Regressionsmodelle für Zähldaten in SAS

Ralf Minkenberg
Boehringer Ingelheim Pharma GmbH & Co. KG
Binger Str. 173
55216 Ingelheim
ralf.minkenberg@boehringer-ingelheim.com

Zusammenfassung

Die Analyse von Häufigkeiten oder ähnlichen Zähldaten ist ein in vielen Bereichen häufig anzutreffendes Problem, für das auch innerhalb von SAS vielfältige Modelle angeboten werden. Im folgenden werden die verschiedenen Modelle anhand von medizinischen Beispieldaten vorgestellt und miteinander verglichen. Sowohl verschiedene Möglichkeiten der Modellwahl als auch statistische Tests zur Unterstützung werden vorgestellt. Die verschiedenen Regressionsmodelle sind in verschiedenen SAS-Prozeduren aufrufbar und die entsprechende Syntax sowie die zugehörige Ausgabe werden erläutert.

Schlüsselwörter: Regression, Zähldaten, Poisson-Verteilung, Over-Dispersion

1 Zähldaten

In vielen Anwendungsbereichen werden wiederkehrende Ereignisse beobachtet, deren Abhängigkeit von verschiedenen Faktoren modelliert werden soll. Zunächst ist hierbei nur die Zeit bis zum ersten Auftreten dieses Ereignisses von Interesse. Mit Hilfe verschiedener Ansätze innerhalb der Überlebenszeitanalyse lässt sich diese Fragestellung beantworten. Ist jedoch nicht die Zeit bis zum ersten Auftreten eines Ereignisses von Interesse und reicht es auch nicht aus, nur das Auftreten oder Nicht-Auftreten eines Ereignisses (unabhängig von dessen Häufigkeit) zu analysieren, können Modelle zur Analyse von Zähldaten zur Anwendung kommen. Bei diesen Modellen wird die Anzahl, mit der ein bestimmtes Ereignis auftritt, in Abhängigkeit von einem oder mehreren Faktoren analysiert. Es werden Regressionsmodelle angewandt, bei denen die Häufigkeit des Ereignisses als abhängige Größe ins Modell eingeht.
Beispiele für Zähldaten lassen sich in vielen Bereichen finden. Beispielsweise kann in der Medizin die Abhängigkeit der Anzahl Krankenhausaufenthalte eines beobachteten Patientenkollektivs in Abhängigkeit von demographischen, sozio-ökonomischen und medizinischen Merkmalen betrachtet werden. Auch die Analyse der Anzahl bestimmter unerwünschter Ereignisse in einer klinischen Studie in Abhängigkeit von der Behandlung und weiterer Kovariablen ist denkbar. Im Versicherungswesen kann zum Beispiel versucht werden, die Anzahl an Schadensfällen anhand weiterer Variablen mit Hilfe von Zähldatenmodellen zu erklären. Ähnliche Analysen sind im Bankwesen beispielsweise bei der Anzahl an nicht zurückgezahlten Kreditraten verwendbar.

Die folgenden Modellanpassungen und -vergleiche basieren auf einem Beispiel aus einer mehrjährigen klinischen Studie, in der kardiovaskuläre Risikopatienten beobachtet und präventiv behandelt wurden. Die Gesamtanzahl an Krankenhausaufenthalten innerhalb der ca. fünfjährigen Beobachtungszeit wird im folgenden in Abhängigkeit von der Behandlung („Verum“ oder „Placebo“) betrachtet. Eine Verallgemeinerung auf mehrere Einflussgrößen ist problemlos möglich. In der Studie ergab sich eine mittlere Zahl an Krankenhausaufenthalten von 1,5 (SD: 2,5) bei Placebo-Patienten und von 0,8 (SD: 2,1) bei Verum-Patienten. Die kumulativen Häufigkeiten sind in Abbildung 1 dargestellt.

