

Wahrscheinlichkeitsrechnung und Statistik für Biologen Lineare Modelle

Dirk Metzler

23. Juni 2021

Inhaltsverzeichnis

1	Regression zur Mitte	1
2	Multiple Regression	2
2.1	Beispiel: Artenreichtum an Sandstränden	4
2.2	Beispiel: Wirksamkeit von Therapien	12
3	Modellwahl: AIC und Kreuzvalidierung	13
3.1	Beispiel: (Schnabel-)Größen der Darwin-Finken	13
3.2	Beispiel: Wasserflöhe	17

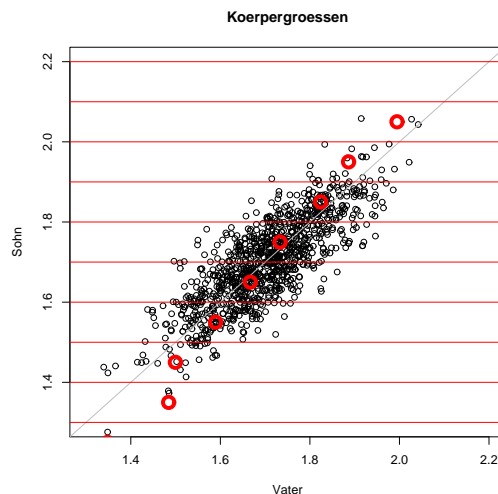
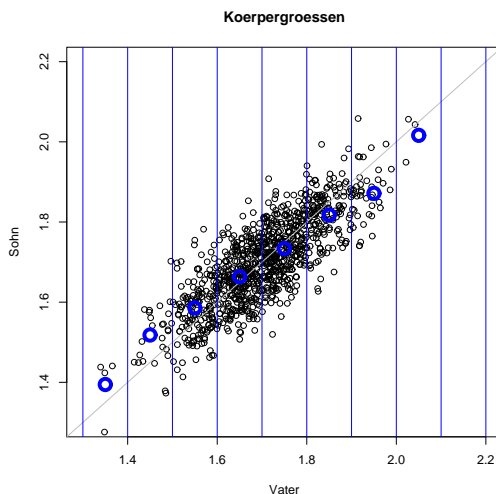
1 Regression zur Mitte

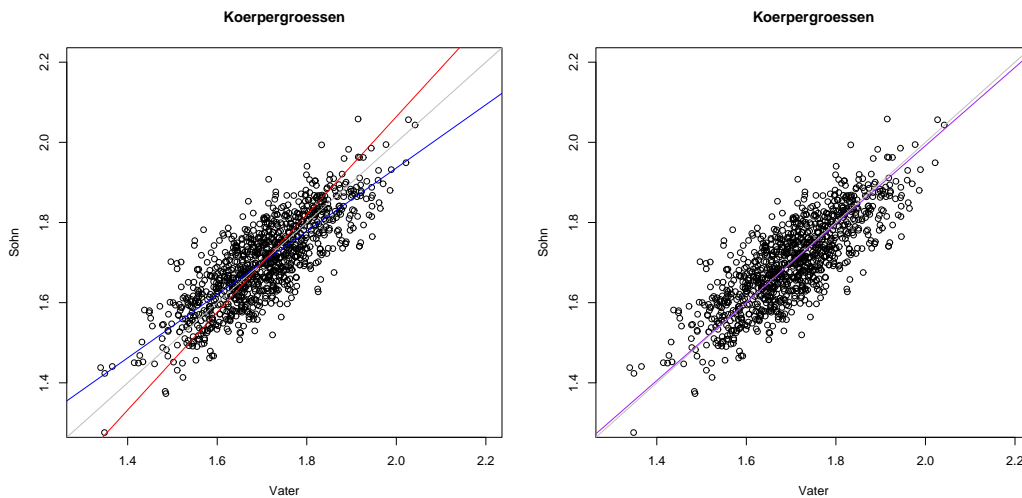
Herkunft des Worts “Regression”

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.





