Forschungsmethodik: Interventions- und Evaluationsforschung theoretisch begründen

Vorlesung Biostatistik, WS 2026/2027

Dr. Marcus Wagner, IMISE

Vorlesung V am 10.11.2026

Inhaltsverzeichnis

8. Lineare Regression

8.1. Regression bei vollständiger Information

8.2. Regression für Stichproben

9. Allgemeine lineare Modelle

9.1. Grundgedanken

9.2. Einbindung von Kovariaten

9.3. GLM für Stichproben

9.4. Bewertung der Modellgüte

9.5. Interaktionsterme

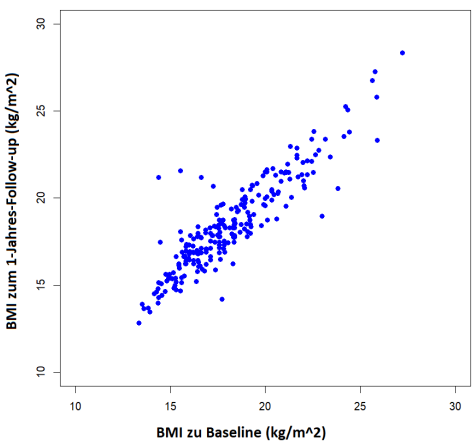
Scatterplot und Ausgleichsgerade

• **Beispiel 5.1.:** *Entwicklung des BMI bei den Teilnehmern der Schülersportstudie: Kann man einen linearen Zusammenhang zwischen den BMI-Werten $K(i)$ zu Baseline und $Z(i)$ zum 1-Jahres-Follow-up herstellen? ($1 \leqslant i \leqslant N$, $N = 232$)*

• **Visualisierung** *als Scatterplot (Streudiagramm)*

• **Ausgleichsgerade** $Z \sim c_0 + c_1 \cdot K$
als einfachstes lineares Modell

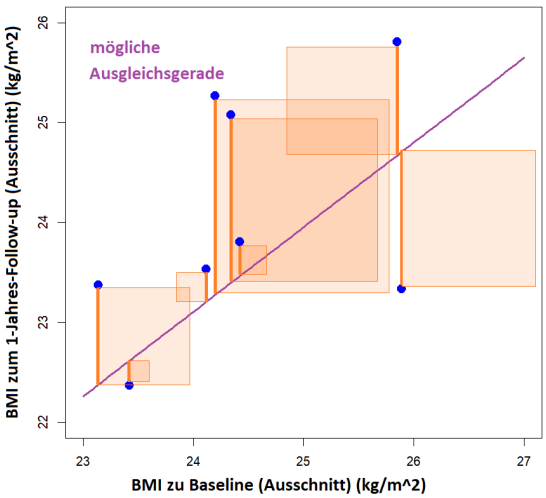
• *Die unbekannten Koeffizienten c_0 (Absolutglied) und c_1 (Anstieg) sind diejenigen Lösungen des überbestimmten Gleichungssystems*

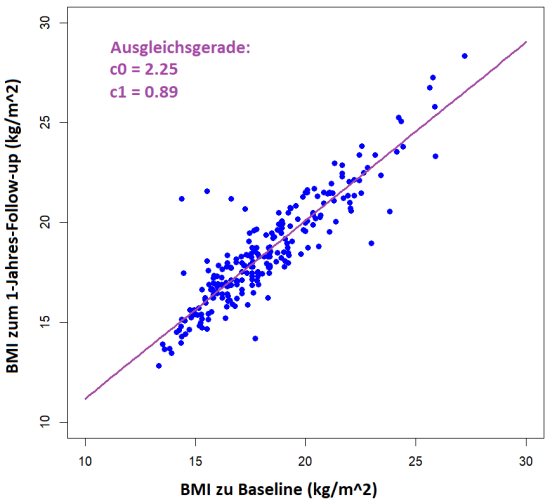


$$Z(1) = c_0 + c_1 \cdot K(1) + \varepsilon(1)\,,$$
$$Z(2) = c_0 + c_1 \cdot K(2) + \varepsilon(2)\,,$$
$$\vdots \qquad\qquad \vdots$$

welche die Bedingung $\varepsilon(1)^2 + \varepsilon(2)^2 + \ldots = \mathrm{Min}\,!$ erfüllen.

Lineare Regression: Geradengleichung





- *Koeffizienten:* $c_1 = \sigma_{ZK}/(\sigma_K)^2$, $c_0 = \mu_Z - c_1 \cdot \mu_K$

  mit $\sigma_{ZK} = \frac{1}{N} \cdot \sum_{i=1}^{N} (Z(i) - \mu_Z) \cdot (K(i) - \mu_K)$

- *Im Beispiel:* $c_0 = 2.25$, $c_1 = 0.89$; *Ausgleichsgerade:* $Z \sim 2.25 + 0.89 \cdot K$

- **Interpretation:** *BMI nach 1 Jahr* $= 2.25 + 0.89 \cdot$ *BMI zu Baseline* $+$ *Restglied*.

Lineare Regression: Bewertung des Modells

• **Wie stark ist der lineare Zusammenhang zwischen Z und K ?**

Definition 5.2.: *(Pearsonscher Korrelationskoeffizient) Der Pearsonsche Korrelationskoeffizient ist die Maßzahl* $\varrho_{ZK} = \dfrac{\sigma_{ZK}}{\sigma_Z \cdot \sigma_K} \in [-1\,,\,1]$. *Der lineare Zusammenhang ("Korrelation") zwischen Z und K ist um so stärker, je näher* $|\,\varrho_{ZK}\,|$ *an Eins liegt. Kommen die Daten aus einer Stichprobe, so sind $\widehat{\sigma}_{ZK}$, $\widehat{\sigma}_Z$ und $\widehat{\sigma}_K$ einzusetzen.*

