EX

Herman Persson

null

Introduktion

Okänd inflation i data är inte det mest vanligt förekommande problemet då det ofta kan försummas eller helt förebyggas. Men om en existerande inflation inte uppmärksammas kan det få stora konskvenser på en statistisk analys

Ett möjligt förekommande fel vid populationsskattning är att en redan observerad individ fel identifieras som en ny individ. Om detta fel förekommer så kommer den slutgiltiga datan innehålla falska observationer (spöken) som ger en inflation hos det antal individer som endast observerats en gång.

Metod

In the beginning of the chapter we will mainly focus on the poisson model which assumes that there is no individual heterogeneity.

The poisson model

Truncated distributions

Suppose we have a population of size N from which we receive samples and that we are able to distinguish from which individual each sample is from. If there is no individual heterogeneity and the probability that a sample comes from an individual is the same for all individuals in the population, then the number of samples per individual of the observed individuals will be zero-truncated poisson (positive poisson) distributed with probability mass function

$$p_+(x;\lambda) = \frac{p(x;\lambda)}{1-p(0;\lambda)} = \frac{\lambda^x e^{-\lambda}}{(1-e^{-\lambda})x!} = \frac{\lambda^x}{(e^{\lambda}-1)x!} \tag{1}$$

where $p(x;\lambda)$ is the probability mass function of the base distribution, poisson with parameter lambda. Now lets assume that for some reason a sample from one individual can be incorrectly identified as a sample from a not previously observed and non existent individual (ghost) with some probability p. For all p bigger than zero this would result in an increased number of observed ones, so called one-inflation, as well as a decreased number of expected observations per individual. To incorporate these conditions into our model we consider the new probability mass function

$$p_{+1}(x;\lambda,p) = \begin{cases} (1-\omega) + \omega p_+(x,\theta) & \text{for } x = 1, \\ \omega p_+(x,\theta) & \text{for } x > 1. \end{cases} \tag{2}$$

In this distrubution ω adjusts the extra mass at $x = 1$ and $\theta = \lambda(1-p)$ is adjusted density parameter with consideration to the reduced number of expected observation per individual. The expected number of extra ones generated by each individual will be $p\lambda$ and therefore $\omega = 1/(1+p\lambda)$. p_{+1} is the probability mass function of the one-inflated positive poisson model (OIPP). If we denote the number of observations with value i as f_i and the highest observed value as m we get the likelihood function

$$L_{p_{+1}}(\lambda, p; x) = [(1-\omega) + p_+(1,\theta)]^{f_1} \prod_{i=2}^{m} [\omega p_+(x_i,\theta)]^{f_i}.$$

If we are uninterested in the exact values of p and λ another useful way to look at data would be to ignore all the individuals observed only once and use a one-truncated positive poisson model (OTPP). By using this approach we will increase the uncertainty of our estimates but avoid some bias which can occur when using the OIPP model. In the case of the OTPP model we get the probability density function

$$p_{++}(x;\theta) = \frac{p(x;\theta)}{1-p(0;\theta)-p(1;\theta)} = \frac{\theta^x}{(e^\theta - \theta - 1)x!}, \qquad x = 2, 3, ...$$

with the corresponding likelihood function

$$L_{p_{++}}(\theta; x) = \prod_{i=2}^{m} p_{++}(x_i,\theta)^{f_i}.$$

The third and final poisson based distribution we will look at is positive poisson model (PP). The reason for this is for comparisons against our other two models, mainly to look at the effect on population estimates when existing inflation for some reason is not taken into account and to see how well our other models compare when there is no inflation, ie how well would our other models function as a precaution against possible inflation. The PP models probability mass function can be seen in (1) and its corresponding likelihood is

$$L_{p_+} = \prod_{i=1}^{m} p_+(x_i,\theta)^{f_i}$$

Negative binomial model

Base distrubution

By expanding our poisson model to a negative binomial model we will be able to account for individual heterogeneity. We can introduce individual heterogeneity for individual i by letting the parameter of its poisson distrubution $\theta_i = \lambda_i(1-p)$ be distributed according to a Gamma distribution with shape parameter $\alpha = k\lambda$ and rate parameter $\beta = k/(1-p)$ where $k > 0$, ie

$$\theta_i = \lambda_i(1-p) \sim \Gamma(k\lambda, \frac{k}{1-p}).$$

This means that $E[\theta_i] = \lambda(1-p)$ and $Var[\theta_i] = \lambda(1-p)^2/k$ and we are able to account for over dispersion and individual heterogenity by adjusting k. Due to ... we know that

$$X_i \sim \Gamma(r, \frac{1-q}{q}) \qquad \Rightarrow \qquad Po(X_i) \overset{d}{=} NBin(r,q)$$

