Count Transformation Models: The cotram Package

Sandra Siegfried and **Torsten Hothorn**
Universität Zürich

Abstract

The **cotram** package offers a ready-to-use R implementation of count transformation models, providing a simple but flexible approach for the regression analysis of count responses arising from various, and possibly complex, data-generating processes. In this unified maximum-likelihood framework count models can be formulated, estimated, and evaluated easily. Specific models in the class can be flexibly customised by the choice of the link function and the parameterisation of the transformation function. Interpretation of explanatory variables in the linear predictor is possible at the scales of the discrete odds ratio, hazard ratio, or reverse time hazard ratio, or as conditional mean of transformed counts. The implemented methods for the model class further provide simple tools for model evaluation. The package simplifies the use of transformation models for modelling counts, while ensuring appropriate settings for count data specifically. Extension to the formulated models can be made by the inclusion of response-varying effects, strata-specific transformation functions, or offsets, based on the underlying infrastructure of the **tram** and **mlt** R add-on packages, which further ensure the correct handling of the likelihood for censored or truncated observations.

Keywords: conditional distribution function, conditional quantile function, count regression, deer-vehicle collisions, transformation model.

1. Introduction

Count transformation models are a novel model class, offering a flexible and data-driven approach to regressing count data. The diverse set of models in the class, as proposed and discussed in Siegfried and Hothorn (2020), are tailored to analyse count responses from various underlying data-generating processes in a unified maximum-likelihood framework. The R add-on package **cotram** features the implementation of the proposed model class, providing a simple and user-friendly interface to fit and evaluate count transformation models. The package is built using the general infrastructure of the R add-on packages **tram** (Hothorn and Barbanti 2022) and **mlt** (Hothorn 2020, 2022) for likelihood-based inference and further extensions to the implemented model specifications.

Count transformation models arise from the direct modelling of the conditional discrete distribution function capturing changes governed by a linear predictor $\boldsymbol{x}^\top\boldsymbol{\beta}$. The models in the class can be represented by the general formulation of the conditional distribution function for any y

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y \mid \boldsymbol{x}) = \mathbb{P}(Y \leq y \mid \boldsymbol{x}) = F\left(\alpha\left(\lfloor y \rfloor\right) - \boldsymbol{x}^\top\boldsymbol{\beta}\right), \quad y \in \mathbb{R}^+ \tag{1}$$

with specific models originating from the choice of the different link functions $g = F^{-1}$. The model class includes models with a logit, complementary log-log (cloglog), log-log, and probit link and thus offers interpretability of the linear predictor at various scales. The framework allows evaluating and interpreting the models in a discrete way, while using a computationally attractive, low-dimensional, continuous representation. The models are designed to simultaneously estimate the transformation function α and the regression coefficients $\boldsymbol{\beta}$ optimising the exact discrete log-likelihood. Simultaneous estimation of the parameters (developed by Hothorn *et al.* 2018) is performed based on the underlying infrastructure provided by the **mlt** package (Hothorn 2022).

All models in the class (1) can be fitted using the general function call

```
R> cotram(<formula>, method = <link>, ...)
```

with `<formula>` being any R formula featuring counts as the response and the right hand side as series of terms determining a linear predictor. The specific models in the class can be fitted by choosing one of the link functions for `method = <link>`. The set of models specified by the different link functions and the interpretation of the explanatory variables in the linear predictor $\boldsymbol{x}^\top\boldsymbol{\beta}$ are outlined in more detail below.

The package further offers `predict()` and `plot()` functions to assess and illustrate the estimated linear predictor, conditional distribution and density function, quantiles and the estimated transformation function, both as step-functions and continuously (setting `smooth = TRUE`). Functionalities for model interpretation and evaluation, such as `summary()`, `coef()`, `confint()`, and `logLik()` are available in this framework.

2. Discrete Hazards Cox Count Transformation Model

The count transformation model with complementary log-log link function $g = F^{-1}$ (`method = "cloglog"`) offers a discrete version of the Cox proportional hazards model with fully parameterised transformation function α and interpretation of the linear predictor as discrete hazard ratio. The model explains the effects of the exponentiated linear predictor $\exp(-\boldsymbol{x}^\top\boldsymbol{\beta})$ on observed counts as multiplicative changes in discrete hazards $\mathbb{P}(Y = y \mid Y \geq y, \boldsymbol{x})$, comparing the conditional cumulative hazard function $\log(1 - F_{Y|\boldsymbol{X}=\boldsymbol{x}})$ with the baseline cumulative hazard function $\log(1 - F_Y)$, with $\boldsymbol{x}^\top\boldsymbol{\beta} = 0$.