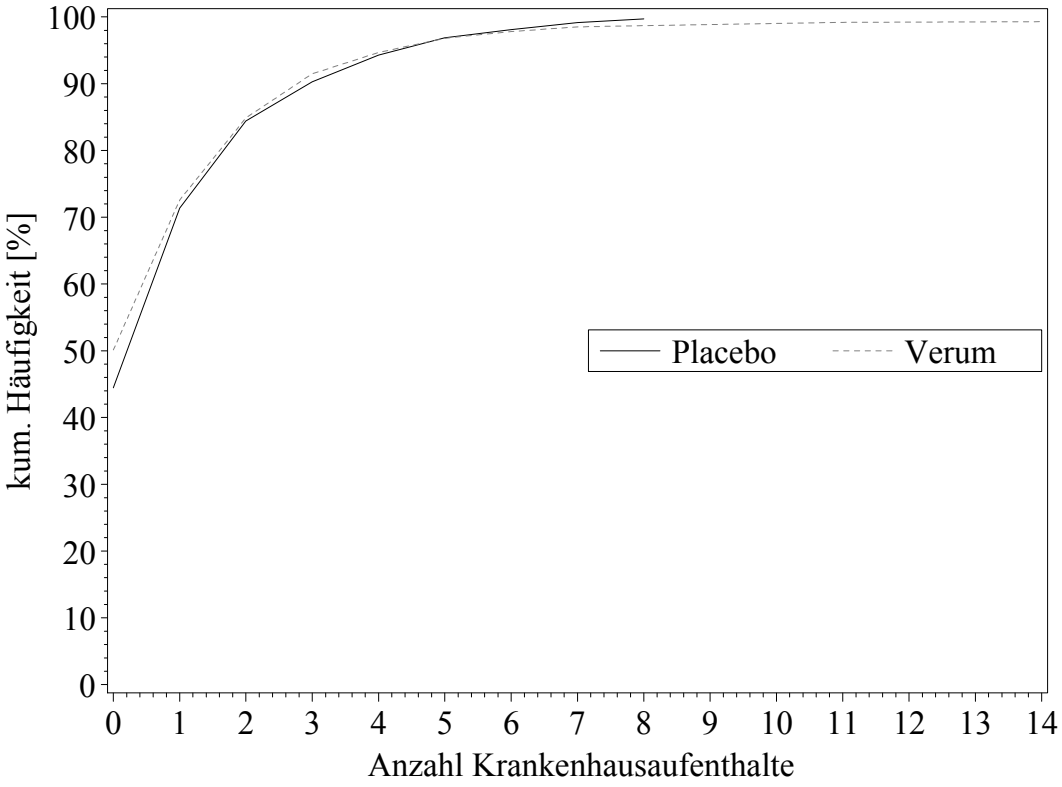

Abbildung 1: Kumulative Häufigkeiten der Anzahl Krankenhausaufenthalte

2 Poisson-Regression: ein einfaches Standardmodell

Bei der Analyse von Zähldaten findet zunächst meistens ein Modell auf Basis der Poisson-Verteilung Anwendung. Es wird dabei angenommen, dass jede beobachtete Häufigkeit y_i einer Poisson-Verteilung mit bedingtem Mittelwert μ_i abhängig von Werten X_i für eine Beobachtung i entstammt. Dies bedeutet also:

$$f\left(Y_i = y_i \mid X_i\right) = \frac{e^{-\mu_i}\,\mu_i^{y_i}}{y_i!}$$

Für die Poisson-Verteilung, oft auch als Verteilung der seltenen Ereignisse genannt, gilt, dass Mittelwert und Varianz gleich sind: $\mu_i = \mathrm{E}(y_i) = \mathrm{Var}(y_i)$.

Eine graphische Darstellung verschiedener Poisson-Verteilungen ist in Abbildung 2 zu finden.

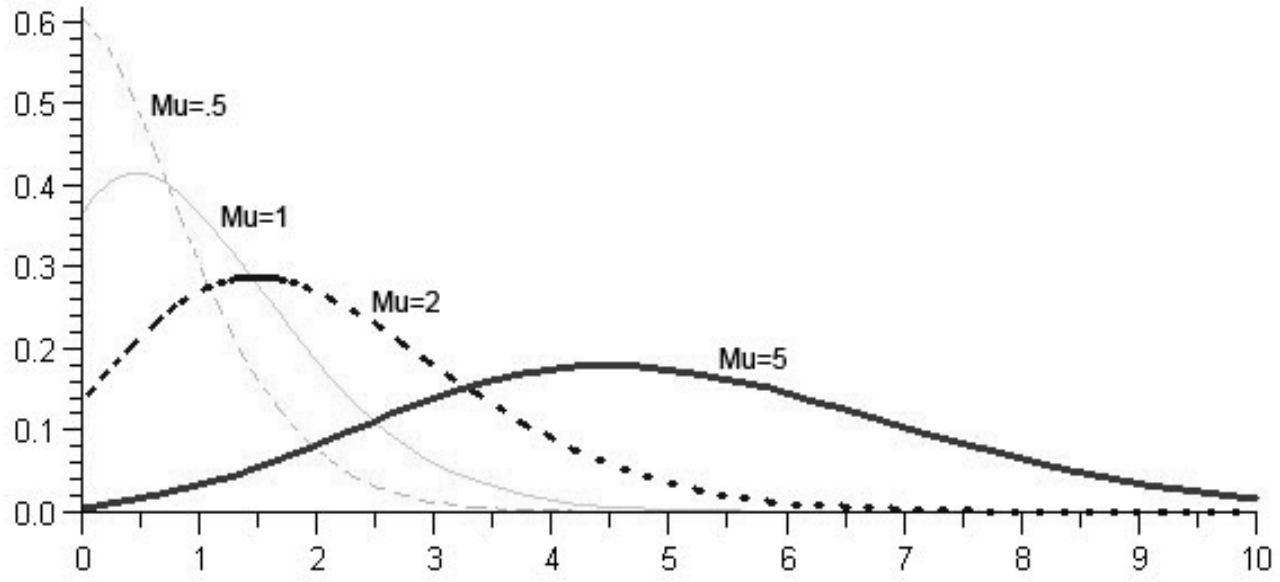

Abbildung 2: Verschiedene Poisson-Verteilungen

Für die beobachteten Häufigkeiten wird nun eine Regressionsanalyse unter Annahme einer Poisson-Verteilung für die abhängige Variable durchgeführt.

In SAS stehen verschiedene Prozeduren für eine solche Regressionsanalyse zur Verfügung. Mit PROC GENMOD sieht ein entsprechender Aufruf folgendermaßen aus:

```
proc genmod data=daten;
  class treat;
  model count = treat / dist=poisson link=log;
  output  out = poi_out  predicted = p;
run;
```

Der entsprechende Output dieses Programms liefert unter anderem folgende Ergebnisse:

```
Criteria For Assessing Goodness Of Fit
Criterion                     DF           Value          Value/DF

Deviance                    5924      12796.0838          2.1600
Scaled Deviance             5924      12796.0838          2.1600
Pearson Chi-Square          5924      16424.3916          2.7725
Scaled Pearson X2           5924      16424.3916          2.7725
Log Likelihood                        -5867.9192
Full Log Likelihood                  -10169.7006
AIC (smaller is better)               20343.4012
AICC (smaller is better)              20343.4033
BIC (smaller is better)               20356.7754

Algorithm converged.
```