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.

2 Multiple Regression

Was, wenn wir die Stoffwechselrate als Funktion von Herzfrequenz und Blutdruck vorhersagen wollen?

```
> data.frame(metabolicrate,heartbeat,bloodpressure)
  metabolicrate heartbeat bloodpressure
1    80.31883    53.25858     97.19762
2    85.19011    56.43013     98.84911
3    91.96019    53.92183    107.79354
4    80.89438    50.46988    100.35254

> ( mod <- lm(metabolicrate ~ heartbeat + bloodpressure) )

Call:
lm(formula = metabolicrate ~ heartbeat + bloodpressure)

Coefficients:
(Intercept)    heartbeat  bloodpressure
   -72.577         0.982         1.035
```

Das Modell sagt also anhand der (rein fiktiven Daten) aus, dass die Stoffwechselrate ungefähr bei folgendem Wert liegt:

$$-72.577 + 0.982 \cdot \text{heartbeat} + 1.035 \cdot \text{bloodpressure}$$

Multiple Regression

und somit $f(\beta) = \langle y - \mathbf{X}\beta, y - \mathbf{X}\beta \rangle = \|y - \mathbf{X}\beta\|^2$.

Gesucht ist also $\hat{\beta}$, so dass $\hat{y} = \mathbf{X}\hat{\beta}$ minimalen euklidischen Abstand zu y hat.
Die Matrix \mathbf{X} heißt auch Modellmatrix; hier ein Beispiel:

```
> data.frame(metabolicrate,heartbeat,bloodpressure)
  metabolicrate heartbeat bloodpressure
1    80.31883    53.25858     97.19762
2    85.19011    56.43013     98.84911
3    91.96019    53.92183    107.79354
4    80.89438    50.46988     100.35254
> mod <- lm(metabolicrate ~ heartbeat + bloodpressure)
> model.matrix(mod)
  (Intercept) heartbeat bloodpressure
1           1    53.25858     97.19762
2           1    56.43013     98.84911
3           1    53.92183    107.79354
4           1    50.46988     100.35254
```

Analytische Lösung: Wir minimieren nun f , indem wir die Nullstelle der Ableitung, also des Gradienten

$$\frac{\partial f(\beta)}{\partial \beta} := \left(\frac{\partial f(\beta)}{\partial a}, \frac{\partial f(\beta)}{\partial b_1}, \dots, \frac{\partial f(\beta)}{\partial b_p} \right) = -2(y - \mathbf{X}\beta)^T \mathbf{X}$$

suchen (T steht für ‘transponiert’).

Wenn $\mathbf{X}^T \mathbf{X}$ invertierbar ist (was in der Regel der Fall ist, wenn die Stichprobengröße okay ist), dann hat $(y - \mathbf{X}\hat{\beta})^T \mathbf{X} = (0, \dots, 0)$ die eindeutige Lösung

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$

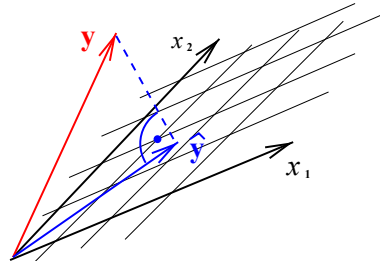
Geometrischer Lösungsweg: $f(\beta) = \|y - \mathbf{X}\beta\|^2$ minimieren bedeutet, dass $\hat{y} = \mathbf{X}\hat{\beta}$ die Projektion von y auf den von den Vektoren x_0, x_1, \dots, x_N aufgespannten Raum sein soll (mit $x_0 = (1, \dots, 1)^T$). Also muss $y - \mathbf{X}\hat{\beta}$ auf jedem x_i senkrecht stehen,

d.h.

$$\forall i : \langle y - \mathbf{X}\hat{\beta}, x_i \rangle = 0,$$

und damit

$$(y - \mathbf{X}\hat{\beta})^T \mathbf{X} = (0, \dots, 0),$$



woraus sich wieder die Lösung $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$ ergibt.

