

Multiple Regression (1)

- Einführung I -

Mit einem Korrelationskoeffizienten und der einfachen linearen Regression können nur **bivariate** Zusammenhänge zwischen **zwei Variablen** untersucht werden. Benutzt man statt dessen **mehrere Variablen** zur Vorhersage, begibt man sich auf das Gebiet der **multivariaten Analyse**. Im Folgenden werden die Grundprinzipien der **multiplen Regression** dargestellt.

Die multiple Regression erlaubt es, **Kausalmodelle** zu untersuchen, mit denen sich folgende Fragen beantworten lassen:

- Wie gut sagt ein Satz von mehreren unabhängigen Variablen **gemeinsam** eine abhängige Variable vorher?
- Wie sehr kann eine **bestimmte Variable** die Vorhersage einer abhängigen Variablen noch **verbessern**, wenn zugleich andere Variablen zur Vorhersage benutzt werden?
- Wie groß ist der Effekt einer Variablen auf eine abhängige Variable, wenn der **Zusammenhang dieser Variablen mit anderen unabhängigen Variablen** des Modells berücksichtigt wird?
- Eignet sich eine bestimmte Variable auch dann zur Vorhersage von Y, wenn eine Reihe von Einflüssen, die in Alternativerklärungen eine Rolle spielen, **statistisch kontrolliert** wird?

Multiple Regression (2)

- Einführung II -

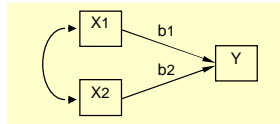
Bei der multiplen Regression

- kennt man den Ausprägungsgrad **mehrerer** Variablen X und möchte den Ausprägungsgrad einer Variablen Y vorhersagen.
- Die Variable, die vorhergesagt werden soll (AV), wird als das **Kriterium** bezeichnet. Die Variablen aufgrund derer vorhergesagt wird (UVs), werden als **Prädiktoren** bezeichnet.
Beispiele:
 - Lassen sich Gewalteinrichtungen (y) aus erfahrener Elterngewalt in der Kindheit (x_1), dem Bildungsgrad (x_2) und dem Migrantenstatus (x_3) vorhersagen?
 - Die Rückfallgeschwindigkeit (y) soll aus der Strafhärte (x_1), der Berufsausbildung des Straftäters (x_2) und seiner Drogenabhängigkeit (x_3) vorhergesagt werden.
- Eine multiple lineare Regressionsanalyse ist nur dann sinnvoll, wenn die Prädiktoren X_j mit dem Kriterium Y **korreliert** sind.
- Es soll diejenige **Linearkombination** gefunden werden, die den multiplen Zusammenhang zwischen X und Y optimal beschreibt.

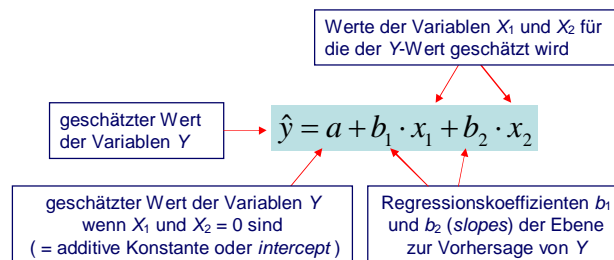
Multiple Regression (3)

- Einführung III -

Das **konzeptuelle Modell** zur Vorhersage einer abhängigen Variablen Y durch zwei unabhängige Variablen X ist (man beachte die Korrelation zwischen X_1 und X_2):



Die mathematische Funktion zur Vorhersage ist eine **Gleichung**, in der die Werte der X-Variablen zur Schätzung der Y-Werte linear kombiniert werden:



Multiple Regression (4)

- Einführung IV -

Die Koeffizienten b_1 und b_2 sind **Gewichtszahlen**. Sie geben an, mit welchem Gewicht der jeweilige Prädiktor in die Vorhersage eingeht. Daraus ergibt sich eine **Linearkombination**, die eine „gewichtete Summe“ der X-Variablen darstellt.

Die Regressionsgewichte werden bei einer Regression nach dem Kriterium der kleinsten Quadrate (*OLS-regression*) so geschätzt, dass sich mit der Linearkombination der Prädiktoren die **höchste Korrelation** zwischen den geschätzten und den beobachteten Werten ergibt.

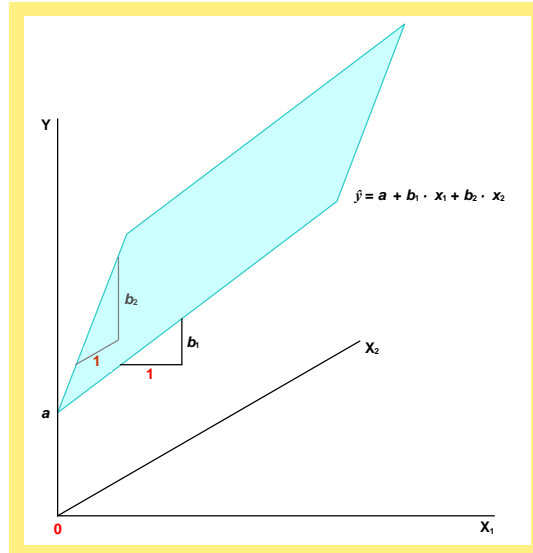
Im Fall von **zwei unabhängigen Variablen** liegen die vorhergesagten Y-Werte auf einer **Ebene**, die durch die Regressionskoeffizienten b_1 und b_2 sowie die Konstante a beschrieben wird.