Die beiden fett hervorgehobenen Werte bei „Deviance“ und „Pearson Chi-Square“ geben an, ob die Annahme der Gleichheit von Mittelwert und Varianz erfüllt ist oder nicht. Wären geschätzte Mittelwert und Varianz der beobachteten Daten gleich, so sollten diese Werte = 1 sein. Diese Annahme ist bei den Beispieldaten (und in sehr vielen realen Situationen) nicht erfüllt. Später werden Lösungen für diese sog. Over-Dispersion diskutiert.
Der Einfluss der unabhängigen Variablen im Modell lässt sich aus folgendem Teil des Outputs erkennen:

```
Analysis Of Parameter Estimates

                      Standard  Wald 95% Confidence  Chi-
Parameter DF  Estimate   Error        Limits       Square Pr > ChiSq

Intercept  1    0.1561  0.0170   0.1228   0.1893  84.62     <.0001
TREAT      1   -0.0499  0.0243  -0.0976  -0.0022   4.20     0.0405
Scale      0    1.0000  0.0000   1.0000   1.0000

NOTE: The scale parameter was held fixed.
```

Mit der im SAS/ETS-Modul bereitgestellten Prozedur COUNTREG kann eine Poisson-Regressionsanalyse besonders einfach programmiert werden:

```
proc countreg data=daten;
  model count = treatn / dist=poisson;
run;
```

Der zugehörige Output liefert die gleichen Ergebnisse:

```
The COUNTREG Procedure

             Model Fit Summary

Dependent Variable                    COUNT
Number of Observations                 5926
Data Set                          WORK.DATEN
Model                                Poisson
Log Likelihood                       -10170
Maximum Absolute Gradient         1.72127E-7
Number of Iterations                      4
Optimization Method           Newton-Raphson
AIC                                   20343
SBC                                   20357
```

```
Algorithm converged.

                        Parameter Estimates

                                    Standard                  Approx
Parameter     DF      Estimate         Error    t Value    Pr > |t|

Intercept      1      0.056341      0.038801       1.45      0.1465
TREATN         1      0.049865      0.024337       2.05      0.0405
```

3 Berücksichtigung unterschiedlicher Beobachtungszeiten

Eine verbesserte Anpassung der beobachteten Zähldaten ist oft möglich, wenn die verschiedenen Beobachtungszeiten mit im Modell berücksichtigt werden. Da in der vorliegenden Beispielstudie die gesamte Beobachtungszeit über fünf Jahre betragen konnte, eine zu beachtende Anzahl von Patienten jedoch nicht die gesamte Zeit an der Studie teilnahmen, schien dies auch hier sinnvoll.
Die Berücksichtigung eines solchen Offsets abhängig von der tatsächlichen Beobachtungsdauer jedes Patienten lässt sich in beiden im vorigen Kapitel beschriebenen Prozeduren mittels der Option OFFSET= im MODEL-Befehl verwirklichen. Als Offset wird meistens nicht die originale Beobachtungszeit, sondern deren Logarithmus benutzt. Der Prozeduraufruf und Output für PROC GENMOD (für COUNTREG entsprechend analog) sieht dann so aus:

```
proc genmod data=daten;
  class treat;
  model count = treat / offset=logfu dist=poisson link=log;
run;
```

Output:

```
          Criteria For Assessing Goodness Of Fit

Criterion                 DF           Value        Value/DF

Deviance                5924      15750.3341          2.6587
Scaled Deviance         5924      15750.3341          2.6587
Pearson Chi-Square      5924      48519.2298          8.1903
Scaled Pearson X2       5924      48519.2298          8.1903
Log Likelihood                     -7345.0443
Full Log Likelihood               -11646.8257
AIC (smaller is better)            23297.6515
AICC (smaller is better)           23297.6535
BIC (smaller is better)            23311.0257

Algorithm converged.
```

```
Analysis Of Parameter Estimates

                      Standard  Wald 95% Confidence  Chi-
Parameter DF  Estimate   Error        Limits       Square Pr > ChiSq

Intercept  1   -7.1670  0.0170  -7.2002  -7.1337 178444     <.0001
TREAT      1   -0.0577  0.0243  -0.1054  -0.0100   5.63     0.0177
Scale      0    1.0000  0.0000   1.0000   1.0000

NOTE: The scale parameter was held fixed.
```

Es ist klar zu erkennen, dass unter dem hier verwendeten (realistischen) Modell die Annahme der Gleichheit von Mittelwert und Varianz verletzt ist.

Die Güte der Anpassung der verwendeten Modelle an die tatsächlich beobachteten Daten lässt sich graphisch gut überprüfen. Es werden dazu die beobachteten und die unter der Modellannahme zu erwartenden kumulierten Häufigkeiten dargestellt. Für die vorliegenden Daten ist dies in Abbildung 3 zu sehen.

