

# 6. Vorlesung Statistik II

## Einführung in die Regressionsanalyse



We are happy to share our materials openly:

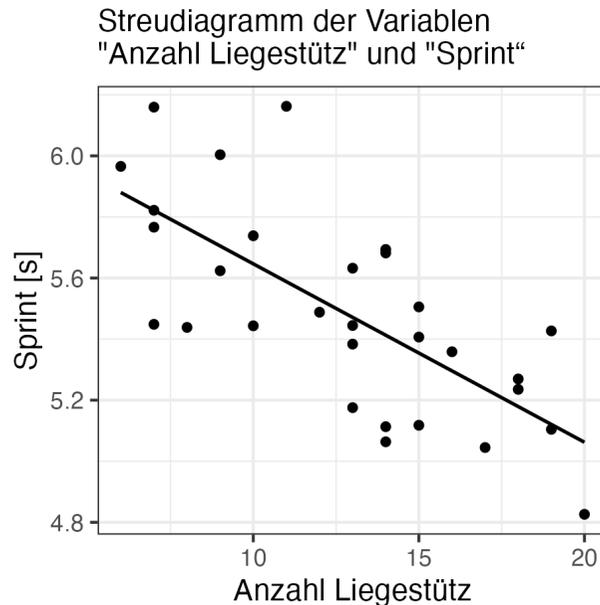
The content of these Open Educational Resources by Lehrstuhl für Psychologische Methodenlehre und Diagnostik, Ludwig-Maximilians-Universität München is licensed under CC BY-SA 4.0. The CC Attribution-ShareAlike 4.0 International license means that you can reuse or transform the content of our materials for any purpose as long as you cite our original materials and share your derivatives under the same license.

- Allen varianzanalytischen Modellen ist gemeinsam, dass die **unabhängigen** Variablen **diskret** sind und die **abhängige** Variable **stetig**.
- In Regressionsmodellen können sowohl die **unabhängigen** Variablen als auch die **abhängige** Variable entweder **stetig** oder **diskret** sein.
- Wir beginnen mit dem einfachsten Regressionsmodell, der so genannten **einfachen linearen Regression** (ELR).
- Bei einer einfachen linearen Regression wird der lineare Zusammenhang zwischen **zwei stetigen Variablen** überprüft.
- Es ist üblich, in Regressionsmodellen die UV als **Prädiktor** und die AV als **Kriterium** zu bezeichnen.

- Sie kennen die Annahmen des Regressionsmodells und sind in der Lage, bedingte Erwartungswerte inhaltlich zu interpretieren.
- Sie sind in der Lage, die Parameter des Modells zu interpretieren und wissen, welche inhaltlichen Fragestellungen damit beantwortbar wären.
- Sie können eine Zentrierung in R praktisch anwenden und wissen, wie Sie sich diese zu Nutze machen können um Parameter anders zu interpretieren als im unzentrierten Modell.
- Sie sind in der Lage, die Schätzung eines Regressionsmodells in R durchzuführen, aus den Outputs die (Intervall-)Schätzungen der Parameter herauszulesen und die geschätzte Regressionsgerade und geschätzte Modellgleichung daraus zu bilden.
- Sie können einfache (ungerichtete) Hypothesentests über die Parameter in R durchführen und inhaltlich interpretieren.
- Sie sind in der Lage Vorhersagewerte und deren Konfidenzintervalle in R zu berechnen und zu interpretieren.

- **Untersuchungsfrage:** Existiert ein linearer Zusammenhang, also eine Abhängigkeit, zwischen der negativen Selbstbewertung und der Depressionsschwere von Personen?
- **AV:** Anzahl der Punkte im Beck-Depressions-Inventar (Testverfahren, das die Schwere der depressiven Symptomatik im klinischen Bereich erfasst)  
Variablenname: „bdi-ges“
- **UV:** Einschätzung der negativen Selbstbewertung (Skala des Fragebogens zur Erfassung irrationaler Einstellungen)  
Variablenname: „fie\_nsb“
- **Beispielitem:** „Ich denke oft, ich bin ein Versager.“

- Zur Überprüfung der Untersuchungsfrage wird aus der Population derjenigen Personen, die akut an einer Depression erkrankt sind, eine **einfache Zufallsstichprobe** vom Umfang  $n$  gezogen.
- Bei jeder Person soll die Depressionsausprägung und die Stärke der negativen Selbstbewertung gemessen werden.
- Beide Variablen stellen **Zufallsvariablen (ZV)** dar.



graphische Darstellung eines linearen  
Zusammenhangs auf Ebene der  
Stichprobe

- Wenn ein linearer Zusammenhang zwischen zwei stetigen Variablen besteht, dann tragen die beiden Variablen einen gemeinsamen Informationsanteil: Die eine Variable kann die andere Variable zum Teil „erklären“.
- Bsp.: Je höher die Anzahl der Liegestütze, desto niedriger die gemessene Zeit im Sprint.
- „Idee“: Formulierung einer Modellgleichung, in der die systematische Komponente als linearer Zusammenhang zwischen einer ZV (dem Prädiktor) und einer anderen ZV (dem Kriterium) dargestellt wird.
- Bemerkung: In Statistik I (VL 3) haben wir bereits gesehen, dass die Stärke eines linearen Zusammenhangs im Sinne der Korrelation mit der Steigung der Geraden im Streudiagramm zusammenhängt. Dies werden wir in der nächsten Vorlesung zu Effektstärken im Rahmen der ELR noch einmal aufgreifen.

- Beispiel aus VL 5: Mittlere Depressionsschwere bei jungen Erwachsenen, Erwachsenen und alten Erwachsenen
- Im einfaktoriellen varianzanalytischen Modell entspricht der Erwartungswert  $\mu_{jung}$  der mittleren Depressionsschwere in der Population der jungen Erwachsenen.
- Angenommen wir definieren eine Zufallsvariable  $X_i$ , die angibt in welcher Population (junge Erwachsene, Erwachsene, alte Erwachsene) sich eine zufällig aus der Gesamtpopulation gezogene Person  $i$  befindet. Dann könnten wir  $\mu_{jung}$  auch als sogenannten bedingten Erwartungswert ausdrücken:

$$E(Y_i|X_i = \text{„junge Erwachsene“}) = \mu_{jung}$$

- Definition: Der **bedingte Erwartungswert**  $E(Y_i|X_i = x_i)$  entspricht dem Erwartungswert der Zufallsvariable  $Y_i$  unter der Bedingung, dass sich die Zufallsvariable  $X_i$  im Wert  $x_i$  realisiert hat.
- Interpretation: Der **bedingte Erwartungswert**  $E(Y_i|X_i = x_i)$  entspricht dem **Mittelwert der Variable  $y$**  von allen **Personen in der Population mit einem konkreten Wert von  $x_i$**  auf der Variable  $x$ .

## Bedingter Erwartungswert (II)

- Die Definition eines bedingten Erwartungswerts funktioniert auch für den Fall, dass  $X_i$  eine stetige Zufallsvariable darstellt, z.B. das Alter in Jahren.
- Im Beispiel:  $E(Y_i|X_i = 50)$  entspricht der mittleren Depressionsschwere von Personen mit einem Alter von 50 Jahren
- Wenn  $X_i$  eine stetige Zufallsvariable darstellt, gibt es theoretisch unendlich viele verschiedene bedingte Erwartungswerte  $E(Y_i|X_i = x_i)$ , da  $X_i$  theoretisch unendlich viele verschiedene Werte annehmen kann.
- Anstatt  $\mu_j$  also mithilfe der möglichen Ausprägungen von  $X_i$  zu indizieren, ist es praktischer, die bedingten Erwartungswerte mit dem Personenindex  $i$  zu kennzeichnen:
- Im Beispiel:  $E(Y_i|X_i = x_i) = \mu_i$  entspricht damit der mittleren Depressionsschwere von Personen in der Population, die das gleiche Alter haben, wie die zufällig in die Stichprobe gezogenen Person  $i$ .

