Logistische Regression

Xuan Son Le (4669361), Freie Universität Berlin

02/04/2018

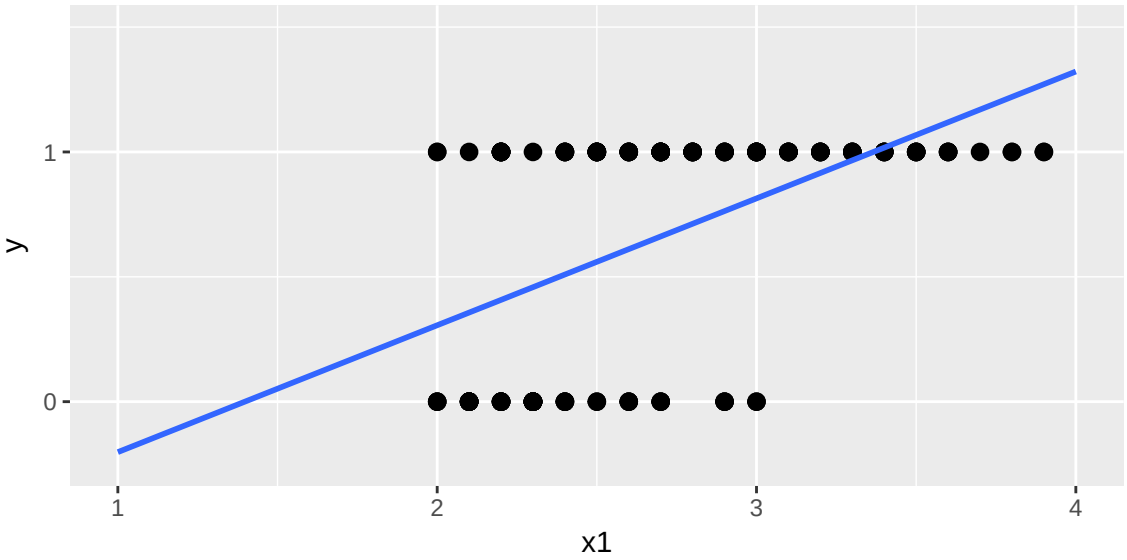
Abstract: Im Rahmen der Abschlussarbeit des Moduls Programmieren mit R im Wintersemester 2017/2018 an der Freie Universität Berlin wird für diese Arbeit die statistische Methode namens binäres Logit-Modell ausgewählt. Diese Arbeit besteht aus zwei großen Hauptteilen: der Theorieteil, wobei die ausgewählte Methode theoretisch vorgestellt wird und der Implementierungsteil, welcher die Erklärung der Funktionalität vom selbst entwickelten Paket beinhaltet. Im Theorieteil wird zunächst ein Überblick über die grundliegende Funktionsweise vom (binären) Logit-Modell widergegeben. Die Grundidee von Generalisierten linearen Modellen wird anschließend kurz eingeführt, bevor der Aufbau vom binären Logit-Modell durch das Maximum Likelihood Verfahren vorgenommen wird. Demzufolge folgt die Interpretation der Koeffizienten und der Lösungsgüte vom binären Logit-Modell. Schließlich werden im Implementierungsteil alle Funktionen vom R-Paket schritterweise vorgestellt.

Keywords: *Logit-Modell, logistische Regression, Paket, R*

---

1 Motivation

Die Anwendung von der klassischen linearen Regression ist für binäre (binomiale oder dichotome) Zielvariable, welche lediglich zwei Werte (ja/nein, männlich/weiblich, erfolgreich/nicht erfolgreich, etc.) annehmen kann, nicht mehr geeignet, da die Zielvariable von der linearen Regression metrisch skaliert ist. Oft wird binäre Variable als 0/1-Variable kodiert, das heißt sie nimmt nur den Wert 0 oder 1 an. Die folgende Grafik stellt den Ansatz graphisch dar, binäre Variable durch lineare Regression zu modellieren:



Graphisch lässt sich festlegen, dass die lineare Regression den Wertebereich [0,1] von binären Responsevariablen sehr schnell verlässt. Wenn die Annahmen von der linearen Regression noch in Betracht gezogen werden, ergeben sich darüber hinaus noch folgende Probleme (vgl. . . .): *

Aus diesen Gründen wird ein ganz anderer Ansatz benötigt, um binäre Zielvariable zu modellieren, nämlich das binäre Logit-Modell, welches ebenfalls als binäre logistische Regression oder binäres logistisches Regressionsmodell bezeichnet werden kann. In der Statistik lassen sich Logit-Modelle noch in multinomiale und kumulative Logit-Modelle aufteilen, je nachdem ob die abhängige Variable multinominal- oder ordinalskaliert sind . Diese Arbeit beschäftigt sich mit dem binären Logit-Modell, welches den Zusammenhang zwischen einer binären abhängigen Variable und einer/mehreren unabhängigen Variablen untersucht. Bei allen Arten von Logit-Modellen können die unabhängigen Variablen beliebig skaliert sein.

