

Logistische Regression

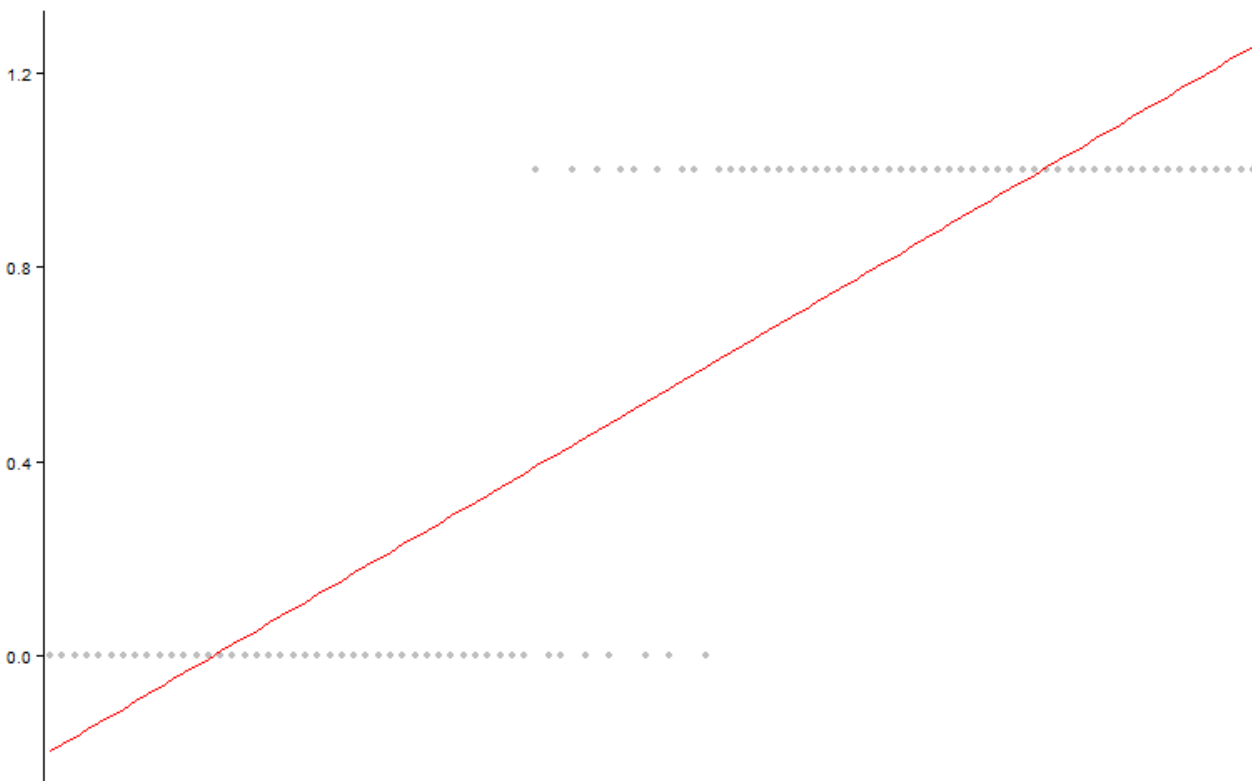
Benjamin Schlegel

11. April 2016

Die logistische Regression kann verwendet werden, wenn die abhängige Variable dichotom ist. Dichotom bedeutet, dass die Variable nur zwei Ausprägungen hat: 0 und 1 (z.B. Mann/Frau; ja/nein). Eine dichotome Variable unterliegt der [Bernoulli-Verteilung](#).

Warum kein lineares Modell?

Würde man eine dichotome Variable mit einem [linearen Modell](#) schätzen, gäbe es mehrere Probleme. Zum einen sind die Fehler nicht [normalverteilt](#) und damit eine der Annahmen der linearen Regression verletzt, weshalb der Schätzer nicht effizient ist. Ein weiteres Problem ist, dass die lineare Regression Werte zurückgibt, die kleiner als 0 oder grösser als 1 sind.



Eine lineare Regression kann also nicht verwendet werden, um eine adäquate Regression für eine dichotome abhängige Variable zu rechnen. Deshalb wird eine Funktion um die lineare Regression gerechnet, die sogenannte Linkfunktion. Diese Art von Regressionen werden generalisierte lineare Modelle genannt. Bei einer dichotomen Variable muss eine Linkfunktion verwendet werden, welche immer Werte zwischen 0 und 1 liefert. Eine dieser Funktionen ist die kumulierte logistische Verteilung. Wird diese verwendet wird von einer logistischen Regression oder Logit geredet. Eine andere Funktion ist die kumulierte Standardnormal-Verteilung. In diesem Fall wird von einer Probit Regression gesprochen. Der einzige Unterschied dieser beiden Funktionen liegt in der Annahme über die Verteilung der Fehler. Beide Regressionen liefern jedoch mehr oder weniger die gleichen Resultate.

