ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

DMATH
Department of Mathematics

Transformationen und multiple lineare Regression

für D-UWIS, D-ERDW und D-AGRL – SS15

Einfache lineare Regression

- $Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \;\; iid$
- $\hat{\beta}_0, \hat{\beta}_1$ minimieren $\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$
- $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}, \;\; \widehat{\beta_0} = \bar{Y}_n - \widehat{\beta_1}\bar{x}_n, \;\; \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} R_i^2$
- $\widehat{\beta_i} \sim \mathcal{N}(\beta_i, \sigma_{\beta_i}^2)$ und man kann zeigen: $\frac{\widehat{\beta_k} - \beta_k}{s.e.(\widehat{\beta_k})} \sim t_{n-2}$

- Modell: $Y_i = \beta_0 + \beta_1 x_i + E_i, E_i \sim \mathcal{N}(0, \sigma^2)\ \ i.i.d.$
- Modell: $Y_i = -19.46 + 5.86 \cdot x_i + E_i, E_i \sim \mathcal{N}(0, 5.43^2)\ \ i.i.d$

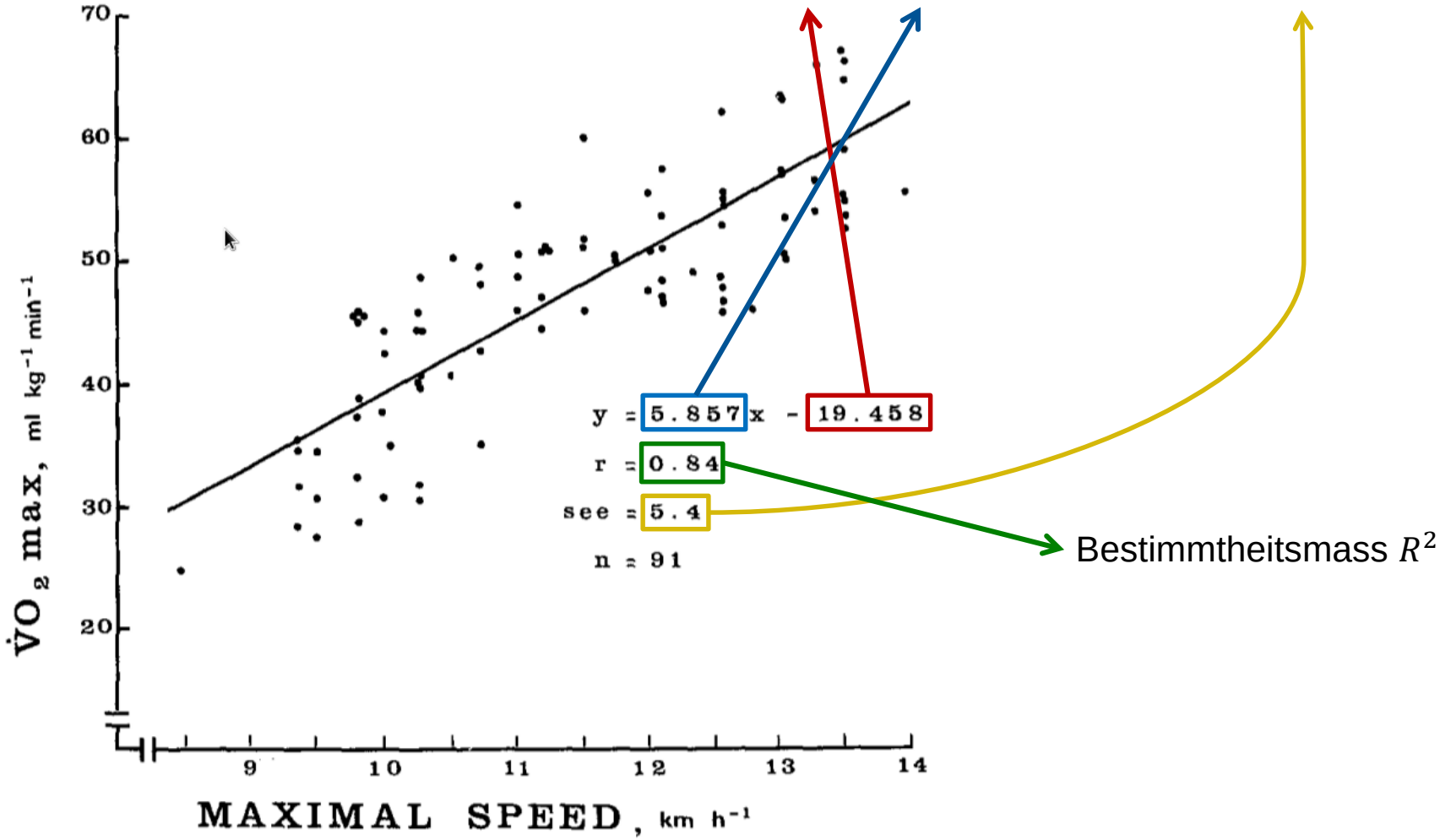
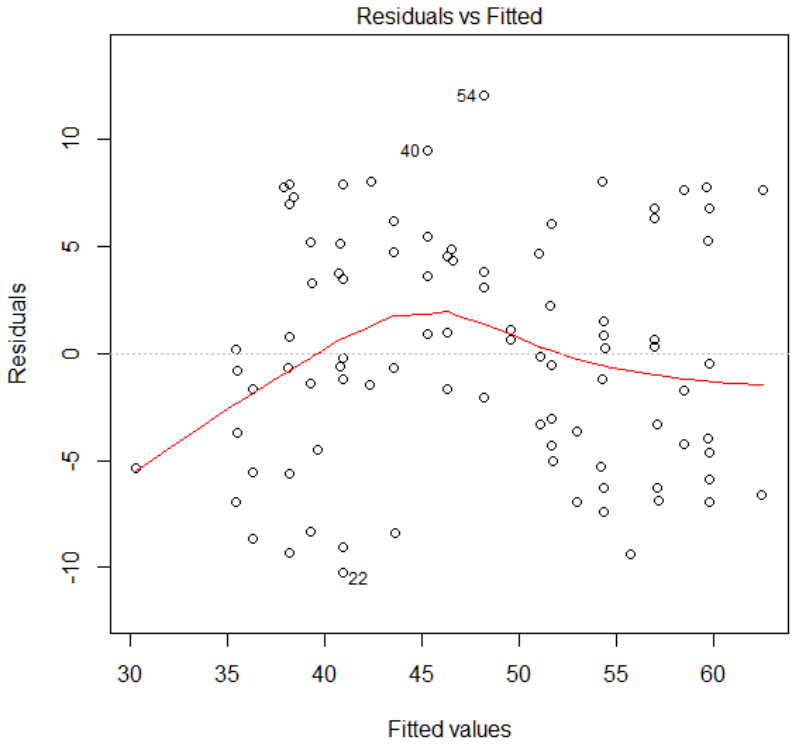
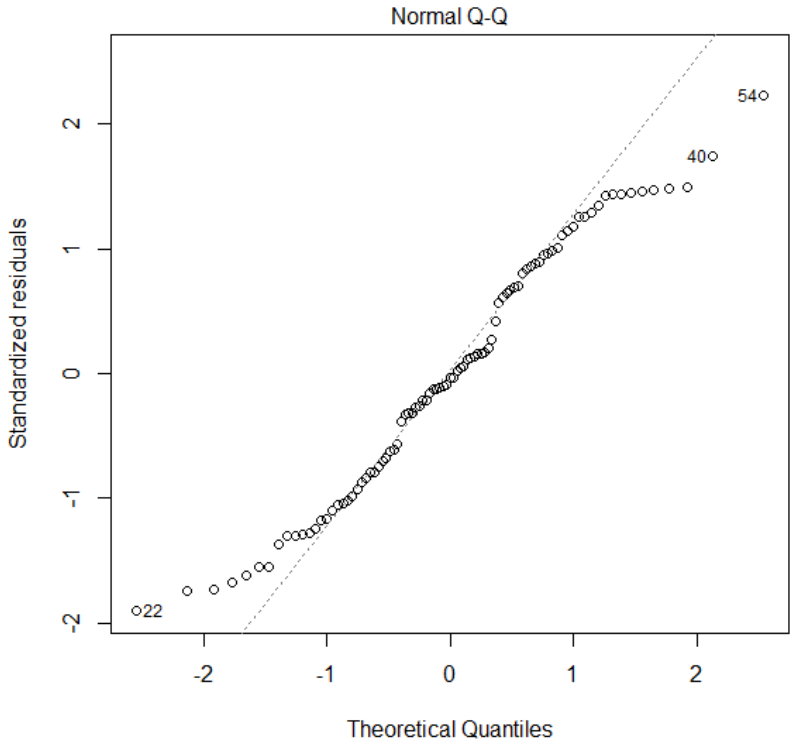


Fig. 2. $\dot{V}O_2$ max as a function of the maximal speed achieved in the 20-m shuttle run test for a total sample of 91 adult subjects. Each point in this figure represents maximal effort

