UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
WYDZIAŁ MATEMATYKI I INFORMATYKI

Piotr Chlebicki

Kierunek: matematyka
Specjalność: statystyka matematyczna i analiza danych
Numer albumu: 456516

Punktowa i przedziałowa estymacja wielkości ukrytej populacji

Point and interval estimation of hidden population

Praca licencjacka
napisana pod kierunkiem
prof. UAM dra hab. Waldemara Wołyńskiego

POZNAŃ 2022

Spis treści

Streszczenie

Tematem niniejszej pracy jest estymacja wielkości populacji w sytuacjach, w których możliwe jest zaobserwowanie tylko jej części. W pracy skupiono się na temacie estymacji przy posiadaniu tylko jednego źródła obserwacji, na przykład rejestr policji dla populacji zajmującej się nielegalną aktywnością. Oczywiście jest to zagadnienie o wiele bardziej rozległe więc warto pamiętać, że opisana zostanie tylko część sposobów estymacji to znaczy jednoźródłowa metoda *capture-recapture*[1].

Początkowo badania typu capture-recapture stosowane były w ekologii w celu oszacowania ilości osobników danego gatunku na danej przestrzeni, w sytuacji gdy niemożliwym lub niepraktycznym jest policzenie wszystkich jednostek. W ciągu ostatnich trzydziestu lat nastąpił wzrost zainteresowania metodą capture-recapture w przypadku obecności jednego źródła. Jednym z powodów było uwzględnienie niejednorodności populacji w nowej klasie estymatorów, wykorzystującej analizę regresji w postaci wektora informacji dodatkowej. Uwzględnienie niejednorodności populacji w modelu pozwala na o wiele dokładniejszą estymację docelowej wielkości, szczególnie jeżeli uwzględnimy, że estymacja korzystająca z jednego źródła jest w stanie wziąć pod uwagę nie tylko czy dana jednostka wystąpiła w źródle, ale także ile razy.

Zaprezentowane w tej pracy metody estymacji znajdują zastosowanie np. w estymacji liczby kierowców prowadzących pojazdy pod wpływem alkoholu Van Der Heijden, M. Cruyff, and Van Houwelingen 2003, ilości osób nielegalnie posiadających broń palną w tej samej pracy, w estymacji liczby narkomanów na terenie miasta M. J. L. F. Cruyff and P. G. M. v. d. Heijden 2008, oraz estymacji liczby imigrantów znajdujących się na terenie danego państwa nielegalnie na terenie danej jednostki/jednostek administracyjnych P. G. v. d. Heijden et al. 2003. Informacje otrzymane w ten sposób mogą chociażby pomóc w ocenie funkcjonalności administracji publicznej, np. pod kątem zwalczania przestępczości. W celu prezentacji funkcjonalności estymatorów posłużymy się częścią danych pochodzących z tych publikacji.

[1] Ponieważ jedynym często używanym tłumaczeniem terminu na język polski jest metoda wielokrotnych złowień, nie pasująca zbytnio do przykładów przedstawionych w pracy ani do szczególnego przypadku jednego rejestru użyta w pracy została anglojęzyczna nazwa metody.

W pierwszym rozdziale krótko omówione zostaną pewne elementarne zagadnienia probabilistyczne, które ułatwią zrozumienie zagadnień poruszanych w dalszej części pracy, a które nie są zazwyczaj omawiane na kursach rachunku prawdopodobieństwa.

W drugim i trzecim rozdziale omówione zostaną konkretne estymatory punktowe, oraz bazujące na nich przedziały ufności. W drugim rozdziale opisane zostaną estymatory wykorzystujące regresje log-liniową, natomiast estymatory znajdujące się w rozdziale trzecim wykorzystują regresje logistyczną, opisane zostaną także starsze estymatory zakładające jednorodność populacji, na podstawie których wyprowadzono modele estymacji uwzględniające zaobserwowaną niejednorodność jednostek.

Czwarty rozdział to omówienie funkcji języka R, które służą w praktycznym zastosowaniu przedstawionych estymatorów.

W piątym rozdziale rozważane modele estymacji zostaną użyte w celu oszacowania populacji imigrantów, którzy przebywają nielegalnie na terenie państwa, w czterech miastach Holandii z wykorzystaniem danych z rejestru policji, oraz ilości brytyjskich farm, które nie oddają próbek ze zdechłych krów do laboratorium w celu potwierdzenia braku nieznanej choroby, w sytuacji w której przyczyny ich śmierci nie mogły zostać zidentyfikowane

Na końcu znajduje się krótkie podsumowanie i dalsza dyskusja, jako że materiał zaprezentowany w pracy nie wyczerpuje tematu estymacji rozmiaru populacji podejściem capture-recapture nawet na bazie jednego źródła.

Praca została zrealizowana w ramach grantu Narodowego Centrum Nauki OPUS 20 nr 2020/39/B/HS4/00941 pt. "Statystyka cudzoziemców bez spisu powszechnego – jakość, integracja danych i estymacja"
(kierownik dr Maciej Beręsewicz)

Abstract

The topic of this thesis is estimation of population size in situations when observing only a part of population is feasible. The focus is placed on estimation from a single source of observation, such as police registry for population that partakes in criminal activities. The subject matter is of course much broader so it is important to take into account that only a fraction of known *capture-recapture* techniques will be described known as single source capture-recapture methods.

Originally capture-recapture type studies were used in ecology with the objective of estimating number of members for a particular species on given space, in situations when counting all members of a species was either not possible or unpractical. In the last thirty years a rise in interest in capture-recapture methods utilising only one source of information, one of the reason for such occurrence was creation of new class of estimators that take into account observed heterogeneity of a population by employing regression analysis. Allowing for heterogeneity of a population allows for much more precise assessment of a target quantity, especially since these methods accommodate information of the number of times a unit has been observed instead of only considering whether a unit was observed or not.

Methods that have been described in chapters two and three find applications in areas such as estimating the number of drivers that have at some point drove under the influence of alcohol Van Der Heijden, M. Cruyff, and Van Houwelingen 2003, number of people who illegally own firearms in the same article, assessing population size of drug users in the territory of a particular city M. J. L. F. Cruyff and P. G. M. v. d. Heijden 2008, or estimating the number of immigrants who illegally reside in part of a country P. G. v. d. Heijden et al. 2003. The information obtained from applying these methods may for example support assessment of functionality of public administration in terms of preventing criminal activity from taking place. To demonstrate the functionality of models presented they will be applied to some of the data used in aforementioned articles.

The first chapter is dedicated to discussing some elementary concepts of probability that will later be useful in description of estimators, but that aren't necessarily mentioned on probability courses.

The second and third chapters describe point and interval estimators for population size. Second chapter is dedicated to models that utilise log-linear regression and third chapter does the same with logistic regression based estimators and it presents older estimators whose generalisation for observed heterogeneity will be used

Fourth chapter focuses on functions of R programming language that will be useful in applying models for estimating population size.

Fifth chapter presents applications of derived models to data sets that originate in publications P. G. v. d. Heijden et al. 2003 Böhning, Vidal-Diez, et al. 2013 concerning immigrants in Netherlands and submissions of British farms to AHVLA.

The last part of this thesis is a short summary of obtained results and further discussion of capture-recapture approaches since the material provided here does not exhaust topic of this method even for single source based estimators

This study was supported by the National Science Center grant OPUS 20 2020/39/B/HS4/00941 "Towards census-like statistics for foreign-born populations – quality, data integration and estimation"
(principal investigator dr Maciej Beręsewicz)

Preliminaria probabilistyczne

Klasyczny estymator Horvitza-Thompsona

Rozważmy proces w którym z pośród N elementowej populacji każda z jednostek $i = 1, 2, \ldots, N$ ma prawdopodobieństwa zostania zaobserwowaną odpowiednio $\pi_i \in (0, 1)$ oraz prawdopodobieństwa obserwacji poszczególnych jednostek są niezależne, i zostało zaobserowane łącznie N_{obs}. W typowej sytuacji zachodzi nierówność $N_{obs} < N$, więc dysponujemy tylko częścią potencjalnej informacji oraz najczęściej nie znamy wartości N. Niech $(I_k)_{i=1}^{N}$ będą zmiennymi indykatorowymi o wartości 1 gdy k-ta jednostka została zaobserwowana i 0 w przeciwnym razie, z definicji wynika, że $I_k \sim b(\pi_k)$. W pracy Horvitz and Thompson 1952 zaproponowany został estymator nieobciążony, który w szczególnym przypadku[2] może zostać zaaplikowany do estymacji wielkości N, wyrażony (w tym przypadku) poprzez

$$\hat{N} = \sum_{k=1}^{N_{obs}} \frac{1}{\pi_k} \tag{1.1}$$

Twierdzenie 1. *Estymator opisany równaniem (1.1) jest nieobciążony.*

Dowód. Pamiętając, że jeżeli zmienna losowa X ma rozkład $b(p)$ to $\mathbb{E}X = p$, wartość oczekiwana estymatora $\hat{N}$ wyraża się poprzez:

$$\begin{aligned} \mathbb{E}\hat{N} &= \mathbb{E}\left(\sum_{k=1}^{N_{obs}} \frac{1}{\pi_k}\right) \\ &= \mathbb{E}\left(\sum_{k=1}^{N} \frac{I_k}{\pi_k}\right) \\ &= \sum_{k=1}^{N} \frac{1}{\pi_k}\mathbb{E}(I_k) = \sum_{k=1}^{N} 1 = N \end{aligned} \tag{1.2}$$

□

[2] W oryginalnej pracy skupiono się na estymacji wartości średniej w warunkach gdy nie znana jest wartość części obserwacji na przykład w sondach ulicznych gdzie część respondentów nie chce odpowiedzieć/odpowiada nie wiem.

W pracy P. G. v. d. Heijden et al. 2003 podana jest interpretacja estymatora $\hat{N}$, jako estymator ilości jednostek, które posiadają dodatnie prawdopodobieństwo zostania zaobserwowanymi.
Jako przedział ufności dla estymatora (1.1) najczęściej przyjmowany jest studentyzowany przedział ufności, taki przedział wymaga estymatora wariancji, analitycznie wariancja $\hat{N}$, przy założeniu braku korelacji zmiennych indykatorowych $I_1, I_2, \ldots, I_N$ wyraża się poprzez:

$$\begin{aligned} \operatorname{var}(\hat{N}) &= \operatorname{var}\left(\sum_{k=1}^{N_{obs}} \frac{1}{\pi_k}\right) \\ &= \operatorname{var}\left(\sum_{k=1}^{N} \frac{I_k}{\pi_k}\right) \\ &= \sum_{k=1}^{N} \frac{1}{\pi_k^2} \operatorname{var}(I_k) = \sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k} \end{aligned} \tag{1.3}$$

Ponieważ końcowa postać wariancji jest nieznana nachodzi potrzeba jej estymacji, nieobciążony estymator wielkości (1.3) to

$$\sum_{k=1}^{N} I_k \frac{1-\pi_k}{\pi_k^2}$$

przy znajomości wartości zmiennych $I_1, \ldots, I_N$ powyższe wyrażenie upraszcza się do:

$$\sum_{k=1}^{N_{obs}} \frac{1-\pi_k}{\pi_k^2}$$

Oczywiście wyprowadzony powyżej estymator wariancji, zakłada, że znane są prawdopodobieństwa $\pi_1, \pi_2, \ldots, \pi_{N_{obs}}$, w dalszych rozważaniach uwzględniona zostanie niepewność związana z wymaganą w większości przypadków estymacją prawdopodobieństw obserwacji.

