Logistische Regression und die Analyse von Proportionen

Jonathan Harrington

Logistische Regression und die abhängige Variable

Mit der logistischen Regression kann geprüft werden, ob Proportionen von einem (oder von mehreren) Faktoren beeinflusst werden.

Die abhängige Variable ist **immer binär**, zB:

glottalisiert vs. nicht-glottalisiert
lenisiert vs. nicht-lenisiert
geschlossen vs. offen
ja vs. nein
True vs. False usw.

Allgemeine Überlegungen

In der logistischen Regression wird eine Regressionslinie an Proportionen durch ein Verfahren genannt 'maximum likelihood' (anstatt least squares) angepasst.

Zusätzlich werden in der logistischen Regression *log-odds* statt Proportionen modelliert

logodds(R) = mF + b

R ist der binäre Response (abhängige Variable), *F* ist der Faktor, *m* und *b* sind die Neigung und Intercept (y-Achse-Abschnitt)

z.B. Response (Abhängige Variable)
20 männlich, 15 weiblich
Logodds(R) = log(20/15)

Hohe vs. tiefe Vokale

Die Anzahl der Sprecher, die in der Standard-Aussprache von England *lost* mit einem hohen und tiefen Vokal ist wie folgt.

Jahr	Hoch	Tief
1950	30	5
1960	18	21
1971	15	26
1980	13	20
1993	4	32
2005	2	34

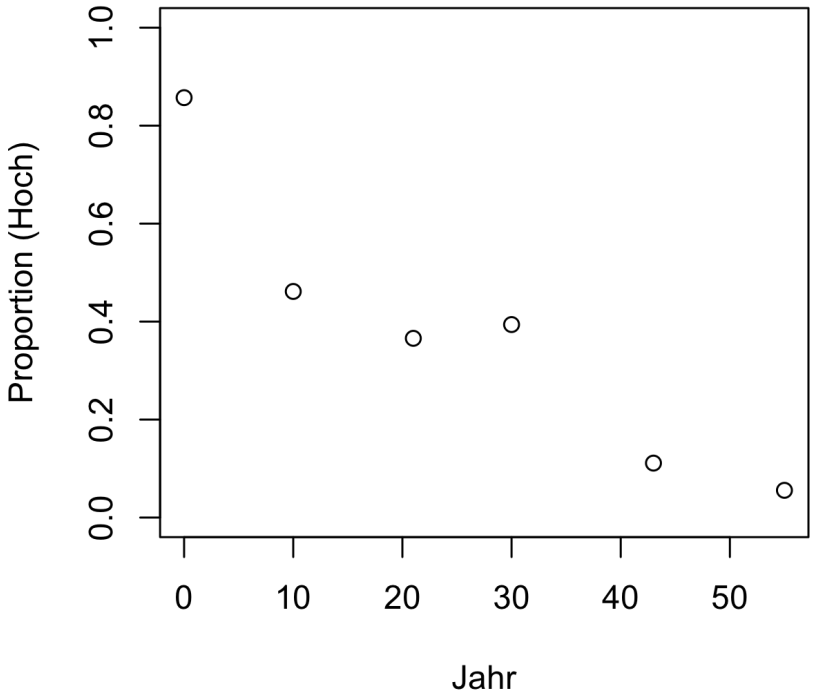
Ändert sich die Proportion der Sprecher, die *lost* mit hohem/tiefen Vokal zwischen 1950 und 2005 (= hat Jahr einen Einfluss auf die Proportionen von Hoch/Tief?)

Data-Frame und Abbildung

```
# Anzahl Hoch
P = c(30, 18, 15, 13, 4, 2)
# Anzahl nicht Hoch
Q = c(5, 21, 26, 20, 32, 34)
# Numerischer Faktor
Jahr = c(1950, 1960, 1971, 1980, 1993, 2005)
# Proportionen
prop = P/(P+Q)
# Data-Frame
lost = data.frame(P, Q, prop, Jahr)

ylim = c(0,1)
plot(prop ~ Jahr, data = lost, ylim=ylim,
xlab="Jahr", ylab="Proportion (Hoch)")
```

