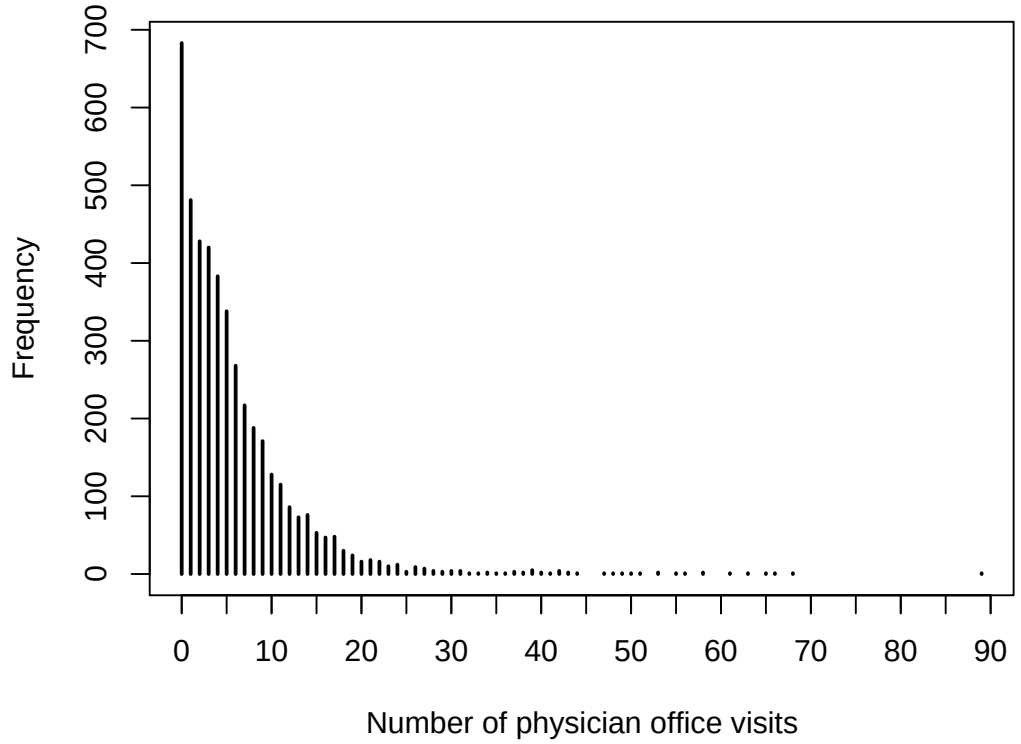


Generalized count data regression in R

Christian Kleiber
U Basel

and

Achim Zeileis
WU Wien

Outline

- Introduction
- Regression models for count data
- Zero-inflation models
- Hurdle models
- Generalized negative binomial models
- Further extensions

Introduction

- Classical count data models (Poisson, NegBin) often not flexible enough for applications in economics and the social sciences.

- Typical problems include overdispersion and excess zeros.

  Also relevant in e.g. fisheries research, medical sciences (DMF teeth index) etc.

- Zero-inflation and hurdle models (Mullahy, *J. Econometrics* 1986, Lambert, *Technometrics* 1992) address excess zeros, implicitly also overdispersion.

  Recent paper on implementation in R:

  Zeileis, Kleiber and Jackman (2008): Regression models for count data in R. *J. Statistical Software*, 27(8). URL `http://www.jstatsoft.org/v27/i8/`

- Generalizations of NegBin have more flexible variance function or additional source of heterogeneity via regressors in shape parameter.

Regression models for count data

Classifications:

- **Classical count data models:**
  - Poisson regression
  - Negative binomial regression (including geometric regression)
  - Quasi-Poisson regression
- **Generalized count data models:**
  - Zero-inflation models
  - Hurdle models
  - NegBin-P model
  - heterogeneous NegBin model (NB-H)
- **Single-index models:** Poisson, quasi-Poisson, geometric, negative binomial, NB-P
- **Multiple-index models:** zero-inflation models, hurdle models, NB-H

Regression models for count data

Count data models in R: (incomplete list!)

