CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

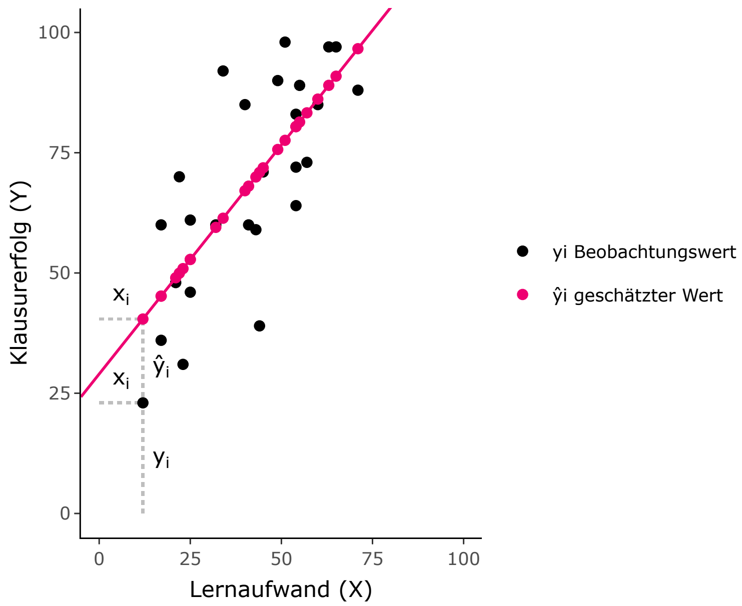
Statistik II

Einheit 4: Einfache lineare Regression (2)

15.05.2025 | Prof. Dr. Stephan Goerigk

Einfache lineare Regression

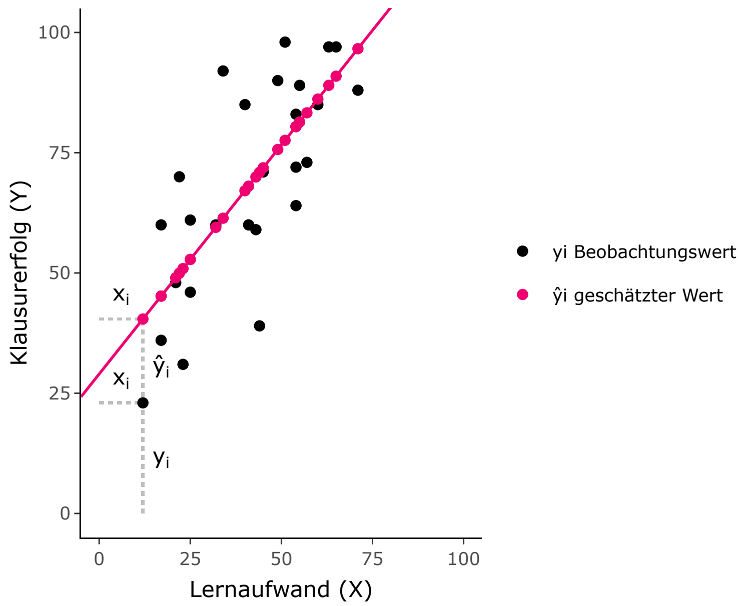
Vorhergesagte Werte (predicted values)

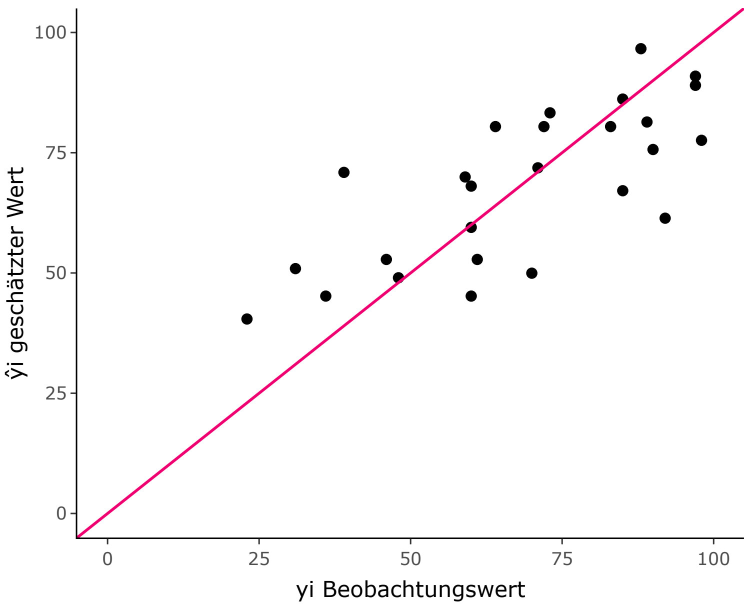


- Der Mittelwert der vorhergesagten Werten entspricht dem Mittelwert der empirischen Verteilung
- Die Regression soll die Abweichungen der tatsächlichen von den vorhergesagten Werten so gering wie möglich halten
- Dabei muss sie jedoch repräsentativ für die ganze Verteilung bleiben
- Der Mittelwert der vorhergesagten Werte darf sich folglich nicht verändern

Einfache lineare Regression

Vorhergesagte Werte (predicted values)

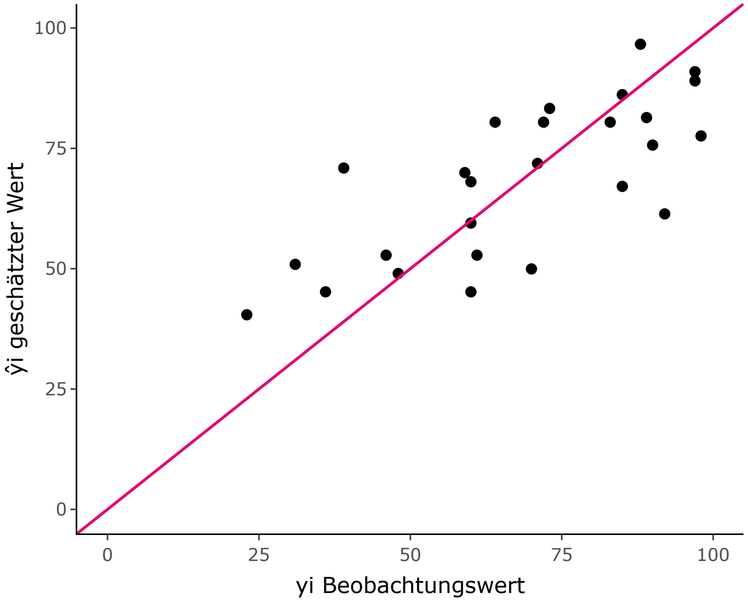




Einfache lineare Regression

Vorhergesagte Werte (predicted values)

- Je mehr die vorhergesagten Werte den tatsächlich beobachteten entsprechen, desto besser ist die Schätzung des Modells
- Häufig werden Modelle zusätzlich an neuen Daten **kreuzvalidiert**, um zu prüfen, wie sehr vorhergesagte Werte mit "neuen Daten" übereinstimmen, die nicht in der ursprünglichen Stichprobe enthalten waren.



