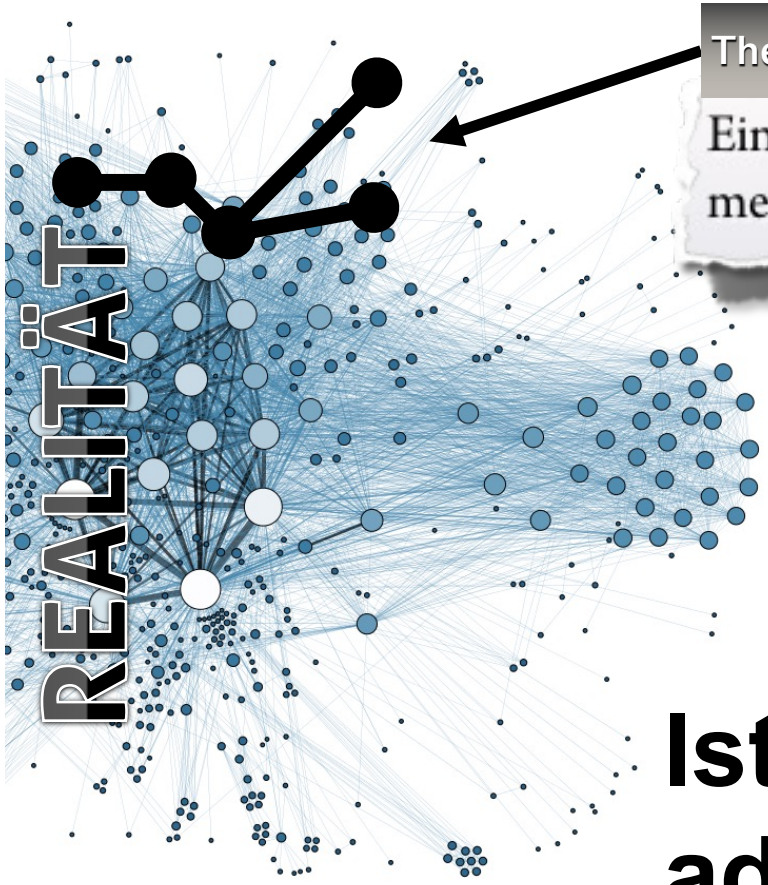
Wiederholung Lineare Regression

Warum Methodenlehre in der Psychologie?

Wie können wir zu Erkenntnissen über die Welt gelangen?

u.a. durch Methodenlehre!



Theorie/Hypothese: Verbales Modell eines Phänomens

Ein Gegenstand im freien Fall beschleunigt immer mehr, bis er auf den Boden trifft.

Ist meine Theorie eine adäquate Abbildung der Realität?

Wie können wir zu Erkenntnissen über die Welt gelangen?

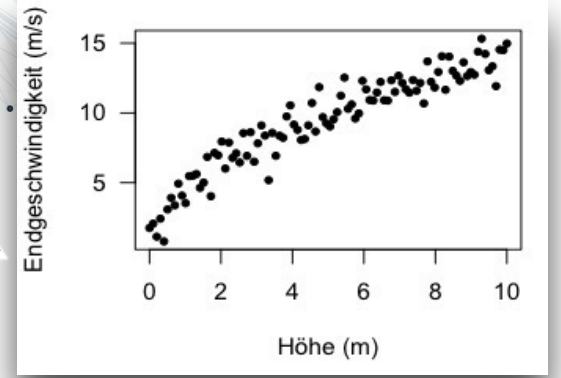
Theorie/Hypothese: Verbales Modell eines Phänomens

Ein Gegenstand im freien Fall beschleunigt immer mehr, bis er auf den Boden trifft.

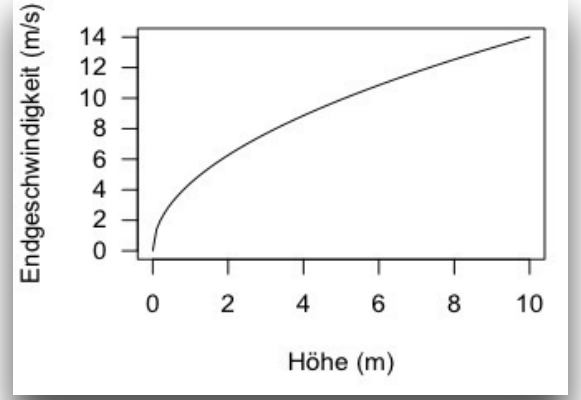
Formales Modell eines Phänomens

$$v(h) = \sqrt{2gh}$$

Die Realität liefert Daten

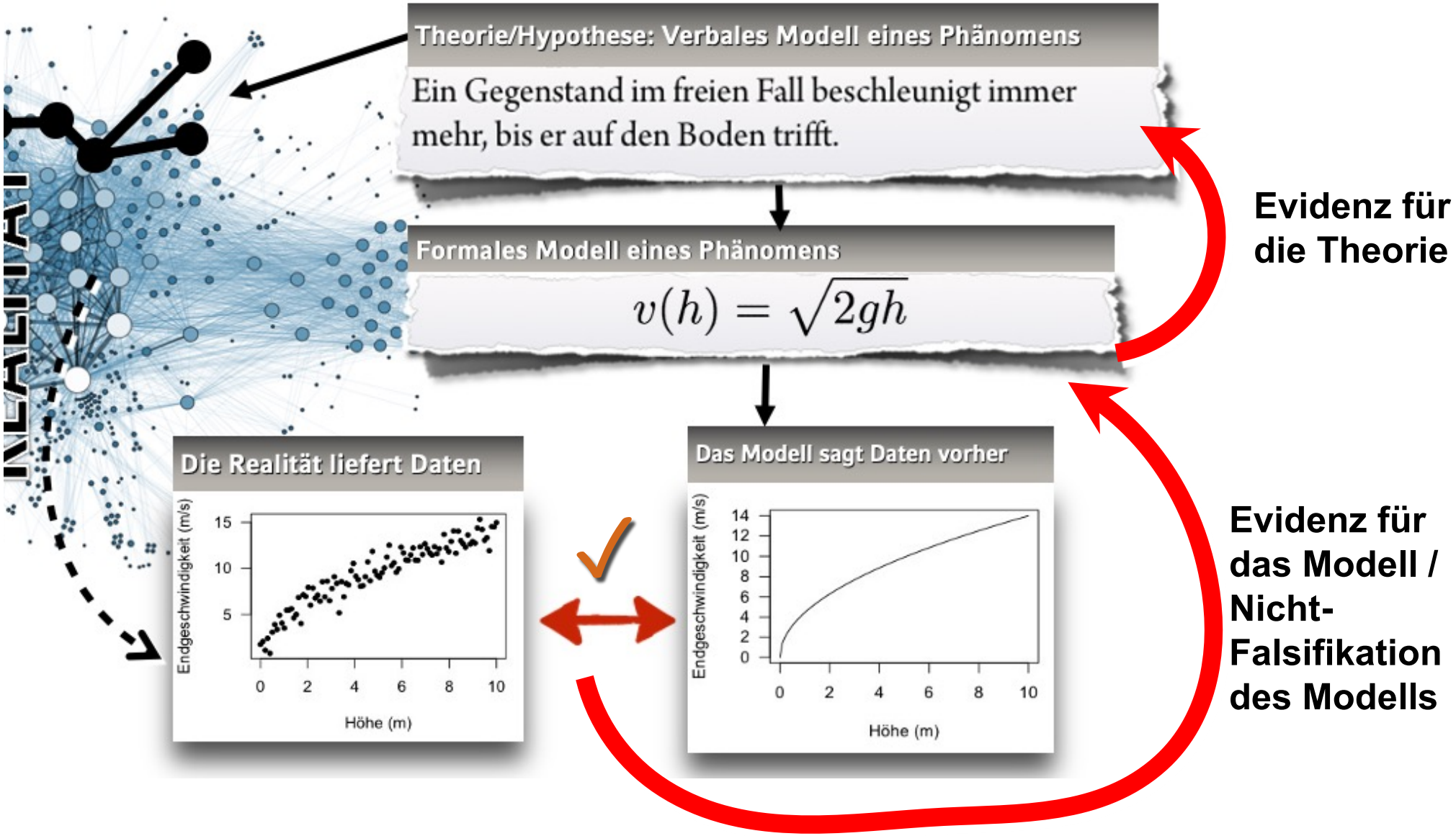


Das Modell sagt Daten vorher



Popper (1935/2005); Meehl (1967; 1990); Dienes (2008); Rouder, Morey, & Wagenmakers (2015)

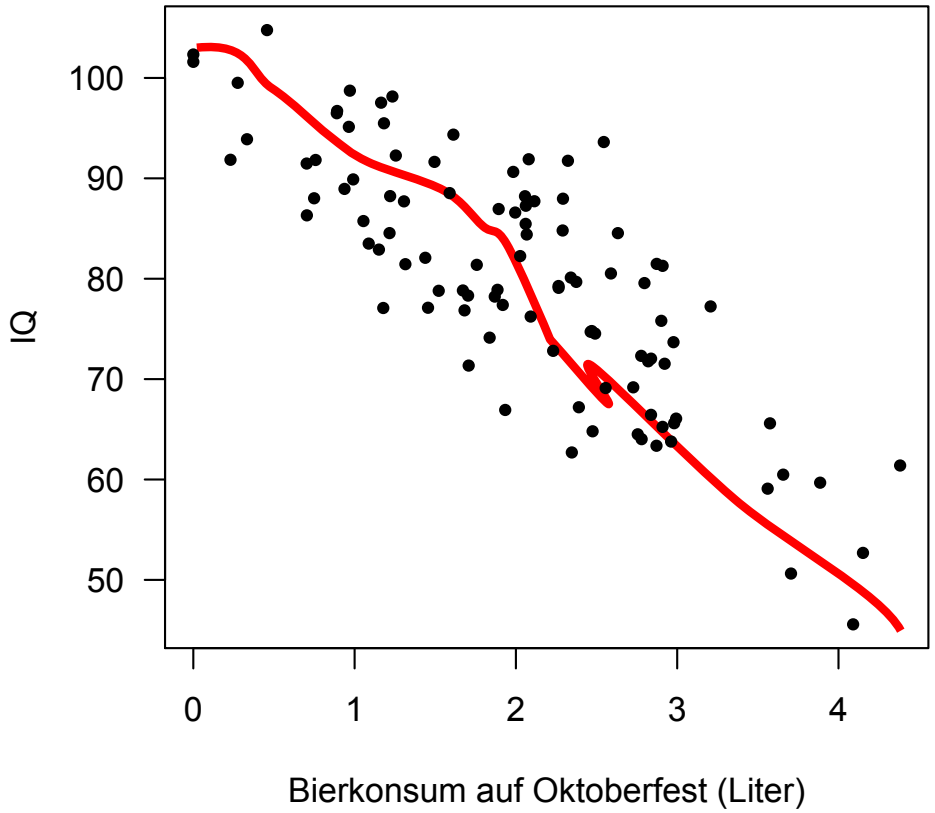
Wie können wir zu Erkenntnissen über die Welt gelangen?



