DHGS
DEUTSCHE HOCHSCHULE
FÜR GESUNDHEIT & SPORT

FORSCHUNGSMETHODEN

REGRESSIONSANALYSE

MULTIPLE LINEARE REGRESSION

C. G. Meyer-Grant

Überblick

DHGS
DEUTSCHE HOCHSCHULE
FÜR GESUNDHEIT & SPORT

1. Allgemeine Fragestellung
2. Regressionsgleichung
3. Standardpartialregressionskoeffizienten
4. Parameterschätzung
5. Hypothesentests
6. Das allgemeine lineare Modell
7. Kontexteffekte
8. Varianzaufklärung
9. Multiple, Partial- und Semipartialkorrelation
10. Modellselektion

Multiple lineare Regression: Allgemeine Fragestellung

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Wie bei der einfachen Regression versucht man auch bei der multiplen Regression die Ausprägung einer interessierenden metrischen abhängigen Variable (das **Kriterium**) vorherzusagen

- Diese Vorhersage beruht bei der multiplen Regressionsanalyse (im Unterschied zur einfachen Regression) auf der Grundlage von **mehreren** unabhängigen Variablen (den **Prädiktoren**)
  - Das Kriterium wird also auf die Prädiktoren zurückgeführt und man sagt daher auch, dass man eine Regression vom Kriterium auf die Prädiktoren durchführt

Multiple lineare Regression: Regressionsgleichung

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Für jeden Fall (zumeist eine Versuchsperson) wird sowohl ein Messwert des Kriteriums (Y) als auch jeweils ein Messwert pro Prädiktor (X_1, X_2, etc.) erhoben
- Analog zur einfachen Regression kann auch die multiple Regression als Vorhersagegleichung dargestellt werden (hier zur Veranschaulichung mit $k = 2$ Prädiktoren), wobei y'_j der anhand aller Prädiktoren vorhergesagte Wert des Kriteriums ist

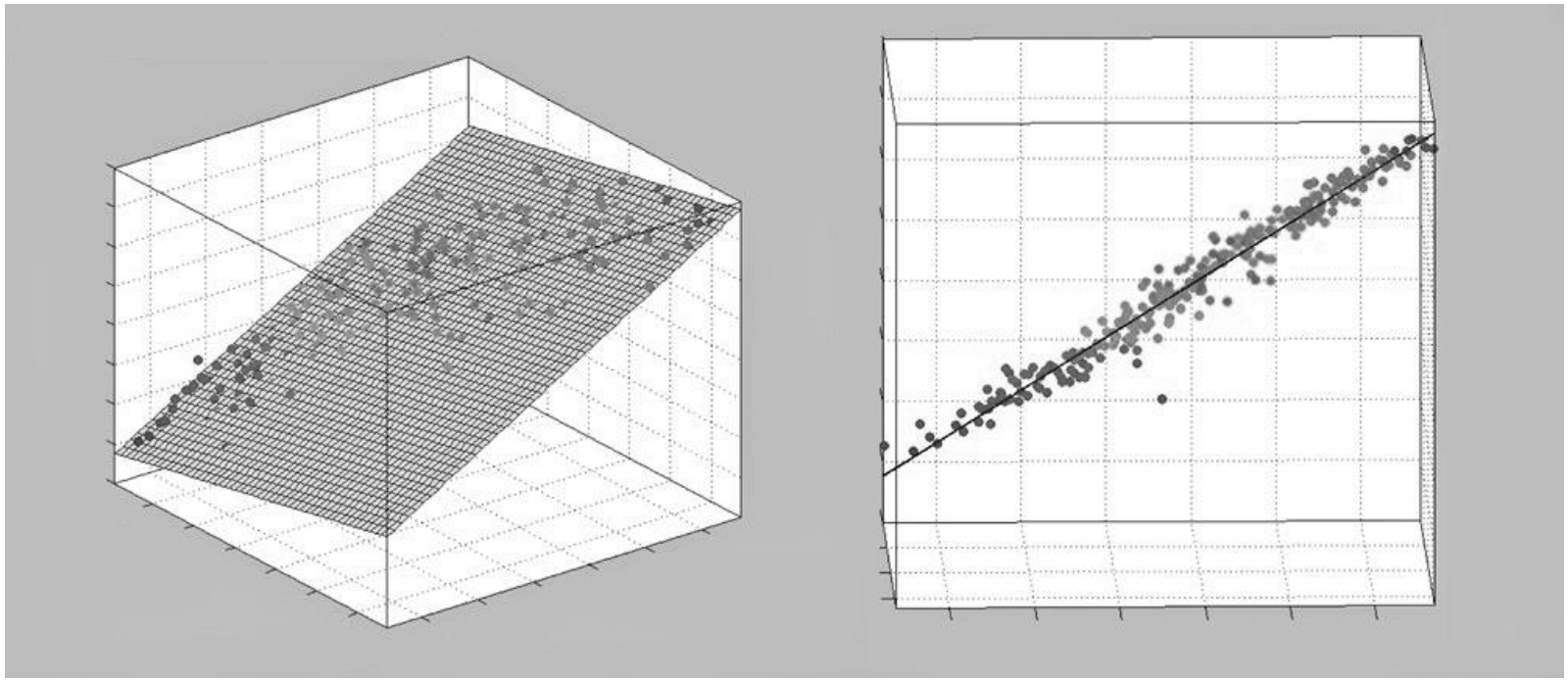
$$y'_j = a + b_1 \times x_{1j} + b_2 \times x_{2j}$$

