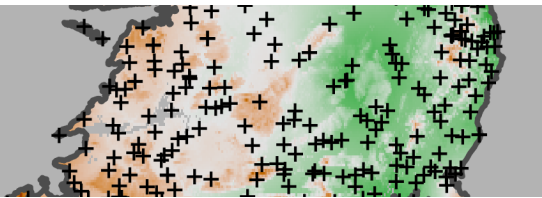
universität
innsbruck

transitreg



Transition Regression for Continuous Data

Reto Stauffer & Nikolaus Umlauf

www.uibk.ac.at/statistics

Count data models

Classical count data regression

Estimate the parameters of a specific distribution.

Count data models

Classical count data regression

Estimate the parameters of a specific distribution.

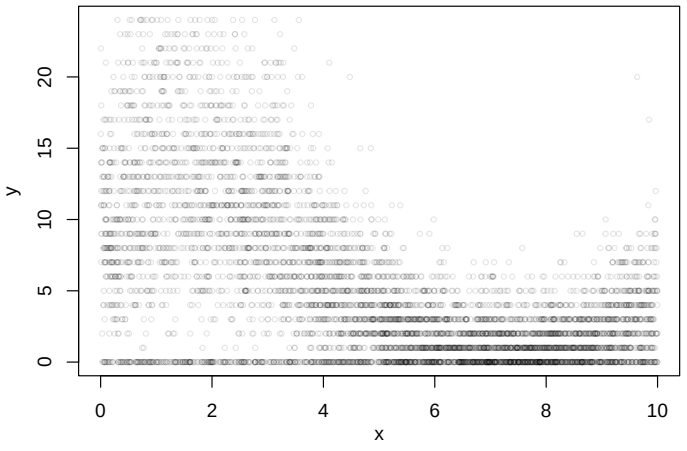
Frequently used models are e.g.,

- Poisson models,
- Negative Binomial models,
- zero-inflated models*,
- or hurdle models*,

*whereof some are designated to account for an excess of zeros (Zeileis & Kleiber 2008, 2024).

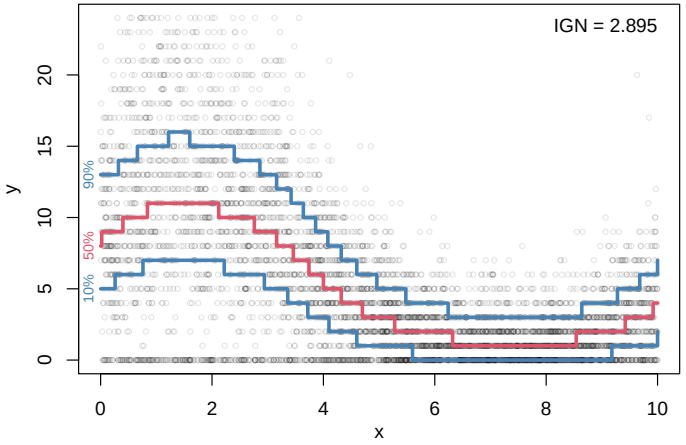
Count data models

Simulated example: $y \sim \text{NBI}(\mu = f(x), \sigma = f(x))$ plus 15% extra 0s.



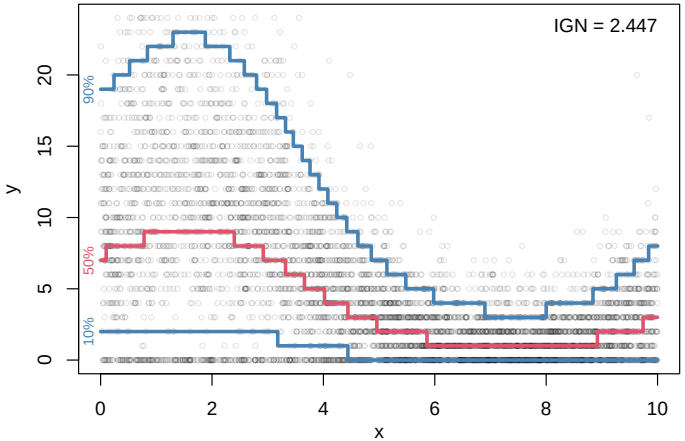
Count data models

Estimated model: $y \sim \text{PO}(\log(\mu) = f(x))$.



