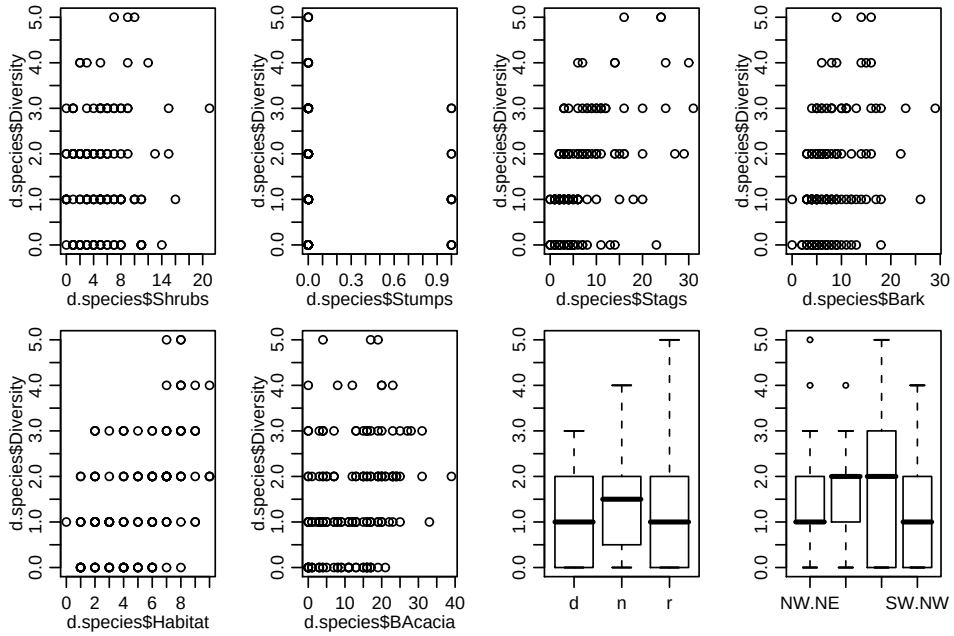
Angewandte Regression — Musterlösungen zur Serie 9

1. **a)** Es werden im Folgenden nur die augenfälligsten Aspekte angesprochen, welche man in den Plots erkennen kann.



Tendenziell werden mehr verschiedene Arten von Beutelratten (`Diversity`) gefunden bei einer höheren Anzahl von Sträuchern, wenig Baumstrünken, vielen hohlen Bäumen, einer mittleren Menge abgenagter Rinde, günstigem Habitat. Welche dieser Aussagen auch statistisch signifikant sind, wird erst die folgende Auswertung zeigen. Ferner scheint die Streuung der Artenvielfalt bei südlicher und westlicher Ausrichtung des Geländes grösser zu sein als sonst. Die Variable `Stumps` wurde vermutlich mit 'Anzahl Baumstrünke' falsch beschrieben, da nur die Werte 0 oder 1 vorkommen. Vermutlich ist diese Variable also binär und gibt nur das Vorhandensein von Baumstrünken (ja, nein) wieder.

b) Da es sich bei der Variable Y_i (`Diversity`) um eine Anzahl Ereignisse pro Gebietseinheit handelt, können wir annehmen, dass diese Poisson-verteilt sind

$$Y_i \sim \mathcal{P}\langle\lambda_i\rangle\,,$$

wobei $\lambda_i = E\langle Y_i\rangle$ der Erwartungswert ist. In diesem Fall ist das Modell der Poisson-Regression angebracht. Als Link-Funktion wählen wir den Logarithmus:

$$g\langle\lambda_i\rangle = \log\langle\lambda_i\rangle = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_m x_i^{(m)}\,.$$