2.1 Beispiel: Artenreichtum an Sandstränden

- 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

Literatur

[ZIS07] 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:

$$\text{richness} = a + b_1 \cdot \text{angle2} + b_2 \cdot \text{NAP} + b_3 \cdot \text{grainsize} + b_4 \cdot \text{humus} + \varepsilon$$

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 = rikz)
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
F-statistic: 10.74 on 4 and 40 DF,  p-value: 5.237e-06
```

- 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:

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

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

z.B. b_7 beschreibt, um wieviel an einer durchschnittlichen Probestelle der Artenreichtum in Woche 4 gegenüber Woche 1 erhöht ist.

in R-Notation:

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

Hier für ein einfacheres Beispiel mit reduziertem Datensatz die von R intern verwendete Modellmatrix:

```
> mod <- lm( richness ~ NAP + factor(week), data = rikz, subset=seq(5,45,by=2) )
> model.matrix(mod)
  (Intercept)    NAP factor(week)2 factor(week)3 factor(week)4
5            1 -0.684            0            0            0
7            1  0.820            0            0            0
9            1  0.061            0            0            0
11           1 -0.976            1            0            0
13           1 -0.201            1            0            0
15           1  0.167            1            0            0
17           1 -0.030            1            0            0
19           1  1.367            1            0            0
21           1  1.117            0            0            1
23           1  0.729            0            0            1
25           1  0.054            0            0            1
27           1 -0.348            0            1            0
29           1 -0.893            0            1            0
31           1  0.883            1            0            0
33           1  1.375            1            0            0
35           1  0.367            1            0            0
37           1 -0.375            0            1            0
39           1  0.170            0            1            0
41           1 -0.356            0            1            0
43           1 -0.002            0            1            0
45           1  0.865            0            1            0
```

```

> 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
- Der geschätzte 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?

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.

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.

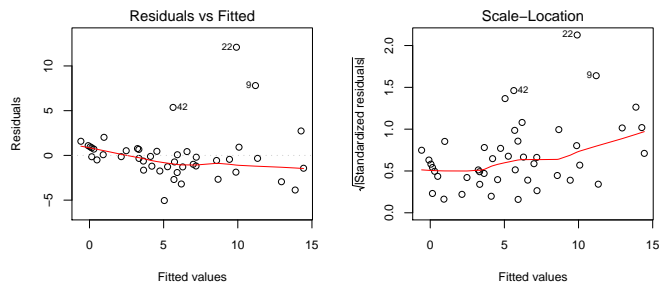
Geht nur, wenn die Modelle eingebettet (engl. nested) sind, d.h. das einfachere Modell lässt sich erzeugen, indem man bei dem komplexeren bestimmte Randbedingungen für die Parameterwerte definiert, in unserem Fall "alle Wocheneffekte sind gleich".

```
> 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

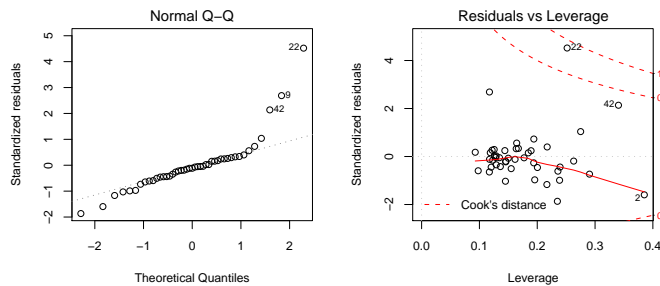
Model 1: richness ~ angle2 + NAP + grainsize + humus
Model 2: richness ~ angle2 + NAP + grainsize + humus + factor(week)
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     40 531.17
2     37 353.66  3    177.51 6.1902 0.00162 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


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

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



`plot(modell)`



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?

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

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

```

Warnhinweis

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 verschiedenen 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.

