

# Statistik

## 09. Generalisierte Lineare Modelle

Roland Schäfer

Institut für Germanistische Sprachwissenschaft  
Friedrich-Schiller-Universität Jena

stets aktuelle Fassungen: <https://github.com/rsling/VL-Deutsche-Syntax>

## 1 Generalisierte Lineare Modelle

- LM und GLM
- GLM Grundlagen
- Maximum Likelihood
- Nominale Unabhängige

- Modellselektion
- Modellevaluation
- Alternativen und Lösungen
- In R

## 2 Nächste Woche | Überblick

GLMs

- Generalisierte Lineare Modelle mit Logit-Link = Logistische Regression
- Regression zur Modellierung dichotomer Abhängiger
- Modellselektion für GLMs
- Modellevaluation für GLMs
- Problemlösungen (Ausblick):  
Zufallseffekte (GLMMs), Kreuzvalidierung, Bootstrapping, GAMs

- Backhaus u. a. 2011
- Zuur u. a. 2009
- Fahrmeir u. a. 2009

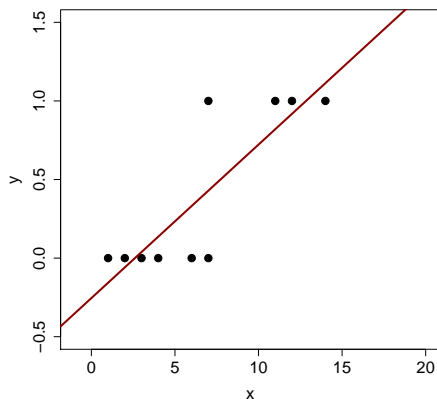
Alternation von Genitiv und Kasusidentität  
in der Maßangabe im Deutschen:

- *Wir trinken eine Flasche guten Wein.* (Agree=1)
- *Wir trinken eine Flasche guten Weines.* (Agree=0)
- Welche Faktoren beeinflussen die Wahl von Agree=1 oder Agree=0?
- Unabhängige hier:
  - ▶ Kasus der Maßangabe (Nom, Akk, Dat)
  - ▶ Definitheit der NP (0, 1)
  - ▶ Maß ist als Zahl geschrieben (0, 1)
- Das Beispiel kommt dann in der R-Session tatsächlich dran.

- LM sagt **kontinuierliche Werte** voraus
- unplausibel für dichotome Abhängige
- auch als Eintrittswahrscheinlichkeit unplausibel (außerhalb  $[0,1]$ )
- **Normalitätsannahmen nicht erfüllt**

# Illustration der Probleme

Datenpunkte einer dichotomen Abhängigen  $y$   
zu einer intervallskalierten Unabhängigen  $x$   
und lineares Modell  $y \sim x$





- Vorhersage der Eintrittswahrscheinlichkeiten
- lineare Kombination der Regressoren wie beim LM
- Linearkombination ergibt die Logits (z):

$$z = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_0$$

Die Logits werden transformiert in Eintrittswahrscheinlichkeiten mittels der **logistischen Funktion** ( $e$  ist die Euler-Konstante):

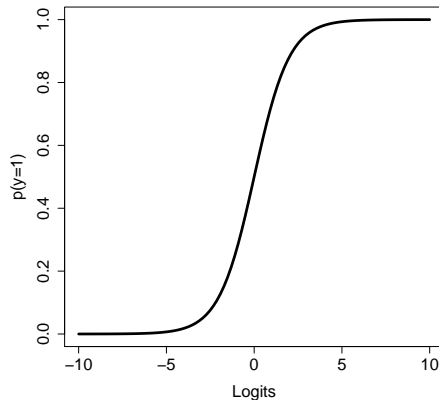
$$\hat{p}(y = 1) = \frac{1}{1 + e^{-z}}$$

Bei der **binären Vorhersage** dann:

$$\hat{y} = \begin{cases} 0 & \text{wenn } \hat{p}(y = 1) \leq 0.5 \\ 1 & \text{wenn } \hat{p}(y = 1) > 0.5 \end{cases}$$