- **stats**: Poisson and quasi-Poisson models via `glm()`
- **MASS**: negative binomial and geometric regression via `glm.nb()`
- **pscl**: zero-inflation and hurdle models via `zeroinfl()` and `hurdle()`
- **AER**: testing for equidispersion via `dispersiontest()`
- **flexmix**: finite mixtures of Poissons via `flexmix()`
- **gamlss**: Poisson-inverse Gaussian (PIG) regression via `gamlss()`

Regression models for count data

Generalized linear models are defined by 3 elements:

- Linear predictor $\eta_i = x_i^\top \beta$ through which $\mu_i = \mathsf{E}(y_i|x_i)$ depends on vectors x_i of observations and β of parameters.

- Distribution of dependent variable $y_i|x_i$ is linear exponential family

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}.$$

- Expected response μ_i and linear predictor η_i are related by monotonic transformation

$$g(\mu_i) = \eta_i,$$

  called **link function**.

Regression models for count data

- **Poisson model:**

$$f(y;\mu) \quad = \quad \frac{\exp(-\mu)\cdot\mu^y}{y!}, \qquad y=0,1,2,\ldots$$

- **Negative binomial model:**

$$f(y;\mu,\theta) \quad = \quad \frac{\Gamma(y+\theta)}{\Gamma(\theta)\cdot y!}\cdot\frac{\mu^y\cdot\theta^\theta}{(\mu+\theta)^{y+\theta}}, \qquad y=0,1,2,\ldots$$

- **Canonical link** is $g(\mu)=\log(\mu)$ for both.

- NegBin is GLM only for fixed θ. Special case: geometric distribution for $\theta=1$.

Regression models for count data

Example: (US National Medical Expenditure Survey [NMES] data for 1987/88)

Available as `NMES1988` in package **AER** (Kleiber and Zeileis 2008).
Originally taken from Deb and Trivedi (*J. Applied Econometrics* 1997).

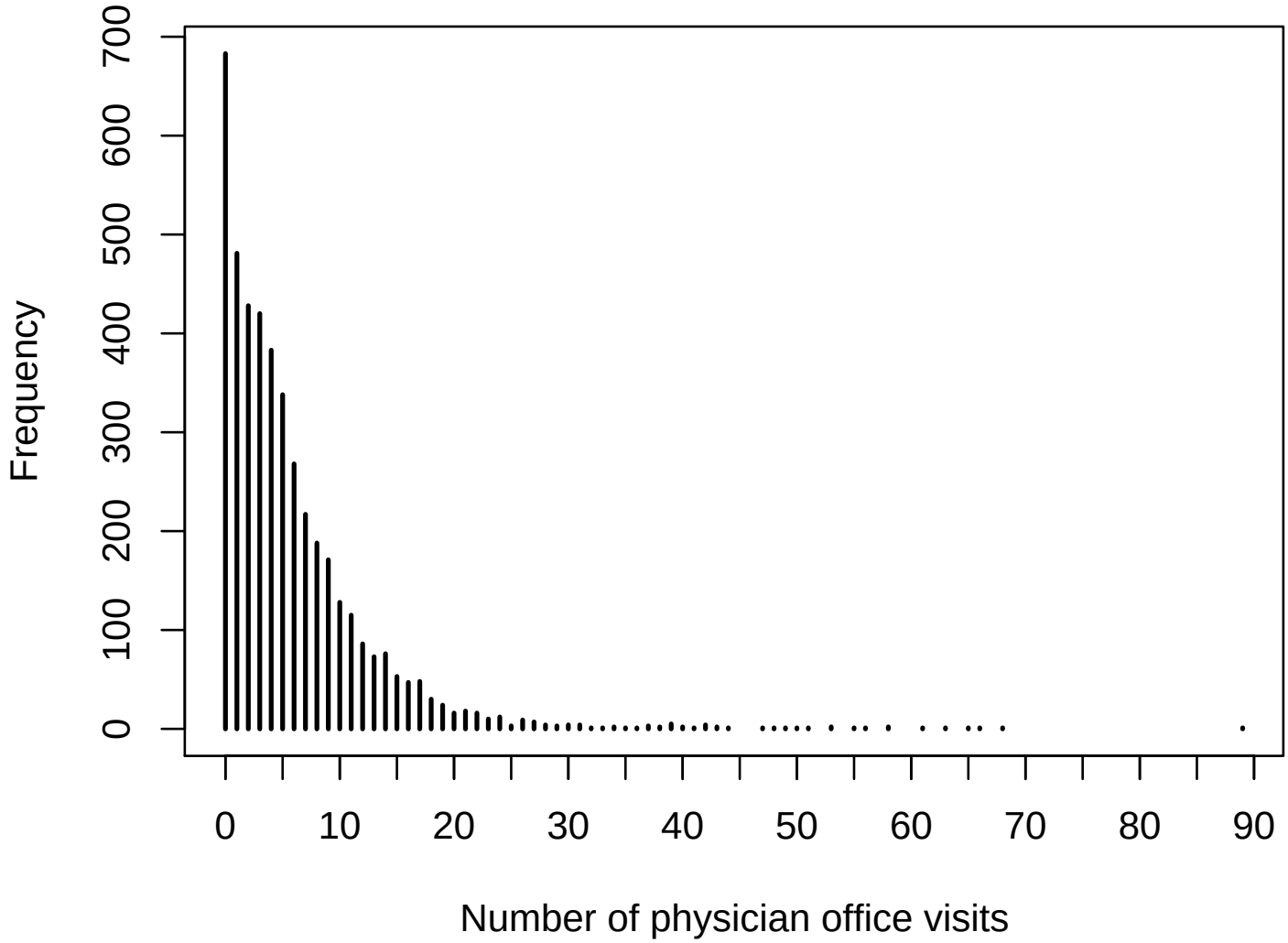
$n = 4406$ individuals, aged 66 and over, covered by Medicare