Jahr	Hoch	Tief
1950	30	5
1960	18	21
1971	15	26
1980	13	20
1993	4	32
2005	2	34



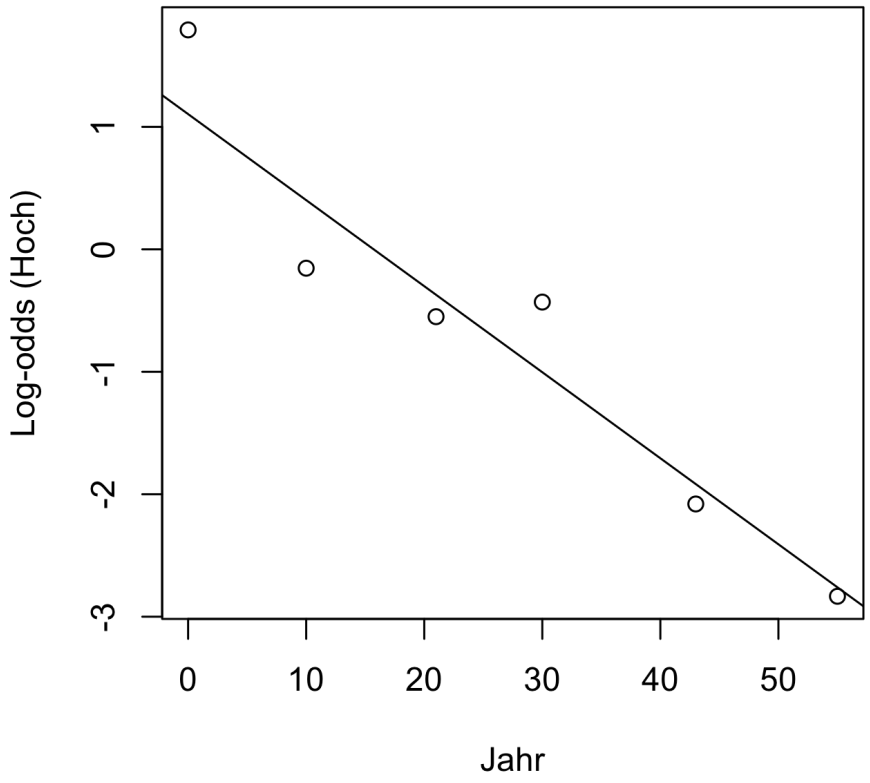
Analyse der Proportionen mit glm()

berechnet eine gerade Linie im Raum log(P/Q)

```
plot(log(P/Q) ~ Jahr, data = lost, ylab="Log-odds (Hoch)")

g = glm(cbind(P, Q) ~ Jahr, binomial, data=lost)

abline(g)
```

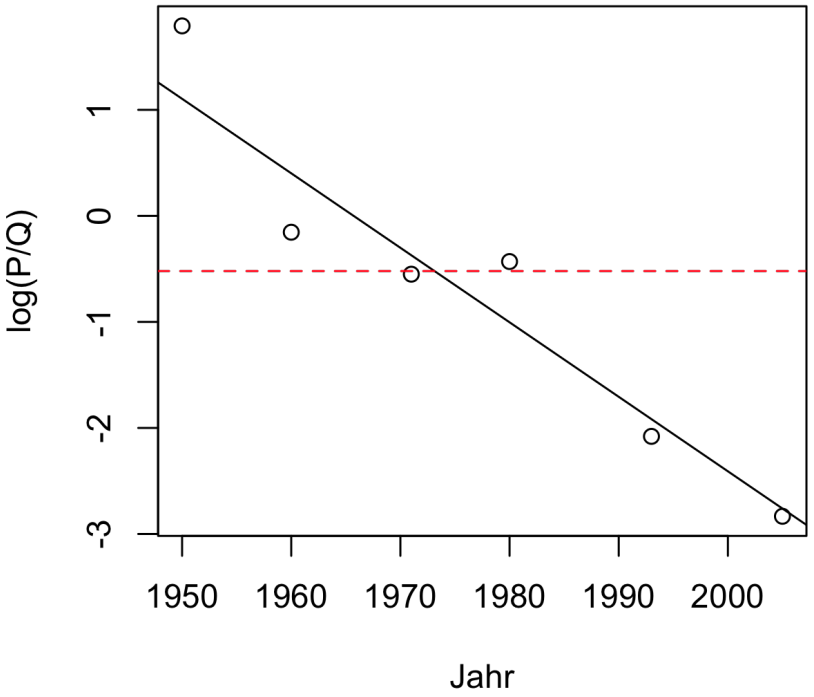


Analyse der Proportionen mit glm()

```
g = glm(cbind(P, Q) ~ Jahr, binomial, data=lost)
anova(g, test="Chisq")
```

```
      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     5     69.363
Jahr   1   61.121         4      8.242 5.367e-15 ***
```

berechnet die Wahrscheinlichkeit, dass Jahr einen Einfluss auf die Proportionen hat – genauer, ob sich diese beiden Linien voneinander signifikant abweichen



Die Proportionen wurden vom Jahr beeinflusst $\chi^2[1] = 61.1$, $p < 0.001$).