Residuenanalyse





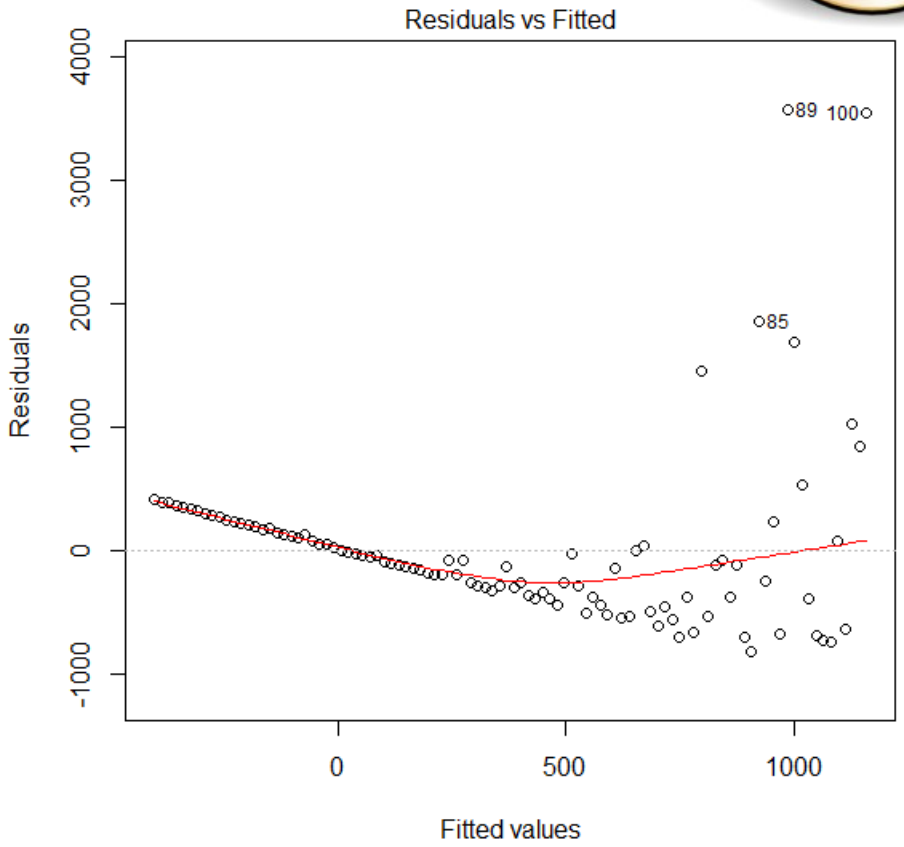
Tukey-Anscombe Plot

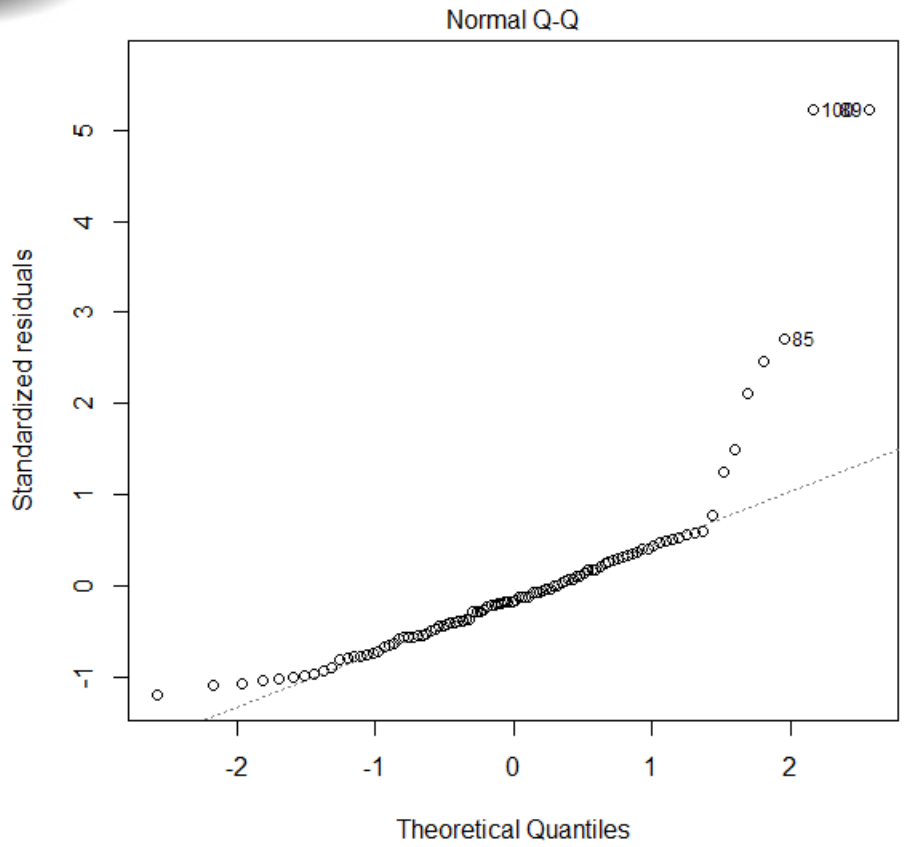
- streuen um “null”
- etwa gleich viel Streuung auf “beiden Seiten”

QQ Plot

- ein bisschen kurzschwänzig
- ist aber gerade noch OK

Residuenanalyse





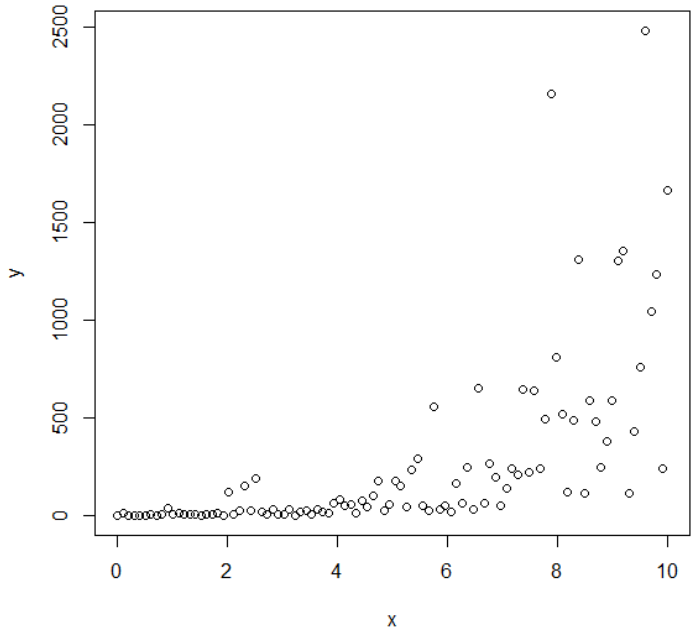
Lernziele heute

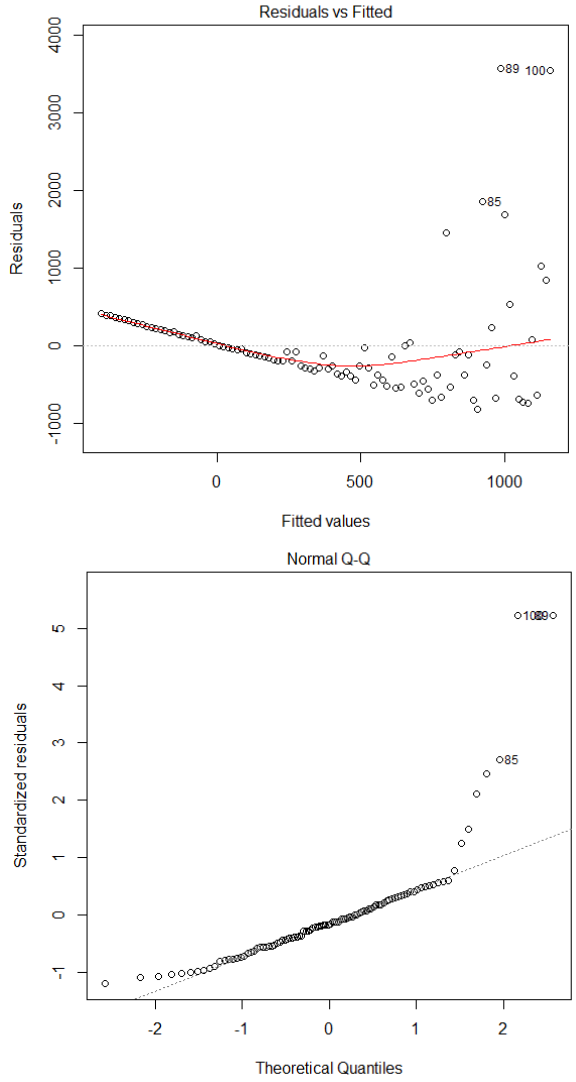
- Transformation von Daten
- Multiple lineare Regression
- Korrelation ≠ Kausalität

Hausaufgaben

- Skript: Kapitel 5.3 lesen
- Serie 13 lösen
- Quiz 13 bearbeiten

Schlechte Residuen





- Daten sind nicht normalverteilt
- Fehler ist nicht konstant

…falls Residuen schlecht aussehen

- oft hilft es x und/oder y zu transformieren

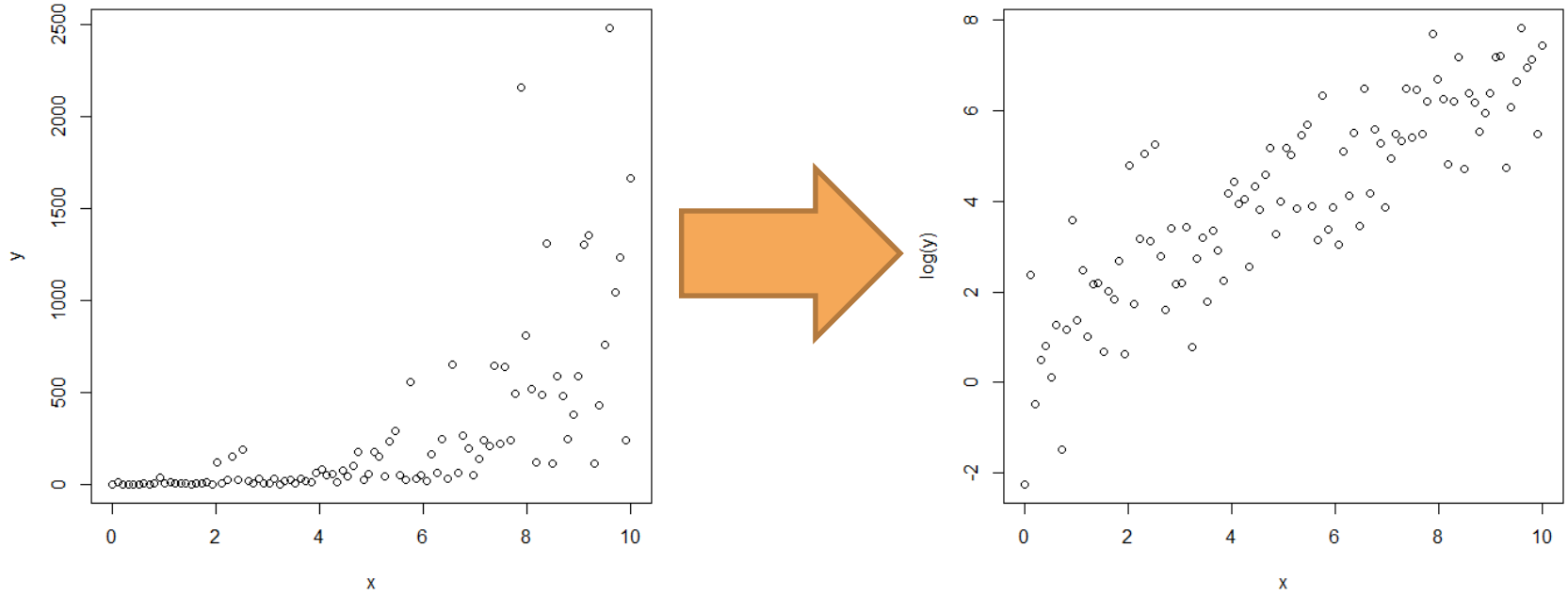
Achtung: Interpretation der neuen Parameter