4 Over-Dispersion

Bei den Analysen der Beispieldaten fällt auf, dass die Annahme der Gleichheit von Mittelwert und Varianz, wie bei der Poisson-Verteilung gefordert, deutlich verletzt ist. Dieses Phänomen, welches in vielen Beispielen zur Poisson-Regressionsanalyse beobachtet wird, bezeichnet man als Over-Dispersion.
Es ist zunächst relativ einfach möglich, mittels eines statistischen Tests zu überprüfen, ob Over-Dispersion vorliegt oder nicht. Cameron und Trivedi (1996) schlagen eine OLS-Regression (OLS: Ordinary Least Square) folgender Art vor:

$$\frac{(y_i - \mu_i)^2 - y_i}{\mu_i} = \alpha\mu_i + e_i$$

Hierbei sei μ_i Exp(X,β)-verteilt und e_i bezeichnet den Fehlerterm. Mit folgendem kleinen SAS-Programm kann dieser Test durchgeführt werden:

```
data ols; set poi_out;
   dep = ((count - p) ** 2 - count) / p;
run;
 proc reg data=ols;
   model dep = p / noint;
 run;
```

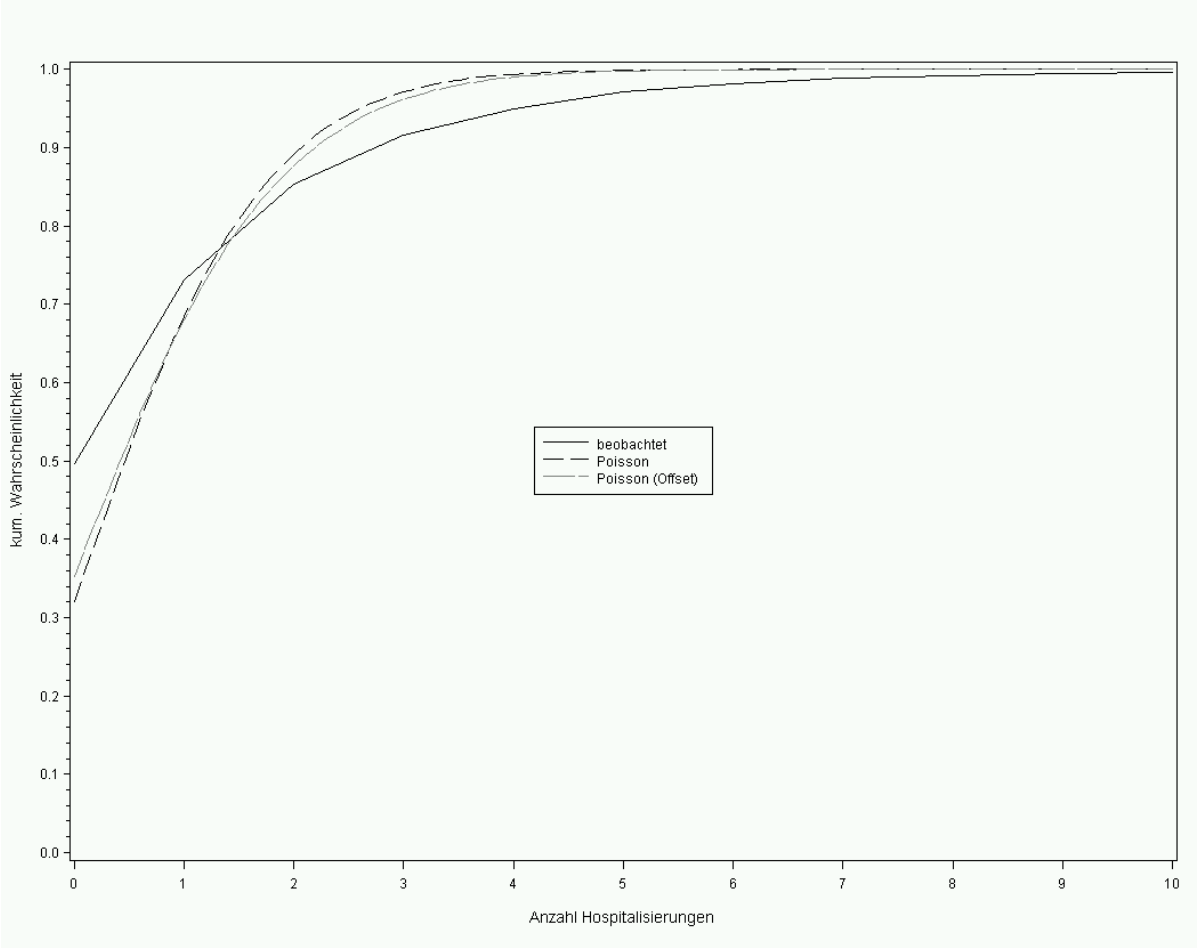

Abbildung 3: Vergleich von beobachteten und theoretischen Wahrscheinlichkeiten

Unter Verwendung der Beispieldaten ergeben sich für die Modelle ohne und mit Offset folgende Outputs:

```
Poisson-Regression
Parameter Estimates

                                  Parameter  Standard
Variable Label           DF       Estimate     Error  t Value  Pr > |t|

P        Predicted Value  1        1.55617   0.14092    11.04    <.0001

Poisson-Regression mit Offset
Parameter Estimates

                                  Parameter  Standard
Variable Label           DF       Estimate     Error  t Value  Pr > |t|

P        Predicted Value  1        1.56005   0.69211     2.25    0.0242
```

Beide Tests zeigen, dass die Annahme von gleichem Mittelwert und Varianz nicht erfüllt ist.

5 Poisson-Regression mit Dispersionsparameter

Das Problem der Over-Dispersion kann gelöst werden, indem ein sog. Dispersionsparameter ϕ eingeführt wird, für den gilt:

$\mathrm{Var}(y_i) = \phi\, \mu_i$

Im SAS-Output ist dieser Wert ϕ über den Scale-Parameter zu berechnen, da dieser gleich $\sqrt{\phi}$ ist.