Im Unterschied zu der klassischen linearen Regression, welche den wahren Wert einer Zielvariable vorhersagt, interessiert sich das binäre Logit-Modell eher für die Wahrscheinlichkeit, dass die Zielvariable den Wert 1 annimmt. Das Hauptziel vom binären Logit-Modell ist es, die Wahrscheinlichkeit für den Eintritt der Zielvariable vorherzusagen. Dadurch soll die folgende theoretische Fragestellung beantwortet werden: *Wie stark ist der Einfluss von den unabhängigen (erklärenden) Variablen auf die Wahrscheinlichkeit, dass die abhängige (zu erklärende / Response) Variable eintritt beziehungsweise den Wert 1 annimmt?* In der Praxis kann diese Fragestellung beispielsweise so formuliert werden: “Haben Alter, Geschlecht,

Berufe oder andere Merkmale der Kunden Einfluss auf die Wahrscheinlichkeit, dass sie ein Kredit rechtzeitig zurückzahlen?" oder "Lässt sich die Wahrscheinlichkeit, dass es regnet, durch die Temparatur, die Windstärke oder Sonnenstrahlungsintensität vorhersagen?".

2 Das binäre Logit-Modell

2.1 Modellspezifikation

Das Logit-Modell ist eine Methode aus der Algorithmenklasse namens *Generalisierte Lineare Modelle* (engl. generalized linear model, kurz GLM), welche eine Verallgemeinerung des klassischen linearen Regressionsmodells anstrebt. Dazu gehören noch die klassische lineare Regression, Probitmodell und Poisson-Regression. Die Grundidee von GLM ist die Transformation der linearen Regressionsgleichung, so dass der Wertebereich der vorhergesagten Zielvariable dem gewünschten entspricht. Ein GLM besteht aus drei Hauptelementen: die systematische Komponente, die Link-Funktion und die Zufallskomponente.

Gegeben seien n unabhängige Beobachtungen $y_1, y_2, ..., y_n$ der binären Zielvariable $\mathbf{Y}$. Ein Verteilungsmodell für $\mathbf{Y}$ ist die Binomialverteilung: $\mathbf{Y}_i \sim B(1, \pi_i)$ mit $\pi_i = P(Y_i = 1)$

Für diese Arbeit wird $\pi_i = (\pi_1, \pi_2, ..., \pi_n)$ als die Eintrittwahrscheinlichkeit von der einzelnen $\mathbf{Y}_i$ benannt. Weiterhin seien p erklärende Variablen $\mathbf{X}_0, \mathbf{X}_1, .., \mathbf{X}_k$ gegeben mit jeweils n unabhängigen Beobachtungen $\mathbf{X}_j = (x_{1j}, x_{2j}, ..., x_{nj})$ mit j $\in$ {1,2,..,k} gegeben. Daraus ergeben sich p Koeffizienten $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$, welche die Stärke den Zusammenhang zwischen die einzelne erklärende Variable mit der Zielvariable widerspiegeln. Dabei ist es sinnvoll, diese in einer Designmatrix $\mathbf{X}$ zu speichern. Da der Interzept (β_0) ebenfalls geschätzt werden soll, sind alle Werte der ersten Spalte von X gleich Eins, also $x_{10} = x_{20} = ... = x_{n0} = 1$. Zusammengefasst lässt sich die Designmatrix wie folgt darstellen:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

Die dazugehörige lineare Regressionsgleichung lautet: $\mathbf{Y} = \mathbf{X}.\beta + \epsilon$ mit $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3, ..., \epsilon_n)$ als die Abweichung der einzelnen Schätzungen gegenüber dem wahren Wert. Die einzelne Beobachtung lässt sich wie folgt darstellen:

$$y_i = \beta_0 + \beta_1.x_{i2} + \beta_2.x_{i3} + ... + \beta_k.x_{ik} + \epsilon_i \qquad \forall_i = 1, 2, 3, ..., n$$

Um die Werte im Bereich der reellen Zahlen von der linearen Regression auf dem Wertebereich von Wahrscheinlichkeiten zwischen 0 und 1 zu beschränken, sollte die rechte Seite der Gleichung transformiert werden. Das Ziel ist es, eine sinnvolle Verteilungsfunktion (Responsefunktion) zu finden, deren Wertebereich in [0,1] liegt: $\pi_i = F(\beta_0 + \beta_1.x_{i2} + \beta_2.x_{i3} + ... + \beta_k.x_{ik} + \epsilon_i) = F(\eta_i)$.

