Statistik für Digital Humanities

Logistische Regression

Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

25. Mai 2020

[Letzte Aktualisierung: 28/05/2020, 10:25]

Logistische Regression

Logistische Regression:

- $\hat{Y}$ = kategorische Variable
- Gesucht: Wahrscheinlichkeit der Zugehörigkeit einer Gruppe $P(Y)$
- Beispiele: Geschlechtszuordnung anhand Lautstärke der Rülpser

Binäre Logistische Regression:

- 2 Kategorien als Outcome
- Geschlecht (Nominal), Bestanden/Durchgefallen (Ordinal), Ja/nein

Multinominale Logistische Regression: auch polychotome L. R.

- Mehr als 2 Kategorien
- Stadtviertelzugehörigkeit, Notenvorhersage, Lieblingsfarbe

Berechnung

- Verletzung der Annahme *Linearität der Outcomes*
- $\rightarrow$ "Formulieren der multiplen linearen Regression in logarithmischer Art umgeht das Problem"
- $P(Y) = \frac{1}{1+e^{-(b_0+b_1*X_1)}}$
- $P(Y) = \frac{1}{1+e^{-(b_0+b_1*X_1+b_2*X_2+\cdots+b_m*X_m)}}$ Logistische Funktion = Umkehrfunktion des Logit
  `https://de.wikipedia.org/wiki/Logit`
- e = Basis des natürlichen Logarithmus ≈ 2.718
  Eulersche Zahl $\rightarrow y = e^x$ Umkehrfunktion von $y = \ln x$
- $P(Y) \in [0, 1] \rightarrow [\textit{unwahrscheinlich}, \textit{wahrscheinlich}]$
- wir berechnen also nicht mehr Y, sondern dessen Wahrscheinlichkeit / Likelihood

Fitness des Modells

Log-Likelihood

- $Y_i \in \{0,1\} \rightarrow \{nein, ja\}, P(Y) \in [0,1]$
- Log-Likelihood $= \sum_{i=1}^{n} Y_i * \ln P(Y_i) + (1 - Y_i) * \ln(1 - P(Y_i))$
  $\rightarrow$ Für jede Person i
- negativer Wert
- Je größer Betrag, desto schlechter ist der Fit , desto höher unerklärte Varianz

Abweichung

- $-2LL = -2 * LogLikelihood$
- χ^2-Verteilung Chi Square
- Je größer, desto schlechterer Fit

Likelihood-Ratio

- $\chi^2 = 2LL(model) - 2LL(baseline) = -2LL(baseline) - (-2LL(model))$
- Baseline-Modell bei Log. Regression = Modell reduziert auf $Y = b_0$
- $df = k_{model} - k_{baseline} = (Anzahl\ Prädiktoren + 1) - 1 = Anzahl\ Prädiktoren$
- je höher desto größer der Unterschied

Fitness des Modells

R, R^2

- Verschiedene Berechnungen, die verschieden interpretiert werden müssen
- Anders zu behandeln als sonst
- $\rightarrow$ Homser, D.C. & Lemeshow, S. (1989): *Applied logistic regression*
  $\rightarrow$ Cox, D.R. & Snell, D.J. (1989): *The analysis of binary data*
  $\rightarrow$ Nagelkerke, N.J.D. (1991): *A note on a general definition of the coefficient of determination*
- Besser mit höherer Anzahl der Prädiktoren
- Kann als Effektstärke verwendet werden
- Codevorlage Siehe Moodle.

Akaike Information Criterion

- $AIC = -2LL + 2k$ mit $k =$ Anzahl der Prädiktoren
- Je höher desto schlechter

Bayes Information Criterion

- $BIC = -2LL + 2k * log(n)$ mit $n =$ Anzahl der Fälle
- Je höher desto schlechter, normalisiert auf Datenmenge

Fitness der Prädiktoren – Wald Statistik

- $z_{b_i} = \frac{b_i}{SE_b}$
- z-Verteilung (Normalverteilt)
- $H_0 = b_i$ ist signifikant ähnlich 0
- $z_{b_i} > z_{kr} \rightarrow H_0$ unwahrscheinlich
- Achtung: je höher b, desto aufgeblähter SE, desto wahrscheinlicher Typ-2 Fehler False Negative
  $\rightarrow$ Menard, S. (1995): *Applied logistic regression analysis*
- Likelihood Ratio sicherer