  - a ist der sog. *Achsenabschnitt* und entspricht dem Wert, den man für das Kriterium vorhersagen würde, wenn alle Prädiktoren den Wert 0 hätten
  - b_i sind die k Partialregressions- bzw. Partialsteigungskoeffizienten bei k Prädiktoren, die jeweils angeben, wie sich das Kriterium verändert, wenn der zugehörige Prädiktor sich um $+1$ verändert und alle weiteren Prädiktoren konstant gehalten werden bzw. sich nicht verändern

Multiple lineare Regression: Allgemeine Fragestellung

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Da hier nicht nur ein Wert, sondern zwei Werte als Prädiktor zur Verfügung stehen, befinden sich alle vorhergesagten Kriteriumswerte nicht auf einer Geraden, sondern auf einer Ebene



Multiple lineare Regression: Regressionsgleichung

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Da aber in der Regel die erhobenen Prädiktoren (hier z.B. X_1 und X_2) nicht alleine für die Variation des Kriteriums verantwortlich sind – es kann schließlich immer noch weitere mögliche Einflussfaktoren geben, die im aktuellen Design nicht berücksichtigt wurden –, wird ein vorhergesagter Wert des Kriteriums (y_j') meist nicht exakt dem tatsächlich erhobenen Kriteriumswert (y_j) entsprechen

- Um die einzelnen Kriteriumswerte mit der Regressionsgleichung exakt modellieren zu können, benötigt man also (wie bei der einfachen linearen Regression) noch einen individuellen Fehlerterm (e_j), das sog. **Residuum**

$$y_j = a + b_1 \times x_{1j} + b_2 \times x_{2j} + e_j$$

Multiple lineare Regression: Standardisierte Partialregressionskoeffizienten

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Um die Partialsteigungskoeffizienten der Prädiktoren innerhalb des Modells unabhängig von den bei der Messung der Variablen gewählten Einheiten miteinander vergleichen zu können, kann es sinnvoll sein die sog. **Standardpartialregressionskoeffizienten** ($\hat{\beta}_i$) zu berechnen, die häufig auch einfach als **Beta-Gewichte** bezeichnet werden

- Dabei führt man im Prinzip eine Z-Transformation mit dem Kriterium und dem entsprechenden Prädiktor durch und relativiert die Steigung somit an den jeweiligen Varianzen

$$b_i \times \frac{s_{x_i}}{s_y} = \hat{\beta}_i$$

  - In einer Regressionsanalyse mit nur einem Prädiktor ($k = 1$) handelt es sich beim Standardpartialregressionskoeffizienten ($\hat{\beta}$) um die Korrelation r_{yx} zwischen dem Kriterium und dem Prädiktor

Multiple lineare Regression: Parameterschätzung

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Wie bei der einfachen Regression, kennt man auch bei der multiplen Regression in aller Regel die „wahren" Beziehungen zwischen den Prädiktoren und dem Kriterium (und damit auch den Achsenabschnitt und die Partialregressionskoeffizienten) in der Grundgesamtheit nicht
- Man ist also darauf angewiesen diese Modellparameter anhand der erhobenen Stichprobe zu schätzen
  - Die Standardpartialregressionskoeffizienten lassen sich in einem Design mit zwei Prädiktoren mit folgenden Formeln schätzen

$$\hat{\beta}_1 = \frac{r_{x_1y} - r_{x_2y} \times r_{x_1x_2}}{1 - r_{x_1x_2}^2} \qquad \hat{\beta}_2 = \frac{r_{x_2y} - r_{x_1y} \times r_{x_1x_2}}{1 - r_{x_1x_2}^2}$$

    - Hierbei sind $r_{x_1x_2}$, r_{x_1y} und r_{x_2y} die paarweisen Korrelationen der drei Variablen X_1, X_2 und Y

Multiple lineare Regression: Parameterschätzung

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Zunächst kann durch Umstellen der bereits bekannten Formel der Partialregressionskoeffizient geschätzt werden

$$b_1 = \hat{\beta}_1 \times \frac{s_y^2}{s_{x_1}^2} \qquad b_2 = \hat{\beta}_2 \times \frac{s_y^2}{s_{x_2}^2}$$