Einfache lineare Regression

Dichotom nominalskalierte Prädiktoren (UVs)

- Oft nutzen Psycholog:innen die einfache lineare Regression, um eine intervallskalierte AV (Y) mit einer intervallskalierten UV (X) vorherzusagen
- Es kann jedoch auch eine dichotom nominalskalierte Variable als UV verwendet werden

Mathematische Integration:

- nominalskalierte UV lässt sich mathematisch integrieren, indem die beiden Kategorien mit 0 und 1 kodiert werden
- Man spricht dann von einer **Dummy-Kodierung**

UV: Gruppe (nominal dichotom)	UV: Gruppe (dummy-kodiert)	AV: Sorgen (skaliert von 1-12)
Gesund	0	3.44
Gesund	0	3.77
Gesund	0	5.56
Gesund	0	4.07
Gesund	0	4.13
Gesund	0	5.72
Gesund	0	4.46
Gesund	0	2.73
GAD	1	8.31
GAD	1	8.55
GAD	1	10.22
GAD	1	9.36
GAD	1	9.40
GAD	1	9.11
GAD	1	8.44
GAD	1	10.79

Einfache lineare Regression

Dichotom nominalskalierte Prädiktoren (UVs)

$$\hat{y}_i = a + b \cdot x_i + \epsilon_i$$

a : Y-Achsenabschnitt
b : Steigungsparameter

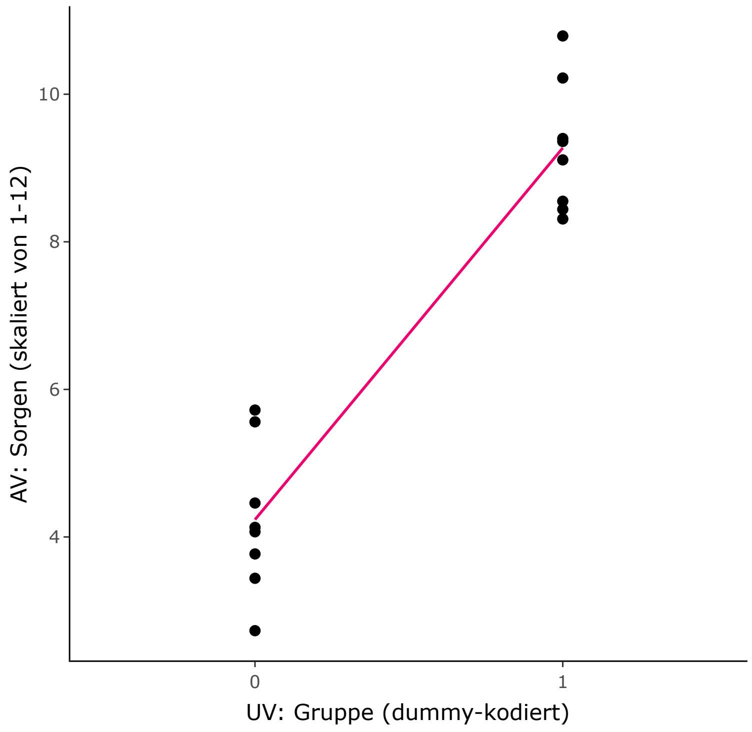
Interpretation:

a : Wert, den Y hat, wenn $X = 0$ ist
b : Veränderung von Y bei Zunahme von X um 1 Einheit

Spezialfall dichotom nominalskalierte UV

a : Mittelwert der mit 0 kodierten Kategorie (Referenz)
b : Veränderung in AV, wenn man von Referenz zur mit 1 kodierten Kategorie "übergeht"

$\rightarrow$ Steigung entspricht genau Mittelwertsdifferenz zwischen beiden Kategorien



Einfache lineare Regression

Dichotom nominalskalierte Prädiktoren (UVs)

Mittelwerte beider Kategorien zum Vergleich:

$$\bar{y}_0 = 4.24$$

$$\bar{y}_1 = 9.27$$

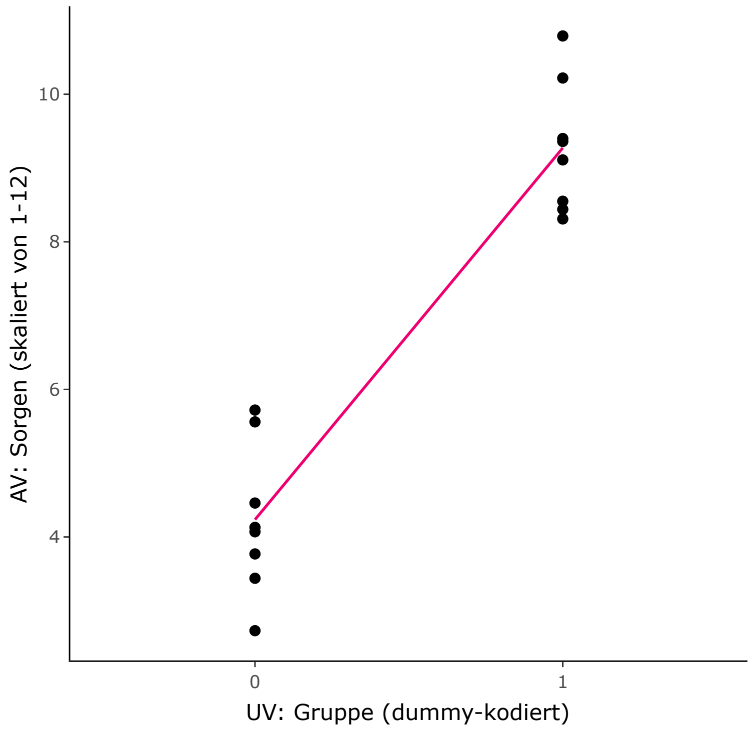
$$\bar{y}_1 - \bar{y}_0 = 5.04$$

Bestimmung Regressionskoeffizienten:

$$b = \frac{\sigma^2_{yx}}{\sigma^2_x} = \frac{1.34}{0.27} = 5.04$$

$$a = \bar{y} - b \cdot \bar{x} = 4.24$$

$\rightarrow a$: Mittelwert der mit 0 kodierten Kategorie (Referenz)
$\rightarrow$ b entspricht genau Mittelwertsdifferenz zwischen beiden Kategorien



Einfache lineare Regression

Dichotom nominalskalierte Prädiktoren (UVs)

Regression vs. unabhängiger t-Test

- Steigung entspricht genau Mittelwertsdifferenz zwischen beiden Kategorien
- unabhängiger t-Test: Prüft Mittelwertsdifferenz zwischen 2 Gruppen

→ Test der Steigung auf Signifikanz gelangt zu **identischem Ergebnis** wie der t-Test

- Grund: Gemeinsame mathematische Fundierung im Allgemeinen Linearen Modell
- Man könnte also auch lediglich mit der Regression Gruppenunterschiede berechnen

Einfache lineare Regression

Dichotom nominalskalierte Prädiktoren (UVs)

Regression vs. unabhängiger t-Test in R

```
summary(lm(Sorgen ~ Gruppe, data = df2))
```

```
##
## Call:
## lm(formula = Sorgen ~ Gruppe, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5050 -0.7406 -0.1338  0.4056  1.5175
##
## Coefficients:
##             Estimate Std. Error t value      Pr(>|t|)
## (Intercept)   4.2350     0.3348   12.65 0.00000000474 ***
## Gruppe        5.0375     0.4735   10.64 0.00000004315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.947 on 14 degrees of freedom
## Multiple R-squared:  0.8899,    Adjusted R-squared:  0.8821
## F-statistic: 113.2 on 1 and 14 DF,  p-value: 0.00000004315
```

```
t.test(Sorgen ~ Gruppe, data = df2, var.equal = T)
```

```
##
##     Two Sample t-test
##
## data:  Sorgen by Gruppe
## t = -10.639, df = 14, p-value = 0.00000004315
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal t
## 95 percent confidence interval:
##  -6.053057 -4.021943
## sample estimates:
## mean in group 0 mean in group 1
##          4.2350          9.2725
```

$\longrightarrow$ t-Wert und p-Wert von Regression und t-Test sind identisch!