With some fairly easy calculations we then get

$$\theta_i \sim \Gamma(k\lambda, \frac{k}{1-p}) \qquad \Rightarrow \qquad Po(\theta_i) \overset{d}{=} NBin(k\lambda, \frac{1-p}{1-p+k}). \tag{3}$$

Notice that we can allow all $k > 0$ by using the extended negative binomial distribution which extends the binomial coefficient to all real-values by using the gamma function.

$$NBin(r,q) \sim \frac{\Gamma(x+r)}{x!\Gamma(r)}(1-q)^x q^r, \qquad r > 0, \quad 0 \leq q \leq 1, \quad x = 0, 1, ...$$

Truncated distributions

Let us denote the probability density function of our base distribution in (4) as $g(x; \lambda, k, p)$. Similar to the poisson model case we are going to use three different distributions for estimating the total population, the zero-truncated one-inflated negative binomial distribution (ZTOINB), zero and one-truncated negative binomial distribution (ZOTNB) and the zero-truncated negative binomial distribution (ZTNB). In the case of the last mentioned ZTNB distribution we assume that $p = 0$ and with size parameter $r = k\lambda$ and probability parameter $q = (1-p)/(1-p+k)$ get the probability density function

$$g_+(x; \lambda, k, p \mid p = 0) = \frac{g(x; \lambda, k, p \mid p = 0)}{1 - g(0; \lambda, k, p \mid p = 0)} = \frac{\frac{\Gamma(x+k\lambda)}{x!\Gamma(k\lambda)}\left(1 - \frac{1}{1+k}\right)^x \left(\frac{1}{1+k}\right)^{k\lambda}}{1 - \left(\frac{1}{1+k}\right)^{k\lambda}}, \qquad x = 1, 2, ...$$

and in the ZOTNB model we get

$$g_{++}(x; \lambda, k, p) = \frac{g(x; \lambda, k, p)}{1 - g(0; \lambda, k, p) - g(1; \lambda, k, p)} = \frac{\frac{\Gamma(x+k\lambda)}{x!\Gamma(k\lambda)}\left(1 - \frac{1-p}{1-p+k}\right)^x \left(\frac{1-p}{1+k-p}\right)^{k\lambda}}{1 - \left(\frac{1-p}{1+k-p}\right)^{k\lambda} - k\lambda\left(1 - \frac{1-p}{1+k-p}\right)\left(\frac{1-p}{1+k-p}\right)^{k\lambda}}, \qquad x = 2, 3, ...$$

with p in the interval $[0, 1]$. In the case of the ZTOINB model we similarly to the OIPP model use our corresponding zero truncated distribution with the addition of ω to adjust the extra mass at 1. So we get the probability density function

$$g_{+1}(x; \lambda, k, p) = \begin{cases} (1-\omega) + \omega g_+(x; \lambda, k, p) & \text{for } x = 1, \\ \omega g_+(x; \lambda, k, p) & \text{for } x > 1. \end{cases}$$

where $\omega = 1/(1 + \lambda(1-p))$ just as the poisson model since $E[g_+] = \lambda(1-p)$ (see proof in appendix).

Population estimation

For each one of our models we want to create an estimate of the total population N, which includes all the unobserved individuals. We denote the number of observed individuals (ghosts included) as n and estimate the total population in the zero-truncated models which ignore inflation as

$$\hat{N}_{f+} = \frac{n}{1 - f(0)}.$$

Where f is our base distrubution and f_+ is the corresponding zero-truncated distrubution which ignores inflation. In the one-truncated models we use a similar estimator

$$\hat{N}_{f++} = \frac{n - f_1}{1 - f(0) - f(1)},$$

where f_1 as before is the number of individuals observed once. In the case of the zero-truncated one-inflaited models we can use the estimator