Using the deer-vehicle collisions data from Hothorn *et al.* (2015), we can fit the Cox count transformation model to the roe deer-vehicle collision counts per day, recorded from 2002 to 2011 in Bavaria, Germany, and obtain the estimated multiplicative temporal changes in "risk" as discrete hazards. The `tvar` variables are sin-cosine transformed times (see Hothorn *et al.* 2015).

```
R> mod_cloglog <- cotram(DVC ~ year + weekday + tvar1 + tvar2 + tvar3 +
+                         tvar4 + tvar5 + tvar6 + tvar7 + tvar8 + tvar9 +
+                         tvar10 + tvar11 + tvar12 + tvar13 + tvar14 +
+                         tvar15 + tvar16 + tvar17 + tvar18 + tvar19 + tvar20,
+                         data = df, method = "cloglog", prob = c(0, .9))
R> logLik(mod_cloglog)

'log Lik.' -16547.59 (df=42)
```

To assess how the risk varies across days and seasons, we can now compute the estimated discrete hazards ratio for each day of the year, based on the predictor values of the year 2011. The results, shown in Figure 1, illustrate the changes in the hazard ratios, relative to baseline on January 1st (note that we plot $\exp(\boldsymbol{x}(\text{day})^\top \boldsymbol{\beta} - \boldsymbol{x}(\text{2011–01–01})^\top \boldsymbol{\beta})$, such that large values correspond to large number of collisions and thus higher risk).

```
R> nd <- model.frame(mod_cloglog)[which(df$year == "2011"), -1]
R> nd$day <- df[which(df$year == "2011"), "day"]
R> nd$weekday <- factor("Monday", levels = levels(nd$weekday))
```

```
R> fit_cloglog <- predict(mod_cloglog, type = "lp", newdata = nd) -
+    predict(mod_cloglog, type = "lp", newdata = nd)[1]
R> xyplot(exp(fit_cloglog) ~ day , data = cbind(nd, fit_cloglog),
+         ylab = "Hazard ratio", xlab = "Day of year", panel = panel)
```

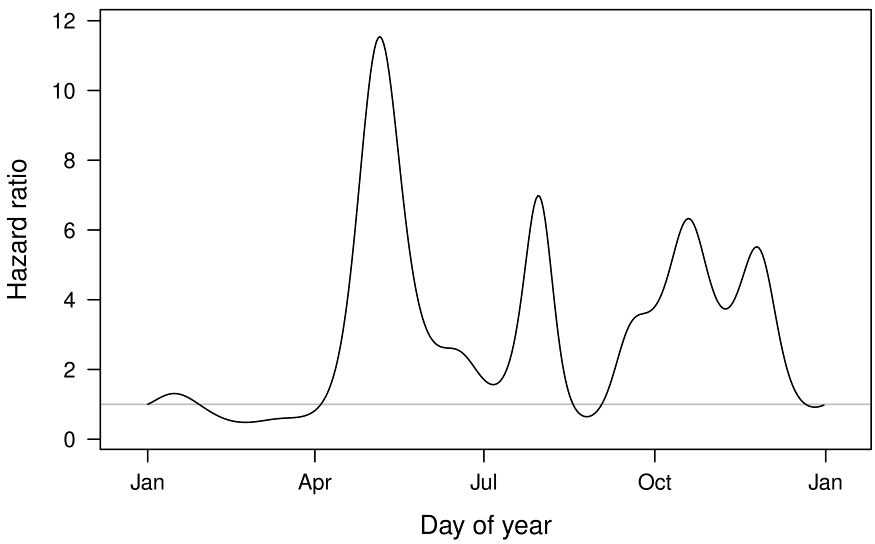

Figure 1: Deer-vehicle collisions. Temporal changes in risk for deer-vehicle collisions across the year as discrete hazard ratios estimated by model `mod_cloglog` with reference: January 1st. The curve indicates, that the hazard ratio is increased associated with animal activity due to search for new habitats and food resources in April and rut season in July and August. The peak in October does not seem to have a clear explanation in terms of increased roe deer activity.