Einfache lineare Regression

Non-lineare Zusammenhänge

- Wie der Name bereits sagt, eignet sich die einfache lineare Regression in erster Linie für lineare Zusammenhänge.
- Ihre Anwendung ist also prinzipiell nur zur Modellierung solcher Zusammenhänge angemessen.

Beispiele für bivariate (zwischen 2 Variablen) non-lineare Zusammenhänge:

- Exponentieller Zusammenhang
- Quadratisches Polynom (parabolischer Zusammenhang)
- Kubisches Polynom
- logarithmischer Zusammenhang

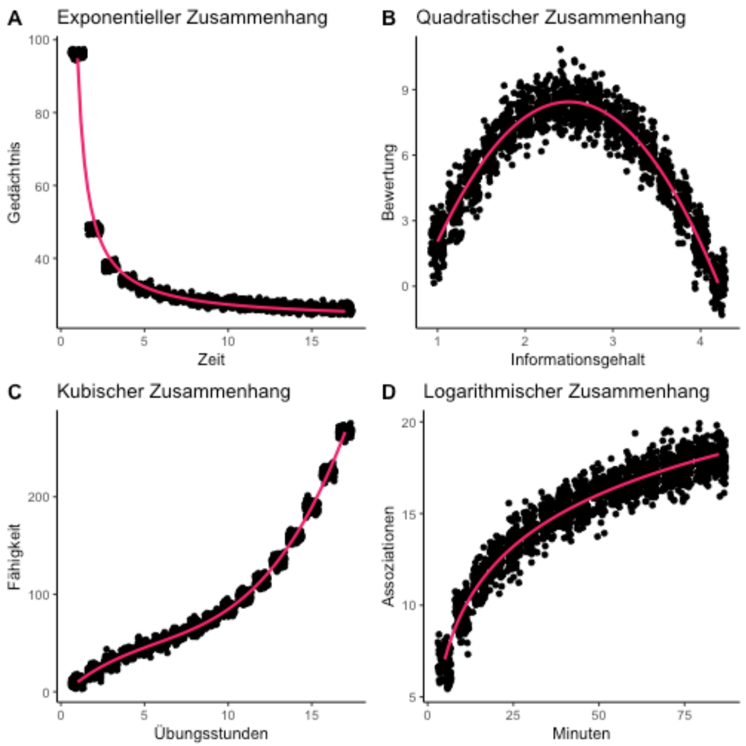
Einfache lineare Regression

Non-lineare Zusammenhänge

- A: $y = a \cdot b^{\cdot 1/x}$
- B: $y = a + b_1 \cdot x + b_2 \cdot x^2$
- C: $y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$
- D: $y = a + b \cdot log(x)$

$\rightarrow$ Es gibt nach wie vor nur die Variablen X und Y

$\rightarrow$ Lediglich die angenommene (modellierte) Beziehung ändert sich



Einfache lineare Regression

Regressionsgewichte

Unstandardisierte Regressionsgewichte:

- Steigungsparameter (b_{yx}) = Regressionsgewicht
- X-Wert wird "gewichtet", sodass entsprechendes Y herauskommt (Verechnungsregel: z.B. mal 2 oder durch 3)
- Steigung (b_{yx}) gibt an, um wie viele Einheiten sich Y in der Originalmetrik (Fragebogenpunkte, Reaktionszeit, Gewicht in mg/g/kg...) verändert, wenn X um 1 Einheit zunimmt
- Steigung in Originalmetrik = unstandardisiertes Regressionsgewicht

$$b_{yx} = \frac{\text{Anzahl Einheiten auf Y}}{\text{pro 1 Einheit X}}$$

Problem mit unstandardisierten Regressionsgewichten:

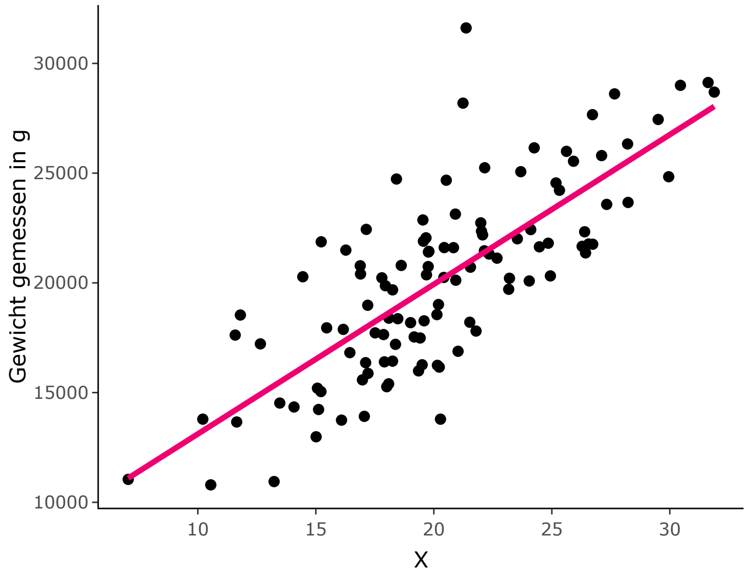
- unstandardisierte Steigungsparameter für 2 Regressionen mit unterschiedlichen Y können nicht hinsichtlich ihrer Größe (Skalierung) verglichen werden
- Beispiel: 1 Einheit Reaktionszeit [in ms] $\neq$ 1 Einheit Fragebogenpunkte [z.B. 1-10]

Einfache lineare Regression

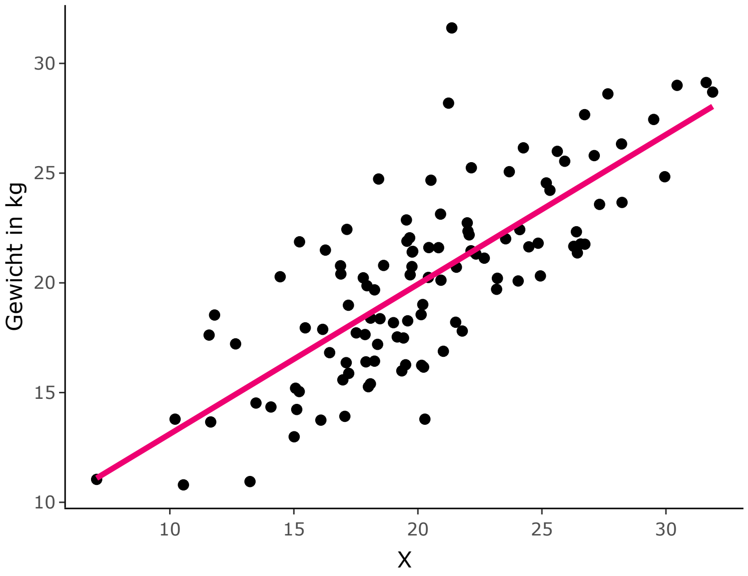
Regressionsgewichte

Selbes Konstrukt (Gewicht) $\rightarrow$ unterschiedliche Originalmetrik von Y $\rightarrow$ unterschiedliche Steigung

