9. Vorlesung Statistik II
Regressionsmodelle mit diskreten Prädiktoren und Interaktionen

Regressionsmodelle mit diskreten Prädiktoren

- Thema der heutigen Vorlesung: Regressionsmodelle mit diskreten Prädiktoren.
- Wir beginnen mit dem einfachsten Fall: Eine stetige AV und ein diskreter Prädiktor mit zwei (kategorialen) Ausprägungen.
- Beispielfragestellung: Welchen Einfluss hat das Land (Deutschland oder USA) auf das (stetige) monatliche Einkommen der dort arbeitenden Personen.

Dummy-Kodierung für diskrete Prädiktoren

- In Regressionsmodellen müssen alle Prädiktoren in Form von Zahlen vorliegen. Diese Schwierigkeit ergibt sich vor allem für diskrete kategoriale Prädiktoren.
- Bevor wir also Regressionsmodelle mit diskreten kategorialen Prädiktoren aufstellen können, müssen wir uns damit beschäftigen, wie wir kategoriale Variablen in sinnvoller Art und Weise in Zahlen umwandeln können. Wir müssen uns also mit der Kodierung von diskreten (kategorialen) Prädiktoren beschäftigen.
- Hier gibt es verschiedene Möglichkeiten. Wir werden die sogenannte Dummy-Kodierung betrachten.
- Bemerkung: Das Problem, wie man diskrete Prädiktoren für Regressionsmodelle am besten in Zahlen umwandelt, tritt nicht nur bei nominalskalierten Prädiktoren auf. Häufig ist auch bei ordinalskalierten (und manchmal auch bei absolutskalierten) Prädiktoren eine Dummy-Kodierung sinnvoll, selbst wenn dabei die natürliche Ordnung der Merkmalsausprägungen nicht berücksichtigt wird. Spezielle Kodierungen für ordinalskalierte Prädiktoren werden wir in dieser Vorlesung nicht behandeln.

Dummy-Kodierung für diskrete Prädiktoren mit zwei (kategorialen) Ausprägungen I

- Im Fall eines diskreten Prädiktors mit zwei (kategorialen) Ausprägungen legen wir zunächst eine der beiden Ausprägungen als Referenzkategorie fest.
- Die Dummy-Kodierung sieht dann folgendermaßen aus:
  - Die Referenzkategorie wird mit 0 kodiert.
  - Die andere Kategorie wird mit 1 kodiert.
- Wir definieren also für jede Person i in der Stichprobe eine Dummy-Variable D_i, die wie folgt definiert ist:

$$D_i = \begin{cases} 1, & \text{falls Person i nicht zur Referenzkategorie gehört} \\ 0, & \text{falls Person i zur Referenzkategorie gehört} \end{cases}$$

- Bemerkung I: Welche Ausprägung wir als Referenzkategorie wählen, ist prinzipiell egal. Es hat jedoch, wie wir sehen werden, Auswirkungen auf die Interpretation der Parameter des Regressionsmodells.
- Bemerkung II: Wir behandeln D_i meist als Zufallsvariable mit Realisation d_i.

Dummy-Kodierung für diskrete Prädiktoren mit zwei (kategorialen) Ausprägungen II

Beispiel:

- Die diskrete kategoriale Variable Land weist die Ausprägungen “Deutschland“ und „USA“ auf.
- Wir wählen „Deutschland“ als Referenzkategorie.
- In diesem Fall ist

$$D_i = \begin{cases} 1, & \text{falls Person i in den USA arbeitet} \\ 0, & \text{falls Person i in Deutschland arbeitet} \end{cases}$$

Regressionsmodell mit einem diskreten Prädiktor mit zwei (kategorialen) Ausprägungen

Mithilfe der Dummy-Variablen D_i können wir nun wie bisher die folgende allgemeine Modellgleichung aufstellen:

$$Y_i = \alpha + \beta \cdot D_i + \varepsilon_i, \qquad \text{mit } \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Die Realisation von Y_i ist der Wert der zufällig gezogenen Person i auf der AV.
- Die Realisation von D_i ist der Wert der zufällig gezogenen Person i auf der Dummy-kodierten UV.
- ε_i ist ein zufälliger Fehler.
- α, β und σ^2 sind Modellparameter.

Interpretation der Parameter α und β I

- Wie können wir die Parameter α und β in diesem Modell interpretieren?
- Hierfür betrachten wir die Modellgleichung getrennt für die beiden Ausprägungen der diskreten (kategorialen) Variable:
  - In der Referenzkategorie, also im Fall von $D_i = 0$ gilt

$$Y_i = \alpha + \beta \cdot D_i + \varepsilon_i = \alpha + \beta \cdot 0 + \varepsilon_i = \alpha + \varepsilon_i$$

    und deshalb

$$E(Y_i) = E(\alpha + \varepsilon_i) = E(\alpha) + E(\varepsilon_i) = \alpha + 0 = \alpha$$

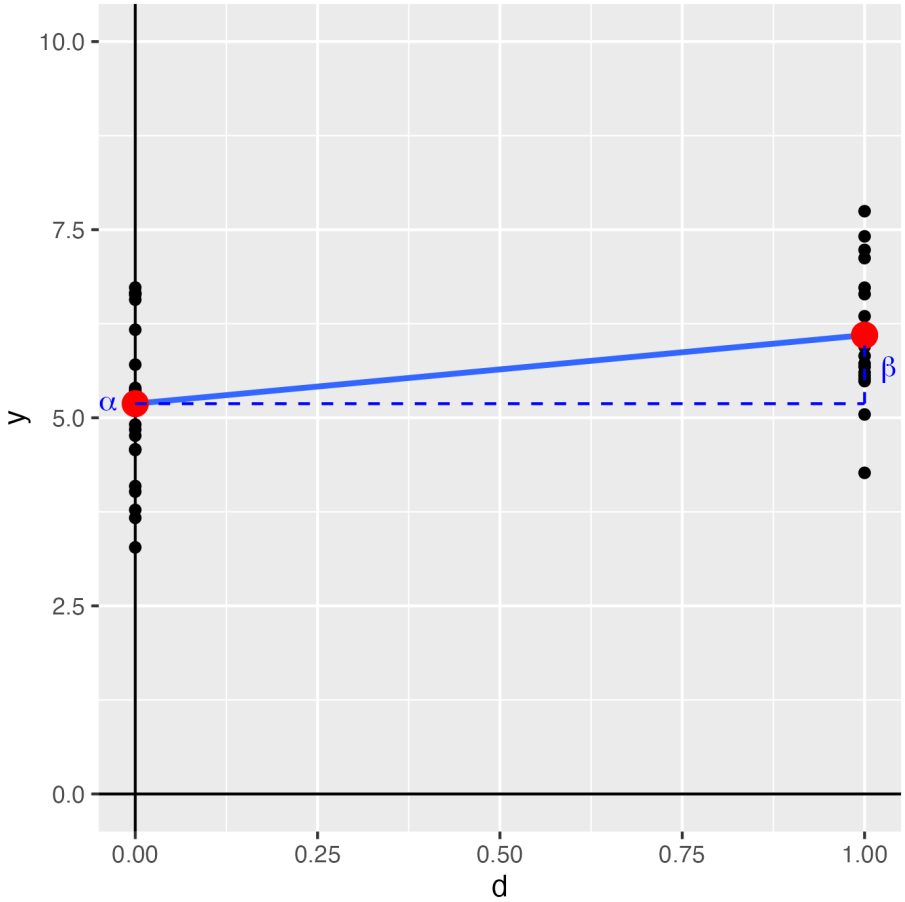
  - In der anderen Kategorie, also im Fall von $D_i = 1$ gilt

$$Y_i = \alpha + \beta \cdot D_i + \varepsilon_i = \alpha + \beta \cdot 1 + \varepsilon_i = \alpha + \beta + \varepsilon_i$$

    und deshalb

$$E(Y_i) = E(\alpha + \beta + \varepsilon_i) = E(\alpha) + E(\beta) + E(\varepsilon_i) = \alpha + \beta + 0 = \alpha + \beta$$

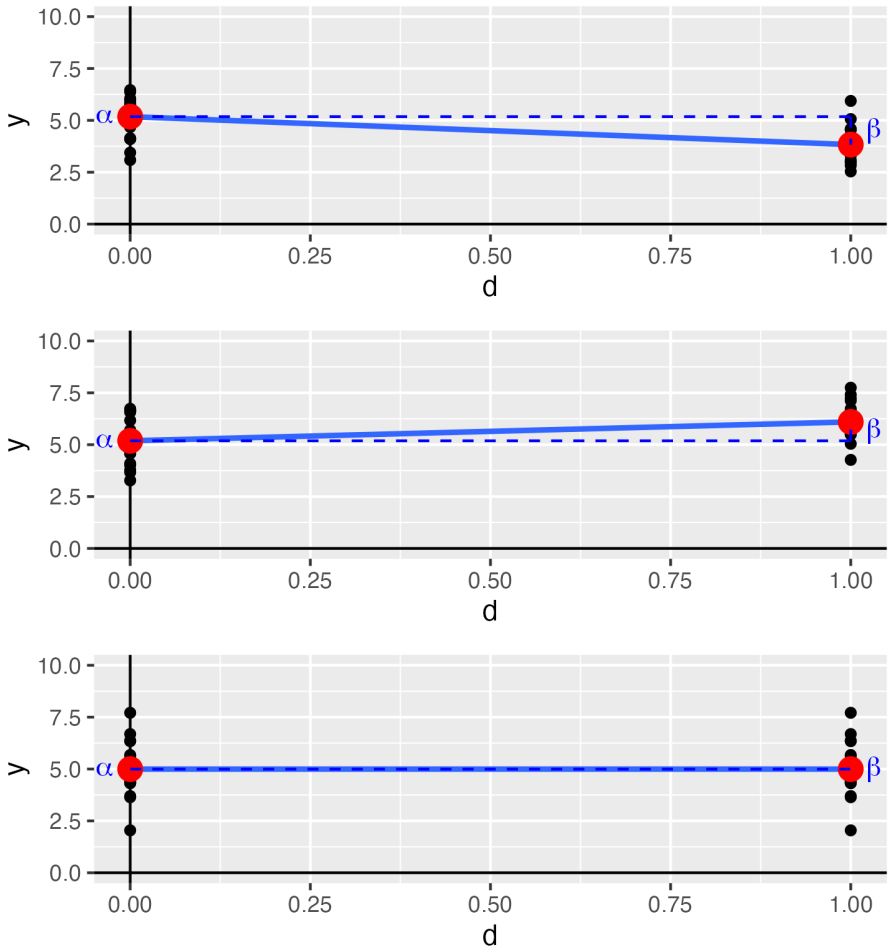
Interpretation der Parameter α und β II



Damit ergibt sich:

- α ist der Erwartungswert der AV in der Referenzkategorie.
- $\alpha + \beta$ ist der Erwartungswert der AV in der anderen Kategorie.
- β ist die Erwartungswertdifferenz der AV zwischen der anderen Kategorie und der Referenzkategorie.

