Logistische Regression

Xuan Son Le (4669361), Freie Universität Berlin

02/04/2018

Contents

1 Motivation **2**

2 Das binäre Logit-Modell **3**
2.1 Modellspezifikation 3
2.2 Maximum Likelihood Schätzung 3
2.3 Intepretation der Koeffizienten 3

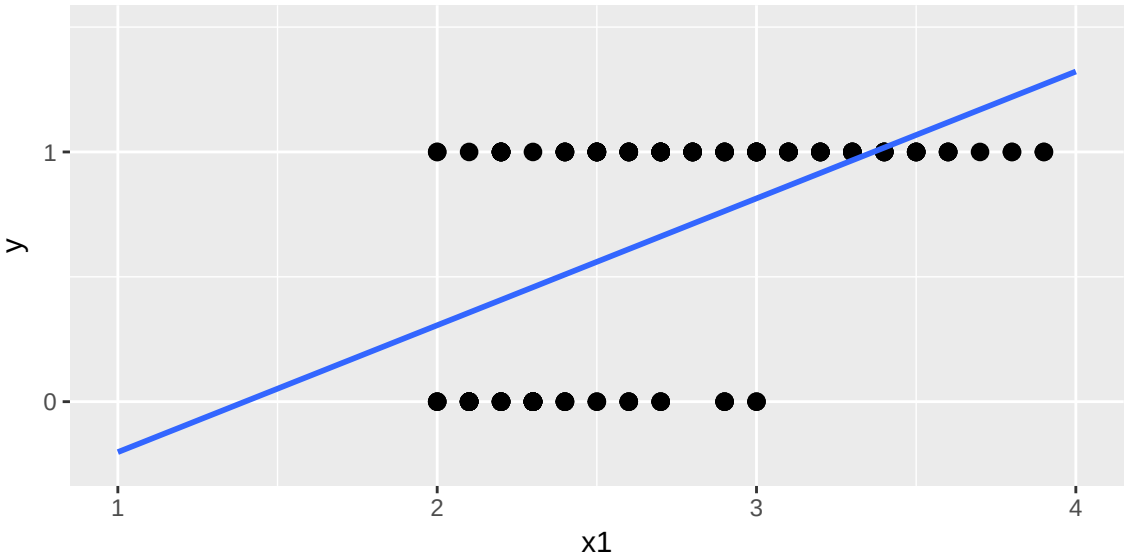
3 Implementierung in R **3**

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:

-
-
-

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*, welche eine Verallgemeinerung des klassischen linearen Regressionsmodells anstrebt. Dazu gehören noch die klassische lineare Regression, Probitmodell und Poisson-Regression.

2.2 Maximum Likelihood Schätzung

Während bei der klassischen linearen Regression die Methode der Kleinsten Quadrate (engl. *method of least squares*) genutzt wird, um die "beste" Regressionslinie zu bestimmen, findet

2.3 Intepretation der Koeffizienten

3 Implementierung in R

Im Folgenden wird die Funktionalität von dem Package ... erklärt, welches zum Ziel setzt, die Grundidee hinter dem binären Logit-Modell programmiert darzustellen.

Ein Beispieldatensatz wird verwendet, um die Richtigkeit und Vollständigkeit der Ergebnisse der implementierten Methode im Vergleich zu der R-Standardmethode für Logit-Modell 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 der Ruf von der angestrebten Institution.

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")
```