6. Vorlesung Statistik II
Einführung in die Regressionsanalyse

Vorbemerkung

- Bislang:
  - Allen varianzanalytischen Modellen ist gemeinsam, dass die **unabhängigen Variablen diskret** sind und die **abhängige Variable stetig**.

- Jetzt:
  - Kennenlernen von regressionsanalytischen Modellen.
  - In Regressionsmodellen können sowohl die **unabhängigen Variablen** als auch die **abhängige Variable** entweder **stetig** oder **diskret** sein.

Begriffe

- 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.

Beispiel

- **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)

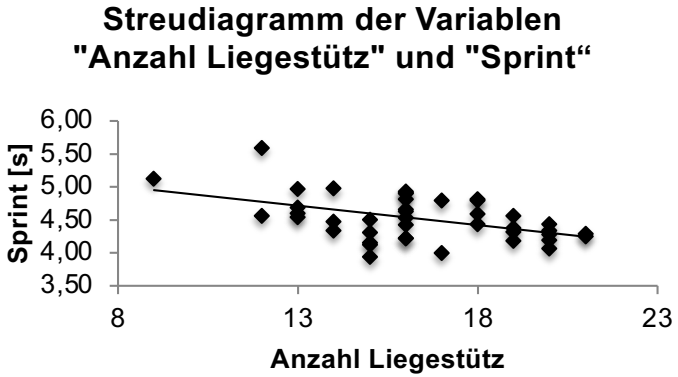
  Variablenname: „bdi-ges“
- UV: Einschätzung der negativen Selbstbewertung (Skala des Fragebogens zur Erfassung irrationaler Einstellungen)

  Variablenname: „fie_nsb“
- Beispielitem: „Ich denke oft, ich bin ein Versager.“

Ausgangslage

- 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** dar.

„Idee“ der Regressionsanalyse



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 welcher 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 zusammen hängt. Dies werden wir in der nächsten Vorlesung zu Effektstärken im Rahmen der ELR noch einmal aufgreifen.

Bedingter Erwartungswert (I)

- Beispiel aus VL 5: Mittlere Depressionsschwere bei jungen Erwachsenen, Erwachsenen und alten Erwachsenen
- Im einfaktoriellen varianzanalytischen Modell entspricht der Erwartungswert μ_{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 μ_{jung} auch als sogenannten bedingten Erwartungswert ausdrücken:

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

- 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: $E(Y_i|X_i = x_i)$ 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 μ_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 bei einer einfachen linearen Regression

- Statistisches Modell der einfachen linearen Regression:

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

- 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**.
- α, β und σ^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 ε_i verursacht wird.
  - ... ein **linearer Zusammenhang** zwischen dem Prädiktor und dem Kriterium besteht.

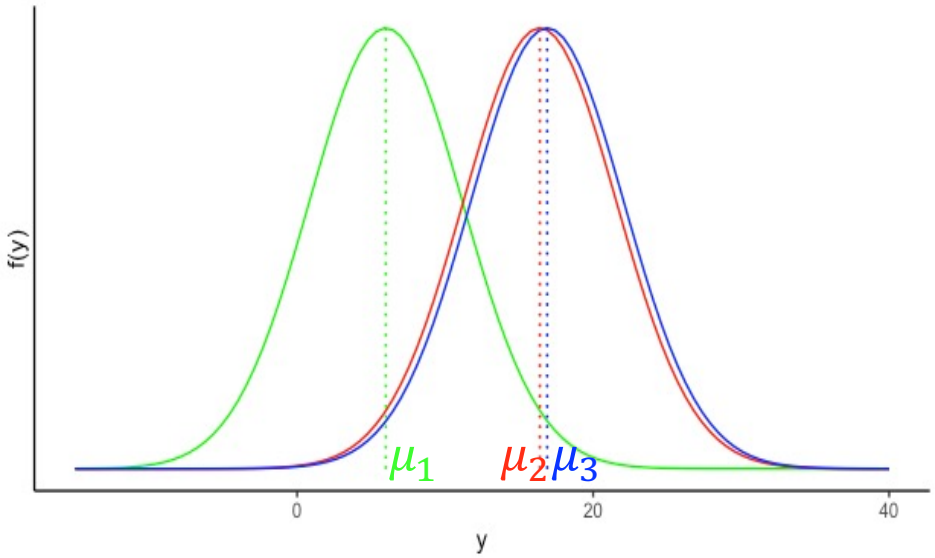
Definition: Fehlervariable ε_i

- Die Fehlervariable ε_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 ε_i ist damit selbst auch eine Zufallsvariable.