Interpretation der Parameter α und β III



Interpretation des Vorzeichens von β:

- Falls der Erwartungswert der AV in der Referenzkategorie größer ist als in der anderen Kategorie, ist $\beta < 0$.
- Falls der Erwartungswert der AV in der Referenzkategorie kleiner ist als in der anderen Kategorie, ist $\beta > 0$.
- Falls der Erwartungswert der AV in der Referenzkategorie gleich groß ist, wie in der anderen Kategorie, ist $\beta = 0$.

Interpretation der Parameter α und β IV

Beispiel: AV monatliches Einkommen. UV Land, Dummy-kodiert mit Referenzkategorie „Deutschland“:

- Allgemeine Modellgleichung:

$$Y_i = \alpha + \beta \cdot D_i + \varepsilon_i$$

- Modellgleichung für Personen in Deutschland, also für $D_i = 0$:

$$Y_i = \alpha + \beta \cdot D_i + \varepsilon_i = \alpha + \beta \cdot 0 + \varepsilon_i = \alpha + \varepsilon_i$$

- Modellgleichung für Personen in den USA, also für $D_i = 1$:

$$Y_i = \alpha + \beta \cdot D_i + \varepsilon_i = \alpha + \beta \cdot 1 + \varepsilon_i = \alpha + \beta + \varepsilon_i$$

- Interpretation der Parameter:
  - α ist das erwartete monatliche Einkommen in Deutschland.
  - $\alpha + \beta$ ist das erwartete monatliche Einkommen in den USA.
  - β ist die Differenz des erwarteten monatlichen Einkommens in den USA und Deutschland.

Interpretation der Parameter α und β V

Gleiches Beispiel mit anderer Referenzkategorie: AV monatliches Einkommen. UV Land, Dummy-kodiert mit Referenzkategorie „USA“:

- Allgemeine Modellgleichung:

$$Y_i = \alpha + \beta \cdot D_i + \varepsilon_i$$

- Modellgleichung für Personen in den USA, also für $D_i = 0$:

$$Y_i = \alpha + \beta \cdot D_i + \varepsilon_i = \alpha + \beta \cdot 0 + \varepsilon_i = \alpha + \varepsilon_i$$

- Modellgleichung für Personen in Deutschland, also für $D_i = 1$:

$$Y_i = \alpha + \beta \cdot D_i + \varepsilon_i = \alpha + \beta \cdot 1 + \varepsilon_i = \alpha + \beta + \varepsilon_i$$

- Interpretation der Parameter:
  - α ist das erwartete monatliche Einkommen in den USA.
  - $\alpha + \beta$ ist das erwartete monatliche Einkommen in Deutschland.
  - β ist die Differenz des erwarteten monatlichen Einkommens in Deutschland und den USA.

Bezug zu Statistik I

- Es besteht ein direkter Zusammenhang zwischen den inferenzstatistischen Verfahren im Rahmen eines Regressionsmodells mit einer Dummy-Variable und den inferenzstatistischen Verfahren, die wir in Statistik I für Erwartungswertdifferenzen kennengelernt haben:
  - Konfidenzintervalle für β sind Konfidenzintervalle für die Erwartungswertdifferenz zwischen der Nicht-Referenzkategorie und der Referenzkategorie .
  - Hypothesentests für β sind Hypothesentests („t-Tests“) für die Erwartungswertdifferenz zwischen der Nicht-Referenzkategorie und der Referenzkategorie.

- Tatsächlich sind die Verfahren mathematisch äquivalent, da in beiden Fällen die gleichen Annahmen getroffen werden.

- Da es sich bei dem Modell mit Dummy-Variable um ein einfaches lineares Regressionsmodell handelt, können wir einfach die normalen inferenzstatistischen Verfahren aus der ELR verwenden.

- Wir werden diese nicht mehr im Detail besprechen, sondern uns lediglich die R-Outputs anschauen.

Beispiel in R I

- Beispiel von oben mit fiktiven Daten (Referenzkategorie „Deutschland“):

```
Call:
lm(formula = Einkommen ~ Land, data = daten)

Residuals:
     Min       1Q   Median       3Q      Max
-1475.41  -157.22    -0.87   167.11  1191.40

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1683.85      18.92   88.98   <2e-16 ***
LandUSA      2000.11      26.13   76.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 345.3 on 698 degrees of freedom
Multiple R-squared:  0.8935,      Adjusted R-squared:  0.8934
F-statistic:  5857 on 1 and 698 DF,  p-value: < 2.2e-16
```

Schätzwert für α, also das erwartete monatliche Einkommen in Deutschland

Schätzwert für β, also die Differenz im erwarteten monatlichen Einkommen zwischen den USA und Deutschland

p-Wert für den Test der Hypothesen

$H_0: \beta = 0$

$H_1: \beta \neq 0$

also dafür, ob es einen Unterschied zwischen den USA und Deutschland im erwarteten monatlichen Einkommen gibt.

Beispiel in R II

- Konfidenzintervall:

```
                2.5 %   97.5 %
(Intercept) 1646.696 1721.001
LandUSA     1948.799 2051.419
```

Konfidenzintervall für α, also das erwartete monatliche Einkommen in Deutschland

Konfidenzintervall für β, also die Differenz im erwarteten monatlichen Einkommen zwischen den USA und Deutschland

- Interpretation:
  - Wir gehen davon aus, dass das erwartete monatliche Einkommen in Deutschland zwischen 1646.696 und 1721.001 Euro liegt.
  - Wir gehen davon aus, dass das erwartete monatliche Einkommen in den USA 1948.799 bis 2051.419 Euro höher ist als das in Deutschland.

Beispiel in R III

`t.test` - Output zum Vergleich:

```
        Two Sample t-test

data:  Einkommen_USA and Einkommen_DE
t = 76.534, df = 698, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1948.799 2051.419
sample estimates:
mean of x mean of y
 3683.957  1683.848
```

p-Wert für den ungerichteten t-Test für unabhängige Stichproben. Entspricht dem p-Wert für den Test bzgl. β auf der vorletzten Folie. Da es sich mathematisch gesehen um den gleichen Test handelt, sind auch die Realisationen der Teststatistiken identisch.

Konfidenzintervall für die Differenz im erwarteten monatlichen Einkommen zwischen den USA und Deutschland. Entspricht dem KI für β auf der letzten Folie.

Zwischengliederung

- Bislang:
  - Regressionsmodelle mit einem diskreten (kategorialen) Prädiktor mit zwei Ausprägungen

- Jetzt:
  - Regressionsmodelle mit einem diskreten (kategorialen) Prädiktor mit mehr als zwei Ausprägungen

Dummy-Kodierung für diskrete Prädiktoren mit mehr als zwei (kategorialen) Ausprägungen I

- Auch im Fall eines diskreten Prädiktors mit $k > 2$ (kategorialen) Ausprägungen legen wir zunächst eine der Ausprägungen als Referenzkategorie fest.
- Dann wird für jede Kategorie j der übrigen $k - 1$ Kategorien eine eigene Dummy-Variable D_{ji} definiert:

$$D_{ji} = \begin{cases} 1, \text{falls Person i zur Gruppe } j \text{ gehört} \\ 0, \text{falls Person i } \textbf{nicht} \text{ zur Gruppe } j \text{ gehört} \end{cases}$$

- Für eine diskrete Variable mit k Ausprägungen gibt es also $k - 1$ Dummy-Variablen.
- Bemerkung I: Welche Ausprägung wir als Referenzkategorie wählen, wirkt sich auch hier lediglich auf die Interpretation der Parameter aus.
- Bemerkung II: Wir behandeln D_{ji} meist als Zufallsvariable mit Realisation d_{ji}.
- Bemerkung III: Aufgrund der Definition der D_{ji} gilt für Personen aus der Referenzkategorie $D_{ji} = 0$ für alle j.
- Bemerkung IV: Aufgrund der Definition der D_{ji} gilt für Personen aus der Kategorie j $D_{ji} = 1$ und $D_{j^*i} = 0$ für alle anderen Kategorien $j^* \neq j$.