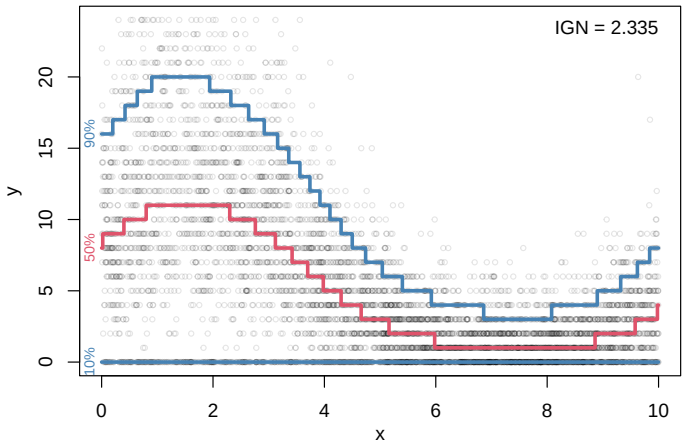
Count data models

Estimated model: $y \sim \text{NBI}(\log(\mu) = f(x), \log(\sigma) = f(x))$.



Count data models

Estimated model: $y \sim \text{ZANBI}(\log(\mu) = f(x), \log(\sigma) = f(x), \text{logit}(\nu) = f(x))$.



Transition regression

Transition models for count data

Modeling the probability $P(\dot{})$ of observation y_i transitioning into higher counts r dependent on covariates $\mathbf{x}$ (Berger & Tutz 2021).

$$P(y_i > r | y_i \geq r, \mathbf{x}_i) \in [0, 1].$$

Transition regression

Transition models for count data

Modeling the probability $P(\dot{})$ of observation y_i transitioning into higher counts r dependent on covariates $\mathbf{x}$ (Berger & Tutz 2021).

$$P(y_i > r | y_i \geq r, \mathbf{x}_i) \in [0, 1].$$

Similar models are also known as ‘continuation ratio models’ or ‘ordered logistic models’ in the logit context.

Transition regression

Core idea

- Estimate the probability of transitioning between intervals (counts)
- No need for prior assumptions about the response distribution
- Account for excesses in any count(s) (e.g., excess zeros)
- Highly flexible in both its parameters and the response distribution
- Allows retrieval of quantiles, mean, median, mode, . . .

Transition regression

The transition probability $P(\cdot)$ for count data is defined as

$$P(y_i > r | y_i \geq r, \mathbf{x}_i) = F(\eta_{ir}(\boldsymbol{\alpha})), \quad r = 0, 1, 2, \ldots$$

where $F(\cdot)$ is a CDF (e.g., logistic or probit) and r represents the counts, with an additive predictor

$$\eta_{ir}(\boldsymbol{\alpha}) = \theta_r + \sum_{j=1}^{k} f_j(\mathbf{x}_i, r; \beta).$$

The parameters $\boldsymbol{\alpha} = (\theta^\top, \beta^\top)$ include count-specific intercepts θ_r and (possibly) smooth functions $f_j(\cdot)$. For i.i.d. observations, let π_{ir} denote the probability that the count response equals r, i.e., $P(y_i = r | \mathbf{x}_i)$. These probabilities are computed recursively as

$$\pi_{ir} = (1 - F(\eta_{ir}(\boldsymbol{\alpha}))) \prod_{s=0}^{r-1} F(\eta_{is}(\boldsymbol{\alpha})).$$

Transition models

Parameter estimation considers the underlying Markov chain $Y_{i0}, Y_{i1}, \ldots$, where

$$Y_{ir} = 1 - I(y_i = r) \quad \in \quad \{0, 1\}.$$

Simplifies to binary model log-likelihood

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{s=0}^{y_i} \Big[Y_{is} \log(F(\eta_{ir})) + (1 - Y_{is}) \log(1 - F(\eta_{ir})) \Big].$$

$(Y_{i0}, \ldots, Y_{iy_i})^\top = (1, \ldots, 1, 0)$ are created, along with a new covariate $\theta_i = (0, 1, 2, \ldots, y_i)^\top$ to capture count-specific effects $f_j(\mathbf{x}_i, \theta_i)$, or simple count-specific intercepts.