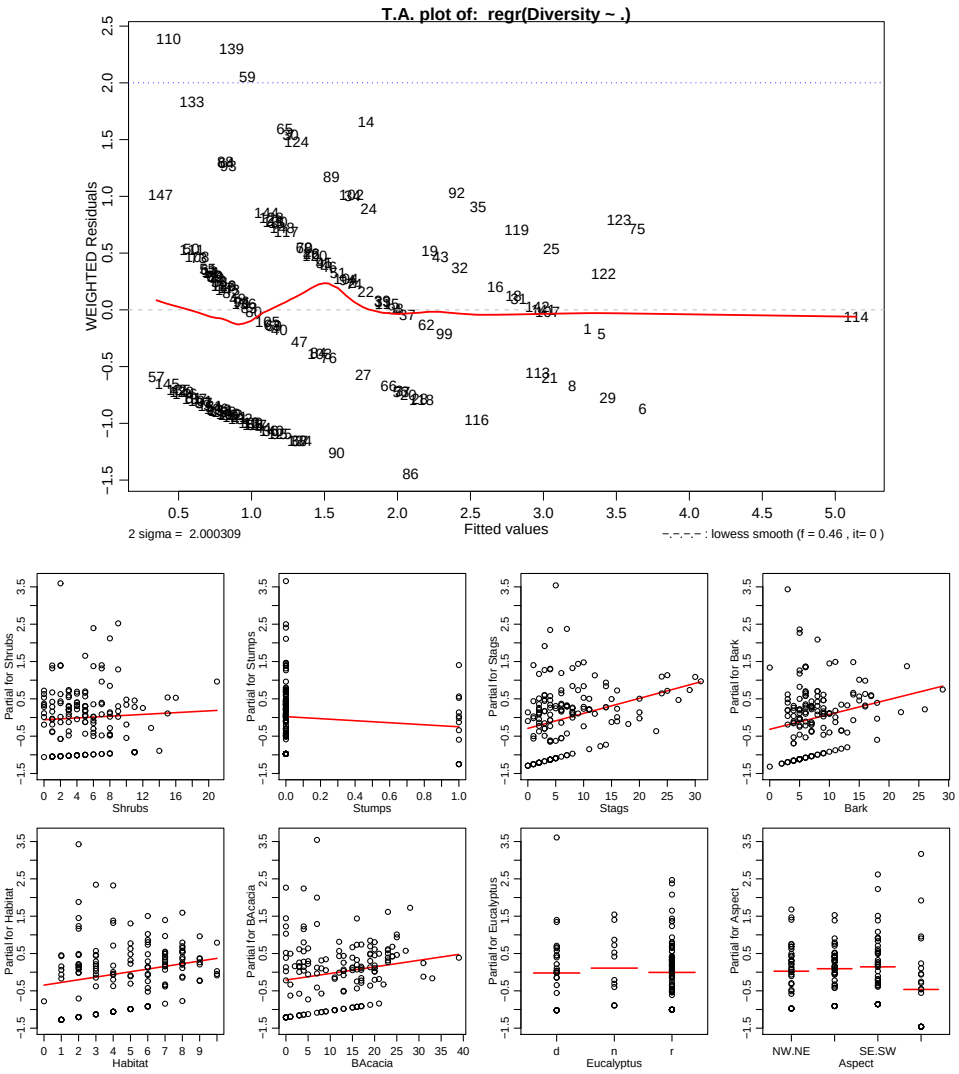
$x^{(1)}, \ldots, x^{(m)}$ sind die erklärenden Variablen und $m = 8$.
Eine andere Möglichkeit wäre der Wurzel-Link.

c) Wir passen das Modell mit der Funktion `glm()` an, wobei wir zuerst alle erklärenden Variablen ins Modell nehmen. In R können wir dies in der Formel bekanntlich mit einem Punkt erreichen, damit wir nicht die Namen aller 8 erklärenden Variablen hinschreiben müssen.

```
> r.p.regr <- regr(Diversity ~ . , family="poisson", data=d.species)
> summary(r.p.regr)
Call:
regr(formula = Diversity ~ ., data = d.species, family = "poisson")
Terms:
                   coef      stcoef     signif       R2.x df p.value
(Intercept) -0.96228763  0.00000000 -1.7367224         NA  1  0.0006
Shrubs       0.01192096  0.04515786  0.2747252 0.20020210  1  0.5886
Stumps      -0.27240588 -0.07666438 -0.4818638 0.06496421  1  0.3249
Stags        0.04022862  0.26634905  1.8159018 0.24169263  1  0.0005
Bark         0.03988606  0.19133894  1.4020069 0.14121939  1  0.0069
Habitat      0.07173483  0.18306310  0.9512575 0.31416291  1  0.0595
BAcacia      0.01763833  0.15198855  0.8417771 0.27087135  1  0.0965
Eucalyptus           NA          NA -1.3252760 0.08382862  2  0.9075
Aspect               NA          NA  1.7709410 0.03533445  3  0.0383

Coefficients for factors:
$Eucalyptus
         d          n          r
0.00000000 0.13026532 0.01534376

$Aspect
      NW.NE       NW.SE       SE.SW       SW.NW
 0.00000000  0.06675529  0.11694626 -0.48890705

          deviance  df      p.value
Model     68.61585  11 2.237150e-10
Residual 118.87372 139           NA
Null     187.48957 150           NA

Family is poisson.  Dispersion parameter taken to be 1.
AIC: 423.67
```