Dummy-Kodierung für diskrete Prädiktoren mit mehr als zwei (kategorialen) Ausprägungen II

- Beispiel: Wir interessieren uns für den Einfluss der kategorialen Variable Haarfarbe mit den Ausprägungen “schwarz“, „braun“ und „blond“ auf die stetige Variable Depressionsschwere.
- Wir wählen „schwarz“ als Referenzkategorie.
- In diesem Fall definieren wir zwei Dummy-Variablen D_{braun_i} und D_{blond_i}:

$$D_{braun\_i} = \begin{cases} 1, & \text{falls Person i braune Haare hat} \\ 0, & \text{falls Person i keine braunen Haare hat} \end{cases}$$

$$D_{blond\_i} = \begin{cases} 1, & \text{falls Person i blonde Haare hat} \\ 0, & \text{falls Person i keine blonden Haare hat} \end{cases}$$

Haarfarbe	D_{braun_i}	D_{blond_i}
Schwarz (Referenzkategorie)	0	0
Braun	1	0
Blond	0	1

Regressionsmodell mit einem diskreten Prädiktor mit mehr als zwei (kategorialen) Ausprägungen

Mithilfe der Dummy-Variablen D_{ji} können wir nun die folgende allgemeine Modellgleichung für diskrete Prädiktoren mit k (kategorialen) Ausprägungen aufstellen:

$$Y_i = \alpha + \beta_1 \cdot D_{1i} + \cdots + \beta_{(k-1)} \cdot D_{(k-1)i} + \varepsilon_i, \qquad \text{mit } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- Die Realisation von Y_i ist der Wert der zufällig gezogenen Person i auf der AV.
- Die Realisation von D_{ji} ist der Wert der zufällig gezogenen Person i auf der j-ten Dummy-Variable.
- ε_i ist ein zufälliger Fehler.
- $\alpha, \beta_1, \ldots, \beta_{(k-1)}$ und σ^2 sind Modellparameter.

Interpretation der Parameter I

Wir betrachten die Modellgleichung wieder getrennt für die Ausprägungen der diskreten (kategorialen) Variable:

- In der **Referenzkategorie**, also im Fall von $D_{ji} = 0$ für alle j, gilt:

$$\begin{aligned} Y_i &= \alpha + \beta_1 \cdot D_{1i} + \cdots + \beta_{(k-1)} \cdot D_{(k-1)i} + \varepsilon_i \\ &= \alpha + \beta_1 \cdot 0 + \cdots + \beta_{(k-1)} \cdot 0 + \varepsilon_i \\ &= \alpha + \varepsilon_i \end{aligned}$$

  und deswegen

$$E(Y_i) = E(\alpha + \varepsilon_i) = E(\alpha) + E(\varepsilon_i) = \alpha + 0 = \boldsymbol{\alpha}$$

- In den **anderen Kategorien**, also im Fall von $D_{ji} = 1$ für das entsprechende j und $D_{j^*i} = 0$ für alle $j^* \neq j$, gilt:

$$\begin{aligned} Y_i &= \alpha + \beta_1 \cdot D_{1i} + \cdots + \beta_{(k-1)} \cdot D_{(k-1)i} + \varepsilon_i \\ &= \alpha + \beta_1 \cdot 0 + \cdots + \beta_j \cdot 1 + \cdots + \beta_{(k-1)} \cdot 0 + \varepsilon_i \\ &= \alpha + \beta_j + \varepsilon_i \end{aligned}$$

  und deswegen

$$E(Y_i) = E\left(\alpha + \beta_j + \varepsilon_i\right) = E(\alpha) + E\left(\beta_j\right) + E(\varepsilon_i) = \alpha + \beta_j + 0 = \boldsymbol{\alpha + \beta_j}$$

Interpretation der Parameter II

Damit ergibt sich:

- α ist der Erwartungswert der AV in der Referenzkategorie.
- $\alpha + \beta_j$ ist der Erwartungswert der AV in Kategorie j.
- β_j ist die Erwartungswertdifferenz der AV zwischen der Kategorie j und der Referenzkategorie.

Interpretation der Parameter III

Beispiel: AV Depression. UV Haarfarbe mit den Ausprägungen „schwarz“, „braun“ und „blond“ mit Referenzkategorie „schwarz“

Allgemeine Modellgleichung:

$$Y_i = \alpha + \beta_{braun} \cdot D_{braun\_i} + \beta_{blond} \cdot D_{blond\_i} + \varepsilon_i$$

Interpretation der Parameter IV

- Modellgleichung für Personen mit **schwarzen** Haaren,
  also für $D_{braun\_i} = 0$ und $D_{blond\_i} = 0$:

$$\begin{aligned} Y_i &= \alpha + \beta_{braun} \cdot D_{braun_i} + \beta_{blond} \cdot D_{blond_i} + \varepsilon_i \\ &= \alpha + \beta_{braun} \cdot 0 + \beta_{blond} \cdot 0 + \varepsilon_i \\ &= \boldsymbol{\alpha + \varepsilon_i} \end{aligned}$$

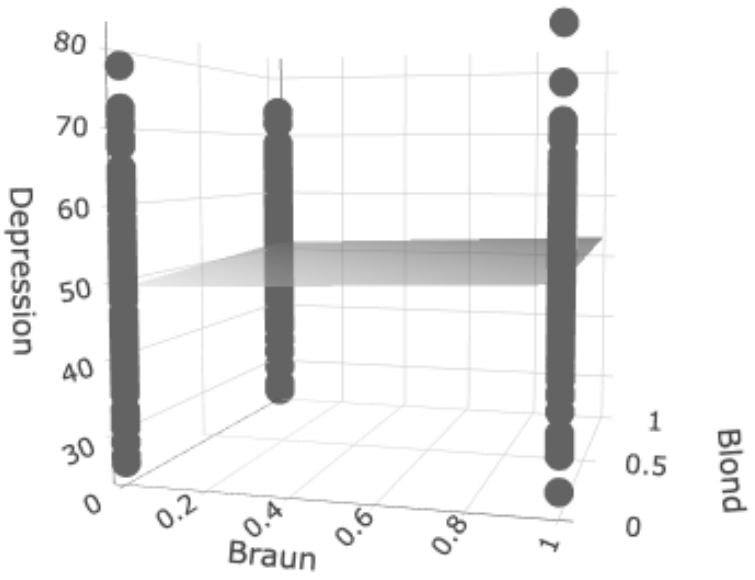
- Modellgleichung für Personen mit **braunen** Haaren,
  also für $D_{braun\_i} = 1$ und $D_{blond\_i} = 0$:

$$\begin{aligned} Y_i &= \alpha + \beta_{braun} \cdot D_{braun_i} + \beta_{blond} \cdot D_{blond_i} + \varepsilon_i \\ &= \alpha + \beta_{braun} \cdot 1 + \beta_{blond} \cdot 0 + \varepsilon_i \\ &= \boldsymbol{\alpha + \beta_{braun} + \varepsilon_i} \end{aligned}$$

- Modellgleichung für Personen mit **blonden** Haaren,
  also für $D_{braun\_i} = 0$ und $D_{blond\_i} = 1$:

$$\begin{aligned} Y_i &= \alpha + \beta_{braun} \cdot D_{braun_i} + \beta_{blond} \cdot D_{blond_i} + \varepsilon_i \\ &= \alpha + \beta_{braun} \cdot 0 + \beta_{blond} \cdot 1 + \varepsilon_i \\ &= \boldsymbol{\alpha + \beta_{blond} + \varepsilon_i} \end{aligned}$$

Interpretation der Parameter V



Interpretation der Parameter:

- α ist die erwartete Depressionsschwere der Personen mit schwarzen Haaren.
- $\alpha + \beta_{braun}$ ist die erwartete Depressionsschwere der Personen mit braunen Haaren.
- $\alpha + \beta_{blond}$ ist die erwartete Depressionsschwere der Personen mit blonden Haaren.
- β_{braun} ist die Differenz der erwarteten Depressionsschwere von Personen mit braunen Haaren und Personen mit schwarzen Haaren.
- β_{blond} ist die Differenz der erwarteten Depressionsschwere von Personen mit blonden Haaren und Personen mit schwarzen Haaren.

Bezug zum einfaktoriellen varianzanalytischen Modell

- Es besteht ein direkter Zusammenhang zwischen den inferenzstatistischen Verfahren im Rahmen eines Regressionsmodells einer diskreten Variable in Dummy-Kodierung und den inferenzstatistischen Verfahren, die wir im Rahmen des einfaktoriellen varianzanalytischen Modells kennengelernt haben.
- Zum Beispiel:
  - Der Omnibus-Test mit den Hypothesen
    - H_0: $\beta_j = 0$ für alle j
    - H_1: $\beta_j \neq 0$ für mindestens ein j

    ist äquivalent zum Omnibustest des einfaktoriellen varianzanalytischen Modells.
  - Konfidenzintervalle für β_j im Regressionsmodell sind äquivalent zu Konfidenzintervallen für $\mu_{Kategorie\_j} - \mu_{Ref}$ im varianzanalytischen Modell.
- Tatsächlich sind die beiden Modelle mathematisch äquivalent und es gibt für jedes inferenzstatistische Verfahren in einem Modell ein entsprechendes äquivalentes Verfahren in dem anderen Modell.

- Da es sich bei dem Modell mit mehreren Dummy-Variablen um ein multiples lineares Regressionsmodell handelt, können wir einfach die normalen inferenzstatistischen Verfahren für die MLR verwenden.
- Wir werden diese nicht im Detail besprechen, sondern uns lediglich die R-Outputs anschauen.