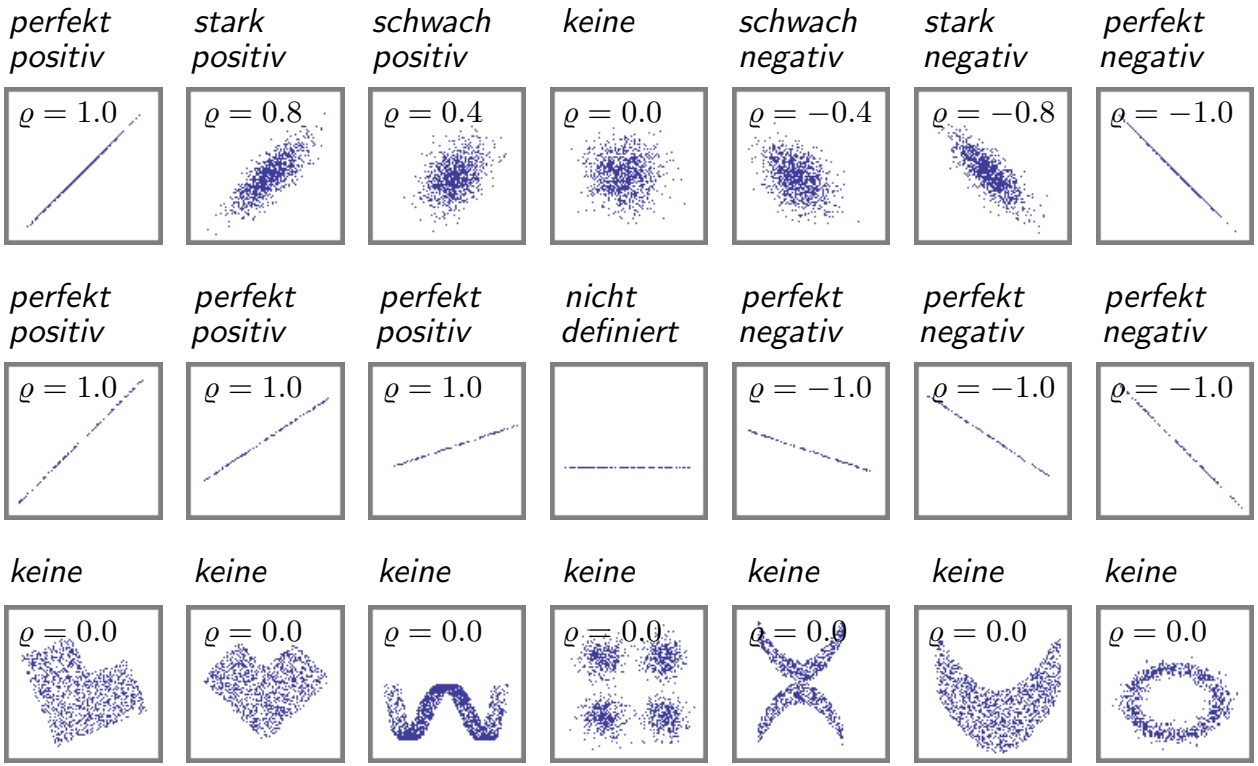
• **Güte des linearen Regressionsmodells:** *Sie wird ausgedrückt durch*

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \big(\, Z(i) - (\, c_0 + c_1 \cdot K(i)\,)\,\big)^2}{\sum_{i=1}^{N} (\, Z(i) - \mu_Z\,)^2} = 1 - \frac{\sum_{i=1}^{N} \varepsilon(i)^2}{\sum_{i=1}^{N} (\, Z(i) - \mu_Z\,)^2} = (\varrho_{ZK})^2\,.$$

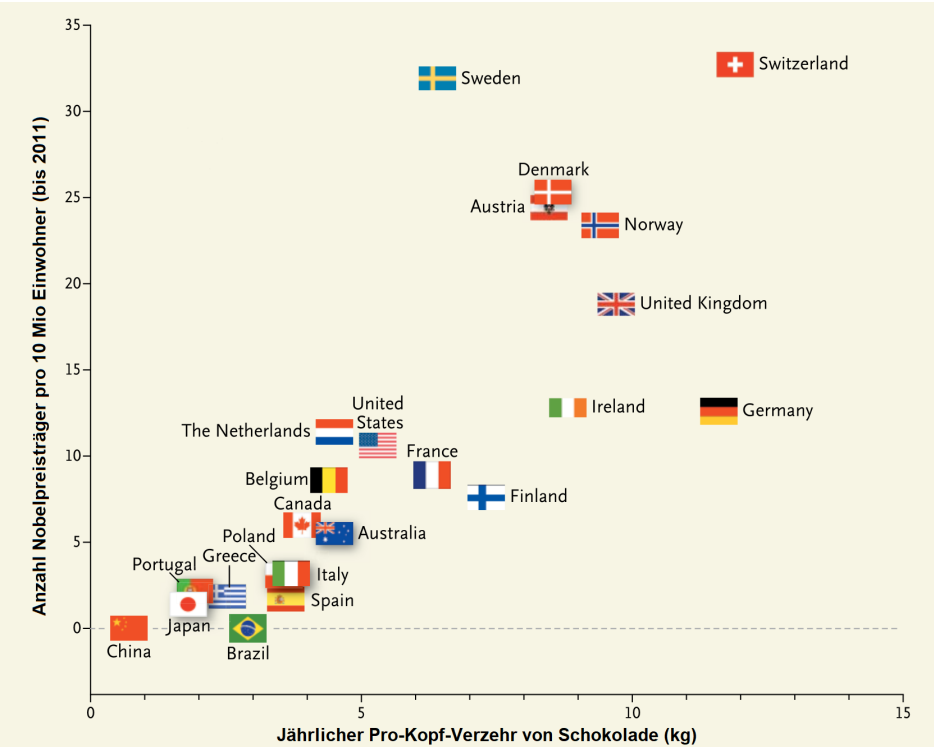
R^2 ist der Anteil der Gesamt-Varianz der Zielgröße Z, der durch das lineare Regressionsmodell "erklärt" wird. Je näher R^2 an Eins liegt, desto besser ist die Modellgüte.

• *Im Beispiel ist $\varrho_{ZK} = 0.903$ und $R^2 = 0.815$ (starke Korrelation, hohe Modellgüte).*

Schwache und starke Korrelation



Beispiel: Nobelpreis und Schokolade



- **Beispiel 5.3.**
  [Messerli 2012]

- *Pro-Kopf-Schokoladenverzehr und die Rate der Nobelpreisträger sind stark positiv korreliert. Wie kann diese Tatsache erklärt werden?*

Korrelation: was bedeutet sie (nicht)?

• **Fünf mögliche Gründe** *für eine Korrelation zwischen Merkmalen Z und K (nach* [Dubben/Beck-Bornholdt (2020)] *):*

1) K ist die Ursache von Z.

2) Z ist die Ursache von K.

3) K und Z haben eine gemeinsame Ursache.

4) Die Korrelation beruht auf einem systematischen Fehler.

5) Es handelt sich um einen Zufallsbefund. (Das ist sogar dann möglich, wenn ein Hypothesentest für die Stärke der Korrelation signifikant ausfällt!)

• *Welcher Grund zutrifft, kann im Rahmen der Statistik nicht entschieden werden!*

• **Der häufigste Fehlschluß:**

Aus dem (statistisch korrekten) Nachweis einer starken Korrelation wird auf einen Ursache-Wirkungs-Zusammenhang geschlossen, der aber in Wirklichkeit nicht besteht.