- Möchte man nun den Punkt ermitteln, an dem die Regressionsebene die Y-Achse schneidet, denn genau darum handelt es sich bei a, verwendet man folgende Formel

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

  - Da die beste Vorhersage des Kriteriums bei durchschnittlicher Ausprägung aller Prädiktoren genau der Durchschnitt des Kriteriums sein muss, verläuft die Regressionsebene immer durch den Punkt $P(\bar{y}, \bar{x}_1, \bar{x}_2)$
  - Wenn man von diesem Punkt um $\bar{x}_1$ und $\bar{x}_2$ auf der Regressionsebene „zurückgeht", befindet man sich genau an dem Wert den Y' annimmt, wenn sowohl X_1 als auch X_2 gleich 0 sind

Multiple lineare Regression: Parameterschätzung

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Mithilfe der so geschätzten Modellparameter kann man nun eine Vorhersagegleichung aufstellen, die für jede mögliche Ausprägung der Prädiktoren einen bestimmten Kriteriumswert vorhersagt

$$y'_j = a + b_1 \times x_{1j} + b_2 \times x_{2j}$$

- Bildlich gesprochen wird dabei eine Ebene so in die Daten gelegt, dass die Summe der quadrierten Abweichungen der tatsächlichen Kriteriumswerte von ihren Vorhersagewerten minimal wird (die sog. Methode der kleinsten Quadrate)
  - → Es handelt sich also um ein Verfahren zur Minimierung der Residualquadratsumme (QS_e)

Multiple lineare Regression: Hypothesentests

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Die zentrale Frage bei einer multiplen Regression ist – ähnlich wie bereits bei der einfachen Regression –, ob die Prädiktoren jeweils etwas zu der Vorhersage des Kriteriums beitragen bzw. ob der Partialsteigungskoeffizient für den Prädiktor in der Population ungleich 0 ist
  - Die H_0 für den Prädiktor i würde also hier lauten

$$H_0: \quad b_i = \hat{\beta}_i = 0$$

- Ein Partialsteigungskoeffizient (b_i) ist daher immer dann signifikant von 0 verschieden, wenn sich auch der jeweilige Standardpartialregressionskoeffizient ($\hat{\beta}_i$) signifikant von 0 unterscheidet

  → Die Hypothesentest sind also äquivalent

Multiple lineare Regression: Hypothesentests

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Ob der Standardregressionskoeffizient des Prädiktors signifikant (von 0 verschieden) ist, lässt sich mit folgender Teststatistik prüfen

$$t_{n-k-1} = \frac{\hat{\beta}_i}{sta.\,err.(\hat{\beta}_i)}$$

  - Unter Annahme der H_0 ist diese Teststatistik t-verteilt mit $n - k - 1$ Freiheitsgraden und muss daher mit dem kritischen t-Wert der entsprechenden t-Verteilung für das gewünschte α-Niveau verglichen werden
  - Wobei es sich bei dem Ausdruck im Nenner (sta.err.$(\hat{\beta}_i)$) um den Standardfehler von $\hat{\beta}_i$ handelt
  - Auf die genaue Berechnung von sta.err.$(\hat{\beta}_i)$ soll hier nicht weiter eingegangen werden

Multiple lineare Regression: Hypothesentests

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Wichtiger ist es zu verstehen, was es inhaltlich bedeutet, wenn ein Partialregressionskoeffizient signifikant wird

  - Die Berücksichtigung des entsprechenden Prädiktors im Modell ermöglicht eine überzufällige Verbesserung der systematischen Vorhersage durch das Modell gegenüber einem einfacheren Modell, in dem alle übrigen Prädiktoren (also alle bis auf den interessierenden Prädiktor) enthalten sind
  - Nur wenn das individuelle Erklärungspotential eines Prädiktors im Verhältnis zu den übrigen Prädiktoren des Modells groß genug ist, wird der entsprechende t-Test signifikant

  → Ein und derselbe Prädiktor kann also in einigen Modellen signifikant werden und in anderen nicht, da neben dem interessierenden Prädiktor auch alle übrigen Prädiktoren berücksichtigt werden

Das ALM der multiplen linearen Regressionsanalyse

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Ebenso wie bei der einfachen linearen Regression, kann auch bei der multiplen linearen Regression eine Dekompositionsmatrix (hier beispielhaft für die ersten zwei Fälle dargestellt) erstellt werden

$$Y_j = \mu(a) + \left[\sum_{i=1}^{k} b_i(x_{ij} - \bar{x}_i)\right] + S\left(\sum_{i=1}^{k} b_i \times x_{ij}\right)$$

$$Y_1 = \mu(a) + \left[\sum_{i=1}^{k} b_i(x_{i1} - \bar{x}_i)\right] + S\left(\sum_{i=1}^{k} b_i \times x_{i1}\right)$$

$$Y_2 = \mu(a) + \left[\sum_{i=1}^{k} b_i(x_{i2} - \bar{x}_i)\right] + S\left(\sum_{i=1}^{k} b_i \times x_{i2}\right)$$

...