Beispiel in R I

Bsp. mit fiktiven Daten (Referenzkategorie „schwarze Haare“):

```
Call:
lm(formula = Depression ~ Braun + Blond, data = daten2)

Residuals:
     Min       1Q   Median       3Q      Max
-26.1260  -6.4446  -0.1669   6.2391  30.2899

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.447      1.022  48.386   <2e-16 ***
Braun          1.942      1.445   1.344    0.180
Blond          1.844      1.445   1.276    0.203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.22 on 297 degrees of freedom
Multiple R-squared:  0.00766,     Adjusted R-squared:  0.0009772
F-statistic: 1.146 on 2 and 297 DF,  p-value: 0.3192
```

Schätzwert für α, also die erwartete Depressionsschwere der Personen mit schwarzen Haaren

Schätzwert für β_{braun}, also die Differenz der erwarteten Depressionsschwere von Personen mit braunen Haaren und Personen mit schwarzen Haaren

Schätzwert für β_{blond}, also die Differenz der erwarteten Depressionsschwere von Personen mit blonden Haaren und Personen mit schwarzen Haaren

p-Werte für
$H_0: \alpha = 0$
$H_1: \alpha \neq 0$

bzw.
$H_0: \beta_j = 0$
$H_1: \beta_j \neq 0$

p-Wert für den Omnibustest:
$H_0: \beta_j = 0$ für alle j
$H_1: \beta_j \neq 0$ für mindestens ein j

Beispiel in R II

Konfidenzintervalle:

```
                2.5 %    97.5 %
(Intercept) 47.4362983 51.458625
Braun       -0.9022204  4.786208
Blond       -0.9998519  4.688576
```

Konfidenzintervall für α, also die erwartete Depressionsschwere der Personen mit schwarzen Haaren

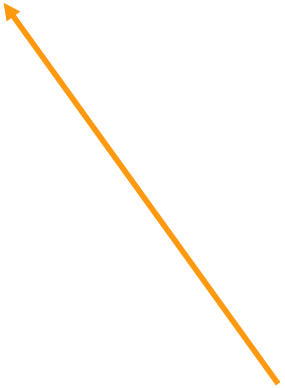
Konfidenzintervall für β_{braun}, also die Differenz der erwarteten Depressionsschwere von Personen mit braunen Haaren und Personen mit schwarzen Haaren

Konfidenzintervall für β_{blond}, also die Differenz der erwarteten Depressionsschwere von Personen mit blonden Haaren und Personen mit schwarzen Haaren

Beispiel in R III

aov -Output mit Omnibustest zum Vergleich:

```
             Df Sum Sq Mean Sq F value Pr(>F)
Haarfarbe     2    239   119.7   1.146  0.319
Residuals   297  31018   104.4
```



gleicher p-Wert und gleiche Realisation der Teststatistik wie beim Omnibustest im Rahmen des Regressionsmodells auf der vorletzten Folie.

Zwischengliederung

- Bislang:
  - Regressionsmodelle mit einem diskreten Prädiktor mit zwei (kategorialen) Ausprägungen
  - Regressionsmodelle mit einem diskreten Prädiktor mit mehr als zwei (kategorialen) Ausprägungen

- Jetzt:
  - Regressionsmodelle mit mehreren diskreten (kategorialen) Prädiktoren

Regressionsmodelle mit mehreren diskreten Prädiktoren

- Prinzipiell lassen sich Regressionsmodelle mit beliebig vielen jeweils Dummy-kodierten diskreten Prädiktoren aufstellen.
- Beispiel-Fragestellung: Welchen Einfluss haben Land und Haarfarbe auf die (stetige) Depressionsschwere.
- Oft sind die Parameter dieser Regressionsmodelle jedoch schwierig zu interpretieren (vor allem, wenn wir Interaktionen mit in das Modell aufnehmen wollen).
- Da es für jeden dieser Fälle ein äquivalentes mehrfaktorielles varianzanalytisches Modell gibt, lassen sich Fragestellungen bzgl. des Einflusses mehrerer diskreter (kategorialer) Variablen daher häufig einfacher im varianzanalytischen Kontext untersuchen.
- Aber: Falls wir lediglich AV-Werte vorhersagen wollen und uns nicht für die Interpretation der Parameter des Modells interessieren, ist das Regressionsmodell einfacher zu handhaben.

Zwischengliederung

- Bislang:
  - Regressionsmodelle mit einem diskreten Prädiktor mit zwei (kategorialen) Ausprägungen
  - Regressionsmodelle mit einem diskreten Prädiktor mit mehr als zwei (kategorialen) Ausprägungen
  - Regressionsmodelle mit mehreren diskreten (kategorialen) Prädiktoren

- Jetzt:
  - Regressionsmodelle mit stetigen und diskreten (kategorialen) Prädiktoren

Regressionsmodelle mit stetigen und diskreten Prädiktoren

- Wir betrachten nun Regressionsmodelle, die sowohl stetige, also auch diskrete Prädiktoren enthalten.
- Wir beschränken uns zunächst auf den einfachsten Fall: Eine stetige AV mit einem stetigen Prädiktor und einem diskreten Prädiktor mit zwei (kategorialen) Ausprägungen.
- Beispielfragestellung: Welchen Einfluss haben die (stetige) Bildung und das Land (Deutschland oder USA) auf das (stetige) monatliche Einkommen.

Dummy-Kodierung des diskreten Prädiktors

- Für die diskrete (kategoriale) Variable wählen wir wieder eine Referenzkategorie und definieren die folgende Dummy-Variable:

$$D_i = \begin{cases} 1, & \text{falls Person i nicht zur Referenzkategorie gehört} \\ 0, & \text{falls Person i zur Referenzkategorie gehört} \end{cases}$$

- Beispiel Land mit Referenzkategorie „Deutschland“:

$$D_i = \begin{cases} 1, & \text{falls Person i in den USA arbeitet} \\ 0, & \text{falls Person i in Deutschland arbeitet} \end{cases}$$

Regressionsmodell mit einem stetigen Prädiktor und einem diskreten Prädiktor mit zwei Ausprägungen: Modell 1

Erstes Modell: Aufnehmen der Dummy-Variablen D_i und des stetigen Prädiktors X_i in die folgende allgemeine Modellgleichung:

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \varepsilon_i, \qquad \text{mit } \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Die Realisation von Y_i ist der Wert der zufällig gezogenen Person i auf der AV.
- Die Realisation von X_i ist der Wert der zufällig gezogenen Person i auf der stetigen UV.
- Die Realisation von D_i ist der Wert der zufällig gezogenen Person i auf der Dummy-kodierten UV.
- ε_i ist ein zufälliger Fehler.
- α, β_1, β_2 und σ^2 sind Modellparameter.

Interpretation der Parameter I

- Damit ergibt sich für jede Ausprägung auf der diskreten Variable ein ELR-Modell:
  - Für Personen aus der Referenzkategorie, also für $D_i = 0$:

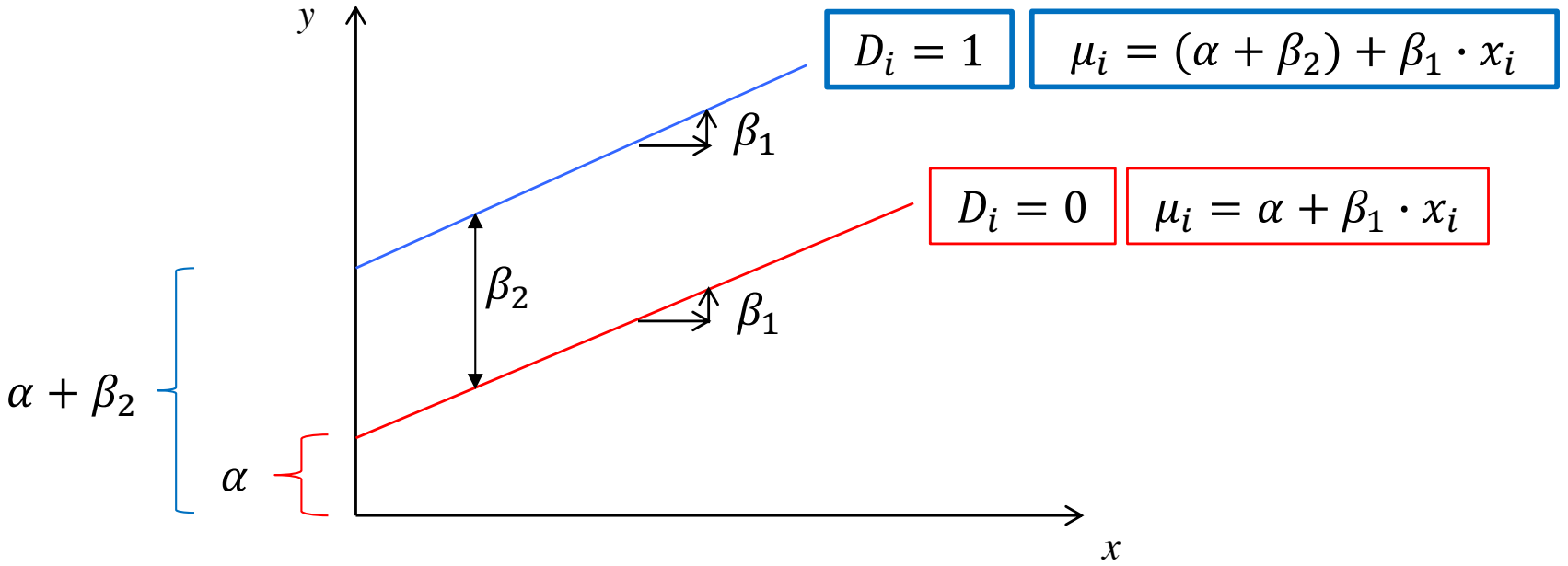
    $$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \varepsilon_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot 0 + \varepsilon_i = \alpha + \beta_1 \cdot X_i + \varepsilon_i$$

  - Für Personen aus der anderen Kategorie, also für $D_i = 1$:

    $$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \varepsilon_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot 1 + \varepsilon_i = (\alpha + \beta_2) + \beta_1 \cdot X_i + \varepsilon_i$$

- Damit ergibt sich die folgende Interpretation der Parameter:
  - α ist der Intercept in der Referenzkategorie.
  - $\alpha + \beta_2$ ist der Intercept in der anderen Kategorie.
  - β_1 ist der Steigungsparameter in beiden Kategorien. Die Regressionsgeraden in den beiden Kategorien sind also parallel.
  - β_2 ist zunächst nur die Differenz der Intercepts der anderen Kategorie und der Referenzkategorie. Aufgrund der Parallelität der Regressionsgeraden ist β_2 aber auch der erwartete Unterschied in der AV zweier Personen aus unterschiedlichen Kategorien, aber mit gleicher Ausprägung auf der stetigen UV.