Zur Erinnerung: Veranschaulichung des ANOVA-Modells

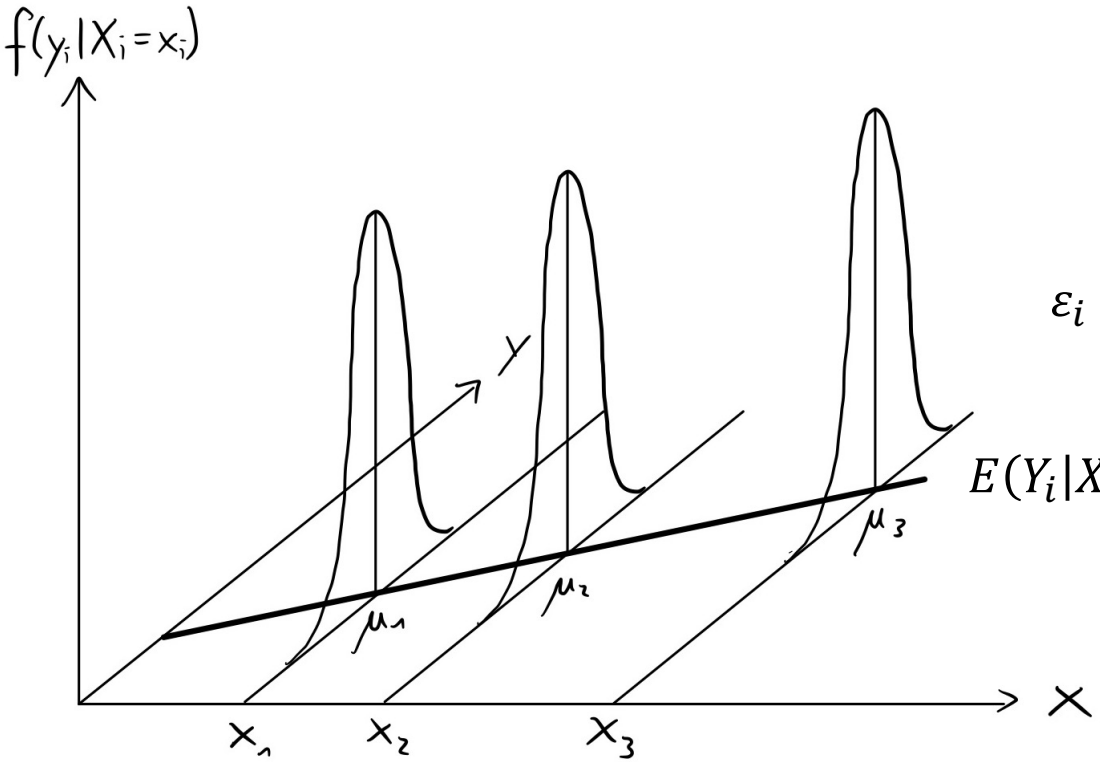


- Bsp.: Mittlere Depression bei jungen Erwachsenen ($j = 1$), Erwachsenen ($j = 2$) und alten Erwachsenen ($j = 3$)
- Erste Darstellungsform:

$$Y_{ij} = \mu_j + \varepsilon_{ij},$$
$$\text{mit } j = 1, 2, 3 \text{ und } i = 1, \dots, n_j$$

- Veranschaulichung der Verteilungen der Depression in den 3 Altersgruppen:
- Annahmen des varianzanalytischen Modells: Eine Normalverteilung pro Population
- Die Erwartungswerte der Normalverteilungen dürfen sich unterscheiden.
- Die Varianzen der Normalverteilungen müssen gleich sein.