Das ALM der multiplen linearen Regressionsanalyse

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Hier wird ersichtlich, dass bei der üblichen spaltenweise Berechnung der Quadratsummen der einzelnen Modellkomponenten nur eine gemeinsame Quadratsumme für alle Partialregressionskoeffizienten ermittelt werden kann
  - Da $QS_{\left[\sum_{i=1}^{k} b_i(x_{ij}-\bar{x}_i)\right]}$ etwas sperrig anmutet, wird im Folgenden QS_{Modell} verwendet

- Der letzte Term des Modells entspricht (wie auch bei der einfachen Regression) dem Residuum
  - Anstelle von $QS_{S\left(\sum_{i=1}^{k} b_i \times x_{ij}\right)}$ wird im Folgenden QS_{Fehler} verwendet
  - Bei dieser Quadratsumme handelt es sich damit um die Residualquadratsumme QS_e

Das ALM der multiplen linearen Regressionsanalyse

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Jeder aus den Daten geschätzte Parameter „beansprucht“ dabei erneut je einen Freiheitsgrad, womit $df_{Modell} = k$ gilt, da für dieses Modell zusätzlich zum Achsenabschnitt k Partialregressionskoeffizienten geschätzt wurden
- Für den Fehlerterm verbleiben demnach $n - df_{\mu(a)} - df_{Modell}$ Freiheitsgrade (im Folgenden als df_{Fehler} bezeichnet)

- Mit diesen Informationen lassen sich erneut mittlere Quadrate und F-Werte bestimmen, wobei der entsprechende F-Test die Signifikanz des gesamten Vorhersagemodells (also aller Prädiktoren gemeinsam) überprüft

Die ANOVA-Tabelle für eine lineare Regressionsanalyse

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Alle Ergebnisse dieser Berechnungen können ebenfalls in einer ANOVA-Tabelle dargestellt werden

Quelle	df	QS	MQ	F
μ	1	$QS_{\mu(a)}$	$MQ_{\mu(a)}$	$F_{\mu(a)}$
Modell	df_{Modell}	QS_{Modell}	MQ_{Modell}	F_{Modell}
Fehler	df_{Fehler}	QS_{Fehler}	MQ_{Fehler}	
Total	n	QS_Y		
Kor. Total	$n-1$	$QS_Y - QS_{\mu(a)}$	σ_y^2	

- Mit dem F-Test für die Modellkomponente wird diejenige Nullhypothese getestet, die besagt, dass tatsächlich alle Partialregressionskoeffizienten gleich 0 sind
- Wird F_{Modell} nicht signifikant, so leistet das Modell mit allen Prädiktoren auch keine systematisch besseren Vorhersagen als das triviale Modell $y_i = a + e_i$

Multiple lineare Regression: Kontexteffekte

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Bei Regressionsdesigns mit mehr als einem Prädiktor kann es zu Phänomenen kommen, die zunächst nicht unbedingt intuitiv zugänglich sind

- Wenn es mehrere Prädiktoren gibt, dann kann es (neben der Beziehung zwischen den Prädiktoren und dem Kriterium) zu verschieden gearteten Beziehungen von den Prädiktoren untereinander kommen

- Dies kann wiederum verschiedene Implikationen auf die Interpretation der Regressionsergebnisse haben

Multiple lineare Regression: Redundanz

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Ein erstes scheinbar paradoxes Ergebnismuster einer multiplen Regression ist im wesentlichen eine Kombination der beiden folgenden Befunde

  - Ein bestimmter Prädiktor (X_2) liefert einen signifikanten Beitrag zur Vorhersage des Kriteriums (Y), wenn man ihn als einzigen Prädiktor betrachtet

  → X_2 scheint ein guter Prädiktor zu sein

  - Derselbe Prädiktor liefert hingegen keinen signifikanten Beitrag für die Vorhersage von Y, wenn dieser Prädiktor (X_2) in einem multiplen Regressionsdesign mit dem weiteren Prädiktor X_1 enthalten ist

  → X_2 scheint ein schlechter Prädiktor zu sein

Multiple lineare Regression: Redundanz

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Dieses Phänomen tritt dann auf, wenn die beiden Prädiktoren **redundant** sind – sie also im wesentlichen dieselbe Information zur Vorhersage von Y enthalten

- Die Vorhersage des Modells mit beiden Prädiktoren ist daher auch nicht wesentlich besser als die Vorhersage des Modells mit X_1 alleine