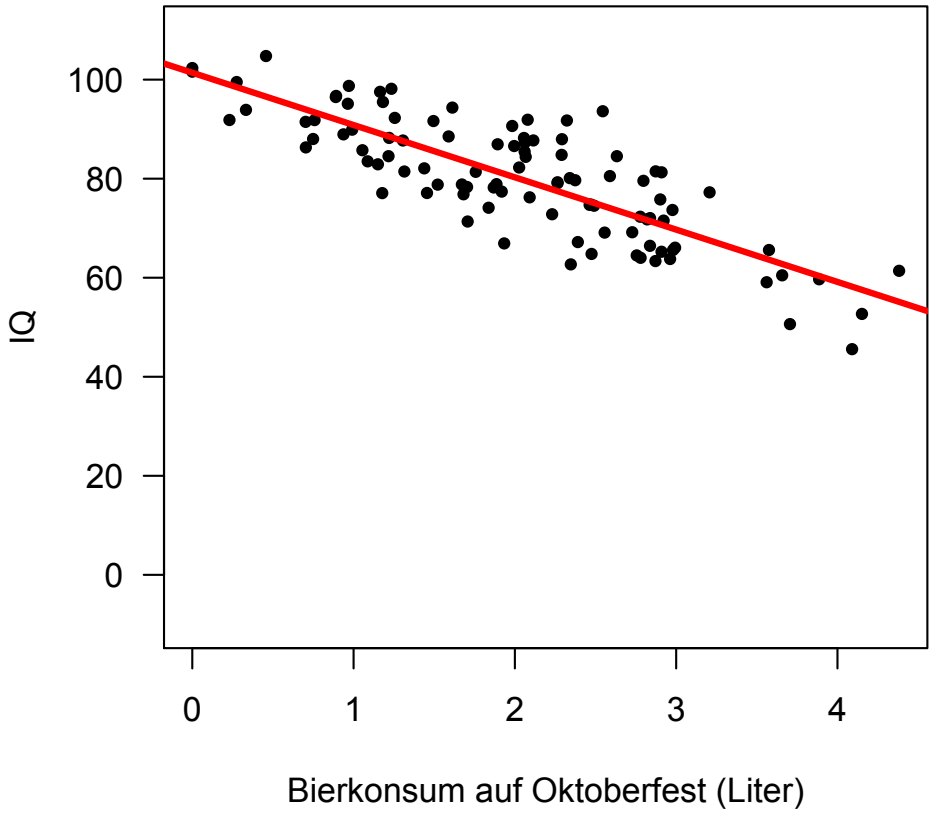
Popper (1935/2005); Meehl (1967; 1990); Dienes (2008); Rouder, Morey, & Wagenmakers (2015)

Zusammenfassung Lineare Regression

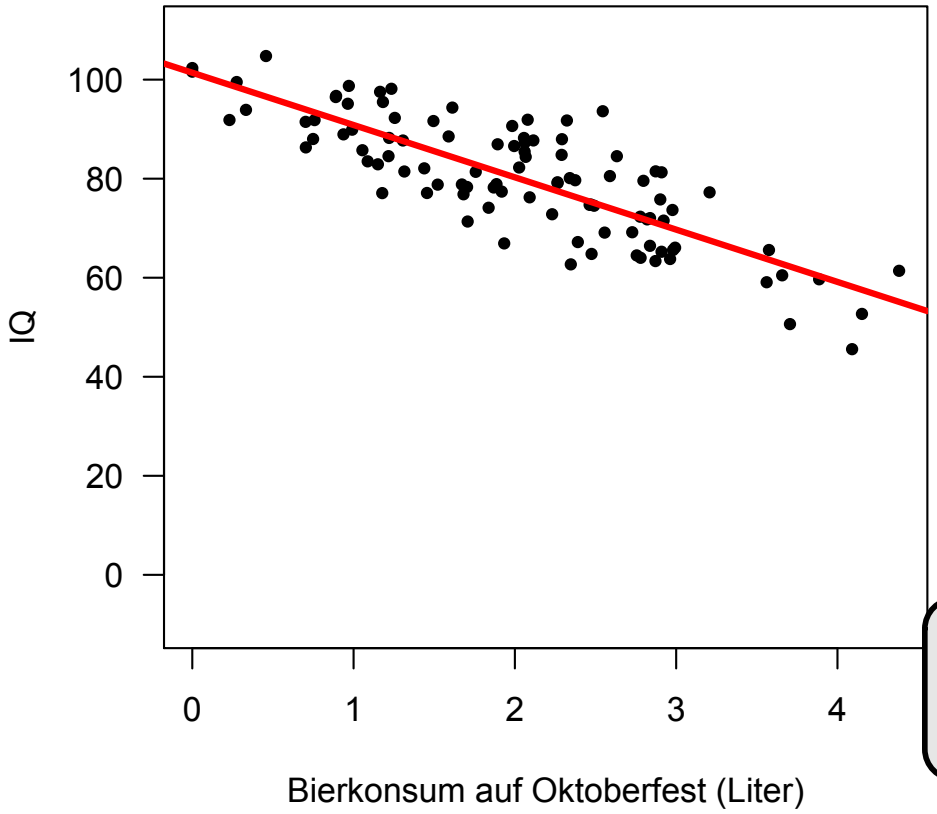
Eine wichtige Klasse von Modellen: Die linearen Modelle



Eine wichtige Klasse von Modellen: Die linearen Modelle



Eine wichtige Klasse von Modellen: Die linearen Modelle



Wie kann man eine Gerade mathematisch beschreiben?

$Y = a + bX$

$IQ = a + b*Liter$

a = Achsenabschnitt (engl.: intercept)

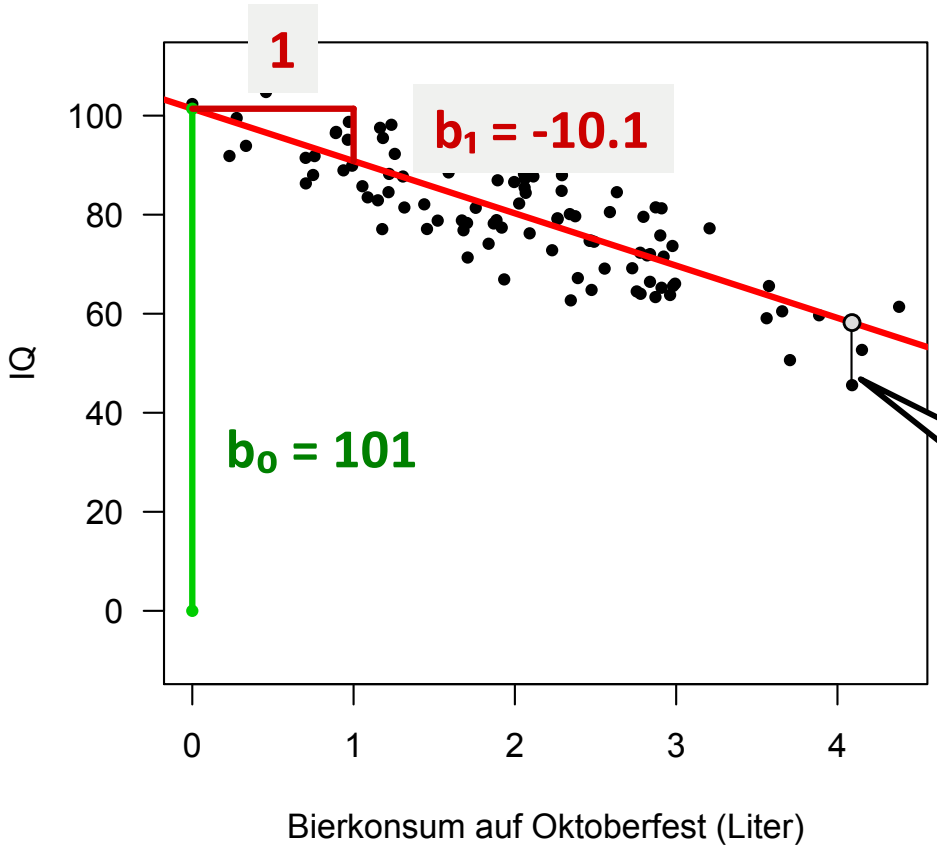
b = Steigung (engl.: slope)

Eine wichtige Klasse von Modellen: Die linearen Modelle

Modellparameter (Achsenabschnitt, bzw, Intercept, eventuell als α bekannt)

Modellparameter (Steigungsparameter, Slope)