Lineare Regression für Stichproben

• **Daten aus einer Stichprobe** $\mathcal{S}$ *vom Umfang* N*: Was ändert sich?*

— *Ausgleichsgerade:* $Z \sim \widehat{c}_0 + \widehat{c}_1 \cdot K$

— *Koeffizienten:* $\widehat{c}_1 = \widehat{\sigma}_{ZK}/(\widehat{\sigma}_K)^2$, $\widehat{c}_0 = \widehat{\mu}_Z - \widehat{c}_1 \cdot \widehat{\mu}_K$

mit $\widehat{\sigma}_{ZK} = \dfrac{1}{N-1} \cdot \sum_{i=1}^{N} (Z(i) - \widehat{\mu}_Z) \cdot (K(i) - \widehat{\mu}_K)$

— *empirischer Korrelationskoeffizient:* $\widehat{\varrho}_{ZK} = \widehat{\sigma}_{ZK}/(\widehat{\sigma}_Z \cdot \widehat{\sigma}_K)$

— *Bestimmtheitsmaß:* $\widehat{R}^2 = (\widehat{\varrho}_{ZK})^2$

• **Erste Frage:** *Mit den Daten aus der Stichprobe* $\mathcal{S}$ *postulieren wir einen linearen Zusammenhang* $Z \sim \widehat{c}_0 + \widehat{c}_1 \cdot K$ *zwischen den Variablen* Z *und* K*. Kann man daraus auf das Bestehen eines linearen Zusammenhangs* $Z \sim c_0 + c_1 \cdot K$ *zwischen* Z *und* K *in der Gesamtpopulation* $\mathcal{P}$ *schließen? Antwort: globaler Hypothesentest*

• **Zweite Frage:** *Wenn ja, wie genau sind die Schätzungen* $\widehat{c}_0$ *und* $\widehat{c}_1$ *für* c_0 *und* c_1*? Antwort: Post-hoc-Tests und Konfidenzintervalle für die Koeffizienten*

Regressionsmodell: zweistufiges Verfahren

• **Situation:** *Gegeben sind zwei metrische Variablen Z, K in einer Gesamtpopulation $\mathcal{P}$ und eine Stichprobe $\mathcal{S} \subset \mathcal{P}$ vom Umfang N. Z heißt abhängige Variable oder Zielvariable, K heißt unabhängige Variable oder Kovariate. Wir behaupten, daß in $\mathcal{P}$ die lineare Gleichung $Z = c_0 + c_1 \cdot K + \varepsilon$ besteht.*

• **Schritt 1: Globaltest:** *Gibt es in $\mathcal{P}$ wirklich einen linearen Zusammenhang zwischen den Variablen Z und K?*
— gegen die Nullhypothese $c_1 = 0$
— zum Signifikanzniveau α

• **Schritt 2: Post-hoc-Tests:** *Welche Koeffizienten tragen tatsächlich zum linearen Zusammenhang bei, und welche nicht?*
— nur dann, wenn der Globaltest signifikant ausfiel,
— hier werden t-Tests gegen die Nullhypothesen $c_j = 0$ durchgeführt,
— aber mit korrigiertem Signifikanzniveau $\alpha/2$.

• **Beachte:** *Die lineare Gleichung kann selbst dann formal richtig bleiben, wenn zwischen Z und K überhaupt kein Zusammenhang besteht: nämlich für $c_1 = 0$.*

Verteilungen der geschätzten Koeffizienten

• **Lehrsatz 5.4.:** *(Verteilungen von $\widehat{c}_0$ und $\widehat{c}_1$) In $\mathcal{P}$ bestehe der lineare Zusammenhang $Z = c_0 + c_1 \cdot K + \varepsilon$ zwischen Z und K. Wir ziehen aus $\mathcal{P}$ immer wieder unabhängige Stichproben mit stets demselben Umfang $N > 2$. Dann gilt unter weiteren technischen Voraussetzungen (siehe Anmerkungen):*

1) Die geschätzten Koeffizienten $\widehat{c}_0$ und $\widehat{c}_1$ unterliegen jeweils einer Normalverteilung mit Mittelwert c_0 bzw. c_1.

2) Die globale Prüfgröße $\widehat{T} = (\widehat{c}_1)^2/(\widehat{\sigma}_1)^2$ unterliegt einer Fisher-Verteilung mit $df_1 = 1$ und $df_2 = (N-2)$ Freiheitsgraden.

3) Die Prüfgrößen $\widehat{T}_0 = \widehat{c}_0/\widehat{\sigma}_0$ und $\widehat{T}_1 = \widehat{c}_1/\widehat{\sigma}_1$ unterliegen jeweils einer t-Verteilung mit $df = (N-2)$ Freiheitsgraden.

4) Die empirischen Standardabweichungen von $\widehat{c}_0$ und $\widehat{c}_1$ berechnen sich zu

$$\widehat{\sigma}_0 = \frac{\widehat{\sigma}}{\widehat{\sigma}_K} \cdot \frac{1}{N} \cdot \sqrt{\sum_{i=1}^{N} K(i)^2}\,; \;\; \widehat{\sigma}_1 = \frac{\widehat{\sigma}}{\widehat{\sigma}_K} \cdot \frac{1}{\sqrt{N}} \quad \textit{mit} \quad \widehat{\sigma} = \sqrt{\frac{1}{N-2} \sum_{i=1}^{N} \widehat{\varepsilon}(i)^2}\,.$$

Beispiel: Globaler Hypothesentest

• **Zurück zu Beispiel 5.1.:** *jetzt als Stichprobe $\mathcal{S} \subset \mathcal{P}$ betrachtet.*

• **Welcher Test?** *Wir führen einen F-Globaltest gegen die Nullhypothese der Unabhängigkeit von Z und K in der Gesamtpopulation $\mathcal{P}$ durch (alle etwa 11-jährigen Schulkinder in Sachsen).*

• **Vorbereitungen:**

— *Nullhypothese $\mathcal{H}_0$: "Es gilt $c_1 = 0$." / Alternativhypothese $\mathcal{H}_A$: "Es gilt $c_1 \neq 0$."*

— *Freiheitsgrade der Fisher-Verteilung: (df_1, df_2) $= (1, 230)$*

— *Signifikanzniveau: $\alpha = 0.05$, dazu gehört $Q_{95}(F, 1, 230) = 3.8822$*

— *Berechnung von $\widehat{\sigma}_1$ nach Satz 5.4., 4): $\widehat{\sigma}_1 = 0.028$*

• **Prüfgröße:** $\widehat{T} = (\widehat{c}_1)^2/(\widehat{\sigma}_1)^2 = (0.89)^2/(0.028)^2 = 1010.33 \Longrightarrow p < 0.0005$

• **Testentscheidung:** *$\mathcal{H}_0$ ist zu verwerfen und $\mathcal{H}_A$ anzunehmen: In $\mathcal{P}$ ist das Bestehen einer Korrelation zwischen Z und K, also zwischen den BMI-Werten vom Folgejahr und vom laufenden Jahr, signifikant zum 5-%-Niveau ($p < 0.0005$).*

Beispiel: Koeffizententests und Konfidenzintervalle

• **Zweiseitiger t-Test für c_0 (post-hoc):** *gegen $\mathcal{H}_0$: "$c_0 = 0$" zum Signifikanzniveau $\alpha' = 0.05/2$ mit $Q_{1.25}(t, 230) = -2.26$, $Q_{98.75}(t, 230) = 2.26$*

Prüfgröße: $\widehat{T}_0 = \widehat{c}_0/\widehat{\sigma}_0 = 2.25/0.51 = 4.41$, dazu gehört $Q_{99.9999}(t, 230)$ und $p = 2 \cdot 1 \cdot 10^{-6}$, damit ist $c_0 \neq 0$ signifikant zum 5-%-Niveau ($p < 0.0005$).