3. Logistic Count Transformation Model

Odds ratios are often used in practice to compare two different configurations of the set of explanatory variables $\boldsymbol{x}$. Conveniently, for the class of count transformation models we can obtain the estimated effects on this scale by specifying a logit link. The exponentiated linear predictor $\exp(-\boldsymbol{x}^\top \boldsymbol{\beta})$ estimated by such a logistic count transformation model can be interpreted as odds ratio

$$\frac{\mathbb{P}(Y \leq y \mid \boldsymbol{x})}{\mathbb{P}(Y > y \mid \boldsymbol{x})} = \frac{\mathbb{P}(Y \leq y)}{\mathbb{P}(Y > y)} \exp(-\boldsymbol{x}^\top \boldsymbol{\beta}),$$

comparing the conditional odds of a configuration $\boldsymbol{x}$ with the baseline odds $F_Y/1-F_Y$ (with $\boldsymbol{x}^\top \boldsymbol{\beta} = 0$). The response-varying intercept $\alpha(y)$ cancels out in the odds ratio, resulting in an estimate, which can be interpreted simultaneously across all cut-offs y.

To explain the temporal risk of roe deer-vehicle collisions on the odds ratio scale, the only modification to the model formulation of Section 2 required, is the link specification in the function call as `method = "logit"`.

```
R> mod_logit <- cotram(DVC ~ year + weekday + tvar1 + tvar2 + tvar3 +
+                        tvar4 + tvar5 + tvar6 + tvar7 + tvar8 + tvar9 +
+                        tvar10 + tvar11 + tvar12 + tvar13 + tvar14 +
+                        tvar15 + tvar16 + tvar17 + tvar18 + tvar19 + tvar20,
+                        data = df, method = "logit", prob = c(0, .9))
R> logLik(mod_logit)

'log Lik.' -16319.55 (df=42)
```

Comparison of the log-likelihoods of the fitted model and the Cox count transformation model from Section 2 shows almost matching values, with a slight improvement in model fit, when replacing the cloglog with the logit link.

We now could further assess the effect of the factor `year` on the deer-vehicle collision counts by computing the odds ratios (small values correspond to moving the distribution to the right and thus to larger number of collisions) along with the likelihood-based confidence intervals.

```
R> years <- grep("year", names(coef(mod_logit)), value = TRUE)
R> coef <- exp(-coef(mod_logit)[years])
R> ci <- exp(-confint(mod_logit)[years,])
R> round(cbind(coef, ci), 3)

          coef 2.5 % 97.5 %
year2003 0.595 0.765  0.463
year2004 0.337 0.433  0.262
year2005 0.306 0.395  0.238
year2006 0.407 0.526  0.315
year2007 0.156 0.203  0.120
year2008 0.096 0.124  0.075
year2009 0.105 0.136  0.081
year2010 0.090 0.116  0.069
year2011 0.097 0.126  0.075
```

Plotting the estimated conditional distribution functions of model `mod_logit` in Figure 2, demonstrates the linear shift in $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ guided by the different levels of the factor `year`.

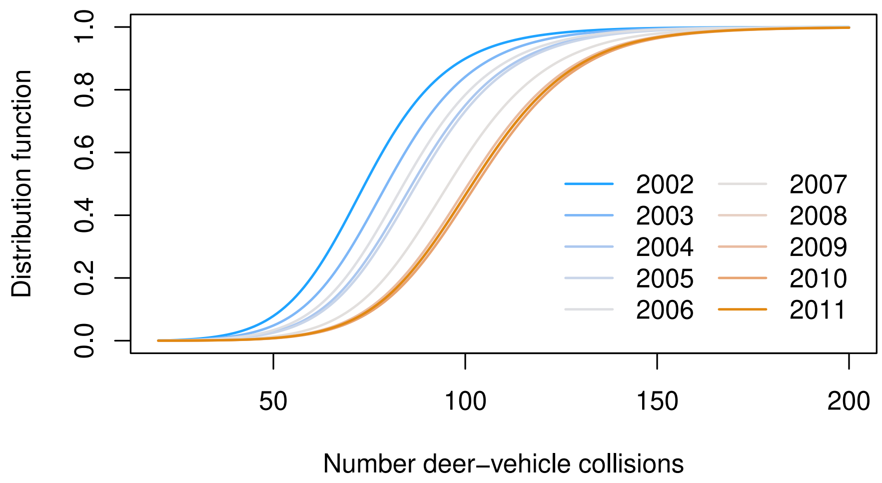

Figure 2: Deer-vehicle collisions. Illustration of the estimated conditional distribution functions of each year between 2002 and 2011.