All other covariates are duplicated accordingly.

Transition models

The data set to model the transition probabilities therefore looks as follows:

```
> print(df)
  index y theta Y        x
1     1 3     0 1 8.861246
2     1 3     1 1 8.861246
3     1 3     2 1 8.861246
4     1 3     3 0 8.861246
5     2 2     0 1 7.609823
6     2 2     1 0 7.609823
7     3 2     0 1 1.494579
8     3 2     1 0 1.494579
```

Transition models

The data set to model the transition probabilities therefore looks as follows:

```
> print(df)
  index y theta Y        x
1     1 3     0 1 8.861246
2     1 3     1 1 8.861246
3     1 3     2 1 8.861246
4     1 3     3 0 8.861246
5     2 2     0 1 7.609823
6     2 2     1 0 7.609823
7     3 2     0 1 1.494579
8     3 2     1 0 1.494579
```

After the inflation, any standard binary regression method can be used (e.g.,):

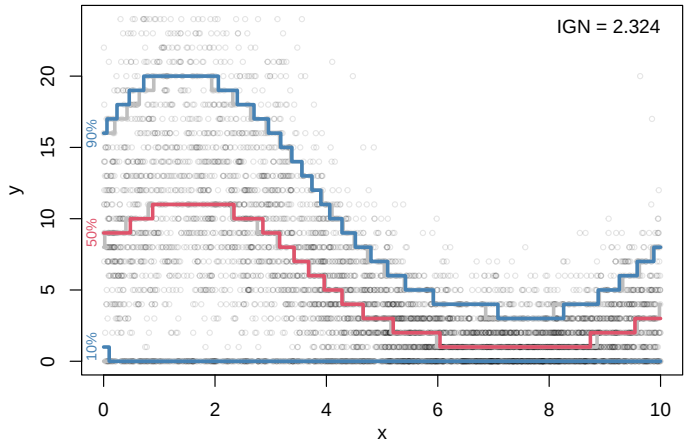
```
> b <- glm(Y ~ as.factor(theta) + x, data = df, family = binomial)
```

Transition models

Transitreg model: `transitreg(y ~ theta0 + te(theta, x, k = 15), ...)`

Transition models

Transitreg model: `transitreg(y ~ theta0 + te(theta, x, k = 15), ...)`

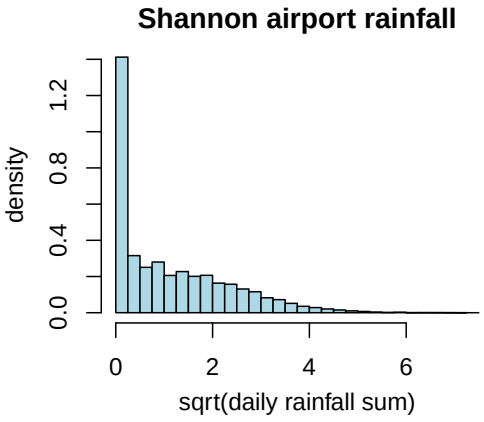


Extension for continuous response

Idea: Define intervals and generate pseudo-counts.