• **Zweiseitiger t-Test für c_1 (post-hoc):** *gegen $\mathcal{H}_0$: "$c_1 = 0$" zum Signifikanzniveau $\alpha' = 0.05/2$ mit $Q_{1.25}(t, 230) = -2.26$ und $Q_{98.75}(t, 230) = 2.26$*

Prüfgröße: $\widehat{T}_1 = \widehat{c}_1/\widehat{\sigma}_1 = 0.89/0.028 = 31.78$, damit ist $c_1 \neq 0$ auch signifikant zum 5-%-Niveau ($p < 0.0005$).

• **95-%-Konfidenzintervall für $\widehat{c}_0 = 2.25$:**

$$\left[\, \widehat{c}_0 - \widehat{\sigma}_0 \cdot Q_{97.5}(t, 230)\,,\, \widehat{c}_0 + \widehat{\sigma}_0 \cdot Q_{97.5}(t, 230) \,\right] \;=\; [\,1.24\,,\,3.26\,]$$

• **95-%-Konfidenzintervall für $\widehat{c}_1 = 0.89$:**

$$\left[\, \widehat{c}_1 - \widehat{\sigma}_1 \cdot Q_{97.5}(t, 230)\,,\, \widehat{c}_1 + \widehat{\sigma}_1 \cdot Q_{97.5}(t, 230) \,\right] \;=\; [\,0.84\,,\,0.94\,]$$

Allgemeine lineare Modelle (GLM)

• **Verallgemeinerung des Regressionsmodells:** *Gegeben sind eine metrische Variable Z ("abhängige Variable") und zwei oder mehr Kovariaten K_j ("unabhängige Variablen"). Kann man einen linearen Zusammenhang $Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + ...$ zwischen den Variablen herstellen? Und welche Variablen K_j spielen dabei überhaupt eine Rolle?*

• **Beispiel 5.5.:** *(Rehabilitation und Gehtest) Die abhängige Variable Z ist die Wegstrecke im 6-Minuten-Gehtest beim Follow-up, die möglichen unabhängigen Variablen sind die Wegstrecke zu Baseline (metrisch), das Alter (metrisch), der Studienarm (binär) und die NYHA-Klasse zu Baseline (kategorial, dreistufig).*

• **Beispiel 5.6.:** *(Training im Diskuswerfen) In zwei Trainingszentren (Freiburg und Ulm) wird je eine Gruppe gebildet, die nach Methode A oder B trainiert. Wie hängt der Zuwachs bei der maximalen Wurfweite nach 3 Monaten Training (abhängige Variable Z) von den binären Kovariaten Zentrum und Methode ab?*

Allgemeine lineare Modelle (GLM) (II)

• **Variablen:** *metrische Zielvariable Z ("abhängige Variable"), Kovariaten K_1, K_2, K_3, ... ("unabhängige Variablen") und Residuum ε mit jeweils N Beobachtungen (zunächst vollständige Information). Wir behaupten, daß ein linearer Zusammenhang besteht:*

• **Lineare Modellgleichung:** $Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + c_3 \cdot K_3 + \ ... \ + \varepsilon$.

• **Ermittlung der Koeffizienten** c_0, c_1, c_2, c_3, ... *als diejenigen Lösungen des überbestimmten linearen Gleichungssystems*

$$\begin{aligned}
Z(1) &= c_0 + c_1 \cdot K_1(1) + c_2 \cdot K_2(1) + c_3 \cdot K_3(1) + \ ... \ + \varepsilon(1)\,, \\
Z(2) &= c_0 + c_1 \cdot K_1(2) + c_2 \cdot K_2(2) + c_3 \cdot K_3(2) + \ ... \ + \varepsilon(2)\,, \\
Z(3) &= c_0 + c_1 \cdot K_1(3) + c_2 \cdot K_2(3) + c_3 \cdot K_3(3) + \ ... \ + \varepsilon(3)\,, \\
&\ \ \vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots
\end{aligned}$$

welche die Bedingung $\varepsilon(1)^2 + \varepsilon(2)^2 + \varepsilon(3)^2 + \ ... \ = \mathrm{Min}\,!$ erfüllen (dies überlassen wir dem Computer).

• **Interpretation:** *Das GLM postuliert einen deterministischen linearen Zusammenhang $Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + c_3 \cdot K_3 + \ ...$ und trennt ihn vom zufälligen Fehler ε.*

Allgemeine lineare Modelle (GLM) (III)

• **Wichtigste Modellvoraussetzungen:**

— Die Kovariaten K_1, K_2, K_3, ... sind linear unabhängig (keine läßt sich als Linearkombination der anderen darstellen).
— Die Residuen ε sind (annähernd) normalverteilt mit $\mu = 0$ und $\sigma > 0$ und paarweise unabhängig (analog zur einfachen Regression).

• **Spezialfälle des GLM:**

— Einfache Regression: genau eine metrische Variable als Kovariate

— Multiple Regression: zwei oder mehr metrische Variablen als Kovariaten

— Einfaktorielle ANOVA: alle Kovariaten ergeben sich aus einer einzigen kategorialen Variable

— Zwei- oder mehrfaktorielle ANOVA: die Kovariaten ergeben sich aus zwei oder mehr kategorialen Variablen

• *Manche Autoren bezeichnen nur die metrischen unabhängigen Variablen als Kovariaten (alle anderen heißen dann "Faktoren" oder "Faktorstufen").*

Einbindung von Kovariaten

- **Aufbau einer Modellgleichung** *für Beispiel 5.5., drei von vier möglichen Variablen:*
  $Z \sim \ ...$ Z *— Wegstrecke im Follow-up*
- **Absolutglied** *("Intercept"): wird immer berücksichtigt.*
  $Z \sim c_0 + ...$
- **Metrische Variablen:** *unmittelbar als Kovariaten einbinden.*
  $Z \sim c_0 + c_1 \cdot K_1 + ...$ K_1 *— Wegstrecke zu Baseline*
- **Binäre Variablen:** *Einbindung einer sog. Indikatorvariablen mit Werten 0 oder 1 als Kovariate. Eine der beiden Stufen wird "Referenzlevel" und bekommt den Wert Null.*

  $Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + ...$ K_2 *— Studienarm, definiert als*

  $$K_2 = \begin{cases} 1 \mid & \text{Teilnehmer gehört zu Studienarm B.} \\ 0 \mid & \text{Teilnehmer gehört zu Studienarm A (Referenzlevel);} \end{cases}$$

