Wahrscheinlichkeitsrechnung und Statistik für Biologen
Lineare Modelle

Noémie Becker & Dirk Metzler

http://evol.bio.lmu.de/_statgen

Sommersemester 2013

1. Regression zur Mitte

2. Multivariate Regression
   - Beispiel: Artenreichtum an Sandstränden
   - Beispiel: Wirksamkeit von Therapien

3. Modellwahl: AIC und Kreuzvalidierung
   - Beispiel: (Schnabel-)Größen der Darwin-Finken
   - Beispiel: Wasserflöhe

4. Klausur

Inhalt

Herkunft des Worts “Regression”

Wieso Regression=Rückkehr, Rückschritt?

Herkunft des Worts “Regression”

Wieso Regression=Rückkehr, Rückschritt?

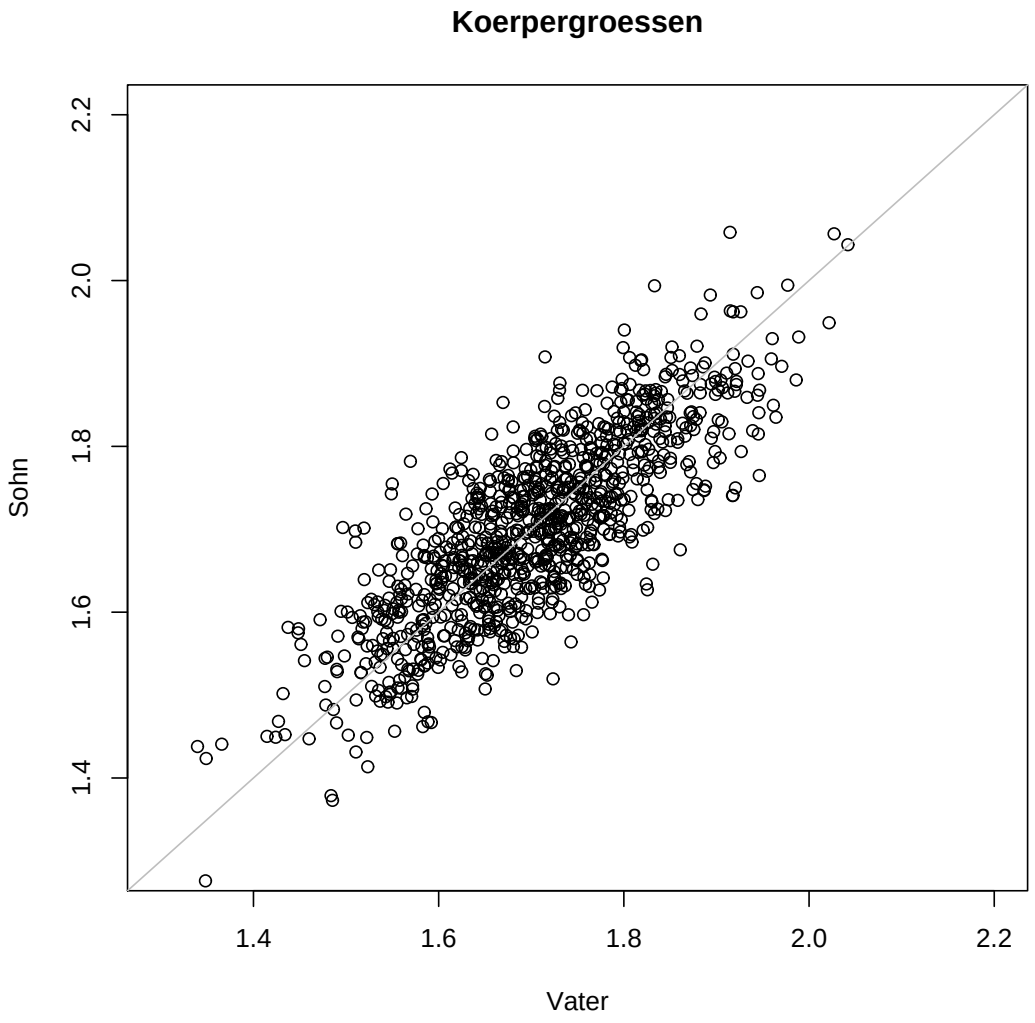
Sir Francis Galton (1822–1911): Regression toward the mean.

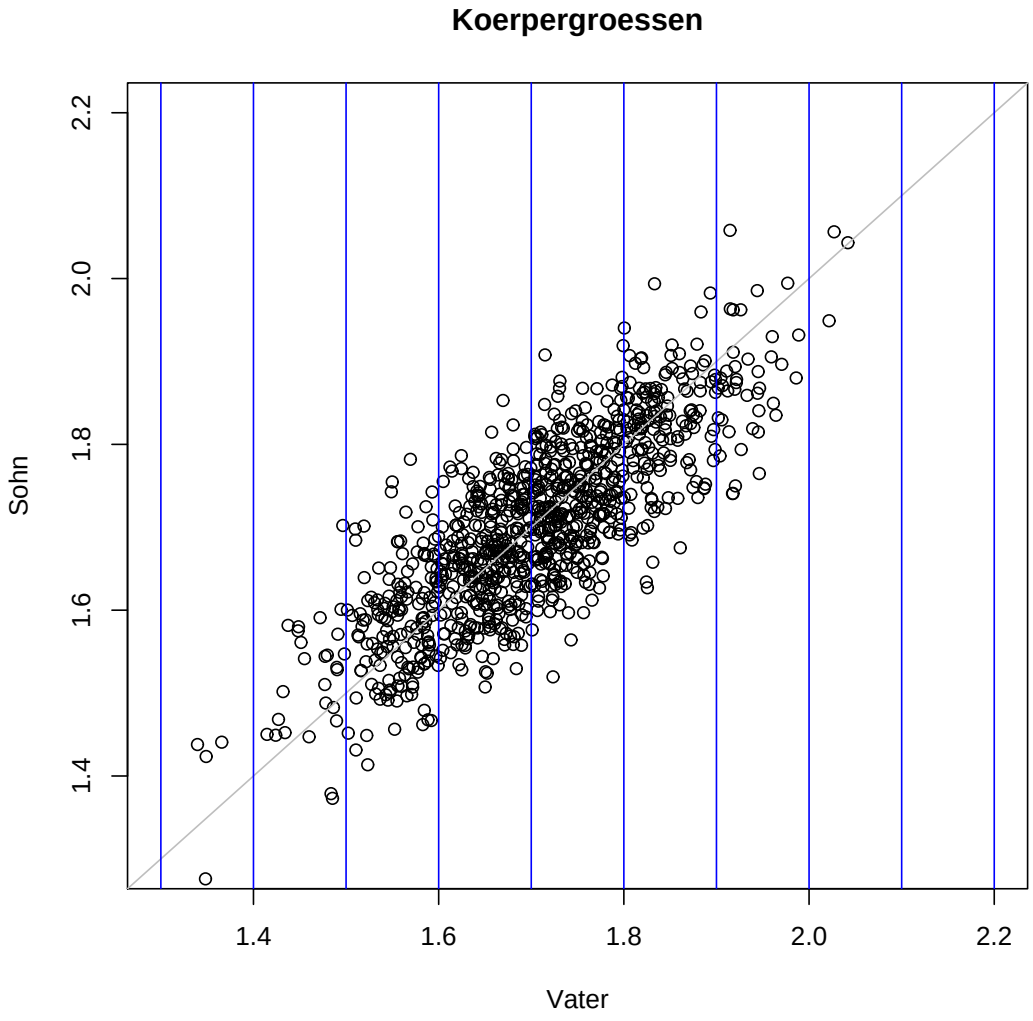
Herkunft des Worts “Regression”

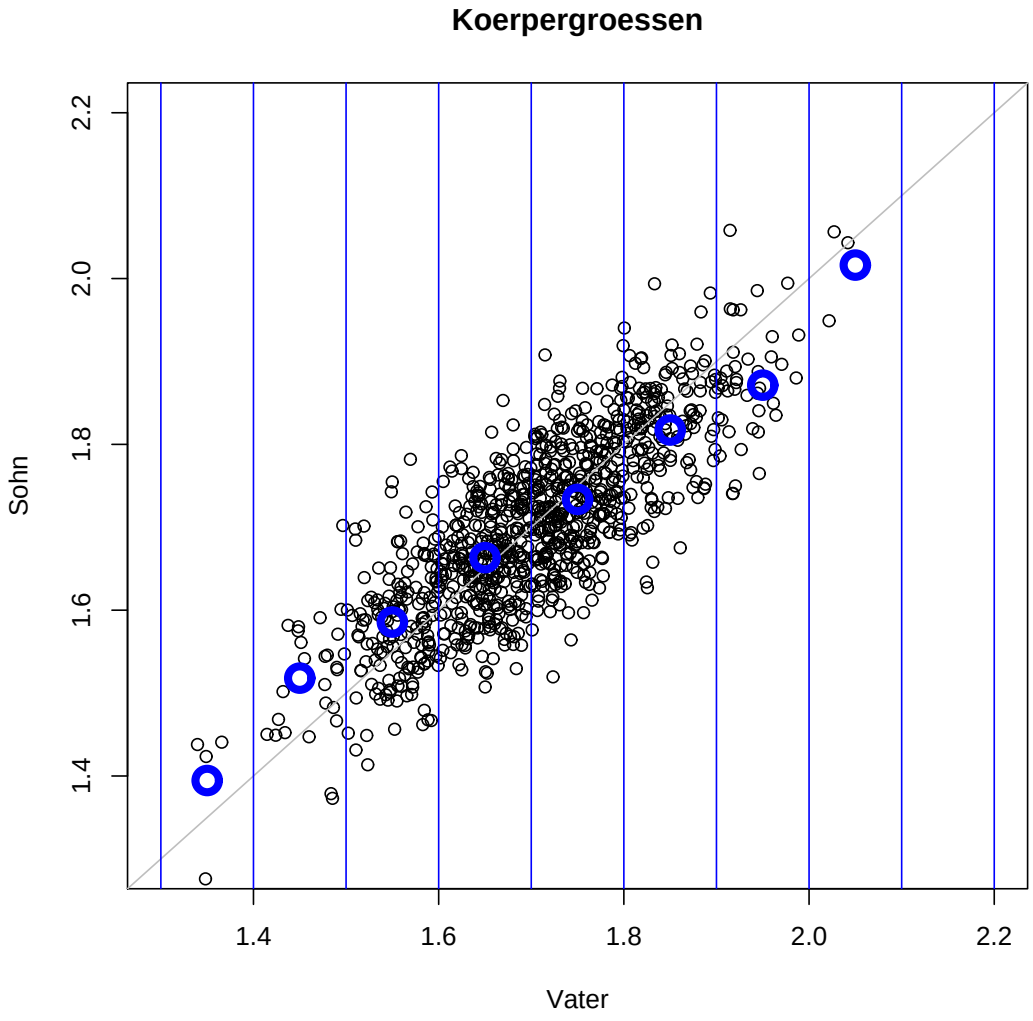
Wieso Regression=Rückkehr, Rückschritt?

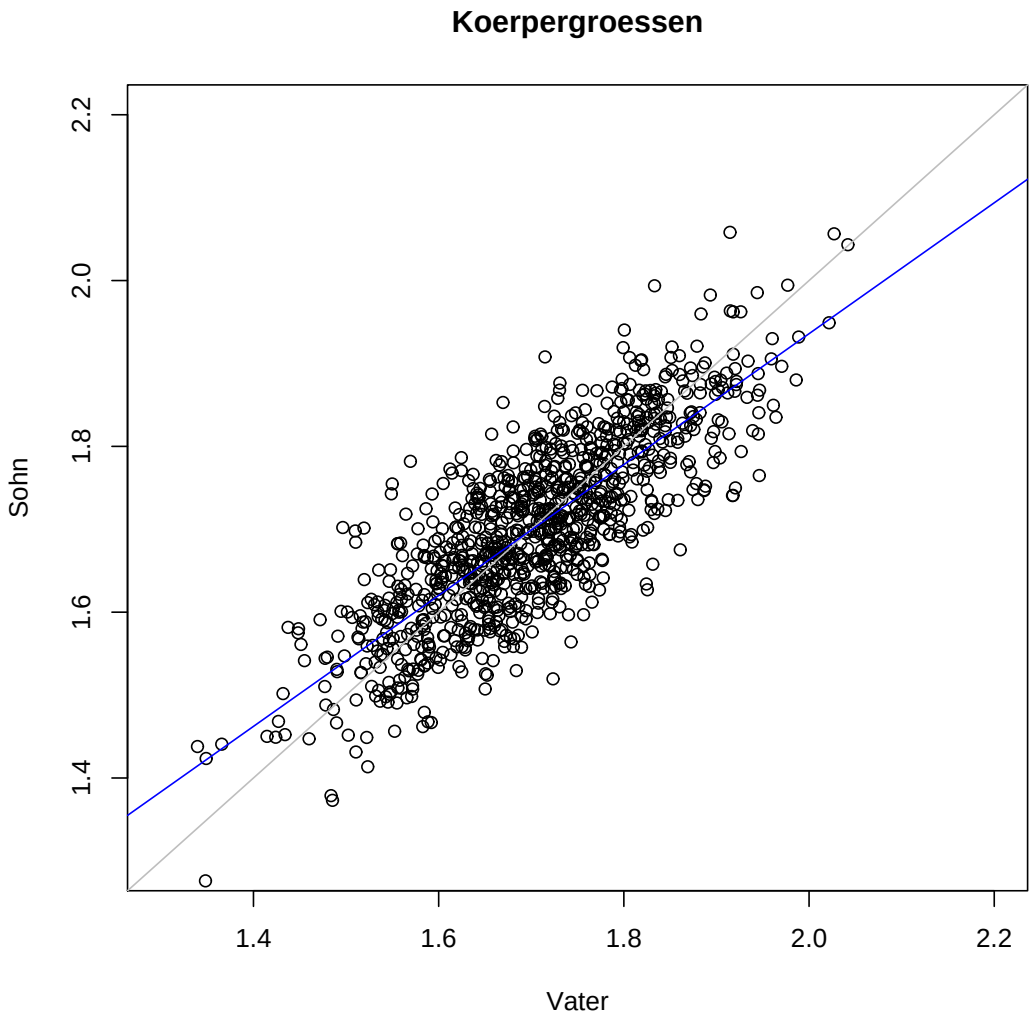
Sir Francis Galton (1822–1911): Regression toward the mean.

