

Verallgemeinertes lineares Modell

Logistische Regression, Poisson Regression und negative Binomialregression¹

Bei einem OLS (*ordinary least squares* = kleinste Quadrate) Regressionsmodell wird angenommen, dass die Residuen sowohl normal verteilt als auch homoskedastisch sind. Für manche AVs gilt dies nicht, z.B. wenn die AV dichotom ist (Fall vs. Nicht-Fall, z.B. Prävalenzen) oder eine (diskrete) Zählvariable seltener Ereignisse darstellt (z.B. Inzidenzen). In diesen Fällen ist das OLS Regressionsmodell nicht effizient (die geschätzten Parameter sind verzerrt) und statistische Signifikanztests sind ungenau (die Standardfehler der Parameter werden unterschätzt).

Um abhängigen Variablen gerecht zu werden, deren Residuen die Annahmen eines OLS Regressionsmodells verletzen, wurde eine Klasse von statistischen Verfahren entwickelt, die als *verallgemeinertes lineares Modell* (*generalized linear model*) bezeichnet wird.

Merkmale der OLS Regression

Die OLS Regression ist durch drei Merkmale gekennzeichnet:

1. *Die algebraische Form des Modells*: Die OLS Regression ist ein lineares Modell, da es linear in den Parametern (oder in den Koeffizienten) ist, es besteht aus einer linearen Kombination der Prädiktoren (der Summe der Prädiktoren multipliziert mit ihren Regressionskoeffizienten).
2. *Die Fehlerstruktur bzw. die Verteilung der Residuen*: Die Residuen sind normal verteilt. Das bedeutet auch, dass der Mittelwert und die Varianz der Residuen unabhängig von einander sind. Diese Unabhängigkeit ist die Voraussetzung dafür, dass die Daten homoskedastisch sind, d.h., die bedingte Varianz der beobachteten Werte der abhängigen Variablen Y ist konstant für alle Werte der vorhergesagten Werte (der linearen Kombination der Prädiktoren) \hat{Y} .
3. *Die Einheit (Skalierung) der vorhergesagten Werte \hat{Y} in Relation zur Skala der beobachteten Werte der abhängigen Variablen Y* : Die vorhergesagten Werte \hat{Y} haben die gleiche Einheit (Skalierung) wie die beobachteten Werte Y .

Das verallgemeinerte lineare Modell

Das OLS Regressionsmodell ist ein Spezialfall des verallgemeinerten linearen Modells. Mit letzterem können Regressionsmodelle mit unterschiedlichen Arten von abhängigen Variablen formuliert werden. Wie in der OLS Regression können alle diese Modelle in einer Form ausgedrückt werden, die *linear in den Parametern* (oder Koeffizienten) ist. Dies gilt auch für Variablen, deren *Residuen nicht normal verteilt* und deshalb nicht homoskedastisch sind. Sie erlauben statt dessen, dass die Varianz der Residuen ($Y - \hat{Y}$) vom Wert der beobachteten Variablen Y abhängig ist. Zusätzlich kann die *Form (Skalierung) der vorhergesagten Werte \hat{Y} anders als die Form der beobachteten Werte Y* sein.

Drei häufig benutzte Fälle des verallgemeinerten linearen Modells sind die *logistische Regression* für dichotome, polychotome, oder ordinal skalierte abhängige Variablen sowie

¹ Angelehnt an Cohen et al. (2003, pp. 479-535)

die *Poissonregression* und die *negative Binomialregression* für positive Zählvariablen seltener Ereignisse.

Die Varianten des verallgemeinerten linearen Modells sind durch zwei Merkmale gekennzeichnet: Die *Varianzfunktion* und die *Verknüpfungsfunktion (link function)*:

a) *Varianzfunktion*

Die große Flexibilität des verallgemeinerten linearen Modells hängt damit zusammen, dass die Annahme normal verteilter Residuen zur *Familie von exponentialer Wahrscheinlichkeitsverteilungen* erweitert wird. Ist z.B. die AV dichotom (logistische Regression), folgen die Residuen einer Binomialverteilung. Besteht sie dagegen aus einer Zählung von Ereignissen in einem bestimmten Zeitintervall (Poisson- oder negative Binomialregression), folgen die Residuen einer Poissonverteilung (oder negativen Binomialverteilung). Diese und andere Mitglieder der exponentiellen Familie (z.B. Gamma oder inverse Gaussverteilung) haben allgemein die Eigenschaft, dass Mittelwert und Varianz der Werte nicht unabhängig sind bzw. dass die Varianz vom Mittelwert abhängt. Die Gauss- oder Normalverteilung ist ein Spezialfall der exponentiellen Familie, hier sind Mittelwert und Varianz von einander unabhängig.