# Darstellung des Effekts der Logit-Transformation

Die transformierten Logits als  $\hat{p}(y = 1)$ :



- Interpretation der Koeffizienten nur **indirekt** möglich
- $\beta_i$  positiv  $\Rightarrow$  positiver Einfluss auf  $\hat{p}(y = 1)$
- $\beta_i$  negativ  $\Rightarrow$  negativer Einfluss auf  $\hat{p}(y = 1)$
- Stärke des Einflusses: **nicht linear**
- linearer Einfluss nur auf die Logits, nicht auf  $\hat{p}(y = 1)$

- Chance (Odds):  $o(y=1) = \frac{p(y=1)}{1-p(y=1)}$
- Die Chancen des Modells verteilen sich (zum Glück) einfach:

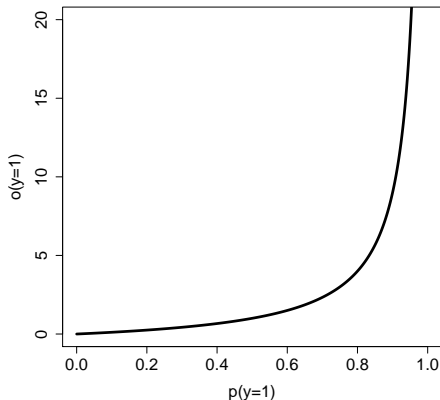
$$o(y=1) = \frac{p(y=1)}{1-p(y=1)} = e^z$$

Beachte:  $\ln(e^z) = z = \text{Logits}$

- Die Chance liegt offensichtlich in  $[0, \infty]$ .
- Mit steigender Wahrscheinlichkeit gehen die Odds gegen  $\infty$ .
- Bei einem Logit von 3 ist die Chance für  $y = 1$  doppelt so hoch wie bei einem Logit von 1.5 usw.

# Beziehung zwischen Wahrscheinlichkeit und Odds

In der Interpretation stellen die Odds die Linearität her, die den Wahrscheinlichkeiten bei der log. Regression fehlen.



Für die Interpretation der einzelnen Koeffizienten  $\beta_i$   
im Sinne eines Chancenverhältnisses:

$$or(y = 1|x_i) = e^{\beta_i}$$

In Worten: Steigt  $x_i$  (intervallskaliert!) um eine Einheit,  
dann steigt die Chance für  $y = 1$  um  $e^{\beta_i}$ .

Ein Chancenverhältnis von 1 entspricht einem Koeffizienten 0,  
also einem ohne jeglichen Effekt.

Beziehungen zwischen den Maßen  
sowie ihre Wertebereiche.

Einzel-Koeffizient		Gesamtmodell		
Koeffizient	Chancenverhältnis	Logit	Chance	$\hat{p}(y = 1)$
$\beta > 0$	$e^{\beta} > 1$	steigt um $\beta x$	steigt um $e^{\beta x}$	steigt
$\beta < 0$	$e^{\beta} < 1$	sinkt um $\beta x$	sinkt um $e^{\beta x}$	sinkt
$[-\infty, +\infty]$	$[0, +\infty]$	$[-\infty, +\infty]$	$[0, +\infty]$	$[0, 1]$



- Es gibt keine direkte Lösung für die Koeffizientenberechnung.
- Das Schätzverfahren funktioniert iterativ.
- Es kommt der sog. Maximum-Likelihood-Schätzer zum Einsatz.