- Bsp: $\log(y)$ statt y
  - vorher: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,
    wenn x durch $x + 1$ ersetzt wird, ändert sich Y zu $Y + \beta_1$

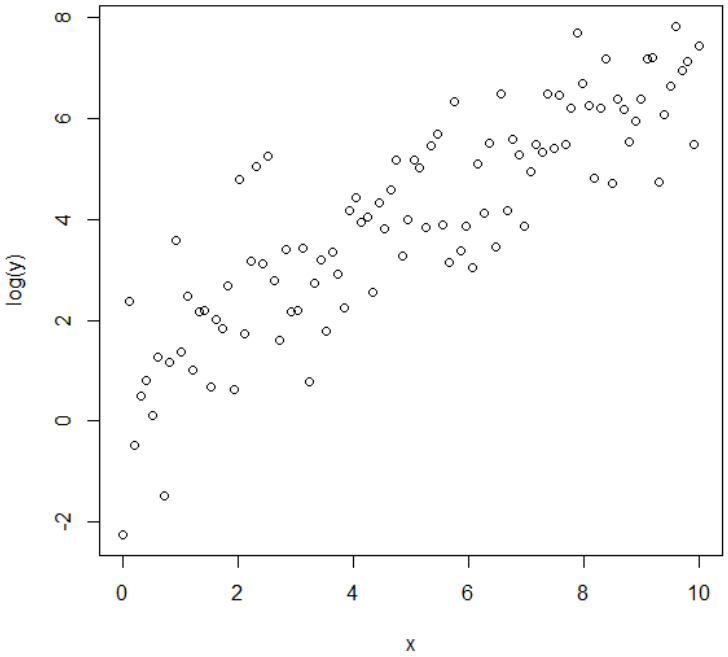
    x um eine Einheit erhöhen...

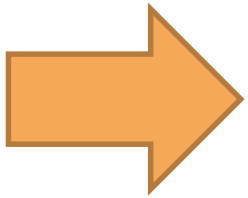
  - nachher: $\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i \leftrightarrow Y_i = \exp(\beta_0 + \beta_1 x_i + \varepsilon_i)$
    wenn x durch $x + 1$ ersetzt wird, ändert sich Y zu $Y \cdot \exp(\beta_1)$

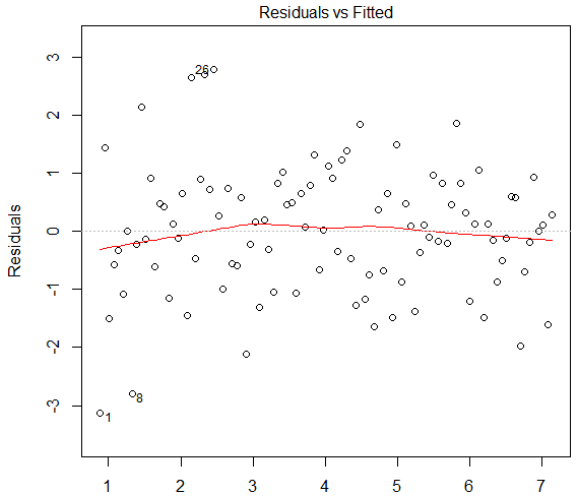
y log-transformieren...

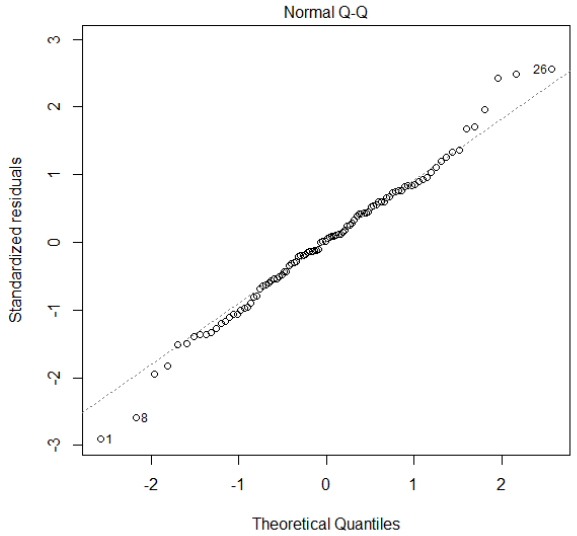


...bessere Residuen erhalten





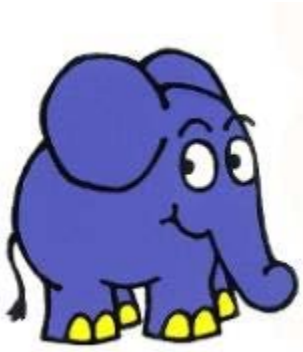




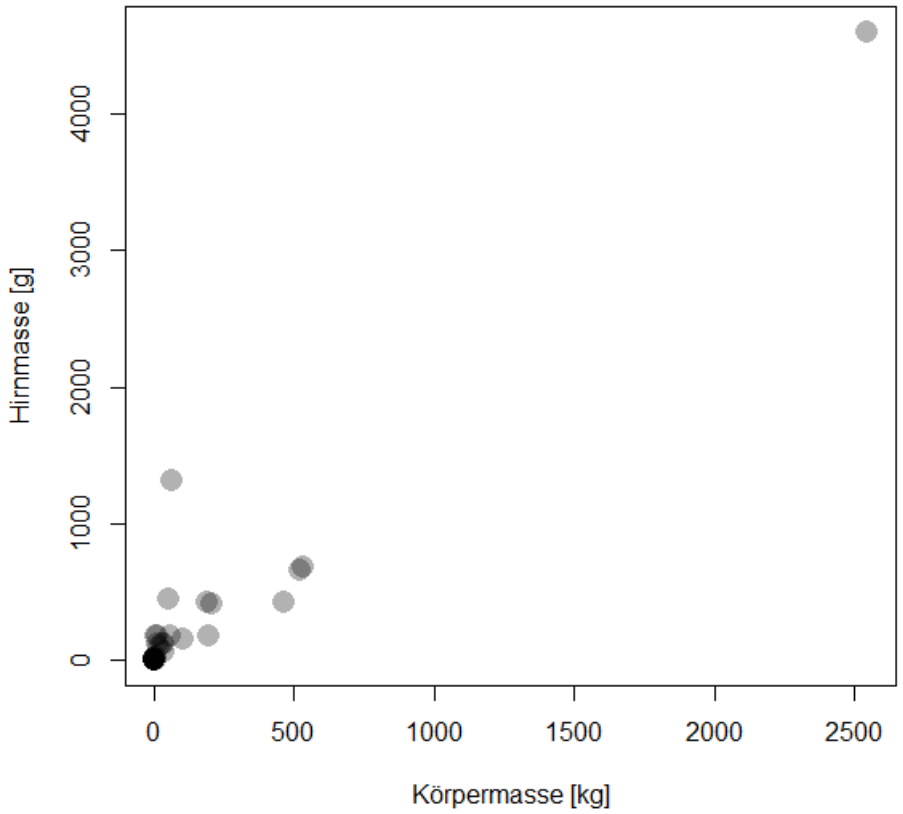
- $\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$
- $Y_i = \exp(\beta_0 + \beta_1 x_i + \varepsilon_i)$

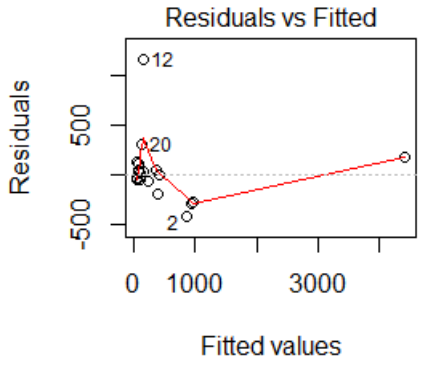
Bsp. Hirnmasse vs Körpermasse

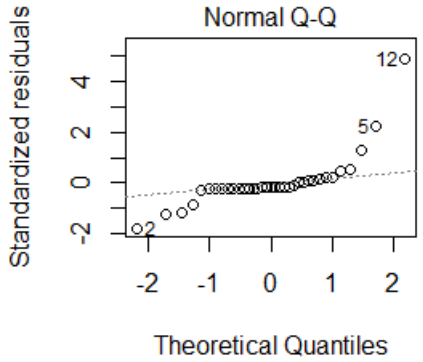
- Frage:
  - Gibt es einen Zusammenhang zwischen der Körpermasse und der Hirnmasse?