4. Discrete Reverse Time Hazards Count Transformation Model

Specifying a count transformation model with log-log link $g = F^{-1}$ we get the model formulation

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y \mid \boldsymbol{x}) = \mathbb{P}(Y \leq y \mid \boldsymbol{x}) = \exp\left(-\exp\left(\alpha(\lfloor y \rfloor) - \boldsymbol{x}^\top \boldsymbol{\beta}\right)\right)$$

with interpretation of the linear predictor $\exp(\boldsymbol{x}^\top \boldsymbol{\beta})$ as discrete reverse hazard ratio with multiplicative changes in $\log(F_Y)$. To fit the model, we again only need to adapt the model specification in terms of the link function by setting `method = "loglog"`.

```
R> mod_loglog <- cotram(DVC ~ year + weekday + tvar1 + tvar2 + tvar3 +
+                         tvar4 + tvar5 + tvar6 + tvar7 + tvar8 + tvar9 +
+                         tvar10 + tvar11 + tvar12 + tvar13 + tvar14 +
+                         tvar15 + tvar16 + tvar17 + tvar18 + tvar19 + tvar20,
+                         data = df, method = "loglog", prob = c(0, .9))
R> logLik(mod_loglog)
```

```
'log Lik.' -16436.54 (df=42)
```

For further assessment we could evaluate the discrete conditional density of a set of $\boldsymbol{x}$. Figure 3 illustrates the estimated density function in terms of the predictor values recorded on 2002–01–01 along with the actually observed deer-vehicle collision count.

```
R> nd <- model.frame(mod_loglog)[1,]
```

```
R> plot(mod_loglog, type = "density", newdata = nd, q = 0:150, col = col,
+       xlab = "Number of deer-vehicle collisions", ylab = "Density function")
R> abline(v = nd$DVC)
```

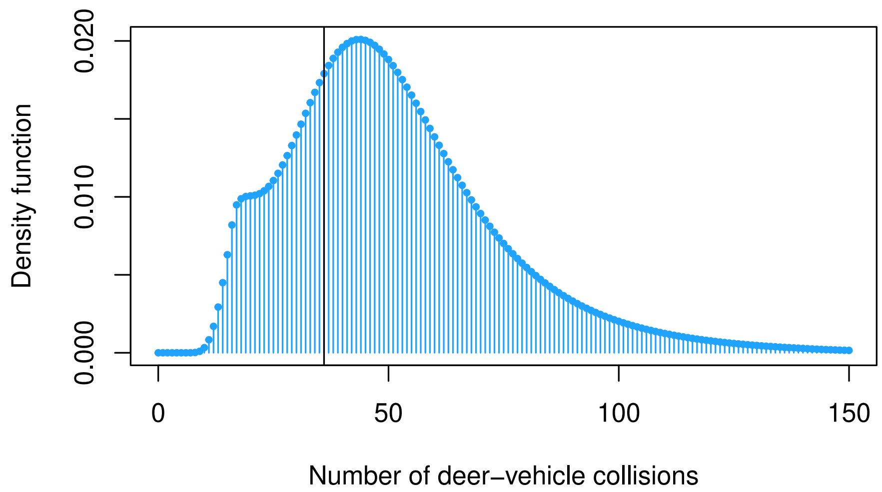

Figure 3: Deer-vehicle collisions. Estimated discrete density function for model `mod_loglog` with the actual observed count shown as vertical black line.

5. Probit Count Transformation Model

When applying a count transformation model with a probit link (`method = "probit"`) we can interpret the estimated effects as changes in the conditional mean of the transformed counts $\mathbb{E}(\alpha(y) \mid \boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$. This interpretation is the same, as obtained from fitting a normal linear regression model on a priori transformed counts, by *e.g.* a log or square-root transformation. However, for the probit count transformation model, as implemented in the **cotram** package, the transformation of the response y was not heuristically chosen, as in a least-squares approach, but estimated from data by optimising the exact count log-likelihood.

```
R> mod_probit <- cotram(DVC ~ year + weekday + tvar1 + tvar2 + tvar3 +
+                       tvar4 + tvar5 + tvar6 + tvar7 + tvar8 + tvar9 +
+                       tvar10 + tvar11 + tvar12 + tvar13 + tvar14 +
+                       tvar15 + tvar16 + tvar17 + tvar18 + tvar19 + tvar20,
+                       data = df, method = "probit", prob = c(0, .9))
R> logLik(mod_probit)
```

```
'log Lik.' -16311.51 (df=42)
```