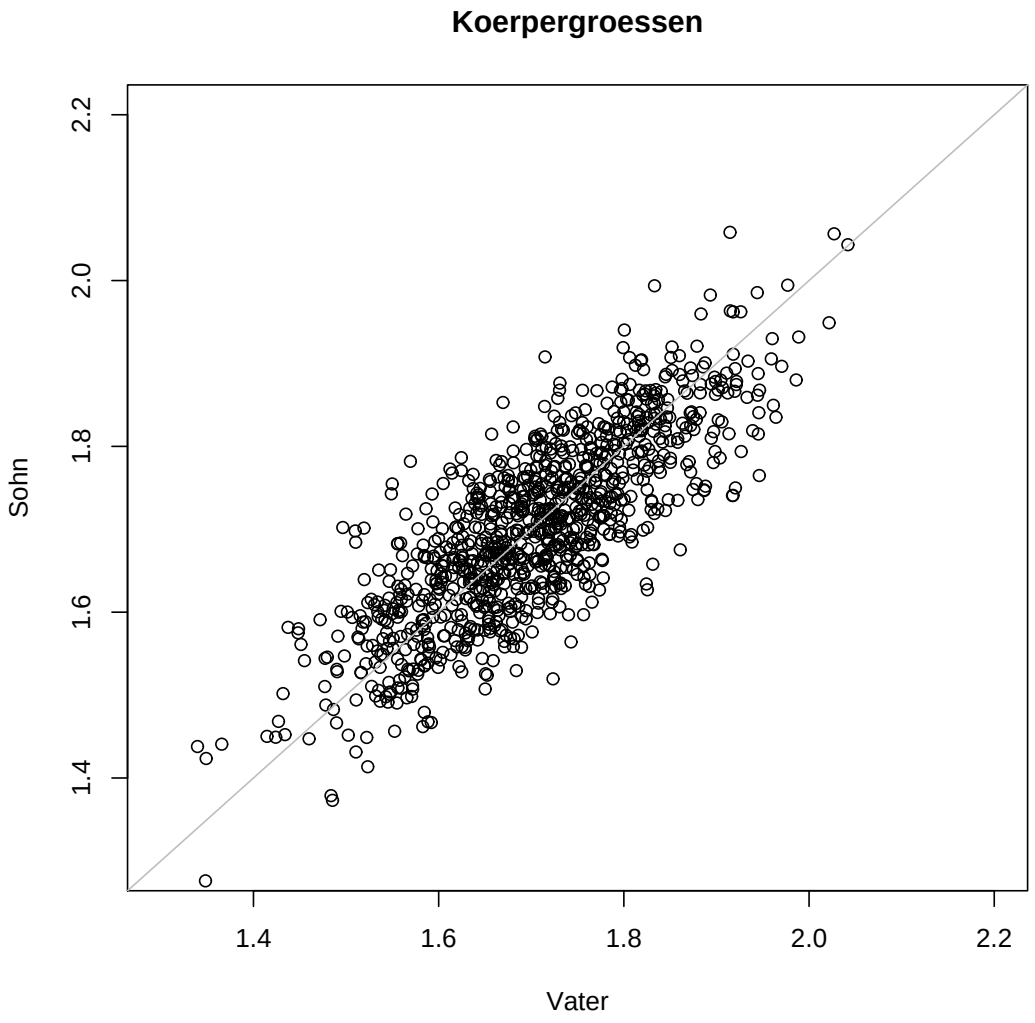
Große Väter haben Söhne, die im Schnitt etwas kleiner werden als sie selbst. Söhne kleiner Väter werden im Schnitt etwas größer als ihre Väter.

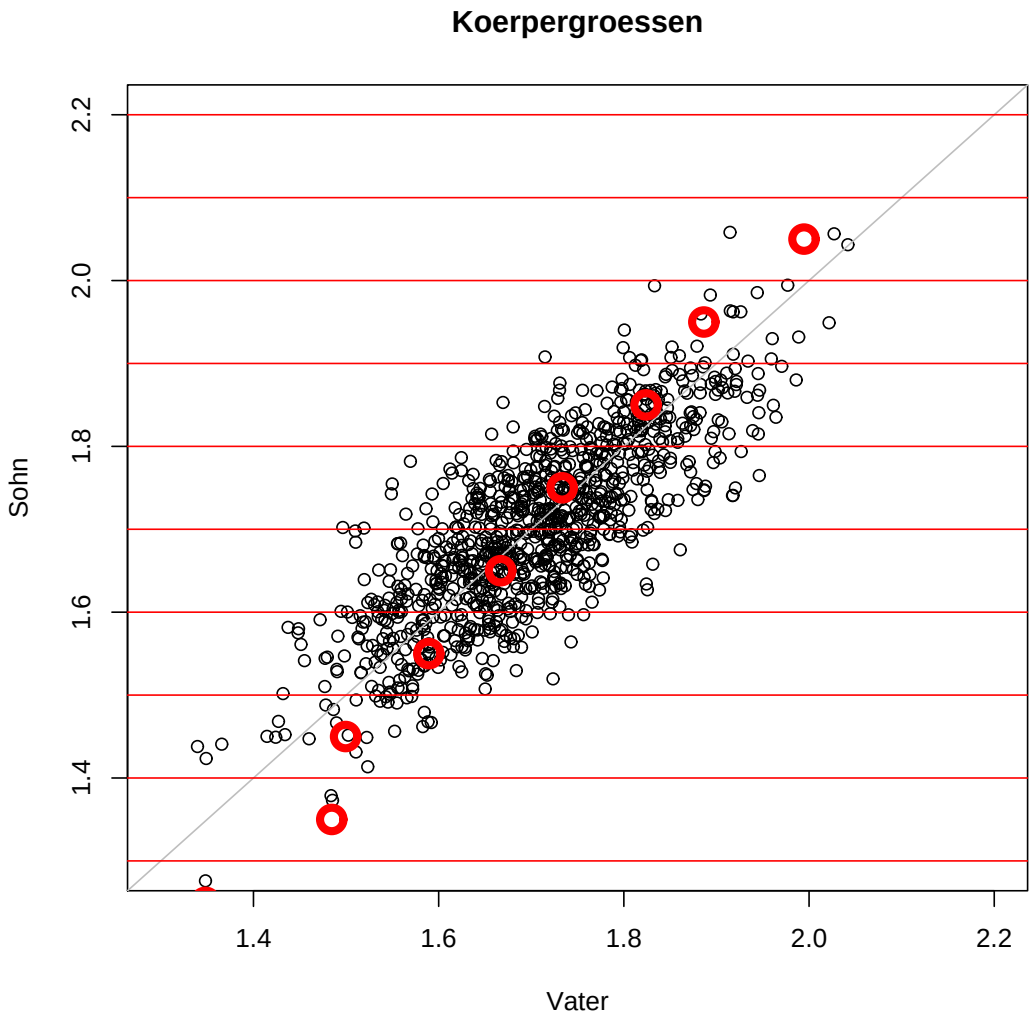

Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2

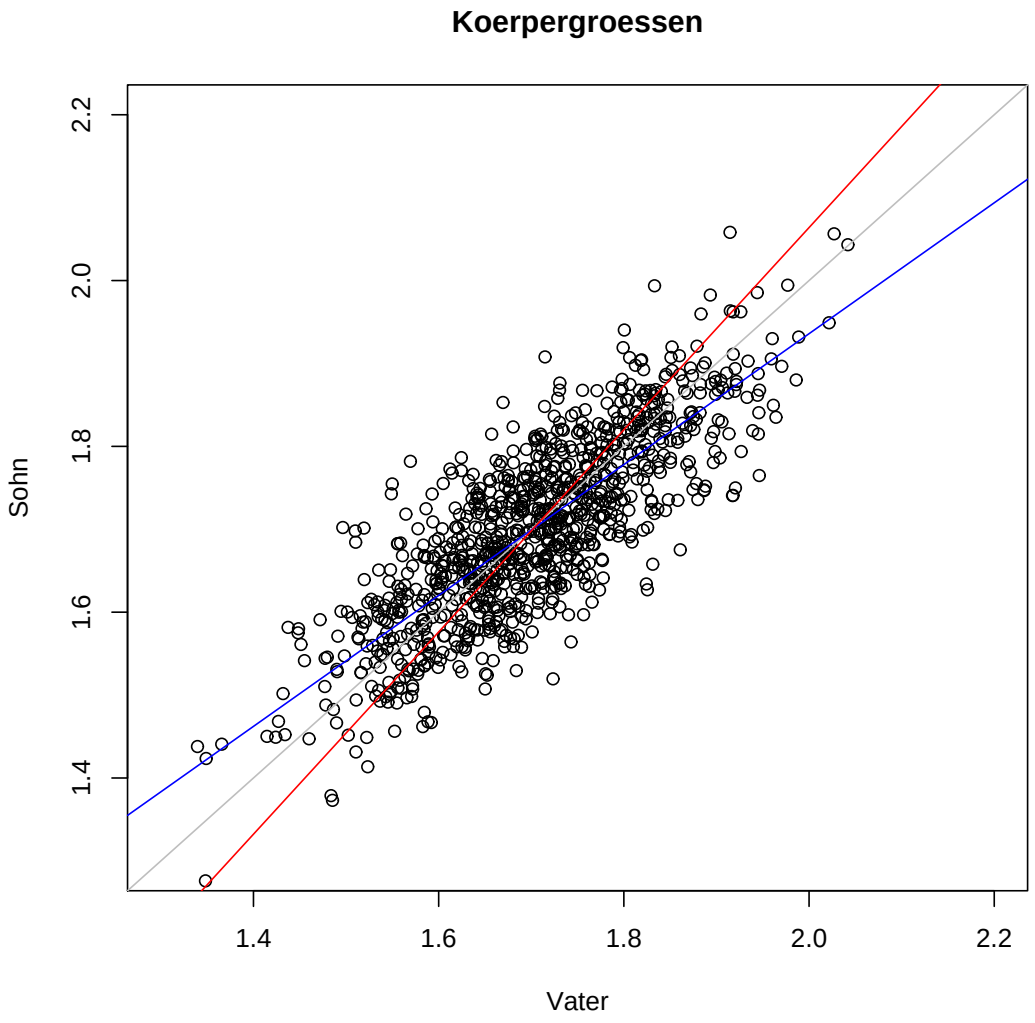

Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2

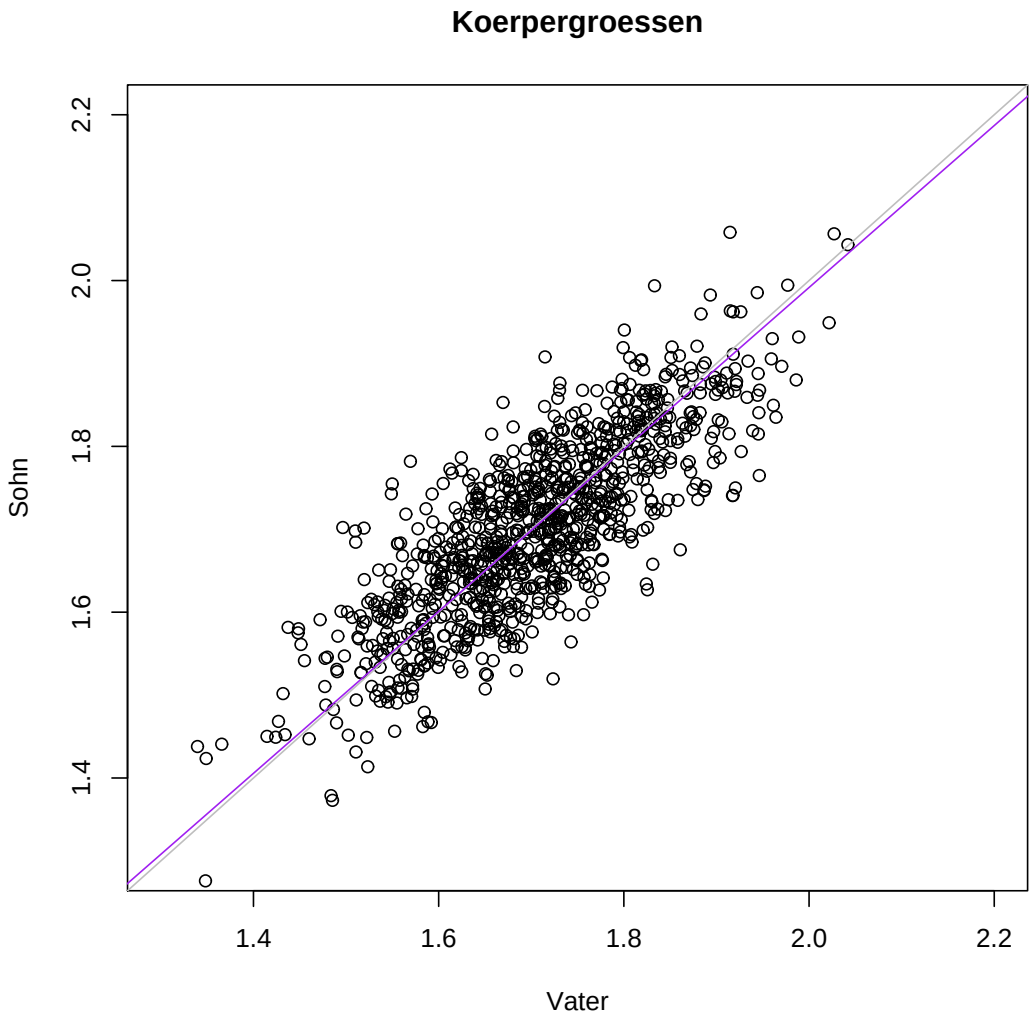

Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2


Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2


Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2


Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2


Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2


Koerpergroessen
Sohn
Vater
1.4
1.6
1.8
2.0
2.2

Ähnliche Effekte

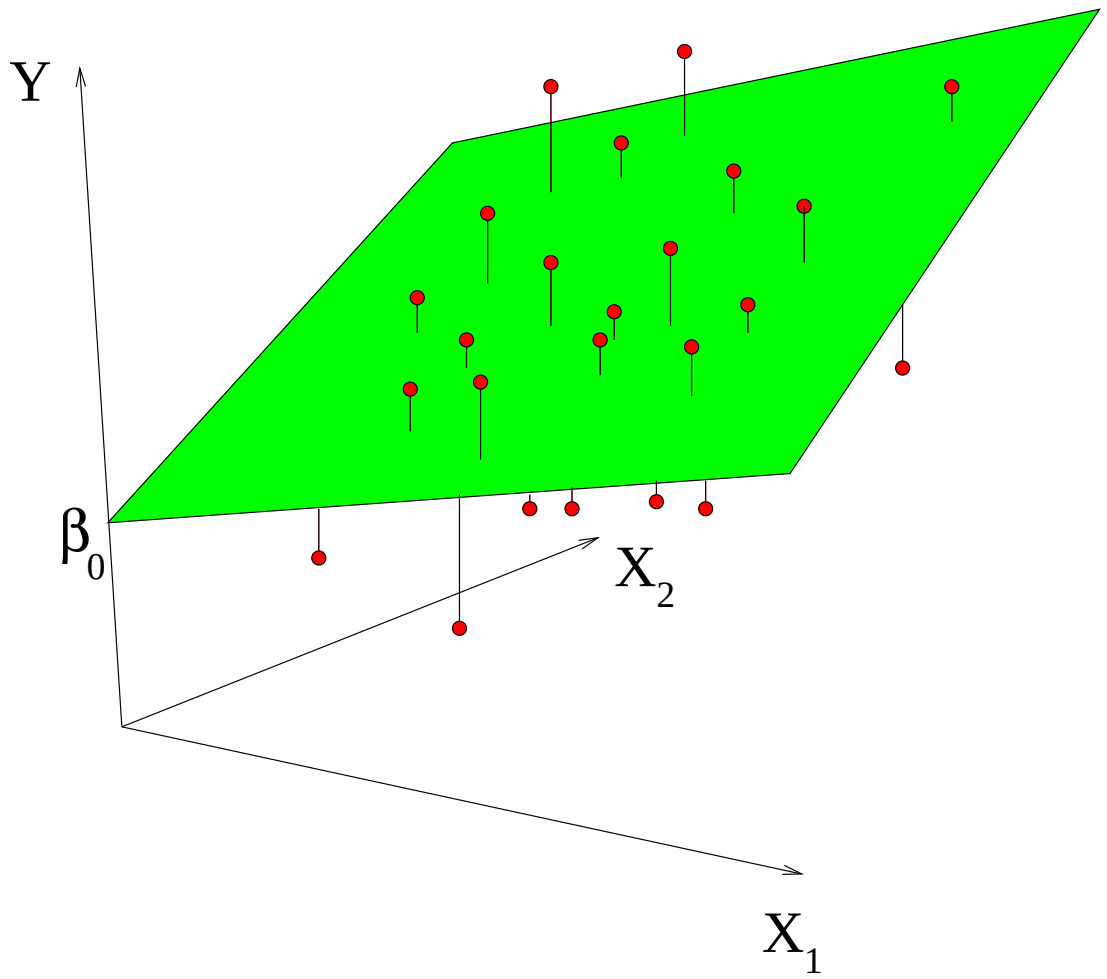
- Im Sport: der beste Sportler einer Saison wird in der nächsten Saison die hohen Erwartungen nicht erfüllen können.