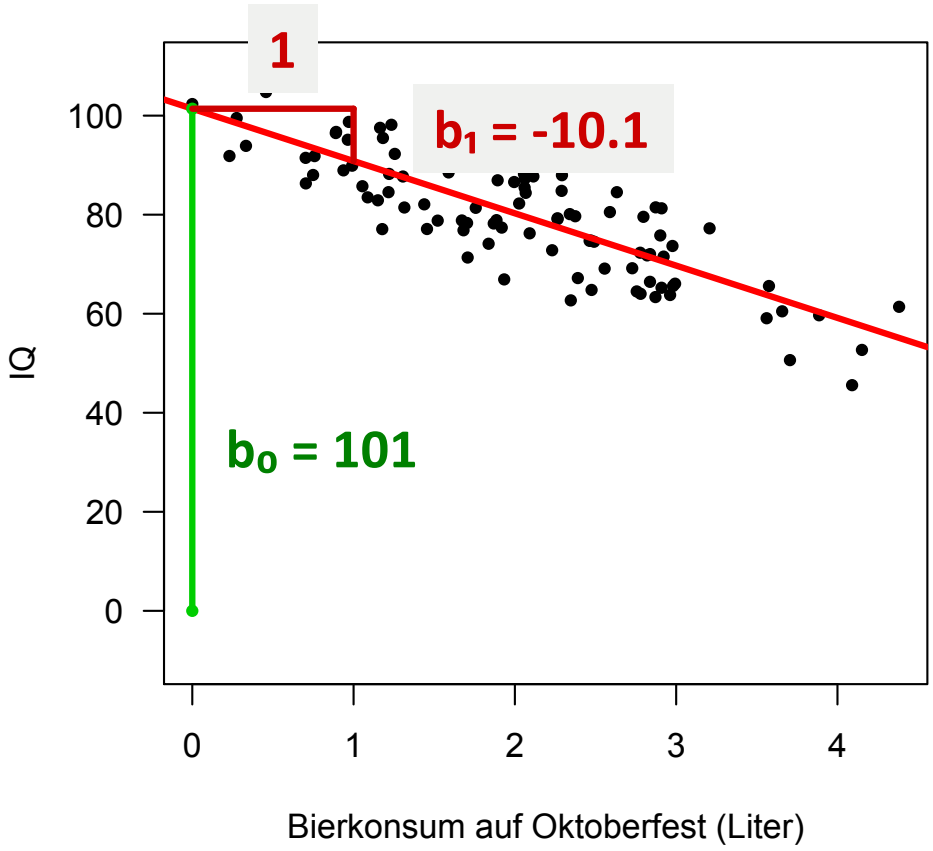
$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i \ , \ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$
$$E(Y_i \mid X_i = x_i) = \mu_i = \beta_0 + \beta_1 \cdot x_i$$
$$\hat{y}_i = b_0 + b_1 \cdot x_i$$



Vorhergesagter Wert (hier nun b0 und b1, da die konkret geschätzten Werte eingesetzt werden)

Residuum: $e_i = y_i - \hat{y}_i$

Eine wichtige Klasse von Modellen: Die linearen Modelle



Dach = vorhergesagter Wert

$$\widehat{IQ}_i = 101 - 10.1 * x_i$$

- Der Achsenabschnitt beschreibt den vorhergesagten Wert, wenn alle Prädiktorvariablen 0 sind
- Der Steigungsparameter gibt an, wie stark die Gerade steigt bzw. fällt. Die Steigung der Geraden ist an jeder Stelle konstant.
- Beispiel: wie stark sinkt der IQ **im Mittel**, wenn der Bierkonsum um eine Einheit zunimmt.
- Beispiel: Jemand der 2 Maß mehr trinkt, hat im Mittel einen IQ, der um 20.2 Einheiten geringer ist.

Wie findet man die optimal passende Gerade?

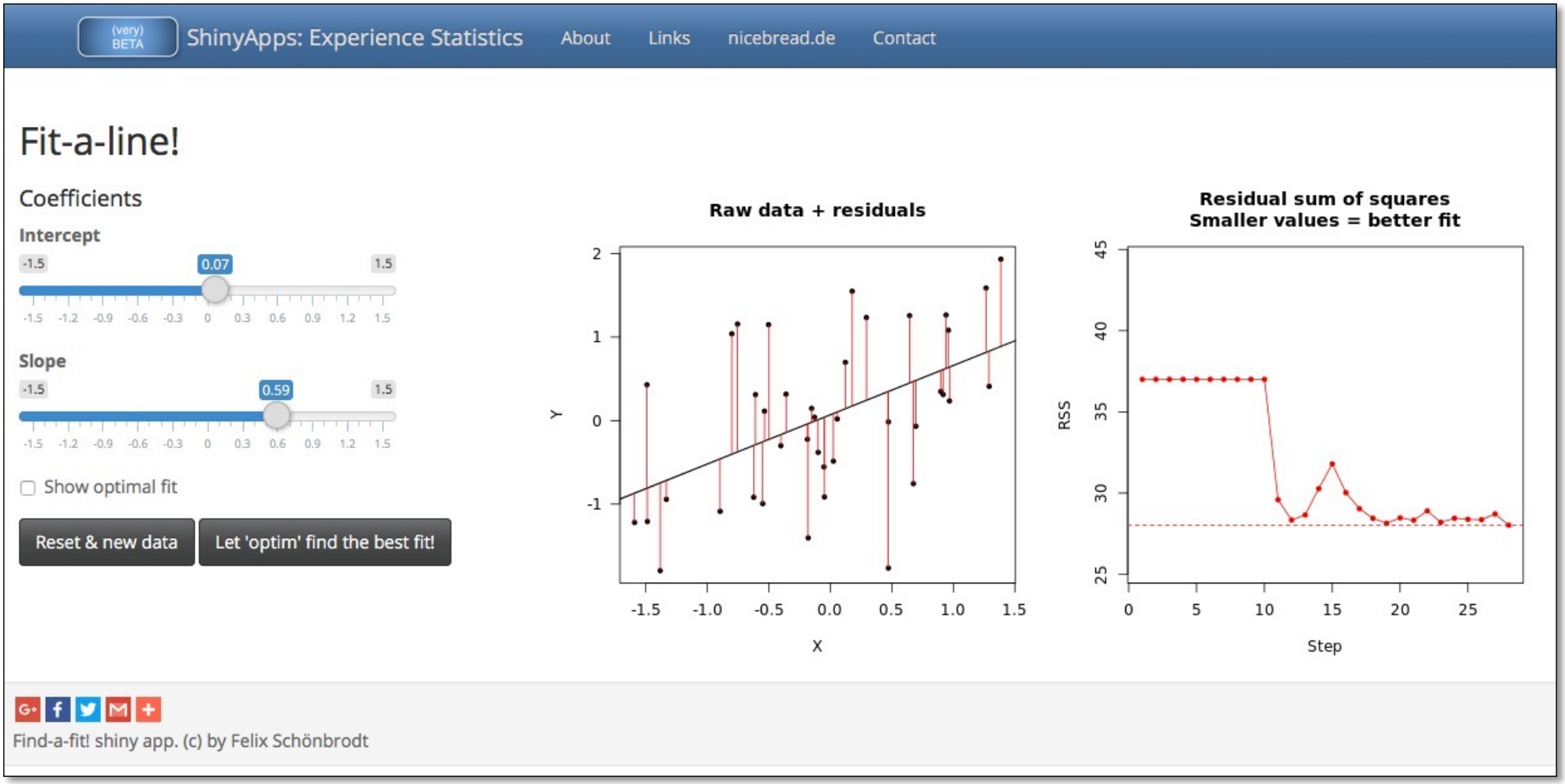
- Eine optimale Modellierung der Daten im Sinne der Regressionsanalyse besteht in der Minimierung der Fehlervariable ε: Über alle Datenpunkte betrachtet soll der Vorhersagefehler so klein wie möglich sein.
- Möglichkeit 1: Minimierung der **quadrierten Abweichung** von vorhergesagtem und tatsächlichen Wert (RSS = residual sum of squares):

$$\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

- → ermöglicht analytische Berechnung der optimalen Modellparameter
- Aber: man könnte z.B. auch die **absolute Abweichung** optimieren. Oder die **maximale Abweichung** minimieren, ...

Wie findet man die optimal passende Gerade?

Iterativ durch Parameteroptimierung: https://shiny.psy.lmu.de/felix/lmfit/



Exkurs: Analytische Lösung der Least-Squares-Optimierung

Wir suchen die Werte für A und B, für die f(A,B) minimiert wird:

$$f(A,B) = \sum_{i=1}^{n} [Y_i - (A + BX_i)]^2$$

$$\sum_{i=1}^{n} \frac{1}{n}(-1)(2)(Y_i - A - BX_i) = 0$$

$$-2 \cdot \frac{1}{n} \sum_{i=1}^{n} (Y_i - A - BX_i) = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{n}(nA + B \sum_{i=1}^{n} X_i)$$

$$\bar{Y} = A + B\bar{X}$$

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

$$\sum_{i=1}^{n} \frac{1}{n}(-X_i)(2)(Y_i - A - BX_i) = 0$$

$$B \sum_{i=1}^{n} X^2{}_{ii} = \sum_{i=1}^{n} X_i Y_i - \left((\bar{Y} - B\bar{X}) \sum_{i=1}^{n} X_i \right)$$

$$B\left(\sum_{i=1}^{n} X_i^2 - n \cdot \bar{X}^2 \right) = \sum_{i=1}^{n} X_i Y_i - n \cdot \bar{X} \cdot \bar{Y}$$

$$B = \frac{\sum_{i=1}^{n} X_i Y_i - n \cdot \bar{X} \cdot \bar{Y}}{\sum_{i=1}^{n} X_i^2 - n \cdot \bar{X}^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Ausmultiplikation von $\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$, ergibt