Interpretation der Parameter II



Interpretation der Parameter III

Beispiel: AV monatliches Einkommen, kategoriale UV Land mit Referenzkategorie „Deutschland“, stetige UV Bildung:

- Allgemeine Modellgleichung:

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \varepsilon_i$$

- Modellgleichung für Personen in Deutschland:

$$Y_i = \alpha + \beta_1 \cdot X_i + \varepsilon_i$$

- Modellgleichung für Personen in den USA:

$$Y_i = (\alpha + \beta_2) + \beta_1 \cdot X_i + \varepsilon_i$$

Interpretation der Parameter IV

Interpretation der Parameter:

- Interpretation α: Erwartetes monatliches Einkommen einer Person in Deutschland mit einer Bildung von 0 (keine inhaltlich sinnvolle Interpretation).
- Interpretation β_2: Differenz des erwarteten monatlichen Einkommens einer Person in den USA und einer Person in Deutschland mit gleicher Bildung.
- Interpretation β_1: Falls sich die Bildung um eine Einheit erhöht, erhöht sich das erwartete monatliche Einkommen bei Personen in den USA und Personen in Deutschland um β_1 Einheiten.

Problem

Problem: Der Steigungsparameter β_1 ist in beiden Kategorien gleich.

- Das heißt: In dem Modell wird angenommen, dass der Zusammenhang zwischen der stetigen UV und der AV in beiden Kategorien gleich groß ist.
- Diese Annahme kann natürlich verletzt sein. Zum Beispiel könnte es sein, dass die Art des Zusammenhangs zwischen den stetigen Variablen von der diskreten Variable abhängt. In diesem Fall läge eine Interaktion zwischen der diskreten (kategorialen) UV und der stetigen UV vor.
- In vielen Fällen interessieren wir uns gerade für diese Interaktion!
- Also: Erweiterung des Modells um einen Interaktionsterm.

Regressionsmodell mit einem stetigen Prädiktor und einem diskreten Prädiktor mit zwei Ausprägungen: Modell 2

- Zweites Modell: Aufnehmen der Dummy-Variablen D_i und des stetigen Prädiktors X_i und des Produkts aus D_i und X_i in die folgende allgemeine Modellgleichung:

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \beta_3(X_i \cdot D_i) + \varepsilon_i, \qquad \text{mit } \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Die Realisation von Y_i ist der Wert der zufällig gezogenen Person i auf der AV.
- Die Realisation von X_i ist der Wert der zufällig gezogenen Person i auf der stetigen UV.
- Die Realisation von D_i ist der Wert der zufällig gezogenen Person i auf der Dummy-kodierten UV.
- ε_i ist ein zufälliger Fehler.
- α, β_1, β_2, β_3 und σ^2sind Modellparameter.

Interpretation der Parameter I

- Wieder ergibt sich jede Ausprägung auf der diskreten Variable ein ELR-Modell:
- Modell für die Referenzkategorie, also für $D_i = 0$:

$$\begin{aligned} Y_i &= \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \beta_3 (X_i \cdot D_i) + \varepsilon_i \\ &= \alpha + \beta_1 \cdot X_i + 0 + 0 + \varepsilon_i \\ &= \alpha + \beta_1 \cdot X_i + \varepsilon_i \end{aligned}$$

- Modell für die andere Kategorie, also für $D_i = 1$:

$$\begin{aligned} Y_i &= \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \beta_3 (X_i \cdot D_i) + \varepsilon_i \\ &= \alpha + \beta_1 \cdot X_i + \beta_2 \cdot 1 + \beta_3 \cdot (X_i \cdot 1) + \varepsilon_i \\ &= (\alpha + \beta_2) + (\beta_1 + \beta_3) \cdot X_i + \varepsilon_i \end{aligned}$$

Interpretation der Parameter II

Damit ergibt sich die folgende Interpretation der Parameter:

- α ist der Intercept in der Referenzkategorie.
- β_1 ist der Steigungsparameter in der Referenzkategorie.
- $\alpha + \beta_2$ ist der Intercept in der anderen Kategorie.
- $\beta_1 + \beta_3$ ist der Steigungsparameter in der anderen Kategorie.
- β_2 ist die Differenz der Intercepts der anderen Kategorie und der Referenzkategorie.
- β_3 ist die Differenz der Steigungsparameter der anderen Kategorie und der Referenzkategorie.

Interpretation der Parameter III

Beispiel: AV monatliches Einkommen, kategoriale UV Land mit Referenzkategorie „Deutschland“, stetige UV Bildung:

- Allgemeine Modellgleichung:

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_i + \beta_3 (X_i \cdot D_i) + \varepsilon_i$$

- Modellgleichung für Personen in Deutschland:

$$Y_i = \alpha + \beta_1 \cdot X_i + \varepsilon_i$$

- Modellgleichung für Personen in den USA:

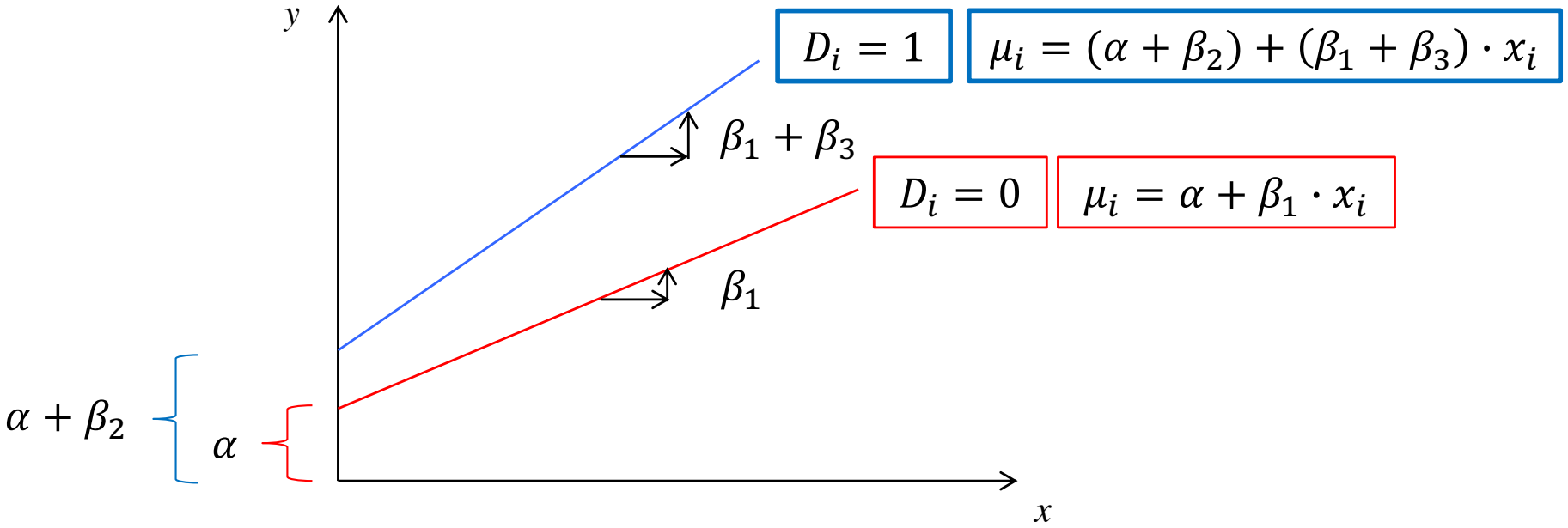
$$Y_i = (\alpha + \beta_2) + (\beta_1 + \beta_3) \cdot X_i + \varepsilon_i$$

Interpretation der Parameter IV

Interpretation der Parameter:

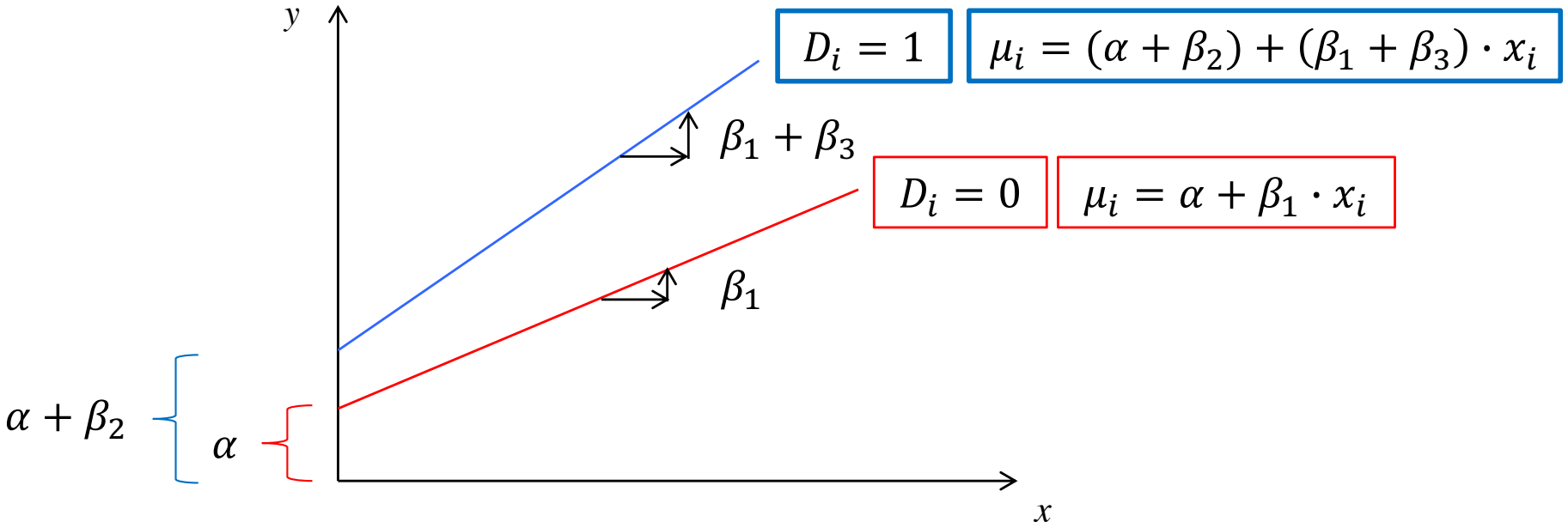
- Interpretation α: Erwartetes monatliches Einkommen einer Person in Deutschland mit einer Bildung von 0 (keine inhaltlich sinnvolle Interpretation).
- Interpretation β_2: Differenz des erwarteten monatlichen Einkommens einer Person in den USA mit einer Bildung von 0 und einer Person in Deutschland mit einer Bildung von 0 (keine inhaltlich sinnvolle Interpretation).
- Interpretation β_1: Falls sich die Bildung um eine Einheit erhöht, erhöht sich das erwartete monatliche Einkommen bei Personen in Deutschland um β_1 Einheiten.
- Interpretation β_3: Falls sich die Bildung um eine Einheit erhöht, erhöht sich das erwartete monatliche Einkommen bei Personen in den USA um β_3 mehr Einheiten als bei Personen in Deutschland.