Objective: model demand for medical care – here defined as number of physician office visits – in terms of covariates.

Variables:

`visits` – number of physician office visits (response)
`hospital` – number of hospital stays
`health` – self-perceived health status
`chronic` – number of chronic conditions
`gender` – gender
`school` – number of years of education
`insurance` – private insurance indicator

Regression models for count data



Zero-inflation models

A mixture of point mass at zero $I_{\{0\}}(y)$ and count distribution $f_{\text{count}}(y; x, \beta)$:

$$f_{\text{zeroinfl}}(y; x, z, \beta, \gamma) = \pi \cdot I_{\{0\}}(y) \; + \; (1 - \pi) \cdot f_{\text{count}}(y; x, \beta)$$

- Probability of observing zero count is inflated with probability π.
- More recent applications have $\pi = f_{\text{zero}}(0; z, \gamma)$.
  Unobserved probability π is modelled by binomial GLM $\pi = g^{-1}(z^\top \gamma)$.
- Regression equation for the mean is (using canonical [= log] link)

$$\mu_i \quad = \quad \pi_i \cdot 0 \; + \; (1 - \pi_i) \cdot \exp(x_i^\top \beta),$$

- Vectors of regressors z_i and x_i need not be distinct.
- Inference for (β, γ, θ) by ML. θ is treated as nuisance parameter.

Zero-inflation models

In R:

- Package **pscl** has function `zeroinfl()`
- Typical call looks like

```
R> dt_zinb <-   zeroinfl(visits ~ . |
+ hospital + chronic + insurance + school + gender,
+ data = dt, dist = "negbin")
```

- Count part specified by `dist` argument, using canonical [= log] link.
- Binary part defaults to `link = "logit"`, other links also available.
- Optimization via `optim()`. Otherweise GLM building blocks are reused.
- Methods include `coef()`, `fitted()`, `logLik()`, `predict()`, `summary()`, `vcov()`.

Zero-inflation models

```
Call:
zeroinfl(formula = visits ~ . | hospital + chronic + insurance +
    school + gender, data = dt, dist = "negbin")


Count model coefficients (negbin with log link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.19372    0.05666   21.07  < 2e-16
hospital         0.20148    0.02036    9.90  < 2e-16
healthpoor       0.28513    0.04509    6.32  2.6e-10
healthexcellent -0.31934    0.06040   -5.29  1.2e-07
chronic          0.12900    0.01193   10.81  < 2e-16
gendermale      -0.08028    0.03102   -2.59   0.0097
school           0.02142    0.00436    4.92  8.8e-07
insuranceyes     0.12586    0.04159    3.03   0.0025
Log(theta)       0.39414    0.03503   11.25  < 2e-16
```

Zero-inflation models

```
Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.0468     0.2686   -0.17   0.8615
hospital      -0.8005     0.4208   -1.90   0.0571
chronic       -1.2479     0.1783   -7.00  2.6e-12
insuranceyes  -1.1756     0.2201   -5.34  9.3e-08
school        -0.0838     0.0263   -3.19   0.0014
gendermale     0.6477     0.2001    3.24   0.0012

Theta = 1.483
Number of iterations in BFGS optimization: 28
Log-likelihood: -1.21e+04 on 15 Df
```