$$\sum_{i=1}^{n} X_i Y_i - n \cdot \bar{X} \cdot \bar{Y}$$

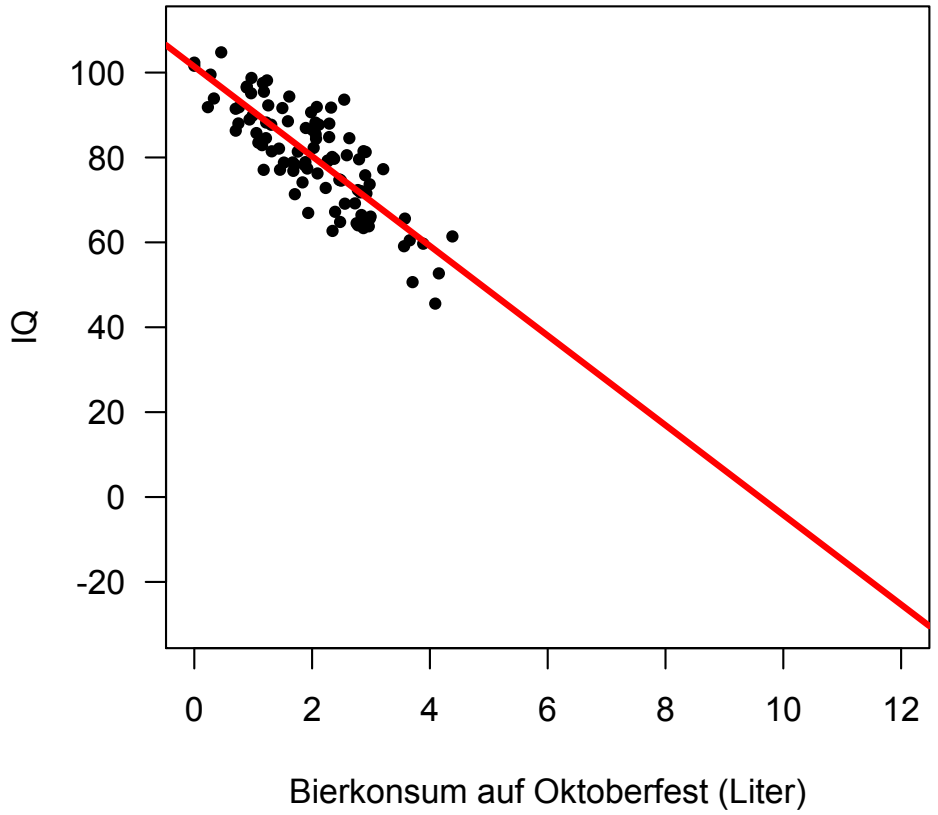
Nebenrechnung:

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^{n} (X^2{}_i - 2X_i\bar{X} + \bar{X}^2) =$$

$$S^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X^2{}_i - 2\bar{X} \sum_{i=1}^{n} X_i + \bar{X}^2 \sum_{i=1}^{n} 1 \right) =$$

$$nS^2 = \sum_{i=1}^{n} X^2{}_i - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^{n} X^2{}_i - n\bar{X}^2$$

Vorsicht bei Extrapolation!



Nach 12 Litern Bierkonsum ...

$$\widehat{IQ} = 101 - 10.1 * x_i$$
$$= 101 - 10.1 * 12 = -20.2$$

- Extrapolation = Daten außerhalb des ursprünglichen Wertebereichs der Prädiktoren vorhersagen
- Extrapolation funktioniert nur unter typischerweise sehr unrealistischen Zusatzannahmen.
- Daher: Im Normalfall keine Extrapolation machen.

Annahmen der (multiplen) linearen Regression

1. Die Zufallsvariablen hängen linear zusammen
2. Alle ε_i sind unabhängig voneinander
   → Diese Annahme wird z.B. in hierarchischen Datenstrukturen verletzt!
3. Die ε_i folgen einer Normalverteilung mit Erwartungswert Null und konstanter Varianz (Homoskedastizität):

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

→ Achtung: das Modell nimmt an, dass **die Fehlervariablen** normalverteilt sind – nicht die Prädiktoren, und auch nicht die abhängige Variable!

→ diese Annahme ist für die Schätzung der Modellparameter meist irrelevant (Gelman & Hill, 2007); eine Verletzung ist auch für die Schätzung der Standardfehler meist unkritisch.

Kategoriale Prädiktoren: Dummy vs. Effektkodierung

Dichotome Variable: Münchner*in

- Bei kategorialen Variablen wird die Ausprägung numerisch kodiert
- Verschiedene Kodierungsschemata am Beispiel „Münchner*in“:
- Dummy-Kodierung:
  - z.B.: keine Münchner*in = 0, Münchner*in = 1
  - Die Gruppe mit der Kodierung 0 ist die Referenzgruppe
- Effektkodierung:
  - z.B.: keine Münchner*in = -1, Münchner*in = 1
  - Bei zwei Gruppen: immer symmetrisch um Null herum definieren
- Die Zahlen sind arbiträr – man könnte auch keine Münchner*in= 0 und Münchner*in = 3 machen. Die Modelle sind mathematisch alle äquivalent, die Interpretation der Koeffizienten aber unterschiedlich!

Mehr Info unter https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-effect-coding/

$$IQ_i = \beta_0 + \beta_1 \cdot Bier_i + \beta_2 \cdot Muc_i + \varepsilon_i$$

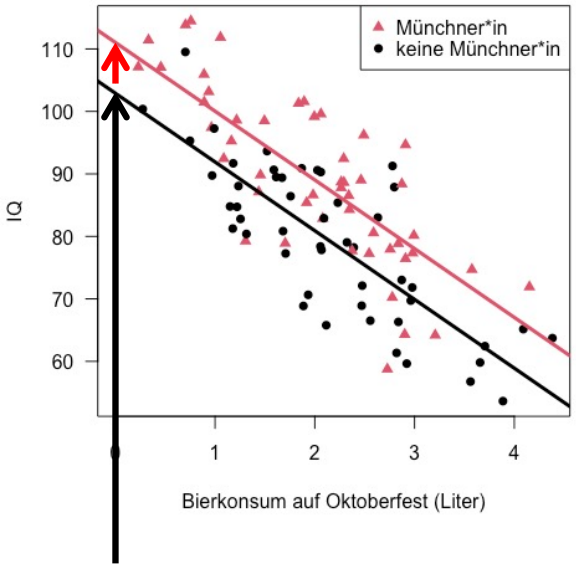
Dummykodierung
!muc=0, muc=1

β₀ β₁ β₂

```
Coefficients:
(Intercept)        bier   muc_dummy1
    102.919     -11.002        8.123
```

Slope beider Gruppen

Unterschied in den Intercepts



Intercept =
Intercept der Referenzgruppe

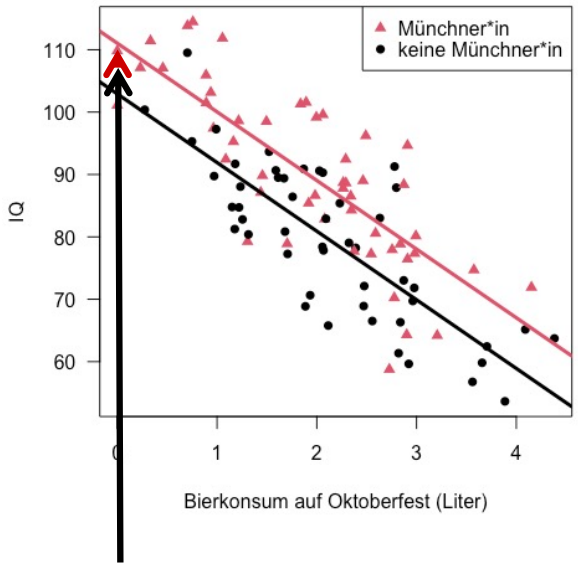
Effektkodierung
!muc=-1, muc=1

β₀ β₁ β₂

```
Coefficients:
(Intercept)        bier   muc_effect
    106.980     -11.002        4.062
```

Slope beider Gruppen

Abweichung beider Gruppen (+/-) vom grand intercept



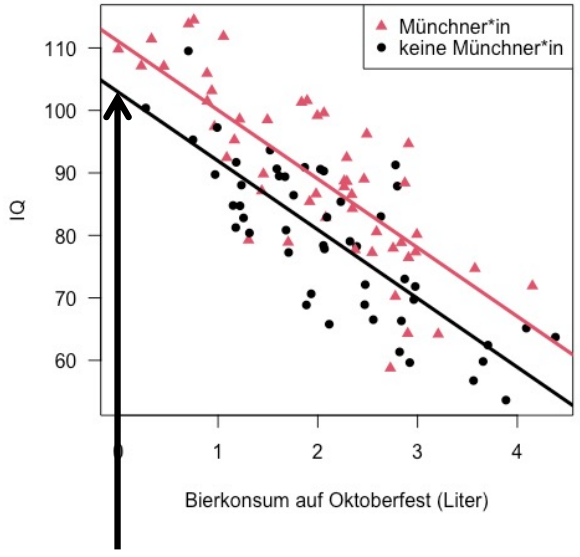
Intercept =
gemitteltes Intercept über beide Gruppen
(ohne Gewichtung nach Fallzahl beider Gruppen)

Seltsame Dummykodierung
!muc=0, muc=3

β₀ β₁ β₂

```
Coefficients:
(Intercept)        bier   muc_dummy2
    102.919     -11.002        2.708
```

Wenn die dummy-kodierte Variable eine Einheit hoch geht, geht das Intercept 2.7 Einheiten hoch. Nur steht bei muc immer 3, also ist das Intercept 3*2.7 = 8.1 höher!



Intercept =
Intercept der Referenzgruppe

$$IQ_i = \beta_0 + \beta_1 \cdot Bier_i + \beta_2 \cdot Muc_i + \varepsilon_i$$

Dummykodierung
!muc=0, muc=1

```
Coefficients:
(Intercept)         bier    muc_dummy1
    102.919      -11.002         8.123
```

Eine Münchner*in trinkt 3.4 Liter Bier.

Wie hoch ist ihr IQ?

$bier = 3.4$
$muc\_dummy1 = 1$

$\widehat{IQ}_i = 102.9 - 11.00 \cdot 3.4 + 8.12 \cdot 1 = 73.6$

Effektkodierung
!muc=-1, muc=1

```
Coefficients:
(Intercept)         bier    muc_effect
    106.980      -11.002         4.062
```

Eine Münchner*in trinkt 3.4 Liter Bier.

Wie hoch ist ihr IQ?

$bier = 3.4$
$muc\_effect = 1$

$\widehat{IQ}_i = 106.98 - 11.00 \cdot 3.4 + 4.06 \cdot 1 = 73.6$

Seltsame Dummykodierung
!muc=0, muc=3

```
Coefficients:
(Intercept)         bier    muc_dummy2
    102.919      -11.002         2.708
```

Eine Münchner*in trinkt 3.4 Liter Bier.

Wie hoch ist ihr IQ?