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
```

2.2 Beispiel: Wirksamkeit von Therapien

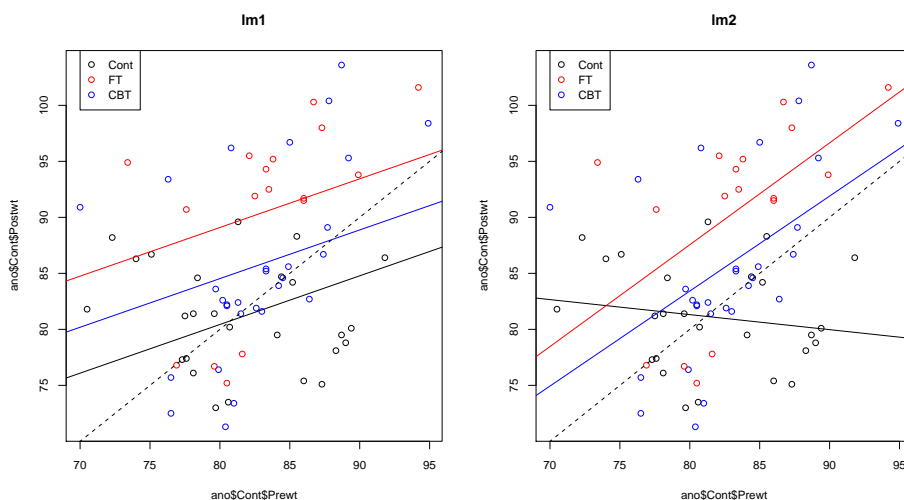
Vergleiche bei jugendlichen Magersuchtpatientinnen den Behandlungserfolg von Familientherapie (FT) und kognitiver Verhaltenstherapie (CBT) mit einer Kontrollgruppe (Cont), indem das Gewicht (in lbs.) vor (Prewt) und nach (Postwt) der Behandlung (Treat) verglichen wird.

Literatur

[HD+93] 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 <- 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.

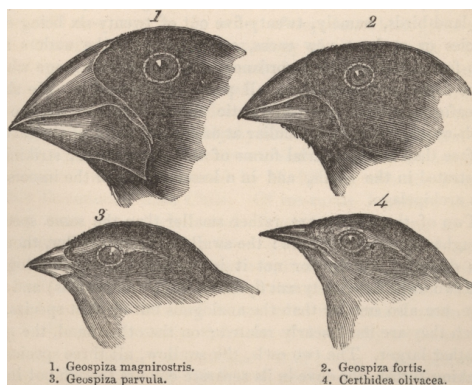
Interpretation: Welche Rolle das Gewicht vor der Behandlung spielt, hängt von der Behandlung ab.
 oder auch: Der Unterschied zwischen den Wirkungen der verschiedenen Behandlungen hängt vom Gewicht vor der Therapie ab.

Was Sie u.a. erklären können sollten

- Herkunft des Worts “Regression” und ähnliche Effekte
- grundlegendes Modell der multiplen Regression, auch in Matrixschreibweise
- geometrische Interpretation des kleinste Quadrate-Schätzers
- Wie passen kategorielle Variablen (factor) in das Modell?
- Worauf genau beziehen sich die p-Werte jeweils in den verschiedenen R-Ausgaben?
- ANOVA für eingebettete lineare Modelle
- Was bedeuten Interaktionsterme mathematisch und praktisch?

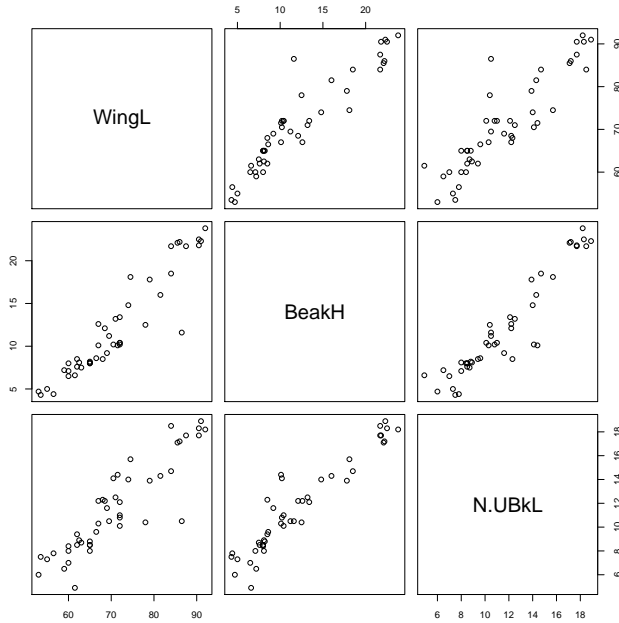
3 Modellwahl: AIC und Kreuzvalidierung

3.1 Beispiel: (Schnabel-)Größen der Darwin-Finken



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.



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?