Regression mit Y in g (b = 864.4)



Regression mit Y gemessen in kg (b = 0.8644)



Einfache lineare Regression

Regressionsgewichte

Standardisierte Regressionsgewichte:

- Ziel: einheitliche Metrik für Vergleiche erhalten
- Standardisiertes Regressionsgewicht wird oft als β bezeichnet ("beta-Gewicht")
- Vorgehen: Regressionsgewicht muss von Originalmetrik des untersuchten Merkmals (Y) bereinigt werden
  - Zähler- und Nennereinheiten werden an der Streuung von Y und X relativiert

$$b_{yx} = \frac{\frac{\text{Anzahl Einheiten auf Y}}{\sigma_y}}{\frac{\text{pro 1 Einheit X}}{\sigma_x}} = b \cdot \frac{\frac{1}{\sigma_y}}{\frac{1}{\sigma_x}} = b \cdot \frac{\sigma_y}{\sigma_x}$$

Interpretation:

- Standardisiertes Regressionsgewicht (β) ist unabhängig von Originalmetrik
- Es drückt aus, um wie viele Standardabweichungen sich Y verändert, wenn X um eine Standardabweichung zunimmt.
- Sonderfall einfache Regression (nur 1 UV): β ist identisch mit Pearson-Korrelation $(r) \rightarrow$ Wertbereich -1 bis +1

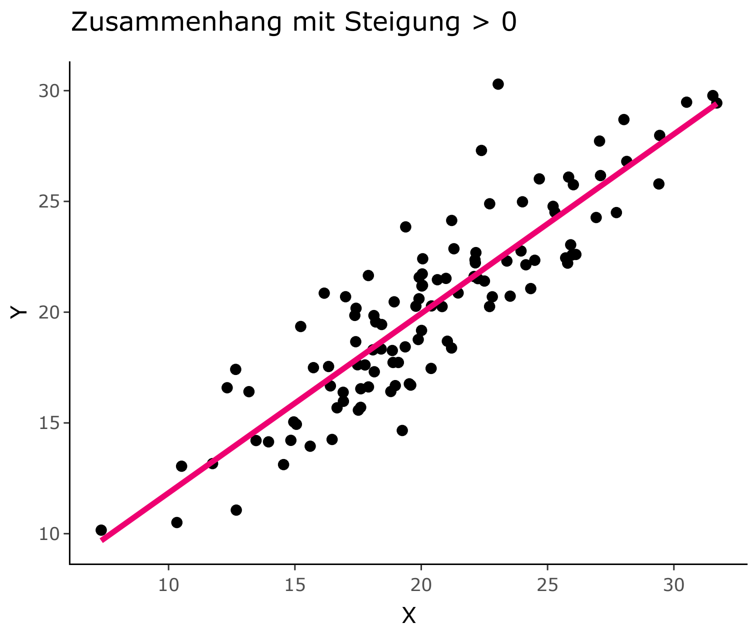
Einfache lineare Regression

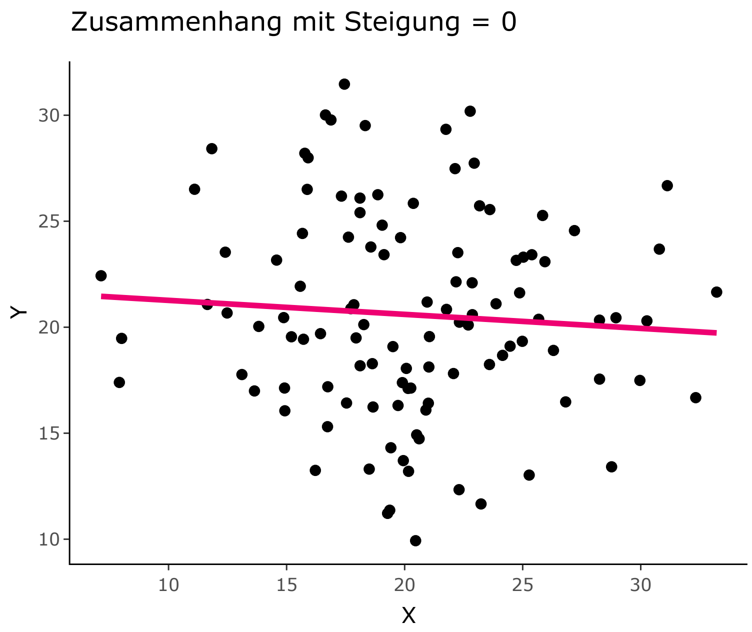
Signifikanztest für Regressionskoeffizienten

- Mit der Regression kann z.B. überprüft werden, ob überhaupt ein linearer Zusammenhang zwischen AV und UV besteht.
- In der Nullhypothese wird in diesem Fall die Aussage formuliert, dass der lineare Zusammenhang zwischen der UV und der AV gleich null ist.
- Die statistischen Hypothesen für diesen Fall lauten:
    - $H_0 : \beta = 0$
    - $H_1 : \beta \neq 0$
- Allgemeiner Fall:
    - $H_0 : \beta = \beta_0$
    - $H_1 : \beta \neq \beta_0$
- mit $\beta_0 = a$ (Y-Achsenabschnitt)

Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten





Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

- Zur Beurteilung, ob X (UV) Y (AV) statistisch bedeutsam vorhersagt, rechnen wir einen Signifikanztest, der wie der t-Test funktioniert (Wald-Test)
- Prüfgröße ist t-verteilt mit $df = N - 2$ Freiheitsgraden
- Sie wird gebildet, indem der unstandardisierte Regressionskoeffizient b durch seinen Standardfehler geteilt wird (an diesem relativiert wird)

$$t = \frac{b}{s_b}$$

- Standardfehler (s_b) schätzt Streuung des Regressionskoeffizienten um den Populationsmittelwert (wie beim t-Test)

Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Beispiel: Vorhersage Leistung im Verkehrstest (AV) aus IQ (UV):

1. Regressionsgerade aufstellen
2. Standardschätzfehler ermitteln $\left(\hat{\sigma}_{(y|x)}\right)$
3. Standardfehler der Steigung (s_b) ermitteln
4. empirischen t-Wert (t_{emp}) berechnen
5. Entscheidungsregel: Vergleich empirischer t-Wert vs. kritischer t-Wert (t_{krit})

ID	UV: IQ	AV: Testleistung (skaliert von 1-10)
1	110	4
2	112	5
3	100	7
4	91	2
5	125	9
6	99	3
7	107	5
8	112	3
9	103	6
10	117	8
11	114	4
12	106	4
13	129	7
14	88	3
15	94	4
16	107	5
17	108	4
18	114	7
19	115	6
20	104	5

Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Standardfehler der Steigung

Für die Berechnung des Standardfehlers der Steigung (b_{yx}) ermitteln wir den Standardschätzfehler:

$$\hat{\sigma}_{(y|x)} = \sqrt{\frac{n \cdot s_y^2 - n \cdot b^2 \cdot s_x^2}{n-2}}$$