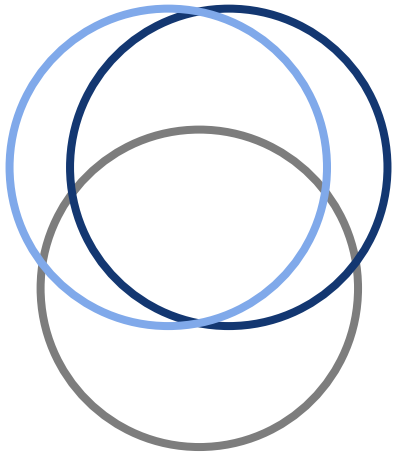
    - Man könnte sich hierzu z.B. vorstellen, die Körpergröße der Tochter (Y) anhand der Körpergröße der Mutter (X_1) und der Körpergröße der Großmutter mütterlicherseits (X_2) vorherzusagen

Multiple lineare Regression: Redundanz

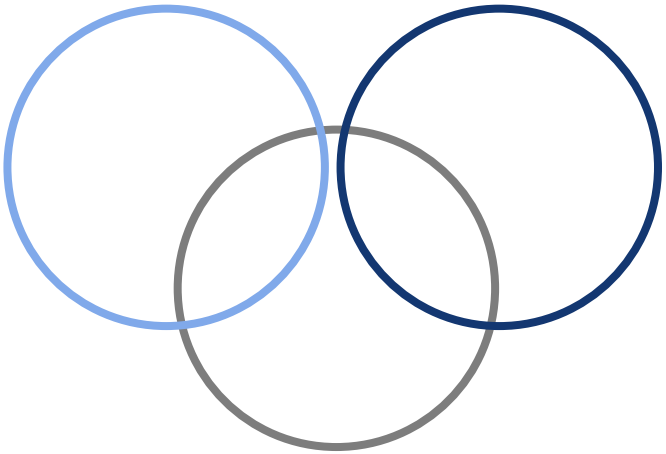
DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

Y
X_1
X_2

Redundanz



Keine Redundanz



Multiple lineare Regression: Multikollinearität

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Bei der multiplen Regression können die einzelnen Regressionskoeffizienten (z.B. b_1) als ein Maß für die Änderung des Kriteriums interpretiert werden, wenn der entsprechende Prädiktor (z.B. X_1) um eine Einheit wächst und alle weiteren Prädiktoren konstant gehalten werden

- Es ist zwar theoretisch immer möglich eine Prädiktorvariable (z.B. X_1) in einer Regressionsgleichung bei Konstanz aller anderen Prädiktorvariablen zu verändern, wenn aber die Prädiktoren untereinander in einer engen lineare Beziehungen zueinander stehen, kann es praktisch unsinnig sein die übrigen Prädiktoren zu „fixieren“
  - In der Praxis würden die anderen Prädiktoren sich ja entsprechend ihrer jeweiligen Beziehung zu X_1 ebenfalls verändern

Multiple lineare Regression: Multikollinearität

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Sind die Prädiktorvariabeln (oder zumindest einige von Ihnen) sehr eng miteinander verbunden, spricht man von Multikollinearität
- Die Regressionsergebnisse sind dann nicht mehr eindeutig zu interpretieren
  - Es wird dann typischerweise unmöglich die separaten Wirkungen der einzelnen Prädiktoren in der Regressionsgleichung zu schätzen
  - Die Regressionskoeffizienten sind dann auch sehr viel anfälliger für Stichprobenfehler, was sich sowohl auf die statistischen Analysen als auch auf die Vorhersagegenauigkeit des Regressionsmodells auswirkt
    - → Es kommt zu beträchtlichen Veränderungen im Regressionsmodell, wenn sich z.B. die Daten geringfügig verändern oder erklärende Variablen eingefügt bzw. weggelassen werden

Multiple lineare Regression: Suppression

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Ein weiteres scheinbar paradoxes Ergebnismuster einer multiplen Regression ist im wesentlichen eine Kombination der folgenden Befunde

  - Ein bestimmter Prädiktor (X_2) steht in keiner bzw. nahezu keiner Beziehung (Nullkorrelation) zum Kriterium (Y) und liefert daher auch keinen signifikanten Beitrag zur Vorhersage von Y, wenn man ihn als einzigen Prädiktor betrachtet

  → X_2 scheint ein schlechter Prädiktor zu sein

  - Derselbe Prädiktor liefert hingegen einen signifikanten Beitrag für die Vorhersage von Y, wenn dieser Prädiktor (X_2) in einem multiplen Regressionsdesign mit dem weiteren Prädiktor X_1 enthalten ist

  → X_2 scheint ein guter Prädiktor zu sein

Multiple lineare Regression: Suppression

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Dieses Phänomen wird **Suppression** genannt, da hier einer der Prädiktoren (X_2) in dem anderen Prädiktor (X_1) Fehleranteile (also solche Anteile, die nichts mit dem Kriterium gemeinsam haben) „unterdrückt" oder genauer gesagt extrahiert und dadurch indirekt den Zusammenhang zwischen X_1 und dem Kriterium erhöht