Dyskretne rozkłady prawdopodobieństwa lewostronnie ucięte

Wprowadzenie

Ponieważ w dalszych rozważaniach odnośnie estymatorów wielkości populacji korzystać będziemy z zmiennych losowych tworzonych poprzez obcięcie możliwych wartości zmiennej do podzbioru nośnika tej zmiennej, zaprezentowane są pewne informacje odnośnie zmiennych utworzonych w ten sposób, w kontekście wykorzystywanym w dalszej części pracy.

Definicja 1. *Niech X będzie zmienną losową dyskretną zdefiniowaną na przestrzeni probabilistycznej $(\Omega, \mathcal{F}, \mathbb{P})$ będzie przestrzenią probabilistyczną spełniającą:*

$$\begin{aligned} \forall k \in \mathbb{N}_0 \text{[3]} : \mathbb{P}(X = k) > 0 \\ \forall k \notin \mathbb{N}_0 : \mathbb{P}(X = k) = 0 \end{aligned} \tag{1.4}$$

O zmiennej losowej $X|X > c$, dla dowolnego $c \in [0, \infty)$ powiemy, że ma rozkład lewostronnie ucięty, lub ucięty w wartościach $0, 1, \ldots, \lfloor c \rfloor$, ponadto z własności prawdopodobieństwa warunkowego, zachodzi:

$$\begin{aligned} &\mathbb{P}(X = k|X > c) = \frac{\mathbb{P}((X = k) \cap (X > c))}{\mathbb{P}(X > c)} = \\ &\frac{\mathbb{P}(X = k)}{1 - \mathbb{P}(X \leqslant c)} = \frac{\mathbb{P}(X = k)}{1 - \sum_{i=0}^{\lfloor c \rfloor} \mathbb{P}(X = i)} && \textit{Gdy: } k > c \\ &0 && \textit{Gdy: } k \leqslant c \end{aligned} \tag{1.5}$$

Najważniejsze momenty

$$\begin{aligned} \mathbb{E}(X|X > c) &= \sum_{k=\lfloor c \rfloor + 1}^{\infty} \left(k \frac{\mathbb{P}(X = k)}{1 - \sum_{a=0}^{\lfloor c \rfloor} \mathbb{P}(X = a)} \right) \\ &= \frac{1}{1 - \sum_{k=0}^{\lfloor c \rfloor} \mathbb{P}(X = k)} \sum_{k=\lfloor c \rfloor + 1}^{\infty} (k\mathbb{P}(X = k)) \\ &= \frac{1}{1 - \sum_{k=0}^{\lfloor c \rfloor} \mathbb{P}(X = k)} \left(\sum_{k=0}^{\infty} (k\mathbb{P}(X = k)) \right. \\ &\qquad \left. - \sum_{k=0}^{\lfloor c \rfloor} (k\mathbb{P}(X = k)) \right) \\ &= \frac{1}{1 - \sum_{k=0}^{\lfloor c \rfloor} \mathbb{P}(X = k)} \left(\mathbb{E}(X) - \sum_{k=0}^{\lfloor c \rfloor} (k\mathbb{P}(X = k)) \right) \end{aligned} \tag{1.6}$$

[3] Przyjmijmy, że $0 \notin \mathbb{N}$, oraz $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$

$$\begin{aligned}
\mathbb{E}(X|X>c)^2 &= \sum_{k=\lfloor c\rfloor+1}^{\infty}\left(k^2\frac{\mathbb{P}(X=k)}{1-\sum_{a=0}^{\lfloor c\rfloor}\mathbb{P}(X=a)}\right)\\
&= \frac{1}{1-\sum_{k=0}^{\lfloor c\rfloor}\mathbb{P}(X=k)}\sum_{k=\lfloor c\rfloor+1}^{\infty}\left(k^2\mathbb{P}(X=k)\right)\\
&= \frac{1}{1-\sum_{k=0}^{\lfloor c\rfloor}\mathbb{P}(X=k)}\left(\sum_{k=0}^{\infty}\left(k^2\mathbb{P}(X=k)\right)\right. \qquad (1.7)\\
&\qquad\qquad \left.-\sum_{k=0}^{\lfloor c\rfloor}\left(k^2\mathbb{P}(X=k)\right)\right)\\
&= \frac{1}{1-\sum_{k=0}^{\lfloor c\rfloor}\mathbb{P}(X=k)}\left(\mathbb{E}\left(X^2\right)-\sum_{k=0}^{\lfloor c\rfloor}\left(k^2\mathbb{P}(X=k)\right)\right)
\end{aligned}$$

Najprostszy sposób na wyprowadzenie wariancji zmiennej $(X|X>c)$ to skorzystanie ze znanego wzoru dla operatora wariancji:

$$\operatorname{var}(X|X>c) = \mathbb{E}(X|X>c)^2 - \left(\mathbb{E}(X|X>c)\right)^2 \qquad (1.8)$$

W przypadku rozkładów uciętych w zerze wyrażenia (1.5)-(1.8) upraszczają się do

$$\begin{aligned}
\mathbb{P}(X=k|X>0) &= \frac{\mathbb{P}(X=k)}{1-\mathbb{P}(X=0)} \qquad \text{Gdy: } k>0\\
\mathbb{E}(X|X>0) &= \frac{\mathbb{E}(X)}{1-\mathbb{P}(X=0)}\\
\mathbb{E}(X|X>0)^2 &= \frac{\mathbb{E}\left(X^2\right)}{1-\mathbb{P}(X=0)}\\
\operatorname{var}(X|X>0) &= \frac{\mathbb{E}\left(X^2\right)}{1-\mathbb{P}(X=0)} - \left(\frac{\mathbb{E}(X)}{1-\mathbb{P}(X=0)}\right)^2\\
&= \frac{\operatorname{var}(X)-\mathbb{P}(X=0)\mathbb{E}\left(X^2\right)}{\left(1-\mathbb{P}(X=0)\right)^2}
\end{aligned}$$

Rozdział 2

Modele logarytmiczno-liniowe

Model Poissona

Motywacja

Model Poissona ucięty w zerze, przy obecności informacji dodatkowej, został po raz pierwszy zaprezentowany w pracy P. G. v. d. Heijden et al. 2003, w celu uwzględnienia niejednorodności populacji, w tym modelu zmienna Y_i to liczba obserwacji i-tej jednostki. We wcześniej wykorzystywanych modelach wykorzystywano nierealistyczne założenie jednorodności populacji, które w sposób istotny zaniżało wartość estymatora rozmiaru populacji. Ważną własnością modelu jest fakt, że nie wymaga aby informacja dodatkowa była uwzględniona w postaci zmiennych dyskretnych lub kategorycznych, zmienne ciągłe także mogą zostać użyte. Założenia w modelu Poissona uciętym w jedynce są następujące:

1. Zmienna zależna Y jest generowana przez rozkład Poissona, w kontekście metody capture-recapture to założenie implikuje, że obserwacja jednostki nie wpływa na zmianę zachowania w sposób, który mógłby zmienić prawdopodobieństwo obserwacji następny raz.
2. Brak występowania niezaobserwowanej niejednorodności, innymi słowy wszystkie ważne informacje dodatkowe są uwzględnione w danych
3. Prawdopodobieństwo obserwacji poszczególnych jednostek jest niezależne.
4. Szacowany rozmiar populacji jest stały, założenie to jest także obecne w tradycyjnych podejściach capture-recapture.
5. Logarytm wartości oczekiwanej zmiennej Y przy znajomości wektora informacji dodatkowej $\boldsymbol{x}$ jest pewną liniową kombinacją elementów wektora $\boldsymbol{x}$ - założenie to wymagane jest z powodu wykorzystania regresji log-liniowej.

Opis modelu regresji

Niech Y_i będzie zmienną losową o rozkładzie Poissona z parametrem λ_i. Jeżeli i-ta jednostka została zaobserwowana to $Y_i > 0$, w efekcie cecha $Y_i|Y_i > 0$ opisująca liczbę obserwacji ma rozkład Poissona ucięty w zerze opisanym przez funkcję prawdopodobieństwa:

$$\mathbb{P}(Y_i|Y_i > 0, \lambda) = \frac{\mathbb{P}(Y_i|\lambda)}{\mathbb{P}(Y_i > 0|\lambda)} = \frac{\lambda^{y_i} \exp(-\lambda)}{(1 - \exp(-\lambda))y_i!} = \frac{\lambda^{y_i}}{(\exp(\lambda) - 1)y_i!} \tag{2.1}$$

Przy log-liniowym modelu regresji z funkcją łączącą $\ln(\cdot)$ opisanym, przez

$$\eta_i = \ln \lambda_i = \ln\left(\mathbb{E}(Y_i|\lambda_i)\right) = \boldsymbol{x}_i\boldsymbol{\beta} = \sum_{k=0}^{p} \beta_k x_{ik} \tag{2.2}$$

gdzie $\boldsymbol{x}_i$ jest wektorem zmiennych dodatkowych dla i-tej jednostki $\boldsymbol{x}_i = [x_{i1}, x_{i2} \ldots, x_{ip}]$, jeżeli w modelu uwzględniamy wyraz wolny wtedy
$\forall i : x_{i1} = 1$. Niech $\boldsymbol{X}$ oznacza macierz modelu postaci:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_{(0)} & \boldsymbol{x}_{(1)} & \cdots & \boldsymbol{x}_{(p)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_{N_{obs}} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_{obs}1} & x_{N_{obs}2} & \ldots & x_{N_{obs}p} \end{pmatrix}$$

Gdzie $\boldsymbol{x}_{(k)}$ są wektorami kolumnowymi.
Logarytm funkcji wiarygodności dla zaobserwowanych jednostek może zostać zapisany jako funkcja wektora parametru regresji $\boldsymbol{\beta}$:

$$\begin{aligned} \ell &= \sum_{i=1}^{N_{obs}} \Big(y_i(\boldsymbol{x}_i\boldsymbol{\beta}) - \exp(\boldsymbol{x}_i\boldsymbol{\beta}) - \ln\left(1 - \exp(-\exp(\boldsymbol{x}_i\boldsymbol{\beta}))\right) - \ln\left(y_i!\right) \Big) \\ &= \sum_{i=1}^{N_{obs}} \Big(y_i(\boldsymbol{x}_i\boldsymbol{\beta}) - \ln\left(\exp(\exp(\boldsymbol{x}_i\boldsymbol{\beta})) - 1\right) - \ln\left(y_i!\right) \Big) \end{aligned} \tag{2.3}$$

