3 Regressionsmodelle für Zähldaten

Aufgabe 20

Ein Bernoulli-Experiment mit Erfolgswahrscheinlichkeit π wird so oft wiederholt, bis sich r Erfolge einstellen. Die Experimente werden dabei unabhängig voneinander durchgeführt. Die Anzahl der hierfür notwendigen Versuche werde mit X bezeichnet.

(a) Wie ist X verteilt? Geben Sie die Wahrscheinlichkeitsfunktion von X an.

(b) Welche Verteilung ergibt sich für den Spezialfall $r = 1$?

Gegeben sei $Y|\lambda \sim Po(\lambda)$ und $\lambda \sim Ga(a, b)$.

(c) Berechnen Sie die Wahrscheinlichkeitsfunktion von Y und benennen Sie die Verteilung.

(d) Leiten Sie die Verteilung von $\lambda|Y$ her.

Aufgabe 21

Der Radweg an der Ludwigstraße vor dem Institut für Statistik ist berüchtigt dafür, dass dort viele Radfahrer entgegen der vorgeschriebenen Fahrtrichtung fahren. Um diesen Sachverhalt zu analysieren, beobachtete ein Statistiker 24 Mal diesen Radweg für eine bestimmte, jeweils unterschiedliche Zeit und notierte sich die Anzahl der Falschfahrer während dieser Zeit. Er vermerkte dabei auch, ob die jeweilige Beobachtung während der Vorlesungszeit oder in der vorlesungsfreien Zeit gemacht wurde.

Der Datensatz `Bike` (siehe Homepage) enthält folgende Variablen:

Variable	Beschreibung
`lecture`	Vorlesungszeit ja (`1`) / nein (`0`)
`time`	Beobachtungsdauer (in Minuten)
`y`	Anzahl der Falschfahrer

(a) Erläutern Sie anhand dieses Beispiels die Bedeutung eines Offsets.

(b) Berechnen Sie ein Poisson- sowie ein Quasi-Poisson-Modell, jeweils mit Offset und vergleichen Sie die Modell-Outputs.

Aufgabe 22

Im `R`-Paket `pscl` steht der Datensatz *bioChemists* zur Verfügung. Im Folgenden soll untersucht werden, welche Merkmale die Anzahl der von Ph.D.-Studenten geschriebenen Artikel beeinflussen.

Der Datensatz enthält folgende Variablen:

Variable	Beschreibung
`art`	Anzahl der geschriebenen Artikel
`fem`	Geschlecht
`mar`	Familienstand
`kid5`	Anzahl an Kindern jünger als 6 Jahre
`phd`	Ansehen des Instituts
`ment`	Anzahl der geschriebenen Artikel des Betreuers

(a) Untersuchen Sie obige Fragestellung zunächst paarweise mittels deskriptiver Analysen.

(b) Für die Untersuchung mithilfe von Regressionsmodellen nehmen Sie als Verteilung für den Response eine Poisson-Verteilung an. Wann sollten Ihnen Zweifel an der damit einhergehenden Äquidispersionseigenschaft kommen?

(c) Fitten Sie ein Poisson- sowie ein Quasi-Poisson-Modell mittels der Funktion `glm()`.

(d) Fitten Sie ein Negativbinomial-Modell mittels der Funktion `glm.nb()`!

(e) Erläutern Sie kurz das Prinzip einer Zero-Inflated Poisson-Regression und geben Sie (allgemein) die zugehörige Wahrscheinlichkeitsverteilung an.

(f) Fitten Sie Zero-Inflated Poissonmodelle mittels der Funktion `zeroinfl()` aus dem Paket `pscl`. Variieren Sie dabei, indem Sie für die Verteilung in der Responder-Population Poissonverteilung bzw. Negativbinomialverteilung verwenden und zur Modellierung, ob jemand zur Responder-Population gehört, ein Logit- bzw. Probit-Modell annehmen.

(g) Erläutern Sie kurz das Prinzip des Hurdle-Modells und geben Sie (allgemein) die zugehörige Wahrscheinlichkeitsverteilung an.

(h) Fitten Sie Hurdle-Modelle mittels der Funktion `hurdle()` aus dem Paket `pscl`. Variieren Sie dabei die angenommenen Verteilungen, indem Sie für f_1 (hurdle part) Poissonverteilung bzw. Binomialverteilung (Logit-Link) verwenden und für f_2 (parent process) Poissonverteilung bzw. Negativbinomialverteilung annehmen.

(i) Diskutieren Sie Möglichkeiten des Modellvergleichs. Für welches der betrachteten Modelle entscheiden Sie sich?