Im allgemeinen Fall von p unabhängigen Variablen ergeben sich die Regressionskoeffizienten b_1 bis b_p . Die vorhergesagten Y-Werte liegen dann in einem **($p+1$)-dimensionalen Raum**.

Multiple Regression (5)

- Einführung V -

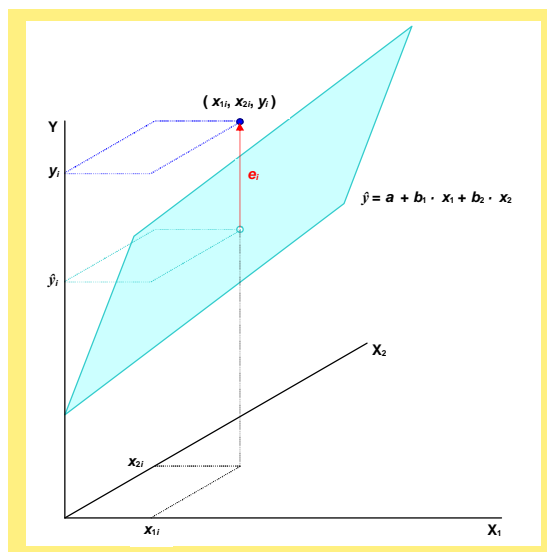
Regressionsebene mit den Steigungen b_1 und b_2 und additiver Konstanter a :



Multiple Regression (6)

- Einführung VI -

Vorhersage eines Y -wertes anhand der Regressionsebene:



Multiple Regression (7)

- Einführung VII -

Obwohl auch hier (wie im Fall der einfachen Regression mit nur einer unabhängigen Variablen) die Regressionskoeffizienten und die Konstante so gewählt werden, dass die Vorhersagefehler minimal sind, werden nicht alle empirischen Werte auf dieser Ebene liegen. Wie im Fall einer einfachen Regression ist der **Schätzfehler** die Distanz zwischen dem geschätzten Y-Wert und dem tatsächlichen Y-Wert des Wert (hier des Tripels (x_{1i}, x_{2i}, y_i)):

$$E = (Y - \hat{Y})$$

Die mittels der Regressionsgleichung minimierte **Schätzfehlervarianz** berechnet sich wie bei der einfachen linearen Regression als das **arithmetische Mittel der quadrierten Residuen** bzw. der quadrierten Abweichungen der geschätzten von den beobachteten Werten:

$$S_{\text{Fehler}}^2 = \overline{e_i^2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{N} = \min$$

Multiple Regression (8)

- Einführung VIII -

Die Gleichungen, um im Fall von zwei unabhängigen Variablen die Koeffizienten a , b_1 und b_2 aus den Rohwerten zu berechnen, sind komplex (und werden bei mehr als zwei Prädiktoren noch komplexer):

$$b_1 = \frac{\sum_{i=1}^n [(x_{1i} - \bar{x}_1) \cdot (y_i - \bar{y})] \cdot \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^n [(x_{2i} - \bar{x}_2) \cdot (y_i - \bar{y})] \cdot \sum_{i=1}^n [(x_{1i} - \bar{x}_1) \cdot (x_{2i} - \bar{x}_2)]}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 \cdot \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^n [(x_{1i} - \bar{x}_1) \cdot (x_{2i} - \bar{x}_2)] \right)^2}$$

$$b_2 = \frac{\sum_{i=1}^n [(x_{2i} - \bar{x}_2) \cdot (y_i - \bar{y})] \cdot \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 - \sum_{i=1}^n [(x_{1i} - \bar{x}_1) \cdot (y_i - \bar{y})] \cdot \sum_{i=1}^n [(x_{1i} - \bar{x}_1) \cdot (x_{2i} - \bar{x}_2)]}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 \cdot \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^n [(x_{1i} - \bar{x}_1) \cdot (x_{2i} - \bar{x}_2)] \right)^2}$$

$$a = \bar{y} - b_1 \cdot \bar{x}_1 - b_2 \cdot \bar{x}_2$$

D.h., zuerst werden die Regressionsgewichte b und dann a bestimmt.

Multiple Regression (9)

- Einführung IX -

Wenn man für die Berechnung der Regressionsgewichte b von **z-standardisierten Werten** ausgeht, vereinfachen sich die Formeln drastisch. Es ergeben sich schließlich Formeln, für die nur noch die Korrelationen aller beteiligter Variablen sowie deren Standardabweichungen benötigt werden.