- **Drei- oder mehrstufige kategoriale Variablen:** *Für eine kategoriale Variable mit M Stufen werden $(M-1)$ Indikatorvariablen ("Dummy-Variablen") als Kovariaten eingeführt. Das Referenzlevel wird dadurch kodiert, daß alle Dummies gleichzeitig Null werden.*

Einbindung von Kovariaten (II)

• **Drei- oder mehrstufige kategoriale Variablen** *beim Aufbau einer Modellgleichung für Beispiel 5.5. (Fortsetzung):*

$Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + c_3 \cdot K_3 + c_4 \cdot K_4 + \ldots$ K_3, K_4 — *definiert als*

$$K_3 = \begin{cases} 1 \mid & \textit{NYHA-Klasse II;} \\ 0 \mid & \textit{sonst.} \end{cases}$$

$$K_4 = \begin{cases} 1 \mid & \textit{NYHA-Klasse III;} \\ 0 \mid & \textit{sonst.} \end{cases}$$

Die NYHA-Klasse I ist Referenzlevel: K_3 *und* K_4 *werden beide gleichzeitig Null.*

• **Residuum:** *zufälliges Restglied, normalverteilt mit* $\mu = 0$ *und* $\sigma > 0$

$Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + c_3 \cdot K_3 + c_4 \cdot K_4 + \varepsilon$ ε — *Residuum*

• **Die Modellgleichung in Worten:**

Wegstrecke zu Follow-up $= c_0 +$ $c_1 \cdot$ *Wegstrecke zu Baseline* $+ c_2 \cdot$ *[Studienarm B ?]*
$+ c_3 \cdot$ *[NYHA-Klasse II ?]* $+ c_4 \cdot$ *[NYHA-Klasse III ?]* $+$ *Restglied.*

GLM für Stichproben

• **Daten aus einer Stichprobe** $\mathcal{S}$ *vom Umfang* N*: Was ändert sich?*

— *Modell:* $Z \sim \widehat{c}_0 + \widehat{c}_1 \cdot K_1 + \widehat{c}_2 \cdot K_2 + \ ... \ + \widehat{c}_p \cdot K_p + \varepsilon$

— *Koeffizienten: Insbesondere gilt* $\widehat{c}_0 = \widehat{\mu}_Z - \widehat{c}_1 \cdot \widehat{\mu}_{K_1} - \widehat{c}_2 \cdot \widehat{\mu}_{K_2} - \ ... \ - \widehat{c}_p \cdot \widehat{\mu}_{K_p}$.

— *Einbindung der Kovariaten: wie bisher*

— *Modellvoraussetzungen: wie bisher; zusätzlich die folgende: Alle empirischen Standardabweichungen* $\widehat{\sigma}_{K_1}$, $\widehat{\sigma}_{K_2}$, ... , $\widehat{\sigma}_{K_p}$ *sind von Null verschieden.*

• **Erste Frage:** *Mit den Daten aus der Stichprobe* $\mathcal{S}$ *postulieren wir einen linearen Zusammenhang* $Z \sim \widehat{c}_0 + \widehat{c}_1 \cdot K_1 + \widehat{c}_2 \cdot K_2 + \ ... \ + \widehat{c}_p \cdot K_p$ *zwischen den Variablen* Z *und* K_j*. Kann man daraus auf das Bestehen eines linearen Zusammenhangs* $Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + \ ... \ + c_p \cdot K_p$ *zwischen* Z *und* K_j *in der Gesamtpopulation* $\mathcal{P}$ *schließen? Antwort: globaler Hypothesentest*

• **Zweite Frage:** *Wenn ja, wie genau sind dann die Schätzungen* $\widehat{c}_0$, $\widehat{c}_1$, $\widehat{c}_2$, ... , $\widehat{c}_p$ *für* c_0, c_1, c_2, ... , c_p*? Antwort: Post-hoc-Tests und Konfidenzintervalle für die Koeffizienten*

GLM: zweistufiges Verfahren

• **Situation:** *In einer Gesamtpopulation $\mathcal{P}$ betrachten wir eine metrische Variable Z und Kovariaten K_1, ... , K_p. Es liegt eine Stichprobe $\mathcal{S}$ vom Umfang N mit Messungen aller Variablen vor. Wir behaupten, daß in $\mathcal{P}$ die lineare Gleichung $Z \sim c_0 + c_1 \cdot K_1 + ... + c_p \cdot K_p + \varepsilon$ besteht.*

• **Schritt 1: Globaltest:** *Gibt es in $\mathcal{P}$ wirklich einen linearen Zusammenhang zwischen den Variablen Z und K_1, ... , K_p?*
— gegen die Nullhypothese $c_1 = c_2 = ... = c_p = 0$ (dann bleibt die lineare Gleichung zwar formal richtig, wird aber bedeutungslos),
— zum Signifikanzniveau α

• **Schritt 2: Post-hoc-Tests:** *Welche Koeffizienten tragen tatsächlich zum linearen Zusammenhang bei, und welche nicht?*
— nur dann, wenn der Globaltest signifikant ausfiel,
— hier werden t-Tests gegen die Nullhypothesen $c_j = 0$ durchgeführt,
— aber mit korrigiertem Signifikanzniveau $\alpha/(p+1)$.

GLM: Verteilungen der geschätzten Koeffizienten

• **Lehrsatz 5.7.:** *(Verteilungen von $\widehat{c}_0$, $\widehat{c}_1$, ... , $\widehat{c}_p$) In $\mathcal{P}$ bestehe der lineare Zusammenhang $Z \sim c_0 + c_1 \cdot K_1 + ... + c_p \cdot K_p + \varepsilon$. Wir ziehen aus $\mathcal{P}$ immer wieder unabhängige Stichproben mit stets demselben Umfang $N > p + 1$. Dann gilt unter den technischen Voraussetzungen von p. 5-14 und p. 5-17:*

1) Die geschätzten Koeffizienten $\widehat{c}_0$, $\widehat{c}_1$, ... , $\widehat{c}_p$ unterliegen jeweils Normalverteilungen mit Mittelwerten c_0, c_1, ... , c_p.

2) Die globale Prüfgröße $\widehat{T} = \widehat{\sigma}^2_{\text{erklärt}} / \widehat{\sigma}^2_{\text{residual}}$ unterliegt einer Fisher-Verteilung mit $df_1 = p$ und $df_2 = (N - p - 1)$ Freiheitsgraden (Berechnung in den Anmerkungen).

3) Die Prüfgrößen $\widehat{T}_0 = \widehat{c}_0/\widehat{\sigma}_0$, $\widehat{T}_1 = \widehat{c}_1/\widehat{\sigma}_1$, ... , $\widehat{T}_p = \widehat{c}_p/\widehat{\sigma}_p$ unterliegen jeweils t-Verteilungen mit $df = (N - p - 1)$ Freiheitsgraden.