(Siehe auch letzte Folie)

Logistische Regression und kategoriale Variablen

Die Verteilung der t-Glottalisierung nach Gender war wie folgt:

	glottalisiert	nicht-glottalisiert
m	110	90
w	82	108

Hat Gender einen Einfluss auf die Proportion glottalisiert?

```
P = c(110, 82)
Q = c(90, 108)
prop = P/(P+Q)
Gender = factor(c("m", "w"))

glot = data.frame(P, Q, prop, Gender)
```

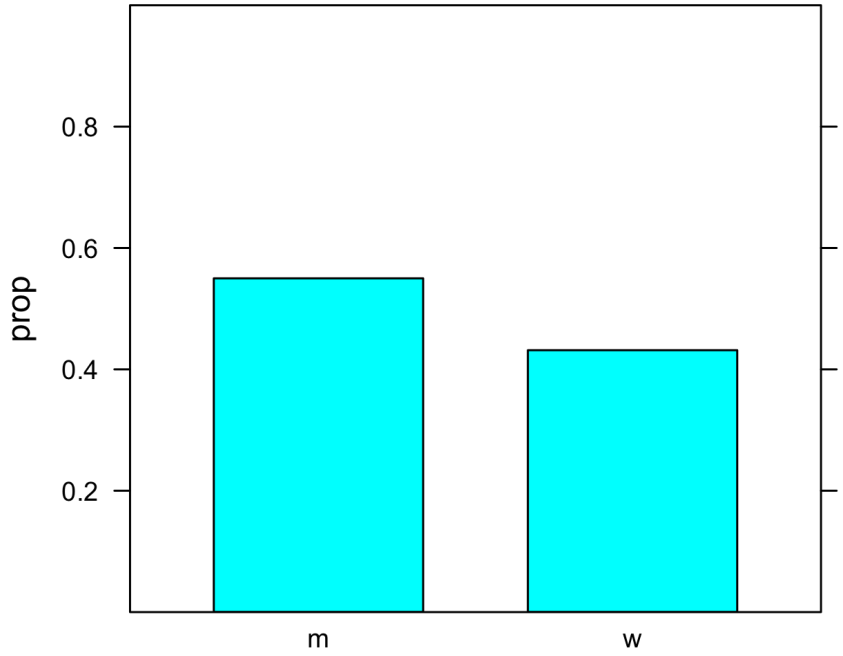
Abbildung

```
plot(prop ~ Gender, data = glot,
ylim=ylim)
```

besser

```
library(lattice)
```

```
barchart(prop ~ Gender,
data = glot, ylim=ylim)
```



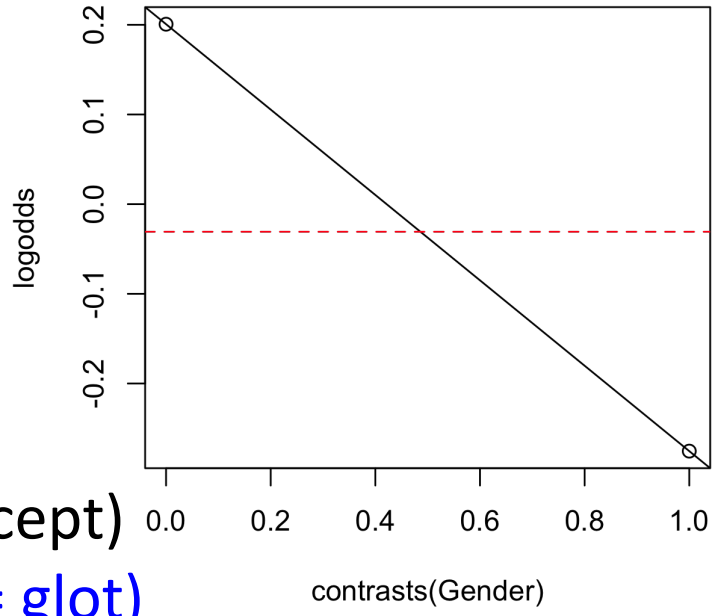

```
g = glm(cbind(P, Q) ~ Gender, binomial, data=glot)
anova(g, test="Chisq")
```

```
        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                        1     5.4801
Gender  1   5.4801          0   3.73e-14   0.01923 *
```