Ähnliche Effekte

- Im Sport: der beste Sportler einer Saison wird in der nächsten Saison die hohen Erwartungen nicht erfüllen können.
- In der Schule: Wenn die 10 % schlechtesten Schüler Nachhilfe bekommen und im nächsten Schuljahr im Schnitt besser sind, beweist das noch nicht den Nutzen des Nachhilfeunterrichts.

Inhalt

Multivariate Regression



Multivariate Regression

Problem: Sage Y aus X_1, X_2,. . . ,X_m voraus.

Multivariate Regression

Problem: Sage Y aus $X_1, X_2, \ldots, X_m$ voraus.
Beobachtungen:

$$\begin{array}{lcl} Y_1 & , & X_{11}, X_{21}, \ldots, X_{m1} \\ Y_2 & , & X_{12}, X_{22}, \ldots, X_{m2} \\ \vdots & & \vdots \\ Y_n & , & X_{1n}, X_{2n}, \ldots, X_{mn} \end{array}$$

Multivariate Regression

Problem: Sage Y aus $X_1, X_2, \ldots, X_m$ voraus.
Beobachtungen:

$$\begin{array}{lcl} Y_1 & , & X_{11}, X_{21}, \ldots, X_{m1} \\ Y_2 & , & X_{12}, X_{22}, \ldots, X_{m2} \\ \vdots & & \vdots \\ Y_n & , & X_{1n}, X_{2n}, \ldots, X_{mn} \end{array}$$

Modell: $Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_m \cdot X_m + \varepsilon$

Multivariate Regression

Problem: Sage Y aus $X_1, X_2, \ldots, X_m$ voraus.
Beobachtungen:

$$\begin{array}{lcl} Y_1 & , & X_{11}, X_{21}, \ldots, X_{m1} \\ Y_2 & , & X_{12}, X_{22}, \ldots, X_{m2} \\ \vdots & & \vdots \\ Y_n & , & X_{1n}, X_{2n}, \ldots, X_{mn} \end{array}$$

Modell: $Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_m \cdot X_m + \varepsilon$
Gleichungssystem zum Bestimmen von a, b_1, b_2, $\ldots$, b_m:

$$\begin{array}{lllllllllllll} Y_1 & = & a & + & b_1 \cdot X_{11} & + & b_2 \cdot X_{21} & + & \ldots & + & b_m \cdot X_{m1} & + & \varepsilon_1 \\ Y_2 & = & a & + & b_1 \cdot X_{12} & + & b_2 \cdot X_{22} & + & \ldots & + & b_m \cdot X_{m2} & + & \varepsilon_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ Y_n & = & a & + & b_1 \cdot X_{1n} & + & b_n \cdot X_{2n} & + & \ldots & + & b_m \cdot X_{mn} & + & \varepsilon_n \end{array}$$

Modell:

$$
\begin{array}{lllllllllllll}
Y_1 & = & a & + & b_1 \cdot X_{11} & + & b_2 \cdot X_{21} & + & \ldots & + & b_m \cdot X_{m1} & + & \varepsilon_1 \\
Y_2 & = & a & + & b_1 \cdot X_{12} & + & b_2 \cdot X_{22} & + & \ldots & + & b_m \cdot X_{m2} & + & \varepsilon_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
Y_n & = & a & + & b_1 \cdot X_{1n} & + & b_n \cdot X_{2n} & + & \ldots & + & b_m \cdot X_{mn} & + & \varepsilon_n
\end{array}
$$

Zielvariable Y
Erklärende Variablen $X_1, X_2, \ldots, X_m$
Zu schätzende Parameter $a, b_1, \ldots, b_m$
Unabhängige normalverteilte Störungen $\varepsilon_1, \ldots, \varepsilon_m$ mit unbekannter Varianz σ^2.

Inhalt

- Von welchen Faktoren hängt der Artenreichtum an einem Stück Strand ab?
- Daten aus einer Studie des niederländischen National Institute for Coastal and Marine Management Rijkswaterstaat/RIKZ
- siehe auch
  -  Zuur, Ieno, Smith (2007) *Analysing Ecological Data.* Springer

	richness	angle2	NAP	grainsize	humus	week
1	11	96	0.045	222.5	0.05	1
2	10	96	-1.036	200.0	0.30	1
3	13	96	-1.336	194.5	0.10	1
4	11	96	0.616	221.0	0.15	1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
21	3	21	1.117	251.5	0.00	4
22	22	21	-0.503	265.0	0.00	4
23	6	21	0.729	275.5	0.10	4
.	.	.	.	.	.	.
.	.	.	.	.	.	.
43	3	96	-0.002	223.0	0.00	3
44	0	96	2.255	186.0	0.05	3
45	2	96	0.865	189.5	0.00	3

Bedeutung der Variablen

richness	Anzahl Arten, die an der Probestelle gefunden wurden.
angle2	Hangneigung des Strandes an der Probestelle
NAP	Höhe der Probestelle im Vergleich zur mittleren Wasserhöhe
grainsize	Durchschnittliche Größe der Sandkörner
humus	Anteil an organischem Material
week	in welcher der 4 Wochen wurde die Stelle beprobt

(Viele weitere Variablen im Originaldatensatz)

Modell 0:

$$\begin{aligned} \text{richness} &= a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ &\quad + b_4 \cdot \text{humus} + \varepsilon \end{aligned}$$

Modell 0:

$$\begin{aligned} \text{richness} &= a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ &\quad + b_4 \cdot \text{humus} + \varepsilon \end{aligned}$$

in R-Notation:

`richness ~ angle2 + NAP + grainsize + humus`

```
> modell0 <- lm(richness ~ angle2+NAP+grainsize+humus,
+                    data = rikz)
> summary(modell0)
Call:
lm(formula = richness ~ angle2 + NAP + grainsize + humus, data
Residuals:
    Min      1Q  Median      3Q     Max
-4.6851 -2.1935 -0.4218  1.6753 13.2957
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.35322    5.71888   3.209  0.00262 **
angle2      -0.02277    0.02995  -0.760  0.45144
NAP         -2.90451    0.59068  -4.917 1.54e-05 ***
grainsize   -0.04012    0.01532  -2.619  0.01239 *
humus       11.77641    9.71057   1.213  0.23234
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 3.644 on 40 degrees of freedom
Multiple R-squared: 0.5178, Adjusted R-squared: 0.4696
```

- z.B. die -2.90451 ist der Schätzer für b_2, den Vorfaktor von `NAP`

- z.B. die -2.90451 ist der Schätzer für b_2, den Vorfaktor von `NAP`
- Der p-Wert `Pr(>|t|)` bezieht sich auf die Nullhypothese, dass der wahre Parameterwert 0 sein könnte, d.h. dass die entsprechende erklärende Variable, z.B. `NAP` dann keinen Einfluß auf die Zielgröße (hier den Artenreichtum) hätte.

- z.B. die -2.90451 ist der Schätzer für b_2, den Vorfaktor von `NAP`
- Der p-Wert `Pr(>|t|)` bezieht sich auf die Nullhypothese, dass der wahre Parameterwert 0 sein könnte, d.h. dass die entsprechende erklärende Variable, z.B. `NAP` dann keinen Einfluß auf die Zielgröße (hier den Artenreichtum) hätte.
- `NAP` wird als hochsignifikant bewertet, `grainsize` ist ebenfalls signifikant.

- z.B. die -2.90451 ist der Schätzer für b_2, den Vorfaktor von `NAP`
- Der p-Wert `Pr(>|t|)` bezieht sich auf die Nullhypothese, dass der wahre Parameterwert 0 sein könnte, d.h. dass die entsprechende erklärende Variable, z.B. `NAP` dann keinen Einfluß auf die Zielgröße (hier den Artenreichtum) hätte.
- `NAP` wird als hochsignifikant bewertet, `grainsize` ist ebenfalls signifikant.
- Hat die Woche einen signifikanten Einfluß?

- z.B. die -2.90451 ist der Schätzer für b_2, den Vorfaktor von `NAP`
- Der p-Wert `Pr(>|t|)` bezieht sich auf die Nullhypothese, dass der wahre Parameterwert 0 sein könnte, d.h. dass die entsprechende erklärende Variable, z.B. `NAP` dann keinen Einfluß auf die Zielgröße (hier den Artenreichtum) hätte.
- `NAP` wird als hochsignifikant bewertet, `grainsize` ist ebenfalls signifikant.
- Hat die Woche einen signifikanten Einfluß?
- Es soll nicht die Nummer 1,2,3,4 der Woche mit einem Vorfaktor verrechnet werden, sondern die Zahlen werden als nicht-numerischer Faktor gesehen, d.h. jede Woche bekommt einen Parameter, der angibt, wie sehr stark die Artenzahl in der entsprechenden Woche erhöht oder vermindert ist.

- z.B. die -2.90451 ist der Schätzer für b_2, den Vorfaktor von `NAP`
- Der p-Wert `Pr(>|t|)` bezieht sich auf die Nullhypothese, dass der wahre Parameterwert 0 sein könnte, d.h. dass die entsprechende erklärende Variable, z.B. `NAP` dann keinen Einfluß auf die Zielgröße (hier den Artenreichtum) hätte.
- `NAP` wird als hochsignifikant bewertet, `grainsize` ist ebenfalls signifikant.
- Hat die Woche einen signifikanten Einfluß?
- Es soll nicht die Nummer 1,2,3,4 der Woche mit einem Vorfaktor verrechnet werden, sondern die Zahlen werden als nicht-numerischer Faktor gesehen, d.h. jede Woche bekommt einen Parameter, der angibt, wie sehr stark die Artenzahl in der entsprechenden Woche erhöht oder vermindert ist.
- In R wird dazu `week` in einen `factor` umgewandelt.

Modell 0:

$$\begin{aligned} \text{richness} &= a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ &\quad + b_4 \cdot \text{humus} + \\ &\quad b_5 \cdot I_{\texttt{week}=2} + b_6 \cdot I_{\texttt{week}=3} + b_7 \cdot I_{\texttt{week}=4} + \varepsilon \end{aligned}$$

Dabei ist $I_{\texttt{week}=k}$ eine sog. Indikatorvariable, die 1 ist, falls $\texttt{week}= k$ und sonst 0.

Modell 0:

$$\begin{aligned} \text{richness} &= a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ &\quad + b_4 \cdot \text{humus} + \\ &\quad b_5 \cdot I_{\texttt{week}=2} + b_6 \cdot I_{\texttt{week}=3} + b_7 \cdot I_{\texttt{week}=4} + \varepsilon \end{aligned}$$

Dabei ist $I_{\texttt{week}=k}$ eine sog. Indikatorvariable, die 1 ist, falls $\texttt{week}= k$ und sonst 0.

z.B. b_7 beschreibt, um wieviel an einer durchschnittlichen Probestelle der Artenreichtum in Woche 3 gegenüber Woche 1 erhöht ist.

Modell 0:

$$\begin{aligned} \texttt{richness} &= a + b_1 \cdot \texttt{angle2} + b_2 \cdot \texttt{NAP} + b_3 \cdot \texttt{grainsize} + \\ &\quad + b_4 \cdot \texttt{humus} + \\ &\quad b_5 \cdot I_{\texttt{week}=2} + b_6 \cdot I_{\texttt{week}=3} + b_7 \cdot I_{\texttt{week}=4} + \varepsilon \end{aligned}$$

Dabei ist $I_{\texttt{week}=k}$ eine sog. Indikatorvariable, die 1 ist, falls $\texttt{week}= k$ und sonst 0.

z.B. b_7 beschreibt, um wieviel an einer durchschnittlichen Probestelle der Artenreichtum in Woche 3 gegenüber Woche 1 erhöht ist.

in R-Notation:

```
richness ~ angle2 + NAP + grainsize + humus +
factor(week)
```

```
> modell <- lm(richness ~ angle2+NAP+grainsize+humus
+                    +factor(week), data = rikz)
> summary(modell)
  .
  .
  .

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.298448   7.967002   1.167 0.250629
angle2          0.016760   0.042934   0.390 0.698496
NAP            -2.274093   0.529411  -4.296 0.000121 ***
grainsize       0.002249   0.021066   0.107 0.915570
humus           0.519686   8.703910   0.060 0.952710
factor(week)2 -7.065098   1.761492  -4.011 0.000282 ***
factor(week)3 -5.719055   1.827616  -3.129 0.003411 **
factor(week)4 -1.481816   2.720089  -0.545 0.589182
---
```

- In Wochen 2 und 3 waren also signifikant weniger Arten zu finden als in Woche 1, die hier als “Bezugspunkt” dient

- In Wochen 2 und 3 waren also signifikant weniger Arten zu finden als in Woche 1, die hier als “Bezugspunkt” dient
- Der geschätze Achsenabschnitt `Intercept` entspricht also der zu erwartenden Artenzahl in Woche 1 an einer Probestelle, an der alle anderen Parameter 0 sind.

- In Wochen 2 und 3 waren also signifikant weniger Arten zu finden als in Woche 1, die hier als “Bezugspunkt” dient
- Der geschätze Achsenabschnitt `Intercept` entspricht also der zu erwartenden Artenzahl in Woche 1 an einer Probestelle, an der alle anderen Parameter 0 sind.
- eine alternative Darstellung ohne `Intercept` nimmt 0 als Bezugspunkt. Eine “-1” in der `R`-Notation repräsentiert “kein Intercept”.

```
> modell.alternativ <- lm(richness ~ angle2+NAP+
+              grainsize+humus+factor(week)-1, data = rikz)
> summary(modell.alternativ)
  .
  .
  .
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
angle2          0.016760   0.042934   0.390 0.698496
NAP            -2.274093   0.529411  -4.296 0.000121 ***
grainsize       0.002249   0.021066   0.107 0.915570
humus           0.519686   8.703910   0.060 0.952710
factor(week)1   9.298448   7.967002   1.167 0.250629
factor(week)2   2.233349   8.158816   0.274 0.785811
factor(week)3   3.579393   8.530193   0.420 0.677194
factor(week)4   7.816632   6.522282   1.198 0.238362
```

die *p*-Werte beziehen sich hier auf die Frage ob die vier geschätzten Achsenabschnitte für die einzelnen Wochen signifikant von 0 verschieden sind.