In SAS können nun sowohl der unter „Deviance“ als auch der unter „Pearson Chi-Square“ angegebene Wert = 1 gesetzt werden, um ein Modell mit Dispersionsparameter zu berechnen. Die beiden entsprechenden Optionen zum MODEL-Befehl lauten DSCALE bzw. PSCALE. Die Outputs bei Verwendung beider Optionen sehen folgendermaßen aus:

Output mit DSCALE:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5924	15750.3341	**2.6587**
Scaled Deviance	5924	5924.0000	**1.0000**
Pearson Chi-Square	5924	48519.2298	**8.1903**
Scaled Pearson X2	5924	18249.0045	3.0805
Log Likelihood		-2762.6108	
Full Log Likelihood		-4380.5925	
AIC (smaller is better)		8765.1850	
AICC (smaller is better)		8765.1871	
BIC (smaller is better)		8778.5593	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-7.1670	0.0277	-7.2212	-7.1128	67116	<.0001
TREAT	1	-0.0577	0.0397	-0.1355	0.0201	2.12	**0.1458**

```
Scale      0    1.6306  0.0000   1.6306   1.6306

NOTE: The scale parameter was estimated by the square root of
DEVIANCE/DOF.
```

Output mit PSCALE:

```
            Criteria For Assessing Goodness Of Fit

Criterion                     DF           Value        Value/DF

Deviance                    5924      15750.3341          2.6587
Scaled Deviance             5924       1923.0518          0.3246
Pearson Chi-Square          5924      48519.2298          8.1903
Scaled Pearson X2           5924       5924.0000          1.0000
Log Likelihood                        -2762.6108
Full Log Likelihood                   -4380.5925
AIC (smaller is better)                8765.1850
AICC (smaller is better)               8765.1871
BIC (smaller is better)                8778.5593

Algorithm converged.

Analysis Of Parameter Estimates

                     Standard  Wald 95% Confidence  Chi-
Parameter DF  Estimate   Error        Limits       Square Pr > ChiSq

Intercept  1   -7.1670  0.0277  -7.2212  -7.1128  67116     <.0001
TREAT      1   -0.0577  0.0696  -0.0788   0.1942   0.69     0.4072
Scale      0    2.8619  0.0000   2.8619   2.8619

NOTE: The scale parameter was estimated by the square root of
Pearson's Chi-Square/DOF.
```

Bei beiden Modellen verändert sich der p-Wert aufgrund des größeren Standardfehlers des Schätzers für die Behandlung von unterhalb 0,05 auf Werte oberhalb 0,05. Leider sind in der Literatur keine Hinweise zu finden, wann welche der beiden Optionen sinnvoller ist. Aufgrund der recht deutlichen Unterschiede in den Ergebnissen sind nähere Untersuchungen hier sicherlich notwendig.

6 Negative Binomial-Regression

Neben den in den bisherigen Kapiteln vorgestellten Modellen auf Basis einer Poisson-Verteilung können auch andere Verteilungsannahmen zugrunde gelegt werden, die oft eine größere Flexibilität haben.
Häufig kann eine Regressionsanalyse auf Basis einer negativen Binomialverteilung sinnvoll sein. In diesem Modell wird der Zusammenhang zwischen Mittelwert und Varianz flexibler dargestellt: $\mathrm{Var}(y_i) = \mu_i + k\,\mu_i^2$
Diese Beziehung berücksichtigt u.a. auch Over-Dispersion. Eine graphische Darstellung verschiedener Poisson-Verteilungen ist in Abbildung 4 zu finden.

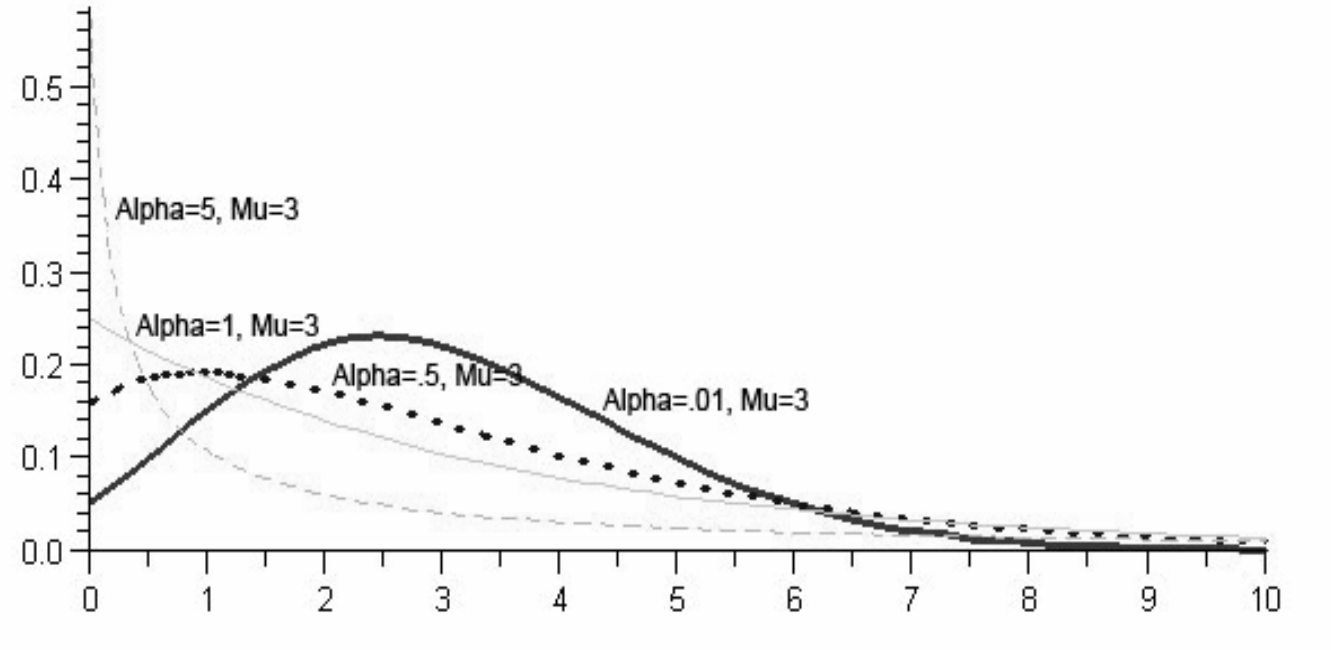

Abbildung 4: Verschiedene negativ bionomial-Verteilungen

Innerhalb der Prozedur GENMOD in SAS können solche Modelle analysiert werden:

```
proc genmod data=daten;
   class treat;
   model count = treat / offset=logfu dist=nb link=log;
run;
```