Hurdle models

Hurdle model combines

- Count part $f_{\text{count}}(y; x, \beta)$ (count left-truncated at $y = 1$)
- Zero hurdle part $f_{\text{zero}}(y; z, \gamma)$ (count right-censored at $y = 1$)

$$f_{\text{hurdle}}(y; x, z, \beta, \gamma) = \begin{cases} f_{\text{zero}}(0; z, \gamma) & \text{if } y = 0, \\ (1 - f_{\text{zero}}(0; z, \gamma)) \cdot \frac{f_{\text{count}}(y; x, \beta)}{1 - f_{\text{count}}(0; x, \beta)} & \text{if } y > 0 \end{cases}$$

Inference for parameters (β, γ, θ) by ML. θ is treated as nuisance parameter.

Logit and censored geometric models as hurdle part both lead to same likelihood, and thus to identical estimates.

If same regressors $x_i = z_i$ are used one can test $\beta = \gamma$ – is hurdle needed or not?

Hurdle models

In R:

- Package **pscl** has function `hurdle()`
- Typical call is

```
R> dt_hurdle <-   hurdle(visits ~ . |
+ hospital + chronic + insurance + school + gender,
+ data = dt, dist = "negbin")
```

- Count part specified by `dist` argument, using canonical [= log] link.
- Binary part defaults to `zero.dist = "binomial"` with `link = "logit"`, other links and distributions also available.
- Optimization via `optim()`. Otherweise GLM building blocks are reused.
- Methods include `coef()`, `fitted()`, `logLik()`, `predict()`, `summary()`, `vcov()`.

Hurdle models

```
Call:
hurdle(formula = visits ~ . | hospital + chronic + insurance +
    school + gender, data = dt, dist = "negbin")


Count model coefficients (truncated negbin with log link):
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.19770    0.05897   20.31  < 2e-16
hospital         0.21190    0.02140    9.90  < 2e-16
healthpoor       0.31596    0.04806    6.57  4.9e-11
healthexcellent -0.33186    0.06609   -5.02  5.1e-07
chronic          0.12642    0.01245   10.15  < 2e-16
gendermale      -0.06832    0.03242   -2.11    0.035
school           0.02069    0.00453    4.56  5.0e-06
insuranceyes     0.10017    0.04262    2.35    0.019
Log(theta)       0.33325    0.04275    7.79  6.5e-15
```

Hurdle models

```
Zero hurdle model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.0159     0.1378    0.12  0.90788
hospital       0.3184     0.0911    3.50  0.00047
chronic        0.5478     0.0436   12.57  < 2e-16
insuranceyes   0.7457     0.1003    7.43  1.1e-13
school         0.0571     0.0119    4.78  1.7e-06
gendermale    -0.4191     0.0875   -4.79  1.7e-06

Theta: count = 1.396
Number of iterations in BFGS optimization: 16
Log-likelihood: -1.21e+04 on 15 Df
```

Generalized negative binomial models

NegBin-P model: (Winkelmann and Zimmermann 1991, Greene 2008)

Negative binomial in standard parametrization has variance function

$$\mathsf{Var}(y_i|x_i) = \mu_i \left(1 + \frac{1}{\theta}\mu_i\right)$$

Special case of

$$\mathsf{Var}(y_i|x_i) = \mu_i \left(1 + \frac{1}{\theta}\mu_i^{P-1}\right)$$

Common versions are $P = 1, 2$, called **NB1** and **NB2**.

Can also estimate P, this gives **NB-P** model.

Generalized negative binomial models

NegBin-H model: (Greene 2007)

Further generalization to multiple index model via

$$\mathsf{Var}(y_i|x_i) = \mu_i \left(1 + \frac{1}{\theta_i}\mu_i^{P-1}\right)$$

with $\theta_i = \exp(z_i^\top \gamma)$.

R implementation of NB-P and NB-H by D. Cueni (M.S. thesis, U Basel 2008).

Optimization via `nlminb()`.