Wie testen wir, ob sich die Wochen unterscheiden?

Wie testen wir, ob sich die Wochen unterscheiden?

Z.B.: Wie wir im vorletzten Modell gesehen haben, sind Wochen 2 und 3 verschieden von Woche 1.

Wie testen wir, ob sich die Wochen unterscheiden?

Z.B.: Wie wir im vorletzten Modell gesehen haben, sind Wochen 2 und 3 verschieden von Woche 1. Der p-Wert bezieht sich aber auf die Situation eines Einzeltests.

Wie testen wir, ob sich die Wochen unterscheiden?

Z.B.: Wie wir im vorletzten Modell gesehen haben, sind Wochen 2 und 3 verschieden von Woche 1. Der p-Wert bezieht sich aber auf die Situation eines Einzeltests.

Wenn wir aber jedes Paar der vier Wochen vergleichen, führen wir $\binom{4}{2} = 6$ Test durch.

Wie testen wir, ob sich die Wochen unterscheiden?

Z.B.: Wie wir im vorletzten Modell gesehen haben, sind Wochen 2 und 3 verschieden von Woche 1. Der p-Wert bezieht sich aber auf die Situation eines Einzeltests.

Wenn wir aber jedes Paar der vier Wochen vergleichen, führen wir $\binom{4}{2} = 6$ Test durch.

Bonferroni-Korrektur: Multipliziere jeden p-Wert mit der Anzahl der durchgeführten Tests, in diesem Fall 6.

Bonferroni-Korrektur

Problem: Wenn man viele Tests durchführt, werden immer einige dabei sein, die Signifikanz anzeigen, auch wenn die Nullhypothese eigentlich gilt.

Bonferroni-Korrektur

Problem: Wenn man viele Tests durchführt, werden immer einige dabei sein, die Signifikanz anzeigen, auch wenn die Nullhypothese eigentlich gilt.

Beispiel: Führt man 20 Tests durch, mit Daten, die die Nullhypothese eigentlich erfüllen, wird im Schnitt ein Test Signifikanz auf dem 5%-Niveau anzeigen.

Bonferroni-Korrektur

Problem: Wenn man viele Tests durchführt, werden immer einige dabei sein, die Signifikanz anzeigen, auch wenn die Nullhypothese eigentlich gilt.

Beispiel: Führt man 20 Tests durch, mit Daten, die die Nullhypothese eigentlich erfüllen, wird im Schnitt ein Test Signifikanz auf dem 5%-Niveau anzeigen.

Bonferroni-Korrektur: Multipliziere alle p-Werte mit der Anzahl der Tests n. Wenn eines der Ergebnisse das Signifikanzniveau unterschreitet, verwirf die Nullhypothese

Bonferroni-Korrektur

Problem: Wenn man viele Tests durchführt, werden immer einige dabei sein, die Signifikanz anzeigen, auch wenn die Nullhypothese eigentlich gilt.

Beispiel: Führt man 20 Tests durch, mit Daten, die die Nullhypothese eigentlich erfüllen, wird im Schnitt ein Test Signifikanz auf dem 5%-Niveau anzeigen.

Bonferroni-Korrektur: Multipliziere alle p-Werte mit der Anzahl der Tests n. Wenn eines der Ergebnisse das Signifikanzniveau unterschreitet, verwirf die Nullhypothese

Nachteil: Konservativ: Häufig werden Abweichungen von der Nullhypothese nicht erkannt (Fehler zweiter Art).

Alternative: Teste ob es einen Wocheneffekt gibt, indem Du mit mit Varianzanalyse (ANOVA, ANalysis Of VAriance) ein Modell mit und eins ohne den Wocheneffekt vergleichst.

Alternative: Teste ob es einen Wocheneffekt gibt, indem Du mit mit Varianzanalyse (ANOVA, ANalysis Of VAriance) ein Modell mit und eins ohne den Wocheneffekt vergleichst.

Geht nur, wenn die Modelle eingebettet (engl. nested) sind, d.h. das einfachere Modell lässt sich erzeugen, indem man bei dem komplexeren bestimmte Randbdingungen für die Parameterwerte definiert, in unserem Fall “alle Wocheneffekte sind gleich”.

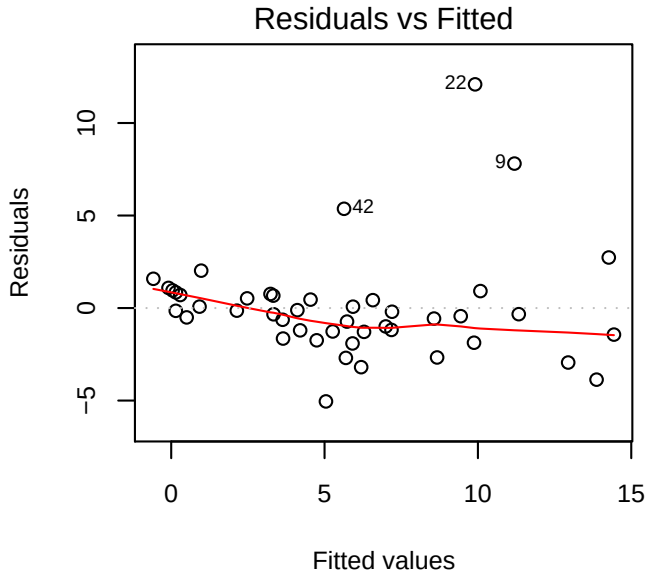
```
> modell0 <- lm(richness ~ angle2+NAP+grainsize+humus,
+                    data = rikz)
> modell <- lm(richness ~ angle2+NAP+grainsize+humus
+                           +factor(week), data = rikz)
> anova(modell0, modell)
Analysis of Variance Table

Model 1: richness ~ angle2 + NAP + grainsize + humus
Model 2: richness ~ angle2 + NAP + grainsize + humus + factor
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     40 531.17
2     37 353.66  3    177.51 6.1902 0.00162 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

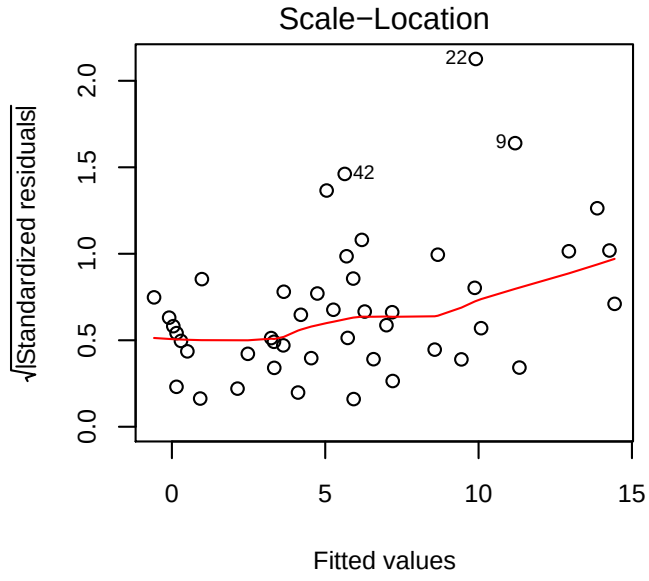
Also verwerfen wir die Nullhypothese, dass die Wochen keinen Effekt haben, mit dem p-Wert 0.00162.

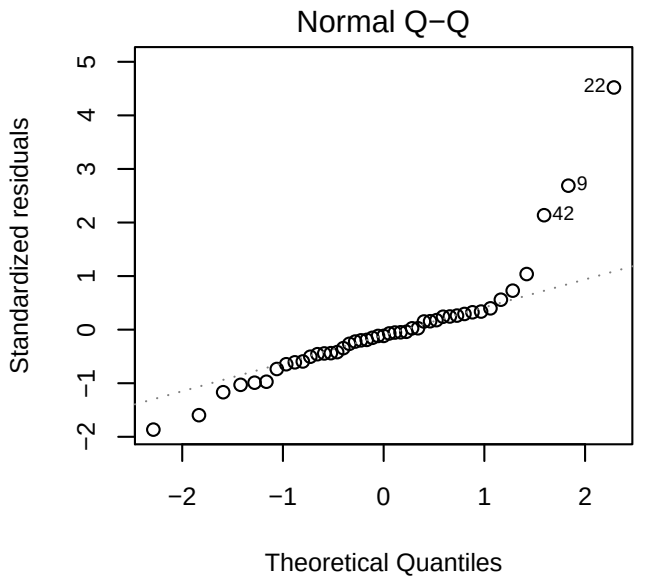
Also verwerfen wir die Nullhypothese, dass die Wochen keinen Effekt haben, mit dem p-Wert 0.00162.

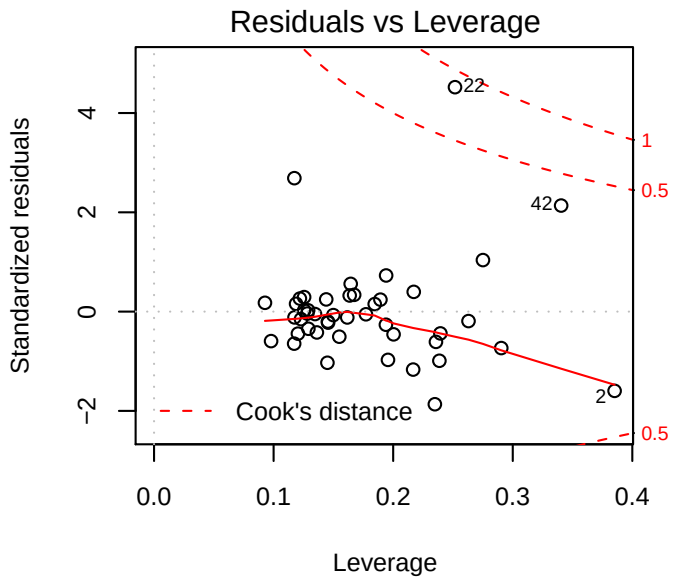
Aber Moment mal! Das können wir nur guten Gewissens tun, wenn das komplexere Modell gut passt. Das überprüfen wir graphisch.

```
plot(modell)
```


Residuals vs Fitted
Residuals
Fitted values
22
9
42


Scale−Location
√|Standardized residuals|
Fitted values
22
9
42


Normal Q−Q
Standardized residuals
Theoretical Quantiles
22
9
42


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
22
42
2
1
0.5
0.5

Als Ausreißer werden uns die Proben 22, 42, und 9 angezeigt.

Als Ausreißer werden uns die Proben 22, 42, und 9 angezeigt.

Können wir die durch Hinzunahme weiterer Parameter besser erklären oder handelt es sich um “echte Ausreißer”, die atypisch sind? Dann sollte man sie evtl. von der Analyse ausschließen und gesondert untersuchen.

Gibt es eine Interaktion zwischen NAP und angle2?

Gibt es eine Interaktion zwischen NAP und angle2?

$$\begin{aligned} \text{richness} = \; & a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ & + b_4 \cdot \text{humus} + \\ & + b_5 \cdot I_{\texttt{week=2}} + b_6 \cdot I_{\texttt{week=3}} + b_7 \cdot I_{\texttt{week=4}} \\ & b_8 \cdot \text{angle2} \cdot \text{NAP} + \varepsilon \end{aligned}$$

in R-Notation:

```
richness ~ angle2 + NAP + angle2:NAP+grainsize + humus
+ factor(week)
```

Gibt es eine Interaktion zwischen NAP und angle2?

$$\begin{aligned} \text{richness} = & \; a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + \\ & + b_4 \cdot \text{humus} + \\ & + b_5 \cdot I_{\texttt{week=2}} + b_6 \cdot I_{\texttt{week=3}} + b_7 \cdot I_{\texttt{week=4}} \\ & b_8 \cdot \text{angle2} \cdot \text{NAP} + \varepsilon \end{aligned}$$

in R-Notation:

```
richness ~ angle2 + NAP + angle2:NAP+grainsize + humus
+ factor(week)
```

oder auch so abgekürzt:

```
richness ~ angle2*NAP+grainsize + humus + factor(week)
```

```
> modell3 <- lm(richness ~ angle2*NAP+grainsize+humus
+                    +factor(week), data = rikz)
> summary(modell3)
[...]
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.438985   8.148756   1.281 0.208366
angle2         0.007846   0.044714   0.175 0.861697
NAP           -3.011876   1.099885  -2.738 0.009539 **
grainsize      0.001109   0.021236   0.052 0.958658
humus          0.387333   8.754526   0.044 0.964955
factor(week)2 -7.444863   1.839364  -4.048 0.000262 ***
factor(week)3 -6.052928   1.888789  -3.205 0.002831 **
factor(week)4 -1.854893   2.778334  -0.668 0.508629
angle2:NAP     0.013255   0.017292   0.767 0.448337
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Warnhinweise und Nebenwirkungen

Wendet man den R-Befehl `anova` auf ein einzelnes Modell an, werden die Variablen in der Reihenfolge, in der sie angegeben wurden, nach und nach hinzugefügt und die *p*-Werte beziehen sich jeweils darauf, ob das Modell durch das Hinzufügen dieses Parameters signifikant besser wird. Es wird also nur mit dem Modell verglichen, das aus den vorherigen Parametern besteht. Im Gegensatz dazu beziehen sich die *p*-Werte, die von `summary` oder dem Befehl `dropterm` aus der Bibliothek `MASS` ausgegeben werden immer auf einen Vergleich zwischen dem gegebenen Modell und einem Modell, bei dem ausschließlich die entsprechende Variable auf 0 gesetzt wird. Daher hängen die von `anova` gegebenen *p*-Werte von der Eingabereihenfolge ab, bei `summary` und `dropterm` aber nicht. Diese verschienen Optionen gibt es auch in anderen Statistik-Software-Paketen. Bei einigen muss man sich zwischen “Typ I”, “Typ II” und “Typ III” und zum Teil weiteren Anova-Typen entscheiden. In Zweifelsfällen fragen Sie Ihre(n) Hausstatistiker oder Übungsgruppenleiter(in).

