Übung zur Vorlesung Kategoriale Daten **Blatt 5**
Gerhard Tutz, Moritz Berger **WiSe 15/16**

2 Regression mit binärem Response

Aufgabe 19

Analyse der Samen der Sommerwurz (Datensatz orob2)

```
# Vorbereitungen
library(aod)
data(orob2)
?orob2
orob2

##    seed     root  n  y
## 1   O75     BEAN 39 10
## 2   O75     BEAN 62 23
## 3   O75     BEAN 81 23
## 4   O75     BEAN 51 26
## 5   O75     BEAN 39 17
## 6   O75 CUCUMBER  6  5
## 7   O75 CUCUMBER 74 53
## 8   O75 CUCUMBER 72 55
## 9   O75 CUCUMBER 51 32
## 10  O75 CUCUMBER 79 46
## 11  O75 CUCUMBER 13 10
## 12  O73     BEAN 16  8
## 13  O73     BEAN 30 10
## 14  O73     BEAN 28  8
## 15  O73     BEAN 45 23
## 16  O73     BEAN  4  0
## 17  O73 CUCUMBER 12  3
## 18  O73 CUCUMBER 41 22
## 19  O73 CUCUMBER 30 15
## 20  O73 CUCUMBER 51 32
## 21  O73 CUCUMBER  7  3

attach(orob2)
response <- cbind(y,n-y)
```

(a) Binomialmodelle

```
# Binomial-Modell nur mit unterschiedlicher Gattung
Bin1 <- glm(response ~ seed, family=binomial)
summary(Bin1)

##
## Call:
## glm(formula = response ~ seed, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4707  -1.5670   0.2426   1.4173   4.1140
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1214     0.1233  -0.984    0.325
## seed075       0.2379     0.1493   1.594    0.111
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 96.175  on 19  degrees of freedom
## AIC: 176.77
##
## Number of Fisher Scoring iterations: 3
```

```
# Binomial-Modell mit unterschiedlicher Gattung und Wurzel-Extrakten
Bin2 <- glm(response ~ seed + root, family=binomial)
summary(Bin2)
```

```
##
## Call:
## glm(formula = response ~ seed + root, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3919  -0.9949  -0.3744   0.9831   2.4766
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.7005     0.1507  -4.648 3.36e-06 ***
## seed075        0.2705     0.1547   1.748   0.0804 .
## rootCUCUMBER   1.0647     0.1442   7.383 1.55e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 39.686  on 18  degrees of freedom
## AIC: 122.28
##
## Number of Fisher Scoring iterations: 4
```

```
# Binomial-Modell mit unterschiedlicher Gattung  und Wurzel-Extrakten
# plus Wechselwirkung
Bin3 <- glm(response ~ seed*root, family=binomial)
summary(Bin3)
```

```
##
## Call:
## glm(formula = response ~ seed * root, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01617  -1.24398   0.05995   0.84695   2.12123
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.4122     0.1842  -2.238   0.0252 *
## seed075               -0.1459     0.2232  -0.654   0.5132
## rootCUCUMBER           0.5401     0.2498   2.162   0.0306 *
## seed075:rootCUCUMBER   0.7781     0.3064   2.539   0.0111 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 33.278  on 17  degrees of freedom
## AIC: 117.87
##
## Number of Fisher Scoring iterations: 4
```

Vergleich der Modelle mittels AIC:

```
AIC(Bin1, Bin2, Bin3)

##      df      AIC
## Bin1  2 176.7715
## Bin2  3 122.2821
## Bin3  4 117.8740
```

Vergleich der Modelle mittels Devianz-Analyse:

```
anova(Bin1,Bin2,Bin3)

## Analysis of Deviance Table
##
## Model 1: response ~ seed
## Model 2: response ~ seed + root
## Model 3: response ~ seed * root
##   Resid. Df Resid. Dev Df Deviance
## 1        19     96.175
## 2        18     39.686  1   56.489
## 3        17     33.278  1    6.408

1-pchisq(Bin1$dev - Bin2$dev, 1)

## [1] 5.651035e-14

1-pchisq(Bin2$dev - Bin3$dev, 1)

## [1] 0.01136007

1-pchisq(Bin1$dev - Bin3$dev, 2)

## [1] 2.198242e-14
```

⇒ Entscheidung für Modell Bin3!