Estymację wektora $\boldsymbol{\beta}$ dokonuje się za pomocą metody największej wiarygodności, która jest równoważna maksymalizacji wielkości (2.3), w praktyce używając do tego metod numerycznych takich jak algorytm Nethona-Raphsona. Ponieważ do estymacji $\boldsymbol{\beta}$ potrzebny jest gradient oraz błędy standardowe elementów wektora $\hat{\boldsymbol{\beta}}$ (będącego estymacją wektora $\boldsymbol{\beta}$) wykorzystuje się hesjan, analitycznie wyznaczony hesjan i gradient funkcji ℓ wyprowadzone są poniżej:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \sum_{i=1}^{N_{obs}} \left(y_i x_{i0} - x_{i0} \exp(\boldsymbol{x}_i\boldsymbol{\beta}) - x_{i0} \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\exp(\exp(\boldsymbol{x}_i\boldsymbol{\beta})) - 1} \right) \\ \vdots \\ \sum_{i=1}^{N_{obs}} \left(y_i x_{ip} - x_{ip} \exp(\boldsymbol{x}_i\boldsymbol{\beta}) - x_{ip} \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\exp(\exp(\boldsymbol{x}_i\boldsymbol{\beta})) - 1} \right) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{N_{obs}} \left(y_i x_{i0} - x_{i0} \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1-\exp(-\exp(\boldsymbol{x}_i\boldsymbol{\beta}))} \right) \\ \vdots \\ \sum_{i=1}^{N_{obs}} \left(y_i x_{ip} - x_{ip} \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1-\exp(-\exp(\boldsymbol{x}_i\boldsymbol{\beta}))} \right) \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{x}_{(0)}^T\boldsymbol{y} - \boldsymbol{x}_{(0)}^T\boldsymbol{\mu} \\ \vdots \\ \boldsymbol{x}_{(p)}^T\boldsymbol{y} - \boldsymbol{x}_{(p)}^T\boldsymbol{\mu} \end{pmatrix} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}) \tag{2.4}$$

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = -\boldsymbol{X}^T \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\lambda}} \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\beta}} \tag{2.5}$$

Gdzie:

$$\boldsymbol{\mu} = \frac{\boldsymbol{\lambda}}{\mathbf{1} - \exp(-\boldsymbol{\lambda})} \tag{2.6}$$

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\lambda}} = \mathrm{Diag}\left(\frac{1}{1-\exp(-\lambda_i)} - \frac{\lambda_i \exp(-\lambda_i)}{(1-\exp(-\lambda_i))^2} \right)_{i \in \{1,\ldots,N_{obs}\}} \tag{2.7}$$

$$= \mathrm{Diag}\left(\frac{1}{1-\exp(-\exp(\boldsymbol{x}_i\boldsymbol{\beta}))} - \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})\exp(-\exp(\boldsymbol{x}_i\boldsymbol{\beta}))}{(1-\exp(-\exp(\boldsymbol{x}_i\boldsymbol{\beta})))^2} \right)_{i \in \{1,\ldots,N_{obs}\}}$$

$$\frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \exp(\boldsymbol{X}\boldsymbol{\beta}) \tag{2.8}$$

Estymator wielkości populacji

Przy zadanej wartości wektora $\boldsymbol{\beta}$ estymator Horvitza-Thompsona, dla wielkości populacji N w tym modelu wyraża się poprzez:

$$\hat{N} = \sum_{k=1}^{N} I_k \frac{1}{1-\exp(-\exp(\boldsymbol{x}_k\boldsymbol{\beta}))} \tag{2.9}$$

Gdzie I_k są zmiennymi indykatorowymi przyjmującymi wartość 1 jeżeli jednostka k została zaobserwowana i zero w przeciwnym przypadku, czyli:

$$I_k = \begin{cases} 1 & \text{gdy: } Y_k > 0 \\ 0 & \text{gdy: } Y_k = 0 \end{cases} \qquad I_k \sim b\left(e^{-e^{\boldsymbol{x}_k\beta}}\right)$$

Dodatkowo wielkość

$$\hat{\boldsymbol{f}}_0 = \hat{N} - N_{obs} = \sum_{k=1}^{N} I_k \frac{\exp(-\exp(\boldsymbol{x}_k\boldsymbol{\beta}))}{1-\exp(-\exp(\boldsymbol{x}_k\boldsymbol{\beta}))}$$

estymuje wielkość niezaobserwowanej populacji, interpretowana jako estymator jednostek niezaobserwowanych, ale o dodatnim prawdopodobieństwie zostania

zaobserwowanymi. Estymator z równania (2.9) upraszcza się, przy znajomości wartości zmiennych indykatorowych, do postaci:

$$\hat{N}|I_1, \dots, I_N = \sum_{k=1}^{N_{obs}} \frac{1}{1 - \exp\left(-\exp\left(\boldsymbol{x}_k\boldsymbol{\beta}\right)\right)}$$

Gdzie sumowanie odbywa się po jednostkach zaobserwowanych.
Ponieważ $\hat{N}$ jest sumą zmiennych o rozkładzie Bernouliego pomnożonych przez wartości stałe można założyć, że ma on rozkład asymptotycznie normalny, zatem o ile liczba zaobserwowanych jednostek nie jest zbyt mała jako przedział ufności, na poziomie istotności $(1-\alpha)\cdot 100\%$ dla statystyki $\hat{N}$ można przyjąć studentyzowany przedział ufności o postaci

$$\left(\hat{N} - z\left(1-\frac{\alpha}{2}\right)\sqrt{\text{var}\left(\hat{N}\right)}, \hat{N} + z\left(1-\frac{\alpha}{2}\right)\sqrt{\text{var}\left(\hat{N}\right)}\right) \tag{2.10}$$

Gdzie $z(w)$ jest kwantylem rzędu w z rozkładu standaryzowanego normalnego, dla $\alpha = 0.05$, $z\left(1-\frac{\alpha}{2}\right) \approx 1.96$. Zagadnienie konstrukcji przedziału ufności dla estymatora $\hat{N}$ zostało więc sprowadzone do zagadnienia estymacji wariancji estymatora $\hat{N}$.
W celu wyprowadzenia wariancji estymatora $\hat{N}$ skorzystamy z metody opisanej w P. G. v. d. Heijden et al. 2003. Korzystając z prawa pełnej wariancji następująca relacja pomiędzy wariancją $\hat{N}$ a momentami zmiennej $\hat{N}|I_1, \dots, I_N$ jest prawdziwa

$$\text{var}(\hat{N}) = \text{var}\left(\mathbb{E}(\hat{N}|I_1, \dots, I_N)\right) + \mathbb{E}\left(\text{var}\left(\hat{N}|I_1, \dots, I_N\right)\right) \tag{2.11}$$

Ponieważ przy ustalonych wartościach zmiennych indykatorowych wartość (2.8) jest stała, $\mathbb{E}(\hat{N}|I_1, \dots, I_N)$ jest zmienną losową, której wartość zależy od wartości zmiennych indykatorowych. Wariancja tej zmiennej losowej wyrażona jest poprzez:

$$\begin{aligned}
\text{var}\left(\mathbb{E}(\hat{N}|I_1, \dots, I_N)\right) &= \text{var}\left(\sum_{k=1}^{N} I_k \frac{1}{1 - \exp\left(-\exp\left(\boldsymbol{x}_k\boldsymbol{\beta}\right)\right)}\right) \\
&= \sum_{k=1}^{N} \text{var}(I_k) \frac{1}{\left(1 - \exp\left(-\exp\left(\boldsymbol{x}_k\boldsymbol{\beta}\right)\right)\right)^2} \\
&= \sum_{k=1}^{N} \frac{\exp\left(-\exp\left(\boldsymbol{x}_k\boldsymbol{\beta}\right)\right)}{1 - \exp\left(-\exp\left(\boldsymbol{x}_k\boldsymbol{\beta}\right)\right)}
\end{aligned} \tag{2.12}$$

W typowej sytuacji wartość (2.12) jest nieznana, jej nieobciążony estymator wyrażony jest poprzez

$$\sum_{k=1}^{N} I_k \frac{\exp\left(-\exp\left(\boldsymbol{x}_k\boldsymbol{\beta}\right)\right)}{\left(1 - \exp\left(-\exp\left(\boldsymbol{x}_k\boldsymbol{\beta}\right)\right)\right)^2} = \sum_{k=1}^{N_{obs}} \frac{\exp\left(-\exp\left(\boldsymbol{x}_k\boldsymbol{\beta}\right)\right)}{\left(1 - \exp\left(-\exp\left(\boldsymbol{x}_k\boldsymbol{\beta}\right)\right)\right)^2}$$

Druga z wartości obecnych w równaniu (2.11) w metodzie van der Heidena estymowana jest wielowymiarową metodą delty

$$\left(\frac{\partial(\hat{N}|I_1,\dots,I_N)}{\partial\boldsymbol{\beta}}\right)^T \mathrm{Cov}(\boldsymbol{\beta}) \left(\frac{\partial(\hat{N}|I_1,\dots,I_N)}{\partial\boldsymbol{\beta}}\right) \tag{2.13}$$

dla faktycznej wartości $\boldsymbol{\beta}$. Ponieważ wektor $\boldsymbol{\beta}$ jest estymowany, oraz macierz kowariancji dla wektora $\boldsymbol{\beta}$, przy zastąpieniu ich poprzez ich estymatory dostajemy:

$$\left(\frac{\partial(\hat{N}|I_1,\dots,I_N)}{\partial\boldsymbol{\beta}}\right)^T \left(-\frac{\partial^2\ell}{\partial\boldsymbol{\beta}^T\partial\boldsymbol{\beta}}\right)^{-1} \left(\frac{\partial(\hat{N}|I_1,\dots,I_N)}{\partial\boldsymbol{\beta}}\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \tag{2.14}$$

Gdzie odwrotność hesjanu funkcji $-\ell$ jest estymatorem otrzymanym za pomocą nierówności Cramera-Rao, estymacja macierzy kowariancji tym sposobem staje się dokładniejsza przy wzroście ilości obserwacji, jeżeli wektor $\boldsymbol{\beta}$ jest estymowany metodą największej wiarygodności.
Pochodna zmiennej $\hat{N}|I_1,\dots,I_N$ występująca w (2.13), to

$$-\boldsymbol{X}^T \left(\frac{\exp\left(\boldsymbol{X}\boldsymbol{\beta}-\exp(\boldsymbol{X}\boldsymbol{\beta})\right)}{\left(\mathbf{1}-\exp(-\exp\left(\boldsymbol{X}\boldsymbol{\beta}\right))\right)^2}\right)$$

Ponieważ wyrażenie (2.14) jest wartością stałą, jest nie zmiennicza względem operatora $\mathbb{E}$.
Zmienna $\hat{N}|I_1,\dots,I_N$ jest funkcją wektora $\boldsymbol{\beta}$ z powodu użycia metody delta do estymacji wariancji wymagane jest aby estymator $\hat{\boldsymbol{\beta}}$ był asymptotycznie normalny.
Podstawiając wartości (2.12) (2.14) do równania (2.11) dostajemy estymator wariancji estymatora $\hat{N}$, co daje nam możliwość konstruowania przedziałów ufności.