- Statistisches Modell der einfachen linearen Regression:

$$Y_i = \alpha + \beta \cdot X_i + \varepsilon_i, \quad \text{wobei } \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

- $Y_i$  ist eine **ZV**. Ihre Realisation  $y_i$  repräsentiert den Variablenwert der zufällig gezogenen Person  $i$  auf der **AV**.
- $X_i$  ist eine **ZV**. Ihre Realisation  $x_i$  repräsentiert den Variablenwert der zufällig gezogenen Person  $i$  auf der **UV**.
- $\alpha, \beta$  und  $\sigma^2$  sind die **Modellparameter**. Sie sind **unbekannte Konstanten**.
- Es gilt für die bedingten Erwartungswerte:

$$E(Y_i | X_i = x_i) = \mu_i = \alpha + \beta \cdot x_i$$

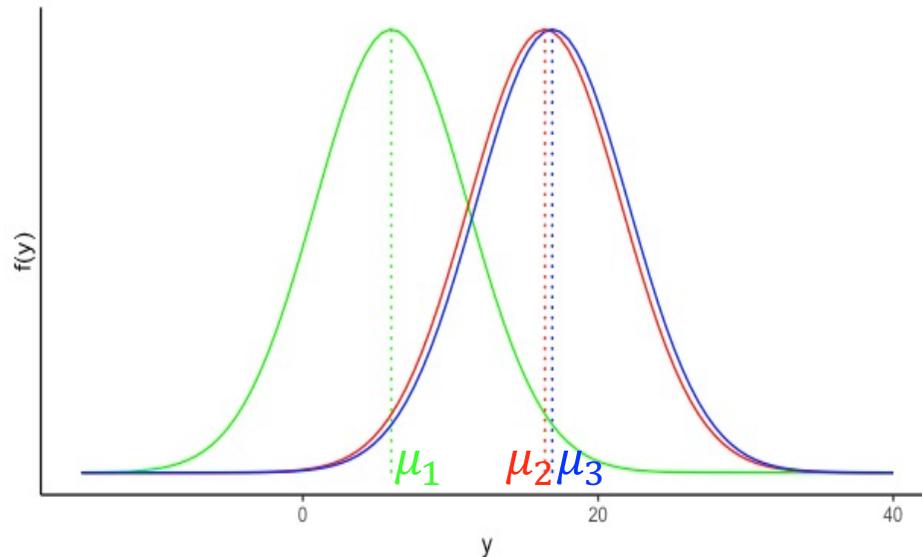
- D.h. die Modellgleichung impliziert, dass...
  - ... die Abweichung der ZV  $Y_i$  von ihrem bedingten Erwartungswert an der Stelle  $x_i$  einzig durch die unsystematische Fehlervariable  $\varepsilon_i$  verursacht wird.
  - ... ein **linearer Zusammenhang** zwischen dem Prädiktor und dem Kriterium besteht.

## Definition: Fehlervariable $\varepsilon_i$

- Die Fehlervariable  $\varepsilon_i$  ist in der ELR allgemein definiert als die Differenz zwischen der Zufallsvariable  $Y_i$  von ihrem bedingten Erwartungswert an der Stelle  $x_i$ :

$$\varepsilon_i = Y_i - E(Y_i | X_i = x_i)$$

- Die Fehlervariable  $\varepsilon_i$  ist damit selbst auch eine Zufallsvariable.



- Bsp.: Mittlere Depression bei jungen Erwachsenen ( $j = 1$ ), Erwachsenen ( $j = 2$ ) und alten Erwachsenen ( $j = 3$ )

- Erste Darstellungsform:

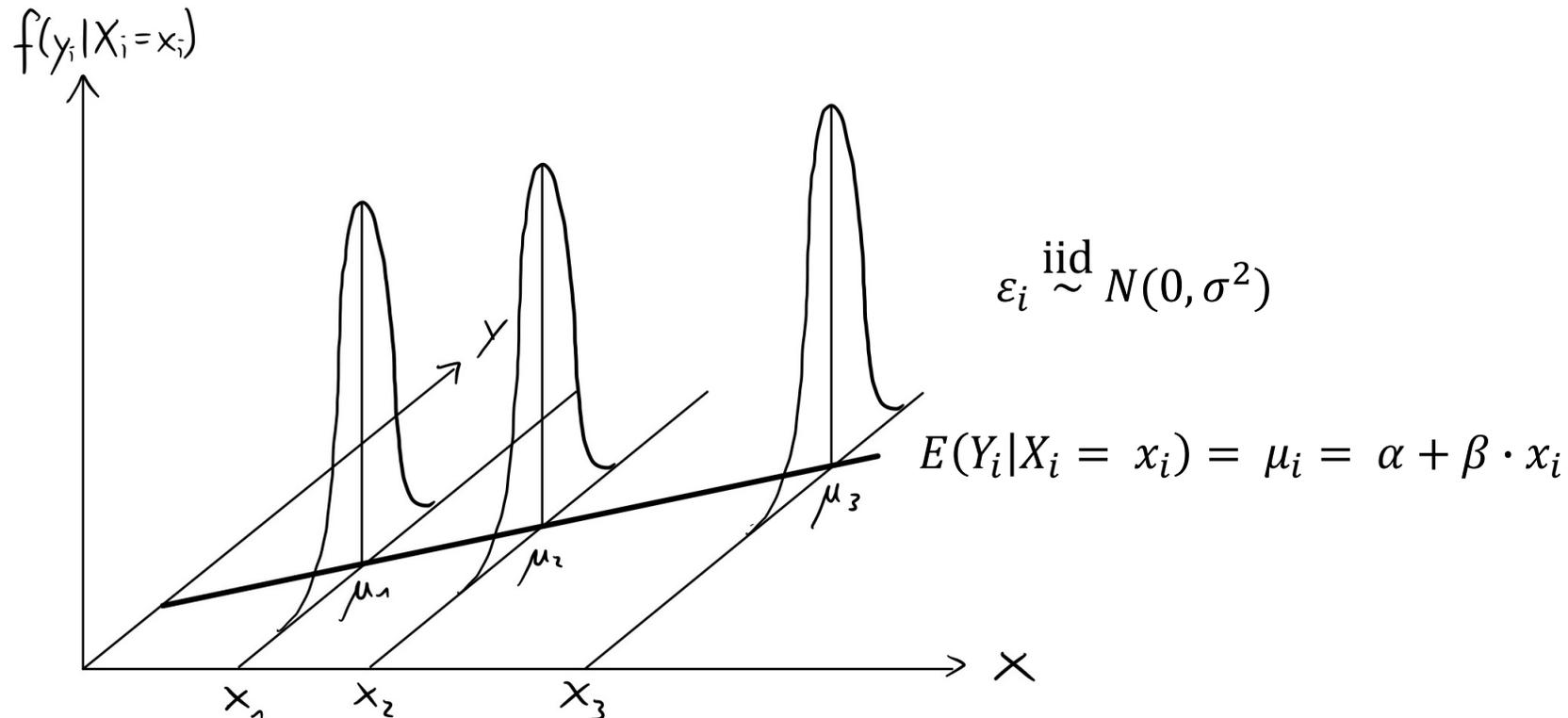
$$Y_{ij} = \mu_j + \varepsilon_{ij},$$

mit  $j = 1, 2, 3$  und  $i = 1, \dots, n_j$

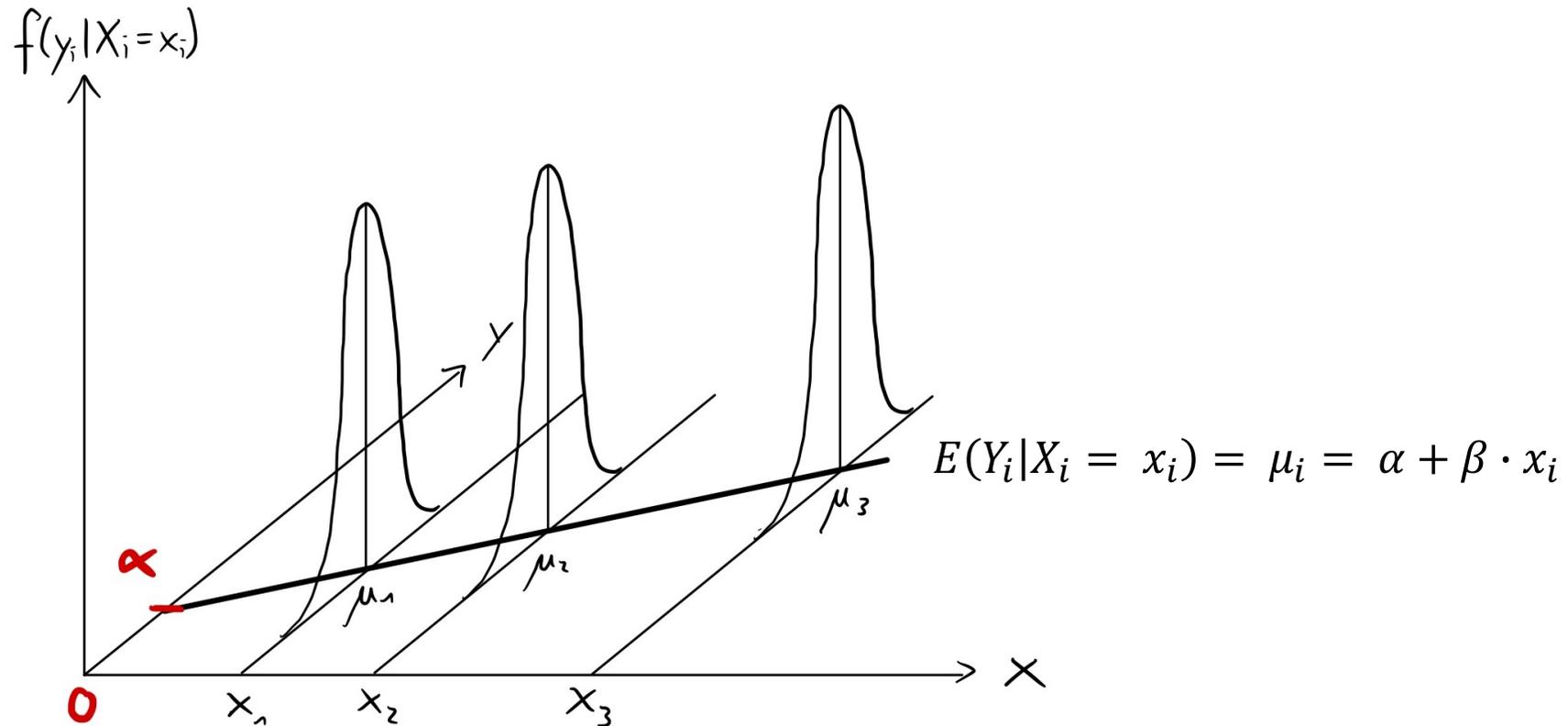
- Veranschaulichung der Verteilungen der Depression in den 3 Altersgruppen:
- Annahmen des varianzanalytischen Modells:

$$Y_i \stackrel{\text{ind}}{\sim} N(\mu_j, \sigma^2)$$

- Die Erwartungswerte der Normalverteilungen dürfen sich unterscheiden.
- Die Varianzen der Normalverteilungen müssen gleich sein.