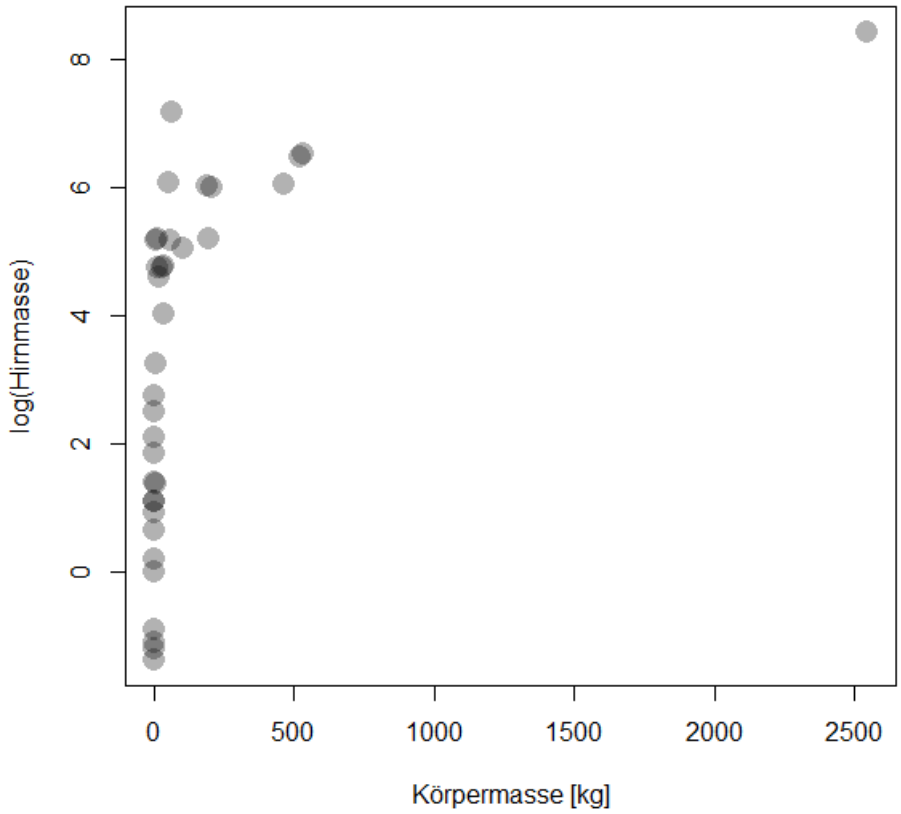
Bsp. Hirnmasse vs Körpermasse

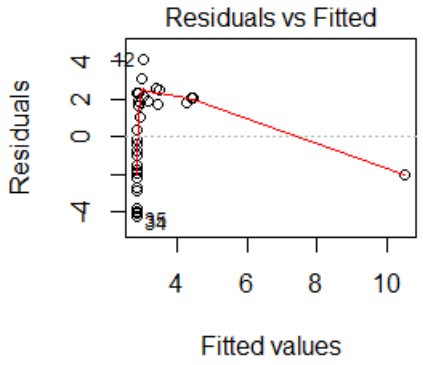


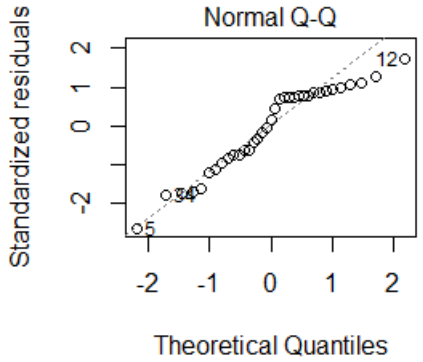




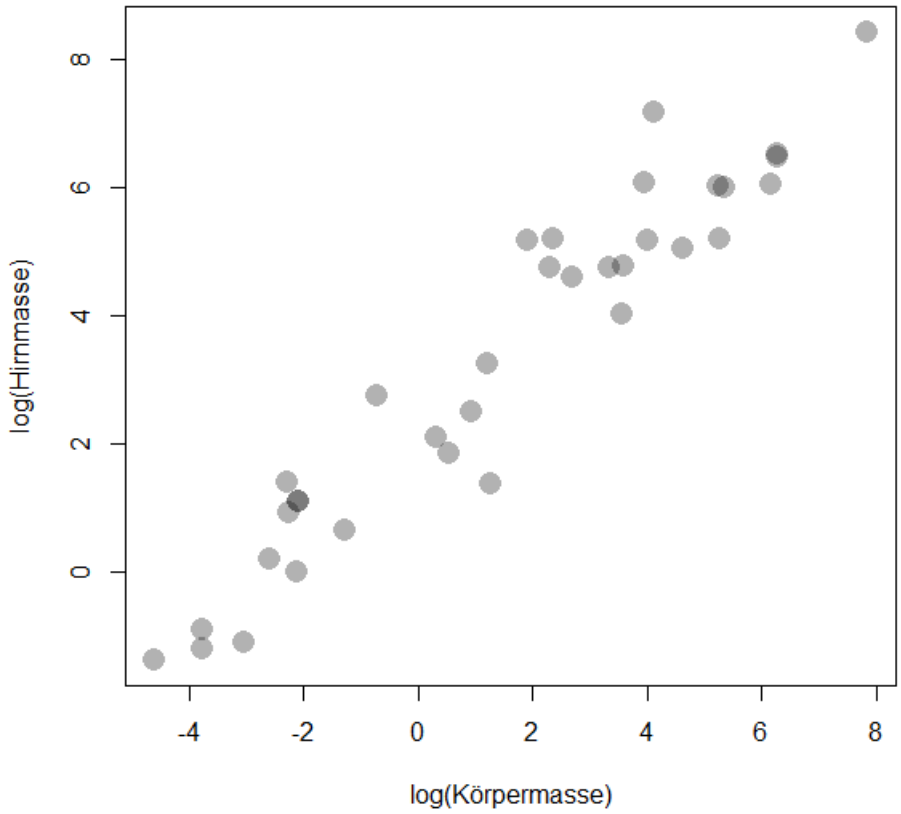
Bsp. log(Hirnmasse) vs Körpermasse

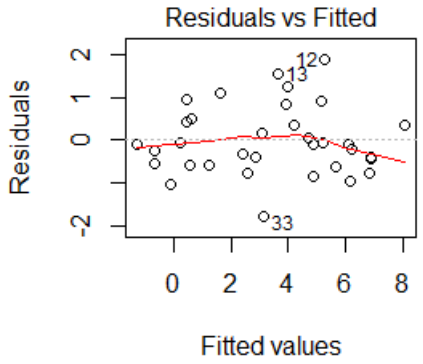


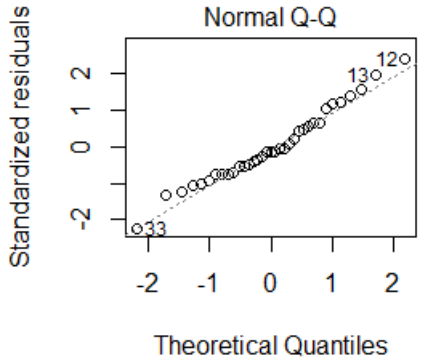




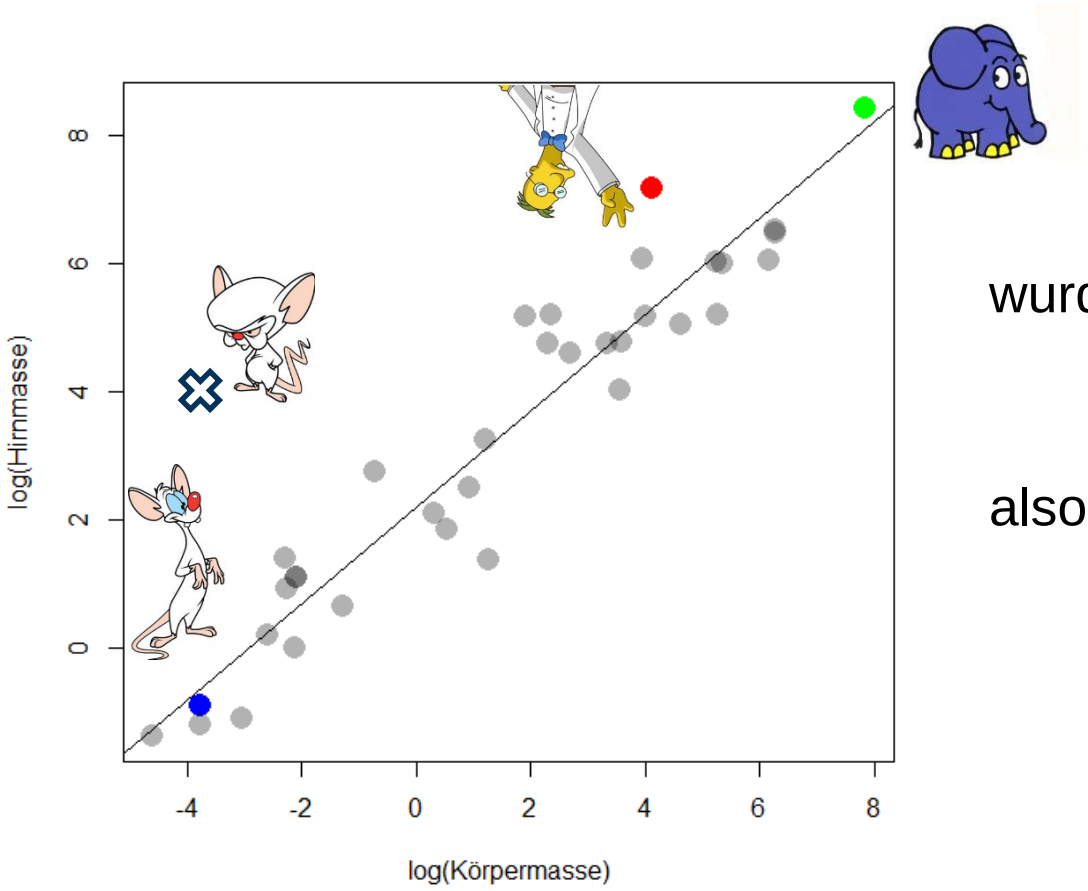
Bsp. log(Hirnmasse) vs log(Körpermasse)