Model ujemny dwumianowy

Motywacja

Drugie z wymienionych założeń w modelu Poissona to brak niezaobserwowanej niejednorodności, ponieważ podczas procesu kolekcji danych brak informacji odnośnie tego, które zmienne są istotne lub brak możliwości otrzymania informacji dodatkowej dla wystarczającej liczby obserwacji nie jest szczególnie rzadkim przypadkiem, pojawia się potrzeba wprowadzenia modelu, który jest w stanie poradzić sobie z niezaobserwowaną niejednorodnością. W pracy M. J. L. F. Cruyff and P. G. M. v. d. Heijden 2008 autorzy argumentują, że model ujemny dwumianowy (ucięty w zerze), rozumiany jako model Poissona z nadmierną dyspersją, jest sensownym rozwiązaniem w wyżej opisanych przypadkach, ponieważ bardzo restrykcyjne założenie obecne w rozkładzie Poissona związane z momentami: $\mathbb{E}(X) = \mathrm{var}(X)$ przy obecności niezaobserwowanej niejednorodności będzie naruszone.

Opis modelu regresji

Rozkład prawdopodobieństwa

Ujemny dwumianowy rozkład prawdopodobieństwa ma kilka używanych parametryzacji, w celu estymacji N skorzystamy z rozkładu Poissona z nad dyspersją o funkcji prawdopodobieństwa postaci:

$$\mathbb{P}(Y|\lambda,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})y!}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right)^{\alpha^{-1}}\left(\frac{\lambda}{\alpha^{-1}+\lambda}\right)^{y} \tag{2.15}$$

Gdzie parametr $\alpha \in (0,\infty)$ nazywany jest w tym kontekście parametrem dyspersji, a Γ jest funkcją Gamma postaci:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

Wariancja i wartość oczekiwana zmiennej losowej Y o takim rozkładzie wyrażona jest poprzez:

$$\mathbb{E}(Y) = \lambda \qquad \mathrm{var}(Y) = \lambda + \alpha\lambda^2 = \lambda(1+\alpha\lambda) > \lambda \tag{2.16}$$

Nadmierna dyspersja występuje, gdy $\mathrm{var}(Y) > \mathbb{E}(Y)$ widać, że nie występuje ona dla zmiennych o rozkładzie Poissona ale występuje dla powyższej parametryzacji modelu ujemnego dwumianowego. Ponadto zachodzi następujące twierdzenie:

Twierdzenie 2. *Gdy $\alpha \to 0^+$, funkcja prawdopodobieństwa rozkładu ujemnego dwumianowego jest zbieżna do funkcji prawdopodobieństwa rozkładu Poissona.*

Dowód. Aby udowodnić powyższe twierdzenie skorzystamy ze wzoru Strilinga, mówiącego, że

$$\Gamma(x+1) \sim_a \sqrt{2\pi x}\left(\frac{x}{e}\right)^x \tag{2.17}$$

Gdzie $\sim_a$ jest równością asymptotyczną funkcji zdefiniowaną, w przypadku gdy granice funkcji istnieją, poprzez:

$$f(x) \sim_a g(x) \Longleftrightarrow \lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$$

Fakt : Relacja $\sim_a$ jest relacją równoważności.
Z wzoru Stirlinga zachodzi:

$$\begin{aligned}
\Gamma\left(y+\alpha^{-1}\right) &\sim_a \sqrt{2\pi(y+\alpha^{-1}-1)}\left(\frac{\alpha^{-1}+y-1}{e}\right)^{y+\alpha^{-1}-1} \\
\Gamma\left(\alpha^{-1}\right) &\sim_a \sqrt{2\pi(\alpha^{-1}-1)}\left(\frac{\alpha^{-1}-1}{e}\right)^{\alpha^{-1}-1}
\end{aligned} \tag{2.18}$$

Wykorzystując podstawowe własności granic funkcji rzeczywistych dostajemy:

$$\begin{aligned}
\frac{\Gamma\left(y+\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)} &\sim_a \frac{\sqrt{2\pi(y+\alpha^{-1}-1)}}{\sqrt{2\pi(\alpha^{-1}-1)}}\left(\frac{\alpha^{-1}+y-1}{e}\right)^{y+\alpha^{-1}-1} \\
&\quad \cdot \left(\frac{e}{\alpha^{-1}-1}\right)^{\alpha^{-1}-1} \\
&= \sqrt{1+\frac{y}{\alpha^{-1}-1}}\left(\frac{y+\alpha^{-1}-1}{e}\right)^{y}\left(1+\frac{y}{\alpha^{-1}-1}\right)^{\alpha^{-1}-1} \\
\Longrightarrow \frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})y!}\left(\frac{\lambda}{\alpha^{-1}+\lambda}\right)^{y} &\sim_a \frac{1}{y!}\sqrt{1+\frac{y}{\alpha^{-1}-1}}\left(\frac{\lambda^y}{e^y}\right) \\
&\quad \cdot \left(\frac{y+\alpha^{-1}-1}{\lambda+\alpha^{-1}}\right)^{y}\left(1+\frac{y}{\alpha^{-1}-1}\right)^{\alpha^{-1}-1}
\end{aligned} \tag{2.19}$$

Ponieważ:

$$\begin{aligned}
\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right)^{\alpha^{-1}} = \left(1+\frac{\lambda}{\alpha^{-1}}\right)^{-\alpha^{-1}} &\longrightarrow \exp(-\lambda) \\
\left(1+\frac{y}{\alpha^{-1}-1}\right)^{\alpha^{-1}-1} &\longrightarrow \exp(y) \\
\left(\frac{y+\alpha^{-1}-1}{\lambda+\alpha^{-1}}\right)^{y} &\longrightarrow 1 \\
\sqrt{1+\frac{y}{\alpha^{-1}-1}} &\longrightarrow 1
\end{aligned} \tag{2.20}$$

gdy $\alpha^{-1} \to \infty, \alpha \to 0^+$. Uwzględniając więc (2.19) i (2.20) w (2.15) dostajemy zatem granicę (2.15) przy $\alpha \to 0^+$

$$\lim_{1/\alpha \to \infty} \frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})y!} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right)^{\alpha^{-1}} \left(\frac{\lambda}{\alpha^{-1}+\lambda}\right)^{y} = e^{-\lambda}\frac{1}{y!}\frac{\lambda^y}{e^y}e^y = \frac{\lambda^y e^{-\lambda}}{y!} \quad (2.21)$$

czyli funkcję prawdopodobieństwa rozkładu Poissona. □

Regresja

Niech każdy z elementów Y_i, wektora $\boldsymbol{y}$ zaobserwowanej zmiennej zależnej, będzie opisany funkcją prawdopodobieństwa:

$$\begin{aligned}\mathbb{P}(y_i|y_i>0,\lambda_i,\alpha) =& \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})y_i!}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)^{\alpha^{-1}} \\ &\cdot \left(\frac{\lambda_i}{\alpha^{-1}+\lambda_i}\right)^{y_i} \frac{1}{1-(1+\alpha\lambda_i)^{-\alpha^{-1}}}\end{aligned} \quad (2.22)$$

Gdzie $1-(1+\alpha\lambda_i)^{-\alpha^{-1}}$ jest prawdopodobieństwem zdażenia $y_i = 0$ w modelu nieuciętym.
Logarytm funkcji wiarygodności, przy modelu regresji opisanym przez równanie (2.22), dany jest poprzez:

$$\begin{aligned}\ell(\boldsymbol{\beta},\alpha) = \sum_{k=1}^{N_{obs}} \Bigg(& \ln\Gamma(y_k+\alpha^{-1}) - \ln\Gamma(\alpha^{-1}) - \ln(y_k!) \\ & - \left(y_k+\alpha^{-1}\right)\ln\left(1+\alpha\exp(\boldsymbol{x}_k\boldsymbol{\beta})\right) + y_k\ln\left(\alpha\exp(\boldsymbol{x}_k\boldsymbol{\beta})\right) \\ & - \ln\left(1-(1+\alpha\exp(\boldsymbol{x}_k\boldsymbol{\beta}))^{-\alpha^{-1}}\right)\Bigg)\end{aligned} \quad (2.23)$$

W celu skrócenia zapisu dla gradientu i hesjanu modelu skorzystamy z zapisu:

$$\boldsymbol{S} = \begin{pmatrix} \frac{1}{1+\alpha\lambda_1} \\ \vdots \\ \frac{1}{1+\alpha\lambda_i} \\ \vdots \\ \frac{1}{1+\alpha\lambda_{N_{obs}}} \end{pmatrix} \quad \boldsymbol{\alpha}^{-1} = \begin{pmatrix} \alpha^{-1} \\ \vdots \\ \alpha^{-1} \\ \vdots \\ \alpha^{-1} \end{pmatrix} \text{ wektor długości } N_{obs}$$

$$\boldsymbol{G} = \begin{pmatrix} \frac{1}{1-(1+\alpha\lambda_1)^{-\alpha^{-1}}} \\ \vdots \\ \frac{1}{1-(1+\alpha\lambda_i)^{-\alpha^{-1}}} \\ \vdots \\ \frac{1}{1-(1+\alpha\lambda_{N_{obs}})^{-\alpha^{-1}}} \end{pmatrix} \frac{1}{\boldsymbol{S}} = \begin{pmatrix} (1+\alpha\lambda_1) \\ \vdots \\ (1+\alpha\lambda_i) \\ \vdots \\ (1+\alpha\lambda_{N_{obs}}) \end{pmatrix}$$

Analityczny gradient:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \left(\left(\boldsymbol{y} + \frac{\exp\left(\boldsymbol{X}\boldsymbol{\beta}\right) - \boldsymbol{y}}{\boldsymbol{S}^{-\alpha^{-1}+1}} \right) \times \boldsymbol{G} \right) \tag{2.24}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \alpha} = \sum_{k=1}^{N_{obs}} \Bigg(& \frac{\partial}{\partial \alpha} \left(\ln \Gamma \left(y_k + \alpha^{-1} \right) - \ln \Gamma \left(\alpha^{-1} \right) \right) + \alpha^{-2} \ln\left(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta})\right) \\
& + \frac{(1 + \exp(\boldsymbol{x}_k \boldsymbol{\beta})\alpha)^{-\alpha^{-1}} \left(\frac{\ln(1+\alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))}{\alpha^2} - \frac{\exp(\boldsymbol{x}_k \boldsymbol{\beta})}{(1+\exp(\boldsymbol{x}_k \boldsymbol{\beta})\alpha)\alpha} \right)}{1 - (1 + \exp(\boldsymbol{x}_k \boldsymbol{\beta})\alpha)^{-\alpha^{-1}}} \\
& + \frac{y_k}{\alpha} - \frac{(y_k + \alpha^{-1}) \exp(\boldsymbol{x}_k \boldsymbol{\beta})}{1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta})} \Bigg)
\end{aligned}$$