Graphische Darstellung der Prüfstatistik

Das gleiche wie für den numerischen Faktor (voriges Beispiel): in diesem Fall wird der kategoriale Faktor als 0 (m) und 1 (w) umkodiert.

```
logodds = with(glot, log(P/Q))
x = with(glot, contrasts(Gender))
plot(x, logodds)
abline(g)
```




```
# Berechnung ohne Faktor (nur mit Intercept)
g2 = glm(cbind(P, Q) ~ 1, binomial, data = glot)
abline(g2, col="red", lty=2)
```

Was ist die Wahrscheinlichkeit, dass sich diese Linien voneinander abweichen?

20 Sprecher, 11 aus Bayern, 9 aus Schleswig-Holstein produzierten *Sohn*. Ein Hörer beurteilte wie oft der erste Laut als /s/ oder /z/ produziert wurde. Hat Dialekt einen Einfluss auf die Urteile?

Lang- und Kurzformat

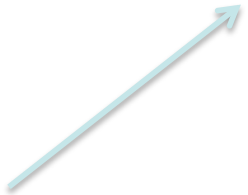
lang.df

	Urteil	D
1	z	BY
2	z	BY
3	z	BY
4	z	BY
5	s	BY
6	s	BY
7	z	BY
8	s	BY
9	z	BY
10	z	BY
11	z	BY
12	s	SH
13	z	SH
14	s	SH
15	s	SH
16	s	SH
17	z	SH
18	s	SH
19	s	SH
20	s	SH

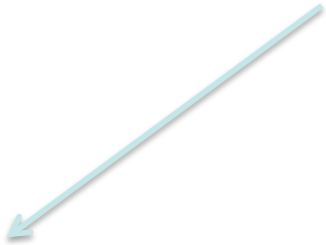
kurz.df

	P	Q	D
1	8	3	BY
2	2	7	SH

g2 = glm(cbind(P, Q) ~ D, binomial, data=kurz.df)



Äquivalente Ergebnisse

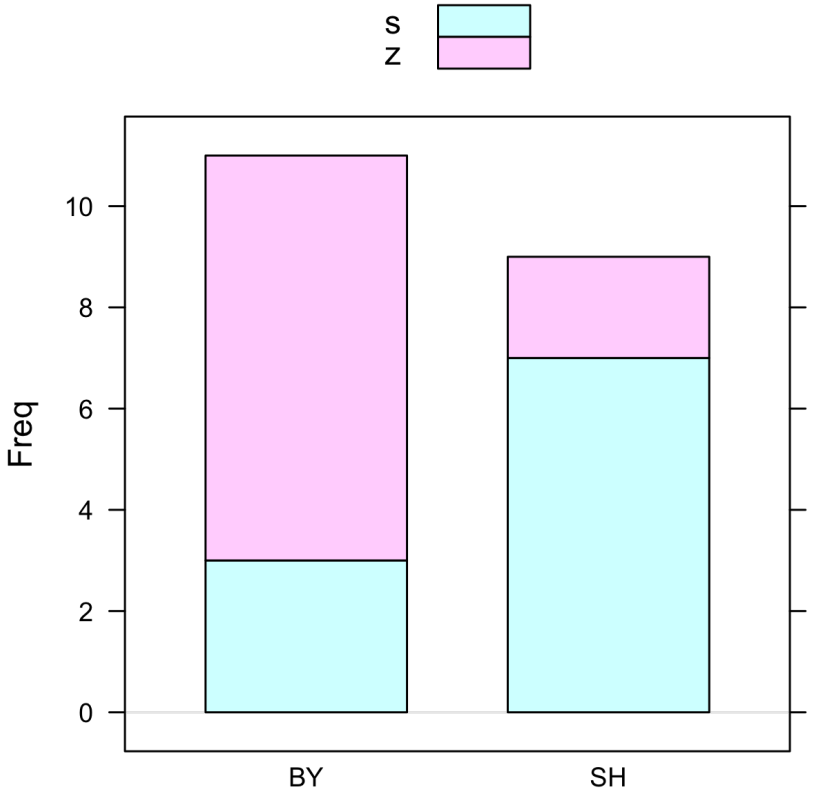


g1 = glm(Urteil ~ D, binomial, data=lang.df)