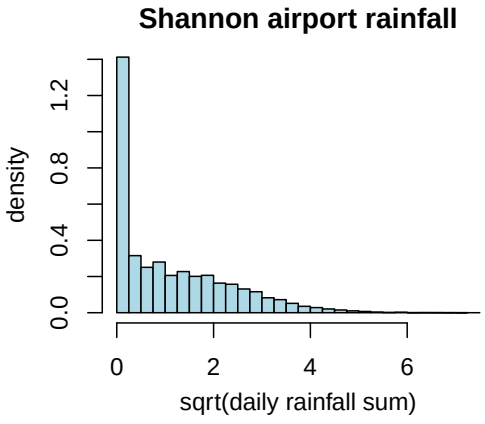
Extension for continuous response

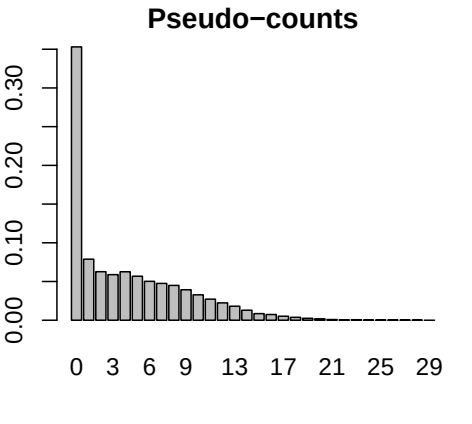
Idea: Define intervals and generate pseudo-counts.



Extension for continuous response

Idea: Define intervals and generate pseudo-counts.





Extension for continuous responses

Inspired by histogram binning, the response $y_i \in \mathbb{R}$, $i = 1, \ldots, n$ is divided into $m - 1$ (potentially non-uniformly spaced) intervals using

$$\zeta_1, \zeta_2, \ldots, \zeta_m,$$

where each interval $[\zeta_r, \zeta_{r+1})$ corresponds to a discrete count r. Each observation y_i is assigned a pseudo count $\tilde{y}_i$.

For a continuous response with CDF $F(y)$, the discretization approximates the probability of y_i falling into each interval as

$$P(\zeta_r \leq y_i < \zeta_{r+1}) = F(\zeta_{r+1}) - F(\zeta_r).$$

Extension for continuous responses

Probabilities are encoded as transformed counts $\tilde{y}_i$, so the transition model uses

$$P(\tilde{y}_i = r) = P(\zeta_r \leq y_i < \zeta_{r+1})$$

to approximate the discrete likelihood.

The model estimates the probability of transitions between counts

$$P(\tilde{y}_i > r \mid \tilde{y}_i \geq r, \mathbf{x}_i) = F(\eta_{ir}(\boldsymbol{\alpha}))$$

and recursively computes

$$P(\tilde{y}_i = r, \mathbf{x}_i) = P(\tilde{y}_i = r \mid \tilde{y}_i \geq r, \mathbf{x}_i) \prod_{s=0}^{r-1} P(\tilde{y}_i > s \mid \tilde{y}_i \geq s, \mathbf{x}_i).$$

Extension for continuous responses

- Let r denote the unique index such that $y_i \in [\zeta_r, \zeta_{r+1})$.
- For any value $y_i \in [\zeta_r, \zeta_{r+1})$, the CDF can be approximated by

$$\hat{F}(y_i) = \sum_{s=0}^{r-1} P(\tilde{y}_i = s) + \frac{y_i - \zeta_r}{\zeta_{r+1} - \zeta_r} P(\tilde{y}_i = r).$$

- The PDF can be approximated as

$$\hat{f}(y_i) = \frac{P(\tilde{y}_i = r)}{\zeta_{r+1} - \zeta_r}.$$

- The mean and variance are approximated using midpoints $c_r = \frac{\zeta_r + \zeta_{r+1}}{2}$

$$E[Y] = \sum_r c_r \, P(\tilde{y} = r), \quad \text{Var}(Y) = \sum_r c_r^2 \, P(\tilde{y} = r) - \left(\sum_r c_r \, P(\tilde{y} = r) \right)^2.$$

Will there be rain?

Will there be rain?

Rainfall Ireland

- Daily rainfall amounts from Met Éireann
- All available data 1977–2024 from 557 stations
- $\sim$3.9 million individual observations

Goal: Estimating spatio-temporal rainfall climatology.

Covariates: Day of the year and geographical location.