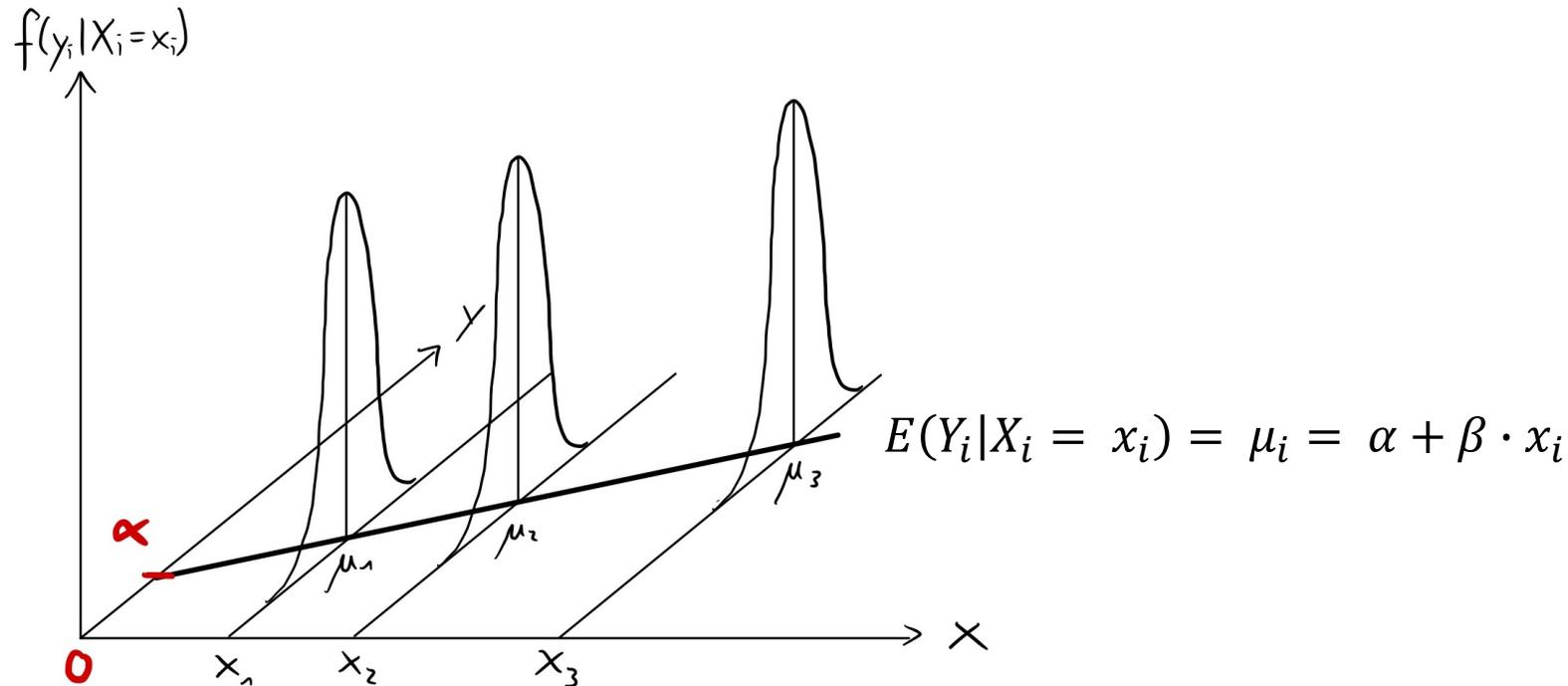


- Der Modellparameter  $\alpha$  wird auch als **Intercept (Achsenabschnitt)** bezeichnet.
- Der Modellparameter  $\beta$  wird auch als **Slope (Steigungsparameter)** bezeichnet.
- Der Modellparameter  $\sigma^2$  wird auch als **Fehlervarianz** bezeichnet.

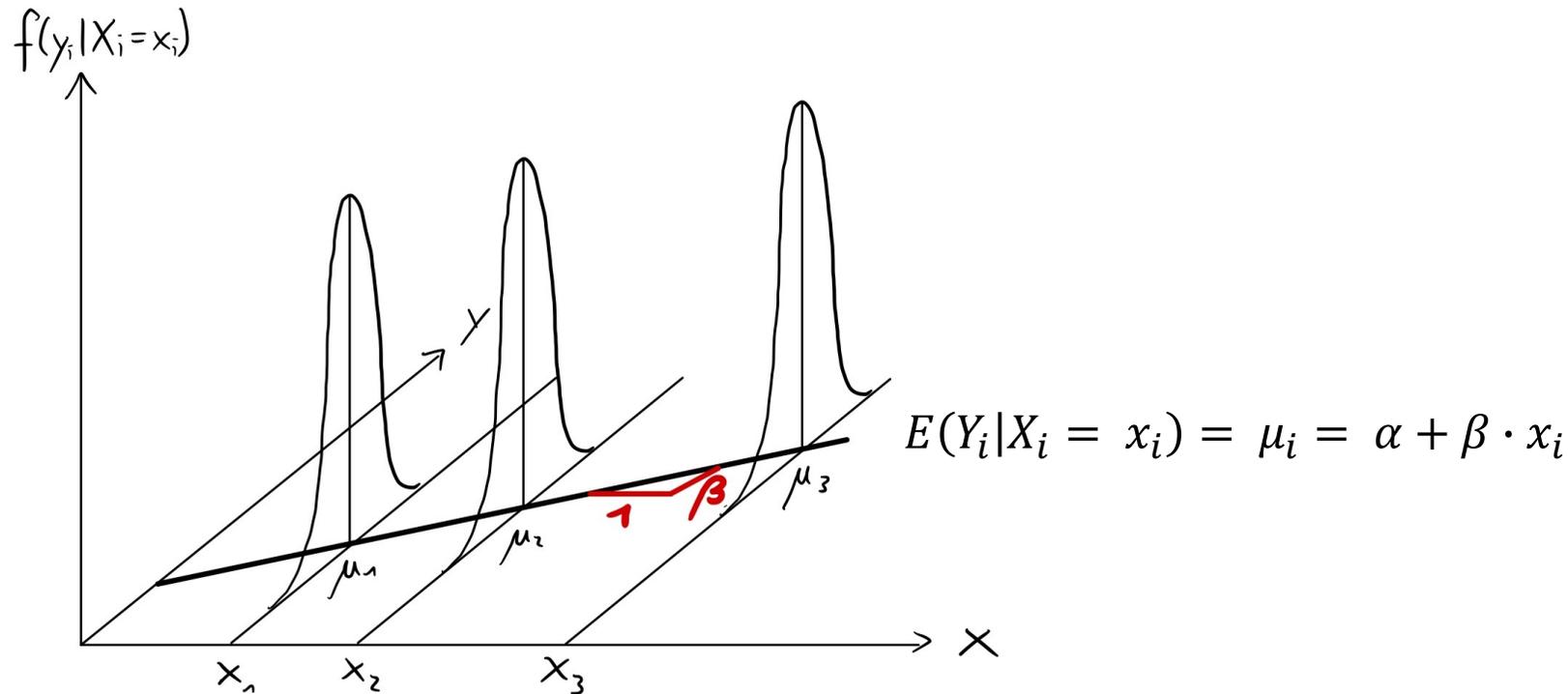


- Der Achsenabschnitt  $\alpha$  gibt an, in welchem  $y$ -Wert die Gerade die  $y$ -Achse schneidet. Man erhält  $\alpha$ , indem man in der Geradengleichung die UV gleich null setzt:

$$E(Y_i | X_i = 0) = \alpha + \beta \cdot 0 = \alpha$$

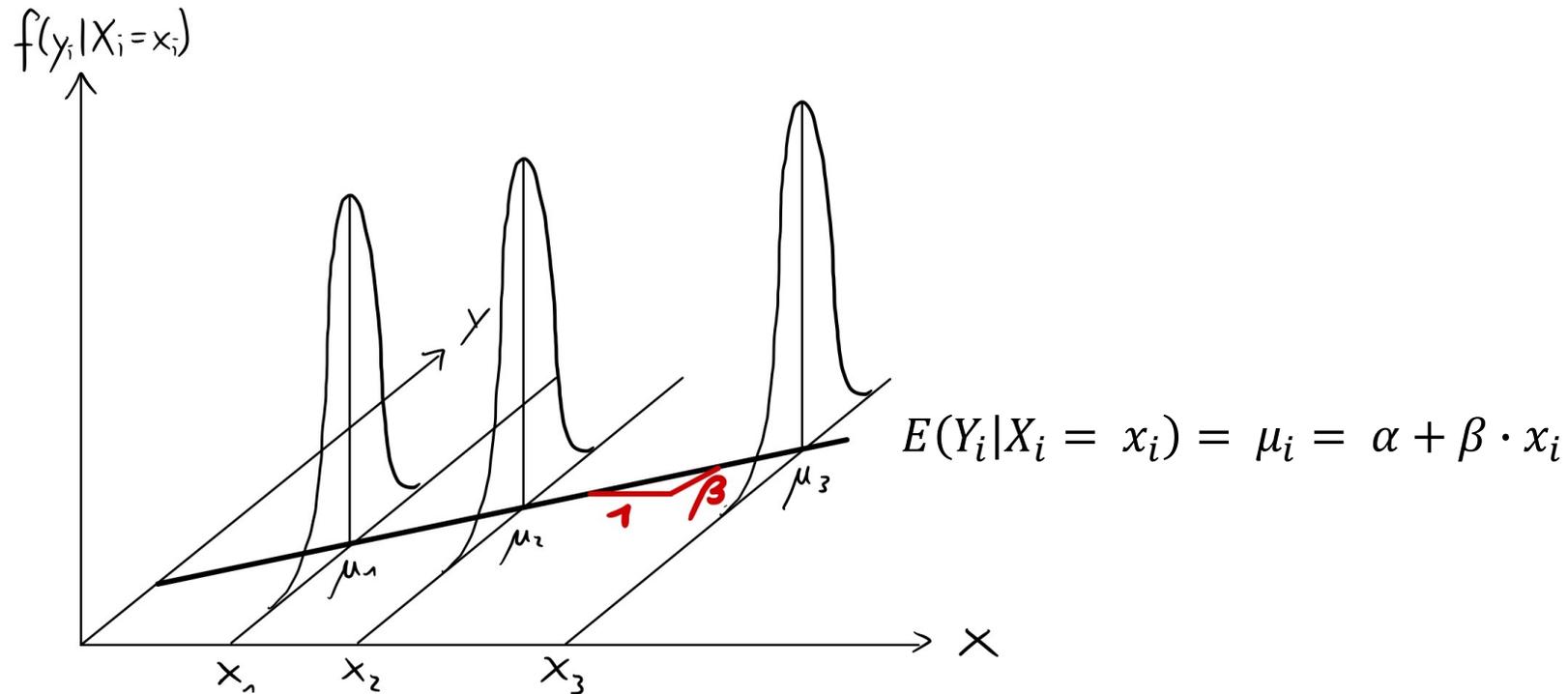


- Bezogen auf unser Beispiel repräsentiert  $\alpha$  den bedingten Erwartungswert (und damit die mittlere Depressionsausprägung) im Falle, dass Personen eine negative Selbstbewertung von  $x_i = 0$  aufweisen.
- Ob der Modellparameter  $\alpha$  sinnvoll interpretiert werden kann, hängt davon ab, ob  $x_i = 0$  einen inhaltlich plausiblen Wert darstellt.
- Durch eine **Zentrierung** des Prädiktors kann sich die Bedeutung von  $x_i = 0$  und damit auch von  $\alpha$  verändern (siehe auch Folie 38ff.).



- Der Steigungsparameter  $\beta$  gibt an, wie stark die Gerade steigt bzw. fällt. Die Steigung der Geraden ist an jeder Stelle  $X_i = x_i$  konstant.
- Ist  $\beta = 0$  ergibt sich eine konstante Funktion: Alle bedingten Erwartungswerte sind gleich dem Achsenabschnitt  $\alpha$ . Die Zufallsvariable  $Y_i$  setzt sich lediglich aus der Konstanten  $\alpha$  und der Fehlervariable  $\varepsilon_i$  zusammen:

$$Y_i = \alpha + 0 \cdot X_i + \varepsilon_i = \alpha + \varepsilon_i$$



- Bezogen auf unser Beispiel gibt der Steigungsparameter  $\beta$  an, wie stark sich der Mittelwert der Depressionsausprägung verändert (erhöht), wenn die negative Selbstbewertung um eine Einheit zunimmt.
- Beispiel: Wenn sich die negative Selbstbewertung um **eine Einheit** erhöht, dann erhöht sich der **bedingte Erwartungswert** und damit der **Mittelwert** der Depressionsschwere um z.B. 1.5 Einheiten.

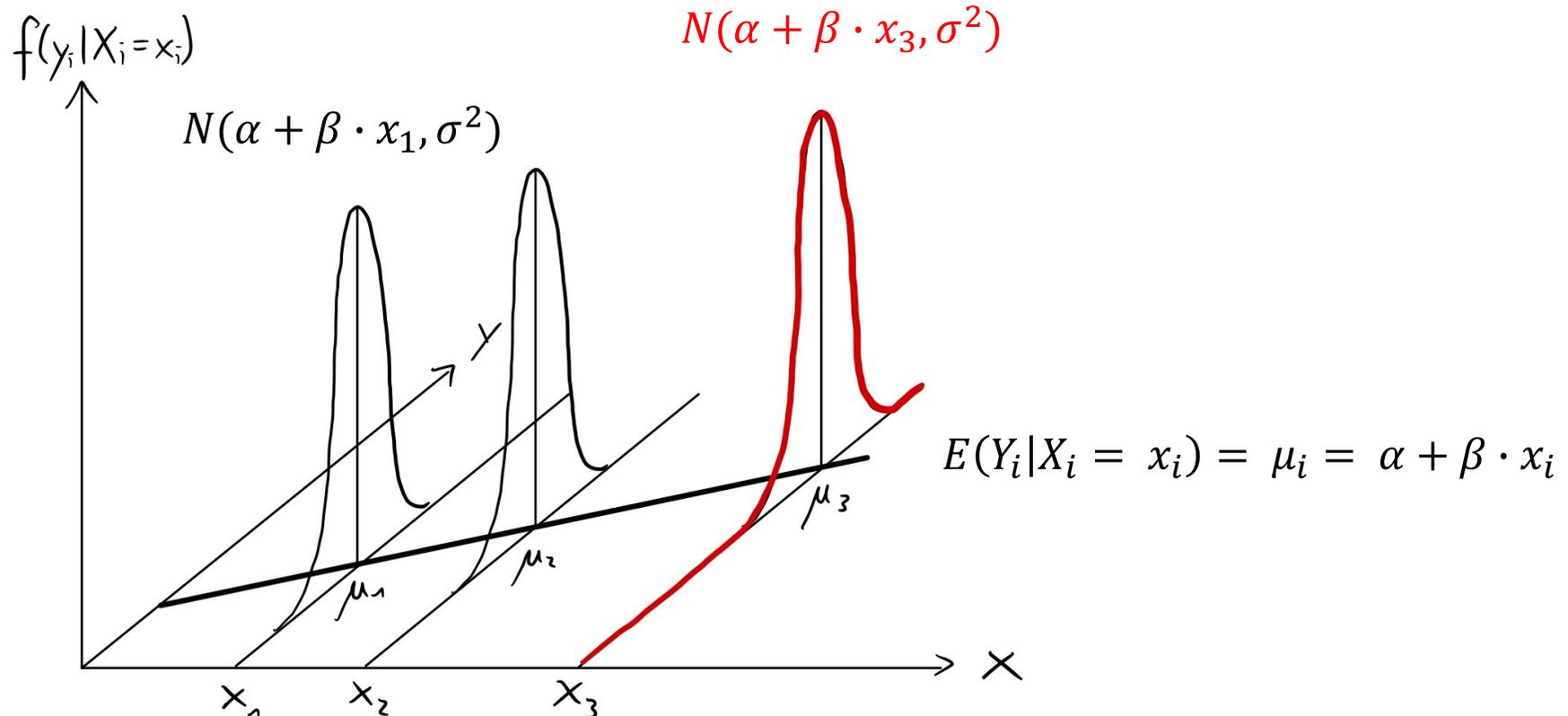
Im Modell der ELR werden die folgenden Annahmen getroffen:

1. Die **UV** und die **AV** hängen **linear** zusammen:

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

2. Alle **Fehler**  $\varepsilon_i$  sind **unabhängig voneinander** und folgen einer **Normalverteilung** mit **Erwartungswert null** und **konstanter Varianz**  $\sigma^2$ :

$$\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

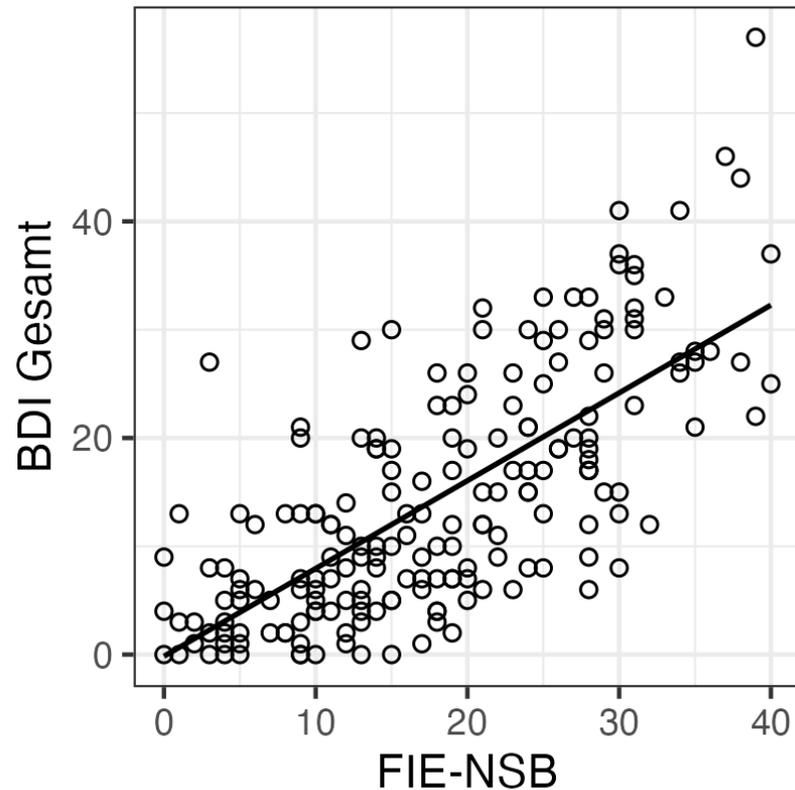


- Aus  $\varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$  ergibt sich für die bedingte Verteilung der AV:  
An jeder beliebigen Stelle auf der x-Achse ist die AV in y-Richtung normalverteilt mit Erwartungswert  $\mu_i = \alpha + \beta \cdot x_i$  und konstanter Varianz  $\sigma^2$ .
- Die Modellannahmen werden wir in der nächsten Vorlesung weiter diskutieren.

- Das Modell der ELR enthält die folgenden Parameter:
  - $\alpha$
  - $\beta$
  - $\sigma^2$
- Alle diese Parameter können geschätzt werden und für alle diese Parameter können statistische Hypothesentests konstruiert werden. Dies gilt auch für Kombinationen der Parameter.
- Je nach konkreter Fragestellung muss entschieden werden, ob und welche Parameter geschätzt werden sollen bzw. ob und welche Hypothesen getestet werden sollen.

- Bislang:
  - Modell der ELR
  
- Jetzt:
  - Parameterschätzung

# Definition: Regressionsgerade und geschätzte Regressionsgerade



- Die **Regressionsgerade** beschreibt den Zusammenhang zwischen UV und AV in der **Population**. Diese „wahre“ Regressionsgerade ist unbekannt und prinzipiell in der Praxis nicht beobachtbar. Sie kann jedoch mithilfe der Stichprobe geschätzt werden.
- Unter der **geschätzten Regressionsgeraden** versteht man die Gerade, die „optimal“ an die Daten in der **Stichprobe** angepasst ist. Sie wird üblicherweise im Streudiagramm der Stichprobendaten von UV und AV eingezeichnet.

Wie in allen statistischen Modellen wird bei der ELR zwischen unbekanntem Modellparametern, Schätzfunktionen und Schätzwerten unterschieden.

**Modellparameter**  
(Bestandteil der  
**unbekannten**  
Regressionsgeraden)

$$\alpha$$

$$\beta$$

$$\sigma^2$$

**Schätzfunktion**

$$\hat{\alpha} = A$$

$$\hat{\beta} = B$$

$$\hat{\sigma}^2 = S^2$$

**Schätzwert**  
(Bestandteil der  
**geschätzten**  
Regressionsgeraden)

$$a$$

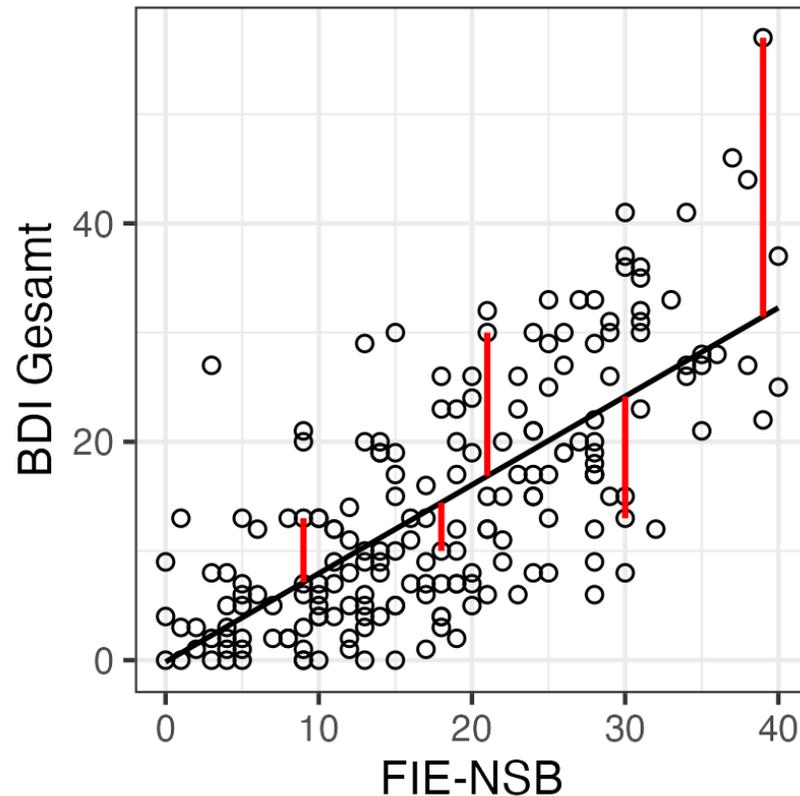
$$b$$

$$s^2$$

- Es lässt sich zeigen, dass erwartungstreue, konsistente und effiziente Schätzfunktionen für die unbekannt Parameter  $\alpha$ ,  $\beta$  und  $\sigma^2$  existieren.
- Die Schätzfunktionen für die unbekannt Parameter sind:

$$\hat{\beta} = B = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\text{cov}(X, Y)}{S_x^2}$$

$$\hat{\alpha} = A = \bar{Y} - B \cdot \bar{X}$$



- Wie können  $A$  und  $B$  gewählt werden, dass sich die Gerade „optimal“ an die Daten anpasst?
- „Idee“: Die Gerade ist dann optimal gewählt, wenn die Abweichungen von der geschätzten Regressionsgeraden in  $y$ -Richtung über alle Personen gesehen minimal ist.
- Diese Bedingung kann formal mithilfe einer Extremwertaufgabe dargestellt werden. In die Berechnung gehen die quadratischen Abweichungen ein, da dies mathematisch gesehen von Vorteil ist.

- Die Methode, Schätzfunktionen durch die Minimierung der quadrierten Abweichungen von der geschätzten Regressionsgeraden zu konstruieren, wird allgemein als **Methode der kleinsten Quadrate** bezeichnet.
- Man erhält durch das Lösen der Extremwertaufgabe für die Schätzfunktion  $A$  bzw. für die Schätzfunktion  $B$ :