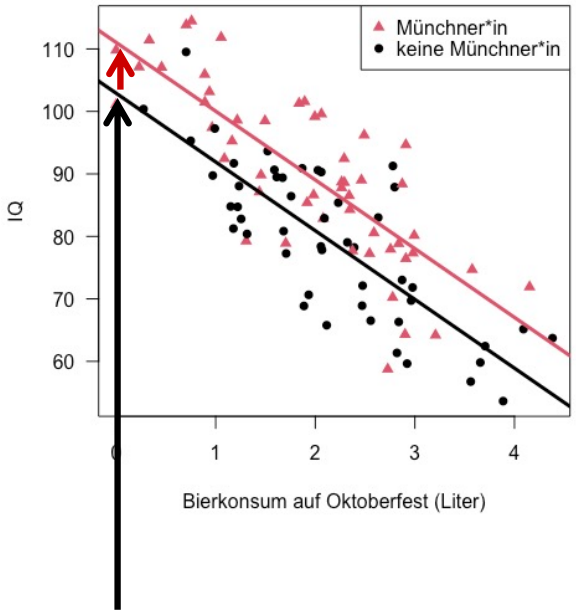
$bier = 3.4$
$muc\_dummy2 = 3$

$\widehat{IQ}_i = 102.9 - 11.00 \cdot 3.4 + 2.71 \cdot 3 = 73.6$

Wechsel der Referenzgruppe

Dummykodierung !muc=0, muc=1

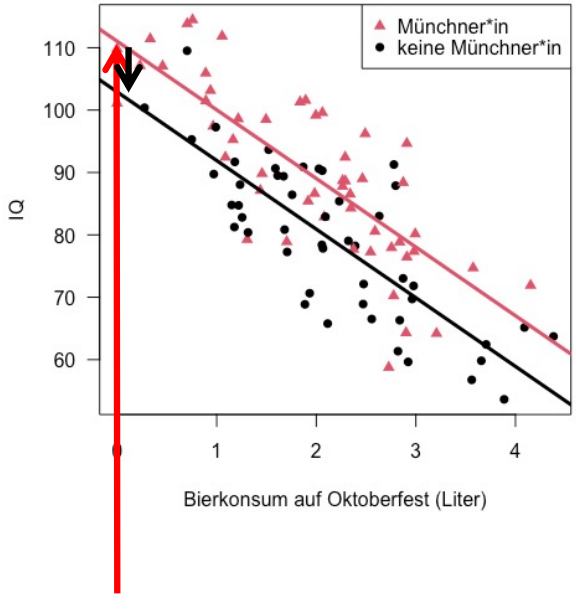
```
Coefficients:
(Intercept)         bier   muc_dummy1
    102.919      -11.002        8.123
```



Intercept =
Intercept der Referenzgruppe (!Muc)

Dummykodierung !muc=1, muc=0

```
Coefficients:
   (Intercept)          bier  muc_dummy1_rev
       111.042       -11.002           -8.123
```



Intercept =
Intercept der Referenzgruppe (Muc)

Dichotome Variable: Münchner*in

- Dummy-Kodierung:
  - z.B.: keine Münchner*in = 0, Münchner*in = 1
  - Die Gruppe mit der Kodierung 0 ist die Referenzgruppe
  - Intercept = Achsenabschnitt der Referenzgruppe
  - Das Regressionsgewicht für `muc` ist der Unterschied im Intercept der anderen Gruppe im Vgl. zur Referenzgruppe
- Effektkodierung:
  - z.B.: keine Münchner*in = -1, Münchner*in = 1
  - Intercept = grand intercept, über beide Gruppen hinweg (ungewichtet falls ungleiche Gruppengrößen – d.h., der Mittelwert der beiden Gruppenmittelwerte, egal wie groß die Gruppen sind)
  - Das Regressionsgewicht für `muc` ist der Unterschied beider Gruppen (+ oder -) im Vergleich zum grand intercept.

Dichotome Variable: Interaktion mit slope (a.k.a. „Moderierte Regression“)

$$IQ_i = \beta_0 + \beta_1 Bier_i + \beta_2 Muc_i + \beta_3 Bier_i \cdot Muc_i + \varepsilon_i$$
$$IQ_i = \beta_0 + \beta_2 Muc_i + (\beta_1 + \beta_3 Muc_i) \cdot Bier_i + \varepsilon_i$$

Dummy-Kodierung: !muc = 0, muc = 1

$$Muc = 0$$
$$IQ_i = \beta_0 + \beta_2 \cdot 0 + (\beta_1 + \beta_3 \cdot 0) \cdot Bier_i + \varepsilon_i$$
$$IQ_i = \beta_0 + \beta_1 \cdot Bier_i + \varepsilon_i$$

- β_0: Intercept der Referenzgruppe
- β_1: Slope der Referenzgruppe

$$Muc = 1$$
$$IQ_i = \beta_0 + \beta_2 \cdot 1 + (\beta_1 + \beta_3 \cdot 1) \cdot Bier_i + \varepsilon_i$$
$$IQ_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot Bier_i + \varepsilon_i$$

- β_0: Intercept der Referenzgruppe
- β_2: Unterschied im Intercept der anderen Gruppe
- β_1: Slope der Referenzgruppe
- β_3: Unterschied im Slope der anderen Gruppe

Dichotome Variable: Interaktion mit slope

$$IQ_i = \beta_0 + \beta_1 Bier_i + \beta_2 Muc_i + \beta_3 Bier_i \cdot Muc_i + \varepsilon_i$$
$$IQ_i = \beta_0 + \beta_2 Muc_i + (\beta_1 + \beta_3 Muc_i) \cdot Bier_i + \varepsilon_i$$

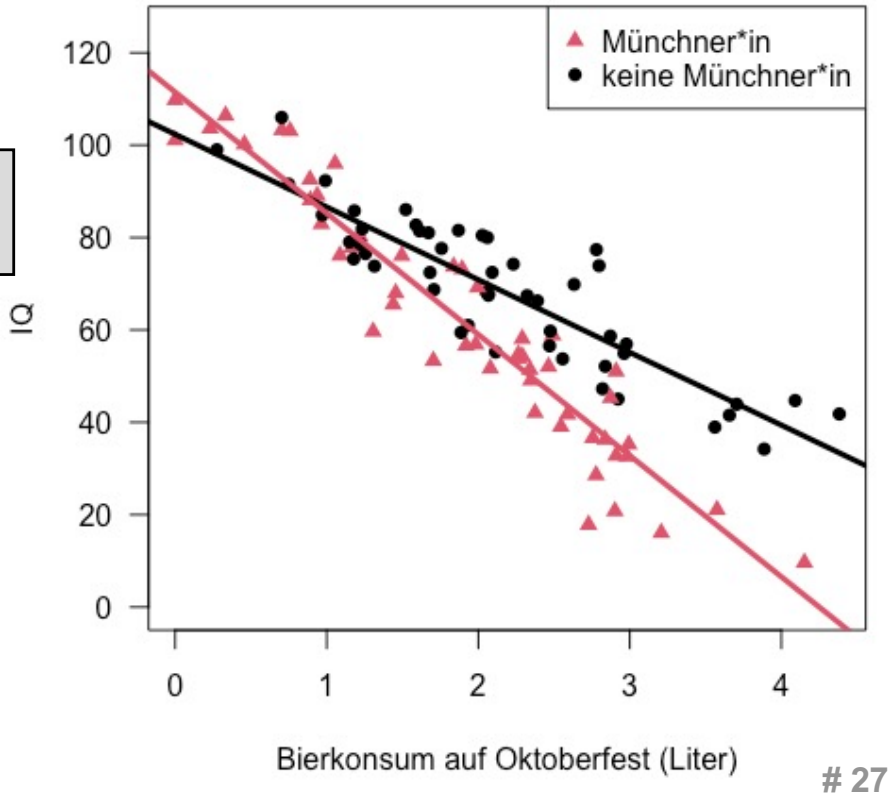
Dummy-Kodierung: !muc = 0, muc = 1

```
Coefficients:
    (Intercept)            bier     muc_dummy1  bier:muc_dummy1
        102.364          -15.744           9.124          -10.491
```

- Intercept !Muc
- Slope !Muc
- Unterschied Intercept Muc zu !Muc
- Unterschied slope Muc zu !Muc

$$Muc = 0$$
$$\widehat{IQ}_i = 102.4 - 15.7 \cdot Bier_i$$

$$Muc = 1$$
$$\widehat{IQ}_i = (102.4 + 9.1) + (-15.7 - 10.5) \cdot Bier_i$$



Dichotome Variable: Interaktion mit slope

$$IQ_i = \beta_0 + \beta_1 Bier_i + \beta_2 Muc_i + \beta_3 Bier_i \cdot Muc_i + \varepsilon_i$$
$$IQ_i = \beta_0 + \beta_2 Muc_i + (\beta_1 + \beta_3 Muc_i) \cdot Bier_i + \varepsilon_i$$

Effektkodierung: !muc = -1, muc = 1

Weder !muc noch muc: Effekt unabhängig von dichotomer Variable.

$$Muc = 0$$
$$IQ_i = \beta_0 + \beta_2 \cdot 0 + (\beta_1 + \beta_3 \cdot 0) \cdot Bier_i + \varepsilon_i$$
$$IQ_i = \beta_0 + \beta_1 \cdot Bier_i + \varepsilon_i$$

- β_0: Gesamt-Intercept
- β_1: Gesamt-Slope

!muc | muc

$$Muc = -1 \; oder \; Muc = +1$$
$$IQ_i = \beta_0 + \beta_2 \cdot (\pm 1) + (\beta_1 + \beta_3 \cdot (\pm 1)) \cdot Bier_i + \varepsilon_i$$
$$IQ_i = (\beta_0 \pm \beta_2) + (\beta_1 \pm \beta_3) \cdot Bier_i + \varepsilon_i$$

- β_0: Gesamt-Intercept
- β_2: Unterschied vom Gesamt-Intercept (*1 oder *-1)
- β_1: Gesamt-Slope
- β_3: Unterschied vom Gesamt-Slope (*1 oder *-1)