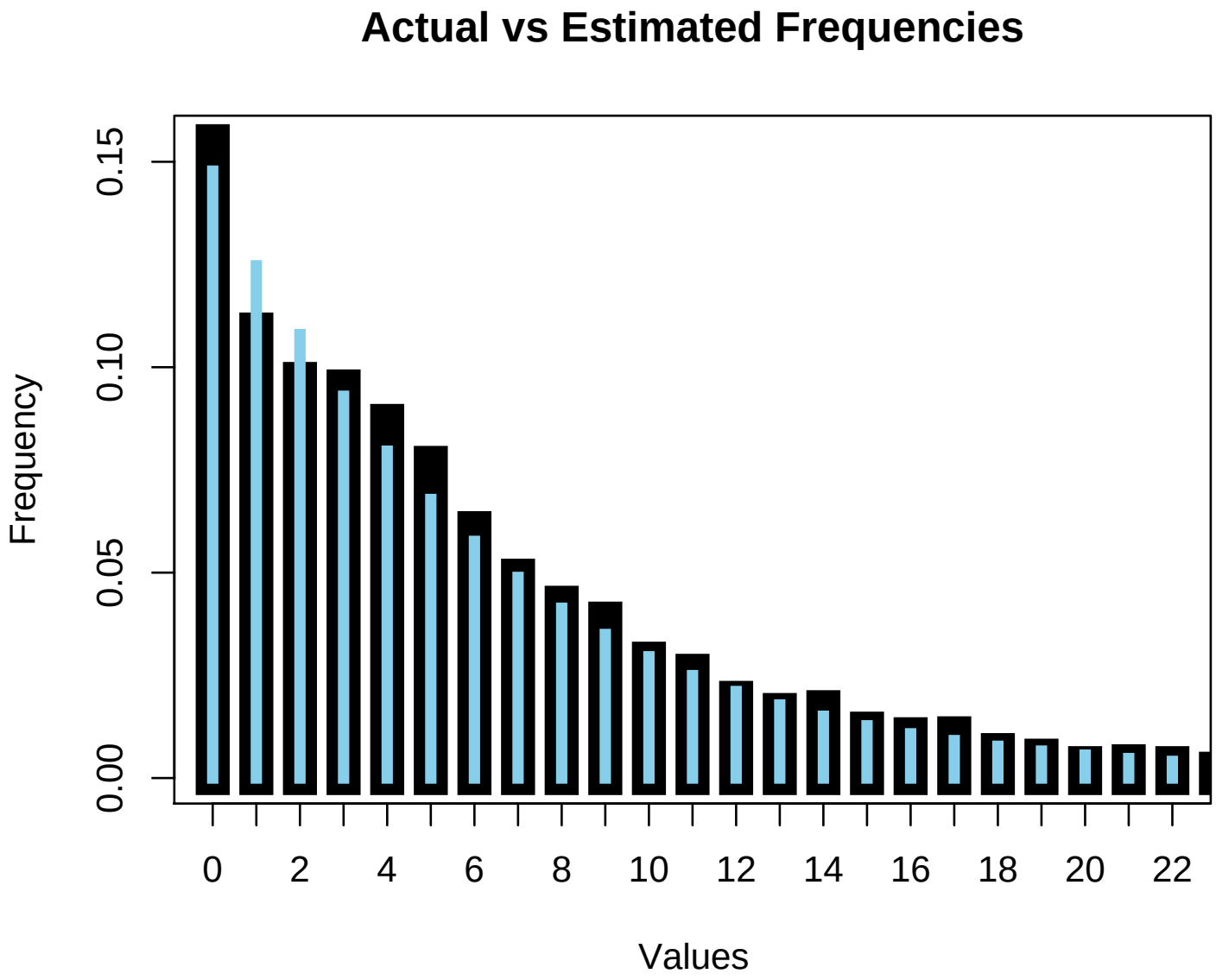
Programs allow for fixing P, thus enabling NB1 regression.

Generalized negative binomial models

Results for 4 models:

```
R> logLik(dt_nb2)

'log Lik.' -12171 (df=9)

R> logLik(dt_hurdle)

'log Lik.' -12090 (df=15)

R> logLik(dt_nbp)

'log Lik.' -12135 (df=10)

R> logLik(dt_nbh)

'log Lik.' -12098 (df=15)
```

Generalized negative binomial models



Further extensions

Welcome additions:

- more on multivariate count data models (**bivpois** has bivariate Poisson models)
- more on finite mixtures (**flexmix** has finite mixtures of Poissons, but not of NegBins).
- count models for panels (to some extent available in **lme4**, **glmmML**, ...)
- further Poisson mixtures
- count models with endogeneity, selectivity, ...