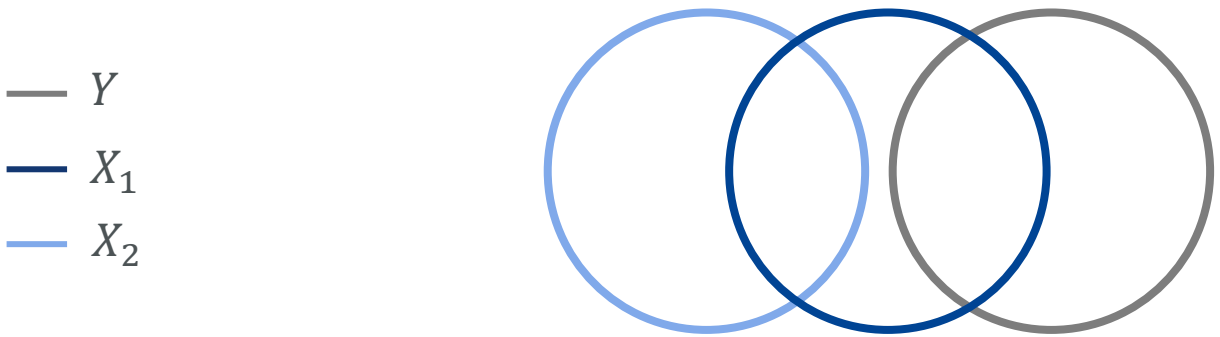
- Die Vorhersage des Modells mit beiden Prädiktoren ist daher wesentlich besser als die Vorhersage des Modells mit X_1 alleine

    - Man könnte sich hierzu z.B. vorstellen, die Körpergröße der Tochter (Y) anhand der der Körpergröße der Mutter (X_1) und der sozioökonomischen Bedingungen unter denen die Mutter aufwuchs (X_2) vorherzusagen

Multiple lineare Regression: Suppression

DHGS
DEUTSCHE HOCHSCHULE
FÜR GESUNDHEIT & SPORT

Suppression



Weitere Kontexteffekte

- Durch Hinzufügen und Entfernen von Prädiktoren verändern sich maßgeblich die Freiheitsgrade des Modells und des Fehlers sowie die Residualvarianz

  - Die df_{Fehler} verringern sich (und die df_{Modell} steigen), wenn ein zusätzlicher Prädiktor aufgenommen wird
  - Dies kann sogar dazu führen, dass zwei Prädiktoren jeweils einzeln in zwei separaten Modellen eine signifikante Vorhersage ermöglichen, ein Modell mit beiden Prädiktoren gemeinsam hingegen nicht signifikant wird

  - Anderseits kann durch zusätzliche Prädiktoren bisher (also ohne diesen Prädiktor) unerklärte Varianz aufgeklärt werden, sodass zuvor nicht signifikante Prädiktoren signifikant werden
  - Zuvor wurden ihre (möglichweise kleinen) Effekte sozusagen vom „Rauschen“ der Fehlervarianz überlagert

Varianzaufklärung

DHGS
DEUTSCHE HOCHSCHULE
FÜR GESUNDHEIT & SPORT

- Wie das η^2 bei der ANOVA entspricht eine quadrierte Korrelation r_{xy}^2 zwischen Prädiktor (X) und Kriterium (Y) dem Anteil der durch den Prädiktor aufgeklärten Varianz an der Gesamtvarianz des Kriteriums um dessen Mittelwert (oder auch $QS_{kor.tot.}$)

- Dies gilt ebenso für sog. multiple, partielle und semipartielle Korrelationen, da es sich dabei grundsätzlich auch um einfache Korrelationen handelt

    - Diese Korrelationen werden allerdings zwischen bestimmten abgeleiteten Variablen gebildet
    - Was sind also multiple, partielle und semipartielle Korrelationen und warum sind sie im Kontext der multiplen Regression von Bedeutung

Multiple lineare Regression: Multiple Korrelation

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Bei einer multiplen linearen Regression (z.B. für den Fall zweier Prädiktoren X_1 und X_2 und einer Kriteriumsvariable Y) lässt sich folgendes lineares Modell aufstellen, wobei e_j den durch X_1 und X_2 nicht erklärbaren Rest (das Residuum) darstellt

$$y_j = a + b_1 \times x_{1j} + b_2 \times x_{2j} + e_j$$

- Die beste Vorhersage (Y_j') für das Kriterium bei gegebenen festen Werten von X_1 und X_2 entspricht folglich

$$y_j' = a + b_1 \times x_{1j} + b_2 \times x_{2j}$$

Multiple lineare Regression: Multiple Korrelation

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

→Die sog. **multiple Korrelation** $R_{y.\,x_1x_2}$ von X_1 und X_2 mit Y ist die einfache Korrelation zwischen den tatsächlichen (Y) und den durch die multiple Regression vorhergesagten (Y') Werten der Kriteriumsvariable