$$A = \bar{Y} - B \cdot \bar{X}$$

$$B = \frac{\text{cov}(X, Y)}{S_x^2}$$

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.15973    1.19345  -0.134    0.894
data$fie_nsb  0.81067    0.05758  14.080 <2e-16 ***
---
```

- Die Schätzwerte  $a$  und  $b$  können im R-Output abgelesen werden
- Der Schätzwert  $a = -0.16$  kann für unser Beispiel inhaltlich nicht sinnvoll interpretiert werden. (Es wäre ja die geschätzte mittlere Depressionsschwere bei einem Wert von 0 in der negativen Selbstbewertung. Beides sind Werte, die auf den jeweiligen Fragebogenskalen gar nicht möglich sind.)
- Der Schätzwert  $b = 0.81$  für den unbekanntem Steigungsparameter  $\beta$  wird wie folgt interpretiert: Steigt die negative Selbstbewertung um eine Einheit (also um einen Punkt), dann schätzen wir, dass sich im Mittel die Depressionsschwere um 0.81 Einheiten (also Punkte) erhöht.
- Bemerkung: da sich inhaltliche Fragestellungen häufig auf den Parameter  $\beta$  bzw. dessen Schätzung  $b$  beziehen, wird dieser häufig auch **Regressionsgewicht** genannt.

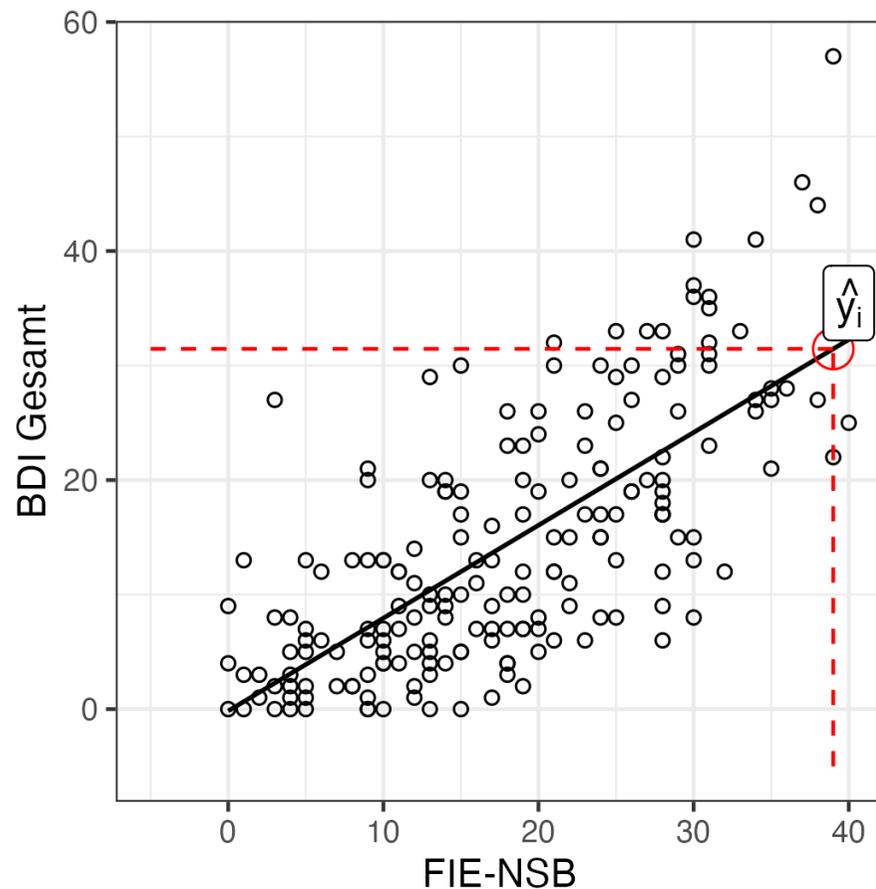
- In der Praxis betrachtet man neben der Fehlervarianz  $\sigma^2$  häufig auch die Standardabweichung der Fehler:  $\sigma = \sqrt{\sigma^2}$ .
- Es lässt sich zeigen, dass erwartungstreue, konsistente und effiziente Schätzfunktionen für die unbekannt Parameter  $\sigma^2$  bzw.  $\sigma$  existieren.
- Die Schätzfunktionen für die unbekannt Parameter sind:

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - (A + B \cdot X_i))^2$$

$$\hat{\sigma} = S = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - (A + B \cdot X_i))^2}$$

- Bemerkung: Die Realisation  $s$  als konkreter Schätzwert für  $\sigma$  wird in der Literatur häufig auch als **Standardschätzfehler** bezeichnet.

## Definition: Vorhersagewert (Synonym: gefitteter Wert)



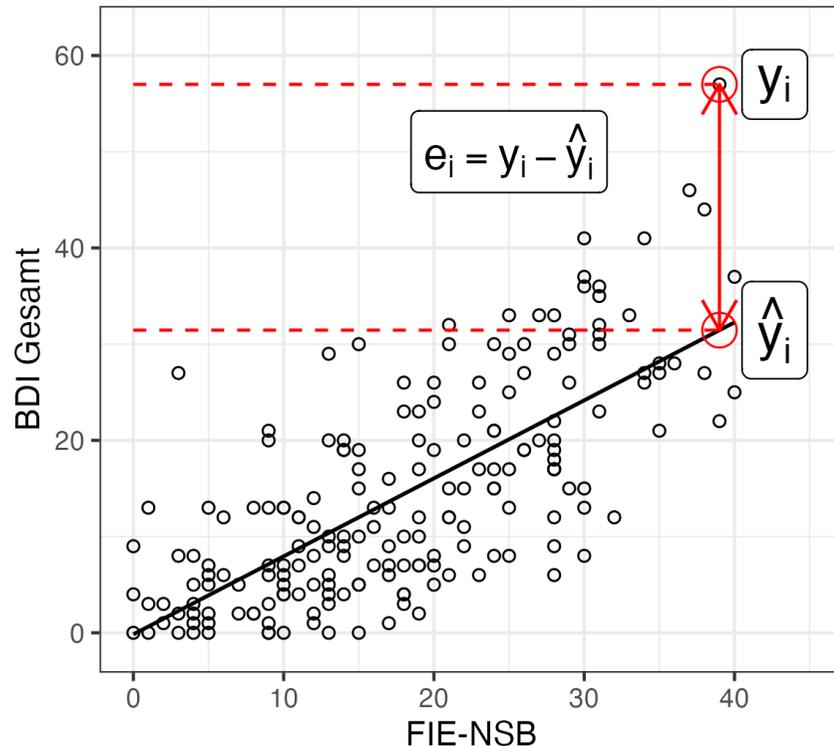
Der Vorhersagewert  $\hat{y}_i$  (Punkte im BDI) einer Person  $i$  mit einer negativen Selbstbewertung von  $x_i = 39$  Punkten.

Zur Vorhersage des AV-Werts einer Person  $i$  wird als konkreter **Vorhersagewert**  $\hat{y}_i$  im Allgemeinen der **Schätzwert** für den **unbekannten bedingten Erwartungswert**  $\mu_i$  verwendet.

Konkreter Vorhersagewert:

$$\hat{y}_i = \hat{\mu}_{i,Wert} = a + b \cdot x_i$$

# Definition: Residuum (Synonym: Residualwert)



- Unter dem **Residuum** versteht man auf der Ebene der **Zufallsvariablen** die Differenz zwischen der Zufallsvariable  $Y_i$  und der Zufallsvariable  $\hat{Y}_i$ .
- Auf Ebene der **Realisationen** versteht man unter dem Residuum die Differenz zwischen dem beobachteten Wert  $y_i$  und dem konkreten Vorhersagewert  $\hat{y}_i$  einer Person  $i$ .
  - Notation für die ZV:  $E_i = Y_i - \hat{Y}_i$
  - Notation für die Realisation der ZV:  
 $e_i = y_i - \hat{y}_i$
- Der Abstand zwischen der konkreten Depressionsausprägung  $y_i$  und dem Vorhersagewert  $\hat{y}_i$  einer Person  $i$  mit einer negativen Selbstbewertung von z.B.  $x_i = 39$  wird durch den konkreten Residualwert  $e_i$  quantifiziert.

- Die Schätzfunktionen für die unbekannte Fehlervarianz  $\sigma^2$  bzw. Standardabweichung  $\sigma$  lassen sich nach Einführung des Residuums einfacher darstellen:

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - (A + B \cdot X_i))^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^n E_i^2$$

$$\hat{\sigma} = S = \sqrt{\frac{\sum_{i=1}^n E_i^2}{n-2}}$$

- Der **Standardschätzfehler** ist folglich:

$$\hat{\sigma}_{\text{Wert}} = s = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}$$

## Schätzung von $\sigma^2$ bzw. $\sigma$ in R

- Der Schätzwert  $s$  kann im R-Output abgelesen werden.
- Inhaltlich ist  $s$  jedoch nur kaum interpretierbar.

```
> fit <-lm(data$bdi_ges~data$fie_nsb)
> summary(fit)
```

Call:

```
lm(formula = data$bdi_ges ~ data$fie_nsb)
```

Residuals:

Min	1Q	Median	3Q	Max
-16.539	-5.459	-1.214	5.053	25.544

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.15973	1.19345	-0.134	0.894
data\$fie_nsb	0.81067	0.05758	14.080	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.879 on 189 degrees of freedom

Multiple R-squared: 0.5119, Adjusted R-squared: 0.5093

F-statistic: 198.2 on 1 and 189 DF, p-value: < 2.2e-16

- Es lässt sich zeigen, dass die Schätzfunktion  $B$  unter der Voraussetzung, dass die Modellannahmen erfüllt sind, einer **Normalverteilung** folgt:

$$B \sim N \left( E(B) = \beta, \text{Var}(B) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

- Dabei ist  $\sigma^2$  wie üblich in der Regel unbekannt, kann aber durch die gerade gezeigte Schätzfunktion  $S^2$  geschätzt werden.
- Wenn wir dann

$$\hat{\text{Var}}(B) = \frac{S^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{1}{n-2} \frac{\sum_{i=1}^n E_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

im Vorgehen der z-Standardisierung verwenden, erhalten wir eine Zufallsvariable

$$Z^* = \frac{B - \beta}{\sqrt{\hat{\text{Var}}(B)}}$$

die einer t-Verteilung mit  $\nu = n - 2$  folgt.

- Die Vorgehensweise bei der Konstruktion eines *KI* für  $\beta$  zum Konfidenzniveau  $1 - \alpha$  mithilfe dieser Zufallsvariable ist analog zur bereits in Statistik I oft angewandten Vorgehensweise (Auflösen von  $P\left(t_{\frac{\alpha}{2}} \leq Z^* \leq t_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$ ).

- Zufälliges KI:

$$I(X_1, \dots, X_n, Y_1, \dots, Y_n) =$$

$$= \left[ B - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{n-2} \cdot \frac{\sum_{i=1}^n E_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}; B + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{n-2} \cdot \frac{\sum_{i=1}^n E_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

- Konkretes KI:

$$I(x_1, \dots, x_n, y_1, \dots, y_n) =$$

$$= \left[ b - t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{n-2} \cdot \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}; b + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{n-2} \cdot \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

## Berechnung des Konfidenzintervalls für $\beta$ in R