Warnhinweise und Nebenwirkungen

Wendet man den R-Befehl `anova` auf ein einzelnes Modell an, werden die Variablen in der Reihenfolge, in der sie angegeben wurden, nach und nach hinzugefügt und die *p*-Werte beziehen sich jeweils darauf, ob das Modell durch das Hinzufügen dieses Parameters signifikant besser wird. Es wird also nur mit dem Modell verglichen, das aus den vorherigen Parametern besteht. Im Gegensatz dazu beziehen sich die *p*-Werte, die von `summary` oder dem Befehl `dropterm` aus der Bibliothek `MASS` ausgegeben werden immer auf einen Vergleich zwischen dem gegebenen Modell und einem Modell, bei dem ausschließlich die entsprechende Variable auf 0 gesetzt wird. Daher hängen die von `anova` gegebenen *p*-Werte von der Eingabereihenfolge ab, bei `summary` und `dropterm` aber nicht. Diese verschienen Optionen gibt es auch in anderen Statistik-Software-Paketen. Bei einigen muss man sich zwischen “Typ I”, “Typ II” und “Typ III” und zum Teil weiteren Anova-Typen entscheiden. In Zweifelsfällen fragen Sie Ihre(n) Hausstatistiker oder Übungsgruppenleiter(in).

Die nachfolgenden Beispiele sollen die Problematik verdeutlichen.

Hier wird zweimal das selbe Modell spezifiziert:

```
> modellA <- lm(richness ~ angle2+NAP+humus
+           +factor(week)+grainsize,data = rikz)
> modellB <- lm(richness ~ angle2+grainsize
+           +NAP+humus+factor(week), data = rikz)
```

Man beachte bei den folgenden Seiten den p-Wert von `grainsize`

```
> anova(modellA)
Analysis of Variance Table

Response: richness
                Df Sum Sq Mean Sq F value    Pr(>F)
angle2           1 124.86  124.86 13.0631 0.0008911 ***
NAP              1 319.32  319.32 33.4071 1.247e-06 ***
humus            1  35.18   35.18  3.6804 0.0627983 .
factor(week)     3 268.51   89.50  9.3638 9.723e-05 ***
grainsize        1   0.11    0.11  0.0114 0.9155704
Residuals       37 353.66    9.56
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> anova(modellB)
Analysis of Variance Table

Response: richness
             Df Sum Sq Mean Sq F value    Pr(>F)
angle2        1 124.86  124.86 13.0631 0.00089 ***
grainsize     1  35.97   35.97  3.7636 0.06003 .
NAP           1 390.11  390.11 40.8127 1.8e-07 ***
humus         1  19.53   19.53  2.0433 0.16127
factor(week) 3 177.51   59.17  6.1902 0.00162 **
Residuals    37 353.66    9.56
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> library(MASS)
> dropterm(modellA,test="F")
Single term deletions

Model:
richness ~ angle2 + NAP + humus + factor(week) + grainsize
             Df Sum of Sq    RSS    AIC  F Value     Pr(F)
<none>                    353.66 108.78
angle2        1      1.46 355.12 106.96     0.15 0.6984
NAP           1    176.37 530.03 124.98    18.45 0.0001 ***
humus         1      0.03 353.70 106.78 0.003565 0.9527
factor(week)3    177.51 531.17 121.08     6.19 0.0016 **
grainsize     1      0.11 353.77 106.79     0.01 0.9155
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> dropterm(modellB,test="F")
Single term deletions

Model:
richness ~ angle2 + grainsize + NAP + humus + factor(week
           Df Sum of Sq    RSS    AIC  F Value     Pr(F)
<none>                  353.66 108.78
angle2      1      1.46 355.12 106.96     0.15 0.6984
grainsize   1      0.11 353.77 106.79     0.01 0.9155
NAP         1    176.37 530.03 124.98    18.45 0.0001 ***
humus       1      0.03 353.70 106.78 0.003565 0.9527
factor(week)3    177.51 531.17 121.08     6.19 0.0016 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> summary(modellA)
[...]
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.298448   7.967002   1.167 0.2506
angle2          0.016760   0.042934   0.390 0.6984
NAP            -2.274093   0.529411  -4.296 0.0001 ***
humus           0.519686   8.703910   0.060 0.9527
factor(week)2 -7.065098   1.761492  -4.011 0.0002 ***
factor(week)3 -5.719055   1.827616  -3.129 0.0034 **
factor(week)4 -1.481816   2.720089  -0.545 0.5891
grainsize       0.002249   0.021066   0.107 0.9155
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> summary(modellB)
[...]
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.298448   7.967002   1.167 0.2506
angle2          0.016760   0.042934   0.390 0.6984
grainsize       0.002249   0.021066   0.107 0.9155
NAP            -2.274093   0.529411  -4.296 0.0001 ***
humus           0.519686   8.703910   0.060 0.9527
factor(week)2 -7.065098   1.761492  -4.011 0.0002 ***
factor(week)3 -5.719055   1.827616  -3.129 0.0034 **
factor(week)4 -1.481816   2.720089  -0.545 0.5891
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

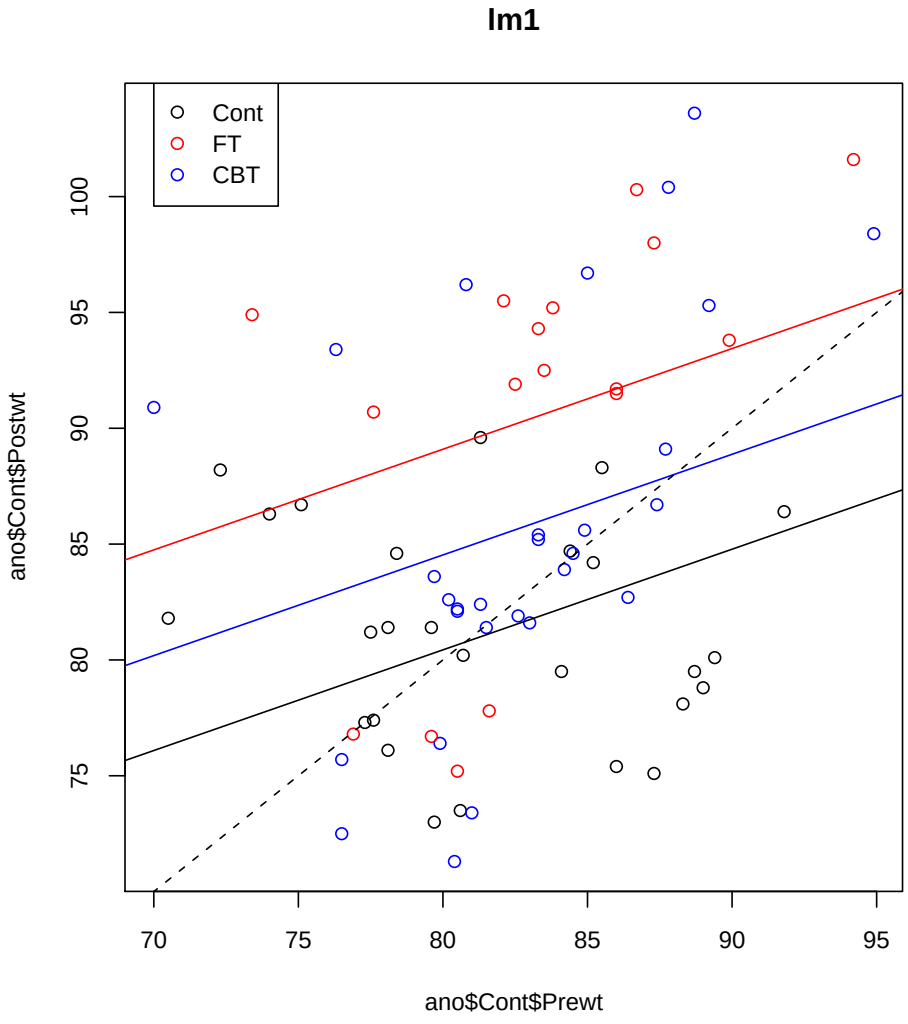
Inhalt

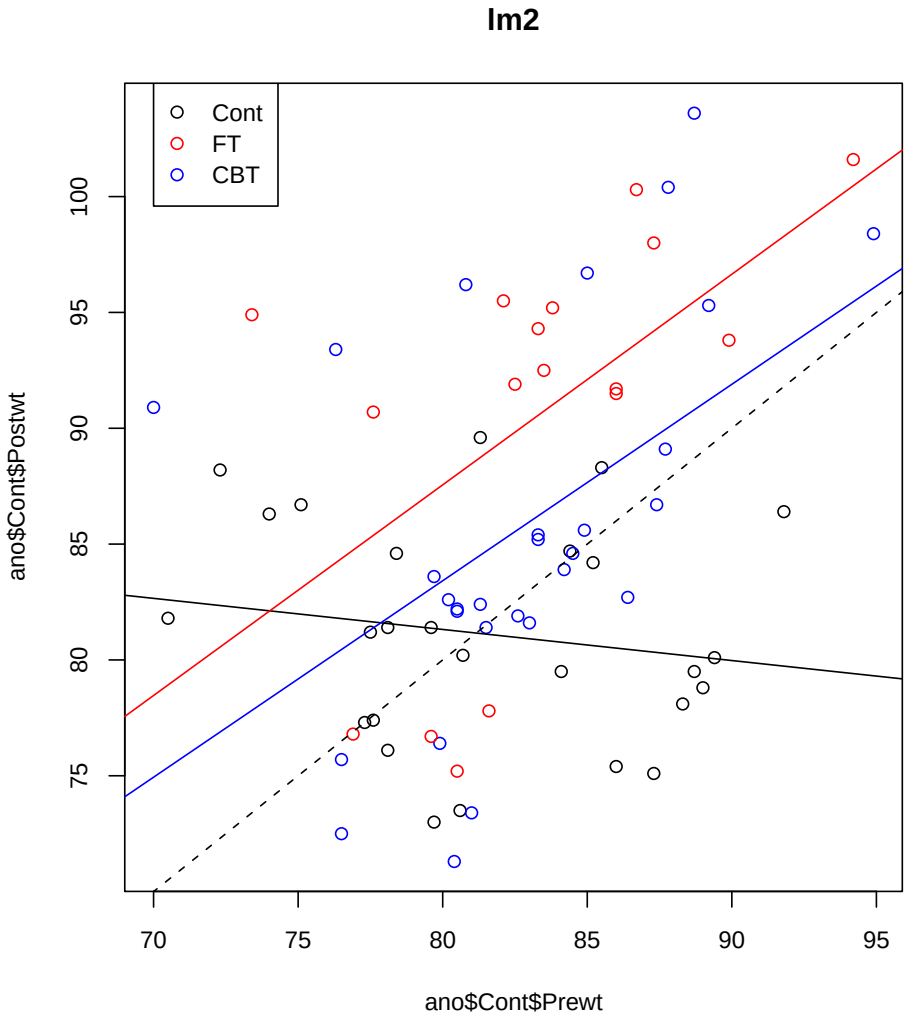
Vergleiche bei jugendlichen Magersuchtpatientinnen den Behandlungserfolg von Familientherapie (FT) und kognitiver Verhaltenstherapie (CBT) mit einer Kontrollgruppe (Cont), indem das Gewicht (in lbs.) vor (Prewt) und nach (Postwt) der Behandlung (Treat) vergleichen wird.

Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. eds (1993) *A Handbook of Small Data Sets.* Chapman & Hall

Modell lm1 Es gibt zusätzlich einen linearen Zusammenhang mit dem Gewicht vor der Therapie. Jede Behandlungsform erhöht (oder vermindert) das Gewicht um einen Wert, der von der Behandlung, aber nicht vom Gewicht vor der Behandlung abhängt.

Modell lm2 Interaktion zwischen Treat und Prwt: Das Gewicht vor der Behandlung wirkt sich bei den verschiedenen Behandlungsarten (einschließlich “keine Therapie”) unterschiedlich stark aus.


lm1
Cont
FT
CBT
ano$Cont$Postwt
ano$Cont$Prewt
70
75
80
85
90
95
75
80
85
90
95
100


lm2
Cont
FT
CBT
ano$Cont$Postwt
ano$Cont$Prewt
70
75
80
85
90
95
75
80
85
90
95
100

```
> lm1 <- lm(Postwt~Prewt+Treat,anorexia)
> lm2 <- lm(Postwt~Prewt*Treat,anorexia)
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: Postwt ~ Prewt + Treat
Model 2: Postwt ~ Prewt * Treat
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     68 3311.3
2     66 2844.8  2     466.5 5.4112 0.006666 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Ergebnis: Das komplexere Modell passt signifikant besser auf die Daten als das eingebettete Modell.

Ergebnis: Das komplexere Modell passt signifikant besser auf die Daten als das eingebettete Modell.

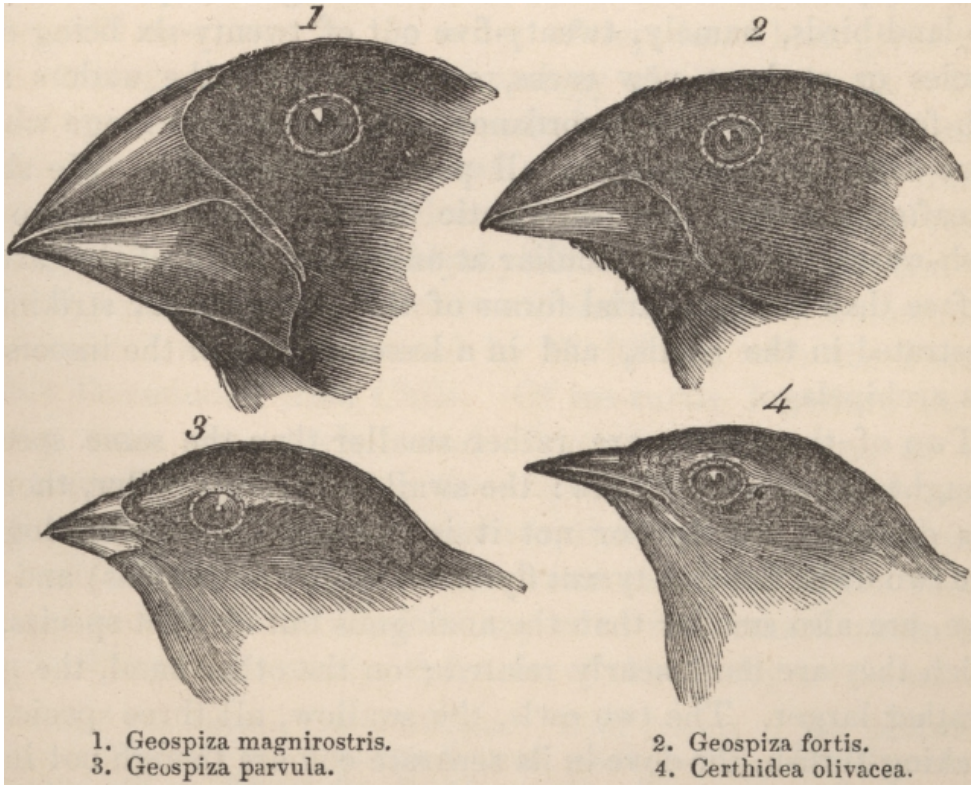
Interpretation: Welche Rolle das Gewicht vor der Behandlung spielt, hängt von der Behandlung ab.

Ergebnis: Das komplexere Modell passt signifikant besser auf die Daten als das eingebettete Modell.