Grafische Veranschaulichung des statistischen Modells

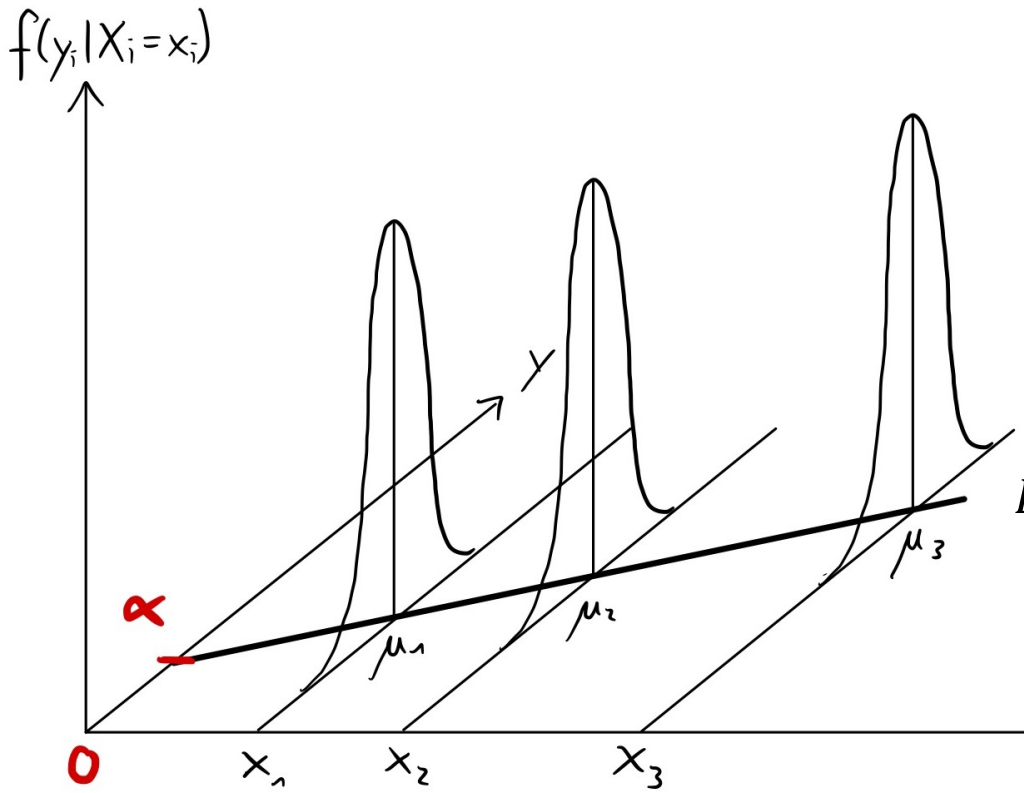


$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

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

- Der Modellparameter α wird auch als **Intercept** bzw. **Achsenabschnitt** bezeichnet.
- Der Modellparameter β wird auch als **Slope** bzw. **Steigungsparameter** bezeichnet.
- Der Modellparameter σ^2 wird auch als **Fehlervarianz** bezeichnet.