```
> confint(fit, level = 0.95)
              2.5 %    97.5 %
(Intercept) -2.5139279 2.1944603
data$fie_nsb 0.6970919 0.9242474
```

- Die plausiblen Werte für den unbekanntem Steigungsparameter  $\beta$  liegen zwischen 0.70 und 0.92.
- Man kann also davon ausgehen, dass sich die mittlere Depressionsschwere zwischen 0.70 und 0.92 Punkte erhöht, falls die negative Selbstbewertung um einen Punkt steigt.

- Die **Schätzung der Modellgleichung**

$$Y_i = \alpha + \beta \cdot X_i + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

drückt unsere „beste Vermutung“ darüber aus, wie der Zusammenhang zwischen der UV und den individuellen Ausprägungen der AV der Personen der Population aussehen könnte.

- Die **Schätzung der Regressionsgeraden**

$$E(Y_i|x_i) = \alpha + \beta \cdot x_i$$

drückt unsere „beste Vermutung“ darüber aus, wie der Zusammenhang zwischen der UV und der **durchschnittlichen** AV in der Population aussehen könnte.

- Die geschätzte Modellgleichung setzt sich zusammen aus der geschätzten Regressionsgeraden und der Fehlervariable sowie der geschätzten Verteilung der Fehlervariablen.
- In unserem Beispiel ist:

$$b = 0.81, a = -0.16, s = 7.88$$

Damit lautet die geschätzte Regressionsgerade

$$\hat{E}(Y_i|x_i) = -0.16 + 0.81 \cdot x_i$$

und die geschätzte Modellgleichung:

$$Y_i = a + b \cdot X_i + \epsilon_i = -0.16 + 0.81 \cdot X_i + \epsilon_i, \quad \text{mit } \epsilon_i \sim N(0, 7.88^2)$$

- Mithilfe der geschätzten Modellgleichung können wir Daten simulieren, die mit den beobachteten Daten aus der Stichprobe vergleichbar sind.

Bsp.: Simulation einer neuen Beobachtung mit  $x_i = 30$ :

```
> set.seed(1)
> -0.16 + 0.81 * 30 + rnorm(n = 1, mean = 0, sd = 7.88)
[1] 19.20354
```

- Bislang:
  - Modell der ELR
  - Parameterschätzung
- Jetzt:
  - Zentrierung der UV

- Durch eine Zentrierung der Prädiktorvariable wird der Bezugsrahmen verändert („das Koordinatensystem wird verschoben“).
- Es wird eine neue Variable berechnet, indem von den Werten der Prädiktorvariable  $X_i$  ein Wert, z.B. der Mittelwert  $\bar{X}$ , abgezogen wird. Diese neue zentrierte Variable ersetzt dann die alte unzentrierte Prädiktorvariable in der Regression.
- Dadurch kann  $\alpha$  bzw.  $a$  sinnvoll interpretiert werden, falls ein Wert von null in der Prädiktorvariable vor der Zentrierung nicht sinnvoll ist, bzw. außerhalb der beobachteten Werte liegt und das Modell in diesem Bereich extrapoliert (im Beispiel eben: negativer Wert für die Variable Depressionsschwere nicht sinnvoll).

- Zentrierung einer Zufallsvariable am Mittelwert:

$$X_{c,i} = X_i - \bar{X}$$

bzw. auf Ebene der Beobachtungen:

$$x_{c,i} = x_i - \bar{x}$$

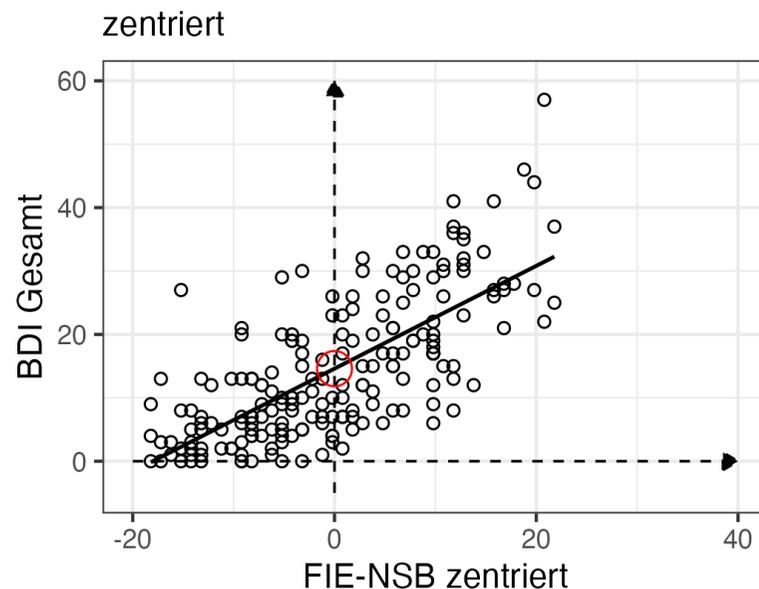
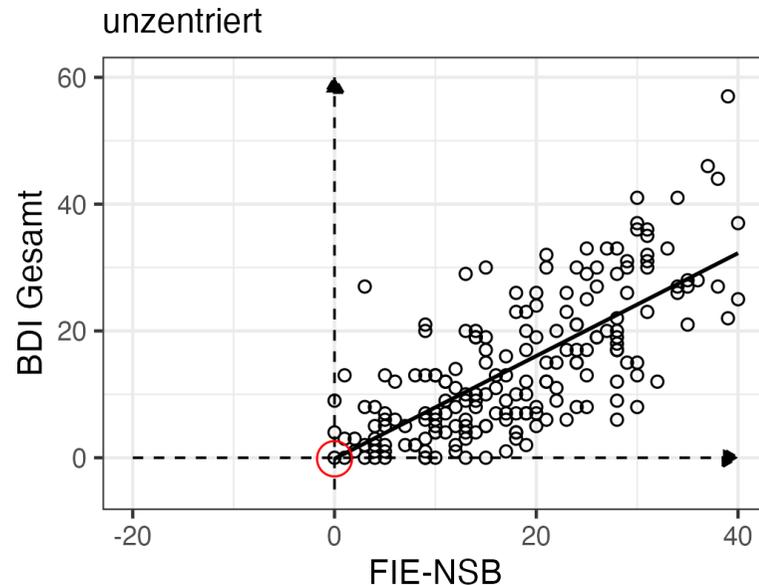
- **Bemerkung:** Statt der Zentrierung am Mittelwert ist es auch möglich, die Prädiktorvariable zu z-standardisieren (mehr dazu in Vorlesung 8), oder sogar auch an einem ganz anderen Wert zu zentrieren als dem Mittelwert.

- Der Schätzwert  $b$  wird durch die Zentrierung der Variable nicht beeinflusst, lediglich  $a$  ändert sich, da sich der Achsenabschnitt nach der Zentrierung auf eine durchschnittliche Ausprägung der unabhängigen Variable bezieht.
- Schätzung der Regression mit unzentriertem Prädiktor:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.15973    1.19345  -0.134    0.894
data$fie_nsb  0.81067    0.05758  14.080 <2e-16 ***
---
```

- Schätzung der Regression mit neuem, zentrierten Prädiktor:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.60209    0.57014  25.61 <2e-16 ***
data$fie_nsb_centered 0.81067    0.05758  14.08 <2e-16 ***
---
```



- Durch die Zentrierung der Prädiktorvariable wird der Bezugsrahmen verändert („das Koordinatensystem wird in Richtung der x-Achse verschoben“).
- Der Schätzwert  $a$  beträgt nach der Zentrierung ungefähr 14.60 (erwarteter BDI bei durchschnittlichem FIE-NSB), während er vor der Zentrierung ca. -0.16 betrug (erwarteter BDI bei einem FIE-NSB Wert von 0).
- Der Schätzwert  $b$  ist in beiden Fällen identisch.

- Bisläng:
  - Modell der ELR
  - Parameterschätzung
  - Zentrierung der UV
- Jetzt:
  - Hypothesentests für  $\beta$

- Im Rahmen des Modells der ELR kann z.B. überprüft werden, ob überhaupt ein linearer Zusammenhang zwischen AV und UV besteht.
- Auf unser Beispiel bezogen: Es soll überprüft werden, ob die Höhe der Depression von der Stärke der negativen Selbstbewertung linear abhängt.
- In der Nullhypothese wird in diesem Fall die Aussage formuliert, dass der lineare Zusammenhang zwischen der Höhe der Depression und der negativen Selbstbewertung der Personen gleich null ist.
- Die statistischen Hypothesen für diesen Fall lauten:
  - $H_0: \beta = 0$
  - $H_1: \beta \neq 0$
- Allgemeiner Fall:
  - $H_0: \beta = \beta_0$
  - $H_1: \beta \neq \beta_0$
- Bemerkung: Besteht eine Vermutung über die Richtung des Effekts (z.B. negative Selbstbewertung führt zu einer Zunahme der Depressionsausprägung), können die Hypothesen mit Hilfe weiterer Verfahren auch gerichtet getestet werden.

- Da

$$\hat{Z}^* = \frac{B - \beta}{\sqrt{\widehat{Var}(B)}} \sim t(n - 2)$$

für das tatsächliche  $\beta$  in der Population gilt (siehe Folie 36), folgt unter der  $H_0: \beta = \beta_0$  also falls  $\beta_0$  das tatsächliche  $\beta$  ist, dass

$$T = \frac{B - \beta_0}{\sqrt{\widehat{Var}(B)}} \sim t(n - 2)$$

- Auf der Basis der Realisation dieser Teststatistik  $T$  und der Verteilung  $t(n - 2)$  können wir unsere Testentscheidung treffen.

# Ungerichteter Hypothesentest für $\beta$ in R