Hesjan:

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \alpha^2} = \sum_{k=1}^{N_{obs}} \Bigg(& \frac{\partial^2}{\partial \alpha^2} \left(\ln \Gamma \left(y_k + \alpha^{-1} \right) - \ln \Gamma \left(\alpha^{-1} \right) \right) + \frac{2 \exp(\boldsymbol{x}_k \boldsymbol{\beta})}{\alpha^2 (1 + \exp(\boldsymbol{x}_k \boldsymbol{\beta})\alpha)} \\
& - \frac{2 \ln(1 + \exp(\boldsymbol{x}_k \boldsymbol{\beta})\alpha)}{\alpha^3} + \frac{(y + \alpha^{-1}) \exp(2\boldsymbol{x}_k \boldsymbol{\beta})}{(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))^2} \\
& + \frac{2 \left(\exp(\boldsymbol{x}_k \boldsymbol{\beta})\alpha - \ln(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta})) \right)}{(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta})) \left((1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))^{\alpha^{-1}} - 1 \right) \alpha^3} \\
& + \frac{(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))^{\alpha^{-1}-1} \left(\frac{\exp(\boldsymbol{x}_k \boldsymbol{\beta})}{\alpha(1+\alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))} - \frac{\ln(1+\alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))}{\alpha^2} \right)}{\alpha^2 \left((1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))^{\alpha^{-1}} - 1 \right)^2} \\
& \cdot \left(\exp(\boldsymbol{x}_k \boldsymbol{\beta})\alpha - \ln(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta})) \right) \\
& + \frac{\exp(\boldsymbol{x}_k \boldsymbol{\beta}) \ln(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))}{(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta})) \left((1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))^{\alpha^{-1}} - 1 \right) \alpha^2} \\
& + \frac{\exp(\boldsymbol{x}_k \boldsymbol{\beta}) \left(\exp(\boldsymbol{x}_k \boldsymbol{\beta})\alpha - \ln(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))(1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta})) \right)}{\alpha^2 (1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))^{\alpha^{-1}} - 1) (1 + \alpha \exp(\boldsymbol{x}_k \boldsymbol{\beta}))^2} \Bigg)
\end{aligned} \tag{2.25}$$

$$\frac{\partial^2 \ell}{\partial \alpha \partial \boldsymbol{\beta}} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \alpha} =$$

$$\begin{aligned}
-\boldsymbol{X}^T \Bigg(& \exp(\boldsymbol{X}\boldsymbol{\beta}) \times (\boldsymbol{y} - \exp(\boldsymbol{X}\boldsymbol{\beta})) \times \boldsymbol{S} \times \boldsymbol{S} - \exp(\boldsymbol{X}\boldsymbol{\beta}) \\
& \times \left(\frac{1}{\boldsymbol{S}^{-1-\alpha^{-1}}} \times \left(\left(1 + \alpha^{-1}\right) \exp(\boldsymbol{X}\boldsymbol{\beta}) \times \boldsymbol{S} \right. \right. \\
& \left. - \alpha^{-2} \ln(\boldsymbol{1} + \alpha \exp(\boldsymbol{X}\boldsymbol{\beta})) \right) \times \boldsymbol{G} \\
& \left. + \frac{\alpha^{-2} \ln \boldsymbol{S} + \alpha^{-1} \exp(\boldsymbol{X}\boldsymbol{\beta}) \times \boldsymbol{S}}{\boldsymbol{S}^{-\alpha^{-1}}} \times \frac{1}{\boldsymbol{S}^{-1-\alpha^{-1}}} \times \boldsymbol{G} \times \boldsymbol{G} \right) \Bigg)
\end{aligned} \tag{2.26}$$

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} = -\boldsymbol{X} \Bigg(& \boldsymbol{x}_{(i)} \times \exp(\boldsymbol{X}\boldsymbol{\beta}) \times \frac{\mathbf{1} + \boldsymbol{y}\alpha^{-1}}{(\mathbf{1} + \alpha \exp(\boldsymbol{X}\boldsymbol{\beta}))^2} \\ & - \boldsymbol{x}_{(i)} \times \exp(\boldsymbol{X}\boldsymbol{\beta}) \times \left(\frac{(\exp(\boldsymbol{X}\boldsymbol{\beta}) - \mathbf{1})}{\boldsymbol{S}^{\alpha^{-1}}} + \mathbf{1} \right) \\ & \times \boldsymbol{S} \times \boldsymbol{S} \times \boldsymbol{G} \times \boldsymbol{G} \Bigg)_{i=0}^{p}\end{aligned} \tag{2.27}$$

W literaturze dotyczącej regresji dobrze udokumentowany jest fakt, że powyższy model jest problematyczny numerycznie, w tym sensie, że algorytmy numeryczne używane do dopasowywania modelu często wychodzą poza dozwolone granice dla parametrów (np. na którymś kroku algorytmu optymalizującego, może się zdażyć, że $\alpha = -1$) z tego powodu zamiast estymować parametr α zazwyczaj zastępuję się go parametrem $t = \ln \alpha$. Pochodne względem zmiennej t uzyskujemy, poprzez aplikację zasady łańcuchowej do (2.26) (2.25) i (2.24) co doprowadza do postaci:

$$\begin{aligned}\frac{\partial \ell}{\partial t} &= \frac{\partial \ell}{\partial \alpha} \frac{\partial \alpha}{\partial t} = \frac{\partial \ell}{\partial \alpha} \underbrace{\exp(t)}_{\alpha} \\ \frac{\partial^2 \ell}{\partial t^2} &= \underbrace{\exp(t)}_{\alpha} \frac{\partial \ell}{\partial \alpha} + \underbrace{\exp(2t)}_{\alpha^2} \frac{\partial^2 \ell}{\partial \alpha^2} \\ \frac{\partial^2 \ell}{\partial t \partial \boldsymbol{\beta}} &= \underbrace{\exp(t)}_{\alpha} \frac{\partial^2 \ell}{\partial \alpha \partial \boldsymbol{\beta}}\end{aligned} \tag{2.28}$$

Wartość $\Gamma(y_k + \alpha^{-1}) - \ln \Gamma(\alpha^{-1})$, łatwiej niż w sposób jawny, obliczyć można poprzez skorzystanie z zależności:

$$\begin{aligned}\ln \Gamma\left(y_k + \alpha^{-1}\right) - \ln \Gamma\left(\alpha^{-1}\right) &= \sum_{j=0}^{y_k - 1} \ln\left(j + \alpha^{-1}\right) \\ \frac{\partial}{\partial \alpha}\left(\ln \Gamma\left(y_k + \alpha^{-1}\right) - \ln \Gamma\left(\alpha^{-1}\right)\right) &= -\alpha^2 \sum_{j=0}^{y_k - 1} \frac{1}{j + \alpha^{-1}} \\ \frac{\partial^2}{\partial \alpha^2}\left(\ln \Gamma\left(y_k + \alpha^{-1}\right) - \ln \Gamma\left(\alpha^{-1}\right)\right) &= \sum_{j=0}^{y_k - 1} \frac{1 + 2j\alpha}{\alpha^2 (1 + j\alpha)^2}\end{aligned} \tag{2.29}$$

która wyprowadzona jest bezpośrednio z własności funkcji gamma:

$$\Gamma(x + 1) = x\Gamma(x)$$

Estymator wielkości populacji

Oznaczmy $p_k(\boldsymbol{\beta}, t) = 1 - (1 + \exp(t + \boldsymbol{x}_k\boldsymbol{\beta}))^{-e^{-t}}$.
Estymator liczby niezaobserwowanych jednostek $\hat{\boldsymbol{f}}_0$ o dodatnim prawdopodobieństwa bycia zaobserwowanymi co najmniej raz oraz estymator jednostek o dodatnim prawdopodobieństwie bycia zaobserwowanymi przynajmniej raz $\hat{N}$, skonstruowany w taki sam sposób, jak w poprzednim przykładzie, przyjmują postacie:

$$\begin{aligned} \hat{\boldsymbol{f}}_0 &= \sum_{k=1}^{N} I_k \frac{1 - p_k(\hat{\boldsymbol{\beta}}, \hat{t})}{p_k(\hat{\boldsymbol{\beta}}, \hat{t})} \\ \hat{N} &= N_{obs} + \hat{\boldsymbol{f}}_0 = \sum_{k=1}^{N} \frac{I_k}{p_k(\hat{\boldsymbol{\beta}}, \hat{t})} \end{aligned} \tag{2.30}$$

Estymacja wariancji statystyki (2.30) odbywa się w ten sam sposób jak w modelu Poissona, z tą różnicą, że mamy dodatkowy parametr t do estymacji, co musi zostać uwzględnione przy użyciu metody δ.

$$\begin{aligned} \operatorname{var}(\hat{N}) &= \mathbb{E}(\operatorname{var}(\hat{N})) + \operatorname{var}(\mathbb{E}(\hat{N})) \\ \operatorname{var}(\mathbb{E}(\hat{N}|I_1, \ldots, I_{N_{obs}})) &\approx \Theta^T \boldsymbol{W}^{-1} \Theta \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, t=\hat{t}} \\ \mathbb{E}(\operatorname{var}(\hat{N}|I_1, \ldots, I_{N_{obs}})) &= \sum_{k=1}^{N} \frac{1 - p_k(\hat{\boldsymbol{\beta}}, \hat{t})}{p_k(\hat{\boldsymbol{\beta}}, \hat{t})} \\ &\approx \sum_{k=1}^{N} I_k \frac{1 - p_k(\hat{\boldsymbol{\beta}}, \hat{t})}{\Big(p_k(\hat{\boldsymbol{\beta}}, \hat{t})\Big)^2} \\ &= \sum_{k=1}^{N_{obs}} \frac{1 - p_k(\hat{\boldsymbol{\beta}}, \hat{t})}{\Big(p_k(\hat{\boldsymbol{\beta}}, \hat{t})\Big)^2} \end{aligned} \tag{2.31}$$