Mit Kenntnis des Standardschätzfehler, errechnet sich der Standardfehler der Steigung:

$$s_b = \frac{\hat{\sigma}_{(y|x)}}{s_x \cdot \sqrt{n}}$$

Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

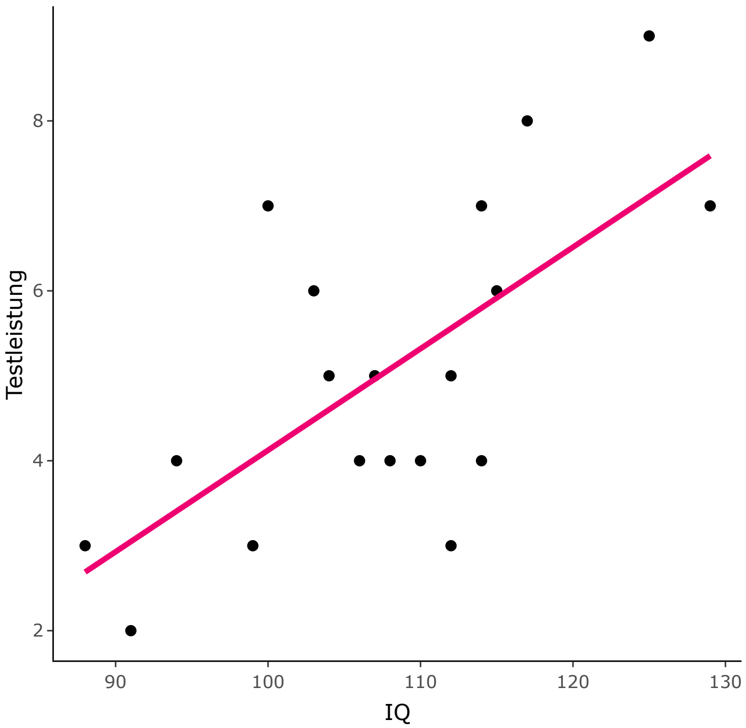
Regressionsgerade aufstellen:

$$b = \frac{\sigma^2_{yx}}{\sigma^2_x} = \frac{12.86}{107.57} = 0.12$$

$$a = \bar{y} - b \cdot \bar{x} = -7.83$$

$$\hat{y} = a + b \cdot x = -7.8 + 0.12 \cdot x$$

VORSICHT: Y-Achsenabschnitt im Graph rechts nicht sichtbar, da definiert als Y wenn $X = 0$ (kein IQ von 0 gemessen)



Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

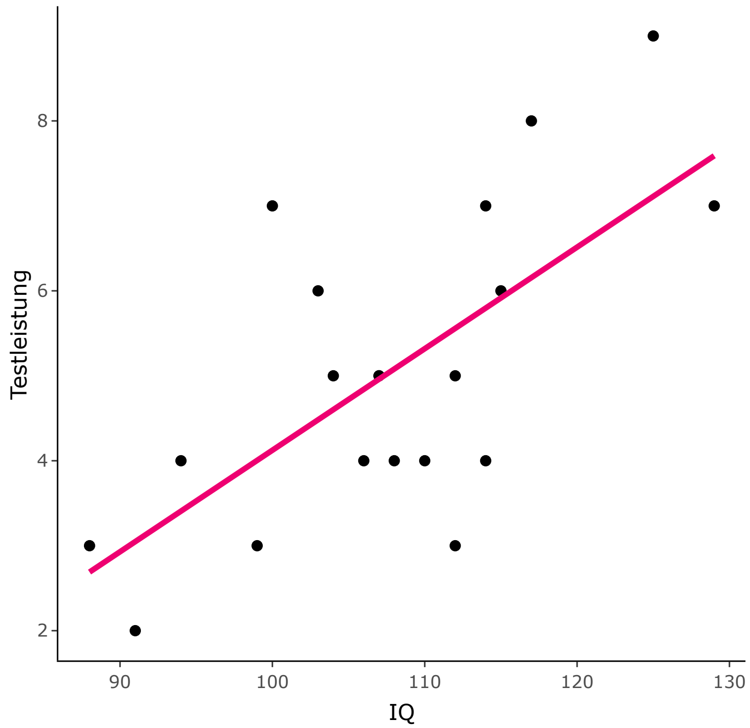
Standardschätzfehler ermitteln:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 107.57$$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = 3.42$$

$$\hat{\sigma}_{(y|x)} = \sqrt{\frac{n \cdot s_y^2 - n \cdot b^2 \cdot s_x^2}{n-2}}$$

$$\hat{\sigma}_{(y|x)} = \sqrt{\frac{20 \cdot 3.42 - 20 \cdot 0.12^2 \cdot 107.57}{18}} = 1.44$$



Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Standardfehler der Steigung ermitteln:

$$s_b = \frac{\hat{\sigma}_{(y|x)}}{s_x \cdot \sqrt{n}}$$

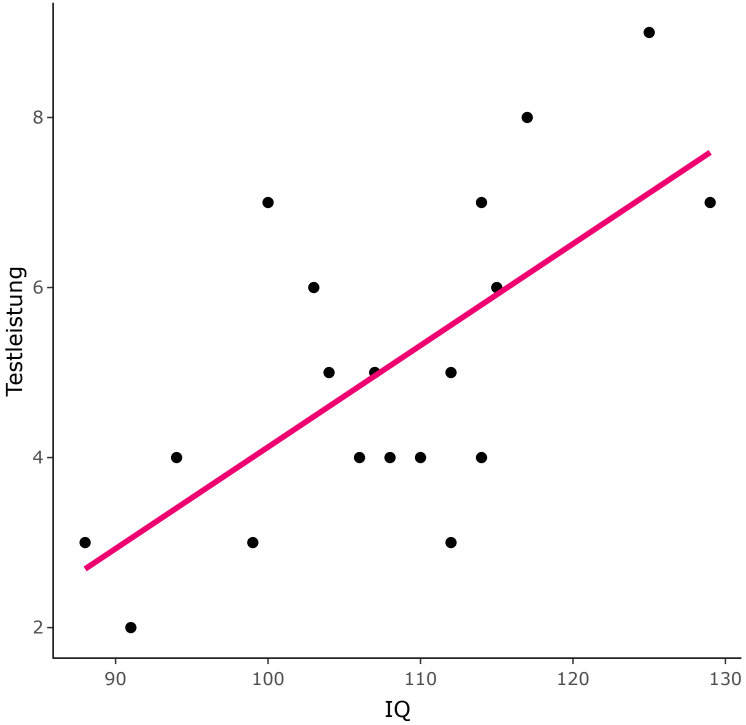
$$s_b = \frac{1.44}{10.37 \cdot \sqrt{20}} = 0.03118$$

Empirischen t-Wert (t_{emp}) ermitteln:

$$t = \frac{b}{s_b} = \frac{0.12}{0.03118} = 3.8$$

Vergleich empirischer vs. kritischer t-Wert:

- $t_{krit,df=18,\alpha=.05} = 1.734 < 3.8$
- $t_{krit} < t_{emp} \rightarrow$ Test ist signifikant.



Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten in R

```
model = lm(Testleistung ~ IQ, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Testleistung ~ IQ, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.55792 -1.01526  0.03963  0.74572  2.87621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.82728    3.37416  -2.320  0.03231 *
## IQ           0.11951    0.03118   3.833  0.00122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.409 on 18 degrees of freedom
## Multiple R-squared:  0.4494,    Adjusted R-squared:  0.4188
## F-statistic: 14.69 on 1 and 18 DF,  p-value: 0.001218
```

Berechnung empirischer t-Wert:

$$t = \frac{b}{s_b} = \frac{0.11951}{0.03118} = 3.8$$

Berechnung Freiheitsgrade:

$$df = N - 2 = 18$$

Entscheidungsregel:

- **Option 1:** Kritischen t-Wert ($df = 18$ und $\alpha = .05$) in t-Tabelle nachsehen

$\rightarrow$ wenn $t_{emp} > t_{krit}$ ist Test signifikant.