Der Output sieht wie folgt aus:

```
Criteria For Assessing Goodness Of Fit

Criterion                     DF           Value        Value/DF

Deviance                    5924       6119.4387          1.0330
Scaled Deviance             5924       6119.4387          1.0330
Pearson Chi-Square          5924      24584.8336          4.1500
Scaled Pearson X2           5924      24584.8336          4.1500
Log Likelihood                        -5248.5751
```

```
Full Log Likelihood                  -9550.3564
AIC (smaller is better)              19106.7129
AICC (smaller is better)             19106.7170
BIC (smaller is better)              19126.7742

Algorithm converged.

Analysis Of Parameter Estimates

                        Standard  Wald 95% Confidence  Chi-
Parameter DF  Estimate    Error        Limits       Square Pr > ChiSq

Intercept  1   -6.9306  0.0338  -6.9968  -6.8644  42102     <.0001
TREAT      1   -0.0498  0.0476  -0.1431   0.0435   1.09     0.2958
Scale      1    2.0800  0.0709   1.9410   2.2190

NOTE: The negative binomial dispersion parameter was estimated by
maximum likelihood.
```

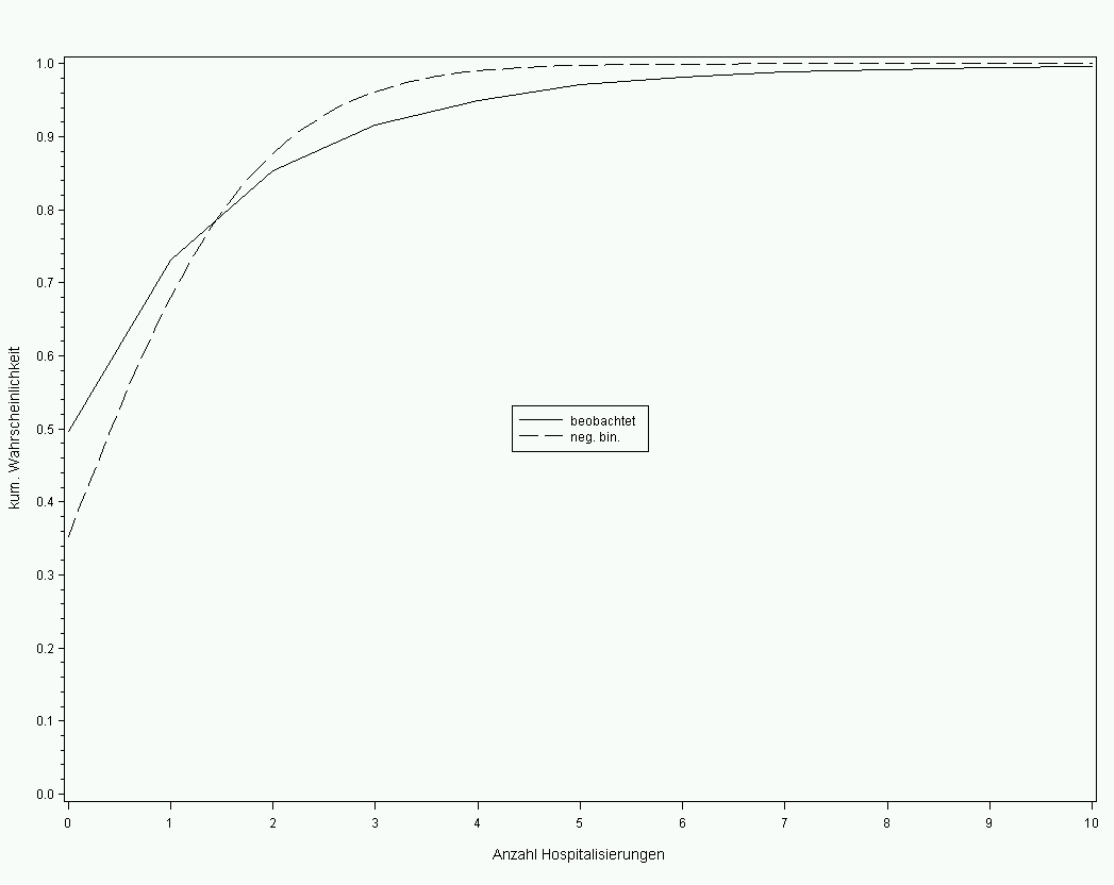

Abbildung 5: Vergleich von beobachteten und theoretischen Wahrscheinlichkeiten

Auch in diesem Modell kann kein Unterschied zwischen den beiden Behandlungen nachgewiesen werden. Die Ergebnisse ähneln daher denen der Poisson-Regression mit Dispersionsparameter. Die Güte der Anpassung lässt sich erneut graphisch veranschaulichen (Abbildung 5).