```
> fit <- lm(data$bdi_ges ~ data$fie_nsb)
> summary(fit)
```

Call:

```
lm(formula = data$bdi_ges ~ data$fie_nsb)
```

Residuals:

Min	1Q	Median	3Q	Max
-16.539	-5.459	-1.214	5.053	25.544

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.15973	1.19345	-0.134	0.894
data\$fie_nsb	0.81067	0.05758	14.080	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.879 on 189 degrees of freedom

Multiple R-squared: 0.5119, Adjusted R-squared: 0.5093

F-statistic: 198.2 on 1 and 189 DF, p-value: < 2.2e-16

Wir gehen davon aus, dass es einen linearen Zusammenhang zwischen der Depressionsausprägung und der negativen Selbstbewertung gibt.

- Bisläng:
  - Modell der ELR
  - Parameterschätzung
  - Hypothesentest für  $\beta$
- Jetzt:
  - Vorhersage und Konfidenzintervall für  $y_i$

## Vorhersage für $y_i$ in der Praxis

- Wie wir bei der Definition des Vorhersagewerts  $\hat{y}_i$  bereits gesehen haben, ist es mit regressionsanalytischen Modellen wie der ELR möglich, konkrete Vorhersagen für die Werte auf der AV bei gegebenen Werten für die UV zu berechnen.
- Diese Eigenschaft ist in der Praxis sehr nützlich und trägt mit zur großen Attraktivität von regressionsanalytischen Verfahren (nicht nur in der Psychologie) bei.

- Wiederholung: Mithilfe der Schätzwerte  $a$  und  $b$  lässt sich für jede Person  $i$  mit einer bestimmten Ausprägung  $x_i$  der Vorhersagewert  $\hat{y}_i$  für ihren Wert  $y_i$  auf der AV bestimmen.
- Bsp.: Berechnung des Vorhersagewerts  $\hat{y}_i$  für eine Person  $i$  mit einer negativen Selbstbeurteilung von  $x_i = 25$  ( $b = 0.81, a = -0.16$ ).

$$\hat{y}_i = a + b \cdot x_i = -0.16 + 0.81 \cdot 25 = 20.09$$

- Interpretation: Man vermutet, dass die mittlere Depressionsschwere von Personen mit einer negativen Selbstbewertung von  $x = 25$  Punkten bei ca. 20 Punkten liegt.
- Bemerkung: Die Person, für die man  $y_i$  vorhersagen will, muss nicht unbedingt in der Stichprobe enthalten sein, mithilfe der man die Parameter des Modells geschätzt hat. Es können auch Werte für  $x_i$  eingesetzt werden, die in der Stichprobe gar nicht beobachtet wurden. Man sollte jedoch sehr vorsichtig damit sein, Vorhersagen für Werte von  $x_i$  zu berechnen, die außerhalb des in der Stichprobe beobachteten Wertebereichs liegen (z.B. extrem hohe oder niedrige Werte für  $x_i$ ), da man nicht überprüfen kann, ob der vom Modell angenommene lineare Zusammenhang auch für nicht beobachtete Wertebereiche von  $x_i$  plausibel ist.

## Konfidenzintervall für $y_i$

- Für  $y_i$  lässt sich bei bekanntem  $x_i$  zudem ein Konfidenzintervall konstruieren.
- Dieses Konfidenzintervall wird in der Literatur oft als **Vorhersageintervall** bezeichnet.
- Die formale Darstellung des Konfidenzintervalls ist komplex und nicht intuitiv verständlich. In seine Breite gehen mehrere Fehlerquellen ein: Die Ungenauigkeit der Schätzung von  $\alpha$  und  $\beta$  sowie die Varianz von  $\varepsilon_i$ .

## Vorhersage und Konfidenzintervall für $y_i$ in R

```
> fit <-lm(bdi_ges~fie_nsb,data)
> summary(fit)

> data_neu<-data.frame(fie_nsb=28)
> data_neu
  fie_nsb
1      28
> predict(fit,data_neu,interval="prediction",level=0.95)
      fit      lwr      upr
1 22.53902 6.91577 38.16226
```

- In unserem Beispiel gehen wir von Personen mit einer negativen Selbstbewertung von  $x_i = 28$  Punkten aus.
- Bei einer negativen Selbstbewertung von  $x_i = 28$  Punkten liegen plausible Werte für die beobachtete Depressionsschwere im BDI zwischen 6.92 und 38.16 Punkten.

(Der konkrete Vorhersagewert beträgt hier 22.54 Punkte)

- Bislang:
  - Modell der ELR
  - Parameterschätzung
  - Hypothesentest für  $\beta$
  - Vorhersage und Konfidenzintervall für  $y_i$
- Jetzt:
  - Übung an einem zweiten Beispiel

## Beispiel 2

- Im Folgenden werden die R-Outputs im Rahmen einer ELR für einen zweiten Prädiktor „fie\_abk“ gezeigt und interpretiert.
- UV („fie\_abk“): Einschätzung der Abhängigkeitskognitionen (Skala des Fragebogens zur Erfassung irrationaler Einstellungen)  
Beispielitem: „Ich brauche es, dass Leute mich mögen.“
- Inhaltliche Alternativhypothese: Die Höhe der mittleren Depression hängt linear von der Stärke der Abhängigkeitskognitionen ab.

```
> fit.2 <- lm(bdi_ges ~ fie_abk, data=dat)
> summary(fit.2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.51937	2.26435	1.554	0.122
fie_abk	0.39860	0.09126	4.368	2.11e-05 ***

- Wie lauten die statistischen Hypothesen bezogen auf  $\beta$ ?
- Testentscheidung bezogen auf  $\beta$  bei einem Signifikanzniveau  $\alpha = 0.005$ ?
- In welchem Wert haben sich  $A$  bzw.  $B$  realisiert?
- Wie werden  $a$  und  $b$  inhaltlich interpretiert?
- Wie groß ist der Schätzwert für den Standardfehler der Schätzfunktion  $B$ ?
- Wie lautet die geschätzte Regressionsgerade?

## Konfidenzintervalle für $y_i$

```
> dat.neu <- data.frame(fie_abk=20:30)
> dat.neu
  fie_abk
1      20
2      21
3      22
4      23
5      24
6      25
7      26
8      27
9      28
10     29
11     30
> predict(fit.2,dat.neu,interval="prediction")
      fit      lwr      upr
1 11.49142 -6.982961 29.96580
2 11.89002 -6.578858 30.35890
3 12.28862 -6.176509 30.75375
4 12.68722 -5.775916 31.15037
5 13.08583 -5.377078 31.54873
6 13.48443 -4.979997 31.94886
7 13.88303 -4.584671 32.35073
8 14.28163 -4.191100 32.75437
9 14.68024 -3.799282 33.15975
10 15.07884 -3.409215 33.56689
11 15.47744 -3.020898 33.97578
```

Interpretation für Personen, deren Abhängigkeitskognition bei  $x = 21$  liegt?

- Output „fie\_nsb“:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.68990	1.09369	1.545	0.124
fie_nsb	0.67990	0.05734	11.858	<2e-16 ***
---				

- Output „fie\_abk“:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.51937	2.26435	1.554	0.122
fie_abk	0.39860	0.09126	4.368	2.11e-05 ***

- Kann man die Schätzwerte  $b_{nsb}$  und  $b_{abk}$  miteinander vergleichen?  
Kann man Aussagen darüber treffen, welche von beiden UVs stärker mit der Depressionsschwere zusammenhängt?
- Diese und weitere Fragen: **Multiple** lineare Regression (siehe Vorlesung 7)

- Mit der Regressionsanalyse lassen sich (lineare) Zusammenhänge zwischen einer oder mehrerer stetiger oder diskreter UVs mit einer stetigen oder diskreten AV modellieren.
- Die Geradengleichung der Regression gibt an, wie die UV mit den bedingten Erwartungswerten der AV zusammenhängt.
- Der Parameter  $\beta$  gibt dabei die Stärke des Zusammenhangs an: Ist  $\beta = 0$  liegt kein Zusammenhang vor. Ist  $\beta > 0$  liegt ein positiver (gleichgerichteter), für  $\beta < 0$  ein negativer (gegengerichteter) Zusammenhang vor.
- Die Interpretation des Intercepts  $\alpha$  hängt davon ab, ob die Ausprägung  $x_i = 0$  der UV inhaltlich sinnvoll interpretierbar ist und sie verändert sich, wenn statt  $x_i$  eine (z.B. am Mittelwert) zentrierte Umrechnung von  $x_i$  in die Regression aufgenommen wird.
- Mit Hypothesentests für  $\beta$  lassen sich inhaltliche Hypothesen über Zusammenhänge zwischen zwei (oder mehreren) Variablen testen.
- Der Vorhersagewert  $\hat{y}_i$  gibt unsere beste Vermutung an, welche Ausprägung wir im Durchschnitt auf der AV für eine Person erwarten, die eine bestimmte Ausprägung der UV  $x_i$  aufweist.