• **95-%-Konfidenzintervalle für $\widehat{c}_0$, $\widehat{c}_1$, ... , $\widehat{c}_p$:**

$$\begin{aligned}
&\left[\, \widehat{c}_0 - \widehat{\sigma}_0 \cdot Q_{97.5}(t, N-p-1)\,,\, \widehat{c}_0 + \widehat{\sigma}_0 \cdot Q_{97.5}(t, N-p-1) \,\right], \\
&\qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\
&\left[\, \widehat{c}_p - \widehat{\sigma}_p \cdot Q_{97.5}(t, N-p-1)\,,\, \widehat{c}_p + \widehat{\sigma}_p \cdot Q_{97.5}(t, N-p-1) \,\right].
\end{aligned}$$

GLM am Beispiel: Globaler Hypothesentest

• **Zurück zu Beispiel 5.5.:** *Daten für $N = 72$ Personen vollständig (mit Follow-up)*

• **Welcher Test?** *F-Test gegen die Nullhypothese der Unabhängigkeit von Z und K_1, K_2, K_3, K_4 in der Gesamtpopulation $\mathcal{P}$ aller etwa 70-jährigen Personen.*

• **Vorbereitungen:**
— *Globale Nullhypothese $\mathcal{H}_0$: "Im linearen Zusammenhang $Z \sim c_0 + \sum_{j=1}^{4} c_j \cdot K_j$ in $\mathcal{P}$ gilt $c_1 = c_2 = c_3 = c_4 = 0$, also sind Z und K_1, K_2, K_3, K_4 unabhängig."*
— *Globale Alternativhypothese $\mathcal{H}_A$: "Im linearen Zusammenhang $Z \sim c_0 + \sum_{j=1}^{4} c_j \cdot K_j$ in $\mathcal{P}$ gibt es mindestens einen Koeffizienten c_j ungleich Null."*
— *Freiheitsgrade der Fisher-Verteilung: $(df_1, df_2) = (4, 67)$*
— *Signifikanzniveau: $\alpha = 0.05$, dazu gehört $Q_{95}(F, 4, 67) = 2.51$.*

• **Prüfgröße:** $\widehat{T} = \widehat{\sigma}^2_{\text{erklärt}} \,/\, \widehat{\sigma}^2_{\text{residual}} = 40.93$, *daraus folgt $p < 0.0005$.*

• **Testentscheidung:** *$\mathcal{H}_0$ ist zu verwerfen und $\mathcal{H}_A$ anzunehmen: Das Bestehen eines linearen Zusammenhangs $Z \sim c_0 + \sum_{j=1}^{4} c_j \cdot K_j$ in $\mathcal{P}$ ist signifikant zum 5-%-Niveau ($p < 0.0005$).*

GLM am Beispiel: Bestimmung der Koeffizienten

Größe	geschätzter Koeffizient	empirische Standardabw.	Prüfgröße (F- bzw. t-Wert)	p-Wert	95-%-Konfidenzintervall
Gesamtmodell			40.93	< 0.0005	
Intercept	147.66	34.10	4.33	< 0.0005	[79.59 , 215.72]
Gehstrecke BAS	0.76	0.07	11.23	< 0.0005	[0.62 , 0.90]
Studienarm B	5.23	11.28	0.46	0.6447	[- 17.30 , 27.75]
NYHA-Klasse II	-45.16	15.61	-2.89	0.0052	[-76.33 , -13.99]
NYHA-Klasse III	-27.40	17.52	-1.56	0.1226	[- 62.37 , 7.57]

• **Zweiseitige t-Tests (post hoc):** *gegen $\mathcal{H}_0$: "$c_0 = 0$" bzw. "$c_j = 0$", $1 \leqslant j \leqslant 4$, zum Signifikanzniveau $\alpha' = 0.05/5$ mit $Q_{0.5}(t, 67) = -2.65$, $Q_{99.5}(t, 67) = 2.65$.*

• **Prüfgrößen:** $\widehat{T}_0 = \widehat{c}_0/\widehat{\sigma}_0 = 4.33$, *dazu gehört $Q_{99.9974}(t, 67)$ und $p = 2 \cdot 2.6 \cdot 10^{-5}$. Ebenso erhält man $\widehat{T}_1$, ... , $\widehat{T}_4$ und die zugehörigen p-Werte.*

• **Testentscheidungen:** *Die Nullhypothesen werden für c_0, c_1 und c_3 verworfen, dagegen für c_2 und c_4 beibehalten. Also sind $c_0 \neq 0$, $c_1 \neq 0$ und $c_3 \neq 0$ jeweils signifikant zum 5-%-Niveau von Null verschieden. Ob die Beiträge mit c_2 und c_4 über $\mathcal{S}$ hinaus auch in $\mathcal{P}$ eine Rolle spielen, bleibt unsicher.*

Was sagt das Ergebnismodell voraus?

- **Ergebnis-Modell:** $Z \sim 147.66 + 0.76 \cdot K_1 + 5.23 \cdot K_2 - 45.16 \cdot K_3 - 27.40 \cdot K_4$

Wegstrecke zu Follow-up $=$ $147.66 + 0.76 \cdot$ *Wegstrecke zu Baseline*

$+ 5.23 \cdot$ *[Studienarm B ?]* $- 45.16 \cdot$ *[NYHA-Klasse II ?]* $- 27.40 \cdot$ *[NYHA-Klasse III ?]*

- **Proband # 1:** *Arm A, NYHA-Klasse II, Strecken zu BAS/FUP 578.0 m bzw. 630.0 m*

$K_1 = 578$, $K_2 = 0$, $K_3 = 1$, $K_4 = 0$

$Z = 147.66 + 0.76 \cdot 578 + 5.23 \cdot 0 - 45.16 \cdot 1 - 27.40 \cdot 0 = 541.78$ *m*

- **Proband # 2:** *Arm B, NYHA-Klasse III, Strecken zu BAS/FUP 590.0 m bzw. 565.0 m*

$K_1 = 590$, $K_2 = 1$, $K_3 = 0$, $K_4 = 1$

$Z = 147.66 + 0.76 \cdot 590 + 5.23 \cdot 1 - 45.16 \cdot 0 - 27.40 \cdot 1 = 573.89$ *m*

- **Proband # 3:** *Arm B, NYHA-Klasse I, Strecken zu BAS/FUP 345.0 m bzw. 433.0 m*

$K_1 = 345$, $K_2 = 1$, $K_3 = 0$, $K_4 = 0$

$Z = 147.66 + 0.76 \cdot 345 + 5.23 \cdot 1 - 45.16 \cdot 0 - 27.40 \cdot 0 = 415.09$ *m*