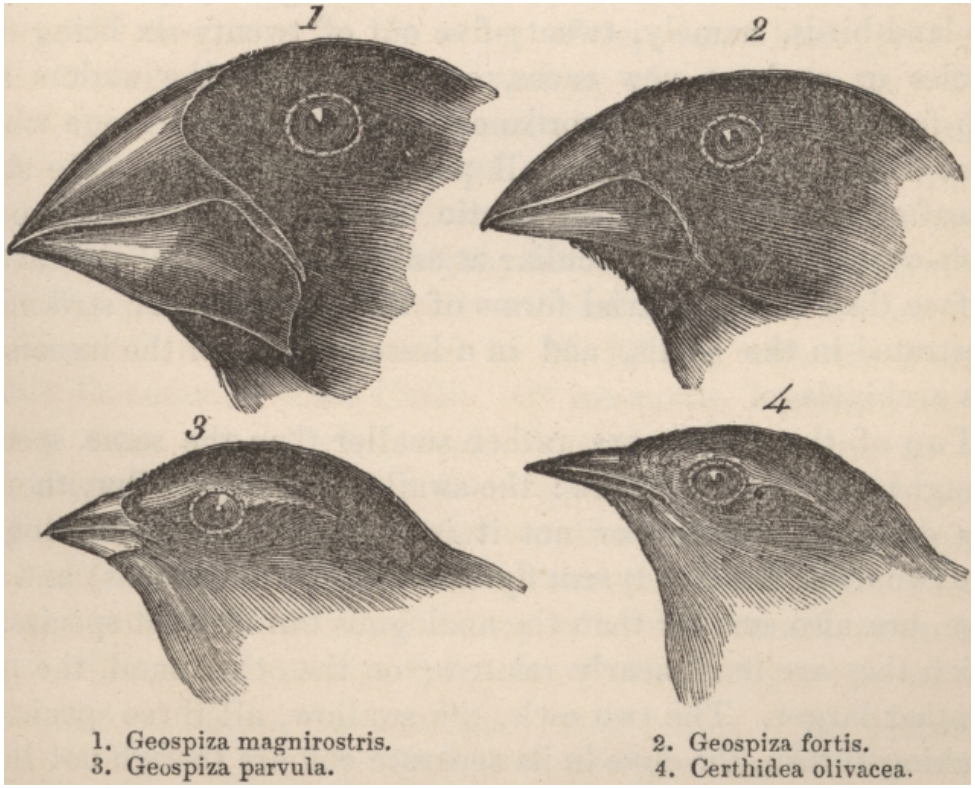
Interpretation: Welche Rolle das Gewicht vor der Behandlung spielt, hängt von der Behandlung ab.
oder auch: Der Unterschied zwischen den Wirkungen der verschiedenen Behandlungen hängt vom Gewicht vor der Therapie ab.

Inhalt

Inhalt

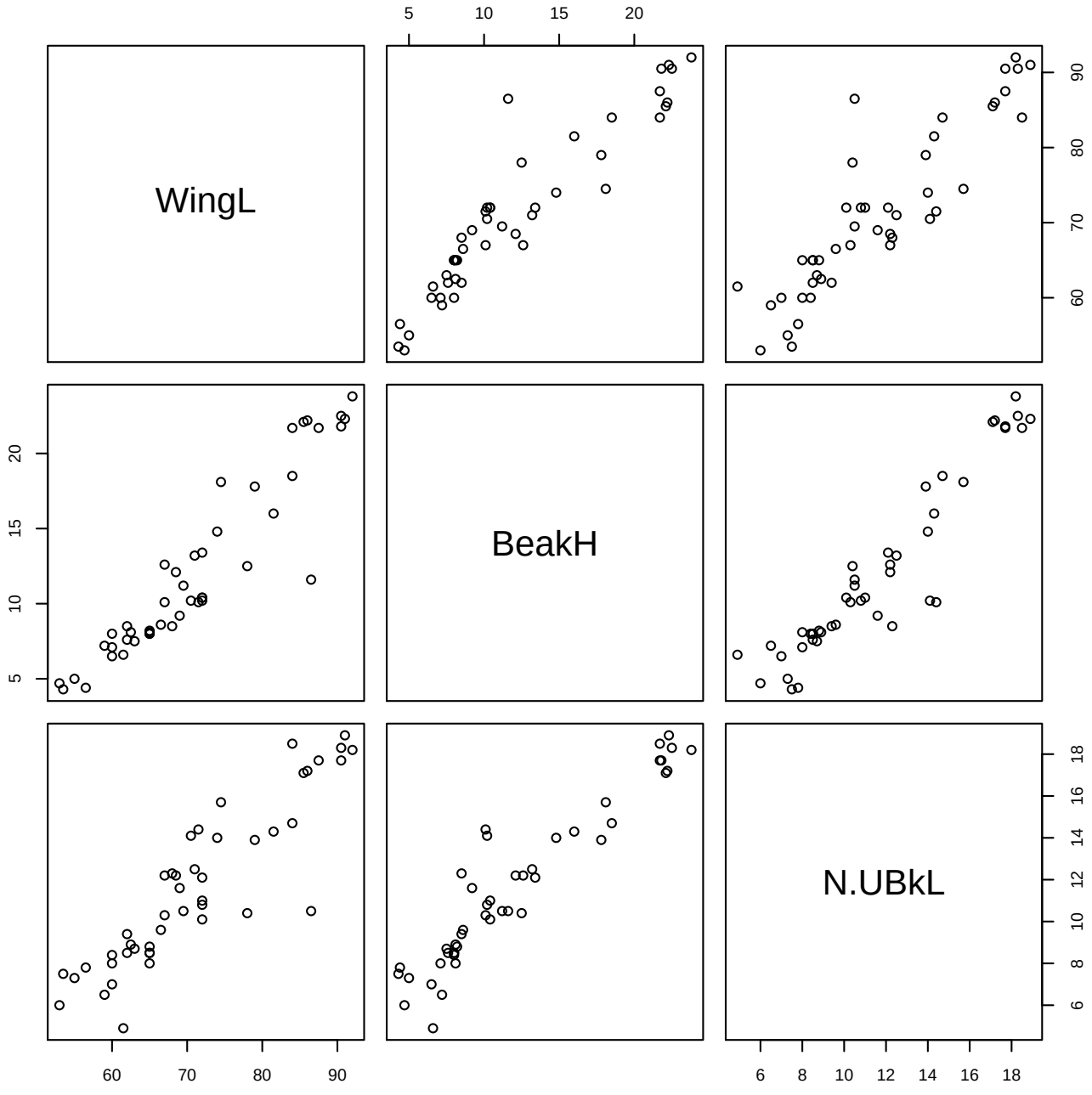


Sie finden den Schnabel eines Darwinfinken. Der Schnabel ist 14 mm lang und 10 mm hoch. Wie gut können Sie die Spannweite des Vogels schätzen?



Sie finden den Schnabel eines Darwinfinken. Der Schnabel ist 14 mm lang und 10 mm hoch. Wie gut können Sie die Spannweite des Vogels schätzen?

Als “Lerndaten” stehen Ihnen Spannweiten (WingL), Schnabelhöhen (BeakH) und Schnabellängen (N.UBkL) von 46 Darwinfinken zur Verfügung.


WingL
BeakH
N.UBkL
5
10
15
20
60
70
80
90
6
8
10
12
14
16
18

Sollen wir nur die Schnabelhöhe, nur die Schnabellänge oder beides einbeziehen?

```
> modH <- lm(WingL~BeakH)
> summary(modH)

Call:
lm(formula = WingL ~ BeakH)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1882 -2.5327 -0.2796  1.8325 16.2702

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.78083    1.33103   37.40   <2e-16 ***
BeakH        1.76284    0.09961   17.70   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.868 on 44 degrees of freedom
Multiple R-squared: 0.8768, Adjusted R-squared: 0.874
F-statistic: 313.2 on 1 and 44 DF,  p-value: < 2.2e-16

> predict(modH,newdata=data.frame(BeakH=10))
       1
67.40924
```

```
> modL <- lm(WingL~N.UBkL)
> summary(modL)

Call:
lm(formula = WingL ~ N.UBkL)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1321 -3.3974  0.4737  2.2966 18.2299

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.5371     2.2884   18.15   <2e-16 ***
N.UBkL        2.5460     0.1875   13.58   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.838 on 44 degrees of freedom
Multiple R-squared: 0.8074, Adjusted R-squared: 0.803
F-statistic: 184.4 on 1 and 44 DF,  p-value: < 2.2e-16

> predict(modL,newdata=data.frame(N.UBkL=14))
       1
77.18117
```

```
> modHL <- lm(WingL~BeakH+N.UBkL)
> summary(modHL)

Call:
lm(formula = WingL ~ BeakH + N.UBkL)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3185 -2.5022 -0.2752  1.5352 16.5893

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.1740     2.2572  21.343  < 2e-16 ***
BeakH         1.5133     0.2999   5.047 8.69e-06 ***
N.UBkL        0.3984     0.4513   0.883    0.382
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.878 on 43 degrees of freedom
Multiple R-squared: 0.879, Adjusted R-squared: 0.8734
F-statistic: 156.2 on 2 and 43 DF,  p-value: < 2.2e-16

> predict(modHL,newdata=data.frame(BeakH=10,N.UBkL=14))
       1
68.88373
```

Welche der drei Vorhersagen 67.4mm, 77.2mm und 68.9mm für die Flügellänge ist am genauesten?

Welche der drei Vorhersagen 67.4mm, 77.2mm und 68.9mm für die Flügellänge ist am genauesten?

Im Modell modHL (mit Schnabellänge und -höhe) ist der Einfluss der Schnabellänge nicht signifikant.

Welche der drei Vorhersagen 67.4mm, 77.2mm und 68.9mm für die Flügellänge ist am genauesten?

Im Modell modHL (mit Schnabellänge und -höhe) ist der Einfluss der Schnabellänge nicht signifikant.

Das muss aber nichts heißen, denn aus Nichtsignifikanz kann man keine Schlüsse ziehen. Die Schnabellänge könnte die Vorhersage verbessern.

Welche der drei Vorhersagen 67.4mm, 77.2mm und 68.9mm für die Flügellänge ist am genauesten?

Im Modell modHL (mit Schnabellänge und -höhe) ist der Einfluss der Schnabellänge nicht signifikant.

Das muss aber nichts heißen, denn aus Nichtsignifikanz kann man keine Schlüsse ziehen. Die Schnabellänge könnte die Vorhersage verbessern.

Sollte man einfach alle verfügbaren Daten einbeziehen?

Welche der drei Vorhersagen 67.4mm, 77.2mm und 68.9mm für die Flügellänge ist am genauesten?

Im Modell modHL (mit Schnabellänge und -höhe) ist der Einfluss der Schnabellänge nicht signifikant.

Das muss aber nichts heißen, denn aus Nichtsignifikanz kann man keine Schlüsse ziehen. Die Schnabellänge könnte die Vorhersage verbessern.

Sollte man einfach alle verfügbaren Daten einbeziehen?

Problem könnte “overfitting” sein: Wenn sehr viele Parameter verfügbar sind, wird das Modell auch an die Zufallsschwankungen angepasst. Die Daten werden sozusagen auswendig gelernt. Vorhersagen für andere Daten werden dann schlechter.

Wir könnten die Modelle anhand der Standardabweichung der ε_i verwenden, die wir aus der Standardabweichung der Residuen r_i schätzen.

Wir könnten die Modelle anhand der Standardabweichung der ε_i verwenden, die wir aus der Standardabweichung der Residuen r_i schätzen.

Dabei müssen wir der Unterschiedlichen Anzahl d an Modellparametern Rechnung tragen, denn für jeden geschätzen Parameter verlieren wir einen Freiheitsgrad:

$$\widehat{\sigma}_\epsilon = \sqrt{\frac{1}{n-d}\sum_i r_i^2} = \sigma_r \cdot \sqrt{\frac{n-1}{n-d}}.$$

Wir könnten die Modelle anhand der Standardabweichung der ε_i verwenden, die wir aus der Standardabweichung der Residuen r_i schätzen.

Dabei müssen wir der Unterschiedlichen Anzahl d an Modellparametern Rechnung tragen, denn für jeden geschätzen Parameter verlieren wir einen Freiheitsgrad:

$$\widehat{\sigma}_\epsilon = \sqrt{\frac{1}{n-d}\sum_i r_i^2} = \sigma_r \cdot \sqrt{\frac{n-1}{n-d}}.$$

Diese Werte werden bei R vom Befehl summary ausgegeben:
modH:

```
Residual standard error: 3.868 on 44 degrees of freedom
```

modL:

```
Residual standard error: 4.838 on 44 degrees of freedom
```

modHL:

```
Residual standard error: 3.878 on 43 degrees of freedom
```

Eine weitere Möglichkeit, die Vorhersagegenauigkeit eines Modells zu beurteilen, ist die *Kreuzvalidierung* (auch *Jackknife* genannt).

Eine weitere Möglichkeit, die Vorhersagegenauigkeit eines Modells zu beurteilen, ist die *Kreuzvalidierung* (auch *Jackknife* genannt).

Idee: Entferne einen der 46 Vögel aus dem Datensatz und passe das Modell an die anderen 45 an. Wie gut kann man mit dem so angepassten Modell die Flügellänge des einen Vogels vorhersagen?

Eine weitere Möglichkeit, die Vorhersagegenauigkeit eines Modells zu beurteilen, ist die *Kreuzvalidierung* (auch *Jackknife* genannt).

Idee: Entferne einen der 46 Vögel aus dem Datensatz und passe das Modell an die anderen 45 an. Wie gut kann man mit dem so angepassten Modell die Flügellänge des einen Vogels vorhersagen?

Wiederhole das für alle 46 Vögel.

Eine weitere Möglichkeit, die Vorhersagegenauigkeit eines Modells zu beurteilen, ist die *Kreuzvalidierung* (auch *Jackknife* genannt).

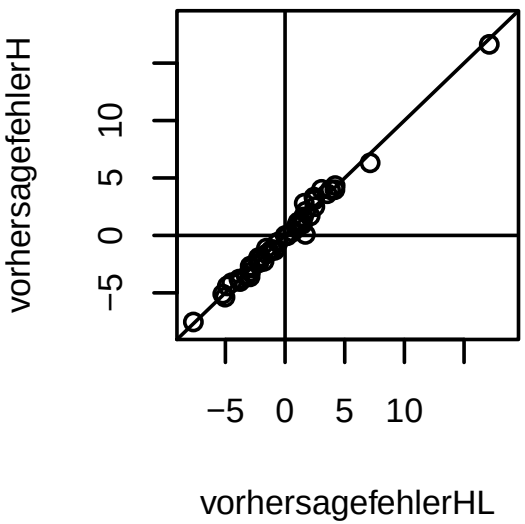
Idee: Entferne einen der 46 Vögel aus dem Datensatz und passe das Modell an die anderen 45 an. Wie gut kann man mit dem so angepassten Modell die Flügellänge des einen Vogels vorhersagen?

