Methoden empirischer Sozialforschung

Verallgemeinerte lineare Modelle
(Logistische Regression)

Verallgemeinerte lineare Modelle

1 binäre und mehrere metrische und kategoriale Variablen

- Wie läßt sich die Abhängigkeit der Erfolgswahrscheinlichkeit einer binären Variablen von mehreren metrischen und kategorialen Variablen beschreiben?
  Läßt sich der Wert einer binären Variablen anhand der Werte anderer (metrischer und kategorialer) Variablen vorhersagen?
  Logistische Regression

Promotion

Ein neues Medikament wird mit verschiedenen Promotionskampagnen bei Ärzten beworben: entweder durch einen Vertreter oder durch die Einladung zu einem Abendessen (`PROMOT`). Die Variable `TRY` erfaßt, ob die Ärzte danach das Produkt ausprobieren wollten oder nicht.

Frage: Ist die Wahrscheinlichkeit, das Produkt auszuprobieren, bei beiden Promotionskampagnen gleich? *(Homogenitätsproblem)*

Anders: Hängt die Wahrscheinlichkeit, das Produkt auszuprobieren, von der Art der Promotionskampagne ab?

Promotion

```
R> PROMOT <- read.csv2("promot.csv")
R> PROMOT

  COUNT TRY         PROMOT
1    58 yes         dinner
2    23  no         dinner
3    47 yes representative
4    38  no representative

R> tab <- xtabs(COUNT ~ PROMOT + TRY, data = PROMOT)
R> tab

                TRY
PROMOT           no yes
  dinner         23 58
  representative 38 47
```

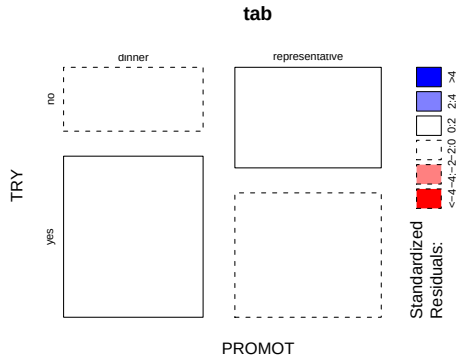
Promotion

Für dieses spezielle Homogenitätsproblem, wo es eine binäre abhängige und eine binäre erklärende Variable gibt, haben wir bereits Methoden kennengelernt: Odds Ratio, Mosaicplot, χ^2-Test.

```
R> prop.table(tab, 1)

                TRY
PROMOT            no       yes
  dinner         0.2839506 0.7160494
  representative 0.4470588 0.5529412

R> (23 * 47)/(38 * 58)

[1] 0.4904719
```

Promotion

```
R> mosaicplot(tab, shade = TRUE)
```



Promotion

```
R> prop.test(tab)

        2-sample test for equality of proportions with continuity correction

data:  tab
X-squared = 4.0715, df = 1, p-value = 0.04361
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.319435375 -0.006781037
sample estimates:
   prop 1    prop 2
0.2839506 0.4470588
```

Logistische Regression

Auch in Situationen, wo die abhängige Variable nicht normalverteilt ist (bzw. zumindest unimodal und symmetrisch), können Regressionsmodelle verwendet werden, sogenannte *verallgemeinerte lineare Modelle (GLMs)*.

Es gibt GLMs für abhängige Variablen, die u.a. binomialverteilt, Poisson-verteilt oder Gamma-verteilt sind. Das GLM mit Binomialverteilung wird auch *logistische Regression* genannt.

Anstatt die Variable direkt zu modellieren werden die abhängige Variable und die erklärenden Variablen mit einer *Link-Funktion* verbunden.

Logistische Regression

In einer Regression kann die Linearkombination der erklärenden Variablen potentiell alle reellen Zahlen annehmen, aber die abhängige Variable kann nur zwei Ausprägungen (Erfolg/Mißerfolg) annehmen. Auch die zugehörige Erfolgswahrscheinlichkeit π nimmt nur Werte in $[0, 1]$ an.

Deshalb modelliert man den Logarithmus des Odds Ratio. Im Beispiel:

$$\log\left(\frac{\text{Wahrscheinlichkeit(nicht probieren)}}{\text{Wahrscheinlichkeit(probieren)}}\right) \quad = \quad \text{Effekt(Promotion)}$$

Logistische Regression

Formal:

$$\log\left(\frac{\pi}{1-\pi}\right) \quad = \quad a + b \cdot x$$

wobei x in diesem Beispiel wieder ein 0/1-Variable ist. Wie auch bei linearen Regressionsmodellen können auf der rechten Seite nun sowohl quantitative Variablen oder qualtitative Variablen (bzw. deren 0/1-Kodierung) und dere Interaktion sthen.

In R werden diese Modelle mit dem Befehl `glm` angepaßt:

```
glm(formula, weights, data, subset, family = binomial, ...)
```

Logistische Regression

```
R> fm <- glm(TRY ~ PROMOT, weights = COUNT, data = PROMOT, family = binomial)
```

Dies erzeugt Objekte der Klasse `"glm"`, für die es wie für `"lm"` Objekte Methoden gibt, die

- eine `summary` erzeugen,
- die Koeffizienten extrahieren (`coef`),
- Modellvergleiche durch `anova` oder `AIC` durchführen,
- sowie `deviance`, `fitted`, `residuals` usw. berechnen.