- **Option 2:** p-Wert mit $\alpha = .05$ vergleichen

$\rightarrow$ wenn $p < .05$ ist Test signifikant.

Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Konfidenzintervall (KI) für den Steigungsparameter

Inferenz:

$$b \pm t_{a/2,n-2} \cdot se(b)$$

To Do:

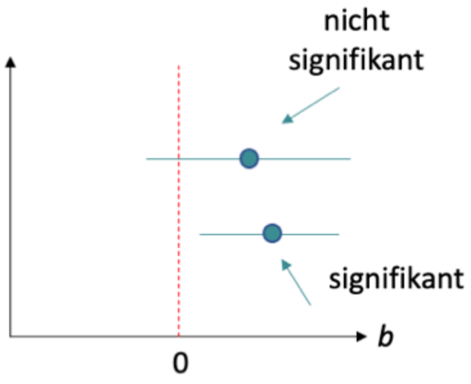
Prüfen ob das in unserer Stichprobe beobachtete b sich verlässlich von 0 unterscheidet.

Strategie:

Wir spannen um b ein **Konfidenzintervall** – je schmaler desto besser:

Es fließen 3 Größen ein:

1) Stärke des Zusammenhangs → erhöht Glaubwürdigkeit
2) Stichprobengröße → erhöht Glaubwürdigkeit
3) Der Streuung um die Gerade → reduziert Glaubwürdigkeit



Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Konfidenzintervall (KI) für den Steigungsparameter

Für die Berechnung des KI ermitteln wir den Standardschätzfehler:

$$\hat{\sigma}_{(y|x)} = \sqrt{\frac{n \cdot s_y^2 - n \cdot b^2 \cdot s_x^2}{n-2}}$$

Mit Kenntnis des Standardschätzfehler, des Signifikanzniveaus $\alpha = .05$ und der Freiheitsgrade $df = N - 2$ lautet das KI für β_{yx}:

$$b_{yx} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}_{(y|x)}}{s_x \cdot \sqrt{n}}$$

Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Konfidenzintervall (KI) für den Steigungsparameter

Beispiel: Vorhersage Leistung im Verkehrstest (AV) aus IQ (UV):

1. Regressionsgerade aufstellen
2. Standardschätzfehler ermitteln
3. KI für Steigungsparameter berechnen (Hypothesentest)
4. Entscheidungsregel: KI enthält die 0 nicht $(\beta \neq 0)$

ID	UV: IQ	AV: Testleistung (skaliert von 1-10)
1	110	4
2	112	5
3	100	7
4	91	2
5	125	9
6	99	3
7	107	5
8	112	3
9	103	6
10	117	8
11	114	4
12	106	4
13	129	7
14	88	3
15	94	4
16	107	5
17	108	4
18	114	7
19	115	6
20	104	5

Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Konfidenzintervall (KI) für den Steigungsparameter

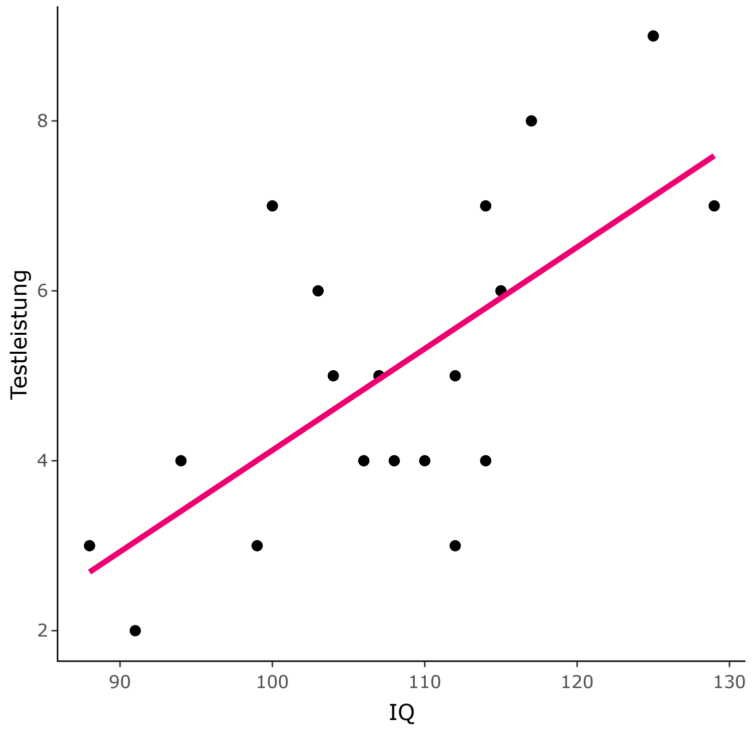
Regressionsgerade aufstellen:

$$b = \frac{\sigma^2_{yx}}{\sigma^2_x} = \frac{12.86}{107.57} = 0.12$$

$$a = \bar{y} - b \cdot \bar{x} = -7.83$$

$$\hat{y} = a + b \cdot x = -7.8 + 0.12 \cdot x$$

VORSICHT: Y-Achsenabschnitt im Graph rechts nicht sichtbar, da definiert als Y wenn $X = 0$ (kein IQ von 0 gemessen)



Einfache lineare Regression

CHARLOTTE FRESENIUS HOCHSCHULE UNIVERSITY OF PSYCHOLOGY

Signifikanztest für Regressionskoeffizienten

Konfidenzintervall (KI) für den Steigungsparameter

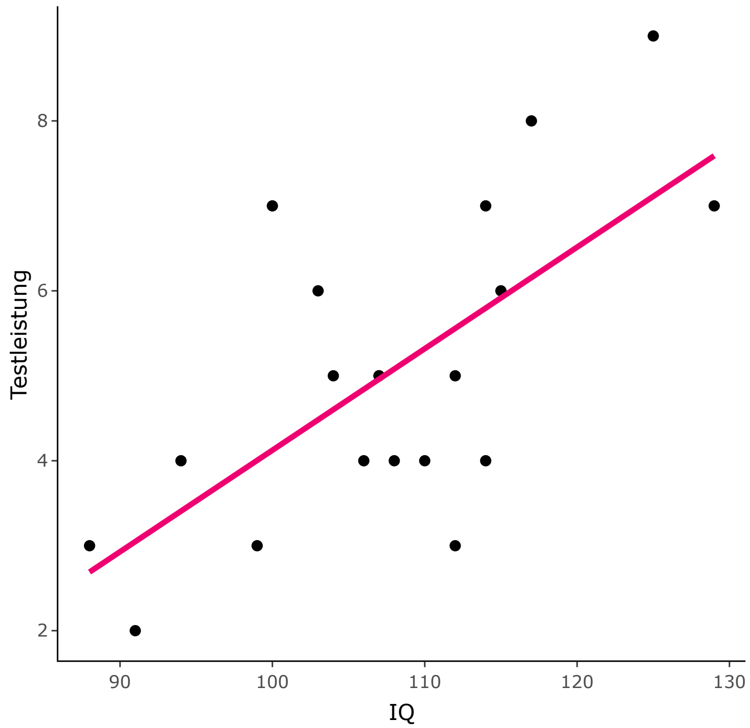
Standardschätzfehler ermitteln:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 107.57$$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = 3.42$$

$$\hat{\sigma}_{(y|x)} = \sqrt{\frac{n \cdot s_y^2 - n \cdot b^2 \cdot s_x^2}{n-2}}$$

$$\hat{\sigma}_{(y|x)} = \sqrt{\frac{20 \cdot 3.42 - 20 \cdot 0.12^2 \cdot 107.57}{18}} = 1.44$$



Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Konfidenzintervall (KI) für den Steigungsparameter

KI für Steigungsparameter berechnen $(\alpha = .05)$:

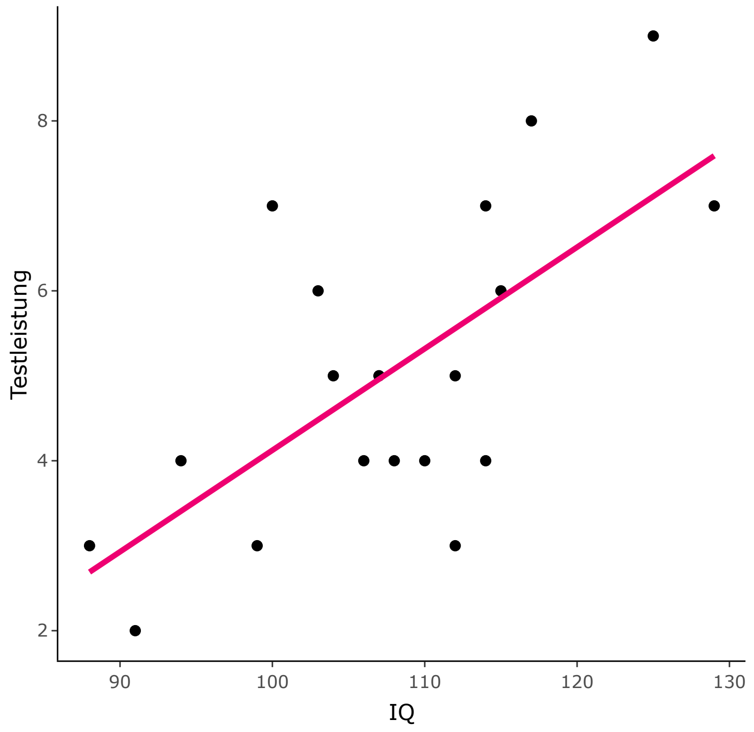
$$b \pm t_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}_{(y|x)}}{s_x \cdot \sqrt{n}}$$

$$0.12 \pm 2.10 \cdot \frac{1.44}{10.37 \cdot \sqrt{20}} = 0.12 \pm 0.07$$

- untere Grenze: $0.12 - 0.07 = 0.05$
- obere Grenze: $0.12 + 0.07 = 0.19$

$\beta_{yx} = 0.12, KI_{95\%}[0.05 - 0.19]$

$\rightarrow$ Da das KI den Wert 0 nicht umschließt, ist β_{yx} signifikant.



Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Konfidenzintervall (KI) für den Steigungsparameter in R

```
model = lm(Testleistung ~ IQ, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = Testleistung ~ IQ, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.55792 -1.01526  0.03963  0.74572  2.87621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.82728    3.37416  -2.320  0.03231 *
## IQ           0.11951    0.03118   3.833  0.00122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.409 on 18 degrees of freedom
## Multiple R-squared:  0.4494,    Adjusted R-squared:  0.4188
## F-statistic: 14.69 on 1 and 18 DF,  p-value: 0.001218
```

```
model = lm(Testleistung ~ IQ, data = df)
confint(model)
```

```
##                   2.5 %     97.5 %
## (Intercept) -14.91612769 -0.7384289
## IQ            0.05400855  0.1850129
```

Ergebnis:

- KI wird automatisch für Y-Achsenabschnitt und Steigung berechnet
- KI umschließt die Regressionskoeffizienten (links bei Estimate angegeben)
- I.d.R. sind wir für den Hypothesentest (H1: "Es besteht ein Zusammenhang zwischen X und Y.") jedoch nur an der Signifikanz der Steigung interessiert

Einfache lineare Regression

Signifikanztest für Regressionskoeffizienten

Konfidenzintervall (KI) für einzelne $\hat{y}$ (vorhergesagte Werte)

- Um die Genauigkeit für einzelne Personen vorhergesagte Werte anzugeben, lässt sich ebenfalls ein KI berechnen
- Hierfür werden jedoch keine zusätzlichen Informationen benötigt, lediglich die Formel sieht etwas anders aus:

Formel für das KI eines einzelnen vorhergesagten Werts:

$$\hat{y}_j = \pm t_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{(y|x)} \cdot \sqrt{\frac{1}{n} + \frac{(x_j - \bar{x})^2}{n \cdot s_x^2}}$$

Einfache lineare Regression

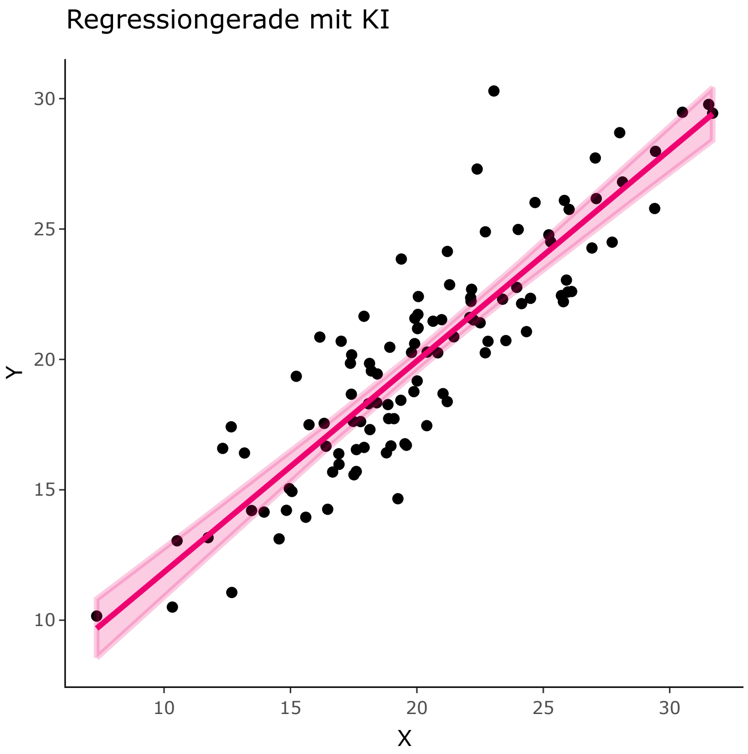
Signifikanztest für Regressionskoeffizienten

Konfidenzintervall (KI) für einzelne $\hat{y}$

Formel für das KI eines einzelnen vorhergesagten Werts:

$$\hat{y}_j = \pm t_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{(y|x)} \cdot \sqrt{\frac{1}{n} + \frac{(x_j - \bar{x})^2}{n \cdot s_x^2}}$$

- Für jeden auf der Gerade liegenden Vorhersagewert wird die obere und untere Grenze des KIs abgebildet



Einfache lineare Regression

Voraussetzungen der einfachen linearen Regression

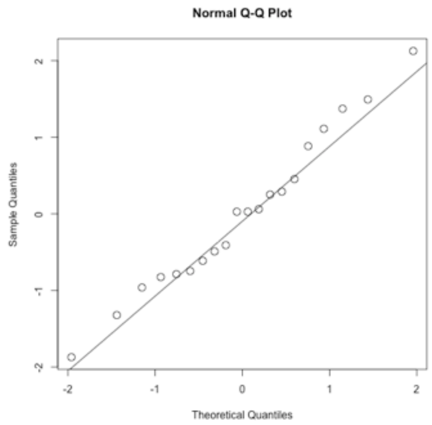
1. Das Kriterium (AV) muss intervallskaliert sein.
2. Der Prädiktor (UV) darf nominal, ordinal und intervallskaliert sein.
3. Die Werte der einzelnen Versuchspersonen müssen unabhängig voneinander sein
4. Der Zusammenhang muss theroretisch linear sein (sonst andere Regressionsmodelle nutzen).
5. Streuungen der zu einem x-Wert gehörenden y-werte müssen über ganzen Wertebereich von X homogen sein (Homoskedastizität).
6. Die Residuen sollten normalverteilt sein.