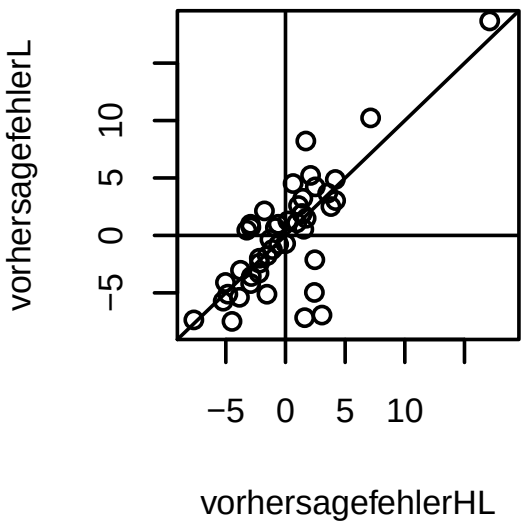
Wiederhole das für alle 46 Vögel.

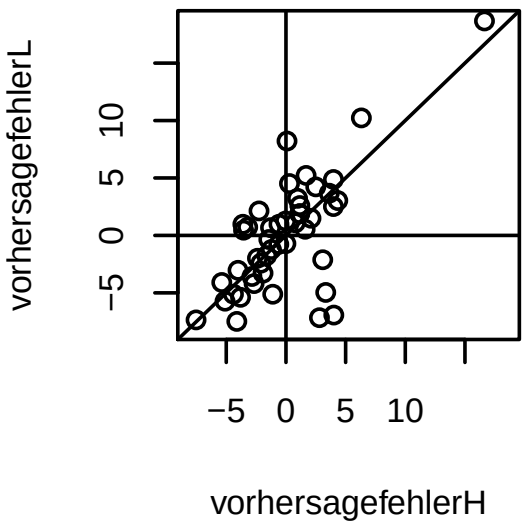
Man muss dann entscheiden, wie Fehler “bestraft” werden. (Ist ein Modell, das häufig kleine Fehler macht besser als eins, das selten große macht?) Wir verwenden hier die Wurzel aus der Summe der quadrierten Fehler.

```
> vorhersagefehlerH <- numeric()
> for (i in 1:46) {
+   selection <- rep(TRUE,46)
+   selection[i] <- FALSE
+   modH.R <- lm(WingL~BeakH,subset=selection)
+   vorhersagefehlerH[i] <- WingL[i]-predict(modH.R,
+                                            finken2[i,])
+ }
> sqrt(sum(vorhersagefehlerH^2))
[1] 26.55519
```

Vergleich der Vorhersagefehler







	Höhe	Länge	Höhe und Länge
σ(Residuen)	3.83	4.78	3.79

	Höhe	Länge	Höhe und Länge
σ(Residuen)	3.83	4.78	3.79
$d =$ (Anzahl Parameter)	2	2	3

	Höhe	Länge	Höhe und Länge
σ(Residuen)	3.83	4.78	3.79
$d =$ (Anzahl Parameter)	2	2	3
σ(Residuen)$\cdot\sqrt{\frac{n-1}{n-d}}$	3.87	4.84	3.88

	Höhe	Länge	Höhe und Länge
σ(Residuen)	3.83	4.78	3.79
$d =$ (Anzahl Parameter)	2	2	3
σ(Residuen)$\cdot\sqrt{\frac{n-1}{n-d}}$	3.87	4.84	3.88
Kreuzvalid.	26.56	33.34	26.68

	Höhe	Länge	Höhe und Länge
σ(Residuen)	3.83	4.78	3.79
$d =$ (Anzahl Parameter)	2	2	3
$\sigma(\text{Residuen}) \cdot \sqrt{\frac{n-1}{n-d}}$	3.87	4.84	3.88
Kreuzvalid.	26.56	33.34	26.68
AIC	259.0	279.5	260.1

	Höhe	Länge	Höhe und Länge
σ(Residuen)	3.83	4.78	3.79
$d =$ (Anzahl Parameter)	2	2	3
$\sigma(\text{Residuen}) \cdot \sqrt{\frac{n-1}{n-d}}$	3.87	4.84	3.88
Kreuzvalid.	26.56	33.34	26.68
AIC	259.0	279.5	260.1

Akaikes Informationskriterium:

$$\mathrm{AIC} = -2 \cdot \log L + 2 \cdot (\mathrm{AnzahlParameter})$$

Bayessches Informationskriterium:

$$\mathrm{BIC} = -2 \cdot \log L + \log(n) \cdot (\mathrm{AnzahlParameter})$$

Dabei ist n die Anzahl der Beobachtungen.

	Höhe	Länge	Höhe und Länge
σ(Residuen)	3.83	4.78	3.79
$d =$ (Anzahl Parameter)	2	2	3
$\sigma(\text{Residuen}) \cdot \sqrt{\frac{n-1}{n-d}}$	3.87	4.84	3.88
Kreuzvalid.	26.56	33.34	26.68
AIC	259.0	279.5	260.1

Akaikes Informationskriterium:

$$\mathrm{AIC} = -2 \cdot \log L + 2 \cdot (\mathrm{AnzahlParameter})$$

Bayessches Informationskriterium:

$$\mathrm{BIC} = -2 \cdot \log L + \log(n) \cdot (\mathrm{AnzahlParameter})$$

Dabei ist n die Anzahl der Beobachtungen. Für $n \geq 8$ ist $\log(n) > 2$ und BIC bestraft jeden zusätzlichen Parameter stärker als AIC. (Mit log ist wie immer der natürliche Logarithmus gemeint.)

Niedrige Werte von AIC und BIC sprechen für das Modell.

Niedrige Werte von AIC und BIC sprechen für das Modell. (Zumindest in R. Manche Programme und Autoren geben AIC und BIC mit umgekehrtem Vorzeichen an.)

Niedrige Werte von AIC und BIC sprechen für das Modell. (Zumindest in R. Manche Programme und Autoren geben AIC und BIC mit umgekehrtem Vorzeichen an.)

AIC basiert auf der Idee, dass ein mit Daten angepasstes Modell bei neuen Daten möglichst präzise Vorhersagen ermöglichen soll. AIC approximiert den Vorhersagefehler für neue Daten.

Niedrige Werte von AIC und BIC sprechen für das Modell. (Zumindest in R. Manche Programme und Autoren geben AIC und BIC mit umgekehrtem Vorzeichen an.)

AIC basiert auf der Idee, dass ein mit Daten angepasstes Modell bei neuen Daten möglichst präzise Vorhersagen ermöglichen soll. AIC approximiert den Vorhersagefehler für neue Daten.

BIC approximiert (bis auf eine Konstante) die logarithmierte a-posteriori-Wahrscheinlichkeit des Modells, wobei a priori alle Modelle als gleich wahrscheinlich angenommen werden.

	Höhe	Länge	Höhe und Länge
σ(Residuen)	3.83	4.78	3.79
$d =$ (Anzahl Parameter)	2	2	3
$\sigma(\text{Residuen}) \cdot \sqrt{\frac{n-1}{n-d}}$	3.87	4.84	3.88
Kreuzvalid.	26.56	33.34	26.68
AIC	259.0	279.5	260.1
BIC	264.4	285.0	267.4

Hier spricht alles dafür, nur die Schnabelhöhe zu berücksichtigen.

Inhalt

Fragestellung: reagieren *Daphnia magna* anders auf das Nahrungsangebot als *Daphnia galeata*?

Fragestellung: reagieren *Daphnia magna* anders auf das Nahrungsangebot als *Daphnia galeata*?

Die Daten wurden im Ökologie-Kurs 2009 erhoben und von Justina Wolinska zur Verfügung gestellt.

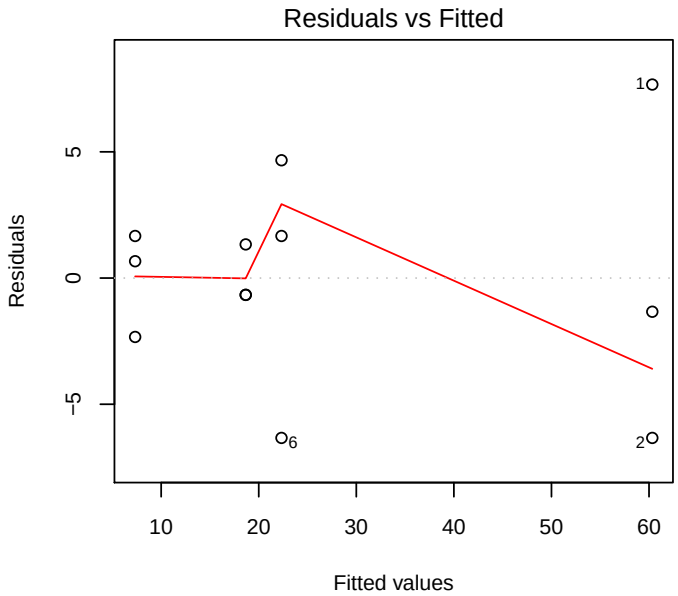
```
> daph <- read.table("daphnia_justina.csv",h=T)
> daph
   counts foodlevel  species
1      68     high     magna
2      54     high     magna
3      59     high     magna
4      24     high   galeata
5      27     high   galeata
6      16     high   galeata
7      20      low     magna
8      18      low     magna
9      18      low     magna
10      5      low   galeata
11      8      low   galeata
12      9      low   galeata
```

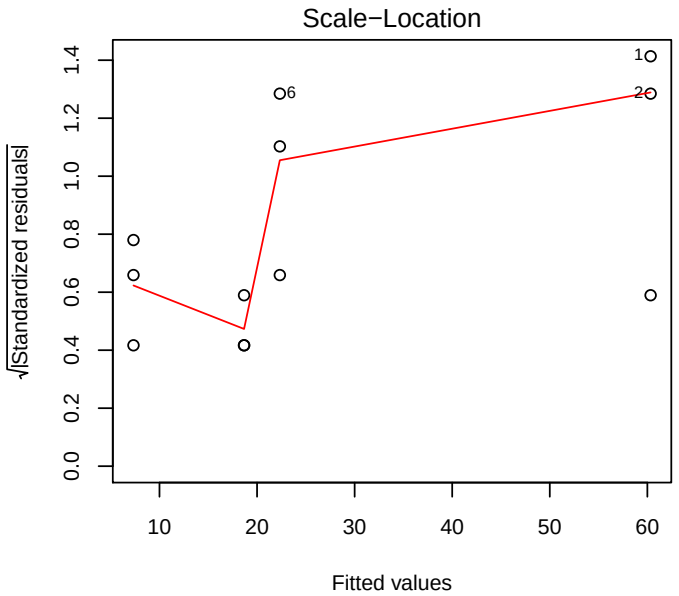
```
> mod1 <- lm(counts~foodlevel+species,data=daph)
> mod2 <- lm(counts~foodlevel*species,data=daph)
> anova(mod1,mod2)
Analysis of Variance Table

Model 1: counts ~ foodlevel + species
Model 2: counts ~ foodlevel * species
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      9 710.00
2      8 176.67  1    533.33 24.151 0.001172 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

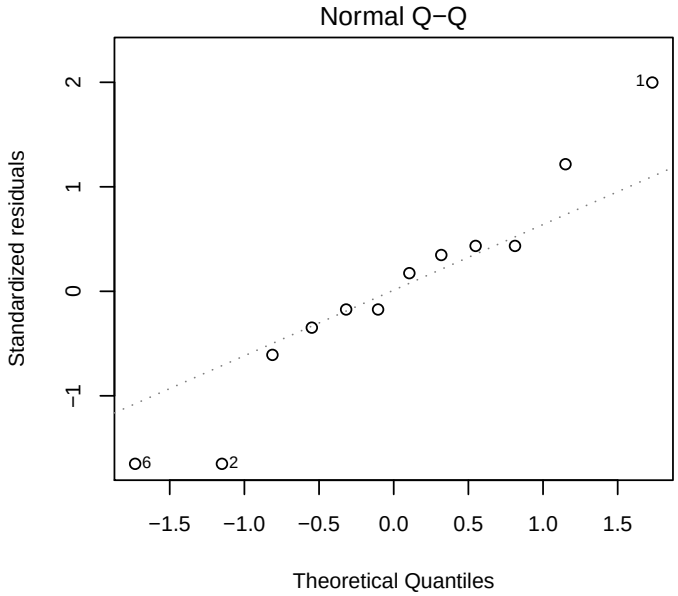
```
> summary(mod2)
[...]
Coefficients:
                          Estimate  Std.Error t.value Pr(>|t|)
(Intercept)                  22.33  2.713    8.232 3.55e-05 ***
countslow                   -15.00  3.837   -3.909  0.00449 **
foodlevelmagna               38.00  3.837    9.904 9.12e-06 ***
countslow:foodlevelmagna -26.67  5.426   -4.914  0.00117 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.699 on 8 degrees of freedom
Multiple R-squared: 0.9643, Adjusted R-squared: 0.9509
F-statistic: 71.95 on 3 and 8 DF,  p-value: 3.956e-06
```

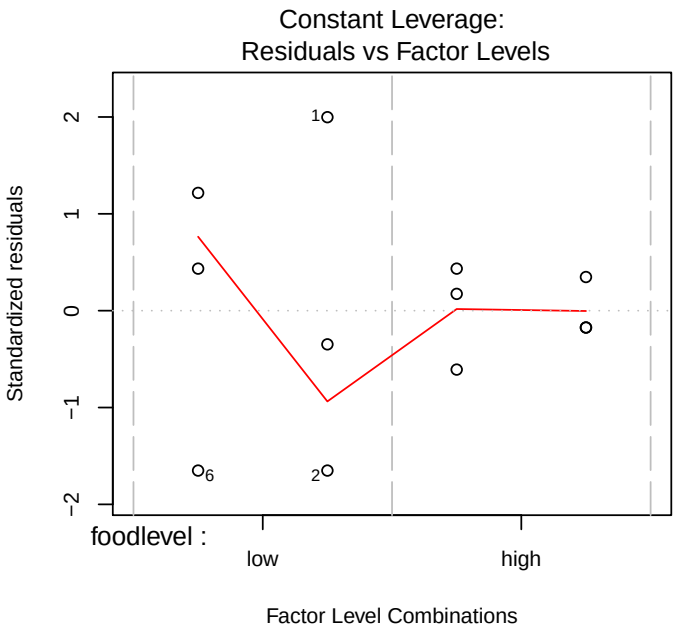
Ergebnis: das komplexere Modell, in dem die verschiedenen Arten auf unterschiedliche Weise auf Nahrungsknappheit reagieren, passt signifikant besser auf die Daten.

Ergebnis: das komplexere Modell, in dem die verschiedenen Arten auf unterschiedliche Weise auf Nahrungsknappheit reagieren, passt signifikant besser auf die Daten. Aber passt es gut genug?


Residuals vs Fitted
Residuals
Fitted values
10
20
30
40
50
60
−5
0
5
1
6
2

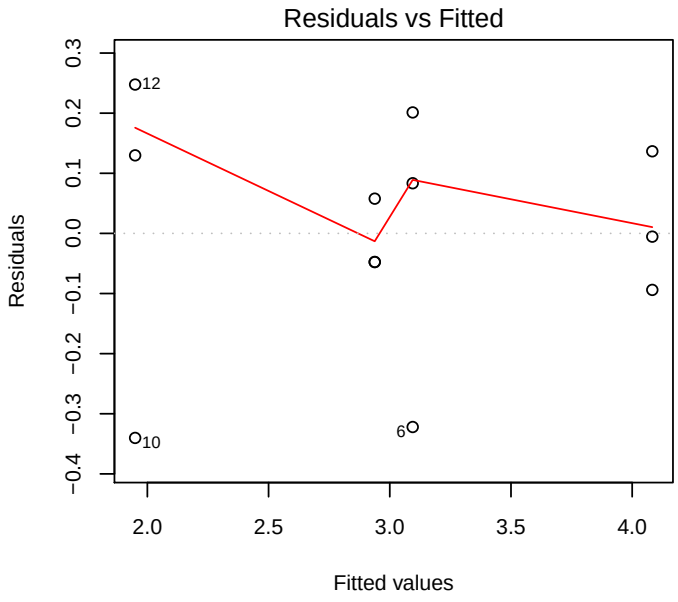

Scale−Location
√|Standardized residuals|
Fitted values
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1
6
2

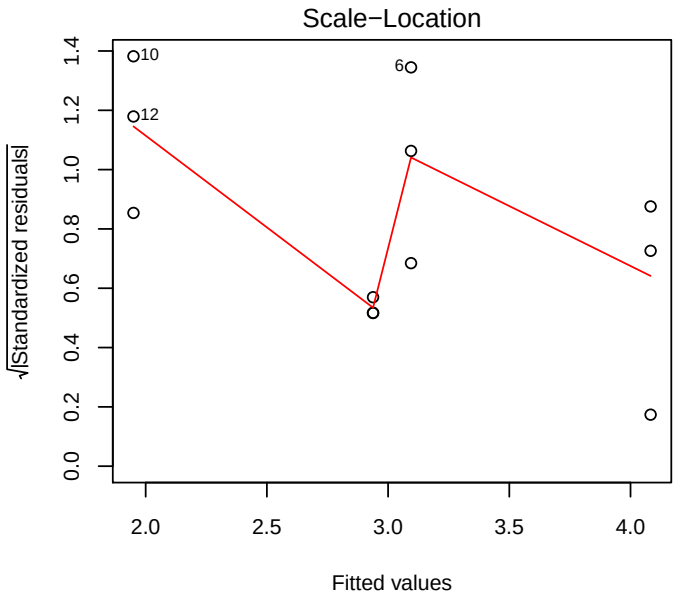

Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−1
0
1
2
1
6
2

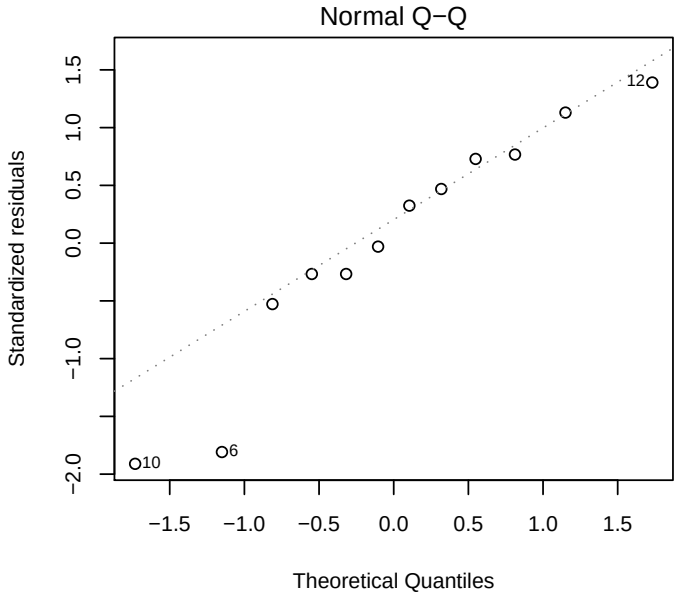

Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
Factor Level Combinations
foodlevel :
low
high
−2
−1
0
1
2
1
6
2

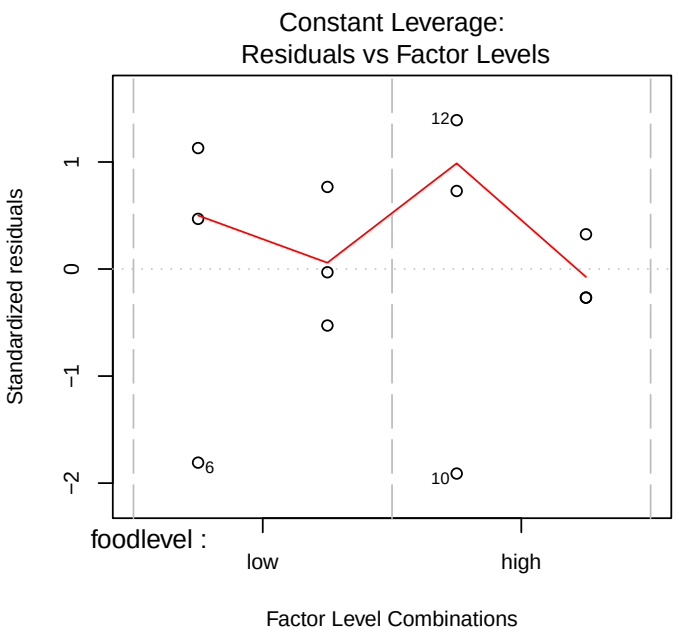
```
> mod3 <- lm(log(counts)~foodlevel+species,data=daph)
> mod4 <- lm(log(counts)~foodlevel*species,data=daph)
> anova(mod3,mod4)
Analysis of Variance Table