Dichotome Variable: Interaktion mit slope

$$IQ_i = \beta_0 + \beta_1 Bier_i + \beta_2 Muc_i + \beta_3 Bier_i \cdot Muc_i + \varepsilon_i$$
$$IQ_i = \beta_0 + \beta_2 Muc_i + (\beta_1 + \beta_3 Muc_i) \cdot Bier_i + \varepsilon_i$$

- Effektkodierung: !muc= -1, muc = 1

```
Coefficients:
    (Intercept)            bier      muc_effect  bier:muc_effect
        106.926         -20.989           4.562           -5.246
```

Grand intercept

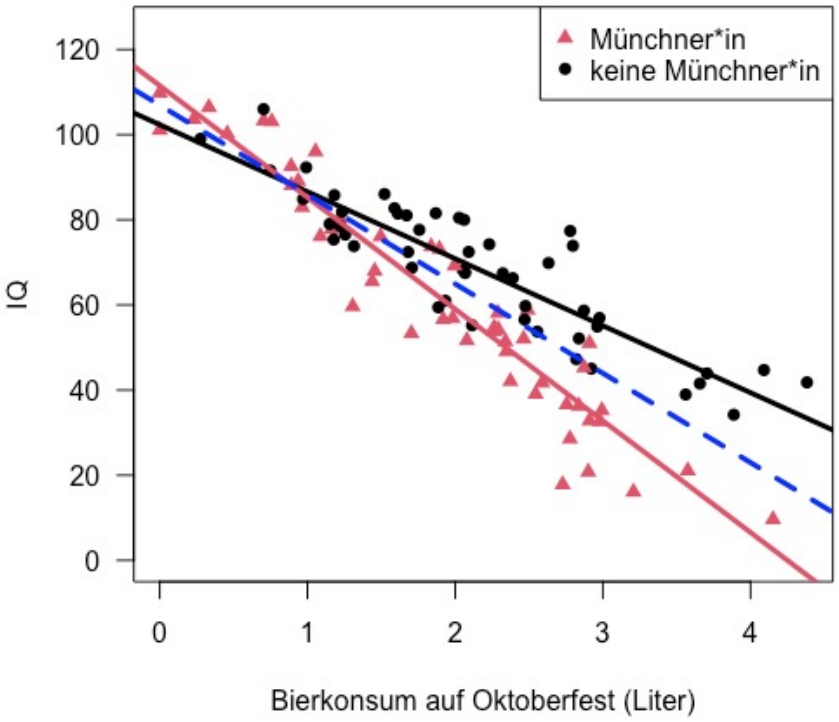
Grand slope

Unterschied Intercept !muc (-1) und muc (+1) zu grand intercept

Unterschied slope !muc (-1) und muc (+1) zu grand slope

$$Muc = 0$$
$$\widehat{IQ}_i = 106.9 - 21.0 \cdot Bier_i$$

$$Muc = \pm 1$$
$$\widehat{IQ}_i = (106.9 \pm 4.6) + (-21.0 \pm (-5.2)) \cdot Bier_i$$



Äquivalente Modelle

Dummykodierung
!muc=0, muc=1

```
Call:
lm(formula = IQ ~ bier * muc_dummy1)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0866  -5.3774  -0.0536   5.2388  18.7691

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      102.364      2.914  35.123  < 2e-16 ***
bier             -15.744      1.247 -12.624  < 2e-16 ***
muc_dummy1         9.124      3.863   2.362   0.0202 *
bier:muc_dummy1  -10.491      1.720  -6.099 2.23e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.06 on 96 degrees of freedom
Multiple R-squared:  0.8748, Adjusted R-squared:  0.8708
F-statistic: 223.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

Effektkodierung
!muc=-1, muc=1

```
Call:
lm(formula = IQ ~ bier * muc_effect)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0866  -5.3774  -0.0536   5.2388  18.7691

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      106.926      1.931  55.365  < 2e-16 ***
bier             -20.989      0.860 -24.405  < 2e-16 ***
muc_effect         4.562      1.931   2.362   0.0202 *
bier:muc_effect   -5.246      0.860  -6.099 2.23e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.06 on 96 degrees of freedom
Multiple R-squared:  0.8748, Adjusted R-squared:  0.8708
F-statistic: 223.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

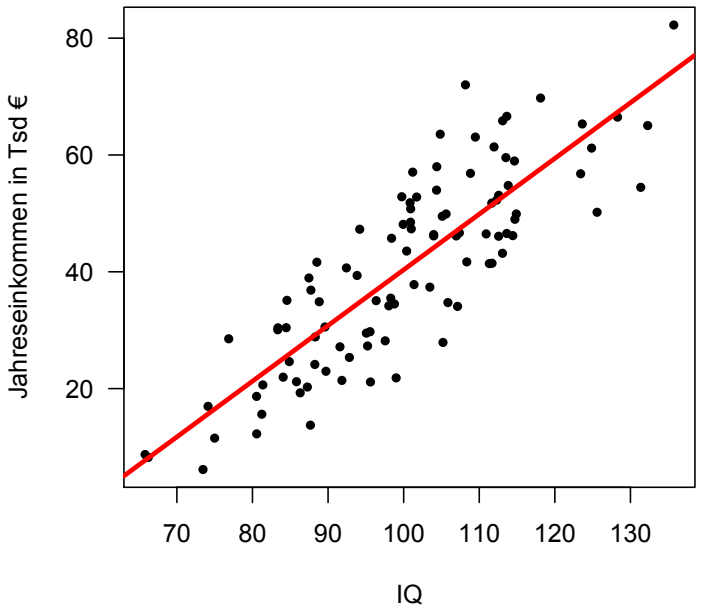
Unterschiedliche Koeffizienten, aber identisches R^2, F-Statistik, Residuen!

Zentrierung von Prädiktorvariablen

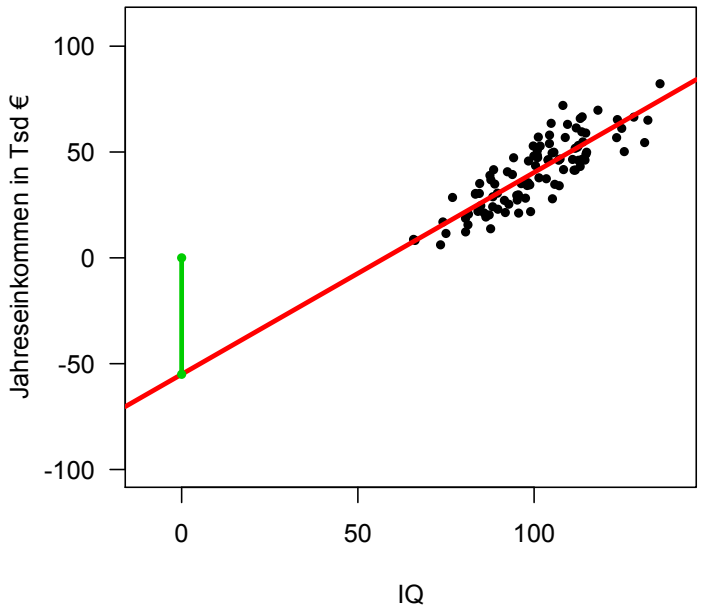
Zentrierung

- Der Achsenabschnitt beschreibt den vorhergesagten Wert, wenn alle Prädiktorvariablen 0 sind
- Ist die 0 ein sinnvoller Wert bei einem Prädiktor?

```
Coefficients:
(Intercept)          IQ
   -55.0262      0.9536
```

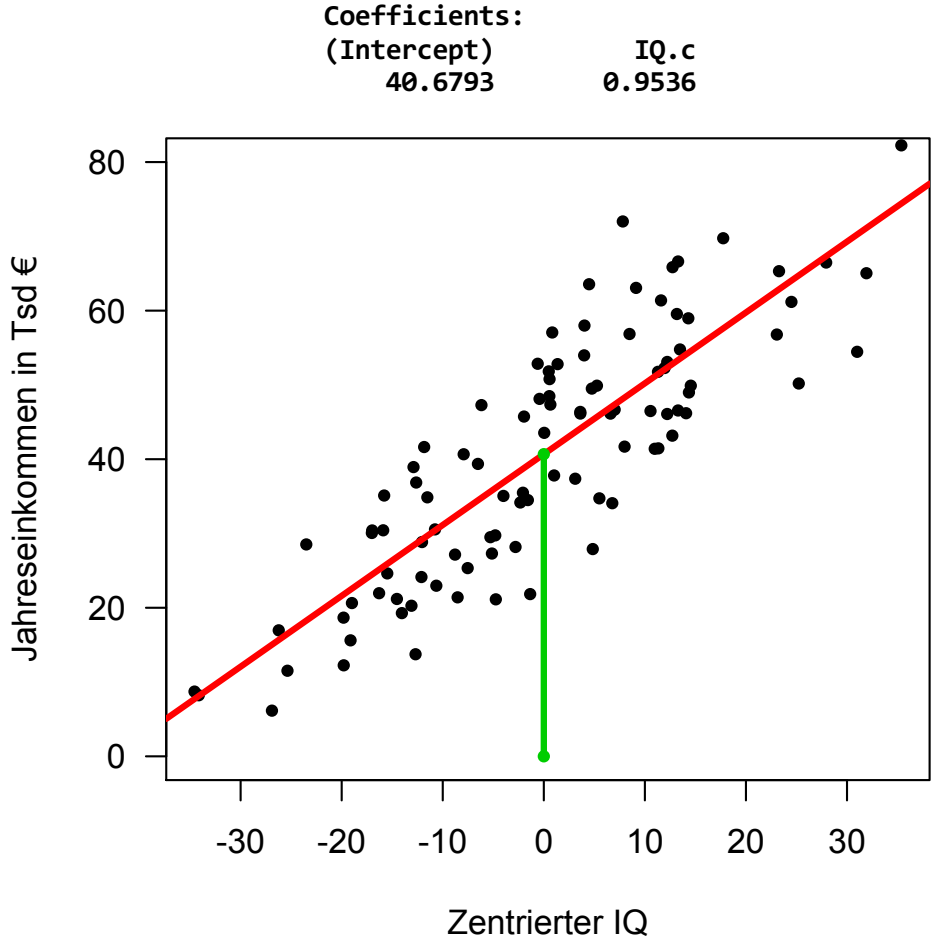


Vorhergesagtes Einkommen bei IQ = 0: -55.026 €



Zentrierung

- Faustregel: Zentriere Prädiktorvariablen immer so, dass die Null einen sinnvollen Wert beschreibt. Möglichkeiten:
  - Auf Stichprobenmittelwert zentrieren (am häufigsten). Dann beschreibt das Intercept den vorhergesagten Wert einer durchschnittlichen Person
  - Auf Populationsmittelwert zentrieren (aus Normdaten)
  - Auf das Minimum der Stichprobe zentrieren
  - Bei Zeitreihen: Auf ersten Messzeitpunkt zentrieren
  - Bei Likert-Skalen: Auf den semantischen Mittelpunkt zentrieren:
    -3 = trifft gar nicht zu;
    0 = unentschieden;
    +3 = trifft voll zu



Äquivalente Modelle

Unterschiedliche Zentrierungen (bzw. lineare Transformationen im Allgemeinen) von Prädiktoren führen zu mathematisch äquivalenten Modellen!
→ Unterschiedliche Koeffizienten, aber identisches R^2, F-Statistik, Residuen!