Kommentar zu den Resultaten:

- **Residuenanalyse:** Der Tukey-Anscombe-Plot zeigt keine Struktur: Die Glättkurve bewegt sich um Null herum. Die Beobachtung 114 mit dem grössten geschätzten Erwartungwert $\widehat{\mu}_i$ sollte ev. genauer untersucht werden. Der Plot der partiellen Residuen (erstellt mit `termplot()`) zeigt ohne Glättung nicht viel.



- **Schrittweise Variablenreduktion mit `step()`** Der Output von `r.step <- step(r.p.regr)` ist ziemlich gross und wird hier aus Platzgründen weggelassen. Kürzer und ebenfalls informativ ist die Tabelle, die unter der Komponente `r.step$anova` gespeichert ist. Sie zeigt eine Auflistung der aus dem Modell entfernten Variablen:

```
          Step Df  Deviance Resid. Df Resid. Dev      AIC
1             NA        NA       139   118.8737 423.6733
2 - Eucalyptus  2 0.1942112       141   119.0679 419.8675
3     - Shrubs  1 0.2903417       142   119.3583 418.1579
4     - Stumps  1 0.8997410       143   120.2580 417.0576
```

  Als erste Variable wird also `Eucalyptus` weggelassen, dann `Shrubs` und als letzte `Stumps`. Nachher kann der AIC-Wert durch Weglassen von weiteren Variablen nicht mehr verkleinert werden. Mit `summary(r.step)` erhält man die übliche Auflistung verschiedener Werte und Resultate des optimierten Modells:

```
> summary(r.step)  ## im Modell belassene Variablen
Call:
regr(formula = Diversity ~ Stags + Bark + Habitat + BAcacia +
    Aspect, data = d.species, family = "poisson")
Terms:
                    coef    stcoef    signif       R2.x df p.value
(Intercept) -0.88768085 0.0000000 -1.919000         NA  1  0.0001
Stags        0.04075724 0.2698490  1.969009 0.18628740  1  0.0002
```

```
Bark         0.04082373 0.1958371  1.650204 0.03159869  1  0.0018
Habitat      0.07636217 0.1948718  1.048230 0.29098885  1  0.0379
BAcacia      0.01413663 0.1218146  0.732490 0.21306962  1  0.1478
Aspect               NA        NA  2.014108 0.02338537  3  0.0220

Coefficients for factors:
$Aspect
      NW.NE       NW.SE       SE.SW       SW.NW
 0.00000000  0.07809185  0.10925403 -0.52426833

          deviance  df    p.value
Model     67.23156   7 5.3475e-12
Residual 120.25802 143         NA
Null     187.48957 150         NA

Family is poisson.  Dispersion parameter taken to be 1.
AIC: 417.06
```

Bemerkung: Die AIC-Werte berechnen sich gemäss der Formel

$$\mathrm{AIC} = \mathrm{const} + D + 2p\,,$$

wobei $p = n - \texttt{Resid.Df}$ die Anzahl Parameter im Modell ist (hier ist $n = 151$). Also z.B.

$$\begin{aligned} 423.6733 &= const + 118.8737 + 2 \cdot 12 \,\text{volles Modell} \\ 419.8675 &= const + 119.0679 + 2 \cdot 10 \,\text{ohne Ecalyptus} \\ 418.1579 &= const + 119.3583 + 2 \cdot 9 \;\text{ohne Shrubs} \\ 417.0576 &= const + 120.2580 + 2 \cdot 8 \;\text{ohne Stumps} \end{aligned}$$

(Es ergibt sich const = 280.7996.)