Model 1: log(counts) ~ foodlevel + species
Model 2: log(counts) ~ foodlevel * species
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      9 0.38041
2      8 0.37856  1 0.0018545 0.0392  0.848
```

```
> summary(mod3)

Call:
lm(formula = log(counts) ~ foodlevel + species, data = daph)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34017 -0.05915  0.02622  0.13153  0.24762

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.0946     0.1028  30.104 2.41e-10 ***
foodlevellow  -1.1450     0.1187  -9.646 4.83e-06 ***
speciesmagna   0.9883     0.1187   8.326 1.61e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2056 on 9 degrees of freedom
Multiple R-squared: 0.9475, Adjusted R-squared: 0.9358
F-statistic: 81.19 on 2 and 9 DF,  p-value: 1.743e-06
```


Residuals vs Fitted
Residuals
Fitted values
12
10
6
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
2.0
2.5
3.0
3.5
4.0


Scale−Location
√|Standardized residuals|
Fitted values
10
6
12
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
2.0
2.5
3.0
3.5
4.0


Normal Q−Q
Standardized residuals
Theoretical Quantiles
12
10
6
−2.0
−1.0
0.0
0.5
1.0
1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Constant Leverage:
Residuals vs Factor Levels
Standardized residuals
Factor Level Combinations
foodlevel :
low
high
12
6
10
−2
−1
0
1

Der qqplot sieht schon etwas besser aus, aber nicht ganz optimal.

Der qqplot sieht schon etwas besser aus, aber nicht ganz optimal.

Das liegt aber auch daran, dass wir es hier bei der Zielvariable `counts` z.T. mit kleinen ganzen Zahlen zu tun haben, auf die die Normalverteilungsannahme eigentlich nicht ganz passt.

Der qqplot sieht schon etwas besser aus, aber nicht ganz optimal.

Das liegt aber auch daran, dass wir es hier bei der Zielvariable `counts` z.T. mit kleinen ganzen Zahlen zu tun haben, auf die die Normalverteilungsannahme eigentlich nicht ganz passt.

Statt des normalen linearen Modells könnte man ein verallgemeinert-lineares Modell vom Typ Poisson mit log-Transformation anwenden, aber das geht über den Inhalt der Vorlesung hinaus.

Der qqplot sieht schon etwas besser aus, aber nicht ganz optimal.

Das liegt aber auch daran, dass wir es hier bei der Zielvariable `counts` z.T. mit kleinen ganzen Zahlen zu tun haben, auf die die Normalverteilungsannahme eigentlich nicht ganz passt.

Statt des normalen linearen Modells könnte man ein verallgemeinert-lineares Modell vom Typ Poisson mit log-Transformation anwenden, aber das geht über den Inhalt der Vorlesung hinaus.

Wir begnügen uns daher mit den normalen linearen Modellen und entscheiden uns für eines der bisher betrachteten.

```
> AIC(mod1,mod2,mod3,mod4)
     df         AIC
mod1  4 91.0188246
mod2  5 76.3268216
mod3  4  0.6376449
mod4  5  2.5790019
```

Die log-linearen Modelle haben deutlich bessere AIC-Werte als die linearen Modelle mit den nicht-transformierten Daten. (Allerdings sollte man AIC-Werten eigentlich nicht vergleichen, wenn die Zielvariable sich unterscheided oder unterschiedlich transformiert ist.)

```
> AIC(mod1,mod2,mod3,mod4)
     df         AIC
mod1  4 91.0188246
mod2  5 76.3268216
mod3  4  0.6376449
mod4  5  2.5790019
```

Die log-linearen Modelle haben deutlich bessere AIC-Werte als die linearen Modelle mit den nicht-transformierten Daten. (Allerdings sollte man AIC-Werten eigentlich nicht vergleichen, wenn die Zielvariable sich unterscheided oder unterschiedlich transformiert ist.)

Die Interaktion in Modell mod4 ist nicht nur nicht-signifikant, das Modell mod3 ohne Interaktion mod3 hat auch einen besseren AIC-Wert.

Vieles spricht also für mod3:

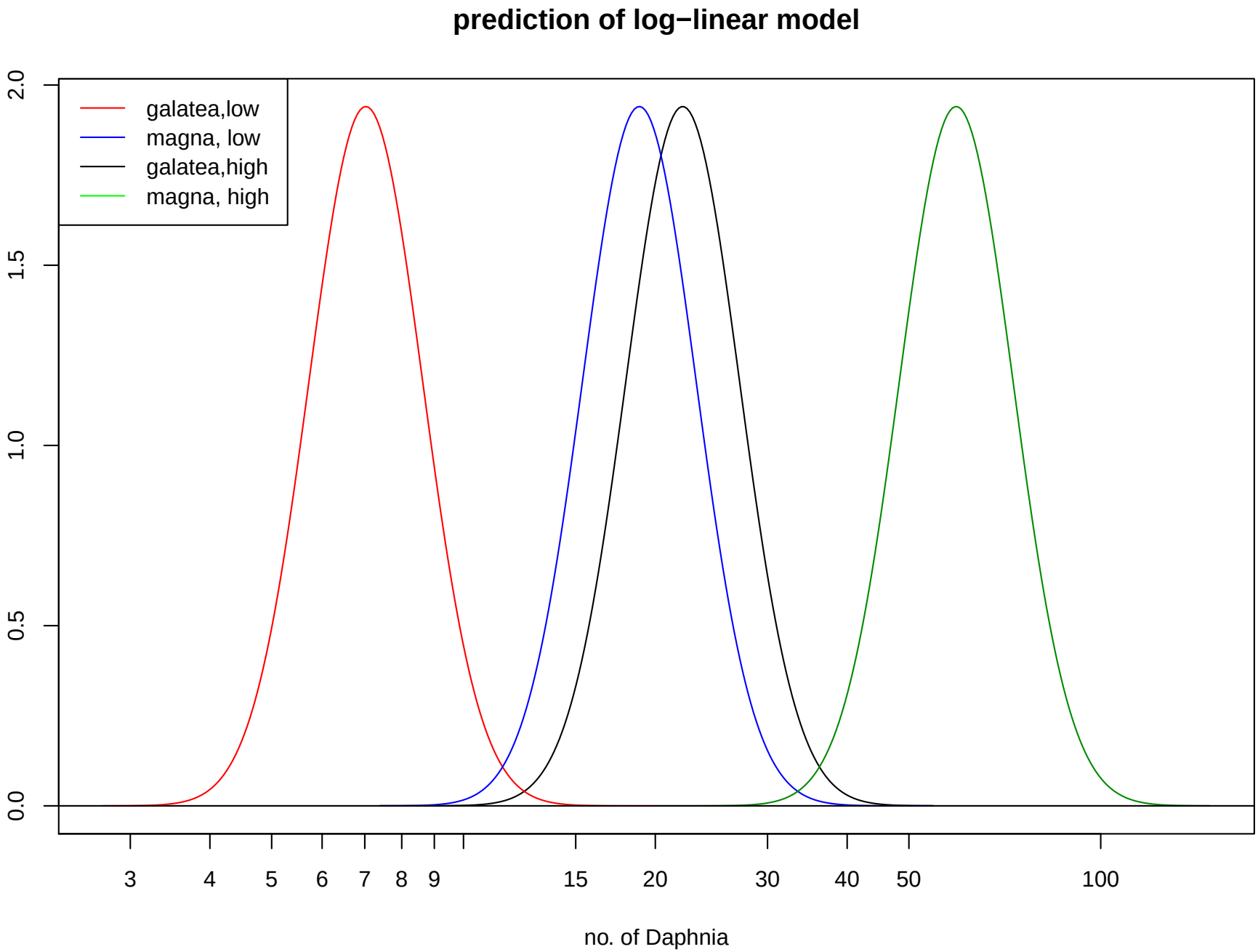
$$\log(\texttt{counts}) = 3.09 - 1.14 \cdot I_{\text{low food}} + 0.99 \cdot I_{\text{magna}} + \varepsilon$$

Vieles spricht also für mod3:

$$\log(\texttt{counts}) = 3.09 - 1.14 \cdot I_{\text{low food}} + 0.99 \cdot I_{\text{magna}} + \varepsilon$$

Anwenden der *e*-Funktion ergibt:

$$\texttt{counts} = 21.98 \cdot 0.32^{I_{\text{low food}}} \cdot 2.69^{I_{\text{magna}}} \cdot e^{\varepsilon}$$


prediction of log−linear model
galatea,low
magna, low
galatea,high
magna, high
0.0
0.5
1.0
1.5
2.0
3
4
5
6
7
8
9
15
20
30
40
50
100
no. of Daphnia

Inhalt

Zur Klausur

Montag, 23. Juli 2013 ab 12 Uhr im Buchner-Hörsaal und im Bayer-Hörsaal.

mitbringen:

1. Studierendenausweis (wegen der Matrikelnummer)
2. Personalausweis (wegen des Lichtbilds)
3. Formelblatt
   - DIN A 4
   - nur eigene Handschrift
   - nichts gedrucktes, nichts kopiertes
   - als Formelblatt gekennzeichnet mit abgeben
4. nicht programmierbaren Taschenrechner ohne Graphik-Funktion und ohne spezielle Statistik-Funktionen.
5. Kugelschreiber und Papier. Schmierpapier, das Sie für Nebenrechnungen verwenden, muss als solches gekennzeichnet und mit der Klausur abgegeben werden.

Klausur: bitte beachten

- Mobiltelefone, Notebooks etc. dürfen nicht mitgeführt werden.
- Jeder Täuschungsversuch führt dazu, dass die Klausur mit “nicht bestanden” bewertet wird.
- Nachklausur: 13. September