Sind Mittelwert und Varianz abhängig voneinander (was bedeutet, dass die Residuen nicht homoskedastisch sind), benötigt man ein Modell der Varianzfunktion, mittels dessen spezifiziert wird, wie die bedingte Wahrscheinlichkeit der beobachteten Werte Y als eine Funktion von \hat{Y} variiert. Z.B. wird bei der Poissonregression angenommen, dass die Residuen für jeden Wert $\hat{\mu}$ (der vorhergesagten Ereignisrate) Poisson verteilt sind, wobei die Varianz ebenfalls gleich $\hat{\mu}$ ist.

b) *Verknüpfungsfunktion*

Im verallgemeinerten linearen Modell ist der Zusammenhang zwischen den beobachteten Werten Y und den vorhergesagten Werten linear in den \hat{Y} Koeffizienten. Während im OLS Regressionsmodell die beobachteten Werte Y und die vorhergesagten Werte \hat{Y} die gleichen Einheiten (Skalierung) haben, ist dies beim verallgemeinerten linearen Modell nicht notwendig der Fall. Z.B. sind die vorhergesagten Werte im logistischen Regressionsmodell Logits (logarithmierte Odds-Ratios) und bei der Poisson- oder negativen Binomialregression logarithmierte Anzahlen. Die *Verknüpfungsfunktion* des verallgemeinerten linearen Modells ist die Transformation, die die vorhergesagten Werte mit den Werten der beobachteten AV verknüpft. Die OLS Regression stellt hier einen Spezialfall dar: Hier ist die Funktion eine Identitätsfunktion, da die vorhergesagten und beobachteten Werte die gleiche Einheit (Skalierung) haben.

Regressionsmodelle, die linear in ihren Koeffizienten sind, deren Residuen einer Varianzfunktion der exponentiellen Familie folgen und in denen eine der verschiedenen Verknüpfungsfunktionen zur Linearisierung des Zusammenhangs der vorhergesagten und beobachteten Werte benutzt werden, sind Mitglieder des verallgemeinerten linearen Modells.

Tabelle 1 zeigt unterschiedliche Regressionsmodelle des verallgemeinerten linearen Modells mit ihren jeweiligen Varianzfunktionen (*family*) sowie Verknüpfungsfunktionen (*link function*).

Tabelle 1: Ausgewählte Regressionsmodelle des verallgemeinerten allgemeinen Modells und mögliche Verknüpfungsfunktionen

Regression	Varianzfunktion			Verknüpfungsfunktion		
	Familie	Bereich der AV	bedingte Varianz der AV	Link	Formel $\eta_i = g(\mu_i)$	Inverse $\mu_i = g^{-1}(\eta_i)$
OLS Regression	Gauss	$[-\infty; +\infty]$	ϕ	Identität	μ_i	η_i
logistische Regression	binomial	$\frac{0,1,2,\dots,n_i}{n_i}$	$\mu_i \cdot (1 - \mu_i)$	Logit	$\ln\left(\frac{\mu_i}{1 - \mu_i}\right)$	$\frac{1}{1 + e^{-\eta_i}}$
				Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Poisson Regression	Poisson	0,1,2,...	μ_i	Log	$\ln(\mu_i)$	e^{η_i}
negative Binomialregression	negativ binomial	0,1,2,...	$\mu_i + \phi \cdot \mu_i^2$	Log	$\ln(\mu_i)$	e^{η_i}
Gamma Regression	Gamma	$[0; +\infty]$	$\phi \cdot \mu_i^2$	Inverse	$\frac{1}{\mu_i}$	$\frac{1}{\eta_i}$

Anmerkung: μ_i ist der Erwartungswert von y_i , η_i der lineare Prädiktor, ϕ ist der Verteilungsparameter

Logistische Regression

[wird ergänzt]

Poissonregression

[wird ergänzt]

Negative Binomialregression

[wird ergänzt]

Literatur

Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ (3rd ed.): Lawrence Erlbaum.