$$\hat{N}_{f_{+1}} = \frac{n - f_1 + \hat{n}_1}{1 - f(0)} \tag{4}$$

where $\hat{n}_1$ is the estimated number of real individuals observed once, ie not including ghosts. In (2) the proportion of real individuals observed once are $\omega p_+(1, \theta)$ and the proportion of ghosts are $1 - \omega$. Our estimator for $\hat{n}_1$ is therefor

$$\hat{n}_1 = f_1 \left(\frac{f_{+1}(1) - \mathbb{P}(\text{"ghost"})}{f_{+1}(1)} \right) = f_1 \frac{\hat{\omega} f_+(1)}{(1 - \hat{\omega}) + \hat{\omega} f_+(1)}$$

which combined with (3) then gives us the estimator of N as

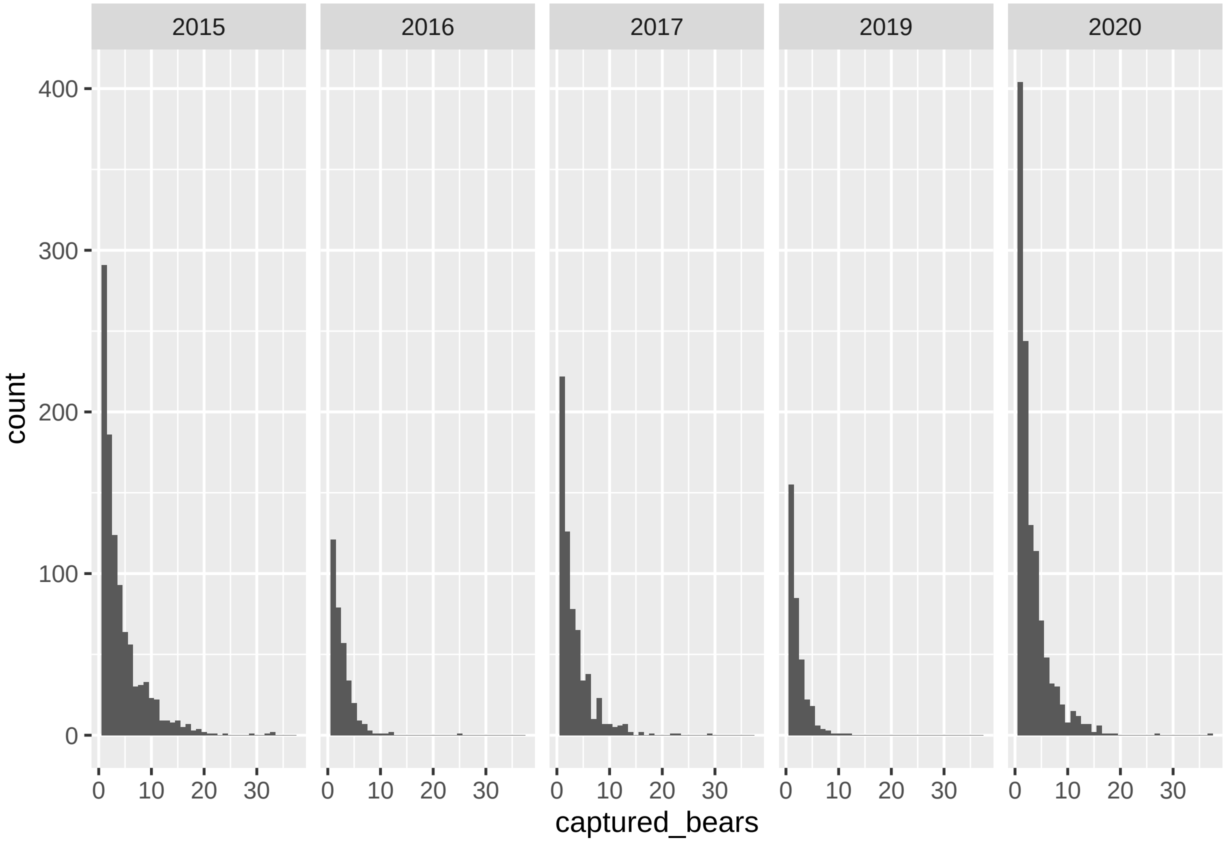
$$\hat{N}_{f_{+1}} = \frac{n - f_1 \left(1 - \frac{\hat{\omega} f_+(1)}{(1 - \hat{\omega}) + \hat{\omega} f_+(1)} \right)}{1 - p(0; \hat{\theta})}.$$

Simulation

By doing a simulation study we will be able to analyse the bias of our estimators for different values of our model parameters. As an estimate for all parameters in our models we will be using the ML-estimates, and since some of the likelihoods cant be solved algebraically they will be maximized numerically. We will create confidence intervals for our estimators using our estimates arranged in order of magnitude.