7 Zweistufige Regressionsmodelle

Neben dem Problem der Over-Dispersion tritt bei vielen Analysen von Zähldaten ein weiteres Problem auf. Da sehr häufig negative Ereignisse (Krankheiten, Schadensfälle, …) beobachtet werden, ist der Anteil von Beobachtungen ohne ein einziges aufgetretenes Ereignis oft überproportional groß. Dieses gehäufte Auftreten von keinem Ereignis kann in speziellen Modellen berücksichtigt werden. Als Beispiel solcher meistens zweistufiger Modelle sollen hier die Hurdle-Regression und die Zero-inflated Regression kurz vorgestellt werden.

7.1 Hurdle-Regression

Das zweistufige Regressionsmodell der Hurdle-Regression modelliert die Wahrscheinlichkeit, ob = 0 oder ≠ 0 Ereignisse auftreten, mittels einer Binomialverteilung. Für positive Anzahlen an Ereignissen wird eine Poisson-Verteilung angenommen. Die entsprechende Dichtefunktion lautet also:

$$f\left(Y_i = y_i \mid X_i = x_i\right) = \begin{cases} \Theta_i & y_i = 0 \\ \dfrac{\left(1-\Theta_i\right)\cdot e^{-\mu_i}\cdot \mu_i^{y_i}}{\left(1-e^{-\mu_i}\right)\cdot y_i!} & y_i > 0 \end{cases}$$

mit $\theta_i = P\,(y_i = 0)$ und $\mu_i = \mathrm{Exp}(X, \beta)$.

In SAS sind solche Modelle mit den Prozeduren NLMIXED, MODEL oder NLIN analysierbar. Im folgenden ist beispielhaft ein entsprechendes Programm für NLMIXED angegeben:

```
proc nlmixed data=daten;
  parms a0=0 a1=0 b0=0 b1=0;
  eta0 = a0 + a1 * treatn;
  exp_eta0 = exp(eta0);
  p0 = exp_eta0 / (1 + exp_eta0);
  etap = b0 + b1 * treatn;
  exp_etap = exp(etap);
  if count eq 0 then ll = log(p0);
  else ll = log(1 - p0) - exp_etap + count * etap -
            lgamma(count + 1) - log(1 - exp(-exp_etap));
  model count ~ general(ll);
  predict exp_etap out=_hdl1(keep=pred count
                             rename=(pred=yhat));
  predict p0 out=_hdl2(keep=pred rename=(pred=p0));
run;
```

Die beiden Stufen des Modells müssen innerhalb der Prozedur entsprechend der Modellvorgaben mit entsprechenden Befehlen beschrieben werden.
Für die Beispieldaten ergibt sich dann folgender Output:

Fit Statistics

-2 Log Likelihood	18503
AIC (smaller is better)	18511
AICC (smaller is better)	18511
BIC (smaller is better)	18538

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Lower	Upper	Gradient
a0	0.07828	0.08224	5926	0.95	0.3412	-0.08294	0.2395	0.000263
a1	-0.06068	0.05197	5926	-1.17	**0.2430**	-0.1626	0.04120	0.0001
b0	0.6191	0.04738	5926	13.07	<.0001	0.5262	0.7120	0.001995
b1	0.02923	0.02964	5926	0.99	**0.3240**	-0.02887	0.08733	0.002785

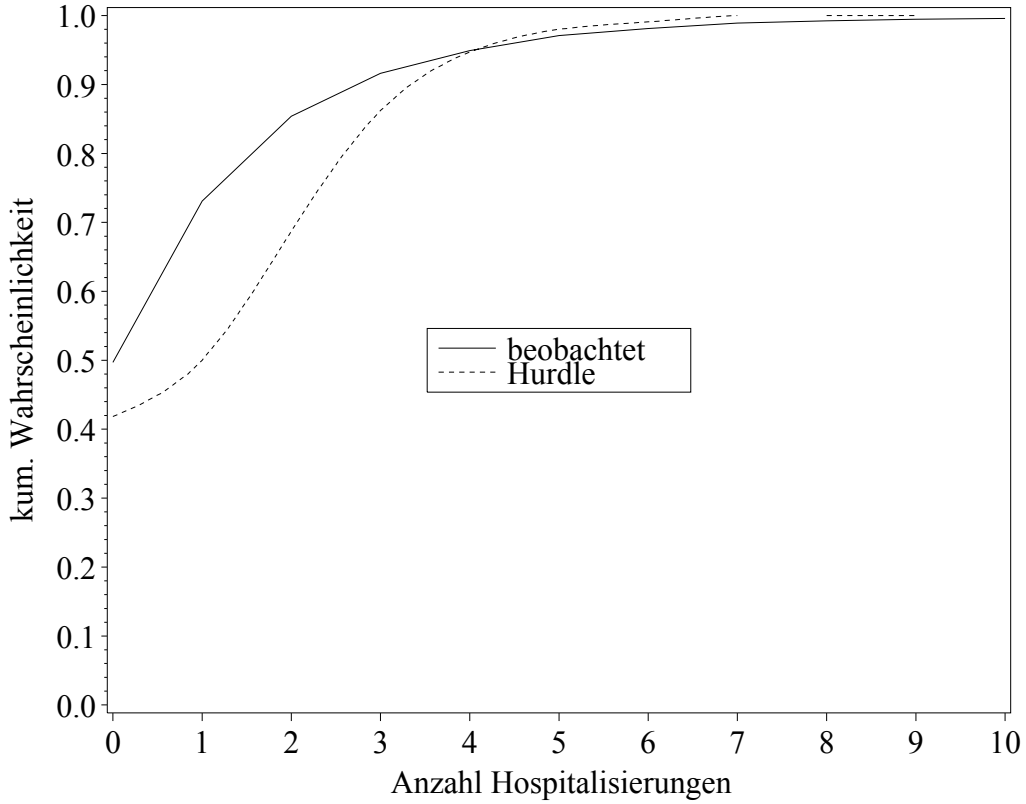

Abbildung 6: Vergleich von beobachteten und theoretischen Wahrscheinlichkeiten

Die graphische Gegenüberstellung von beobachteten mit theoretisch erwarteten Häufigkeiten in Abbildung 6 zeigt, dass für die Beispieldaten keine überzeugende Anpassung mittels einer Hurdle-Regression möglich ist.