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.

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

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{1}{n-d} \sum_i r_i^2} = \sigma_r \cdot \sqrt{\frac{n-1}{n-d}}$$

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

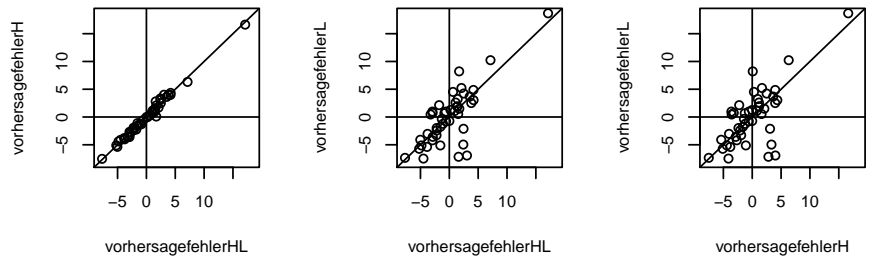
Eine weitere Möglichkeit, die Vorhersagegenauigkeit eines Modells zu beurteilen, ist die *Kreuzvalidierung* (auch *Jackknife* genannt).

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
$\sigma(\text{Residuen})$	3.83	4.78	3.79
$d = (\text{Anzahl Parameter})$	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

Akaiikes Informationskriterium:

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

Bayessches Informationskriterium:

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

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. (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 Mo-

dells, wobei a priori alle Modelle als gleich wahrscheinlich angenommen werden.

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

belhöhe zu berücksichtigen.

3.2 Beispiel: Wasserflöhe

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 ***
foodlevellow        -15.00    3.837   -3.909 0.00449 **
speciesmagna         38.00    3.837    9.904 9.12e-06 ***
foodlevellow:speciesmagna -26.67  5.426   -4.914 0.00117 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

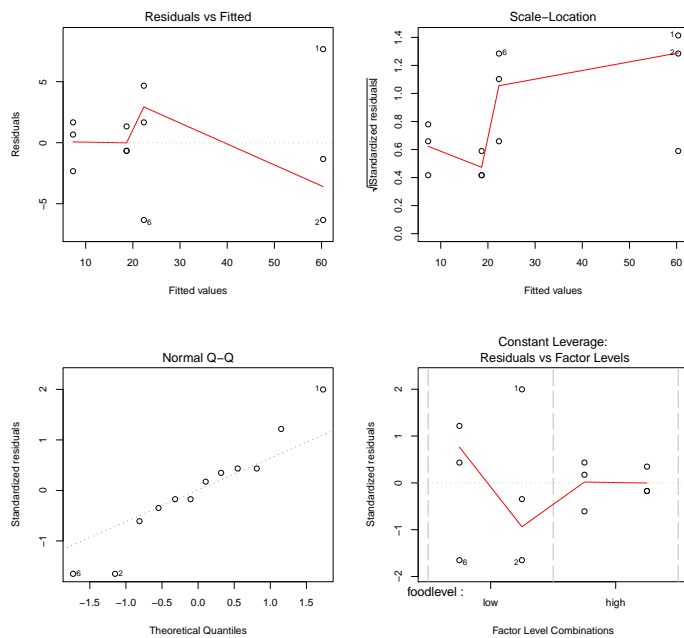
```

```

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. Aber passt es gut genug?