Modellparameter α

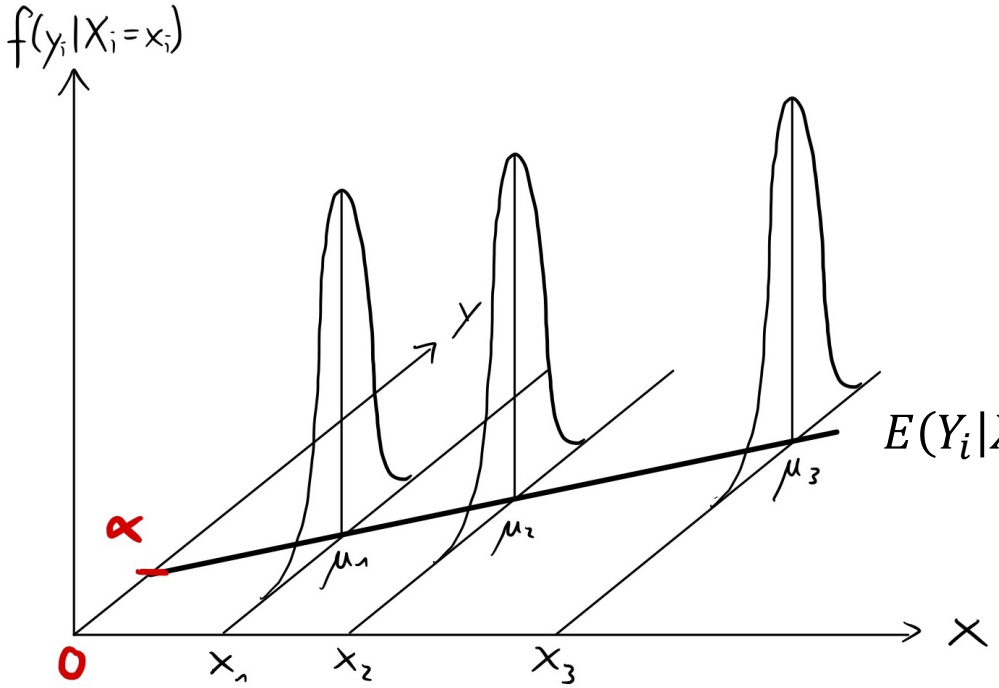


$$E(Y_i|X_i = x_i) = \mu_i = \alpha + \beta \cdot x_i$$

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

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

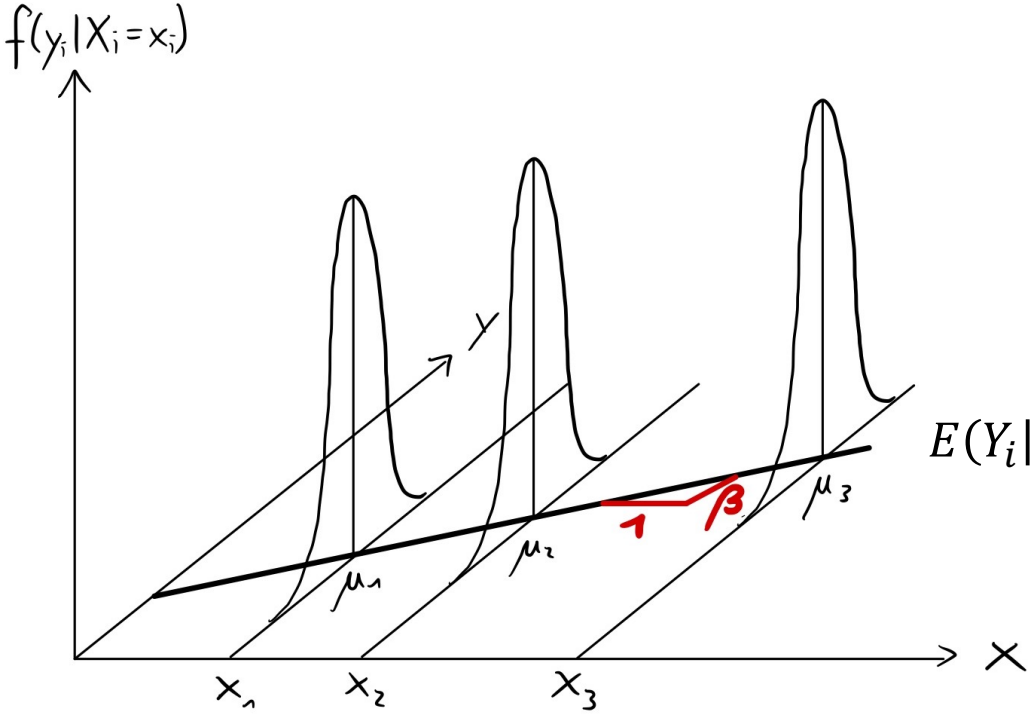
Interpretation des Modellparameter α bezogen auf das Beispiel



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

- Bezogen auf unser Beispiel repräsentiert α die mittlere Depressionsausprägung im Falle, dass Personen keine negative Selbstbewertung aufweisen.
- Ob der Modellparameter α sinnvoll interpretiert werden kann, hängt davon ab, ob $x_i = 0$ einen inhaltlich plausiblen Wert darstellt.
- Nach einer Zentrierung des Prädiktors kann α als die mittlere Depressionsausprägung einer Person mit durchschnittlicher negativer Selbstbewertung interpretiert werden (siehe auch Folie 27).

Modellparameter $\boldsymbol{\beta}$

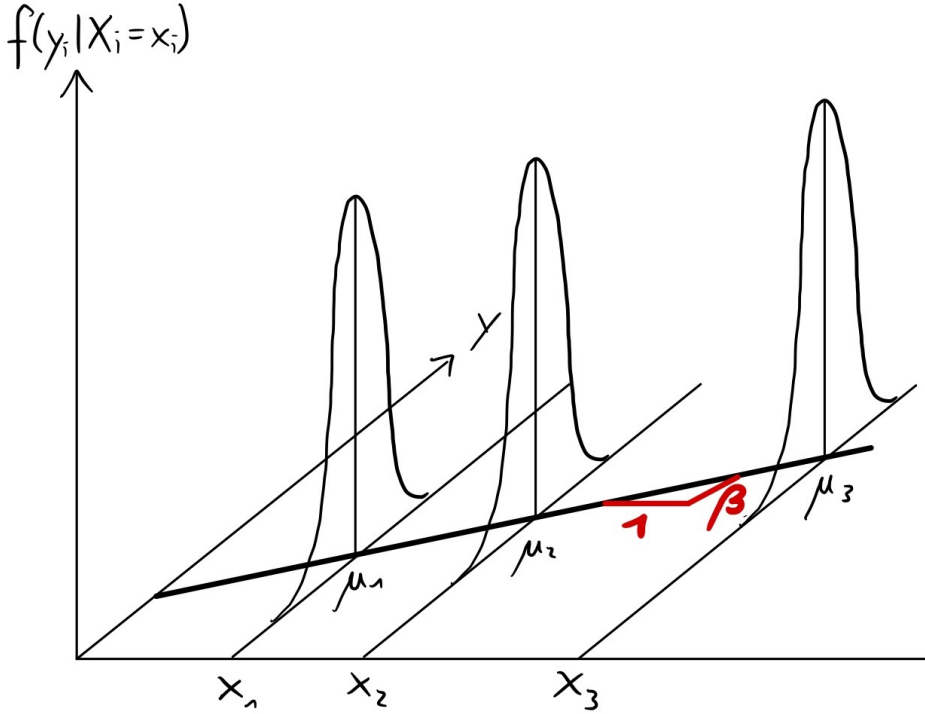


$$E(Y_i|X_i = x_i) = \mu_i = \alpha + \beta \cdot x_i$$

- Der Steigungsparameter β 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 α. Die Zufallsvariable Y_i setzt sich lediglich aus der Konstanten α und der Fehlervariable ε_i zusammen:

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

Interpretation des Modellparameters $\boldsymbol{\beta}$ bezogen auf das Beispiel



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

- Bezogen auf unser Beispiel gibt der Steigungsparameter β an, wie stark sich im Mittel die Depressionsausprägung erhöht, wenn die negative Selbstbewertung um eine Einheit zunimmt.
- Beispiel: Wenn sich die negative Selbstbewertung um **eine Einheit** erhöht, dann erhöht sich **im Mittel** die Depressionsschwere um z.B. 1.5 Einheiten.