Model: Estimate square-root transformed daily precipitation amount using $m = 90$ pseudo-bins with left-censoring (censored at 0).

Will there be rain?

Model in pseudo-code:

$$\begin{aligned}\tilde{y} \sim \quad & \theta_0 + f(\theta) + f(\text{longitude}, \text{latitude}) + f(\theta, \text{longitude}, \text{latitude}) + \\ & f(\text{day}) + f(\theta, \text{day}) + f(\text{altitude}) + f(\theta, \text{altitude}) + f(\text{year})\end{aligned}$$

Will there be rain?

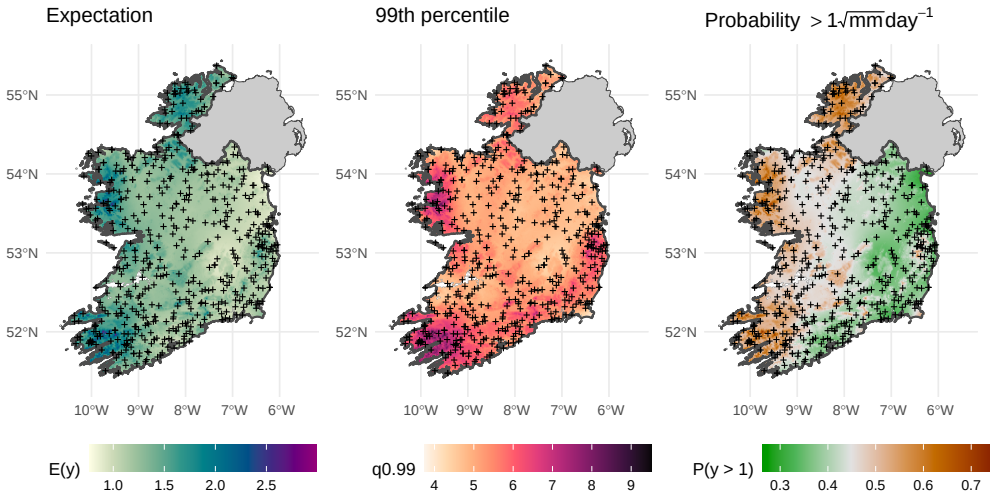
Model in pseudo-code:

$$\begin{aligned}\tilde{y} \sim \quad & \theta_0 + f(\theta) + f(\text{longitude}, \text{latitude}) + f(\theta, \text{longitude}, \text{latitude}) + \\ & f(\text{day}) + f(\theta, \text{day}) + f(\text{altitude}) + f(\theta, \text{altitude}) + f(\text{year})\end{aligned}$$

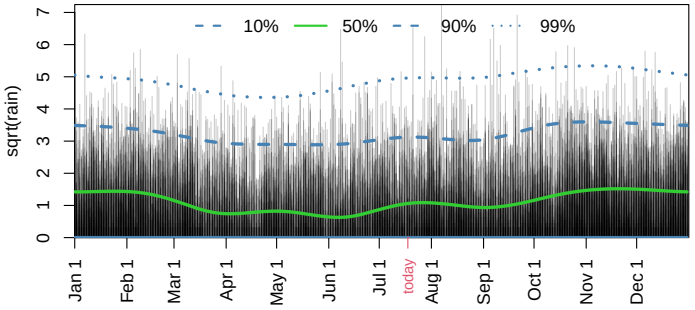
Model in *R*:

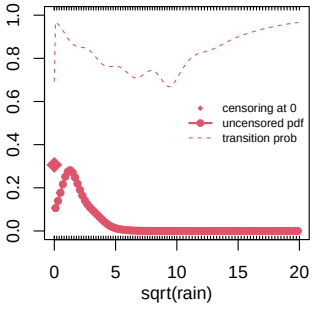
```
> ## Package available via https://retostauffer.r-universe.dev
> library("transitreg")

> ## Model formula
> f <- sqrt(rain) ~ theta0 + ti(theta, k = 12) +
+          ti(lon, lat) + ti(theta, lon, lat, k = c(8, 8, 8)) +
+          ti(day, bs = "cc") + ti(theta, day, bs = c("cr", "cc"), k = c(8, 10)) +
+          ti(alt, bs = "cr", k = 8) + ti(theta, alt, k = c(8, 8)) +
+          ti(year, k = 8)

> ## Breaks for pseudo-binning (non-uniformly distributed)
> bk  <- c(0, seq(0.2, 10, by = 0.2), seq(10.25, 20, by = 0.25))

> ## Estimating model
> mod <- transitreg(f, data = data, breaks = bk, censored = "left")
```

Will there be rain?