Bsp. log(Hirnmasse) vs log(Körpermasse)



$$H = \widehat{\beta_0} + \widehat{\beta_1} \cdot K$$

wurde zu...

$$\log(H) = \widehat{\beta_0} + \widehat{\beta_1} \cdot \log(K)$$

also ist...

$$H = \exp\left(\widehat{\beta_0} + \widehat{\beta_1} \cdot \log(K)\right)$$

$$= \exp\left(\widehat{\beta_0}\right) \cdot \exp\left(\widehat{\beta_1} \cdot \log(K)\right)$$

$$= \hat{a} \cdot K^{\widehat{\beta_1}} = \hat{a} \cdot K^{\hat{b}}$$

$\widehat{\beta_0} = 2.19$ (95%-CI: $[1.89, 2.49]$); $\widehat{\beta_1} = 0.75$ (95%-CI: $[0.67, 0.83]$)

$\hat{a} = \exp\left(\widehat{\beta_0}\right) = 8.94$ (95%-CI: $[6.60, 12.0]$); $\hat{b} = \widehat{\beta_1}$

Übersicht nützlicher Transformationen

- Linearer Zusammenhang
  - $y = a + b \cdot x$
  - Keine Transformation nötig
- Exponentieller Zusammenhang
  - $\log(y) = a + b \cdot x \leftrightarrow y = \exp(a) \cdot \exp(b \cdot x)$
  - $\log$ -Transformation von y
- Polynomieller Zusammenhang
  - $\log(y) = a + b \cdot \log(x) \leftrightarrow y = \exp(a + b \cdot \log(x)) \leftrightarrow y = \exp(a) \cdot x^b$
  - $\log$ -Transformation von y und x

Multiple Lineare Regression

- Wie hängt Energie von Eiweiss, Kohlehydrate und Fett ab?

Multiple Lineare Regression: Interpretation

- Energie (E), Eiweiss (EW), Kohlehydrate (K), Fett (F)
- Modell:

$$E[kcal] = \beta_0 + \beta_1 EW[g] + \beta_2 K[g] + \beta_3 F[g] + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

- Was bedeutet es, wenn in diesem Modell $\beta_3 = 8$?

1. Wenn ein Nahrungsmittel ein Gramm mehr Fett als ein anderes hat, enthält es im Schnitt 8 kcal mehr Energie.
2. Wenn ein Nahrungsmittel ein Gramm mehr Fett als ein anderes hat und gleich viel Eiweiss und Kohlehydrate enthält, enthält es im Schnitt 8 kcal mehr Energie.

Einfache oder Multiple Regression

(Gilt für alle GLMs; hier am Bsp der linearen Regression)

- Einfache Regression: "Totaler Effekt"
  - $y \sim x \rightarrow$ "Wenn sich x um eine Einheit erhöht, erhöht sich y um β_1"
- Multiple Regression: "Bereinigter Effekt"
  - $y \sim x_1 + x_2 \rightarrow$ "Wenn sich x_1 um eine Einheit erhöht und x_2 konstant bleibt, erhöht sich y um β_1"
- Kein "richtig" oder "falsch"; eher zwei verschiedene Sichtweisen auf das gleiche Problem

Vorteil von Multipler Regression

- Andere Einflüsse werden ausgeschaltet
- Bsp: Diskriminierung
  - Einfache Regression:
    Zulassung ~ Geschlecht
  - Multiple Regression:
    Zulassung ~ Geschlecht + Job
- Berühmtes Beispiel:
  - Simpson’s Paradox

Multiple Lineare Regression

- Modell:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i(p-1)} + \varepsilon_i$$
$$\varepsilon_1, \ldots, \varepsilon_n \;\; i.i.d., \mathrm{E}(\varepsilon_i) = 0, \mathrm{Var}(\varepsilon_i) = \sigma^2$$

  - $p \mathrel{\widehat{=}} \#\beta$'s

- Achtung:
  Wenn man bei der Multiplen Linearen Regression eine erklärende Variable weglässt, muss man das Modell neu schätzen
  (alle β's ändern sich jeweils)

- Interpretation der Tests der Koeffizienten und der Residuenplots sind gleich bei der MLR...
- ...bei der Interpretation der Koeffizienten selber besteht ein wesentlicher Unterschied (Clicker Frage von vorhin!)

F-Test

- Hat mindestens eine erklärende Variable einen relevanten Effekt auf die Zielvariable?

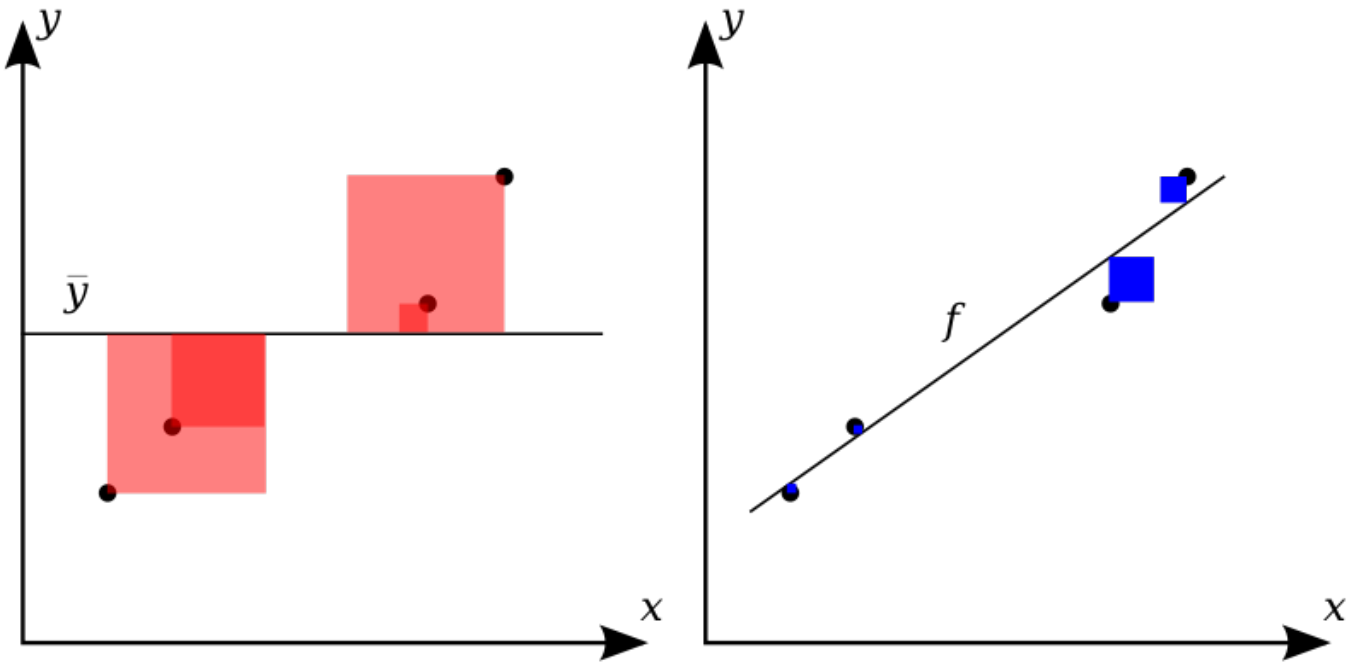
$$\mathcal{H}_0: \beta_1 = \cdots = \beta_{p-1} = 0$$

$$\mathcal{H}_A: \beta_j \neq 0$$

für mindestens ein j, $j = 1, 2, \ldots, p-1$.

- in R: `F-statistic: 362.7 on 1 and 33 DF,  p-value: < 2.2e-16`

Bestimmtheitsmass R^2



$$R^2 = 1 - \frac{SS_E}{SS_Y}$$

R^2 : wie nahe liegen die Punkte an der Regressionsgeraden
(im Vergleich zu der ursprünglichen Streuung der y-Werte)

Energiegehalt von 20 Lebensmitteln

Daten (pro 100g)

Name	kcal	gE	gK	gF
Butter	729	0.5	0.5	82.0
Laetta	370	0.0	4.0	39.0
Mozzarella	257	19.0	1.0	20.0
Cantadou	323	7.0	3.0	32.0
Lc1	105	3.5	15.5	3.0
Emmi	130	4.0	16.0	5.5
Quark	65	12.0	2.5	0.1
LightKaese	249	29.0	2.0	14.0
Banane	93	1.0	22.0	0.0
Zucchini	19	1.6	3.3	0.4
Tomate	17	1.0	2.6	0.2
Kartoffel	86	2.0	19.0	0.1
Brot	282	11.0	53.0	1.5
CremeSchnitte	311	4.5	48.0	11.0
Pizza	227	13.0	31.0	5.0
Schoko	569	7.0	46.0	40.0
Chips	517	7.0	51.0	32.0
Spaghetti	350	12.0	72.2	1.5
Reis	358	5.0	83.0	0.5
Stocki	320	9.0	70.0	1.0

Multiple Lineare Regression

```
lm(formula = kcal ~ gE + gK + gF, data = dat)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.70736    2.10299   0.812    0.429
gE           4.04087    0.14280  28.298  4.3e-15
gK           4.00415    0.03838 104.330  < 2e-16
gF           8.84937    0.05025 176.115  < 2e-16
```