ELR: Annahmen

Im Modell der ELR werden die folgenden Annahmen getroffen:

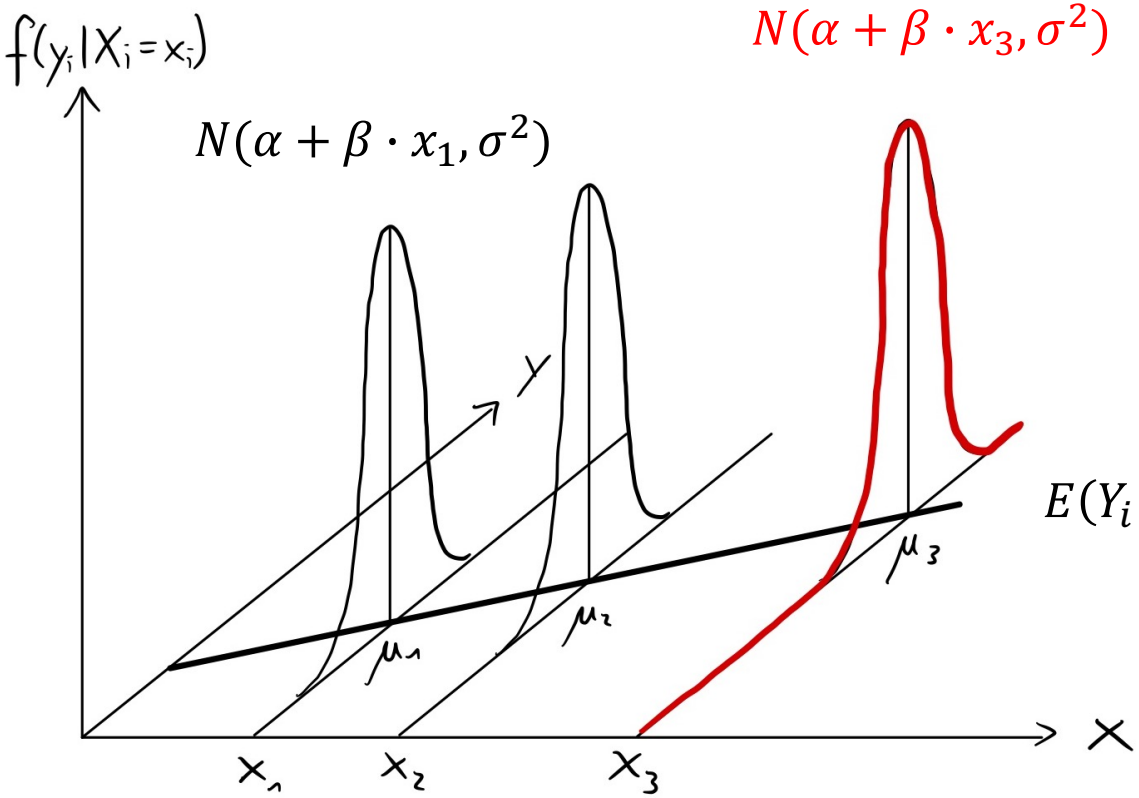
1. Die **UV** und die **AV** hängen **linear** zusammen:

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

2. Alle **Fehler** ε_i sind **unabhängig voneinander** und folgen einer **Normalverteilung** mit **Erwartungswert null** und **konstanter Varianz** σ^2:

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

Veranschaulichung der 2. Modellannahme



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

- Aus $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ 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 $\alpha + \beta \cdot x_i$ und konstanter Varianz σ^2.
- Die Modellannahmen werden wir in der nächsten Vorlesung weiter diskutieren.

ELR: Parameter

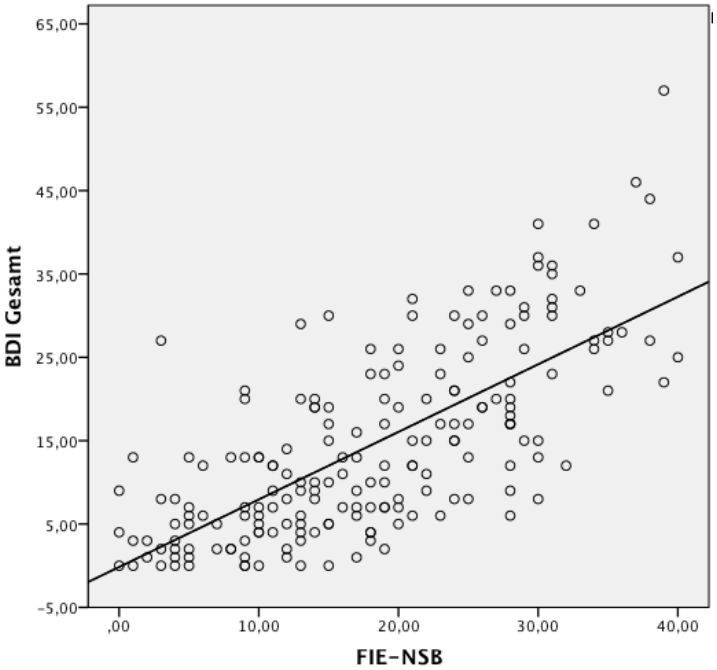
- Das Modell der ELR enthält die folgenden Parameter:
  - α
  - β
  - σ^2
- Alle diese Parameter können geschätzt werden und für alle diese Parameter können statistische Hypothesentests konstruiert werden. Dies gilt auch für Kombinationen der Parameter.
- Je nach konkreter Fragestellung muss entschieden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

Zwischengliederung

- 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.