Einfache lineare Regression

Voraussetzungen der einfachen linearen Regression

Normalverteilung der Residuen:

```
qqnorm(rstandard(model), cex = 1.5)
qqline(rstandard(model))
```




```
model = lm(Testleistung ~ IQ, data = df)
shapiro.test(rstandard(model))
```

```
##
##      Shapiro-Wilk normality test
##
## data:  rstandard(model)
## W = 0.97872, p-value = 0.9165
```

Benchmarks:

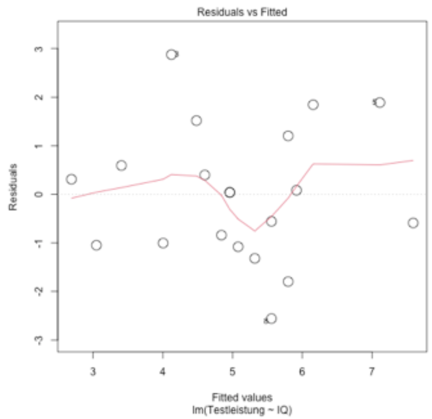
- QQ-Plot: Punkte sollten möglichst auf der 45 Grad Diagonalen liegen
- Shapiro-Wilk Test: p-Wert sollte > als $\alpha = .05$ sein

Einfache lineare Regression

Voraussetzungen der einfachen linearen Regression

Homoskedastizität:

```
model = lm(Testleistung ~ IQ, data = df)
plot(model, 1, cex = 2)
```



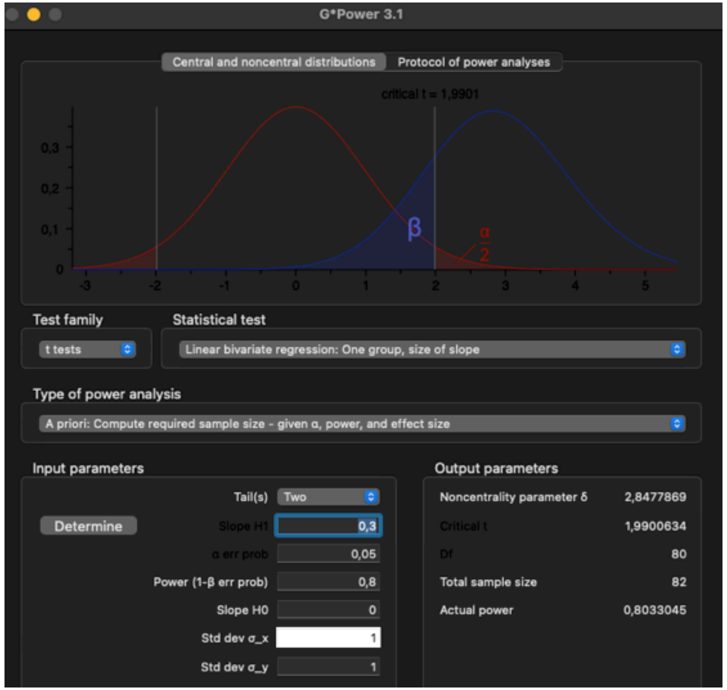
- Plot der standardisierten Residuen gegen die standardisierten vorhergesagten Werte
- Ideal ist eine Punktewolke ohne Systematik (Pattern)
- Die Linie sollte relativ horizontal verlaufen

→ dann ist Homoskedastizitätsannahme gegeben

Einfache lineare Regression

Teststärkeanalyse und Stichprobenumfangsplanung

- Bei einfachen linearen Regressionen nutzt man als Effektstärke für die Stichprobenumfangsplanung oft
    - das standardisierte Regressionsgewicht oder
    - das Bestimmtheitsmaß (R^2)



Einfache lineare Regression

Berichten der Ergebnisse nach APA

Statistischer Bericht: (ausführlich)

Es wurde eine lineare Regression mit dem Prädiktor IQ und dem Kriterium Testleistung durchgeführt. Die Analyse ergab, dass IQ ein signifikanter Prädiktor für die Testleistung ist, β = 0.12, t(18) = 3.83, p = .001. Das Modell erklärte etwa 44.9% der Varianz der Testleistung, R^2 = .42, F(1,18) = 14.69, p = .001.

Eine Prüfung der Normalverteilung der standardisierten Residuen mittels Shapiro-Wilk-Test ergab keinen Hinweis auf eine signifikante Abweichung von der Normalverteilung, W = 0.98, p = .92. Die visuelle Inspektion des Residualplots („Residuals vs. Fitted“) zeigt ebenfalls keine systematische Verletzung der Homoskedastizität.

Einfache lineare Regression

Berichten der Ergebnisse nach APA

Statistischer Bericht: (In Ihrer Klausur)

Wenn Sie in Ihrer Klausur eine Regression händisch berechnen und auf signifikanz prüfen, könnte Ihr Antwortsatz so aussehen:

> Es wurde eine lineare Regression mit dem Prädiktor IQ und dem Kriterium Testleistung durchgeführt. Die Analyse ergab, dass IQ ein signifikanter Prädiktor für die Testleistung ist, β = 0.12, t(18) = 3.83, p = .001. Das Modell erklärte etwa 44.9% der Varianz der Testleistung. **Inhaltliche Interpretation der Modellparameter**
>
> Bei einem IQ von 0 liegt die erwartete Testleistung bei -7.83. Wenn der IQ um eine Einheit zunimmt, steigt die erwartete Testleistung um 0.12 Einheiten.

CHARLOTTE FRESENIUS HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

Take-aways

- Der Mittelwert der vorhergesagten Werte entspricht dem Mittelwert der empirischen Verteilung.
- Nominalskalierte UVs können in der Regression mittels Dummy-Kodierung verwendet werden.
- Das Ergebnis einer Regression mit dichotomer nominalskalierter UV ist äquivalent zum unabhängigen t-Test.
- Während unstandardisierte Regressionsgewichte (b) in der Orginalmetrik der AV angegeben werden, werden standardisierte Regressionsgewichte β in Standardabweichungen (-1 bis +1) angegeben und sind somit über unterschiedliche Modelle hinweg vergleichbar.
- Hypothesentests über Zusammenhänge zwischen Y und X können durchgeführt werden, indem geprüft wird, ob $b \neq 0$ signifikant ist.
- Hypothesen können mittels Wald-Test oder Konfidenzintervall des Steigungsparameters durchgeführt werden.