2. Die Musterlösung von dieser Aufgabe erscheint später als separate Blätter.

3. **a)** Verallgemeinertes lineares Model:

$$\texttt{number}_i \sim \mathcal{P}\langle\lambda_i\rangle\,, \qquad \log\langle\lambda_i\rangle = \beta_0 + \beta_1 \cdot \texttt{time}_i$$

Begründung: Eine Zielvariable, die die Anzahl Todesfälle (= Ereignisse) pro Zeiteinheit (Periode von 3 Monaten) misst, ist typischerweise poissonverteilt. Der kanonische Link der Poissonregression ist der Logarithmus.

b) Schätzung mit R:

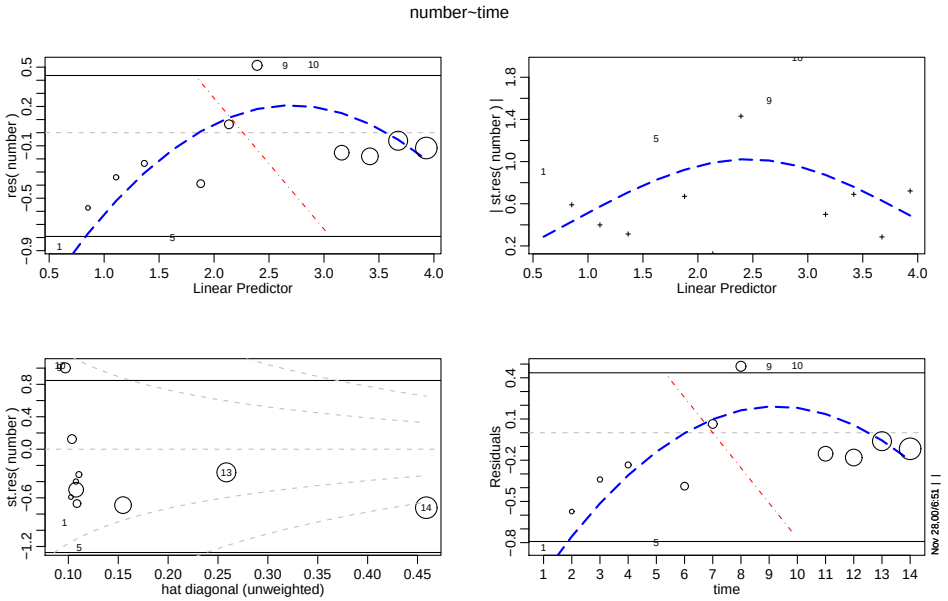
```
> r.glm.regr <- regr(number ~ time, family="poisson", data=d.aids)
> r.glm.regr
Call:
regr(formula = number ~ time, data = d.aids, family = "poisson")

Terms:
                 coef   stcoef    signif R2.x df p.value
(Intercept) 0.3396339 0.000000 0.6205745   NA  1  0.1763
time        0.2565236 1.073115 5.3421156    0  1  0.0000
```

```
          deviance df     p.value
Model     177.61879  1 0.000000000
Residual   29.65352 12 0.003147878
Null      207.27231 13          NA

Family is quasi.poisson.  Dispersion parameter estimated to be  2.471127.
AIC: 86.58
```

Mit `regr` wird der Dispersionsparameter automatisch mitgeschätzt, falls man ihn nicht ausdrücklich mit `calcdisp=F` innerhalb des `regr`-Aufrufs auf 1 setzt. In unserem Beispiel wird er auf beachtliche 2.47 geschätzt. Dadurch vergrössert sich das Vertrauensintervall des Koeffizienten von `time`. Der Parameter bleibt aber trotzdem hochsignifikant. Ebenfalls signifikant ist die Residuendevianz, d.h. das gesättigte Modell ist deutlich besser. Ob das an übergrosser Streuung liegt oder daran, dass das Modell schlecht passt?

c) Plot der (Devianz-) Residuen gegen die Zeit:

```
> plot(r.glm.regr)
```



Der Termplot (rechts unten) zeigt deutlich ein nicht-lineares Verhalten der erklärenden Variablen `time`. Durch eine Transformation mit z.B. dem Logarithmus können wir versuchen, dies zu verbessern.