Odds / Chance

- $odds = \frac{P(event)}{P(no\ \ event)}$
- $P(event) = \frac{1}{1+e^{b_0+b_1*X_1+\cdots+b_m*X_m}}$
- $P(no\ event) = 1 - P(event)$
- Odds Ratio $\Delta odds = \frac{odds(\ddot{A}nderungum1Einheit)}{odds(Original)}$
- $\Delta odds > 1 \rightarrow$ Chance erhöht sich mit Prädiktor
  $\Delta odds < 1 \rightarrow$ Chance sinkt mit Prädiktor
- Konfidenzintervalle von $\Delta odds$ geben zusätzliche Sicherheit
- Wenn KI_U oder KI_O die 1-Grenze überschreiten ist es problematisch

Beispiel

Daten (Siehe R-Projekt im Moodle):

Bestanden	Gelernt	Fachsemester
Nicht Bestanden	Nicht Gelernt	7
Bestanden	Nicht Gelernt	6
Bestanden	Gelernt	8
...	...	...

```
#Daten Einlesen und die ersten 6 Daten anzeigen
bestandenData<-read.delim("logRegBsp.dat", header = TRUE)
head(bestandenData)

#Baseline definieren
bestandenData$Bestanden<-relevel(bestandenData$Bestanden, "Nicht Bestanden")
bestandenData$Gelernt<-relevel(bestandenData$Gelernt, "Nicht Gelernt")
```

Beispiel

```
#Hierarchie mit 2 Modellen
# Nur Gelernt
model.gelernt <- glm(Bestanden ~ Gelernt, data = bestandenData,
    family = binomial())

# Gelernt und Fachsemester
model.gelerntFS <- glm(Bestanden ~ Gelernt + Fachsemester, data = bestandenData,
    family = binomial())

summary(model.gelernt)
summary(model.gelerntFS)
```

Beispiel model.gelernt

```
summary(model.gelernt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5940  -1.0579   0.8118   0.8118   1.3018

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.2877     0.2700  -1.065  0.28671
GelerntGelernt   1.2287     0.3998   3.074  0.00212 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 154.08  on 112  degrees of freedom
Residual deviance: 144.16  on 111  degrees of freedom
AIC: 148.16

Number of Fisher Scoring iterations: 4
```

Beispiel model.gelernt

```
summary(model.gelernt)

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 154.08  on 112  degrees of freedom   #-2LL(Baseline)
                                                        = Abw(Baseline)
Residual deviance: 144.16  on 111  degrees of freedom   #-2LL(Modell)
                                                        = Abw(Modell)
AIC: 148.16
```

- $Abw(Modell) < Abw(Baseline) \rightarrow$ Modell verbessert
- $\chi^2 = Abw(Modell) - Abw(Baseline) = 9.92 > \chi^2_{kr95}(df = 1) = 3.84 \rightarrow$ Verbesserung signifikant

Beispiel model.gelernt

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.2877     0.2700  -1.065  0.28671
GelerntGelernt   1.2287     0.3998   3.074  0.00212 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$P(Best., Nichtgel.) = \frac{1}{1+e^{-(-0.288+(1.229*0))}} = 0.428$
$P(NichtBest., Nichtgel.) = 1 - P(Best., Nichtgel.) = 0.572$
$OriginalOdds = \frac{0.428}{0.572} = 0.748$
$P(Best., Gel.) = \frac{1}{1+e^{-(-0.288+(1.229*1))}} = 0.719$
$P(NichtBest., Gel.) = 1 - P(Best., Nichtgel.) = 0.281$
$Odds(x = 1) = \frac{0.719}{0.281} = 2.559$

Beispiel model.gelernt

$\Delta Odds = \frac{2.56}{0.75} = 3.41$

- $\Delta Odds > 1 \rightarrow$ Die Chance zu bestehen erhöht sich mit Lernen
- Die Chance auf Bestehen wenn man gelernt hat ist 3.42 mal höher als wenn nicht
- 95% Konfidenzintervalle sind ebenfalls $> 1 \rightarrow$ Entweder sehr sicher oder man hat eine der 5% Stichproben erwischt

```
#Konfidenzintervalle
exp(confint(model.gelernt))
```

```
                   2.5 %   97.5 %
(Intercept)    0.4374531 1.268674
GelerntGelernt 1.5820127 7.625545
```

Beispiel model.gelerntFS

```
summary(model.gelerntFS)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6025  -1.0572   0.8107   0.8161   1.3095

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.234660   1.220563  -0.192  0.84754
GelerntGelernt  1.233532   0.414565   2.975  0.00293 **
Fachsemester   -0.007835   0.175913  -0.045  0.96447
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 154.08  on 112  degrees of freedom
Residual deviance: 144.16  on 110  degrees of freedom
AIC: 150.16

Number of Fisher Scoring iterations: 4
```