```

> 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

```

```

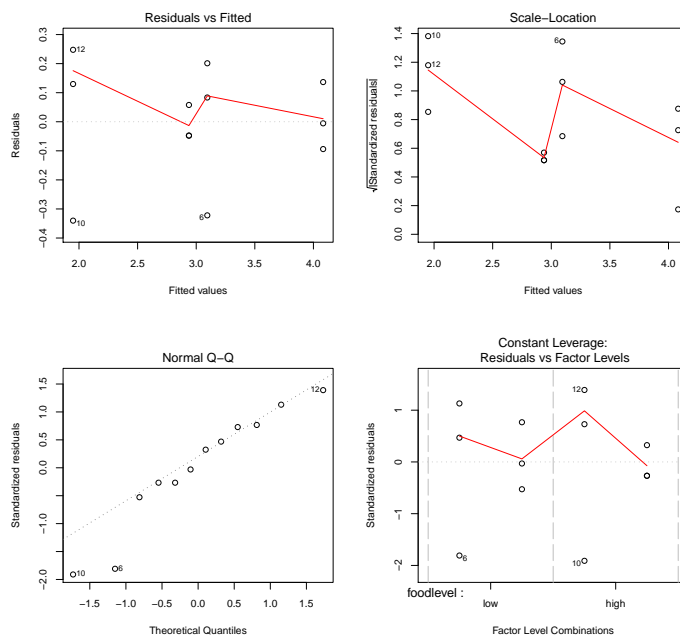
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.0946    0.1028  30.104 2.41e-10 ***
foodlevellow  -1.1450    0.1187  -9.646 4.83e-06 ***
speciesmagna   0.9883    0.1187   8.326 1.61e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

```

```

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

```



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

```

Das Modell mod2 hat einen besseren AIC wert als mod1. Kein Wunder, es war ja auch signifikant besser.

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

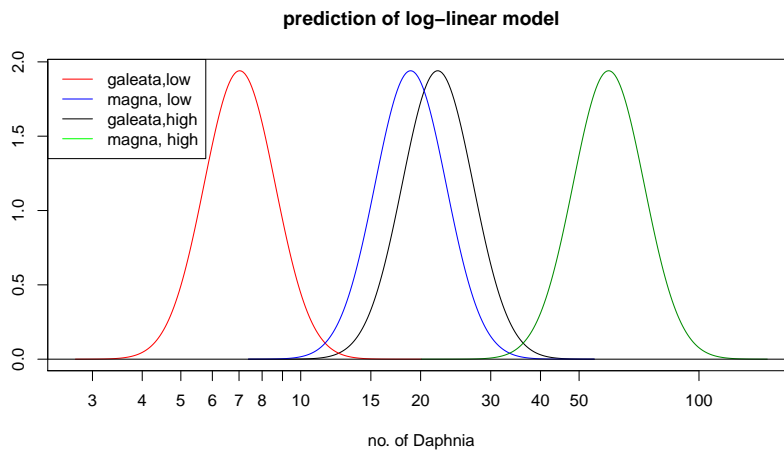
Die AIC-Werte der log-linearen Modelle mod3 und mod4 kann man nicht mit denen der linearen Modelle vergleichen, da die Zielvariable eine andere (weil transformiert) ist.

Vieles spricht also für mod3:

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

Anwenden der e -Funktion ergibt:

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



Was Sie u.a. erklären können sollten

- Modelle als R-Formel, aber auch mathematisch präzise spezifizieren können
- Wieso ist das Modell, das alle Parameter berücksichtigt, nicht immer das beste?
- Freiheitsgrade bei der Residuenvarianz
- Kreuzvalidierung
- AIC und BIC
- Beurteilung eines Modells durch graphische Analyse der Residuen
- Interpretation von Modellen bei log-Skalierung der Zielvariable

Bitte beachten Sie auch die Auflistung aus Seite 13!