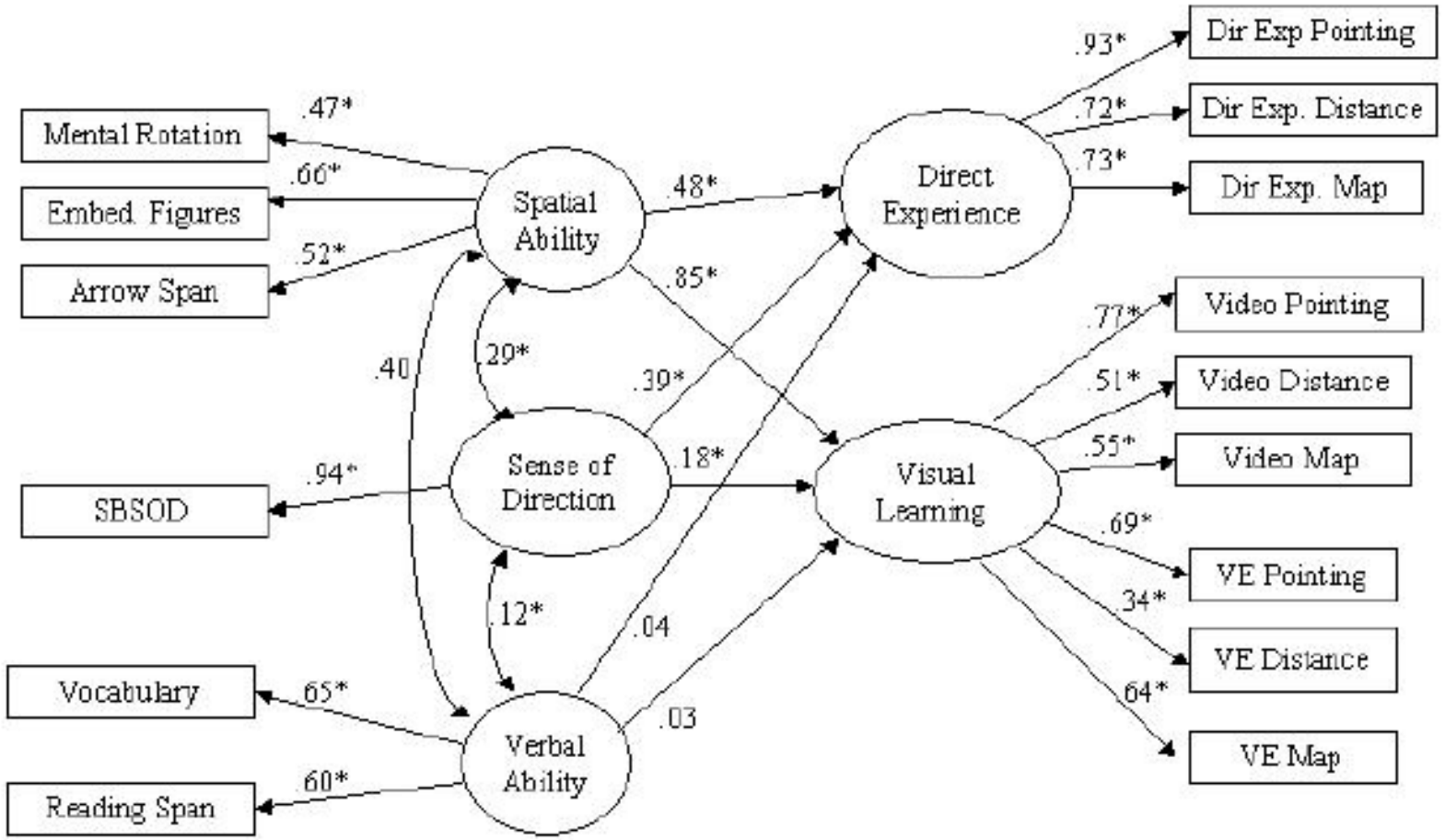
- Alle möglichen Untermengen
  - Für k Prädiktoren 2^k Regressionsanalysen
- Vorwärtsselektion (forward)
  - Zunächst wird der Prädiktor mit der höchsten inkrementellen Validität aufgenommen.
  - Danach iterativ derjenige Prädiktor mit der nächsthöchsten inkrementellen Validität, bis kein Prädiktor mit inkrementeller Validität übrig bleibt.
- Rückwärtsselektion (backward)
  - Zunächst alle potentiellen Prädiktoren aufnehmen.
  - Dann wird der Prädiktor mit der niedrigsten inkrementellen Validität eliminiert.
  - Iterative Wiederholung
- Schrittweise Regression (stepwise)
  - Kombination abwechselnder Vorwärts- und Rückwärtsselektion
  - Nach jeder Inklusion oder Exklusion werden alle beta-Gewichte und Partialkorrelationen neu bestimmt.
  - Vorteil: gut geeignet, um ein Mininum an Variablen zu erzielen
  - Nachteil: häufig nicht sehr stabil in der Kreuzvalidierung
- Bei allen iterativen Verfahren sollte auf jeden Fall eine ***Kreuzvalidierung*** der Regressionsgleichung vorgenommen werden.

Kreuzvalidierung

- Erhebung einer ersten Stichprobe
- Durchführung einer (multiplen) Regression
- Erhebung einer zweiten Stichprobe
- Vorhersage des Kriterium der zweiten Stichprobe durch die Regressionsgleichung aus der ersten Stichprobe
- Vergleich der beobachteten und vorhergesagten Kriteriumswerte der zweiten Stichprobe.
- **Kreuzvalidierung**: Vorhersage des Kriteriums der ersten Stichprobe anhand der Regressionsgleichung aus der zweiten

Strukturgleichungsmodelle

(Beispiel: Hegarty et al. 2006)



Zusammenfassung

- Um die Werte einer Kriteriumsvariablen (y) aus mehreren Prädiktoren (x1, x2, x3…) vorherzusagen, bestimmt man die Gleichung einer multiplen Regression.
- Zunächst erstellt man eine Korrelationsmatrix und ermittelt die Determinationskoeffizienten (R^2)
- Dann werden nacheinander verschiedene Prädiktorvariablen in das Modell (die Gleichung) aufgenommen. Für jede zusätzlich aufgenommene Variable wird geprüft, ob sie genug zusätzliche Varianz aufklärt (F-Test).
- Prinzipiell gilt das Prinzip der Sparsamkeit (Ockham‘s Razor), d.h. mit möglichst wenigen Variablen möglichst viel Varianz aufzuklären.
- Wenn man die Variablen bestimmt hat, die gemeinsam am meisten Varianz aufklären, bestimmt man die Regressionskoeffizienten a und b_i