Gdzie:

$$\Theta = \begin{pmatrix} \frac{\partial}{\partial t} \sum_{k=1}^{N_{obs}} \frac{1}{p(\lambda_k, t)} \\ \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{k=1}^{N_{obs}} \frac{1}{p(\lambda_k, \alpha)} \end{pmatrix} \tag{2.32}$$

$$\boldsymbol{W} = -\begin{pmatrix} \frac{\partial^2 \ell}{\partial t^2} & \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial t} \\ \frac{\partial^2 \ell}{\partial t \partial \boldsymbol{\beta}} & \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} \end{pmatrix} \tag{2.33}$$

$$
\begin{aligned}
\frac{\partial}{\partial\alpha}\sum_{k=1}^{N_{obs}}\frac{1}{p_k(\boldsymbol{\beta},\alpha)} &= \sum_{k=1}^{N_{obs}}\frac{(1+\alpha e^{\boldsymbol{\beta}\boldsymbol{x}_k})^{\alpha^{-1}-1}}{\alpha^2\left(1-(1+\alpha\exp(\boldsymbol{\beta}\boldsymbol{x}_k))^{\alpha^{-1}}\right)^2} \\
&\quad\cdot\left((1+\alpha e^{\boldsymbol{\beta}\boldsymbol{x}_k})\ln(1+\alpha e^{\boldsymbol{\beta}\boldsymbol{x}_k})-\alpha e^{\boldsymbol{\beta}\boldsymbol{x}_k}\right) \\
\frac{\partial}{\partial t}\sum_{k=1}^{N_{obs}}\frac{1}{p_k(\boldsymbol{\beta},t)} &= \exp(t)\frac{\partial}{\partial\alpha}\sum_{k=1}^{N_{obs}}\frac{1}{p(\boldsymbol{\beta},\alpha)} \\
\frac{\partial}{\partial\boldsymbol{\beta}}\frac{1}{p_k(\boldsymbol{\beta},\alpha)} &= -\boldsymbol{x}_k\exp(\boldsymbol{\beta}\boldsymbol{x}_k)\frac{(1+\alpha\exp(\boldsymbol{\beta}\boldsymbol{x}_k))^{\alpha^{-1}-1}}{\left(1-(1+\alpha\exp(\boldsymbol{\beta}\boldsymbol{x}_k))^{\alpha^{-1}}\right)^2} \\
\sum_{k=1}^{N_{obs}}\frac{\partial}{\partial\boldsymbol{\beta}}\frac{1}{p_k(\boldsymbol{\beta},\alpha)} &= -\left(\exp(\boldsymbol{X}\boldsymbol{\beta})\times\boldsymbol{G}^2\times\frac{1}{\boldsymbol{S}^{\alpha^{-1}-1}}\right)\boldsymbol{X}^T
\end{aligned}
\tag{2.34}
$$

Model geometryczny

Motywacja

Zaprezentowany poprzednio model ujemny dwumianowy ma kilka problemów. Najważniejszym jest fakt, że związany z nim model regresji jest jest często trudny pod względem estymacji numerycznej wektora $(t, \boldsymbol{\beta}^T)^T$ co znacznie utrudnia optymalizację komputerową, szczególnie w sytuacji gdy wybrana metoda numeryczna nie jest zbieżna. Dodatkowym problemem może być zbieżność do nieprawidłowej wartości, co może w ogóle nie zostać wykryte. W takich sytuacjach estymator parametru t może zdecydowanie zwiększyć wartość estymatora $\hat{N}$. Poniższy model jest specjalnym przypadkiem modelu ujemnego dwumianowego uzyskanym poprzez wymuszenie $\alpha = 1$ lub równoważnie $t = 0$, który unika w znaczniej większości przypadków wszystkich problemów związanych z metodami numerycznymi, ale nadal pozwala na uwzględnienie części niezaobserwowanej niejednorodności.

Opis modelu regresji

Przy wartości parametru $\alpha = 1$ funkcja prawdopodobieństwa dla rozkładu ujemnego dwumianowego:

$$\begin{aligned}\mathbb{P}(y_i|y_i > 0, \lambda_i, \alpha) =& \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})y_i!}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}} \\ & \cdot \left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i}\right)^{y_i} \frac{1}{1 - (1 + \alpha\lambda_i)^{-\alpha^{-1}}}\end{aligned}$$

upraszcza się do postaci:

$$\mathbb{P}(y_i|y_i > 0, \lambda_i, \alpha = 1) = \left(\frac{1}{1 + \lambda_i}\right)\left(\frac{\lambda_i}{1 + \lambda_i}\right)^{y_i - 1} = \frac{\lambda_i^{y_i - 1}}{(1 + \lambda_i)^{y_i}} \tag{2.35}$$

co jest równoważne parametryzacji $p(1-p)^{y-1}$ jeżeli przyjmiemy $p = \frac{1}{1+\lambda}$. Uproszczony logarytm funkcji wiarygodności przyjmuje poniższą postać:

$$\ell = \sum_{k=1}^{N_{obs}} \Big((y_k - 1)\boldsymbol{x}_k\boldsymbol{\beta} - y_k \ln\left(1 + \exp(\boldsymbol{x}_k\boldsymbol{\beta})\right) \Big) \tag{2.36}$$

Ponieważ nie ma potrzeby estymacji parametru t gradient i hesjan modelu upraszcza się do prostych wyrażeń:

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \boldsymbol{X}^T \left(\frac{\boldsymbol{y} - \exp\left(\boldsymbol{X}\boldsymbol{\beta}\right) - \mathbf{1}}{\mathbf{1} + \exp\left(\boldsymbol{X}\boldsymbol{\beta}\right)}\right) \\ &= \boldsymbol{X}^T \left(\frac{\boldsymbol{y}}{\mathbf{1} + \exp\left(\boldsymbol{X}\boldsymbol{\beta}\right)} - \mathbf{1}\right)\end{aligned} \tag{2.37}$$

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^T \boldsymbol{\beta}} = -\left(\boldsymbol{x}_{(i)} \times \frac{\boldsymbol{y}}{\left(\mathbf{1} + \exp\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)^2} \times \exp\left(\boldsymbol{X}\boldsymbol{\beta}\right) \right)^T \boldsymbol{X} \tag{2.38}$$

Estymator wielkości populacji

Estymator liczby niezaobserwowanych jednostek $\hat{\boldsymbol{f}}_0$ o dodatnim prawdopodobieństwie bycia zaobserwowanymi, oraz estymator całkowitej populacji $\hat{N}$ otrzymane poprzez uproszczenie estymatora (2.30) oraz estymator wariancji powstają poprzez uproszczenie modelu ujemnego dwumianowego. Wielkość p_k upraszcza się do:

$$p_k(\boldsymbol{\beta}) = 1 - (1 + \exp(\boldsymbol{x}_k\boldsymbol{\beta}))^{-1} = \frac{\exp(\boldsymbol{x}_k\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_k\boldsymbol{\beta})}$$

zatem estymatory (2.30) upraszczają się do:

$$\begin{aligned} \hat{\boldsymbol{f}}_0 &= \sum_{k=1}^{N} I_k \frac{1 - p_k(\hat{\boldsymbol{\beta}})}{p_k(\hat{\boldsymbol{\beta}})} = \sum_{k=1}^{N} I_k \frac{\frac{1}{1+\exp(\boldsymbol{x}_k\hat{\boldsymbol{\beta}})}}{\frac{\exp(\boldsymbol{x}_k\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_k\hat{\boldsymbol{\beta}})}} = \sum_{k=1}^{N_{obs}} \frac{1}{\exp(\boldsymbol{x}_k\hat{\boldsymbol{\beta}})} \\ \hat{N} &= N_{obs} + \hat{\boldsymbol{f}}_0 = \sum_{k=1}^{N_{obs}} \frac{1 + \exp(\boldsymbol{x}_k\hat{\boldsymbol{\beta}})}{\exp(\boldsymbol{x}_k\hat{\boldsymbol{\beta}})} \end{aligned} \tag{2.39}$$

Estymator wariancji otrzymany zostaje poprzez uproszczenie równań od (2.31) do (2.34) i opuszczenie pochodnych względem t

$$\begin{aligned} &\operatorname{var}(\mathbb{E}(\hat{N}|I_1, \ldots, I_N)) \approx \\ &\left(\frac{\partial \hat{N}|I_1, \ldots, I_N}{\partial \boldsymbol{\beta}}\right)^T \left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^T \boldsymbol{\beta}}\right)^{-1} \left(\frac{\partial \hat{N}|I_1, \ldots, I_N}{\partial \boldsymbol{\beta}}\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \end{aligned} \tag{2.40}$$

$$\begin{aligned} \mathbb{E}(\operatorname{var}(\hat{N}|I_1, \ldots, I_N)) &= \sum_{k=1}^{N} \frac{1 - p_k(\hat{\boldsymbol{\beta}})}{p_k(\hat{\boldsymbol{\beta}})} \\ &\approx \sum_{k=1}^{N} I_k \frac{1 - p_k(\hat{\boldsymbol{\beta}})}{p_k^2(\hat{\boldsymbol{\beta}})} \\ &= \sum_{k=1}^{N_{obs}} \frac{1 + \exp(\boldsymbol{x}_k\boldsymbol{\beta})}{\exp(2\boldsymbol{x}_k\boldsymbol{\beta})} \end{aligned} \tag{2.41}$$

Pochodna z równania (2.40) to pochodna względem wektora $\boldsymbol{\beta}$ z (2.32) obliczona w punkcie $t = 0$, która upraszcza się do:

$$\frac{\partial \hat{N}|I_1, \ldots, I_N}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^T \exp\left(-\boldsymbol{X}\boldsymbol{\beta}\right) \tag{2.42}$$

Rozdział 3

Modele logistyczne

Modele Chao i Zeltermana

Motywacja

Estymatory Zeltermana i Chao zostały po raz pierwszy wyprowadzone w publikacjach odpowiednio Zelterman 1988 Chao 1989.
Oba estymatory nie wymagały oryginalnie estymacji parametru można było je więc traktować jako estymatory nieparametryczne, chociaż były wyprowadzane z założeniem rozkładu Poissona, zamiast parametru korzystały tylko z wielkości $\boldsymbol{f}_1$ oraz $\boldsymbol{f}_2$ co czyniło je odpornymi na obserwacje odstające.
Dodatkowym powodem dla użycia tych estymatorów, jest podejrzenie, że jednostki niezaobserwowane mogą mieć więcej wspólnego z jednostkami o niskiej frekwencji występowania w zbiorze danych, niż tym z większą.
Przykładowo dla osób, które popełniły kilka wykroczeń na drodze, ale nie zostały złapane można podejrzewać większe podobieństwo z osobami, które zostały złapane raz lub dwa niż z notorycznymi piratami drogowymi.
Jedną z wad klasycznych estymatorów, jest brak uwzględnienia informacji dodatkowej, w nowszych pracach Böhning and P. G. M. v. d. Heijden 2009 oraz Böhning, Vidal-Diez, et al. 2013 uogólniono te estymatory w celu uwzględnienia zaobserwowanej niejednorodności, zachowując odporność na obserwacje odstające.