Interpretation der Parameter V



- Der Modellparameter β_3 quantifiziert den Unterschied in der Steigung zwischen den USA und Deutschland.
- Wenn $\beta_1 + \beta_3 = \beta_1$, also $\beta_3 = 0$, liegt keine Wechselwirkung zwischen Land und Bildung vor: Der lineare Zusammenhang zwischen Einkommen und Bildung ist für Deutschland und die USA gleich groß. Somit kann Modell 1 als Spezialfall von Modell 2 mit $\beta_3 = 0$ aufgefasst werden.

Interpretation der Parameter VI



- Der Modellparameter β_2 ist das erwartete höhere Einkommen bei Personen in den USA im Vergleich zu Personen in Deutschland bei einer Bildung von 0.
- Falls $\beta_2 = 0$, wird für Personen in den USA und in Deutschland mit Bildung 0 das gleiche Einkommen erwartet.
- Falls $\beta_3 = 0$ ist, quantifiziert β_2 (genau wie in Modell 1) Unterschied im erwarteten monatlichen Einkommen zwischen den Ländern bei gleichem Bildungsgrad, da die Geraden in diesem Fall parallel sind.

- Da es sich bei dem Modell mit Dummy-Variable, stetiger UV und Interaktionsterm um ein multiples Regressionsmodell handelt, können wir einfach die normalen inferenzstatistischen Verfahren für die MLR verwenden.
- Wir werden diese nicht im Detail besprechen, sondern uns lediglich die R-Outputs anschauen.

Parameterschätzung in R I

Beispiel von oben mit fiktiven Daten:

```
Call:
lm(formula = Einkommen ~ Bildung * Land, data = daten)

Residuals:
    Min      1Q  Median      3Q     Max
-93.424 -19.115   0.548  20.203  88.021

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
α (Intercept)    702.9113    10.6839  65.791   <2e-16 ***
β1 Bildung         9.9682     0.1073  92.866   <2e-16 ***
β2 LandUSA       -11.5757    14.7772  -0.783    0.434
β3 Bildung:LandUSA  20.0183     0.1475 135.732   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.26 on 696 degrees of freedom
Multiple R-squared:  0.9992,  Adjusted R-squared:  0.9992
F-statistic: 3.041e+05 on 3 and 696 DF,  p-value: < 2.2e-16
```

Schätzwert für α, also das erwartete monatliche Einkommen einer Person in Deutschland mit Bildung 0

Schätzwert für $\boldsymbol{\beta_1}$, also den Steigungsparameter bei Personen in Deutschland

Schätzwert für $\boldsymbol{\beta_2}$, also den Unterschied im erwarteten Monatseinkommen zwischen Personen in den USA und Deutschland mit Bildung 0

Schätzwert für β_3, also die Differenz der Steigungsparameter zwischen den USA und Deutschland

Parameterschätzung in R I

Beispiel von oben mit fiktiven Daten:

```
Call:
lm(formula = Einkommen ~ Land * Bildung, data = daten)

Residuals:
    Min      1Q  Median      3Q     Max
-93.424 -19.115   0.548  20.203  88.021

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
α (Intercept)    702.9113    10.6839  65.791   <2e-16 ***
β1 LandUSA       -11.5757    14.7772  -0.783    0.434
β2 Bildung         9.9682     0.1073  92.866   <2e-16 ***
β3 Bildung:LandUSA 20.0183     0.1475 135.732   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.26 on 696 degrees of freedom
Multiple R-squared:  0.9992,  Adjusted R-squared:  0.9992
F-statistic: 3.041e+05 on 3 and 696 DF,  p-value: < 2.2e-16
```

Schätzwert für α, also das erwartete monatliche Einkommen einer Person in Deutschland mit Bildung 0

Schätzwert für $\boldsymbol{\beta_2}$, also den Steigungsparameter bei Personen in Deutschland

Schätzwert für $\boldsymbol{\beta_1}$, also den Unterschied im erwarteten Monatseinkommen zwischen Personen in den USA und Deutschland mit Bildung 0

Schätzwert für β_3, also die Differenz der Steigungsparameter zwischen den USA und Deutschland

Parameterschätzung in R II

Konfidenzintervalle:

```
                 2.5 %    97.5 %
(Intercept)     681.934718 723.88793
Bildung           9.757479  10.17898
LandUSA         -40.588969  17.43761
Bildung:LandUSA  19.728781  20.30792
```

Konfidenzintervall für α, also das erwartete monatliche Einkommen einer Person in Deutschland mit Bildung 0

Konfidenzintervall für β_1, also den Steigungsparameter bei Personen in Deutschland

Konfidenzintervall für β_2, also den Unterschied im erwarteten Monatseinkommen zwischen Personen in den USA und Personen in Deutschland mit Bildung 0

Konfidenzintervall für β_3, also die Differenz der Steigungsparameter zwischen den USA und Deutschland

Parameterschätzung in R III

Interpretation der KIs für die inhaltlich sinnvoll interpretierbaren Parameter β_1 und β_3:

- Wir gehen davon aus, dass bei Personen in Deutschland das erwartete monatliche Einkommen um 9.76 bis 10.18 Euro steigt, falls die Bildung um einen Punkt steigt.
- Wir gehen davon aus, dass bei Personen in den USA das erwartete monatliche Einkommen pro Punkt Bildung um 19.73 bis 20.31 Euro **mehr** steigt **als** bei den Personen in Deutschland.
- Sowohl bei Personen in Deutschland also auch bei Personen in den USA gehen wir also von einem positiven Zusammenhang zwischen Bildung und Einkommen aus. In den USA scheint dieser jedoch stärker zu sein.

Hypothesentests in R I

Wir betrachten für unser Beispiel die Hypothesentests für die folgenden Hypothesen:

1. Gibt es einen Zusammenhang zwischen Einkommen und Bildung in Deutschland?

$$H_0\colon \beta_1 = 0$$
$$H_1\colon \beta_1 \neq 0$$

2. Ist der Zusammenhang zwischen Einkommen und Bildung in Deutschland und in den USA unterschiedlich groß?

$$H_0\colon \beta_3 = 0$$
$$H_1\colon \beta_3 \neq 0$$

Hypothesentests in R II

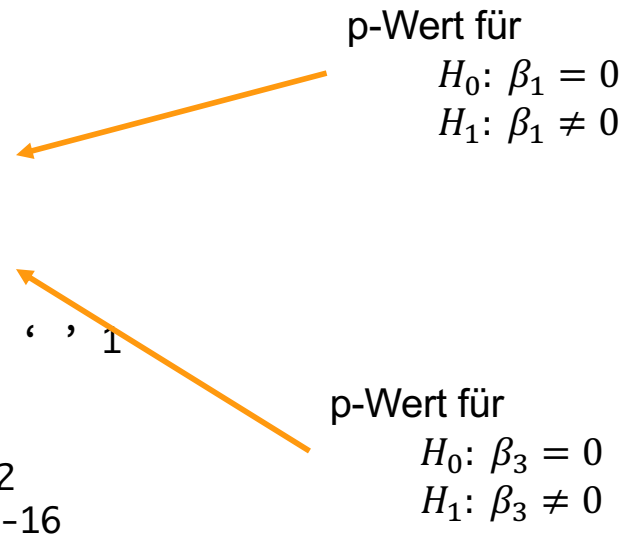
Beispiel von oben mit fiktiven Daten:

```
Call:
lm(formula = Einkommen ~ Bildung * Land, data = daten)

Residuals:
    Min      1Q  Median      3Q     Max
-93.424 -19.115   0.548  20.203  88.021

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     702.9113    10.6839  65.791   <2e-16 ***
Bildung           9.9682     0.1073  92.866   <2e-16 ***
LandUSA         -11.5757    14.7772  -0.783    0.434
Bildung:LandUSA  20.0183     0.1475 135.732   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.26 on 696 degrees of freedom
Multiple R-squared:  0.9992,  Adjusted R-squared:  0.9992
F-statistic: 3.041e+05 on 3 and 696 DF,  p-value: < 2.2e-16
```



Hypothesentests in R III

Interpretation:

- Wir gehen davon aus, dass es einen Zusammenhang zwischen Einkommen und Bildung in Deutschland gibt.
- Wir gehen davon aus, dass der Zusammenhang zwischen Einkommen und Bildung in Deutschland und den USA unterschiedlich groß, also dass eine Interaktion zwischen Bildung und Land vorliegt.

Bemerkungen I

In vielen Fällen ist es sinnvoll, die stetige UV X_i zu z-standardisieren. Dies hat zwei Vorteile:

- Die Parameter α und β_2 können sinnvoll interpretiert werden, da es sich bei einem Wert von 0 auf der stetigen UV dann um den Durchschnittswert der Personen auf der stetigen UV handelt. In unserem Beispiel wäre α dann das erwartete monatliche Einkommen einer Person in Deutschland mit durchschnittlicher Bildung und β_2 der Unterschied im monatlichen Einkommen zwischen einer Person in den USA und einer Person in Deutschland mit durchschnittlicher Bildung.
- Nach der z-Standardisierung ist in vielen Fällen die Korrelationen zwischen X_i, D_i und $X_i \cdot D_i$ geringer und somit auch die Standardfehler der Schätzfunktionen B_1, B_2 und B_3. Man darf dabei jedoch nicht vergessen, dass die Parameter im Modell mit z-standardisierter stetiger UV nicht mehr die gleiche Interpretation aufweisen.