Dummykodierung
!muc=0, muc=1

Bier zentriert auf Gesamtmittelwert

ACHTUNG: Diese Kombination macht wenig Sinn – nur aus didaktischen Gründen gewählt!

```
Call:
lm(formula = IQ ~ bier.c * muc_dummy1)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0866  -5.3774  -0.0536   5.2388  18.7691

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       80.5641     1.1607  69.410  < 2e-16 ***
bier.c           -10.7437     1.2471  -8.615 1.40e-13 ***
muc_dummy1         8.1272     1.6245   5.003 2.55e-06 ***
bier.c:muc_dummy1 -0.4913     1.7201  -0.286    0.776
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.06 on 96 degrees of freedom
Multiple R-squared:  0.6841, Adjusted R-squared:  0.6742
F-statistic:  69.3 on 3 and 96 DF,  p-value: < 2.2e-16
```

Effektkodierung
!muc=-1, muc=1

Bier zentriert auf Mittelwert der Münchner*innen

ACHTUNG: Diese Kombination macht wenig Sinn – nur aus didaktischen Gründen gewählt!

```
Call:
lm(formula = IQ ~ bier.c.w * muc_effect)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0866  -5.3774  -0.0536   5.2388  18.7691

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          85.8711     0.8189 104.864  < 2e-16 ***
bier.c.w            -10.9893     0.8600 -12.778  < 2e-16 ***
muc_effect            4.0914     0.8189   4.996 2.62e-06 ***
bier.c.w:muc_effect  -0.2456     0.8600  -0.286    0.776
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.06 on 96 degrees of freedom
Multiple R-squared:  0.6841, Adjusted R-squared:  0.6742
F-statistic:  69.3 on 3 and 96 DF,  p-value: < 2.2e-16
```

Over- & Underfitting

Modellpassung

- Modellpassung
  - Es gibt mehrere Kandidatenmodelle, die die Zusammenhänge in der Realität beschreiben
  - (darüber hinaus gibt es unendlich viele alternative Modelle)
  - Die Modelle können unterschiedlich flexibel („komplex“) sein
  - Da wir die Realität nicht kennen, gibt es Unsicherheit darüber, welches der Modell das relativ beste ist

$$y_i = \beta_0 + \beta_1 x + \epsilon_i$$

$$y_i = \beta_1 \frac{x_1}{x_2} + \epsilon_i$$

$$y_i = \beta_0 + \beta_1 \frac{x_1}{x_2} + \beta_2 x_3 + \beta_3 \frac{1}{x_4} + \epsilon_i$$

Modellparameter

- Modellparameter = Parameter innerhalb eines konkreten Modells, die variieren können und mithilfe von Daten geschätzt werden müssen
  - In einfachen Modellen (siehe diese Vorlesung) gilt:
    Mehr Modellparameter -> Höhere Flexibilität des Modells
  - In komplexeren Modellen (siehe dieses Semester), wird die Flexibilität manchmal absichtlich durch zusätzliche Modellparameter eingeschränkt
- Schätzung der Parameter: Gegeben der vorgegebenen Modellstruktur – was sind die Parameterausprägungen, die das Modell optimal „auf die Daten einstellen“?
- Manchmal ist ein Modell nicht flexibel genug, um die Daten adäquat modellieren zu können
- Ein gewähltes Modell ist durch die freien Parameter eigentlich eine Modell**familie**

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_i = 1.2 + 0.4x_i + \epsilon_i$$

$$y_i = 3.1 - 3.2x_i + \epsilon_i$$

$$y_i = -5 + 0.05x_i + \epsilon_i$$

Modellpassung & -parameter

- „Ockhams Rasiermesser“
  - Von mehreren möglichen Erklärungen für ein und denselben Sachverhalt ist die einfachste Theorie allen anderen vorzuziehen.
  - Statistischer Kontext: Wenn zwei unterschiedliche Modelle die Daten gleich gut beschreiben, ist das „einfachere“ Modell vorzuziehen
- Aber:
  - Was bedeutet „einfacher“? Weniger Parameter, weniger Annahmen, …
  - Meistens liegt ein Trade-Off zwischen Genauigkeit und Einfachheit vor: Ockhams Rasiermesser nicht direkt anwendbar

Einfacheres Modell

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

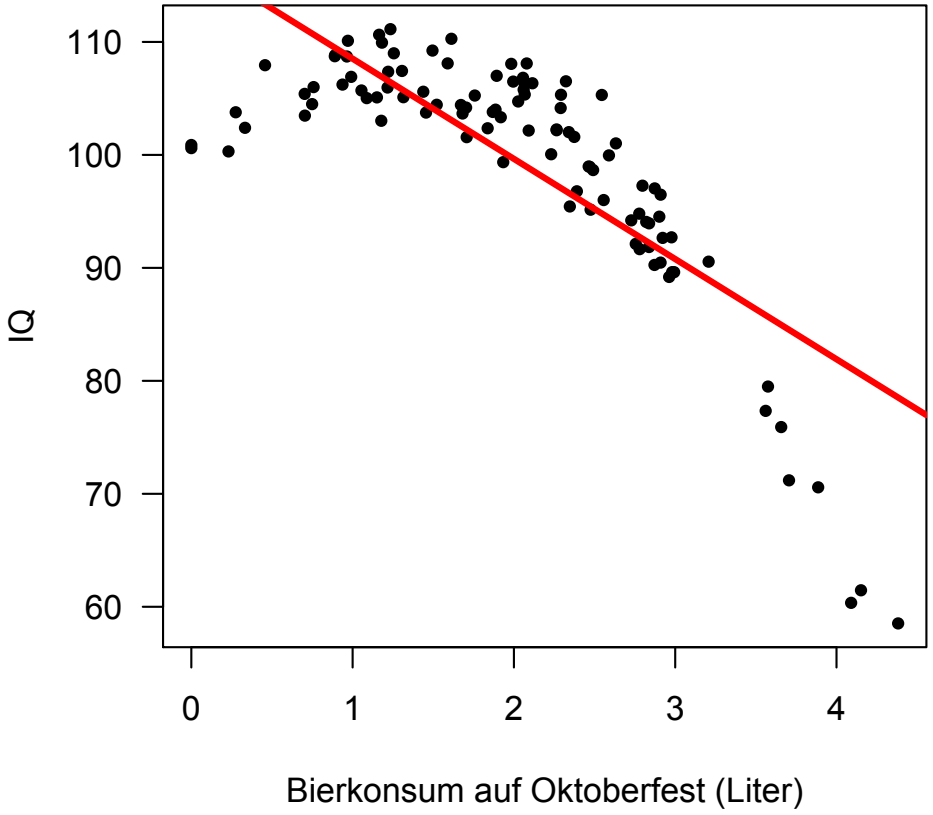
Komplexeres Modell

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

Obwohl der Term quadratisch ist (also eine Kurve abbildet), ist es ein lineares Modell, weil alle Einzelterme linear kombiniert werden

Underfitting

- Underfitting: Ein Modell hat nicht flexibel genug, um die Daten adäquat abbilden zu können
- Problem: Inadequates Abbild der Daten, Schlechte Vorhersagekraft

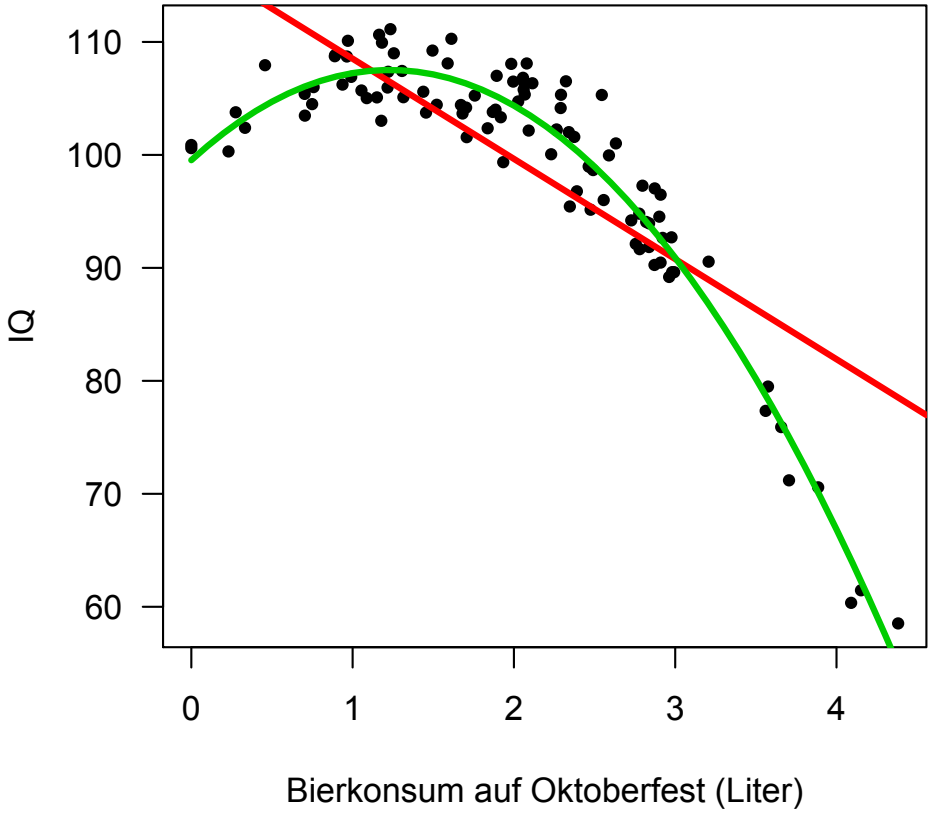



```
lm(formula = y ~ x)

Residual standard error: 6.704 on 98 degrees of freedom
Multiple R-squared:  0.6148, Adjusted R-squared:  0.6108
```