- Es gibt beliebig viele Modelle = Belegungen für die  $\beta$ -Koeffizienten
- Das **wahrscheinlichste Modell angesichts der Beobachtungen** ist zu finden.
- In den Beobachtungsdaten für jeden Fall  $k$ :  $y_k = 1$  oder  $y_k = 0$
- Für jeden Beobachtungswert  $y_k$  betrachtet man:

$$p_k = \left( \frac{1}{1+e^{-z_k}} \right)^{y_k} \cdot \left( 1 - \frac{1}{1+e^{-z_k}} \right)^{1-y_k}$$

$$p_k = \left(\frac{1}{1+e^{-z_k}}\right)^{y_k} \cdot \left(1 - \frac{1}{1+e^{-z_k}}\right)^{1-y_k}$$

- $z_k$  ist der Modell-Logit für die zu  $y_k$  empirische gemessenen  $x$ .
- In den ( ) steht links die vom Model geschätzte Wahrscheinlichkeit  $\hat{p}(y_k)$  und rechts jeweils die Gegenwahrscheinlichkeit dazu  $1 - \hat{p}(y_k)$ .
- Wenn der Modellwert nahe an 0 (z. B. 0.1) und  $y_k = 0$  ist:  
 $p_k = (0.1)^0 \cdot (0.9)^1 = 1 \cdot 0.9 = 0.9$  („gute“ Approximation)
- Wenn der Modellwert bei gleichen empirischen Daten umgekehrt ist:  
 $p_k = (0.9)^0 \cdot (0.1)^1 = 1 \cdot 0.1 = 0.1$  („schlechte“ Approximation)
- Die  $p_k$  messen also die Güte der vom Modell vorhergesagten Wahrscheinlichkeit für jeden beobachteten Datenpunkt.

- Bei unabhängigen Ereignissen  $E_{1..n}$  gilt:  
$$P(E_1 + E_2 + \cdots + E_n) = \prod_i P(E_i)$$
- Die Wahrscheinlichkeit eines Modells (seine „Likelihood“) angesichts aller empirischen Werte  $y_k$  ist also:

$$L = \prod_k p_k$$

- Der Maximum Likelihood-Schätzer maximiert  $L$  für die Belegungen der  $\beta$ -Koeffizienten (= konkurrierende Modelle).

Wie bei der LM-Variante der ANOVA müssen kategoriale Unabhängige mit mehr als zwei Ausprägungen als dichotome Dummy-Variablen kodiert werden.

**Beispiel für dreiwertige Variable A und Dummy-Regressoren  $x_{1..3}$**

	A = 1	A = 2	A = 3
$x_1 =$	1	0	0
$x_2 =$	0	1	0
$x_3 =$	0	0	1

Achtung! De facto gibt es für einen kategorialen Regressor mit  $k$  Ausprägungen nur  $k - 1$  Dummies (s. Abschnitt zum Intercept).

Beispiel für eine als  $x_{1..3}$  dummy-kodierte Unabhängige A und eine intervallskalierte Unabhängige  $x_4$ :

$$\hat{p}(y = 1) = \frac{1}{1 + e^{-z}}$$

$$\text{mit } z = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_0$$

Dabei treten die Werte auf:

- $x_{1..3}$ : 0 oder 1
- Wenn  $x_1 = 1$ , dann  $x_2 = 0$  und  $x_3 = 0$  usw.

(Wh.:) Für die Interpretation der einzelnen Koeffizienten  $\beta_i$   
im Sinne eines Chancenverhältnisses:

$$or(y = 1|x_i) = e^{\beta_i}$$

In Worten für nominale Regressoren bzw. ihr dichotomen Dummies:

Wenn  $x_i = 1$  ( $x_i$  ist dichotom skaliert!),  
dann ist die Chance  $o(y = 1)$  um  $e^{\beta_i}$  höher als bei  $x_i = 0$ .  
Andere Fälle gibt es wegen der dichotomen Skalierung nicht.

- „Intercept“ ( $\beta_0$ ) in GLMs  $\neq$  Schnittpunkt mit y-Achse
- **intervallskalierte Regressoren:**
  - ▶ einfachstes binomiales GLM:  $\hat{p}(y = 1) = \beta_1 x_1 + \beta_0$
  - ▶ Wenn  $x_1 = 0$ , wird  $\beta_0$  vorhergesagt.
- bei **Dummy-Variablen** wird eine zur Referenz-Kategorie:
  - ▶ GLM mit drei Dummies:  $\hat{p}(y = 1) = \beta_{Akk} \cdot x_{Akk} + \beta_{Dat} \cdot x_{Dat} + \beta_{Nom}$
  - ▶ „Alle Regressoren werden 0“ heißt hier, es liegt Nom vor.
  - ▶ Die Dummies modellieren den **Unterschied zwischen Referenz (Nom) und den anderen Fällen**.
  - ▶ Die Referenzkategorie sollte die häufigste sein, besonders bei Interaktionen.