- Die multiple Korrelation kann (für das Beispiel mit $k = 2$ Prädiktoren) auch durch die beteiligten einfachen paarweisen Korrelationen ausgedrückt werden

$$R_{y.\,x_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2 \times r_{x_1x_2} \times r_{yx_1} \times r_{yx_2}}{1 - r_{x_1x_2}^2}}$$

Multiple lineare Regression: Multiple Korrelation

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- $R^2_{y.\,x_1x_2}$ reflektiert damit den Anteil der Gesamtvarianz des Kriteriums (Y), der durch die beiden Prädiktoren (X_1 und X_2) erklärt werden kann

- Natürlich kann durch die Korrelation von Y mit Y' die multiple Korrelation (R) des Kriteriums mit beliebig vielen Prädiktoren gebildet und so selbstverständlich auch der Anteil der durch das entsprechende Modell aufgeklärten Kriteriumsvarianz mittels R^2 berechnet werden

- Dieses R^2 wird gelegentlich auch **Bestimmtheitsmaß** oder **Determinationskoeffizient** genannt und eine alternative Berechnung ist

$$R^2 = \frac{QS_{Modell}}{QS_Y - QS_{\mu(a)}} = 1 - \frac{QS_{Fehler}}{QS_Y - QS_{\mu(a)}}$$

Multiple lineare Regression: Multiple Korrelation

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Um die Nullhypothese zu überprüfen, die besagt, dass dieser multiple Zusammenhang tatsächlich nicht besteht, kann man folgenden empirischen F-Wert berechnen

$$F = \frac{R^2 \times (n - k - 1)}{(1 - R^2) \times k}$$

  - Dieser empirische F-Wert hat k Nenner und $n - k - 1$ Zählerfreiheitsgrade, wobei n der Anzahl der Fälle und k der Anzahl der Prädiktoren entspricht
  - Dieser F-Wert entspricht dem bereits besprochenen F_{Modell} (aus der ANOVA-Tabelle) für ein Modell mit allen Prädiktoren, die in die multiple Korrelation mit dem Kriterium eingegangen sind

Kontexteffekte

- Für $k = 2$ ist es möglich, eine halbwegs übersichtliche und durchaus instruktive Formel für den Standardfehler der Standardpartialregressionskoeffizienten anzugeben

$$sta.err.(\hat{\beta}_1) = \sqrt{\frac{1 - R^2_{y.x_1x_2}}{(1 - r^2_{x_1x_2}) \times (n - k - 1)}}$$

  - Wobei n die Anzahl der Fälle, k die Anzahl der Prädiktoren und $r^2_{x_1x_2}$ die quadrierte Korrelation zwischen den beiden Prädiktoren darstellt
  - $R^2_{y.x_1x_2}$ ist die quadrierte multiple Korrelation zwischen beiden Prädiktoren und dem Kriterium (Anteil aufgeklärter Varianz)

- Hier lässt sich nochmals erkennen, wie die Signifikanz eines Parameters von dem Zusammenhang mit allen übrigen Parametern und der durch das Modell aufgeklärten Gesamtvarianz zusammenhängt

Residualwerte

- Die Variable $Y^* = Y - Y'$ (wobei Y' die lineare Vorhersage von Y durch X darstellt) wird auch als Residuum von Y bezüglich X bezeichnet
  - Man könnte Y^* auch den um X „bereinigten“ Wert von Y nennen
  - Dies lässt sich dadurch verdeutlichen, dass $r_{y^*x} = 0$ gilt

Residualwerte

- Auf diese Weise kann man in einem Regressionsdesign mit zwei Prädiktoren einen Prädiktor (z.B. X_2) sowohl aus dem Kriterium (Y) als auch aus dem anderen Prädiktor (X_1) auspartialisieren

  - Hierbei wäre $X_1^* = X_1 - X_1'$ das Residuum von X_1 bezüglich X_2 (wobei X_1' die lineare Vorhersage von X_1 durch X_2 ist) und $Y^* = Y - Y'$ das Residuum von Y bezüglich X_2 (wobei Y' die lineare Vorhersage von Y durch X_2 ist)

Multiple lineare Regression: Partialkorrelation

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

→Die sog. **Partialkorrelation** $r_{yx_1.x_2}$ von Y mit X_1 unter Auspartialisierung von X_2 ist die einfache Korrelation von Y^* mit X_1^*

- Die partielle Korrelation kann (für das Beispiel mit $k = 2$ Prädiktoren) auch durch die beteiligten einfachen paarweisen Korrelationen ausgedrückt werden

$$r_{yx_1.x_2} = \frac{r_{yx_1} - r_{x_1x_2} \times r_{yx_2}}{\sqrt{1 - r_{x_1x_2}^2} \times \sqrt{1 - r_{yx_2}^2}}$$