Abbildungen

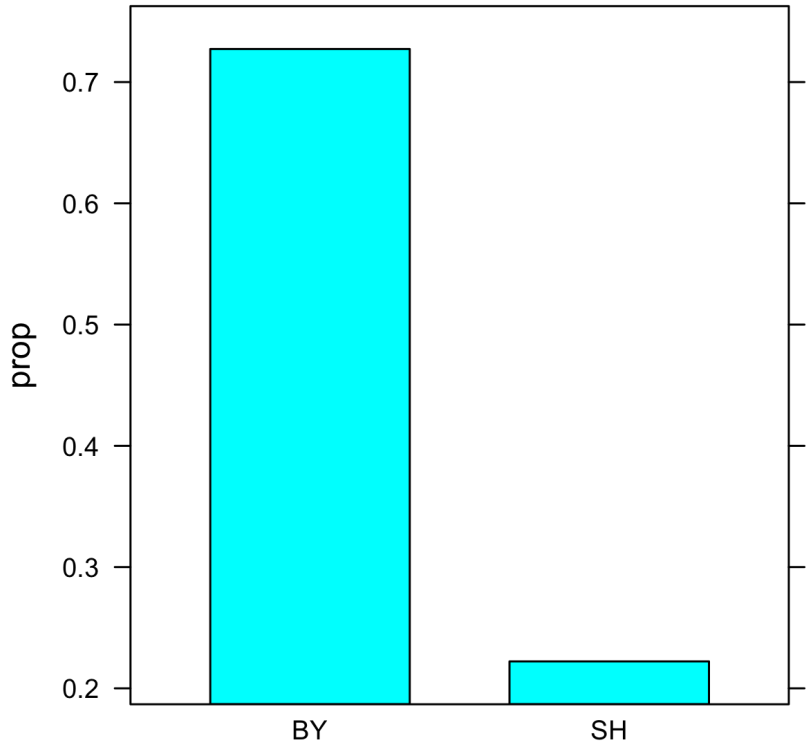
Lang-Format

```
tab = with(lang.df, table(D, Urteil))
barchart(tab, auto.key=T, horizontal=F)
```



Kurz-Format

```
prop = with(kurz.df, P/(P+Q))
barchart(prop ~ D, data = kurz.df)
```



Lang-Format

Für einen Data-frame *d*, mit binärem Response, *R*, und Faktor, *F*:

Tabelle

```
tab = with(d, table(F, R))
```

(*R* immer an letzter Stelle)

Abbildung

```
barchart(tab, auto.key=T, horizontal=F)
```

Modell

```
g = glm(R ~ F, binomial, d)
```

Test

```
anova(g, test="Chisq")
```

Konvertierung in Kurz-Format

Für ein data.frame *d*, mit binärem Response, *R*, und Faktor, *F*. R hat Stufen "J", "N"

logischer Vektor

```
temp = d$R == "J"
```

sum T

```
Pdaten = with(d, aggregate(temp, list(F), sum))
```

sum F

```
Qdaten = with(d, aggregate(!temp, list(F), sum))
```

```
P = Pdaten$x
Q = Qdaten$x
prop = P/(P+Q)
```

Data-Frame (kurz)

```
kurz.df = data.frame(P, Q, prop, F = Pdaten$Group.1)
```

Kurz-Format

Für einen Data-frame *d*, mit P, Q und Proportionen *prop*, und Faktor, *F*:

Abbildung

```
barchart(prop ~ F, data = d)
```

2 Faktoren

```
barchart(prop ~ F1 | F2, data = d)
```

3 Faktoren

```
barchart(prop ~ F1 | F2 * F3, data = d)
```

Modell

```
g = glm(cbind(P, Q) ~ F, binomial, d)
```

Test

```
anova(g, test="Chisq")
```

Logistische Regression und psychometrische Kurven

Sprachsynthese: F2 von einem Vokal wurde in 10 regelmäßigen auditiven Schritten zwischen einem hohen und niedrigen Wert variiert. Diese synthetischen Tokens wurden 5 Mal wiederholt, randomisiert, und einer Versuchsperson präsentiert. Die Versuchspersonen musste in einem forced-choice Test beurteilen, ob der Token /i/ oder /u/ war. Zu welchem F2-Wert liegt die perzeptive Grenze zwischen diesen Vokalen?