Application on bear data

The regions of Sweden which are habituated by brown bears are divided into four parts which are have all been monitored by the department of Environmental Research and Monitoring at the Swedish Museum of Natural History (NRM) since 2015. In each region, hunters are assigned test equipment every five years (maximum one region each year) to collect bear scat-samples, which are then sent to the NRM. These samples are used for genetic identification and NRM can then build up a database of bear individuals in the region. The number of times each bear was observed can be seen in figure 1.



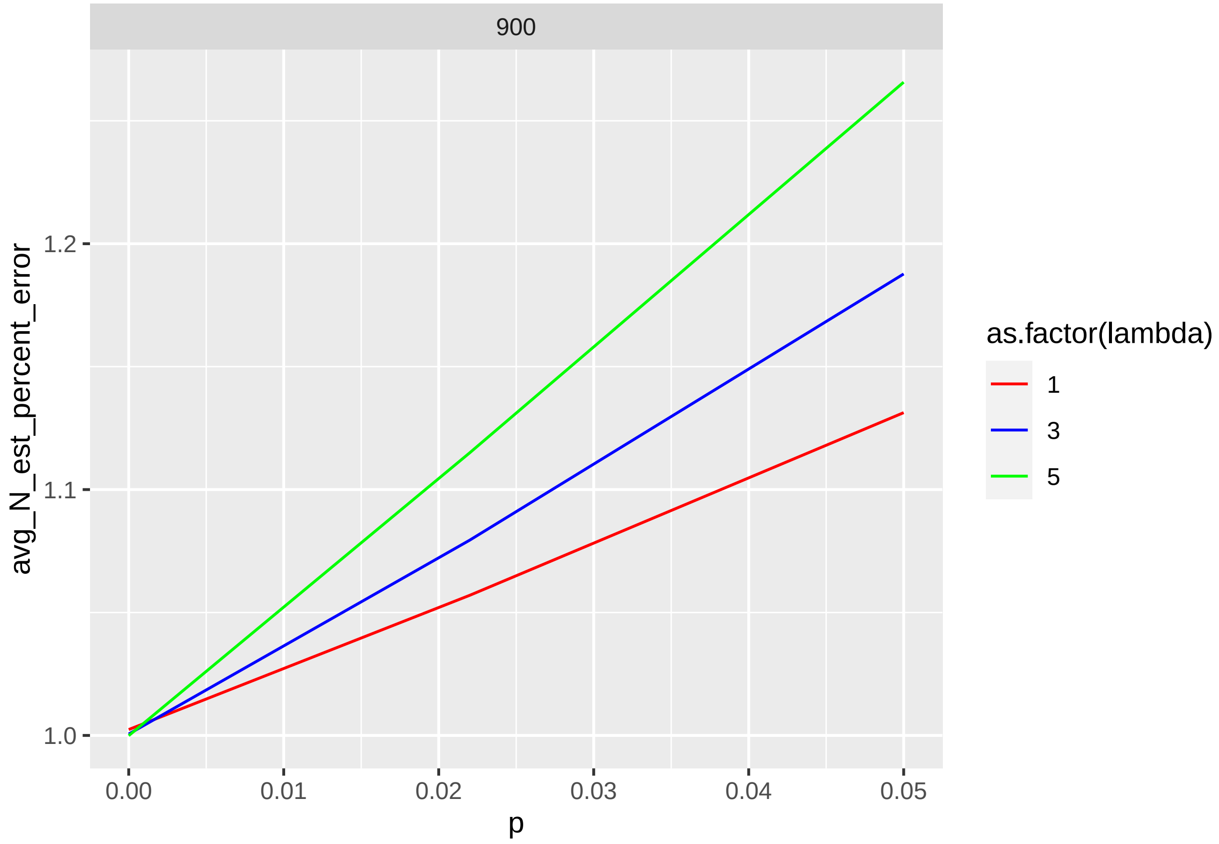
Just by observing the distrubtuion of capture times by year we are able to see that data look evenly distributed over the years. We can also see that most bears were observed just a few number of times and that the most common number of observations is one, for all five years. With our negative binomial models we will be able to see if there seems to be some kind of one inflaions happening aswell as be able to estimate the total number of bears each year.