Beispiel model.gelerntFS

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.234660   1.220563  -0.192  0.84754
GelerntGelernt  1.233532   0.414565   2.975  0.00293 **
Fachsemester   -0.007835   0.175913  -0.045  0.96447
```

- $b_{Fachsemester} = -0.008 \rightarrow$ Geringer Einfluss
- $Pr(> |z|) > 0.05 \rightarrow$ Einfluss nicht signifikant

Beispiel model.gelerntFS

```
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 154.08  on 112  degrees of freedom
Residual deviance: 144.16  on 110  degrees of freedom
AIC: 150.16
```

- AIC hat sich erhöht im Vergleich zu *model.gelernt* →schlechter

```
modelChi <- model.gelernt$deviance - model.gelerntFS$deviance
chidf <- model.gelernt$df.residual - model.gelerntFS$df.residual
chisq.prob <- 1 - pchisq(modelChi, chidf)
modelChi; chidf; chisq.prob

[1] 0.001983528    #Unterschied zwischen model.gelernt und model.gelerntFS
[1] 1
[1] 0.9644765      #>0.5 -> keine sign. Verbesserung im Vergleich zu model.gelernt
```

Auswahl der Prädiktoren

Erzwungen (Alle auf einmal)

- Alle Prädiktoren zum Modell hinzufügen und individuell evaluieren

Schrittweise

- vorwärts/rückwärts/Hybrid: Siehe Folien zu Multipler Regression
- Verbesserung / Verschlechterung feststellbar mit *AIC* und *BIC*

Grundannahmen

- Unabhängigkeit der Fehler, geringe Multikollinearität wie vorher
- Linearer Zusammenhang zwischen stetigen Prädiktoren und $Logit(P(Y)) = log \frac{p}{1-p}$
- Siehe Begleitmaterial im Moodle

Zusätzliche Probleme

- Unvollständige Informationen zu Gruppen
- Vollständige Abgrenzung
- Achtung: *R* und andere Statistikprogramme liefern trotzdem (falsche) Ergebnisse

Problem Unvollständiger Informationen zu Gruppen

Raucher/In	Isst Rosenkohl	Lungenkrebs
Ja	Nein	–
Ja	Ja	Ja
Nein	Ja	Ja
Nein	Nein	Nein

- Auch bei stetigen Variablen wichtig
- Jede Kombination sollte mindestens 1 Eintrag haben und 20% mehr als 5
- Hohe SE als Indikator nutzbar

Problem Vollständiger Abgrenzung

- Eine Kombination aus Prädiktoren sagt Y=1 perfekt voraus, grenzt $Y = 1$ vollständig von $Y = 0$ ab.

Beispiel:

- Druckplatte berechnet $P(Dieb)$ anhand von Gewicht
- Prädiktoren enthalten Diebe und andere Personen $\rightarrow$ OK
- Prädiktoren enthalten Diebe und Katzen $\rightarrow$ Schlecht weil Gewicht sagt $P(Dieb)$ perfekt voraus

Multinominale Logistische Regression

Berechnung

- Paarweise binäre logistische Regression
- Paare je nach Forschungssetup

  $\rightarrow$ A vs. B, B vs. C
  $\rightarrow$ A vs. C, A vs. B
  $\rightarrow$ ...
- Bspw. Abgrenzung gegen sinnvolle Baseline

Beispiel *Erfolg von Anmachsprüchen*

- Daten: Chat-Up Lines.dat
- zusammengefasstes Skript: multinominalRegression.R
- Genauere Beschreibung der Schritte im Moodle

Beispiel Multinominale Logistische Regression

Daten Studie zu Erfolg von Anmachsprüchen

- Outcomes: *keine Reaktion*, *Telefonnummer*, *gemeinsames Frühstück*
- Baseline des Outcomes: *keine Reaktion*
- Prädiktoren pro Spruch: *Lustig*, *Anzüglich*, *Zeigt Gutmütigkeit*, *Gender(EmpfängerIn)*

  Skalierung: *Lustig*, *Anzüglich*, *Zeigt Gutmütigkeit* je von 0, . . . , 10 (nicht, . . . , stark)
  Gender: 1 = female, 0 = male

(Spruch)	Lustig	Anzüglich	Zeigt Gutm.	Gender	Outcome
Na Schnitte, schon belegt?	7	3	1	1	Telefonnummer
. . .	. . .	. . .	. . .	. . .	. . .

Beispiel Multinominale Logistische Regression

Berechnung

- je Binäre Logistische Regression für *keine Reaktion* vs. *Telefonnummer* und *keine Reaktion* vs. *gemeinsames Frühstück*