(b) Beta-Binomial-Modelle

- Definition Beta-Binomial-Modell:

$$\begin{aligned} y_i|D_i &\sim B(n_i, \pi_i) \\ D_i &\sim Be(a_i, b_i) \\ E(Di) &= \pi_i \\ Var(Di) &= \delta\pi_i(1-\pi_i) \end{aligned}$$

- In dem Argument formula von betabin() muss die Zielgröße die Struktur cbind$(y, n-y)$ haben (d.h. selbst wenn response=cbind$(y, n-y)$ gilt, darf man nicht response schreiben)
- ϕ im Output ist das δ aus der Übung/Vorlesung

```
# Beta-Binomial-Modell nur mit unterschiedlicher Gattung
BetaBin1 <- betabin(cbind(y,n-y) ~ seed, random=~1, data=orob2)
summary(BetaBin1)

## Beta-binomial model
## -------------------
## betabin(formula = cbind(y, n - y) ~ seed, random = ~1, data = orob2)
##
## Convergence was obtained after 80 iterations.
##
## Fixed-effect coefficients:
##               Estimate Std. Error   z value Pr(> |z|)
## (Intercept) -2.602e-01  2.262e-01 -1.15e+00 2.500e-01
## seedO75      4.130e-01  2.993e-01  1.38e+00 1.676e-01
##
## Overdispersion coefficients:
##                  Estimate Std. Error   z value  Pr(> z)
## phi.(Intercept) 7.805e-02  3.015e-02 2.589e+00 4.82e-03
##
## Log-likelihood statistics
##    Log-lik      nbpar    df res.   Deviance        AIC       AICc
## -6.355e+01          3         18  5.051e+01  1.331e+02  1.345e+02

# Beta-Binomial-Modell mit unterschiedlicher Gattung und Wurzel-Extrakten
BetaBin2 <- betabin(cbind(y,n-y) ~ seed + root, random=~1, data=orob2)
summary(BetaBin2)

## Beta-binomial model
## -------------------
## betabin(formula = cbind(y, n - y) ~ seed + root, random = ~1,
##     data = orob2)
##
## Convergence was obtained after 167 iterations.
##
## Fixed-effect coefficients:
##                Estimate Std. Error    z value Pr(> |z|)
## (Intercept)  -7.280e-01  1.995e-01 -3.649e+00 2.634e-04
## seedO75       3.427e-01  2.106e-01  1.627e+00 1.038e-01
## rootCUCUMBER  1.011e+00  2.020e-01  5.004e+00 5.625e-07
##
## Overdispersion coefficients:
##                  Estimate Std. Error   z value   Pr(> z)
## phi.(Intercept) 1.937e-02  1.416e-02 1.368e+00 8.569e-02
##
## Log-likelihood statistics
##    Log-lik      nbpar    df res.   Deviance        AIC       AICc
## -5.583e+01          4         17  3.507e+01  1.197e+02  1.222e+02

# Beta-Binomial-Modell mit unterschiedlicher Gattung  und Wurzel-Extrakten
# plus Wechselwirkung
BetaBin3 <- betabin(cbind(y,n-y) ~ seed * root, random=~1, data=orob2)
summary(BetaBin3)

## Beta-binomial model
## -------------------
## betabin(formula = cbind(y, n - y) ~ seed * root, random = ~1,
```

```
##     data = orob2)
##
## Convergence was obtained after 196 iterations.
##
## Fixed-effect coefficients:
##                        Estimate Std. Error    z value Pr(> |z|)
## (Intercept)          -4.456e-01  2.183e-01 -2.041e+00 4.124e-02
## seedO75              -9.612e-02  2.737e-01 -3.512e-01 7.255e-01
## rootCUCUMBER          5.235e-01  2.968e-01  1.764e+00 7.780e-02
## seedO75:rootCUCUMBER  7.962e-01  3.779e-01  2.107e+00 3.514e-02
##
## Overdispersion coefficients:
##                  Estimate Std. Error   z value   Pr(> z)
## phi.(Intercept) 1.236e-02  1.131e-02 1.093e+00 1.373e-01
##
## Log-likelihood statistics
##    Log-lik      nbpar    df res.   Deviance        AIC       AICc
## -5.377e+01          5         16  3.094e+01  1.175e+02  1.215e+02
```

Vergleich der Modelle mittels AIC:

```
AIC(BetaBin1, BetaBin2, BetaBin3)


##          df      AIC     AICc
## BetaBin1  3 133.1054 134.5172
## BetaBin2  4 119.6637 122.1637
## BetaBin3  5 117.5336 121.5336
```