Klasyczne estymatory

Estymator Zeltermana

W artykule Zelterman 1988 autor argumentuje, że dla danych w których zmienna zależna Y prezentuje realizację pewnego rozkład uciętego w zerze, założenie rozkładu Poissona może być poprawne tylko dla niektórych wartości zmiennej, na przykład na niewielkich przedziałach, ale nie dla wszystkich wartości. Jeżeli rozważymy wartości w zakresie $[j, j+1] \cap \mathbb{N}$ dla pewnego $j \in \mathbb{N}$ metoda zaprezentowana przez Zeltermana podprowadzi nas do estymatorów dla λ i N wykorzystujących tylko frekwencje $\boldsymbol{f}_j$ i $\boldsymbol{f}_{j+1}$. Przy założeniu rozkładu Poissona następujące 2 systemy równań są spełnione:

$$\begin{aligned}\frac{\mathbb{P}(Y=j+1|\lambda)}{\mathbb{P}(Y=j|\lambda)} &= \left(\frac{\lambda^{j+1}e^{-\lambda}}{(j+1)!}\right) \Big/ \left(\frac{\lambda^{j}e^{-\lambda}}{j!}\right) \\ &= \frac{\lambda \cancel{\lambda^{j} e^{-\lambda}}}{(j+1)\cancel{j!}} \frac{\cancel{j!}}{\cancel{\lambda^{j} e^{-\lambda}}} = \frac{\lambda}{j+1}\end{aligned} \tag{3.1}$$

$$\begin{aligned}\frac{\mathbb{P}(Y=j+1|\lambda, Y>0)}{\mathbb{P}(Y=j|\lambda, Y>0)} &= \left(\frac{\lambda^{j+1}}{(j+1)!\,(e^{\lambda}-1)}\right) \Big/ \left(\frac{\lambda^{j}}{j!\,(e^{\lambda}-1)}\right) \\ &= \frac{\lambda \cancel{\lambda^{j}}}{\cancel{j!\left(e^{\lambda}-1\right)}(j+1)} \frac{\cancel{j!\left(e^{\lambda}-1\right)}}{\cancel{\lambda^{j}}} = \frac{\lambda}{j+1}\end{aligned} \tag{3.2}$$

Jeżeli zastąpimy prawdopodobieństwa $\mathbb{P}(Y=j+1|\lambda, Y>0), \mathbb{P}(Y=j|\lambda, Y>0)$ ich empirycznymi estymatorami, odpowiednio $\boldsymbol{f}_{j+1}\big/N_{obs}, \boldsymbol{f}_{j}\big/N_{obs}$ zostanie otrzymany estymator dla parametru λ zależny od wyboru wielkości j postaci:

$$\hat{\lambda}_j = (j+1)\frac{\boldsymbol{f}_{j+1}\big/N_{obs}}{\boldsymbol{f}_{j}\big/N_{obs}} = \frac{(j+1)\boldsymbol{f}_{j+1}}{\boldsymbol{f}_{j}} \tag{3.3}$$

Najczęściej przypadek w którym $j=1$ jest rozważany, otrzymany zostanie estymator Zeltermana dla λ $\hat{\lambda}_1 = 2\boldsymbol{f}_2\big/\boldsymbol{f}_1$ co prowadzi do estymatora dla wielkości N poprzez zastąpienie estymatora parametru $\lambda, \exp(-\boldsymbol{x}_k\boldsymbol{\beta})$ przez estymator $\hat{\lambda}_1$ w wyrażeniu (2.9), w końcowej postaci przybierającego formę:

$$\begin{aligned}\hat{\boldsymbol{f}}_0 &= N_{obs}\frac{\exp\left(-\hat{\lambda}_1\right)}{1-\exp\left(-\hat{\lambda}_1\right)} = N_{obs}\frac{\exp\left(-2\frac{\boldsymbol{f}_2}{\boldsymbol{f}_1}\right)}{1-\exp\left(-2\frac{\boldsymbol{f}_2}{\boldsymbol{f}_1}\right)} \\ \hat{N} &= N_{obs} + \hat{\boldsymbol{f}}_0 = \frac{N_{obs}}{1-\exp\left(-2\frac{\boldsymbol{f}_2}{\boldsymbol{f}_1}\right)}\end{aligned} \tag{3.4}$$

W pracy Böhning and P. G. M. v. d. Heijden 2009 wyprowadzony został estymator błędu standardowego statystyki $\hat{\lambda}_1$:

$$\sqrt{4\frac{\boldsymbol{f}_2(\boldsymbol{f}_1+\boldsymbol{f}_2)}{\boldsymbol{f}_1^3}}$$

prowadzący do przedziału ufności dla parametru λ na poziomie istotności $(1-\alpha)\cdot 100\%$ wyrażonego poprzez:

$$\left(\hat{\lambda}_1 - z\left(1-\frac{\alpha}{2}\right)\sqrt{4\frac{\boldsymbol{f}_2(\boldsymbol{f}_1+\boldsymbol{f}_2)}{\boldsymbol{f}_1^3}}, \hat{\lambda}_1 + z\left(1-\frac{\alpha}{2}\right)\sqrt{4\frac{\boldsymbol{f}_2(\boldsymbol{f}_1+\boldsymbol{f}_2)}{\boldsymbol{f}_1^3}}\right) \tag{3.5}$$

Przy założeniu normalności, lub chociaż asymptotycznej normalności, statystyki $\hat{\lambda}_1$.

Estymator Chao

W celu wyprowadzenia estymatora chao przyjmijmy, że λ jest zmienną losową opisaną funkcją prawdopodobieństwa q, $Y|\lambda$ jest zmienną losową o rozkładzie Poissona, oznaczmy łączną funkcję prawdopodobieństwa dla λ i Y poprzez m. Zachodzi wówczas zależność:

$$m(y, \lambda) = \mathbb{P}(Y = y|\lambda)q(\lambda) \tag{3.6}$$

użyta zostanie nierówność Cauchy'ego-Schwarza w postaci:

$$\left(\int_0^\infty f(x)g(x)dx\right)^2 \leqslant \left(\int_0^\infty (g(x))^2 dx\right)\left(\int_0^\infty (f(x))^2 dx\right) \tag{3.7}$$

Postawmy $x = \lambda, f(\lambda) = \sqrt{m(0,\lambda)}, g(\lambda) = \sqrt{2m(2,\lambda)}$, pomiędzy tymi funkcjami zachodzą następujące własności:

$$\begin{aligned} m(0,\lambda) &= \mathbb{P}(y = 0|\lambda)q(\lambda) = e^{-\lambda}q(\lambda) \\ m(2,\lambda) &= \mathbb{P}(y = 2|\lambda) = \frac{e^{-\lambda}\lambda^2}{2}q(\lambda) \\ (m(1,\lambda))^2 &= \left(\mathbb{P}(y = 1|\lambda)q(\lambda)\right)^2 = \lambda^2 e^{-2\lambda}(q(\lambda))^2 \\ &= m(0,\lambda) \cdot 2m(2,\lambda)(q(\lambda))^2 \\ &\implies m(1,\lambda) = f(\lambda)g(\lambda) \end{aligned}$$

Podstawiając powyższe funkcje do nierówności (3.7) otrzymana zostanie zależność:

$$\begin{aligned} (\mathbb{P}(Y = 1))^2 &= \left(\int_0^\infty \lambda e^{-\lambda} q(\lambda)d\lambda\right)^2 \leqslant \\ &\left(2\int_0^\infty \frac{e^{-\lambda}\lambda^2}{2}q(\lambda)d\lambda\right) \cdot \left(\int_0^\infty e^{-\lambda}q(\lambda)d\lambda\right) \\ &= 2\mathbb{P}(Y = 2)\mathbb{P}(Y = 0) \end{aligned} \tag{3.8}$$

Poprzez pomnożenie obu stron nierówności (3.8) przez N_{obs}^2 i zastąpienie wielkości $N_{obs}\mathbb{P}(Y = j)$ przez częstości $\boldsymbol{f}_j$ (3.8) sprowadza się do:

$$\boldsymbol{f}_1^2 \leqslant 2\boldsymbol{f}_2\boldsymbol{f}_0 \implies \frac{\boldsymbol{f}_1^2}{\boldsymbol{f}_2} \leqslant \boldsymbol{f}_0 \tag{3.9}$$

Co daje nam estymatory $\hat{\boldsymbol{f}}_0$ ograniczające $\boldsymbol{f}_0, N$ oddolnie:

$$\hat{\boldsymbol{f}}_0 = \frac{\boldsymbol{f}_1^2}{\boldsymbol{f}_2} \implies \hat{N} = N_{obs} + \frac{\boldsymbol{f}_1^2}{\boldsymbol{f}_2}$$

To podejście do estymacji rozmiaru populacji ma dwa istotne własności, po pierwsze jak już wspomniano $\hat{N}$ ogranicza, przynajmniej teorytyczny, N oddolnie co daje nam wiarygodny estymator, w tym sensie, że możemy całkiem

bezpiecznie założyć, że faktyczna wartość N jest większa niż estymator. Po drugie estymator bierze pod uwagę część niejednorodności, ponieważ dopuszcza się aby λ różniło się pomiędzy jednostkami jako zmienna o rozkładzie opisanym przez q.
Estymator wariancji estymatora $\hat{\boldsymbol{f}}_0$ został wyprowadzony w pracy Chao 1989 jako:

$$\operatorname{var}(\hat{N}) = \boldsymbol{f}_2 \left(\frac{1}{4} \left(\frac{\boldsymbol{f}_1}{\boldsymbol{f}_2} \right)^4 + \left(\frac{\boldsymbol{f}_1}{\boldsymbol{f}_2} \right)^3 + \frac{1}{2} \left(\frac{\boldsymbol{f}_1}{\boldsymbol{f}_2} \right)^2 \right) \tag{3.10}$$

Dokładniejsze porównanie powyższych estymatorów znajduje się w pracy Böhning 2010.

Model Regresji

Opisany poniżej model Regresji użyty zostanie w uogólnionych modelach Chao i Zeltermana.
Niech Y będzie zmienną losową o rozkładzie Poissona. Utwórzmy nową zmienną Z utworzoną ze zmiennej Y ograniczonej do wartości $1, 2$ i spełniającą:

$$Z = \begin{cases} 0 & \text{jeżeli } Y = 1 \\ 1 & \text{jeżeli } Y = 2 \end{cases} \tag{3.11}$$

Ponieważ

$$\frac{\mathbb{P}(Y = 2|\lambda)}{\mathbb{P}(Y = 1|\lambda) + \mathbb{P}(Y = 2|\lambda)} = \frac{\frac{1}{2}\lambda^2 e^{-\lambda}}{\lambda e^{-\lambda} + \frac{1}{2}\lambda^2 e^{-\lambda}} = \frac{\frac{\lambda}{2}}{1 + \frac{\lambda}{2}}$$

Zmienna Z ma rozkład bernoulliego z parametrem $p = \frac{\frac{\lambda}{2}}{1+\frac{\lambda}{2}}$.
Ograniczmy wektory $\boldsymbol{x}_i$ do tych wektorów, które są związane z jednostkami dla których $y_i \in \{1, 2\}$. Rozważmy model regresji z następującą funkcją wiążącą:

$$\operatorname{logit}^4(p_k) = \ln\left(\frac{\lambda_k}{2}\right) = \boldsymbol{x}_k\boldsymbol{\beta} = \eta_k \tag{3.12}$$

Funkcja wiarygodności oraz jej logarytm wyrażają się poprzez:

$$L = \prod_{k=1}^{\boldsymbol{f}_1+\boldsymbol{f}_2} \left(\frac{\exp(\eta_k)}{1+\exp(\eta_k)} \right)^{z_k} \left(\frac{1}{1+\exp(\eta_k)} \right)^{1-z_k} \tag{3.13}$$

$$\ell = \sum_{k=1}^{\boldsymbol{f}_1+\boldsymbol{f}_2} \left(z_k \ln\left(\frac{\exp(\eta_k)}{1+\exp(\eta_k)} \right) + (1 - z_k) \ln\left(\frac{1}{1+\exp(\eta_k)} \right) \right) \tag{3.14}$$