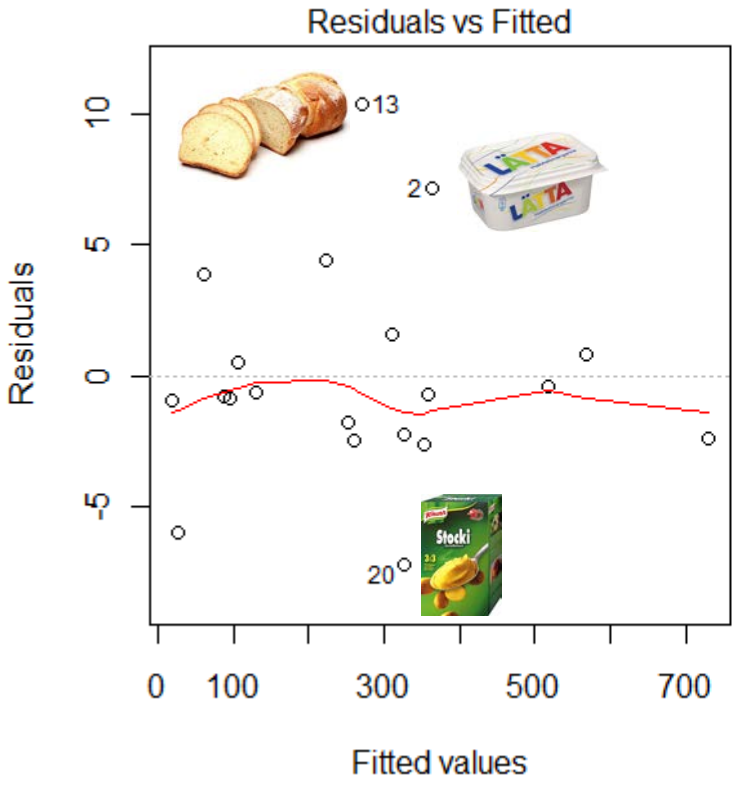
Ein Lebensmittel, das ein Gramm mehr Fett ***aber gleich viel Eiweiss und Kohlenhydrate enthält***, hat im Schnitt 8.8 kcal (95%-VI: [7.8; 9.8]) mehr Energie.

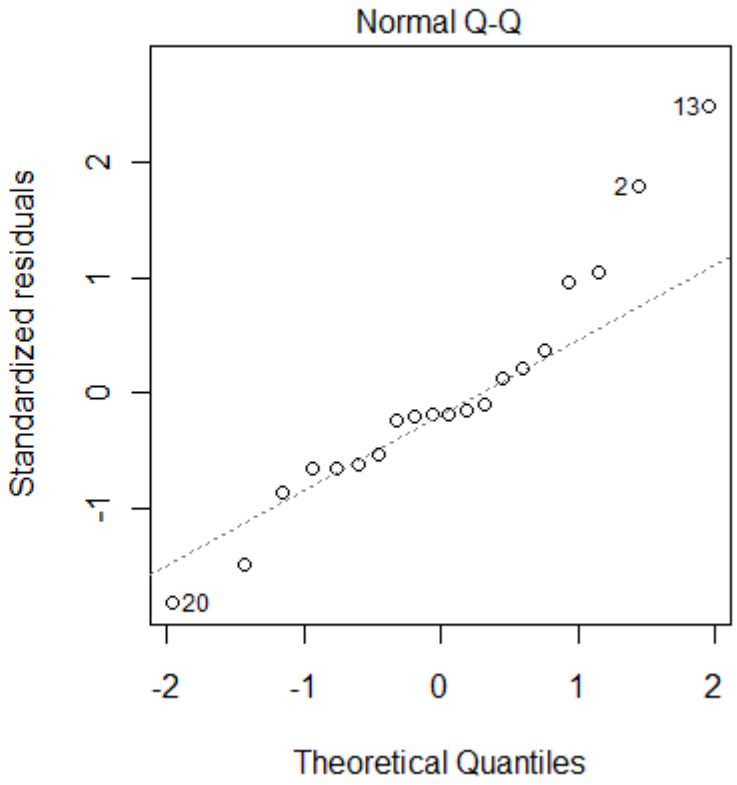
```
Multiple R-squared: 0.9995
```

Die Punkte liegen äusserst genau auf der geschätzten Geraden.

(verglichen mit der ursprünglichen Streuung der Energiewerte)

Residuenanalyse





wie gesagt: «aus Eingeweiden lesen...»

- Sieht ganz nett aus...
- Lätta (2), Brot (13) und Stocki (20) fallen etwas aus dem Rahmen

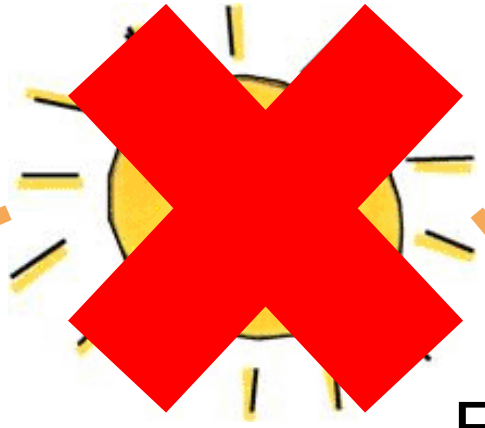
Ursache und Wirkung

Opfer durch Ertrinken

?

Eisverkauf

Ursache und Wirkung



Opfer durch Ertrinken

Eisverkauf

Kausaler Zusammenhang

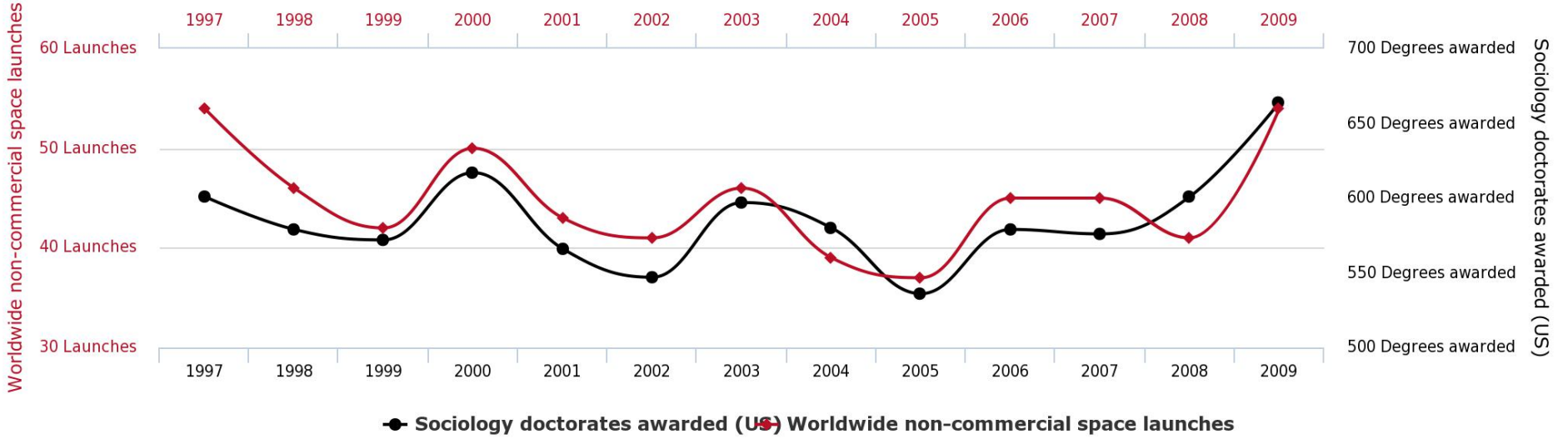
≠

Korrelation

Worldwide non-commercial space launches

correlates with

Sociology doctorates awarded (US)



tylervigen.com

Wie findet man Kausalzusammenhänge?

Randomisiertes, kontrolliertes Experiment

Experiment

Kausaleffekt finden

?

Experiment

Kausaleffekt finden

Experiment

Kausaleffekt finden

Experiment

Kausaleffekt finden

Dünger besser als kein Dünger?

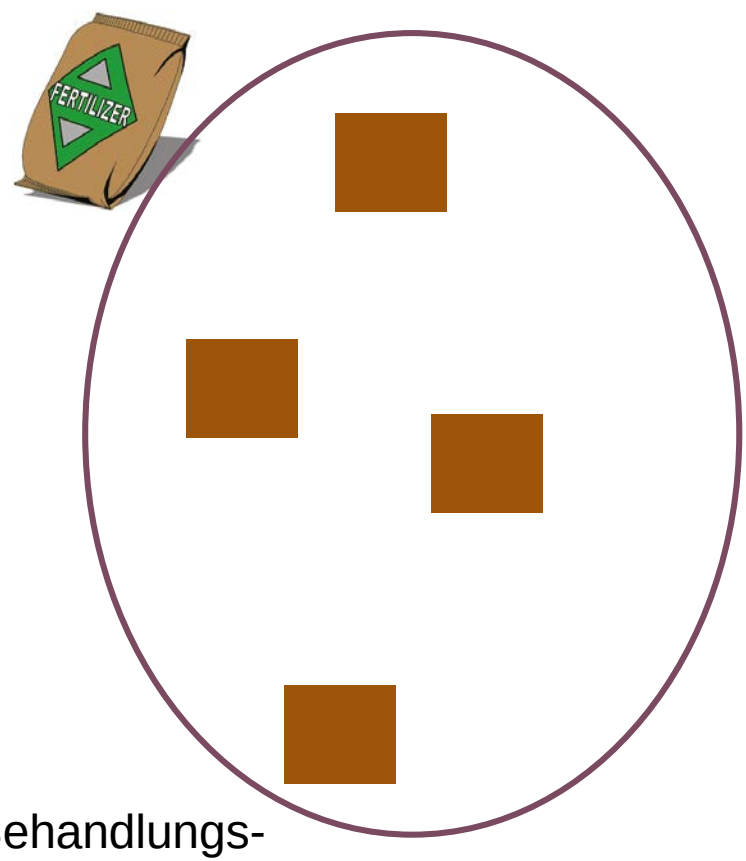
Keine Ahnung!