Der lineare Prädikator $\eta_i = \beta_0 + \beta_1.x_{i2} + \beta_2.x_{i3} + ... + \beta_k.x_{ik} + \epsilon_i$ wird ebenfalls als Linkfunktion genannt, weil dadurch eine Verbindung (Link) zwischen der Eintrittwahrscheinlichkeit und den unabhängigen Variablen erfolgt wird. Für das binäre Logit-Modell wird anstelle der Responsefunktion die standardisierte logistische Verteilung verwendet:

$$F(\eta_i) = Logist(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

Da durch die Responsefunktion die Eintrittwahrscheinlichkeit π_i modelliert werden soll, ergibt sich die Gleichung für das binäre Logit-Modell wie folgt:

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \frac{\exp(\beta_0 + \beta_1.x_{i2} + \beta_2.x_{i3} + ... + \beta_k.x_{ik} + \epsilon_i)}{1 + \exp(\beta_0 + \beta_1.x_{i2} + \beta_2.x_{i3} + ... + \beta_k.x_{ik} + \epsilon_i)}$$

Dabei kann π_i maximal den Wert 1 nehmen, wenn $\exp(\eta_i)$ sehr groß ist und minimal den Wert 0, wenn $\exp(\eta_i)$ sehr nah rechts von 0 liegt. $\exp(\eta_i)$ kann nicht negativ sein. Diese Gleichung erfüllt somit die Anforderung bezüglich dem Wertebereich von Wahrscheinlichkeiten.

Wird die Gleichung nach dem linearen Prädikator η_i gelöst, in dem beide Seite der Gleichung mit $(1 + \exp(\eta_i))$ multipliziert wird, ergibt sich:

$$\pi_i.(1 + \exp(\eta_i)) = \exp(\eta_i)$$

Daraus folgt:

$$\pi_i$$

2.2 Maximum Likelihood Schätzung

Während bei der klassischen linearen Regression die Methode der Kleinsten Quadrate (engl. *method of least squares*) genutzt wird, um eine Regressionslinie zu bestimmen, welche die Summe der quadratischen Abweichungen minimiert, wird bei dem binären Logit-Modell die sogenannte Maximum Likelihood Schätzung eingesetzt.

2.3 Intepretation der Koeffizienten

3 Implementierung in R

Im Folgenden wird die Funktionalität von dem Paket **logitModell** erklärt, welches zum Ziel setzt, die Grundidee hinter dem binären Logit-Modell programmiert darzustellen. Das Paket besteht aus dem R-Code, welcher anhand dem manuell berechneten Maximum Likelihood ein Objekt von der Klasse *logitMod* erstellt und anschließend drei S3 Methoden für diese Klasse (print, summary und plot) definiert, und einer Vignette, welche den R-Code anhand einem konkreten Beispiel ausführt. Dieser Beispieldatensatz wird im Folgenden verwendet, um die Richtigkeit und Vollständigkeit der Ergebnisse der implementierten Methode im Vergleich

zu der R-Standardmethode für Logit-Modell *glm(. . . , family = "binomial")* zu testen. Die binäre Responsevariable heißt *admit*, welche besagt ob ein Kandidat eine Zulassung bekommt. Zudem enthält der Datensatz drei unabhängige Variablen: *gre*, *gpa* (metrisch) und *rank* (kategorial). Der Datensatz soll ein Modell unterstützen, welche die Abhängigkeit von der Wahrscheinlichkeit einer Zulassung von der Abschlussnote, GRE-Note sowie dem Ruf von der angestrebten Institution.

3.1 Beispieldatensatz

Zunächst wird der Datensatz importiert. Dabei wird die Zielvariable aus dem Datensatz entnommen und in einem Vektor gespeichert. Da diese schon als 0/1-Variable vorgegeben wird, besteht es in diesem Fall keine Notwendigkeit, die Zielvariable zu faktorisieren. Der Code funktioniert allerdings ebenfalls mit Zielvariable, welche zum Beispiel als weiblich/männlich oder Erfolg/kein Erfolg kodiert wird und transformiert diese in eine 0/1-Variable.

```
# sei y die eingegebene Zielvariable
if (!(0 %in% y && 1 %in% y)) {
    y <- factor(y, labels = c(0,1))
}
y <- as.numeric(as.character(y))
```

Es muss immer vorab überprüft werden, in welcher Art die Zielvariable eingegeben wird, denn das Maximum Likelihood braucht als Input numerische Vektoren für weitere Berechnungen. Dieser Schritt wird extra gemacht, damit sich das manuelle Modell im Hinblick auf den Input gleich verhält wie das Standardmodell.

3.2 Maximum Likelihood Schätzung

Bevor das eigentliche Logit-Modell erstellt wird, wird in diesem Abschnitt die Implementierung der Maximum Likelihood Schätzung auseinandergesetzt. Der Code dazu ist auf Basis von dem betroffenen theoretischen Teil (siehe Abschnitt . . .) aufgebaut. Schrittweise werden die einzelnen Parameter definiert. Daraus wird in der Newton-Raphson-Schleife das Maximum Likelihood berechnet.

Das gerade ausgeführte Beispiel kann direkt in R geladen werden. Dafür wird in das Paket ein Vignette eingebaut, so dass wenn den folgenden Code ausgeführt wird, wird das Beispiel in der Help-Seite von R angezeigt.

```
setwd("~/Desktop/Uni/Master/WS1718/ProgR/Abschlussarbeit/logisticRegression/Code/logitMo
devtools::install(build_vignettes = TRUE)
vignette("logitModell")
```

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$