Notation und Bezeichnungen: Parameter, Schätzfunktionen und Schätzwerte

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

Modellparameter (unbekannt)	**Schätzfunktion**	**Schätzwert**
α	$\hat{\alpha} = A$	a
β	$\hat{\beta} = B$	b
σ^2	$\hat{\sigma}^2 = S^2$	s^2

Schätzfunktionen für α, β und σ^2

- Es lässt sich zeigen, dass erwartungstreue, konsistente und effiziente Schätzfunktionen für die unbekannten Parameter α, β und σ^2 existieren.
- Die Schätzfunktionen für die unbekannten Parameter sind:

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

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

Schätzfunktionen für σ^2 und σ

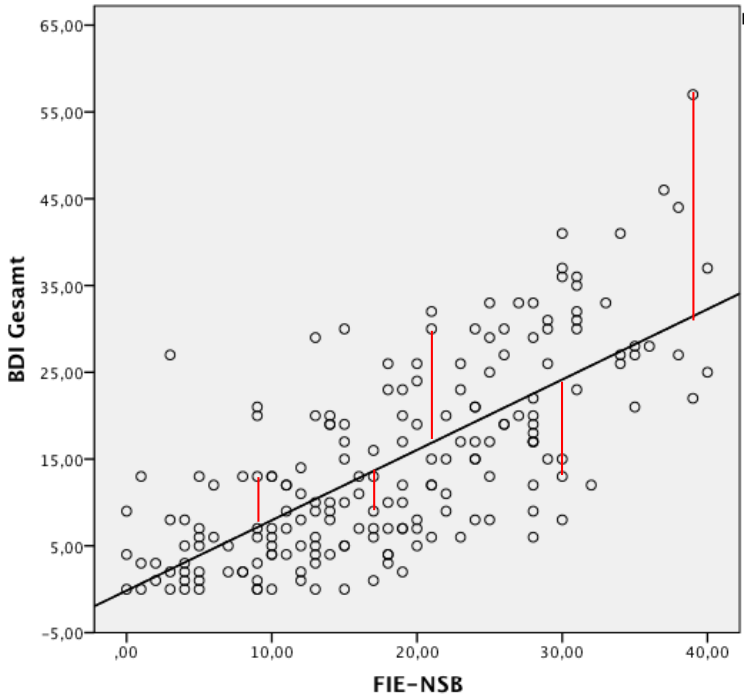
- In der Praxis betrachtet man neben der Fehlervarianz σ^2häufig auch die Standardabweichung der Fehler: $\sigma = \sqrt{\sigma^2}$.
- Es lässt sich zeigen, dass erwartungstreue, konsistente und effiziente Schätzfunktionen für die unbekannten Parameter σ^2 bzw. σ existieren.
- Die Schätzfunktionen für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

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

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

- Die Realisation s als konkreter Schätzwert für σ wird in der Literatur häufig auch als **Standardschätzfehler** bezeichnet.

Exkurs: Bestimmung der Schätzfunktionen A und B (I)



- 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.

Exkurs: Bestimmung der Schätzfunktionen $\boldsymbol{A}$ und $\boldsymbol{B}$ (II)

- 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:

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

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

Schätzung der Modellparameter α und β in R


```
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 unbekannten Steigungsparameter β 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.

Zentrierung des Prädiktors am Mittelwert

- Durch eine Zentrierung der Prädiktorvariable wird der Bezugsrahmen verändert („das Koordinatensystem wird verschoben“).
- Dadurch kann α 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.

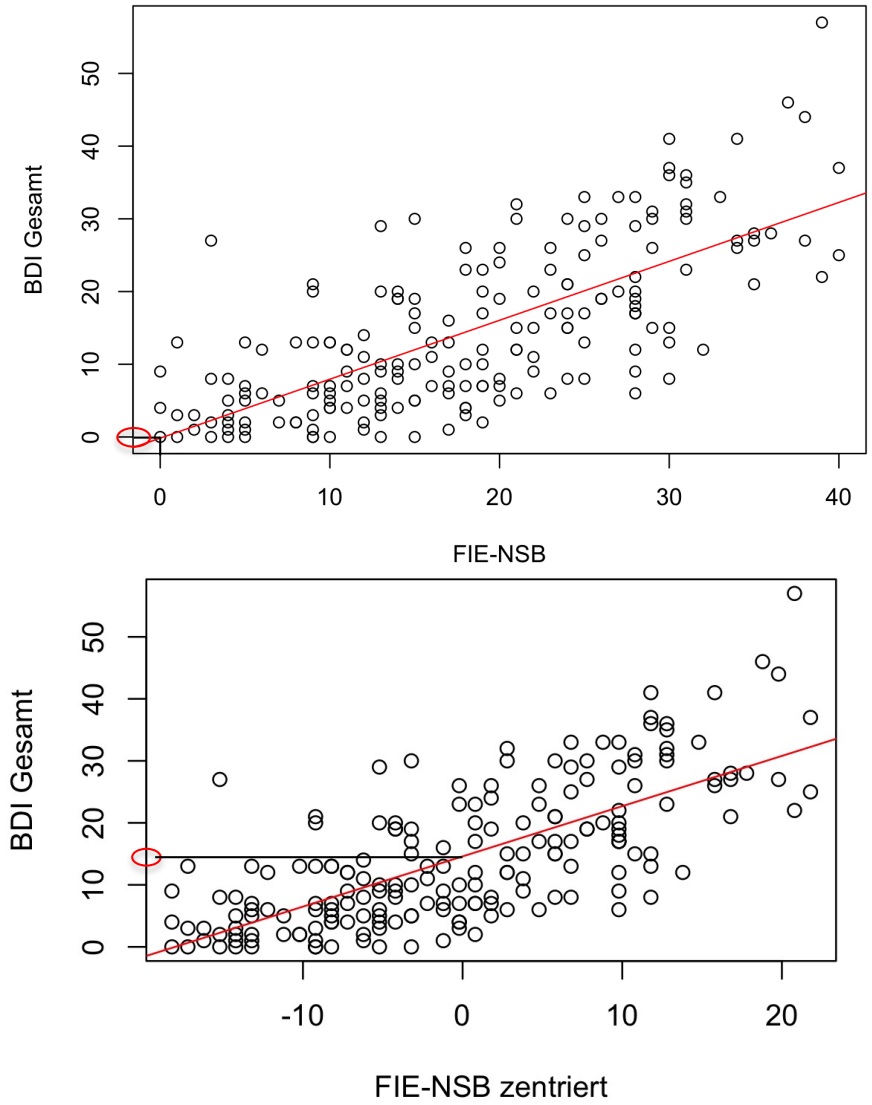
Zentrierung des Prädiktors am Beispiel

- Durch die Zentrierung der Prädiktorvariable wird der Bezugsrahmen verändert („das Koordinatensystem wird verschoben“).
- 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:

```
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 ***
---
```

```
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 ***
---
```

Zentrierung des Prädiktors am Beispiel



- Durch die Zentrierung der Prädiktorvariable wird der Bezugsrahmen verändert („das Koordinatensystem wird 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.

Schätzung von σ^2 bzw. σ in R

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

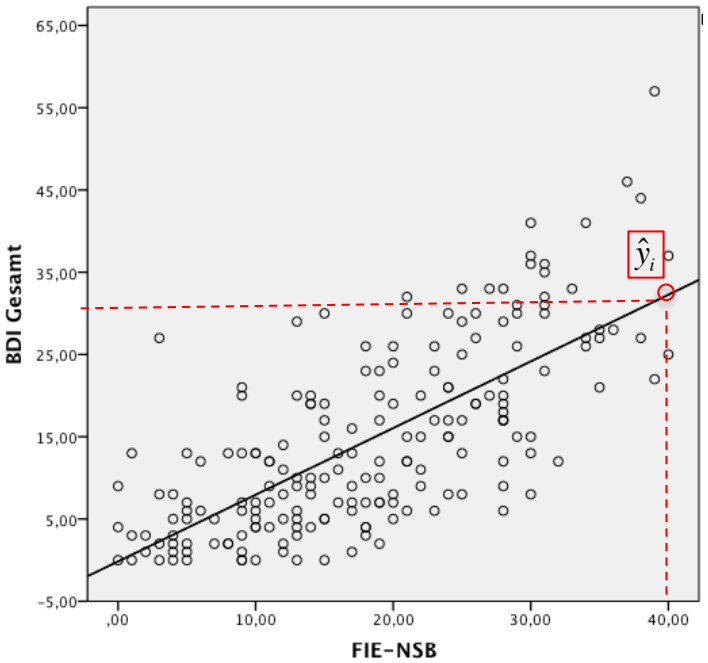
```
> fit <-lm(data$bdi_ges~data$fie_nsb)
> summary(fit)