- nichts wesentlich anderes als in LM
- vereinte Effekte, die über die Einzeleffekte hinausgehen
- bei Interpretationsschwierigkeiten ggf. nachlesen

- Signifikanz wird für das Modell und Koeffizienten bestimmt.
- Allerdings: Signifikanz heißt nicht automatisch Modellgüte.
- Je „weniger signifikant“ ein Regressor, desto wahrscheinlicher kann er ohne Güteverlust entfernt werden.
- Modellselektion: Auswahl des **einfachsten Modells** mit der **größten Modellgüte**.
- Achtung bei dichotomen Dummy-Regressoren:  
Immer **alle** Dummies im Modell lassen oder herausnehmen, die zu einer kategorialen Unabhängigen gehören!

- 1 Weglassen des Regressors mit der geringsten Signifikanz
- 2 Vergleich des vollen und des reduzierten Modells
- 3 bei nicht-signifikantem Unterschied: Regressor weglassen
- 4 von vorne beginnen...

Log-Likelihood-Ratio für Likelihood des vollen ( $L_f$ ) und reduzierten ( $L_r$ ) Modells:

$$LR = (-2 \cdot \ln(L_r)) - (-2 \cdot \ln(L_f))$$

Test: Unter der  $H_0$   $L_r = L_f$  ist die LR  $\chi^2$ -verteilt  
mit  $df = df_f - df_r$  ( $df$  jeweils: Zahl der Regressoren)

Ist die LR größer als der kritische Wert: Regressor im Modell lassen!

Regressoren-Selektion auf Basis des **Akaike Information Criterion**:

- Ablauf wie bei LR-Test
- Maß für Modellvergleich ist das AIC
- Informationstheoretisches Maß:  
**Distanz des Modells zur (geschätzten) absoluten Realität**
- Je kleiner das AIC, desto besser das Modell.
- Achtung: Nur zum Vergleich **eingebetteter Modelle** verwenden, also bei gleichem Datensatz, und wenn das reduzierte Modell eine Teilmenge der Regressoren des vollen enthält.

- Signifikanzbestimmung für einzelne Regressoren
- wie bei LM: **Standardfehler** für jeden Regressor
- darauf basierend: **z-Wert** für jeden Regressor...
- und **z-Test** auf Basis der Normalverteilung

- Log-Likelihood-Ratio-Test für Gesamtheit aller Regressoren
- volles Modell (ggf. nach Eliminierung von Koeffizienten)
- **Nullmodell**, das nur einen konstanten Term zur Vorhersage nutzt
- ähnlich den Modellvergleichen im Kapitel „ANOVA als LM“

- auch Vergleich des vollen Modells und Nullmodells
- Interpretation wie gewohnt: Varianzerklärung

$$\text{Cox \& Snell: } R_C^2 = 1 - \left(\frac{L_0}{L_f}\right)^{\frac{2}{n}}$$

Problem: Geht nicht bis 1!

$$\text{Nagelkerke: } R_N^2 = \frac{R_C^2}{R_{max}^2}$$

$$\text{mit } R_{max}^2 = 1 - (L_0)^{\frac{2}{n}}$$

- gutes GLM  $\Rightarrow$  gute Vorhersagen
- einfache Vorhersagegüte: Anteil der richtigen Vorhersagen
- instruktiv: Vergleich mit „Baseline“  
(= Anteil der richtigen Vorhersagen bei Vorhersage der modalen Kategorie)
- Problem wie bei Fehlerreduktion:  
auch bei starkem Effekt nicht unbedingt Umkehrung der modalen Kategorie