Bemerkungen II

- Das Modell mit stetiger UV und diskreter UV mit Interaktionsterm kann leicht auf diskrete Prädiktoren mit k (kategorialen) Ausprägungen erweitert werden.
- Man nimmt dann einfach den stetigen Prädiktor, alle $k-1$ Dummy-Variablen und die Produkte aller Dummy-Variablen mit dem stetigen Prädiktor in das Modell auf.
- Beispiel für Haarfarbe mit $k=3$ Ausprägungen (schwarz, blond, braun) mit Referenzkategorie schwarz:

$$Y_i = \alpha + \beta_1 \cdot X_i + \beta_2 \cdot D_{braun_i} + \beta_3 \cdot D_{blond_i} + \beta_4 \cdot (X_i \cdot D_{braun_i}) + \beta_5 \cdot (X_i \cdot D_{blond_i}) + \varepsilon_i$$

- Die Interpretation der Parameter erhält man wieder durch die kategorienspezifischen Modellgleichungen:
  - α ist der Intercept in der Referenzkategorie schwarz
  - β_1 ist der Steigungsparameter in der Referenzkategorie schwarz
  - β_2 ist der Unterschied im Intercept zwischen braun und schwarz
  - β_3 ist der Unterschied im Intercept zwischen blond und schwarz
  - β_4 ist der Unterschied im Steigungsparameter zwischen braun und schwarz
  - β_5 ist der Unterschied im Steigungsparameter zwischen blond und schwarz

Zwischengliederung

- Bislang:
  - Regressionsmodelle mit einem diskreten Prädiktor mit zwei (kategorialen) Ausprägungen
  - Regressionsmodelle mit einem diskreten Prädiktor mit mehr als zwei (kategorialen) Ausprägungen
  - Regressionsmodelle mit mehreren diskreten (kategorialen) Prädiktoren
  - Regressionsmodelle mit stetigen und diskreten (kategorialen) Prädiktoren

- Jetzt:
  - Regressionsmodelle mit Interaktion zwischen stetigen Prädiktoren

Regressionsmodell mit zwei stetigen Prädiktoren und einem Interaktionsterm I

Ein Interaktionsterm kann natürlich auch für zwei stetige Prädiktoren gebildet und in das Regressionsmodell aufgenommen werden: Erweiterung der MLR mit dem Produkt aus X_{i1} und X_{i2} (Interaktionsterm) liefert die folgende allgemeine Modellgleichung:

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 (X_{i1} \cdot X_{i2}) + \varepsilon_i, \qquad \text{mit } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- Die Realisation von Y_i ist der Wert der zufällig gezogenen Person i auf der AV.
- Die Realisation von X_{i1} ist der Wert der zufällig gezogenen Person i auf der ersten stetigen UV.
- Die Realisation von X_{i2} ist der Wert der zufällig gezogenen Person i auf der zweiten stetigen UV.
- ε_i ist ein zufälliger Fehler.
- α, β_1, β_2, β_3 und σ^2 sind Modellparameter.

Regressionsmodell mit zwei stetigen Prädiktoren und einem Interaktionsterm II

Alternativ lässt sich die Modellgleichung umformulieren als:

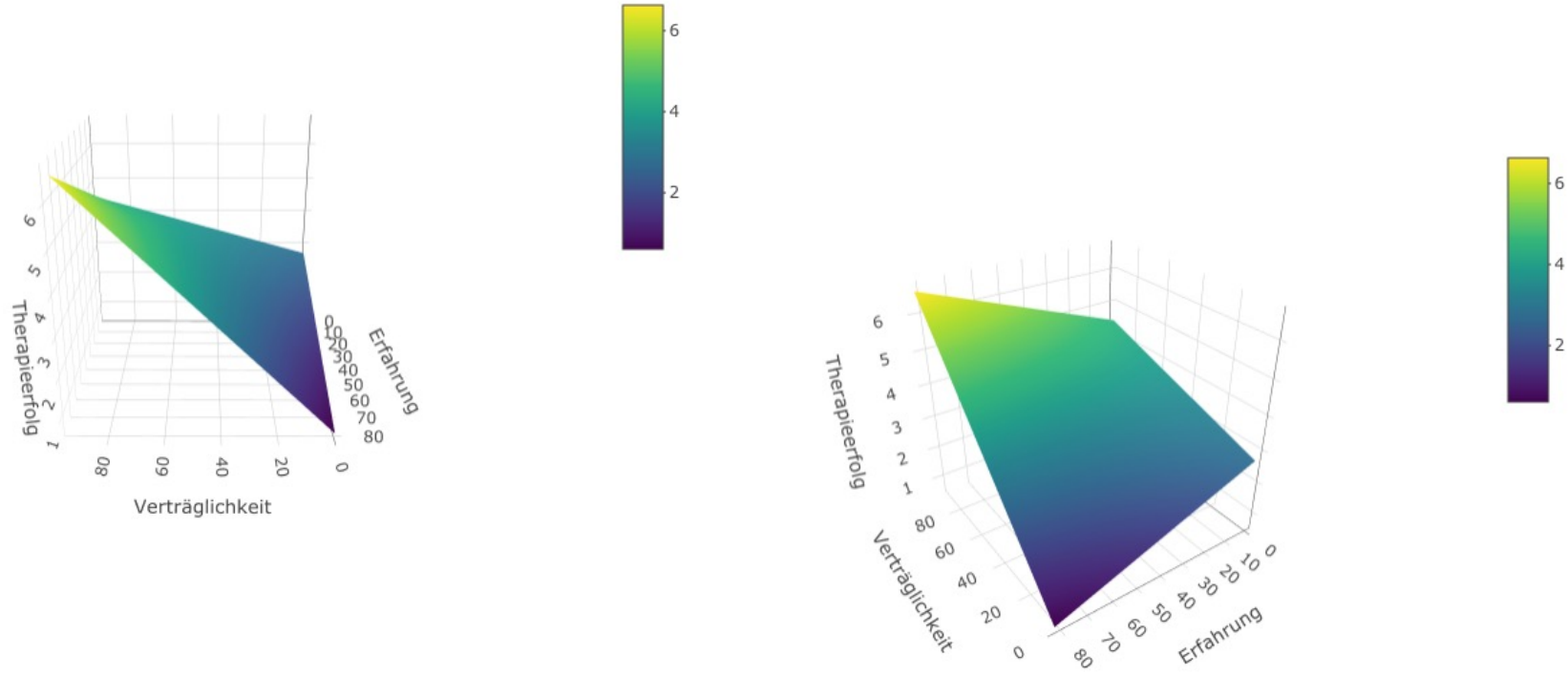
$$Y_i = \alpha + \beta_2 \cdot X_{i2} + (\beta_1 + \beta_3 \cdot X_{i2}) \cdot X_{i1} + \varepsilon_i, \qquad \text{mit } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$(bzw.\ Y_i = \alpha + \beta_1 \cdot X_{i1} + (\beta_2 + \beta_3 \cdot X_{i1}) \cdot X_{i2} + \varepsilon_i), \qquad \text{mit } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- Anhand dieser Umformung erkennt man, dass das Produkt $\beta_3 \cdot X_{i2}$ eine Veränderung der Steigung β_1 angibt.
- Wenn $\beta_3 > 0$ bedeutet das, dass wenn X_{i2} um eine Einheit steigt, die Steigung im Bezug auf die erste UV um β_3 steigt.
- Man spricht dabei auch von einer Moderation, bzw. moderierten Regression: je nachdem wie die zweite UV ausgeprägt ist, fällt der Zusammenhang zwischen der AV und der ersten UV unterschiedlich aus.
- Dabei ist es egal, ob die sogenannte Moderatorvariable stetig ist (hier: X_{i2}) oder diskret (Modell auf Folie 42ff.), wobei die Interpretation im diskreten Fall einfacher ist.
- Ob X_{i1} (lila Variante) oder X_{i2} als „Moderator“ definiert wird ist dabei einzig inhaltlich zu begründen und macht auf Modellebene **keinen** Unterschied!

Regressionsmodell mit zwei stetigen Prädiktoren und einem Interaktionsterm - Interpretation

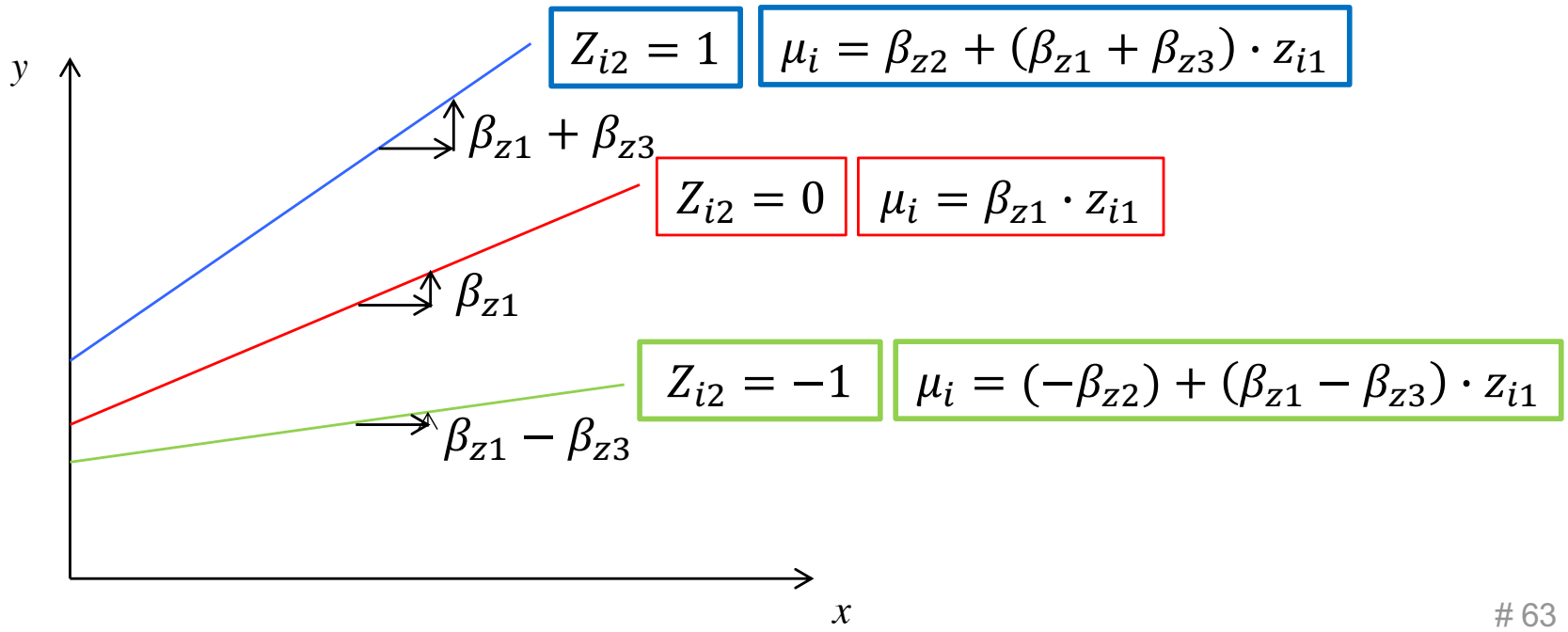
- Graphische Darstellung über Regressionsebene
- Fiktives Beispiel: Je höher die Erfahrung einer Therapeut*in, desto stärker ist der Einfluss ihrer Verträglichkeit auf den Therapieerfolg