Vergleich der Modelle mittels Devianz-Analyse:

```
anova(BetaBin1,BetaBin2,BetaBin3)


## Analysis of Deviance Table (beta-binomial models)
##
## BetaBin1: fixed = cbind(y, n - y) ~ seed; random = ~1
## BetaBin2: fixed = cbind(y, n - y) ~ seed + root; random = ~1
## BetaBin3: fixed = cbind(y, n - y) ~ seed * root; random = ~1
##
##            logL k   AIC  AICc   BIC Resid. dev. Resid. Df              Test
## BetaBin1 -63.55 3 133.1 134.5 136.2       50.51        18
## BetaBin2 -55.83 4 119.7 122.2 123.8       35.07        17 BetaBin1-BetaBin2
## BetaBin3 -53.77 5 117.5 121.5 122.8       30.94        16 BetaBin2-BetaBin3
##          Deviance Df P(> Chi2)
## BetaBin1
## BetaBin2    15.44  1 8.509e-05
## BetaBin3     4.13  1 4.213e-02


1-pchisq(BetaBin1@dev - BetaBin2@dev, 1)


## [1] 8.508813e-05


1-pchisq(BetaBin2@dev - BetaBin3@dev, 1)


## [1] 0.04212658


1-pchisq(BetaBin1@dev - BetaBin3@dev, 2)


## [1] 5.623763e-05
```

⇒ Entscheidung für Modell BetaBin3!

(c) Quasi-Binomial-Modelle

```
# Quasi-Binomial-Modell nur mit unterschiedlicher Gattung
QuasiBin1 <- glm(response ~ seed, family=quasibinomial)
summary(QuasiBin1)

##
## Call:
## glm(formula = response ~ seed, family = quasibinomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4707  -1.5670   0.2426   1.4173   4.1140
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1214     0.2707  -0.448    0.659
## seedO75       0.2379     0.3277   0.726    0.477
##
## (Dispersion parameter for quasibinomial family taken to be 4.818491)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 96.175  on 19  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 3

# Quasi-Binomial-Modell mit unterschiedlicher Gattung und Wurzel-Extrakten
QuasiBin2 <- glm(response ~ seed + root, family=quasibinomial)
summary(QuasiBin2)

##
## Call:
## glm(formula = response ~ seed + root, family = quasibinomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3919  -0.9949  -0.3744   0.9831   2.4766
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.7005     0.2199  -3.186  0.00512 **
## seedO75        0.2705     0.2257   1.198  0.24635
## rootCUCUMBER   1.0647     0.2104   5.061 8.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 2.128368)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 39.686  on 18  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4

# Quasi-Binomial-Modell mit unterschiedlicher Gattung  und Wurzel-Extrakten
# plus Wechselwirkung
QuasiBin3 <- glm(response ~ seed*root, family=quasibinomial)
summary(QuasiBin3)
```

```
##
## Call:
## glm(formula = response ~ seed * root, family = quasibinomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01617  -1.24398   0.05995   0.84695   2.12123
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.4122     0.2513  -1.640   0.1193
## seedO75                -0.1459     0.3045  -0.479   0.6379
## rootCUCUMBER            0.5401     0.3409   1.584   0.1315
## seedO75:rootCUCUMBER    0.7781     0.4181   1.861   0.0801 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.861832)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 33.278  on 17  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Bei Quasi-Binomial-Modelle keine AIC- und Devianz-Bestimmung möglich, da keine Verteilung spezifiziert wird (Achtung! glm() gibt bei Quasi-Modell die Devianzen des gewöhnlichen Binomial-Modells aus!!!)

(d) Vergleich der Outputs

- Schätzer für Koeffizienten im Quasi-Binomial-Modell und im Binomial-Modell sind dieselben, da ϕ in der Schätzgleichung rausfliegt
- Standardfehler im Quasi-Binomial-Modell größer als im Binomial-Modell (siehe Zusammenhang mit ϕ)
- Standardfehler im Beta-Binomial-Modell größer als im Binomial-Modell (Beachte: Nur Overdispersion-Modellierung im Beta-Binomial-Modell möglich)
- Vergleich der Binomial- und der Beta-Binomial-Modelle mit AIC möglich ($\Rightarrow$ Modell BetaBin3!)
- Vergleich der Binomial- und der Beta-Binomial-Modelle mit Devianzen wegen hierarchischer Struktur möglich: Binomial-Modell = Beta-Binomial-Modell bei $\delta = 0$ (und damit $\phi = 1$)

  ```
  1-pchisq(Bin1$dev - BetaBin1@dev, 1)

  ## [1] 1.402289e-11

  1-pchisq(Bin2$dev - BetaBin2@dev, 1)

  ## [1] 0.03162927

  1-pchisq(Bin3$dev - BetaBin3@dev, 1)

  ## [1] 0.1260502
  ```

  $\Rightarrow$ Entscheidung für Modell Bin3!
- Ein Vergleich mit den Quasi-Binomialmodellen über AIC bzw. Devianz ist nicht möglich, da Vtlg nicht bekannt.
- Vergleich sämtlicher Modelle über Kreuzvalidierung: Kriterien z.B. Trefferquote, Kullback-Leibler-Distanz