Logistische Regression

```
R> fm

Call:  glm(formula = TRY ~ PROMOT, family = binomial, data = PROMOT,      weights

Coefficients:
         (Intercept)  PROMOTrepresentative
              0.9249                -0.7124

Degrees of Freedom: 3 Total (i.e. Null);  2 Residual
Null Deviance:            218.3
Residual Deviance: 213.5         AIC: 217.5
```

Logistische Regression

```
R> summary(fm)

Call:
glm(formula = TRY ~ PROMOT, family = binomial, data = PROMOT,
    weights = COUNT)

Deviance Residuals:
     1       2       3       4
 6.225  -7.610   7.463  -7.822

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.9249     0.2464   3.754 0.000174 ***
PROMOTrepresentative  -0.7124     0.3291  -2.165 0.030416 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 218.32  on 3  degrees of freedom
Residual deviance: 213.54  on 2  degrees of freedom
AIC: 217.54

Number of Fisher Scoring iterations: 4
```

Logistische Regression

```
R> anova(fm, test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: TRY

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       3    218.322
PROMOT  1    4.785         2    213.537     0.029
```

Logistische Regression

Die ANOVA in diesem Fall ist (fast) äquivalent zum χ^2-Test ohne Stetigkeitskorrektur:

```
R> prop.test(tab, correct = FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  tab
X-squared = 4.7473, df = 1, p-value = 0.02934
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.30738018 -0.01883623
sample estimates:
   prop 1    prop 2
0.2839506 0.4470588
```

Logistische Regression

Die Koeffizienten des Modells können als log-Odds Ratios interpretiert werden. Durch Anwendung der exp Funktion erhält man also Odds Ratios:

```
R> exp(coef(fm)[2])

PROMOTrepresentative
           0.4904719

R> (23 * 47)/(38 * 58)

[1] 0.4904719
```

Promotion

Zusammenfassend:

Die Art der Promotionsmaßnahme hat einen signifikanten Effekt auf die Bereitschaft von Ärzten ein neues Medikament auszuprobieren ($p = 0.029$). Die Chancen das neue Medikament auszuprobieren sind etwa doppelt (2.039 Mal) so hoch wenn Ärzte zu einem speziellen Abendessen eingeladen wurden verglichen mit jenen Ärzten, die von einem Pharmavertreter besucht wurden.

Auto-Sonderausstattung

Beispiel: Wichtigkeit von Auto-Sonderausstattungen

Personen verschiedener Altersgruppen wurde ersucht zu bewerten, wie viel Wert sie auf bestimmte Sonderausstattungen beim Autokauf legen, insbesonders wurde nach einer Bewertung für die Wichtigkeit von Klimatisierung gefragt.

```
R> AUTO <- read.csv2(file = "autoextras.csv")
R> tab <- xtabs(COUNTS ~ ALTER + KLIMA, data = AUTO)
```

Auto-Sonderausstattung

```
R> tab
```

```
          KLIMA
ALTER      nicht wichtig wichtig
  18-23    66            44
  24-40    26            63
  > 40     13            88
```

```
R> (66 * 63)/(26 * 44)
```

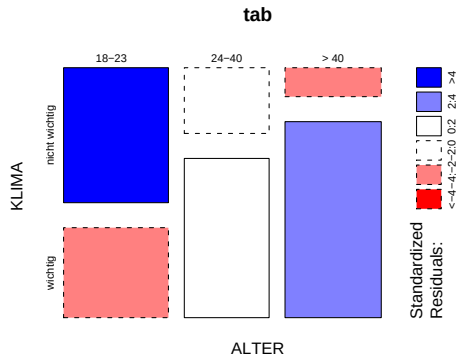
```
[1] 3.634615
```

```
R> (66 * 88)/(13 * 44)
```

```
[1] 10.15385
```

Auto-Sonderausstattung

```
R> mosaicplot(tab, shade = TRUE)
```



Auto-Sonderausstattung

```
R> prop.test(tab)

        3-sample test for equality of proportions without continuity
        correction

data:  tab
X-squared = 53.2693, df = 2, p-value = 2.708e-12
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.6000000 0.2921348 0.1287129
```

Auto-Sonderausstattung

```
R> fm <- glm(KLIMA ~ ALTER, weights = COUNTS, data = AUTO, family = binomial)
R> fm

Call:  glm(formula = KLIMA ~ ALTER, family = binomial, data = AUTO,      weights =

Coefficients:
(Intercept)   ALTER24-40    ALTER> 40
    -0.4055       1.2905       2.3179

Degrees of Freedom: 17 Total (i.e. Null);  15 Residual
Null Deviance:            388.5
Residual Deviance: 333.1         AIC: 339.1

R> exp(coef(fm)[2])

ALTER24-40
  3.634615

R> exp(coef(fm)[3])

ALTER> 40
 10.15385
```

Auto-Sonderausstattung

```
R> summary(fm)

Call:
glm(formula = KLIMA ~ ALTER, family = binomial, data = AUTO,
    weights = COUNTS)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.468  -4.662   2.553   3.527   5.582

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4055     0.1946  -2.083   0.0372 *
ALTER24-40    1.2905     0.3037   4.250 2.14e-05 ***
ALTER> 40     2.3179     0.3552   6.526 6.78e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 388.47  on 17  degrees of freedom
Residual deviance: 333.14  on 15  degrees of freedom
AIC: 339.14

Number of Fisher Scoring iterations: 5
```

Auto-Sonderausstattung

```
R> anova(fm, test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: KLIMA

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     17     388.47
ALTER  2    55.33        15     333.14 9.667e-13
```

Auto-Sonderausstattung

Interpretation:

Die Chance eine Klimaanlage als wichtig zu beurteilen ist für die Altersgruppe “24–40” rund 3.63 mal höher als für die Altersgruppe “18–23” (Referenzgruppe). Für die Altersgruppe “älter als 40” sogar 10 Mal höher. Mit zunehmendem Alter wird die Klimatisierung immer wichtiger.