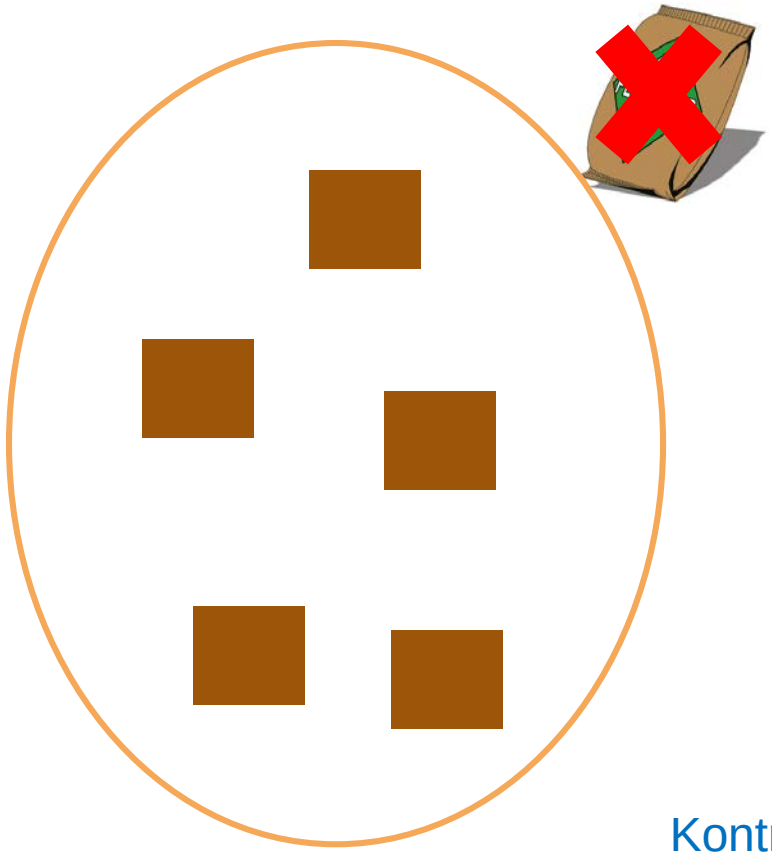
Wie viele rote Blumen hätte es ohne Dünger gegeben?

Brauchen eine Kontrollgruppe

Experiment

Kausaleffekt finden





Behandlungs-gruppe

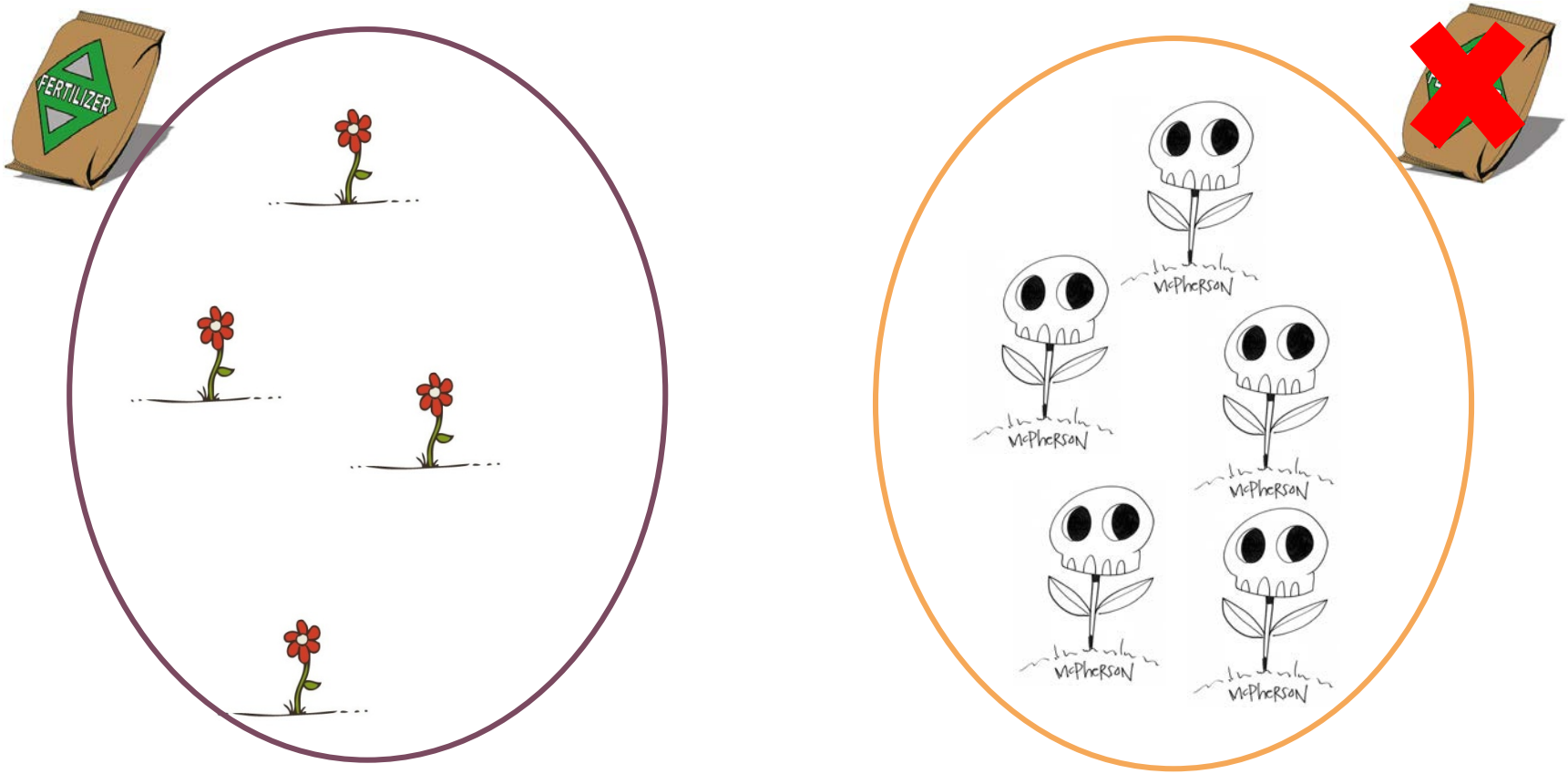
Kontroll-gruppe

Zwei Gruppen von Feldern **in allem gleich**
(Bodenqualität, Wasser, Sonnenlicht, …)

Praxis: Zufällige Zuordnung der Felder

Experiment

Kausaleffekt finden



Ergebnis ist wegen Dünger, weil alles andere gleich war

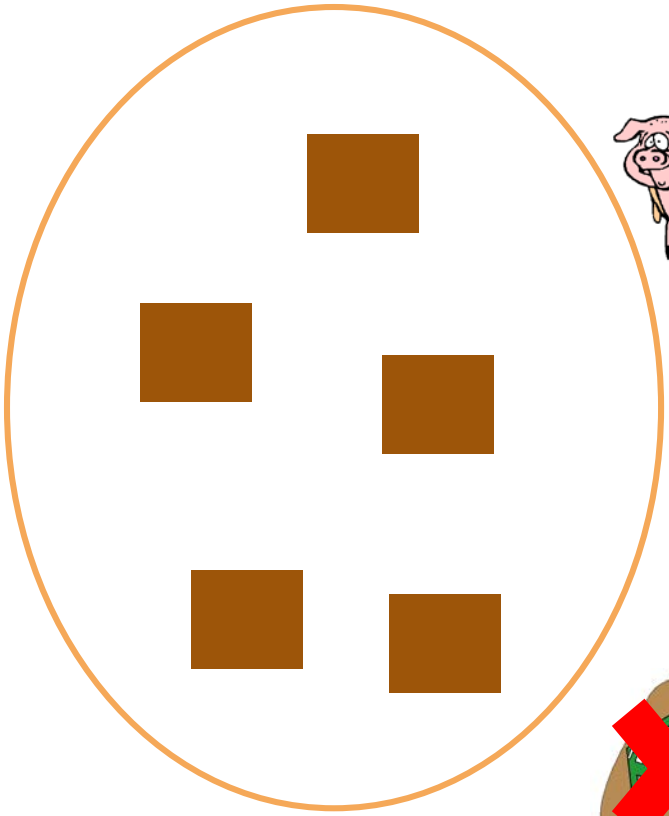
Manchmal sind randomisierte, kontrollierte Experimente nicht machbar

- zu teuer, zu zeitaufwändig (Genexpressionsdaten)
- unethisch, nicht machbar (HIV Behandlung, Rauchen)
- Falls Experiment nicht machbar…

➢ Beobachtungsstudie

Beobachtungsstudie

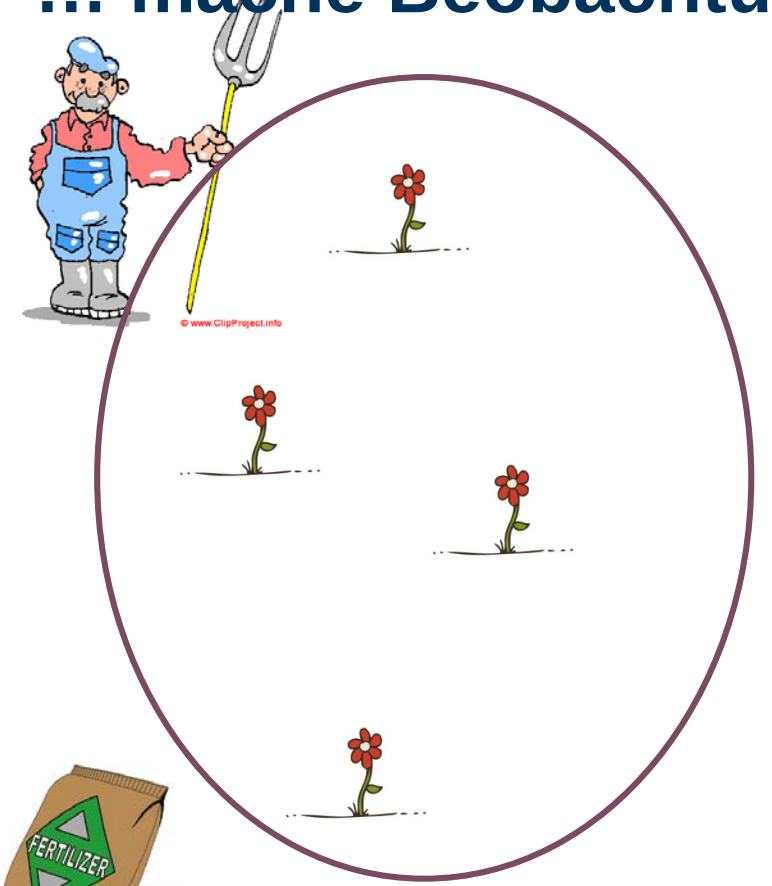
… mache Beobachtungen.