Resultat

Simultaions

Positive poisson model

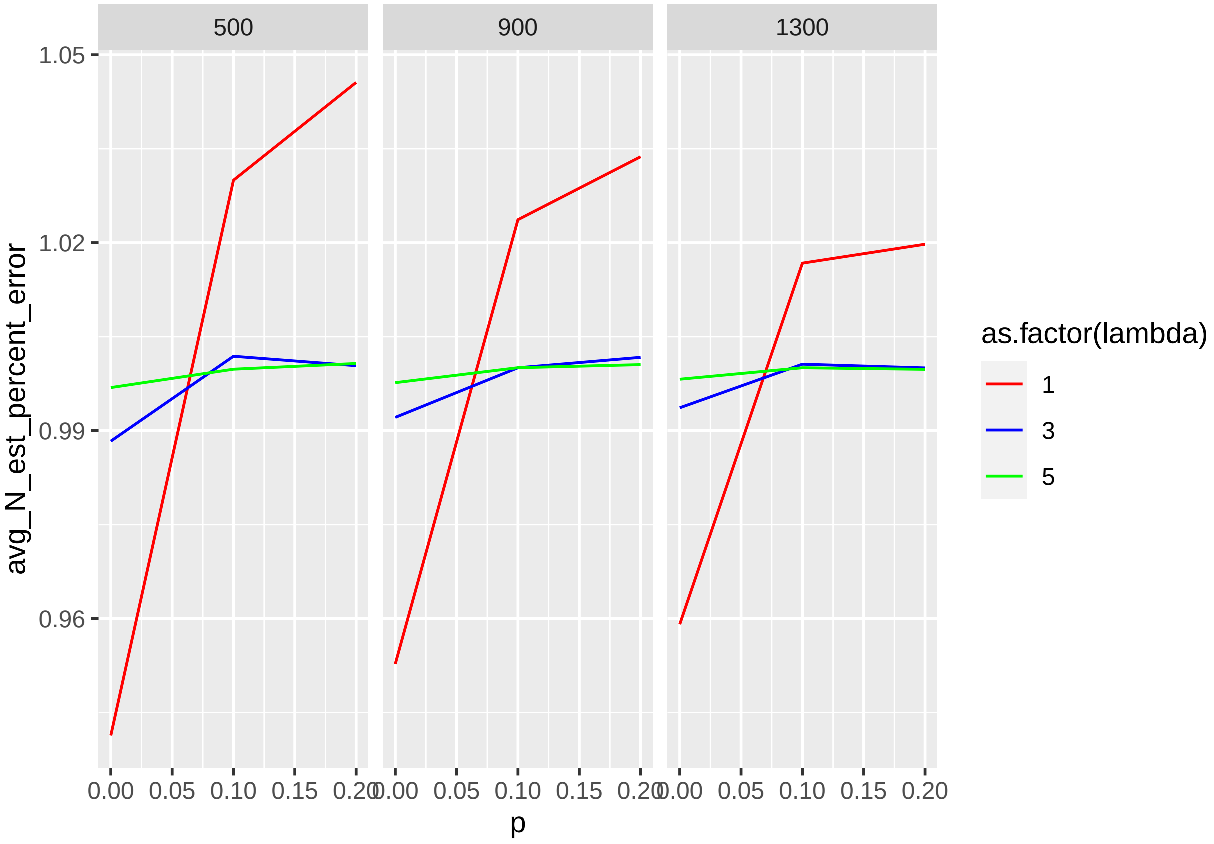
To begin we simulate data from the one-inflated positive poisson (OIPP) distribution and estimate the population with the positive poisson (PP) model, which does not take into account that data is inflated. We simulate 1000 times from all combinations for 3 different population sizes, inflation parameter values and underlying distrubution paramers. For each simulation the undelying population size is estimated using the PP-model population estimator N_{p_+} and the results can been seen in figure . . . .