Multiple lineare Regression: Semipartialkorrelation

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

→Die sog. **Semipartialkorrelation** $r_{y(x_1.\,x_2)}$ ist definiert als die einfache Korrelation von Y mit X_1^*

- Zu beachten ist hierbei, dass aus X_1 der lineare Einfluss von X_2 herausgerechnet ist, aber nicht aus Y

- Auch die semipartielle Korrelation kann (für das Beispiel mit $k = 2$ Prädiktoren) durch die beteiligten einfachen paarweisen Korrelationen ausgedrückt werden

$$r_{y(x_1.\,x_2)} = \frac{r_{yx_1} - r_{x_1x_2} \times r_{yx_2}}{\sqrt{1 - r_{x_1x_2}^2}}$$

Multiple lineare Regression: Semipartialkorrelation

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- $r^2_{y(x_1.\,x_2)}$ reflektiert denjenigen inkrementellen Anteil an der Gesamtaufklärung der Varianz des Kriteriums (Y) durch den Prädiktor X_1, der über die weitere direkte Beziehung von Y zum Prädiktor X_2 hinausgeht
  - Dies lässt sich auch dadurch verdeutlichen, dass $R^2_{y.\,x_1x_2} = r^2_{yx_2} + r^2_{y(x_1.\,x_2)}$ gilt
    - Verzichtet man auf X_1 als Prädiktor, dann wäre die Abnahme von $R^2_{y.\,x_1x_2}$ gleich $r^2_{y(x_1.\,x_2)}$
    - Es handelt sich sozusagen um die „Kosten“ der Nichtberücksichtigung von X_1
    - $r^2_{y(x_1.\,x_2)}$ entspricht damit der Nützlichkeit (U_{X_1}) des Prädiktors X_1

Multiple lineare Regression: Semipartialkorrelation

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Zwischen der Partial- und der Semipartialkorrelation besteht der folgende Zusammenhang

$$r^2_{yx_1.x_2} = \frac{r^2_{y(x_1.x_2)}}{1 - r^2_{yx_2}}$$

  - Die Partialkorrelation ist also nie kleiner als die entsprechende Semipartialkorrelation
  - Sie setzt die zusätzliche, inkrementelle Varianzaufklärung von Y durch X_1 ins Verhältnis zu der Restvarianz, die nach der Regression von Y auf X_2 noch zu erklären verbleibt

Multiple lineare Regression: Modellselektion

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Wie soll man sich im Zweifelsfall für ein Modell entscheiden? Welches Modell ist besser? Welcher Prädiktor ist nötig und welcher überflüssig?

- Generell gilt:
    - Das Gesamtmodell sollte eine signifikante Vorhersage erlauben
    - Jeder Prädiktor eines Modells sollte einen signifikanten Beitrag zu der Vorhersage dieses Modells leisten
    - Die Auswahl der Prädiktoren sollte theoretisch begründbar und plausibel sein

Multiple lineare Regression: Modellselektion

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Je größer die durch das Modell aufgeklärte Varianz (R^2), desto besser ist das Modell in der Lage Y vorherzusagen
    - Allerdings sollte man anmerken, dass durch zusätzliche Prädiktoren der Determinationskoeffizient immer größer wird (oder zumindest nicht kleiner werden kann) und zwar unabhängig davon, ob die zusätzlichen unabhängigen Variablen einen Beitrag zur Erklärungskraft liefern

Multiple lineare Regression: Modellselektion

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Es gibt daher eine Vielzahl von Maßen der Modellgüte die zusätzliche Parameter „bestrafen“ wie z.B. das freiheitsgradbezogene oder korrigierte Bestimmtheitsmaß

$$\bar{R}^2 = 1 - \frac{MQ_{Fehler}}{\sigma_y^2} = 1 - \left(1 - R^2\right) \times \frac{n-1}{df_{Fehler}}$$

- Dies ist plausibel und im Einklang mit der Sparsamkeitsanforderung
  - Ein einfacheres Modell mit vergleichbarer Aussagekraft wie ein komplizierteres ist zu bevorzugen

Multiple lineare Regression: Modellselektion

DHGS
DEUTSCHE HOCHSCHULE FÜR GESUNDHEIT & SPORT

- Ein anderes (sehr aussagekräftiges, wenn auch seltener angegebenes) Bewertungsmaß der Modellgüte ist der **Standardfehler der Regression** (engl. standard error of the regression, SR)

$$SR = \sqrt{MQ_{Fehler}}$$

- Je größer er ist, desto schlechter beschreibt die Regressionsgerade die Verteilung der Messwerte
- Da es sich bei der Wurzelfunktion um eine streng monotone Transformation auf $\mathbb{R}_+$ handelt, kann für den Vergleich von Modellen auch einfach MQ_{Fehler} verwendet werden

Fragen?

DHGS
DEUTSCHE HOCHSCHULE
FÜR GESUNDHEIT & SPORT

Vielen Dank!