Bewertung des Modells

- *Das Modell $Z \sim \widehat{c}_0 + \sum_{j=1}^{p} \widehat{c}_j \cdot K_j$ sei aus einer Stichprobe vom Umfang N entstanden.*
- **Definition 5.8.:** *(Bestimmtheitsmaß $\widehat{R}^2$)*

$$\widehat{R}^2 \;=\; 1 \;-\; \frac{\sum\limits_{i=1}^{N} \big(\, Z(i) - \widehat{c}_0 - \sum\limits_{j=1}^{p} \widehat{c}_j \cdot K_j(i) \, \big)^2}{\sum\limits_{i=1}^{N} (\, Z(i) - \widehat{\mu}_Z \,)^2} \;=\; \frac{\sum\limits_{i=1}^{N} (\, Z(i) - \widehat{\mu}_Z \,)^2 \;-\; \sum\limits_{i=1}^{N} \widehat{\varepsilon}(i)^2}{\sum\limits_{i=1}^{N} (\, Z(i) - \widehat{\mu}_Z \,)^2} \, .$$

$\widehat{R}^2$ ist der Anteil der Gesamt-Varianz der Zielgröße Z, der durch das GLM "erklärt" wird. Je näher $\widehat{R}^2$ an Eins liegt, desto besser ist die (formale) Modellgüte.

- **Definition 5.9.:** *(Akaike-Informationskriterium AIC)*

$$AIC = \ln \Big(\frac{\sum\limits_{i=1}^{N} \big(\, Z(i) - \widehat{c}_0 - \sum\limits_{j=1}^{p} \widehat{c}_j \cdot K_j(i) \, \big)^2}{N} \Big) + \frac{2 \cdot (p+1)}{N} = \ln \Big(\frac{\sum\limits_{i=1}^{N} \widehat{\varepsilon}(i)^2}{N} \Big) + \frac{2 \cdot (p+1)}{N} \, .$$

Das AIC ist nur für den Vergleich von zwei oder mehr Modellen mit gleicher Zielvariable Z geeignet; das Modell mit dem kleinsten AIC hat den besten Informationsgehalt.

Beispiel: Diskuswerfen

• **Beispiel 5.6.:** *Deskriptive Statistik — Verbesserung der maximalen Wurfweite*

Zentrum	Trainingsmethode	Mittelwert	Standardabweichung	Anzahl Teilnehmer
Freiburg	A	- 1.14	2.11	20
	B	+ 2.07	1.05	20
	Gesamt	+ 0.46	2.32	40
Ulm	A	+ 4,53	0.99	20
	B	+ 6.17	0.93	20
	Gesamt	+ 5.35	1.26	40
Gesamt	A	+ 1.69	3.30	40
	B	+ 4.12	2.30	40
	Gesamt	+ 2.91	3.08	80

• **Auf den ersten Blick:** *Methode A erscheint schlechter, in Freiburg sogar Verringerung der Wurfweiten; in Ulm starke Verbesserungen mit beiden Methoden*

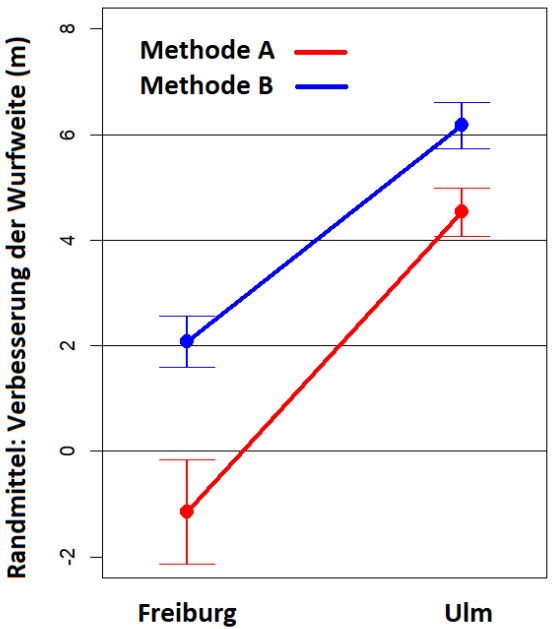
Interaktionsdiagramm bzw. Profilplot

• **Beispiel 5.6.:** *Deskriptive Statistik — Darstellung als Interaktionsdiagramm*

• **Elemente im Profilplot:**

– Skala: Wertebereich der metrischen Variable (Einheit angeben!)
— Punkte: Subgruppen-Mittelwerte
— Balken: Konfidenzintervalle (alle zum gleichen Level, hier 95 %)
– Verbindungslinien: veranschaulichen hier keine Verbesserung oder Verschlechterung, sondern ermöglichen die Beurteilung des Interaktionseffekts zwischen den Subgruppen

• **Interaktionseffekt:** *ist vorhanden, sobald die Verbindungslinien nicht parallel sind*



Bisheriges Modell ohne Interaktionsterme

- **Aufbau der Modellgleichung:** *zwei Indikatorvariablen als Kovariaten*

$$Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + \varepsilon \quad \textit{mit}$$

$$K_1 = \begin{cases} 0 \mid & \textit{Zentrum Freiburg (Referenzlevel);} \\ 1 \mid & \textit{Zentrum Ulm,} \end{cases}$$

$$K_2 = \begin{cases} 0 \mid & \textit{Methode A (Referenzlevel);} \\ 1 \mid & \textit{Methode B.} \end{cases}$$

- **Globaltest und Bestimmung der Koeffizienten:**

Größe	geschätzter Koeffizient	empirische Standardabw.	Prüfgröße (F- bzw. t-Wert)	p-Wert	95-%-Konfidenzintervall
Gesamtmodell			149.30	< 0.0005	
Intercept	-0.75	0.27	-2.74	0.0077	[-1.29 , -0.21]
Zentrum Ulm	4.88	0.32	15.47	< 0.0005	[4.24 , 5.52]
Methode B	2.43	0.32	7.70	< 0.0005	[1.79 , 3.07]

- **Modellgüte:** $\widehat{R}^2 = 0.795$ *(sehr gute Anpassung), AIC*$= 0.7295$

Kontrolle im Profilplot

- *Wie gut paßt das bisherige Modell zu den Daten?*

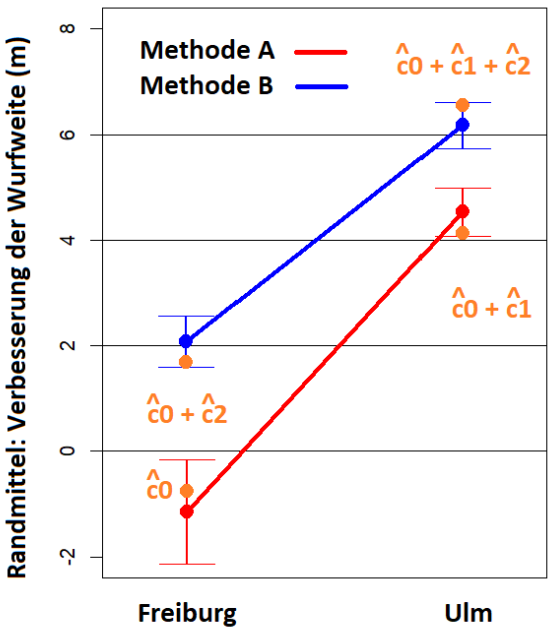
- **Modellgleichung in Worten:**

Verbesserung der Wurfweite

$$= -0.75 + 4.88 \cdot [\textit{Zentrum Ulm ?}] + 2.43 \cdot [\textit{Methode B ?}]$$