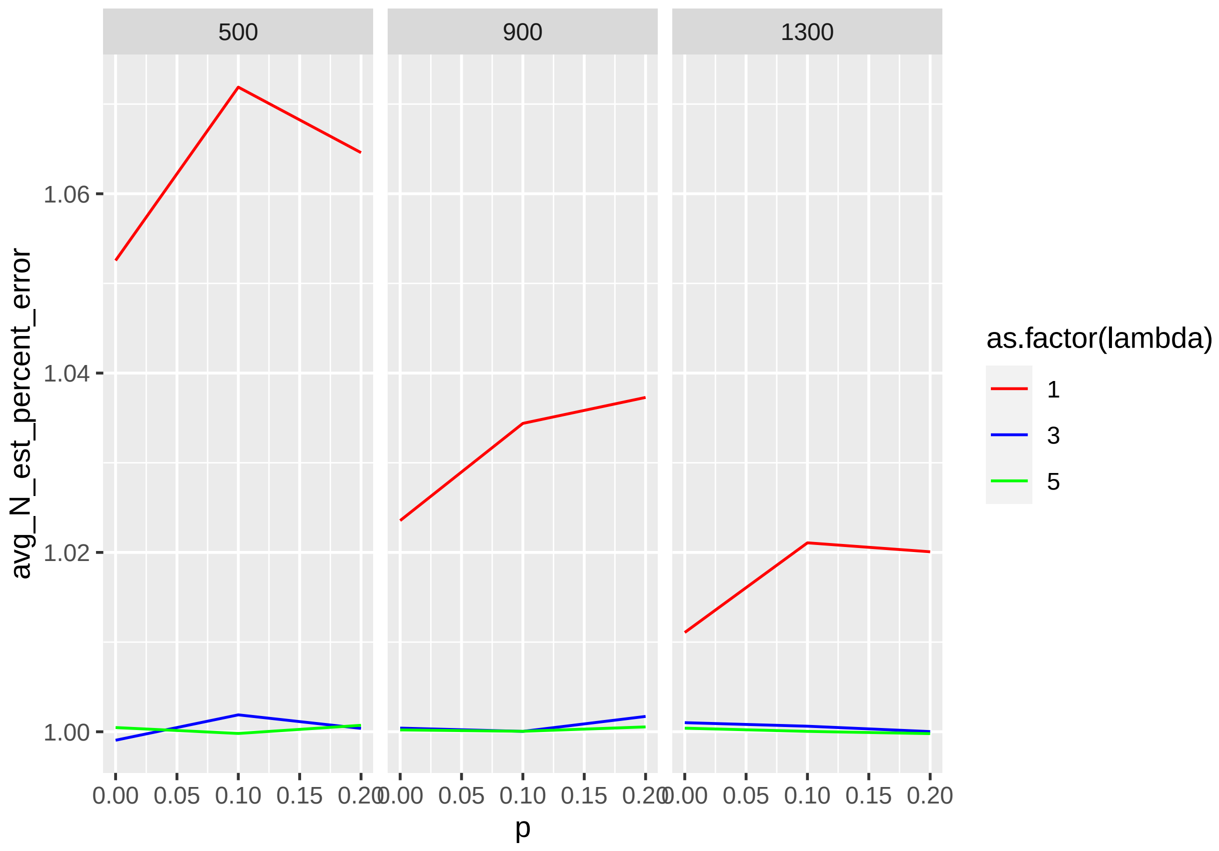


p	N	lambda	type	N_mean_est	conf_90	conf_99	sd
0.0	900	3	OIPP	892.907	[909, 870]	[914, 853]	12.51211
0.0	900	3	OTPP	900.367	[934, 870]	[949, 853]	19.87580
0.2	900	3	OIPP	901.537	[953, 854]	[980, 834]	30.44190
0.2	900	3	OTPP	901.540	[953, 854]	[980, 834]	30.44193

One truncated and inflated positive poisson models

To compare the population estimators of our one-inflated positive poisson (OIPP) and one-truncated positive poisson (OTPP) model we simulate 1000 times from all combinations for 3 different population sizes, inflation parameter values and underlying distrubution paramers. For each simulation the undelying population size is estimated using both the OIPP population estimator N_{p+1} and the OTPP population estimator N_{p++}. The results of our simulations can be seen in figure ....