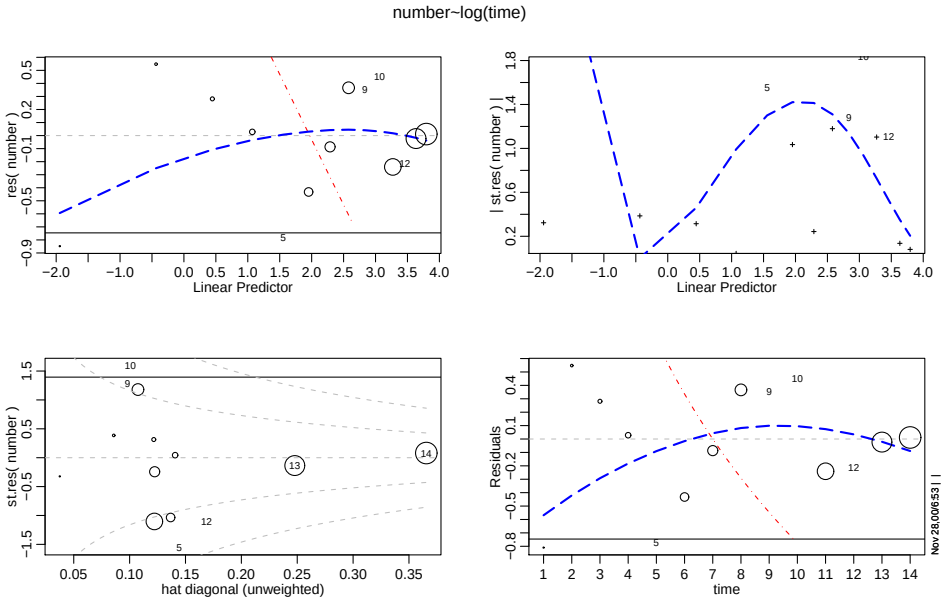
d) "Verbessertes" Modell: (`time`$_i$ logarithmieren)

$$\texttt{number}_i \sim \mathcal{P}\langle\lambda_i\rangle\,, \qquad \log\langle\lambda_i\rangle = \beta_0 + \beta_1 \cdot \log\langle\texttt{time}_i\rangle$$

```
> r.glm.regr.log <- regr(number ~ log(time), family="poisson", data=d.aids)
> summary(r.glm.regr.log)
Call:
regr(formula = number ~ log(time), data = d.aids, family = "poisson")

Terms:
                coef   stcoef    signif R2.x df p.value
```

```
(Intercept) -1.944166 0.000000 -1.744190   NA  1   1e-04
log(time)    2.174784 1.683510  4.641539    0  1   0e+00

          deviance df   p.value
Model     190.18065  1 0.0000000
Residual   17.09166 12 0.1461819
Null      207.27231 13        NA

Family is quasi.poisson.  Dispersion parameter estimated to be  1.424305.
AIC: 74.019
```

Wir sehen, dass sich die übergrosse Streuung zwar stark verkleinert hat, aber immer noch vorhanden ist. Die Variable `log(time)` ist weiterhin signifikant, die Residuendevianz nicht mehr. Unser Modell passt viel besser als das untransformierte, was auch die Residuenplots zeigen. Die Struktur im Termplot ist verschwunden.



Bemerkung: Im Termplot des untransformierten Modelles fällt auf, dass die Residuen bei hohen `time`-Werten nahezu parallel zur Referenzgeraden verlaufen. Mit einem Modell, welches für Zeiten unter acht und über acht verschiedene Steigungen zulässt, erhalten wir ein noch besseres Resultat. Der Dispersionsparameter fällt gar auf 1.11. Die Grenze acht bei der Indikatorvariablen `time.ge8` wurde (von Auge) aus den Daten geschätzt. Deshalb wären die Freiheitsgrade bei den Tests entsprechend zu korrigieren.

```
> d.aids$time.ge8 <- d.aids$time>=8
> r.glm.regr.ind <- regr(number ~ log(time)*time.ge8, family="poisson",
                         data=d.aids)
> r.glm.regr.ind
Call:
regr(formula = number ~ log(time) * time.ge8, data = d.aids,
     family = "poisson")

Terms:
                          coef      stcoef      signif       R2.x df p.value
(Intercept)        -1.7797430  0.0000000 -0.7139333         NA  1  0.1117
```

```
log(time)           1.9006514  1.4713031  1.2971016 0.6519740  1  0.0058
time.ge8            1.7482625  0.9071289  0.5292709 0.8417693  1  0.2718
log(time):time.ge8 -0.4954754 -0.6158232 -0.2889759 0.8779478  1  0.5442

          deviance df   p.value
Model    196.15212  3 0.0000000
Residual  11.12019 10 0.3482309
Null     207.27231 13        NA

Family is quasi.poisson.  Dispersion parameter estimated to be  1.112019.
AIC: 72.047
```