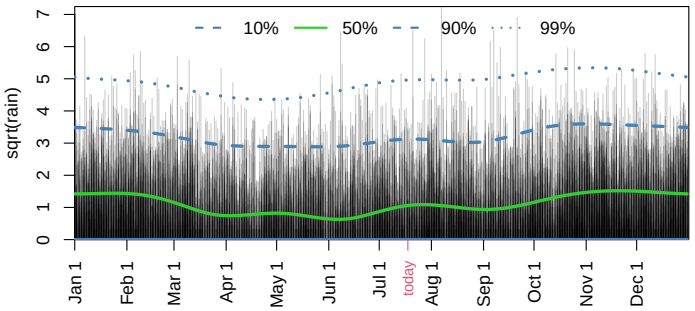


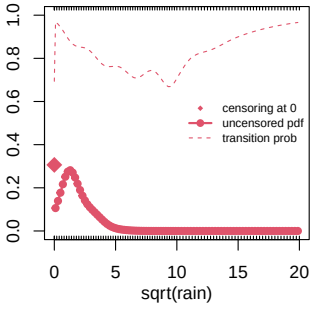
Will there be rain?





Will there be rain?





Climatological estimates for Shannon airport for today (July 17*th*):

- 69.7% probability of getting rain,
- 46.0% chance of more than 1 mm (0.4% for $>$ 31 mm),
- expected daily amount: 1.61 mm (mode: 2.25 mm).

Summary

Summary

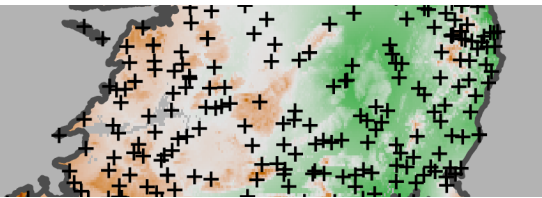
- Extended transition models to continuous responses
- Performs en-par with, e.g., ZANBI, without distributional assumptions
- Allows to retrieve mean, median, mode, variance, quantiles, . . .
- Proven to work well with large data sets

Software

Beta-release available – feel free to test and contribute:

- On GitHub: `https://github.com/retostauffer/transitreg`
- Install via: `https://retostauffer.r-universe.dev`
- Executable toy-example in the proceeding

transitreg



Thank you for your attention!

Reto Stauffer & Nikolaus Umlauf

www.uibk.ac.at/statistics

References I

Umlauf N, Stauffer R (2025). “transitreg: Flexible Transition Models for Probabilistic Learning.” *R* package version 0.2-0. https://retostauffer.r-universe.dev/transitreg, https://github.com/retostauffer/transitreg

Berger M, Tutz G (2021). “Transition Models for Count Data: a Flexible Alternative to Fixed Distribution Models.”, *Statistical Methods & Applications*, **30**, 1259–1283. doi:10.1007/s10260-021-00558-6

Kleiber C, Zeileis A (2016). “Visualizing Count Data Regressions Using Rootograms.” *The American Statistician*, **70**(3), 296–303. doi:10.1080/00031305.2016.1173590

Zeileis A, Kleiber C, Jackman S (2008). “Regression Models for Count Data in *R*.” *Journal of Statistical Software*, **27**(8), 1–25. doi:10.18637/jss.v027.i08

Zeileis A, Kleiber C (2024). “countreg: Count Data Regression.” *R* package version 0.3-0. https://R-Forge.R-project.org/projects/countreg/