Underfitting

- Underfitting: Ein Modell hat nicht flexibel genug, um die Daten adäquat abbilden zu können
- Problem: Inadequates Abbild der Daten, Schlechte Vorhersagekraft
- Das flexiblere Modell mit quadratischem Term „passt besser“




```
lm(formula = y ~ x)

Residual standard error: 6.704 on 98 degrees of freedom
Multiple R-squared:  0.6148, Adjusted R-squared:  0.6108
```

```
lm(formula = y ~ x + I(x^2))

Residual standard error: 2.623 on 97 degrees of freedom
Multiple R-squared:  0.9416, Adjusted R-squared:  0.9404
```

Overfitting

> With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

John von Neumann (1903 - 1957)

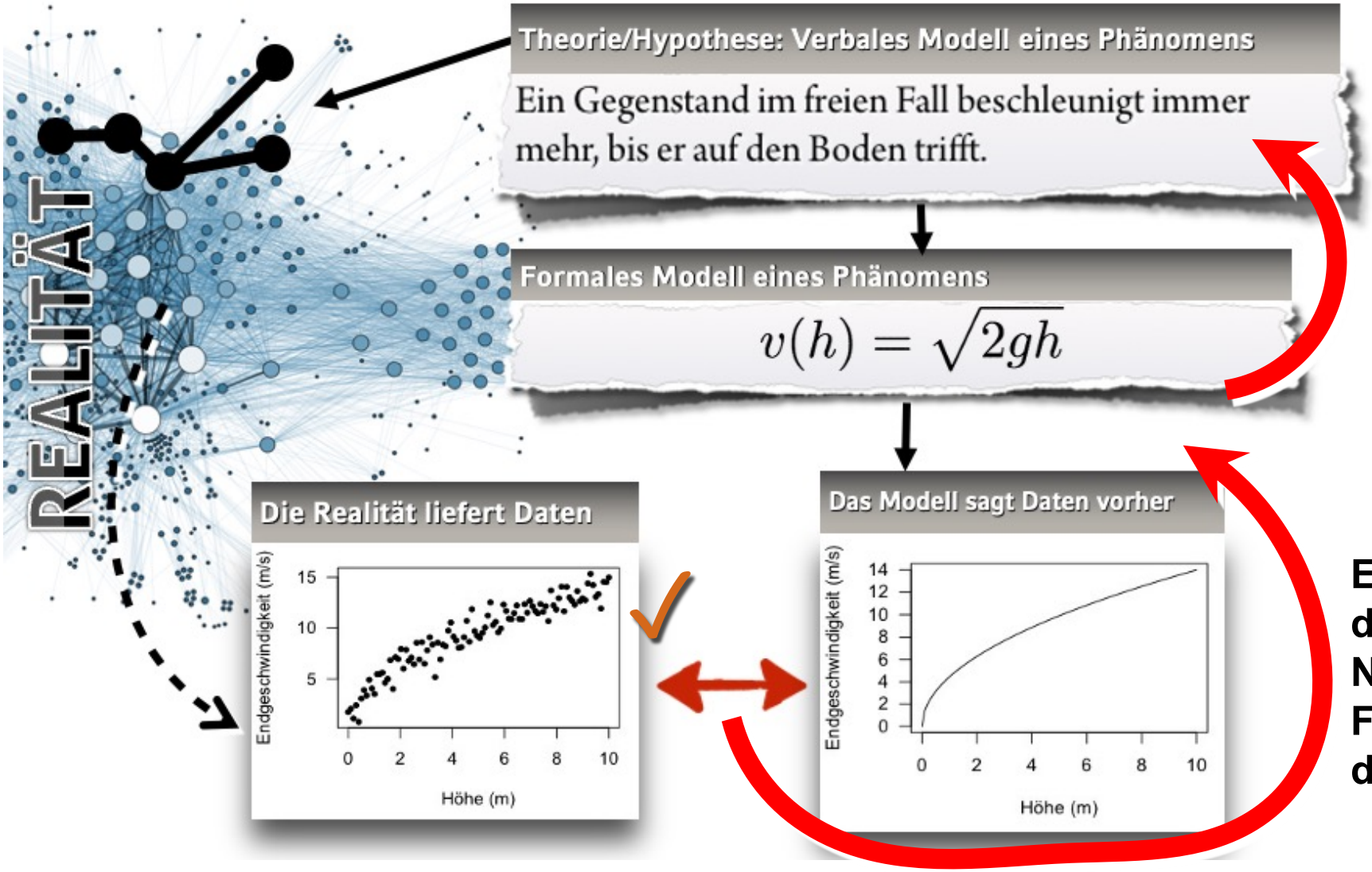
Zitat aus: Dyson, F. (2004). A meeting with Enrico Fermi. Nature, 427(6972), 297–297. Bild: Wikimedia Commons.

Probleme beim Overfitting

- Scheinbar gute Modellpassung und hohe Vorhersagequalität bei dem ursprünglichen Datensatz, auf den das Modell angepasst wurde.
- Wenn das Modell jedoch auf neue Daten aus der gleichen Population angewendet wird, ist Vorhersagekraft deutlich niedriger.
- Warum? Das flexiblere Modell passt sich offenbar auch an spezifische, unsystematische Kleinigkeiten des Datensatzes an.
- Es greift nicht nur das Signal (d.h., den systematischen Anteil) in den Daten ab, sondern fängt an, das Rauschen mit zu modellieren.
- Da das Rauschen in jedem neuen Datensatz anders sein wird, ist die Vorhersagekraft sogar schlechter als bei einem weniger flexiblen Modell.

- Mehr dazu im zweiten Teil des Semesters über Prädiktive Modellierung.

Wie können wir zu Erkenntnissen über die Welt gelangen?

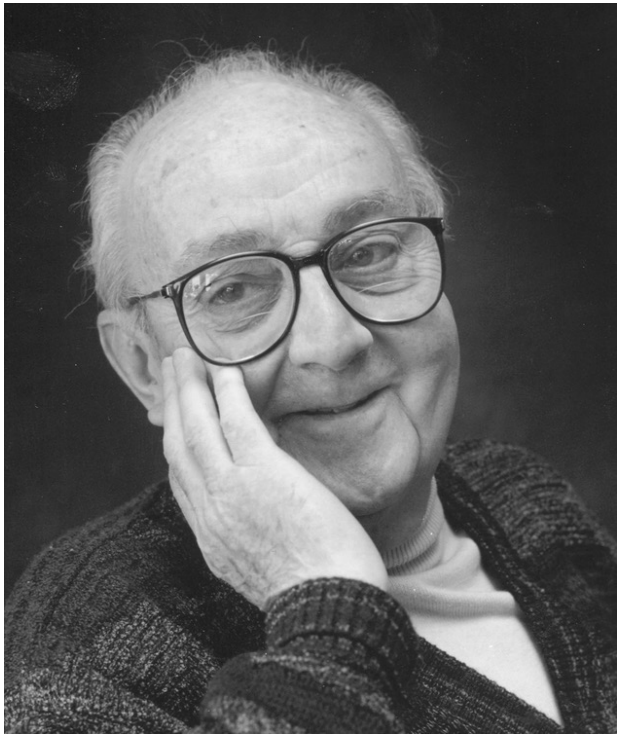


Evidenz für die Theorie

Evidenz für das Modell / Nicht-Falsifikation des Modells

Popper (1935/2005); Meehl (1967; 1990); Dienes (2008); Rouder, Morey, & Wagenmakers (2015)

All models are wrong, but some models are useful.



Georg Box, 1919 - 2013

All models are wrong, but some models are useful.

Exkurs: Ockhams Razor

Warum ist „einfacher“ besser? Was ist die Begründung für Ockham‘s Razor?

- https://aeon.co/essays/are-scientific-theories-really-better-when-they-are-simpler by Elliott Sober
- There‘s no logical justification for Ockham‘s razor (i.e., there is no proof that, even ceteris paribus, the simpler theory is true/closer to truth).
- E.g. evolution depends on random mutations, crystallized randomness: no reason to assume that simpler is (always) closer to truth.
- Can be seen as a **probabilistic** argument: Of two explanations with the same explanatory power, the simpler has a higher probability of being true / closer to the truth.
- Can be seen as a **pragmatic** argument: When the predictive success is the same, it is easier & more tractable for us to retain the simpler theory, without any practical loss. Saves cognitive capacity, less teaching, less memorizing. Do not waste time on irrelevant stuff.
- See it more as a heuristic, not a law

- Concerning in-sample fit, the more complex model always outperforms the simpler model (or is in the boundary case at least equally good). But: when we switch to out-of-sample predictions, the simpler model can be better (even if you do not belief in „true nulls“)
- Does ‚better‘ mean ‚more true‘ or ‚more useful‘? If the latter, Ockham‘s razor can be justified.
- As an antidote to Ockham: Epicurs’ Principle of Multiple Explanations states: “If several theories are consistent with the observed data, retain them all.” (http://cage.ugent.be/~ci/Epicurus.html)
- David MacKay’s online book ITILA (http://www.inference.phy.cam.ac.uk/itila) chapter 28 (http://www.inference.phy.cam.ac.uk/mackay/itprnn/ps/343.355.pdf) gives the clearest justification for Ockham‘s razor I have seen, as a simple consequence of the laws of probability.