	P	Q	prop	F2
1	0	5	0.0	2311
2	0	5	0.0	2176
3	0	5	0.0	2023
4	0	5	0.0	1885
5	0	5	0.0	1770
6	0	5	0.0	1667
7	2	3	0.4	1548
8	4	1	0.8	1437
9	5	0	1.0	1351
10	5	0	1.0	1269

Der Vokal mit F2 = 1548 Hz wurde 2 Mal als /u/, 3 Mal als /i/ beurteilt.

```
F2 = c(2311, 2176, 2023, 1885, 1770, 1667, 1548, 1437, 1351, 1269)
P = c(rep(0, 6), 2, 4, 5, 5)
Q = c(rep(5, 6), 3, 1, 0, 0)
prop = P/(P+Q)
ui = data.frame(P, Q, prop, F2)
```

Psychometrische Kurve und Umkipppunkt

Abbildung F2 x Proportionen

```
plot(prop ~ F2, data = ui, ylab = "Proportion /u/-Urteile")
```

Modell mit F2 als numerischer Faktor

```
g = glm(cbind(P, Q) ~ F2, binomial, ui)
```

Koeffiziente (Intercept, Neigung)

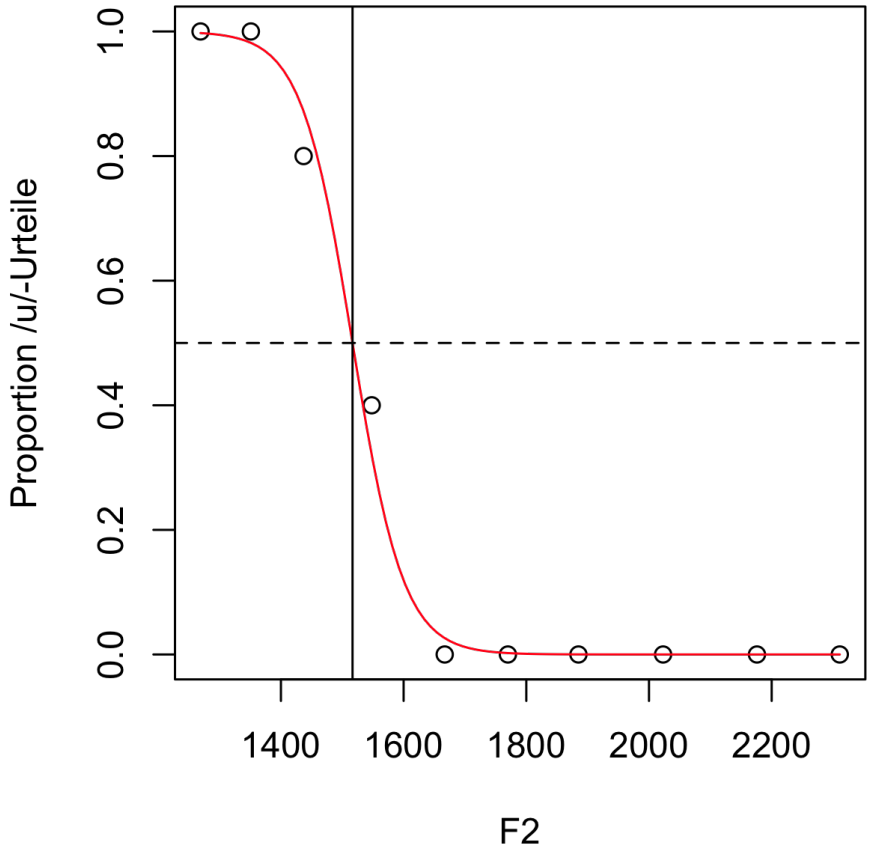
```
k = coef(g)[1]
m = coef(g)[2]
```

Psychometrische Kurve

```
curve(exp(m*x + k)/(1+ exp(m*x+k)),
add=T, col=2)
```

Umkipppunkt überlagern (= den F2-Wert, zu dem die Proportion = 0.5)

```
U = -k/m
abline(v = U)
```



```
      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     5     69.363
Jahr   1   61.121         4      8.242 5.367e-15 ***
```

Wenn sich diese Verhältnisse stark von 1 abweichen (wie hier, also 8.242/4 > 1), dann haben wir 'over-dispersion': die Proportionen können nicht sehr gut durch ein Binomial modelliert werden. In diesem Fall das Modell noch einmal mit `quasibinomial` und einen F-Test durchführen

gq = glm(cbind(P, Q) ~ Jahr, quasibinomial, data = lost)
anova(gq, test="F")

```
      Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                     5     69.363
Jahr   1   61.121         4      8.242 29.655 0.005522 **
```

Die Proportionen wurden vom Jahr beeinflusst F[1,4] = 29.7, p < 0.01).