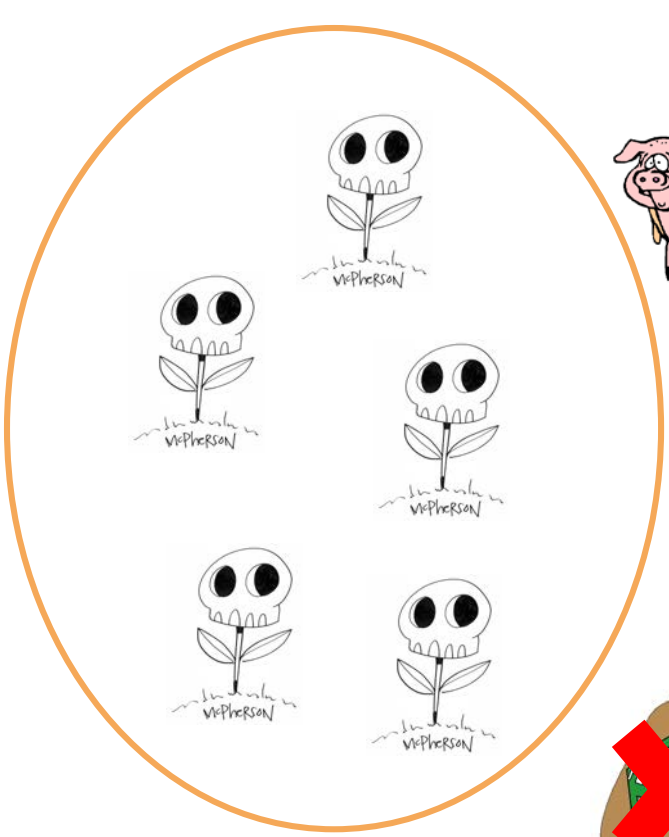


Es ist nicht garantiert, dass beide Gruppen in allen Aspekten gleich sind

Beobachtungsstudie

… mache Beobachtungen.



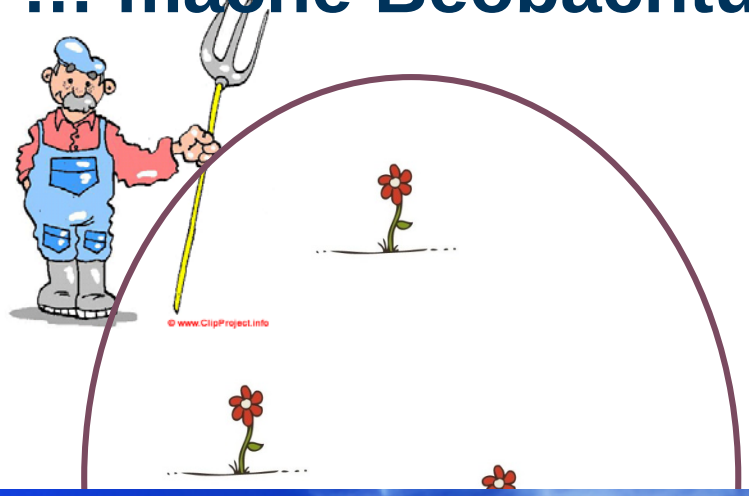


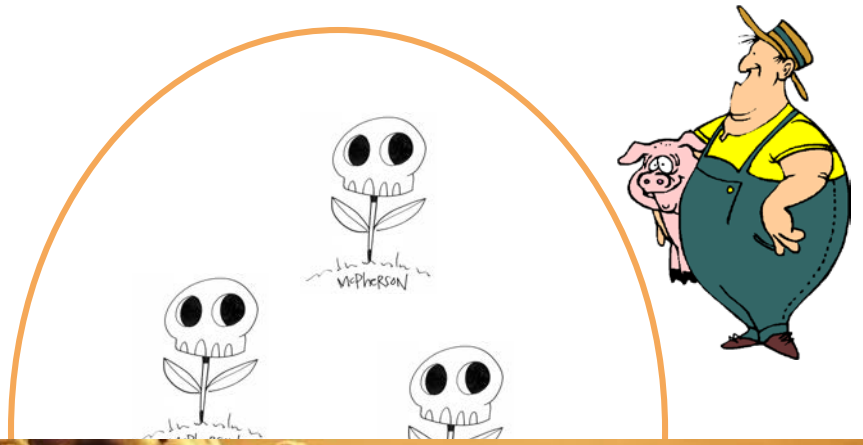
Ist das Ergebnis wegen Dünger?

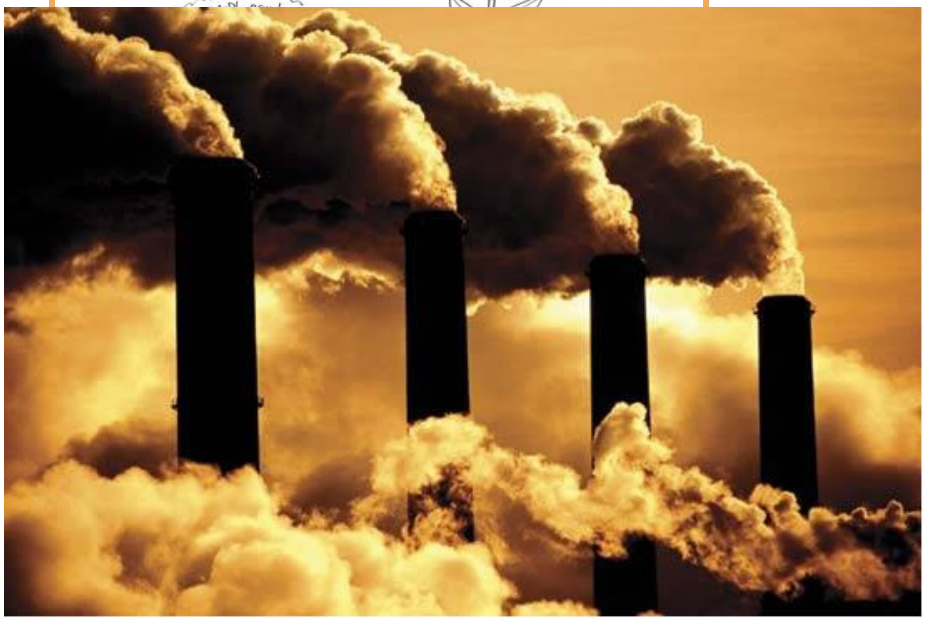
Keine Ahnung!

Beobachtungsstudie

… mache Beobachtungen.

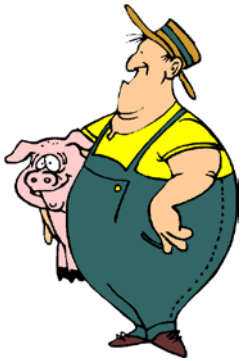






Beobachtungsstudie

Besser: Vergleiche Bauern, die in möglichst vielen Punkten übereinstimmen.

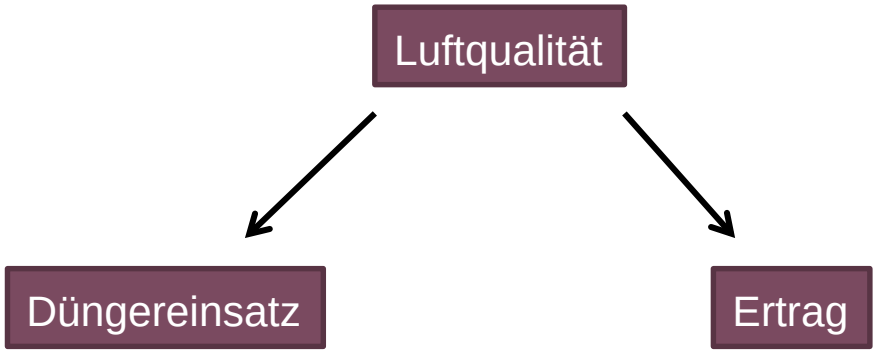


Beobachtungsstudie

Aber: Wir können nie sicher sein, dass es nicht doch noch irgendwelche relevanten Unterschiede zwischen den Gruppen gibt.

Zusammenfassung

- **Randomisierte, kontrollierte Experiment:** Beste Möglichkeit, Daten zu sammeln (“Goldstandard”)
- **Beobachtungsstudie:** Man muss skeptisch sein – kam der Effekt (*viele schöne Blumen*) durch die Behandlung (*Dünger*), oder durch einen Umstand, der in beiden Gruppen unterschiedlich war (*Luftqualität*)?



Zusammenfassung

- Transformation von Daten – bei schlechten Residuenplots
- Multiple lineare Regression – Energie von Nahrungsmitteln
- Korrelation ≠ Kausalität – Eis verursacht keine Badetote

Hausaufgaben

- Skript: Kapitel 5.3 lesen
- Serie 13 lösen
- Quiz 13 bearbeiten