7.2 Zero-inflated Regression

Auch die Zero-inflated Regression ist ein zweistufiges Modell. Für jede Beobachtung i wird ein Prozess 1 mit einer Wahrscheinlichkeit φ_i gewählt. Dieser Prozess generiert nur die Anzahl 0. Ein weiterer Prozess 2 wird mit Wahrscheinlichkeit $1 - \varphi_i$ gewählt. Dieser kann als Poisson- oder negativ-binomial-Modell gewählt werden:

$g(y_i)$ ist hier negativ-binomial

$$P(y_i = 0 \mid X_i = x_i) = \varphi_i + (1-\varphi_i)g(0) \quad \text{Poisson- oder}$$

$$P(y_i \mid X_i = x_i) = (1-\varphi_i)g(y_i) \quad , \quad y_i > 0 \quad \text{verteilt.}$$

Auch hier ist beispielhaft ein entsprechender Prozeduraufruf von NLMIXED wiedergegeben:

```
proc nlmixed data=daten;
  parms a0=0 a1=0 b0=0 b1=0;
  eta0 = a0 + a1 * treatn;
  exp_eta0 = exp(eta0);
  p0 = exp_eta0 / (1 + exp_eta0);
  etap = b0 + b1 * treatn;
  exp_etap = exp(etap);
  if count eq 0 then ll = log(p0 + (1-p0)*exp(-exp_etap));
  else ll = log(1 - p0) - exp_etap + count * etap -
            lgamma(count + 1);
  model count ~ general(ll);
  predict exp_etap out=_zi1(keep=pred count
                            rename=(pred=yhat));
  predict p0 out=_zi2(keep=pred rename=(pred=p0));
run;
```

Für die Beispieldaten sind die Ergebnisse denen bei der Hurdle-Regression angegebenen sehr ähnlich.

8 Vergleich der verschiedenen Modelle

Um für vorhandene Beobachtungen die beste Anpassung zu finden, sollten die verschiedenen Modelle miteinander verglichen werden. Durch den graphischen Vergleich der beobachteten mit den theoretisch erwarteten Häufigkeiten können erste Aufschlüsse über die Güte der Anpassung eines bestimmten Modells erhalten werden. Objektivere Möglichkeiten, verschiedene Modellanpassungen zu bewerten, ergeben sich über das AIC (Akaike-Informationskriterium), welches bei allen Modellen in SAS angegeben wird. In Tabelle 1 sind für die in den vorigen Kapiteln vorgestellten Modellen die AIC-Werte sowie p-Wert und Log-Likelihood-Wert angegeben.

Tabelle 1: Vergleich verschiedener Regressionsmodelle

Modell	p-Wert	Log-Likelihood	AIC
Poisson	0,0405	–10170	20343
Poisson mit Offset	0,0177	–11647	23298
Dispersion	0,1458	–4381	8765
Negativ binomial	0,2958	–9550	19107
Hurdle	0,3240	–9251	18511
Zero-inflated	0,3240	–9251	18511

Ein gegebener AIC-Wert steht für eine bessere Modellanpassung je kleiner der AIC-Wert ist. Für die gegebenen Beispieldaten zeigt sich aus den Tabellenzahlen, dass das Poisson-Regressionsmodell mit Dispersion die beste Anpassung zeigt und somit hier zur Analyse gewählt werden sollte.

9 Zusammenfassung

Für die Analyse von Zähldaten finden sich verschiedene Möglichkeiten der Regressionsanalyse. Im konkreten Fall muss daher die beste Anpassung an vorhandene Daten gefunden werden. Als spezielle Probleme müssen Over-Dispersion und häufiges Auftreten von Nullwerten entsprechend beachtet werden.
In SAS stehen verschiedene Prozeduren zur Analyse von Zähldaten zur Verfügung, z.B. COUNTREG, NLMIXED, GENMOD, MODEL, u.a. Jedoch sind nicht alle Regressionsmodelle in jeder Prozedur durchzuführen, gerade komplexere zweistufige Regressionsmodelle erfordern einen gewissen Programmieraufwand.
Ein Vergleich verschiedener Modelle sollte sowohl graphisch als auch über Werte wie Log-Likelihood und AIC durchgeführt werden. Eine Bevorzugung für bestimmte Modelle kann nicht ausgesprochen werden, je nach Art der Daten muss entsprechend entschieden werden.

Literatur

[1] Lambert, D. (1992), Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing, Technometrics, Vol. 34, No. 1, 1 – 14.

[2] Liu W., Cela J. (2008), Count Data Models in SAS, Proceedings of the SAS Global Forum. Paper 371-2008.

[3] Mullahy, J. (1986), Specification and Testing of Some Modified Count Data Models, Journal of Econometrics, 33, 341-365.

[4] Pedan, A. (2001), Analysis of Count Data Using the SAS System, Proceedings of the 26th Annual SAS Users Group International Conference. Paper 247-26.

[5] Cameron A.C., Trivedi P.K. (1998), Regression Analysis of Count Data, New York: Cambridge University Press.