- zugrundegelegte Verteilung: **Binomialverteilung**
- Überdispersion: Varianz ist größer als für Binomialverteilung angenommen
- mögliche Gründe:
  - ▶ unbeobachtete Heterogenität (fehlende erklärende Variablen)
  - ▶ Gruppenbildung (= Beobachtungen nicht unabhängig)

Schätzung des **Dispersionsparameters**:

$$\hat{\phi} = \sum \left( \frac{R_P}{df_R} \right)^2$$

wobei:  $R_P$  ist das **Pearson-Residual** (hier nicht behandelt) und

$df_R$  die **Residual-Freiheitsgrade**  $n - p$ ,  $p$  die Anzahl der Modellparameter

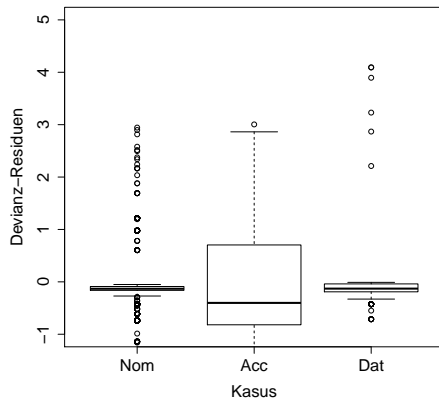
- Problem:  $\hat{\phi}$  deutlich über 1
- Lösung: Schätzung der Parameter bleibt (im Ergebnis) gleich
- aber für die Evaluation der Koeffizienten:
  - ▶ Signifikanzschätzung mit größeren Standardfehlern
  - ▶ t-Verteilung statt Normalverteilung (z-Werte)
- Ein „Quasi-Likelihood-Modell“ folgt im Wesentlichen dieser Strategie.

- (Multi-)kollinearität: Abhängigkeit zwischen Regressoren
- Probleme:  $\beta$ -Fehler, Überanpassung, ungenaue Koeffizientenschätzung
- Test: Varianzinflations-Faktoren (nicht im Detail behandelt)
- Lösungen z. B.: mehr Daten, Regressoren weglassen
- Test des Modells auf Robustheit trotz Kollinearität (z. B. Kreuzvalidierung)

# Varianzhomogenität

Die Residuen werden im GLM zwar anders berechnet, sind aber trotzdem ein Maß für die Varianz.

Die Varianz sollte nicht mit den Regressorausprägungen variieren!



- bei Problemen: Test auf **Robustheit des Modells**
- Idee bei  $k$ -facher Kreuzvalidierung:
  - 1 teile Daten in  $k$  Teile
  - 2 Modellanpassung auf  $k - 1$  von  $k$  Teilen
  - 3 Prüfung der Vorhersage auf verbleibendem Teil
  - 4 Modell ist Robust, wenn die Parameter in der Kreuzvalidierung nicht wesentlich anders geschätzt werden als im Ursprungsmodell
- wenn  $k = n$ : **Leave-One-Out-Kreuzvalidierung**
- verwandtes Verfahren: **Bootstrapping** (mit Zurücklegen)

Einige typische Anwendungsfälle für nicht-binomiale GLMs:

- Zähldaten: **Poisson**
- Zähldaten mit Überdispersion: **negativ-binomial**
- bestimmte Intervalldaten in  $[0, \infty]$ : **Gamma**
- viele Nullen: **zero-inflated** Varianten

Das Vademecum, vor allem für R-Benutzer:  
Zuur u. a. 2009

- typisches gemischtes Modell: mit Zufallseffekten
- Idee: Varianzunterschiede oder Dispersion durch Gruppen
- mögliche Gruppen in linguistischen Experimenten:
  - ▶ Werte von einem Probanden bei Befragung, Rating-Studie
  - ▶ Werte zu einem Lexem bei Korpusstudie
  - ▶ Werte aus einer Textsorte bei Korpusstudie
- ideal: Gruppeneffekte durch zusätzliche normale Regressoren auflösen
- sonst (vereinfacht): Schätzung eines Intercepts pro Gruppe
- Typisch für Zufallseffekte: In der GG sind vermutlich viel mehr Ausprägungen vorhanden, als gemessen (wie z. B. Sprecher oder Lexeme) wurden.