- *Trotz der hohen Anpassungsgüte von* $\widehat{R}^2 = 0.795$ *schätzen die vier Modellgrößen* $\widehat{c}_0$, $\widehat{c}_0 + \widehat{c}_1$, $\widehat{c}_0 + \widehat{c}_2$ *und* $\widehat{c}_0 + \widehat{c}_1 + \widehat{c}_2$ *die vier Randmittel nicht exakt.*

- **Idee:** *Füge eine neue Kovariate hinzu, die eine "Interaktion" zwischen Zentrum und Methode modelliert (durch Multiplikation der binären Kovariaten Zentrum und Methode)*



Interaktionsterme (I)

• **Definition 5.10.:** *(Interaktionsterme) Für ein lineares Modell mit der abhängigen Variable Z und Kovariaten K_1, ... , K_p heißen alle Zusatzterme der Gestalt $c_{j_1,j_2} \cdot K_{j_1} \cdot K_{j_2}$ mit $1 \leqslant j_1 \neq j_2 \leqslant p$ Interaktionsterme.*

• **Ist die Einführung solcher Terme im GLM zulässig?** *Ja, denn das neue Modell ist immer noch linear in den Koeffizienten, und die neue Kovariate $K_{j_1} \cdot K_{j_2}$ ist zwar ein Produkt, aber keine Linearkombination von vorhandenen Kovariaten. Achtung: der Koeffizient $c_{j_1,j_2} \neq c_{j_1} \cdot c_{j_2}$ muß neu berechnet und getestet werden!*

• **Modell mit Interaktionsterm** *für Beispiel 5.6.:*

$$Z \sim c_0 + c_1 \cdot K_1 + c_2 \cdot K_2 + c_3 \cdot (\, K_1 \cdot K_2 \,) + \varepsilon \quad \textit{mit } K_1,\ K_2 \textit{ wie bisher,}$$

$$K_3 = K_1 \cdot K_2 = \begin{cases} 0 \mid & \textit{Zentrum Freiburg \& Methode A;} \\ 0 \mid & \textit{Zentrum Freiburg \& Methode B;} \\ 0 \mid & \textit{Zentrum Ulm \& Methode A;} \\ 1 \mid & \textit{Zentrum Ulm \& Methode B.} \end{cases}$$

Interaktionsterme (II)

• **Globaltest und neue Bestimmung aller Koeffizienten** *fürs erweiterte Modell:*

Größe	geschätzter Koeffizient	empirische Standardabw.	Prüfgröße (F- bzw. t-Wert)	p-Wert	95-%-Konfidenzintervall
Gesamtmodell			109.1	< 0.0005	
Intercept	-1.14	0.30	-3.75	< 0.0005	[-1.28 , -1.00]
Zentrum Ulm	5.67	0.43	13.16	< 0.0005	[5.47 , 5.87]
Methode B	3.22	0.43	7.46	< 0.0005	[3.02 , 3.42]
Interaktionsterm Zentrum Ulm : Methode B	-1.57	0.61	-2.59	0.011	[-1.85 , -1.28]

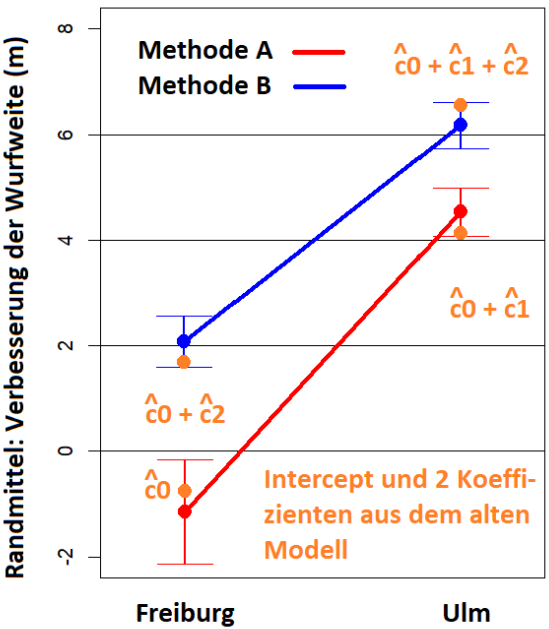
• **Modellgüte:** $\widehat{R}^2 = 0.8115$ *(Verbesserung!), AIC*$= 0.6679$ *(ebenfalls Verbesserung!)*

• **Modellgleichung in Worten:** *Verbesserung der Wurfweite*$= -1.14 + 5.67 \cdot$ *[Zentrum Ulm ?]*$+ 3.22 \cdot$ *[Methode B ?]*$- 1.57 \cdot$ *[Zentrum Ulm : Methode B ?]*

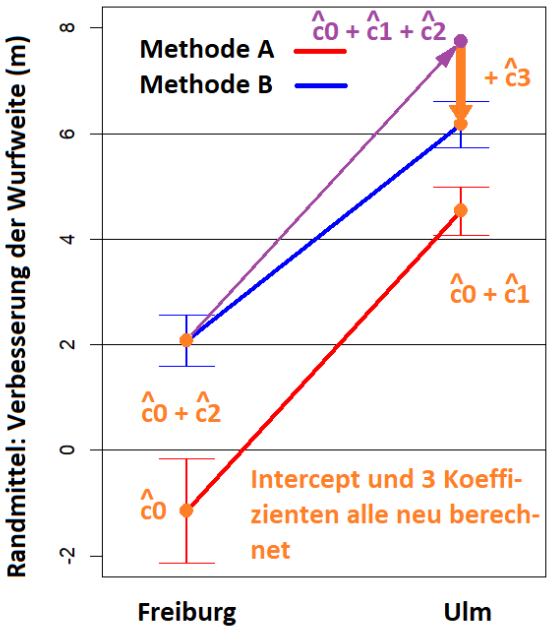
• *Praktische Bedeutung des Interaktionsterms: Der Koeffizient* c_3 *wirkt sich nur auf die Subgruppe "Training in Ulm mit Methode B" aus und korrigiert für genau diese Subgruppe die Addition der "Haupteffekte" Zentrum und Methode.*

Interaktionsterme (III): Vergleich der Profilplots

- *Bisheriges Modell (p. 5-26):*



- *Neues Modell (p. 5-28 f.):*



- *Im Modell mit Interaktionsterm werden alle vier Randmittel exakt reproduziert.*