Man erhält allerdings hierbei zunächst nicht die unstandardisierten Regressionsgewichte b sondern die maßstabsunabhängigen **standardisierten Regressionsgewichte β** , die anschließend mit Hilfe der Standardabweichungen der Variablen in die unstandardisierten Regressionsgewichte b umgewandelt werden können:

$$\beta_1 = \frac{r_{yx1} - r_{yx2} \cdot r_{x1x2}}{1 - r_{x1x2}^2} \quad \beta_2 = \frac{r_{yx2} - r_{yx1} \cdot r_{x1x2}}{1 - r_{x1x2}^2}$$

$$b_1 = \beta_1 \frac{s_y}{s_{x1}} \quad b_2 = \beta_2 \frac{s_y}{s_{x2}}$$

Während die Formeln zur Umwandlung von standardisierten in unstandardisierte Regressionsgewichte allgemeingültig sind, **gelten die Formeln für β_1 und β_2 nur für den Fall von zwei unabhängigen Variablen**. Die Formeln für mehr als zwei unabhängige Variablen sind übersichtlich nur in Matrixschreibweise darstellbar – die Berechnung sollte man Computerprogrammen überlassen.

Multiple Regression (10)

- Unstandardisierte und standardisierte Regressionsgewichte -

Die unstandardisierten Regressionsgewichte b einer multiplen Regression sind maßstabsabhängig. Multipliziert man sie mit dem Verhältnis der Standardabweichungen der jeweiligen UV und der AV, ergeben sich standardisierte Regressionsgewichte β :

$$\beta = b \frac{s_x}{s_y}$$

- Sind die Varianzen aller Variablen gleich (was auch bei **z-Standardisierung** aller Variablen der Fall ist), ist b immer gleich β .
- Die standardisierten Regressionskoeffizienten entsprechen **nur** dann der Korrelation zwischen UV und AV, wenn Korrelationen zwischen den unabhängigen Variablen Null sind (oder wenn das Regressionsmodell nur **eine** UV enthält).
- Die Regressionsgewichte werden in Abhängigkeit aller UV im Modell geschätzt, d.h. sie stellen die „**Nettoeffekte**“ (unter Auspartialisierung der Effekte aller übrigen UV) dar. Sie können deshalb nur im Kontext des jeweiligen Modells interpretiert werden.
- Wenn das Modell nur lineare Terme enthält, können die **standardisierten Regressionsgewichte β** benutzt werden, um den **relativen Einfluss** einer Variablen unter statistischer Kontrolle aller übrigen UV des Modells zu beurteilen. Standardisierte Regressionsgewichte eignen sich **nicht**, um Effekte in Modellen für verschiedene Stichproben zu vergleichen.

Multiple Regression (11)

- Multiple Korrelation und Determinationskoeffizient -

Die Korrelation der durch die multiple Regressionsgleichung vorhergesagten Y-Werte mit den beobachteten Y-Werten wird durch den **multiplen Korrelationskoeffizienten R** erfasst:

$$R = \sqrt{\sum (\beta_i \cdot r_{yi})} = r_{\hat{y}y}$$

R ist immer positiv und nimmt Werte zwischen 0 und 1 an.

Der **multiple Determinationskoeffizient R²** gibt an, welcher Anteil der Varianz der abhängigen Variablen Y aufgrund der Regressionsgleichung (d.h. der gemeinsamen Berücksichtigung aller UV) vorhergesagt bzw. erklärt werden kann:

$$R^2 = \sum (\beta_i \cdot r_{yi}) = \frac{s_{\hat{y}}^2}{s_y^2}$$

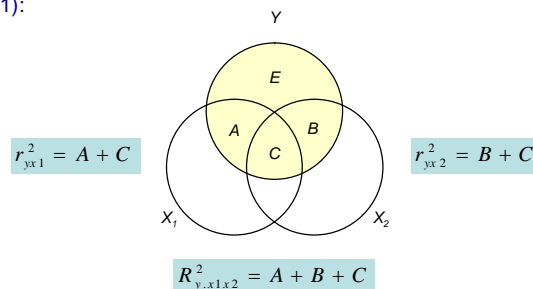
Anhand des multiplen Determinationskoeffizienten R² kann der **Standard-schätzfehler** als Maß der durchschnittlichen Abweichung der beobachteten Y-Werte von den vorhergesagten Y-Werten berechnet werden:

$$s_{\text{fehler}} = s_y \cdot \sqrt{1 - R^2}$$

Multiple Regression (12)

- Erklärte Varianzen -

Die verschiedenen Anteile der erklärten Varianzen in einem multiplen Regressionsmodell lässt sich anhand eines „Venn-Diagramms“ veranschaulichen. Dabei wird von z-standardisierten Variablen ausgegangen (d.h. die Varianz jeder Variablen ist 1):



Man beachte, dass die Fläche **C nicht** als der Teil der Varianz von Y interpretiert werden darf, der von X₁ und X₂ gemeinsam (oder redundant) geschätzt wird! Es ist durchaus möglich, dass C einen negativen Wert annimmt, Varianzen können jedoch niemals negativ sein.

Multiple Regression (13)

- Beispiel (Wirtz & Nachtigall, 2002, S. 176-181) -

Regressionsgleichung:

$$\text{Fitness} = -1.786 - 0.232 \cdot \text{Gewicht} + 0.589 \cdot \text{Lungenvolumen}$$