A simple tool in this framework to check, whether, for example a log transformation, would have been appropriate, is to inspect the estimated transformation function $\alpha(y)$.

```
R> nd <- model.frame(mod_probit)[1, ]
R> trafo_probit <- predict(mod_probit, type = "trafo",
+                          newdata = nd, smooth = TRUE)
```

The variability associated with the estimated transformation functions can be further assessed by an asymptotic confidence band.

```
R> cb_probit <- confband(mod_probit, type = "trafo",
+                        newdata = nd, smooth = TRUE)
```

The results are shown in Figure 4 for both the transformation function and the conditional distribution function.

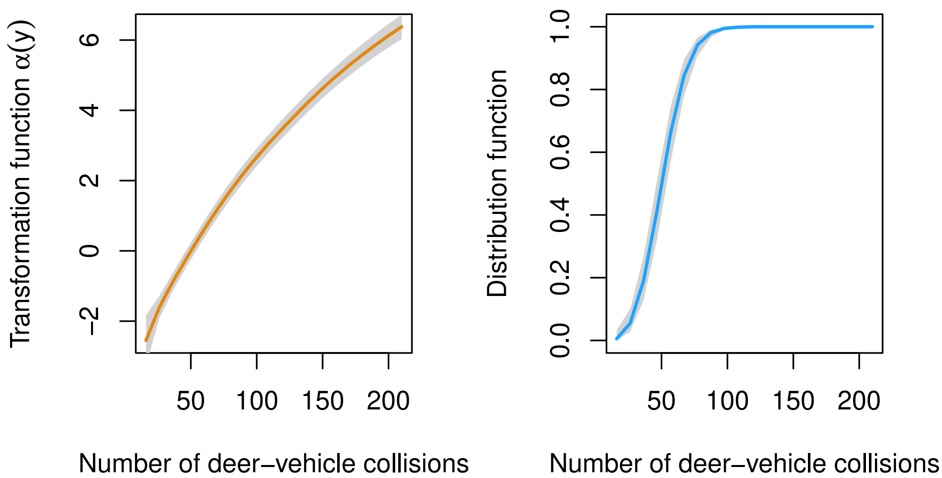

Figure 4: Deer-vehicle collisions. Baseline transformation α and conditional distribution function estimated by the model `mod_probit` along with 95% asymptotic confidence bands.

6. Summary

The implemented models and methods in the **cotram** package offer a unified framework for users to fit and evaluate transformation models for counts, by ensuring the correct handling of the discrete nature of the data. Simplifying the modelling procedure, the models are parameterised under general and empirically tested settings, eliminating the need for overly complicated model specifications.

References

Hothorn T (2020). "Most Likely Transformations: The **mlt** Package." *Journal of Statistical Software*, **92**(1), 1–68. doi:10.18637/jss.v092.i01.

Hothorn T (2022). ***mlt**: Most Likely Transformations.* R package version 1.4-2, URL https://CRAN.R-project.org/package=mlt.

Hothorn T, Barbanti L (2022). ***tram**: Transformation Models.* R package version 0.7-1, URL https://CRAN.R-project.org/package=tram.

Hothorn T, Möst L, Bühlmann P (2018). "Most Likely Transformations." *Scandinavian Journal of Statistics*, **45**(1), 110–134. doi:10.1111/sjos.12291.

Hothorn T, Müller J, Held L, Möst L, Mysterud A (2015). "Temporal Patterns of Deer-vehicle Collisions Consistent with Deer Activity Pattern and Density Increase but not General Accident Risk." *Accident Analysis & Prevention*, **81**, 143–152. doi:10.1016/j.aap.2015.04.037.

Siegfried S, Hothorn T (2020). "Count Transformation Models." *Methods in Ecology and Evolution*, **11**(7), 818–827. doi:10.1111/2041-210X.13383.

Affiliation:

Sandra Siegfried and Torsten Hothorn
Institut für Epidemiologie, Biostatistik und Prävention
Universität Zürich
Hirschengraben 84, CH-8001 Zürich, Switzerland
sandra.siegfried@uzh.ch