GAMs oder „nichtparametrische Regression“

$$\hat{y} = f_1(x_1) + f_2(x_2) + \dots + f_n(x_n) + \beta_0$$

- $f_n$ : besondere Art von Funktion, die geschätzt wird
- Wenn die Funktionen ungefähr linear sind, ist ein GLM genauso gut.
- Interpretation von GAMs: viel schwieriger als GLMs
- letzter Ausweg bei schlechtem GLM



**1** Modell-Anpassung:

```
> m <- glm(y ~ x1+x2*y3, data=mydata, family="binomial")  
> summary(m)
```

**2** Chancenverhältnisse für Koeffizienten:

```
> exp(coef(m))
```

**3** 95%-Konfidenzintervalle für Chancenverhältnisse:

```
> exp(confint(m))
```

**4** Log-Likelihood extrahieren:

```
> logLik(m)
```

**5** Nagelkerke  $R^2$ :

```
> library(fmsb); NagelkerkeR2(m)
```

**6** LR-Test:

```
> m0 <- glm(y ~ 1, data=mydata, family="binomial")  
> lr <- (-2*logLik(m0)) - (-2*logLik(m))  
> pchisq(lr, m$rank-m0$rank)
```

- 7 Modellselektion (wenn nicht von Hand):  
`> drop1(m)`
- 8 Varianzinflationsfaktoren:  
`> library(car); vif(m)`
- 9 Dispersion  $\hat{\phi}$  schätzen:  
`> sum(resid(m, type="pear")^2 / df.residual(m))`
- 10 Vorhersagegüte:  
`> pred <- ifelse(predict(m) <= 0.5, 0, 1)`  
`> tab <- table(pred, mydata$response)`  
`> sum(diag(tab))/sum(tab)`
- 11 Fehlerrate in Kreuzvalidierung (hier  $k = 10$ ):  
`library(boot); cv.glm(mydata, m, K=10)$delta`

Nächste Woche | Überblick

- 1 Statistik, Inferenz und probabilistische Grammatik
- 2 Deskriptive Statistik
- 3 Nichtparametrische Verfahren
- 4 z-Test und t-Test
- 5 ANOVA
- 6 Freiheitsgrade und Effektstärken
- 7 Power
- 8 Lineare Modelle
- 9 Generalisierte Lineare Modelle
- 10 **Gemischte Modelle**

- Backhaus, Klaus, Bernd Erichson, Wulff Plinke & Rolf Weiber. 2011. *Multivariate Analysemethoden*. 13. Aufl. Berlin etc.: Springer.
- Fahrmeir, Ludwig, Thomas Kneib & Stefan Lang. 2009. *Regression – Modelle, Methoden und Anwendungen*. 2. Aufl. Heidelberg etc.: Springer.
- Zuur, Alain F., Elena N. Ieno, Neil Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Berlin etc.: Springer.

## Kontakt

Prof. Dr. Roland Schäfer  
Institut für Germanistische Sprachwissenschaft  
Friedrich-Schiller-Universität Jena  
Fürstengraben 30  
07743 Jena

<https://rolandschaefer.net>  
[roland.schaefer@uni-jena.de](mailto:roland.schaefer@uni-jena.de)

## Creative Commons BY-SA-3.0-DE

Dieses Werk ist unter einer Creative Commons Lizenz vom Typ *Namensnennung - Weitergabe unter gleichen Bedingungen 3.0 Deutschland* zugänglich. Um eine Kopie dieser Lizenz einzusehen, konsultieren Sie

<http://creativecommons.org/licenses/by-sa/3.0/de/> oder wenden Sie sich brieflich an Creative Commons, Postfach 1866, Mountain View, California, 94042, USA.