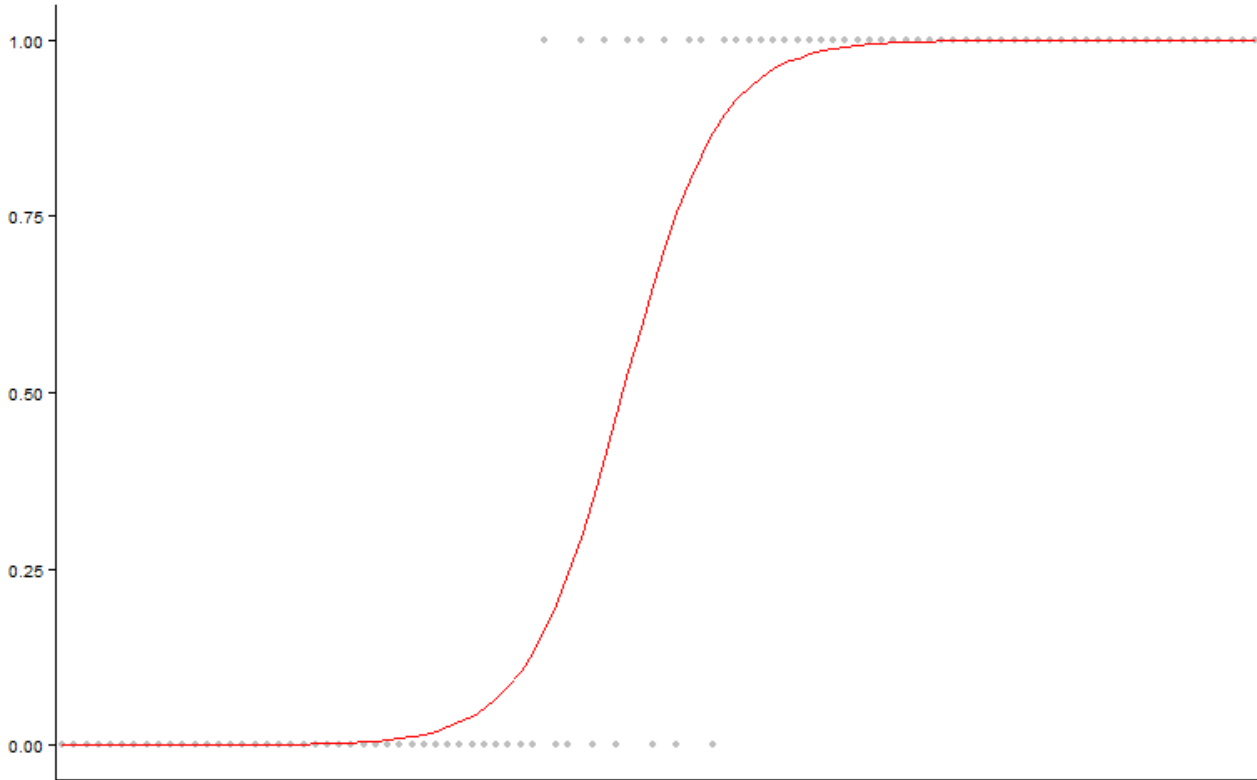
Logit

Auf die Probit Regression geht dieser Artikel nicht weiter ein. Die Linkfunktion für die logistische Regression ist, wie

bereits erwähnt, die kumulierte logistische Verteilung:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Da die logistische Regression kein lineares Modell sondern ein generalisiertes lineares Modell enthält, ist die Regressionslinie keine Gerade mehr.



Anders als bei einem linearen Modell können die Koeffizienten bei einem generalisierten linearen Modell nicht direkt interpretiert werden. Am einfachsten ist die Interpretation von vorausgesagten Wahrscheinlichkeiten und diskreten Änderungen. Bei der logistischen Regression können zudem odds ratio berechnet werden (geht nicht bei Probit).

Weiterführende Literatur

Best, Henning und Christof Wolf (2010): Logistische Regression. In: Christof Wolf und Henning Best (Hrsg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlage. (Deutsch)