Regressionsmodell mit zwei stetigen Prädiktoren und einem Interaktionsterm - Interpretationshilfe

- Die Interpretation einer solchen Interaktion ist aufgrund der Kontinuität des Moderators, bzw. der zweiten UV schwieriger als im dichotomen Fall, weshalb man sich für die Interpretation häufig bestimmte Ausprägungen der Moderatorvariable herausgreift und für diese den Zusammenhang betrachtet:

- Im Falle einer z-standardisierter Variablen, könnte man beispielsweise Regressionsgeraden für die Fälle $Z_{i2} = -1$, $Z_{i2} = 0$ und $Z_{i2} = 1$ anschauen, d.h. den Zusammenhang der ersten UV mit der AV für die Fälle, dass die Moderatorvariable sich in einem Wert eine Standardabweichung unter dem Mittelwert realisiert, im Mittelwert realisiert oder eine Standardabweichung über dem Mittelwert realisiert:



Regressionsmodell mit zwei stetigen Prädiktoren und einem Interaktionsterm in R I

Beispiel: Die Länge des Klinikaufenthalts (in Tagen) von Patient*innen mit Depression wird auf Basis des BDI-Scores (zentriert) und ihres Alters (zentriert) vorhergesagt:

```
Call:
lm(formula = Tage_in_Klinik ~ BDI.c * Alter.c)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8705 -1.8637  0.0018  2.2286  9.1870

Coefficients:
              Estimate  Std. Error  t value     Pr(>|t|)
(Intercept)  97.695319   0.219880  444.313   < 2e-16 ***
BDI.c         3.910981   0.048202   81.137   < 2e-16 ***
Alter.c       1.247919   0.022266   56.047   < 2e-16 ***
BDI.c:Alter.c 0.041616   0.004908    8.478   5.47e-15 ***
---
```

Schätzwert für α, also die erwartete Dauer des Klinikaufenthaltes einer durchschnittlich alten und durchschnittlich depressiven Person

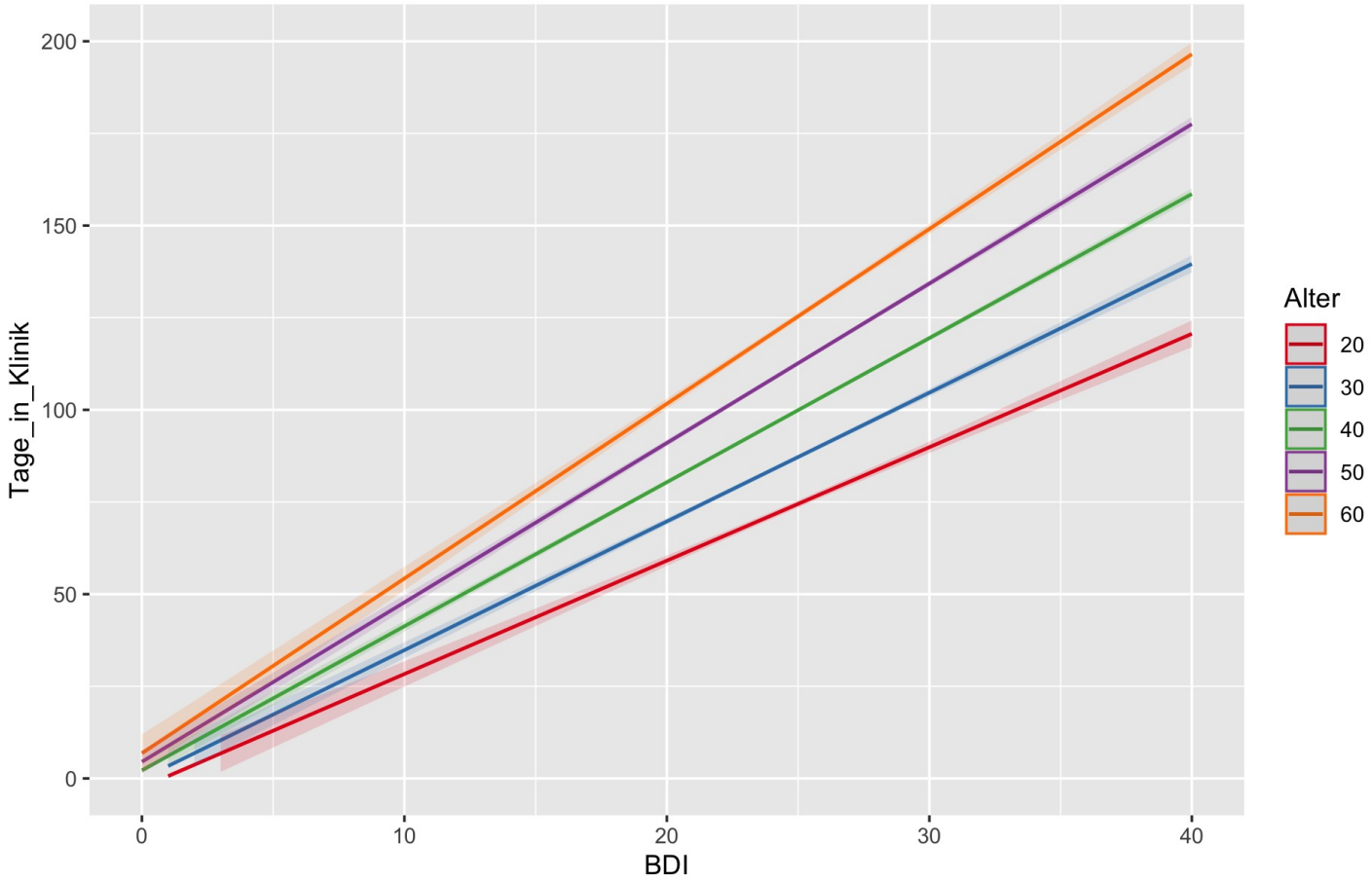
Schätzwert für β_1, also die durchschnittliche Zunahme des Klinikaufenthaltes für eine durchschnittlich alte Person pro Punkt im BDI

Schätzwert für β_2, also die durchschnittliche Zunahme des Klinikaufenthaltes für durchschnittlich depressive Personen pro Lebensjahr

Schätzwert für β_3, also die durchschnittliche Änderung pro Lebensjahr der Zunahme des Klinikaufenthalts pro Punkt im BDI
(Alter und BDI könnten auch getauscht werden)

Regressionsmodell mit zwei stetigen Prädiktoren und einem Interaktionsterm in R II

- Beispiel: Die Länge des Klinikaufenthalts (in Tagen) von Patient*innen mit Depression wird auf Basis des BDI-Scores und des Alters vorhergesagt:



Zwischengliederung

- Bislang:
  - Regressionsmodelle mit einem diskreten Prädiktor mit zwei (kategorialen) Ausprägungen
  - Regressionsmodelle mit einem diskreten Prädiktor mit mehr als zwei (kategorialen) Ausprägungen
  - Regressionsmodelle mit mehreren diskreten (kategorialen) Prädiktoren
  - Regressionsmodelle mit stetigen und diskreten (kategorialen) Prädiktoren
  - Regressionsmodelle mit Interaktion zwischen stetigen Prädiktoren

- Jetzt:
  - Weitere inferenzstatistische Verfahren im Rahmen von Regressionsmodellen

Weitere inferenzstatistische Verfahren im Rahmen von Regressionsmodellen

- Bei der Interpretation der Regressionsmodelle mit stetigen Prädiktoren, diskreten Prädiktoren und Interaktionen haben wir gesehen, dass inhaltliche Fragestellungen oft mit gerichteten Hypothesentests bezüglich mehrerer Parameter oder Kombinationen von Parametern einher gehen.
- Bisher haben wir für einzelne β_j nur ungerichtete Hypothesentests der Form $H_0\colon \beta_j = 0$ bzw. $H_1\colon \beta_j \neq 0$ sowie Konfidenzintervalle für einzelne β_j betrachtet.
- Genau wie in den varianzanalytischen Modelle ist es jedoch auch im Rahmen von Regressionsmodellen möglich, gerichtete oder ungerichtete Hypothesentests sowie Konfidenzintervalle für beliebige Parameter oder Parameterkombinationen zu betrachten. Dies gilt auch für weitere Regressionsmodelle wie zum Beispiel der logistischen Regression (siehe nächste Vorlesung).
- Werden Hypothesentests mit zusammengesetzten Hypothesen durchgeführt, ist eine Korrektur der p-Werte mit der Tukey-Methode möglich.
- Wie auch bei den varianzanalytischen Modellen erfolgt die Berechnung in R mit dem multcomp Paket. Ein Beispiel mit Anleitung finden Sie in Übungsblatt 10.