- Siehe Begleitmaterial für schrittweise Anleitung via R

Beispiel Multinominale Logistische Regression

Ergebnis

	b(SE)	$p(>\|t\|)$	KI_U	Odds Ratio	KI_O
keine Reaktion vs. *Tel.nummer*					
Intercept	-1.78(0.67)	0.008**			
Gutmütigkeit	0.13(0.05)	0.014*	1.03	1.14	1.27
Lustig	0.14(0.11)	0.205	0.93	1.15	1.43
Female	-1.65(0.80)	0.039*	0.04	0.19	0.92
Anzüglich	0.28(0.09)	0.002**	1.11	1.32	1.57
Female x Lustig	0.49(0.14)	0.000***	1.24	1.64	2.15
Female x Anzüglich	-0.35(0.11)	0.004**	0.57	0.71	0.87
keine Reaktion vs. *Frühstück*					
Intercept	-4.29(0.94)	0.000***			
Gutmütigkeit	0.13(0.08)	0.12	0.97	1.14	1.34
Lustig	0.32(0.13)	0.011*	1.08	1.38	1.76
Female	-5.63(1.33)	0.000***	0.00	0.00	0.05
Anzüglich	0.42(0.12)	0.001***	1.20	1.52	1.93
Female x Lustig	1.17(0.20)	0.000***	2.19	3.32	4.77
Female x Anzüglich	-0.48(0.16)	0.004**	0.45	0.62	0.86

Signifikanzcodes: 0:***, 0.001:**, 0.01:*

Beispiel Multinominale Logistische Regression

Interpretation

	b(SE)	$p(>\|t\|)$	KI_U	Odds Ratio	KI_O
...	...	...	...	...	...
keine Reaktion vs. *Frühstück*					
Intercept	-4.29(0.94)	0.000***			
Gutmütigkeit	0.13(0.08)	0.12	0.97	1.14	1.34
Lustig	0.32(0.13)	0.011*	1.08	1.38	1.76
Female	-5.63(1.33)	0.000***	0.00	0.00	0.05
Anzüglich	0.42(0.12)	0.001***	1.20	1.52	1.93
Female x Lustig	1.17(0.20)	0.000***	2.19	3.32	4.77
Female x Anzüglich	-0.48(0.16)	0.004**	0.45	0.62	0.86

- *Lustig* hat signifikanten Einfluss ($p < 0.05$) auf den Unterschied zwischen *keine Reaktion* und *Frühstück*
- Die Chance erhöht sich pro Einheit *Lustig* um 1.38
- *Anzüglich* hat stärkeren Einfluss als *Lustig*
- *Gender=Female* hat negativen Einfluss auf den Unterschied zwischen *keine Reaktion* und *Frühstück*
- In Verbindung mit *Lustig* hat *Gender=Female* positiven Einfluss auf den Unterschied zwischen *keine Reaktion* und *Frühstück*, in Verbindung mit *Anzüglich* negativen

Zusammenfassung

- Logistische Regression:
  - $\rightarrow$ $\hat{Y}$ = kategorische Variable
  - $\rightarrow$ Gesucht: Wahrscheinlichkeit der Zugehörigkeit einer Gruppe $P(Y)$
  - $\rightarrow$ Binär vs. Multinominal
- Fitness
  - $\rightarrow$ Modell: Log-Likelihood, Abweichung, chi^2, (R^2, R), AIC, BIC
  - $\rightarrow$ Prädiktoren: Wald-Statistik
- Odds-Ratio für Modellvergleiche
  - $\rightarrow$ > 1 Chance erhöht sich mit Prädiktor
  - $\rightarrow$ < 1 Chance sinkt sich mit Prädiktor
  - $\rightarrow$ Konfidenzintervalle, die 1-Grenze überschreiten sind problematisch
- Auswahl der Prädiktoren
  - $\rightarrow$ Erzwungen, Schrittweise
- Annahmen & Probleme
  - $\rightarrow$ Wie bei Regression, Linearität des Logit
  - $\rightarrow$ Unvollständige Informationen zu Gruppen, Vollständige Abgrenzung

DH - Beispiele

- Robert Fuchs (2015): *Do Women Use More Intensifiers than Men? Recent Change in the Sociolinguistics of Intensifiers in British English*
- Intensifiers wie *totally*, *really*, *fucking*, *extremely*,...
- Interaktion zwischen Zeit (1994,2014) und je Alter, Geschlecht — Soziale Klasse, Geschlecht und Dialekt
- (Der Fokus auf Gender ist nicht so stark wie es der Titel vermuten lässt)

- Jean M. Twenge, W. Keith Campbell, Brittany Gentile (2012): *Increases in Individualistic Words and Phrases in American Books*
- Individualworte wie *self*, *unique*, *personalize*, *sole*,... gegen Kommunalworte wie *group*, *collective*, *everyone*, *teamwork*,...
- Entwicklung von 1960 bis 2008
- Jahresvorhersage per Regression