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
```

Weitere Begriffe in der Regressionsanalyse

Bevor wir uns mit der Schätzung weiterer interessanter Größen beschäftigen, müssen wir einige neue Begriffe einführen:

- Vorhersagewert (Synonym: gefitteter Wert)
- Residuum (Synonym: Residualwert)

Definition: Vorhersagewert (Synonym: gefitteter Wert)



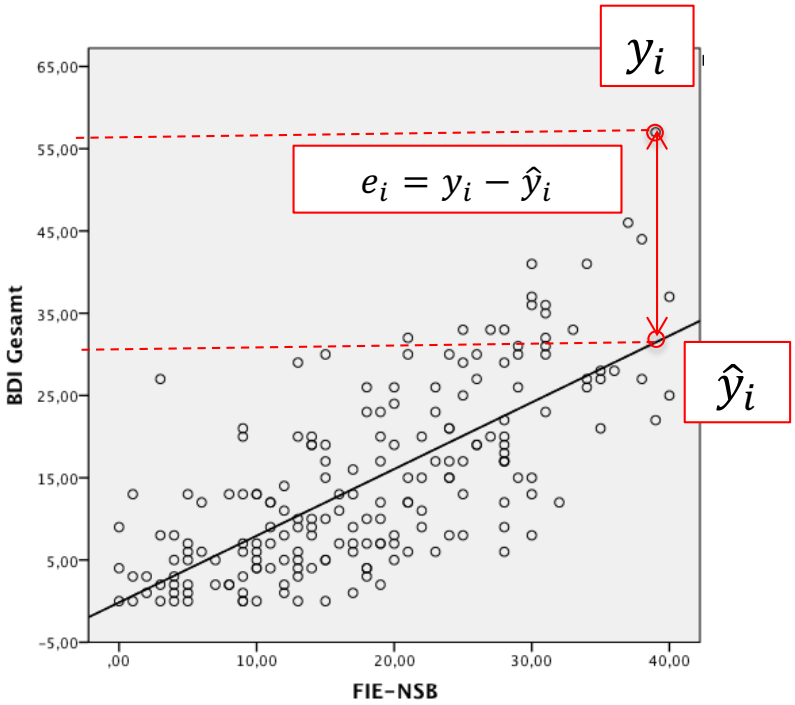
Der Vorhersagewert $\hat{y}_i$ (Punkte im BDI) einer Person *i* mit einer negativen Selbstbewertung von $x_i = 40$ Punkten.

Zur Vorhersage des AV-Werts einer Person *i* wird als konkreter **Vorhersagewert** $\hat{y}_i$ im Allgemeinen der **Schätzwert** für den **unbekannten bedingten Erwartungswert** μ_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 $x_i = 40$ wird durch den konkreten Residualwert e_i quantifiziert.

Darstellung der Schätzfunktion für σ^2 mit Hilfe des Residuums

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

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

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

- Der **Standardschätzfehler** ist folglich:

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

Konfidenzintervall für β (I)

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

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

- Dabei ist σ^2 wie üblich in der Regel unbekannt, kann aber durch die gerade gezeigte Schätzfunktion S^2 geschätzt werden.

- Wenn wir dann

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

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

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

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

Konfidenzintervall für β (II)

- Die Vorgehensweise bei der Konstruktion eines *KI* für β 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 β 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 unbekannten Steigungsparameter β 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.

Schätzung der Modellgleichung und der Regressionsgeraden (I)

- Mit Hilfe der gerade geschätzten Größen können wir jetzt auch noch weitere Schätzungen angeben:
  - Schätzung der Modellgleichung: $Y_i = \alpha + \beta \cdot X_i + \epsilon_i$
  - Schätzung der Regressionsgerade: $E(Y_i|x_i) = \alpha + \beta \cdot x_i$

- Die geschätzte Modellgleichung bzw. Regressionsgerade drückt unsere „beste Vermutung“ aus, wie der Zusammenhang zwischen der UV und der AV in der Population aussehen könnte.

Schätzung der Modellgleichung und der Regressionsgeraden (II)

- 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:

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

- 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
```

Zwischengliederung

- Bislang:
  - Modell der ELR
  - Parameterschätzung
- Jetzt:
  - Hypothesentests für β

Hypothesentests für $\boldsymbol{\beta}$

- 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 auch gerichtet formuliert werden.

Teststatistik

- Da

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

für das tatsächliche β in der Population gilt (siehe Folie 36), folgt unter der $H_0: \beta = \beta_0$ also falls β_0 das tatsächliche β 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.

Hypothesentest für β 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.

Zwischengliederung

- Bislang:
  - Modell der ELR
  - Parameterschätzung
  - Hypothesentest für β
- Jetzt:
  - Vorhersage und Konfidenzintervall für y_i

Vorhersage für y_i in der Praxis

- Wie wir bei der Definition des Vorhersagewerts $\widehat{y_i}$ bereits gesehen haben, ist es mit regressionsanalytischen Modellen wie der ELR sehr leicht 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.

Vorhersage für y_i

- 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 einer Person 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 α und β sowie die Varianz von ε_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 Selbstwertung 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)

Zwischengliederung

- Bislang:
  - Modell der ELR
  - Parameterschätzung
  - Hypothesentest für β
  - 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.

Ergebnisse der Hypothesentestung

```
> 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 β?
- Testentscheidung bezogen auf β 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?

Beispiel 1 und Beispiel 2 im Vergleich

- 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)