500
900
1300
1.05
1.02
0.99
0.96
avg_N_est_percent_error
0.00
0.05
0.10
0.15
0.20
p
as.factor(lambda)
1
3
5



Negative binomial models

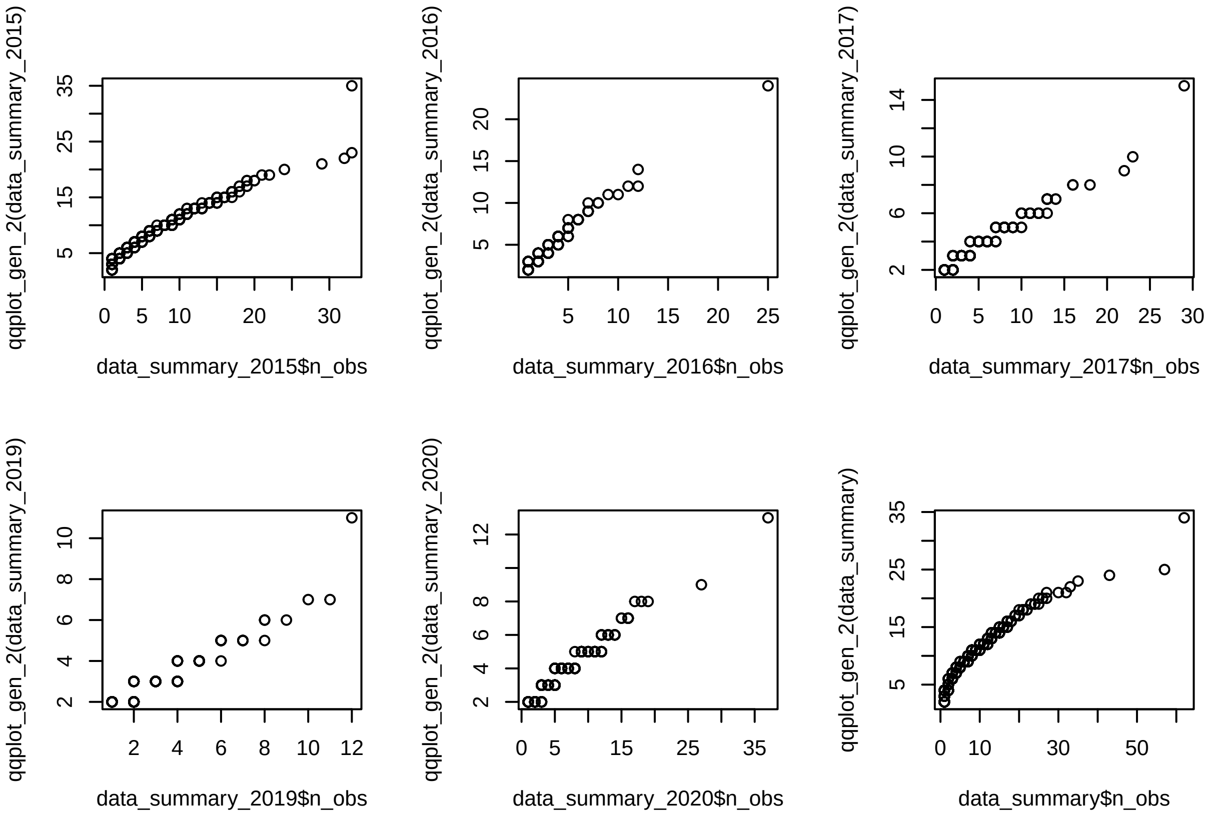
Bear data

We want to know weather or not there seems to be one-inflation present when fitting our bear data to a Negative Binomial model. In tabel

```
## Warning: ‘data_frame()‘ was deprecated in tibble 1.1.0.
## Please use ‘tibble()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was generated.
```

year	n	lambda_hat	k_hat	p_hat	N_ZTOINB	N_ZOTNB	N_ZTNB
All	3076	1.82	0.19	0.00	6558	3070	3490
2015	1016	2.51	0.20	0.00	1721	1054	1311
2016	336	1.72	0.66	0.00	518	390	434
2017	636	1.82	0.28	0.01	1157	596	792
2019	344	1.16	0.71	0.00	669	829	668
2020	1154	1.64	0.27	0.00	2338	1108	1371

TABLE BELOW SHOWS QQ-PLOT OF BEAR DATA WITH REMOVED ONES COMPARED TO THEORETICAL QUANTILES OF NEGBIN DISTRUBUTION WITH PARAMATERS ESTIMATED BY ZOTNB model



Diskussion

negativ inflation

Appendix

Referenser