[4] $\operatorname{logit}(x) = \ln\left(\frac{x}{1-x}\right) \quad \operatorname{logit}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$

$$= \sum_{k=1}^{\boldsymbol{f}_1+\boldsymbol{f}_2} \left(z_k \ln \left(\frac{\exp(\boldsymbol{x}_k\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_k\boldsymbol{\beta})} \right) + (1-z_k) \ln \left(\frac{1}{1+\exp(\boldsymbol{x}_k\boldsymbol{\beta})} \right) \right)$$

Gdzie wektor $\boldsymbol{z}$ został utworzony z wektora $\boldsymbol{y}$ w sposób opisany przez (3.11). Pochodne funkcji ℓ dla regresji logistycznej są dobrze znane i przybierają postacie:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \left(\frac{\exp(\boldsymbol{X}\boldsymbol{\beta}) \times (\boldsymbol{z}-\boldsymbol{1}) + \boldsymbol{z}}{\boldsymbol{1}+\exp(\boldsymbol{X}\boldsymbol{\beta})} \right) = \boldsymbol{X}^T \left(\boldsymbol{z} - \frac{\exp(\boldsymbol{X}\boldsymbol{\beta})}{\boldsymbol{1}+\exp(\boldsymbol{X}\boldsymbol{\beta})} \right) \tag{3.15}$$

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = -\left(\boldsymbol{x}_{(i)} \times \frac{\exp(\boldsymbol{X}\boldsymbol{\beta})}{(\boldsymbol{1}+\exp(\boldsymbol{X}\boldsymbol{\beta}))^2} \right)^T \boldsymbol{X} \tag{3.16}$$

Uogólniony estymator Chao

Estymator rozmiaru populacji uogólniony w pracy Böhning, Vidal-Diez, et al. 2013 wyraża się poprzez:

$$\begin{aligned} \hat{N} &= N_{obs} + \sum_{k=1}^{\boldsymbol{f}_1+\boldsymbol{f}_2} \frac{I_k}{2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right) + 2\exp\left(2\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)} \\ &= N_{obs} - (\boldsymbol{f}_1+\boldsymbol{f}_2) + \sum_{k=1}^{\boldsymbol{f}_1+\boldsymbol{f}_2} I_k \left(1 + \frac{\mathbb{P}\left(y_k=0|\hat{\lambda}_k\right)}{\mathbb{P}\left(y_k=1|\hat{\lambda}_k\right) + \mathbb{P}\left(y_k=2|\hat{\lambda}_k\right)} \right) \end{aligned} \tag{3.17}$$

gdzie $\mathbb{P}$ jest funkcją prawdopodobieństwa nieuciętego rozkładu Poissona[5]. Estymacja wariancji i utworzenie przedziałów ufności odbywa się w ten sam sposób jak w poprzednich wypadkach z tą różnicą, że zmienne indykatorowe mają rozkład z innym parametrem.

$$\operatorname{var}(\hat{N}) = \mathbb{E}(\operatorname{var}(\hat{N}|I_1,\ldots,I_N)) + \operatorname{var}(\mathbb{E}(\hat{N}|I_1,\ldots,I_N)) \tag{3.18}$$

$$\operatorname{var}(\mathbb{E}(\hat{N}|I_1,\ldots,I_N)) \approx \sum_{k=1}^{\boldsymbol{f}_1+\boldsymbol{f}_2} (1-p_k(\hat{\boldsymbol{\beta}})) \left(1 + \frac{\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)}{p_k(\hat{\boldsymbol{\beta}})} \right)^2 \tag{3.19}$$

$$\begin{gathered} \mathbb{E}(\operatorname{var}(\hat{N}|I_1,\ldots,I_N)) \approx \\ \left(\frac{\partial \hat{N}|I_1,\ldots,I_N}{\partial \boldsymbol{\beta}} \right)^T \left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} \right)^{-1} \left(\frac{\partial \hat{N}|I_1,\ldots,I_N}{\partial \boldsymbol{\beta}} \right) \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \end{gathered} \tag{3.20}$$

[5] W tym wypadku przyjmujemy $\hat{\lambda}_k = 2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)$ z równania (3.12)

Gdzie:

$$p_k(\boldsymbol{\beta}) = 2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right) + 2\exp\left(2\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right) \quad I_k \sim b\left(p_k\left(\hat{\boldsymbol{\beta}}\right)\right)$$

$$\frac{\partial \hat{N}|I_1,\ldots,I_N}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^T\left(\frac{2\exp\left(\boldsymbol{X}\boldsymbol{\beta}\right)+4\exp\left(2\boldsymbol{X}\boldsymbol{\beta}\right)}{\left(2\exp\left(\boldsymbol{X}\boldsymbol{\beta}\right)+2\exp\left(2\boldsymbol{X}\boldsymbol{\beta}\right)\right)^2}\right)$$

Uogólniony model Zeltermana

Estymator rozmiaru populacji uogólniony w pracy Böhning and P. G. M. v. d. Heijden 2009 wyraża się poprzez:

$$\begin{aligned}\hat{N} &= \sum_{k=1}^{N}\frac{I_k}{1-\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)}\\ \hat{\boldsymbol{f}}_0 &= \sum_{k=1}^{N} I_k\frac{\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)}{1-\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)}\end{aligned} \tag{3.21}$$

Ważną własnością tego estymatora jest fakt, że pomimo przeprowadzenia regresji tylko na jednostkach o wartościach $y_i \in \{1,2\}$ aczkolwiek sam estymator wykorzystuje wszystkie obserwacje.
Estymacja wariancji i tworzenie przedziałów ufności odbywa się w ten sam sposób z którego korzystaliśmy dla modelu Poissona jedyna różnica polega na tym, że zmienne indykatorowe mają rozkład $I_k \sim b\left(1-\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)\right)$.

$$\operatorname{var}(\hat{N}) = \mathbb{E}(\operatorname{var}(\hat{N}|I_1,\ldots,I_N)) + \operatorname{var}(\mathbb{E}(\hat{N}|I_1,\ldots,I_N)) \tag{3.22}$$

$$\operatorname{var}(\mathbb{E}(\hat{N}|I_1,\ldots,I_N)) = \sum_{k=1}^{N_{obs}}\frac{\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)}{\left(1-\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)\right)^2} \tag{3.23}$$

$$\mathbb{E}(\operatorname{var}(\hat{N}|I_1,\ldots,I_N)) \approx \left(\frac{\partial \hat{N}|I_1,\ldots,I_N}{\partial \boldsymbol{\beta}}\right)^T\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^T\partial \boldsymbol{\beta}}\right)^{-1}\left(\frac{\partial \hat{N}|I_1,\ldots,I_N}{\partial \boldsymbol{\beta}}\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \tag{3.24}$$

Gdzie:

$$\frac{\partial \hat{N}|I_1,\ldots,I_N}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^T\left(\frac{\exp\left(-2\exp\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)}{\left(1-\exp\left(-2\exp\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)\right)^2}\times 2\exp\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)$$

Opis funkcji w R i wybór modelu

SingleRcapture

W celu użycia zaprezentowanych modeli skorzystamy z pakietu singleRcapture w wersji 0.1.0 dla języka R w którym opisane modele są zaimplementowane, zawiera on także dwa zbiory danych na których przedstawione zostanie działanie opisanych modeli. Dodatkowo w celu wizualizacji danych i wyników skorzystano z pakietu ggplot2, będącego częścią biblioteki tidyverse, oraz funkcję matplot z domyślnie ładowanego pakietu graphics. Wymienione pakiety mogą zostać instalowane poprzez wywołanie:

```
devtools::install_github("ncn-foreighners/singleRcapture")
utils::install.packages("tidyverse")
```

Wykorzystane zostały następujące funkcje z pakietu singleRcapture:

— singleRcapture::estimate_popsize - główna funkcja tworząca macierz $\boldsymbol{X}$, estymująca wektor $\boldsymbol{\beta}$ (estymacja odbywa się na poziomie błędu równej podwojonej wartości epsilonu maszynowego około $2.220446 \cdot 10^{-16}$), tworząca estymator $\hat{N}$ oraz przedział ufności związany z tym estymatorem, oraz w razie potrzeby wybierająca potrzebne dane (w przypadku modelu chao $(y_k, \boldsymbol{x}_k)$ spełniające $y_k \in \{1, 2\}$)
— singleRcapture::summary.singleR - metoda dla obiektów klasy singleR zwracanych przez funkcję singleRcapture::estimate_popsize, pozwalająca na przejrzyste spojrzenie na wynik regresji i estymacji
— singleRcapture::marginalFreq - funkcja w celu znalezienia marginalnych frekwencji dla dopasowanego modelu
— singleRcapture::summary.singleRmargin - metoda dla klasy singleRmargin wykorzystana w celu uzyskania wyniku testu zgodności χ^2

Przy wybranej metodzie estymacji wektora $\boldsymbol{\beta}$ funkcja singleRcapture::estimate_popsize deleguje optymalizacje $\ell(\boldsymbol{\beta})$ funkcji stats::optim implementującą metodę numeryczną L-BFGS-B (w przypadku modelu ujemnego dwumianowego metodę Neldera-Meada) będącą modyfika-

cją metody Newthona-Raphsona. Po wywołaniu powyższej funkcji otrzymany zostanie wektor $\hat{\boldsymbol{\beta}}$ spełniający:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big(\ell(\boldsymbol{\beta})\Big) \tag{4.1}$$

gdzie p jest liczbą parametrów w rozważanym modelu.

Kryteria wyboru modelu

Zagadnienie wyboru jednego, możliwie najlepszego, spośród podanych modeli dla konkretnych danych sprowadza się do wyboru najlepszego modelu regresji dla dostępnych danych z ewentualnych uwzględnieniem kwestii spełnienia założeń modelu przez dane. Wybór modelu regresji sam z siebie jest tematem całkiem skomplikowanym i istnieje wiele idei jak takiego wyboru dokonać, nie dokonane zostanie zatem w tej pracy specjalnie rozległe omówienie tematu. Pozostawia to także kwestię wyboru zaobserwowanych zmiennych dodatkowych do odrzucenia lub zachowania w modelu. Zadanie to utrudnione jest w tym przypadku ponieważ regresja dla modeli Zeltermana i Chao przeprowadzona jest na części oryginalnego zbioru danych. W celu wybrania, które zmienne są istotne dla każdego modelu zastosowano podejście polegające na minimalizacji wielkości:

$$^{6}\mathrm{AIC} = 2p - 2\ell(\hat{\boldsymbol{\beta}})$$

znanej jako kryterium informacyjne Akaike, pomagające uniknięcia przetrenowania modelu, ponieważ wraz ze wzrostem liczby parametrów podana wielkość wzrasta, jeżeli te dodatkowe parametry nie są w stanie poprawić wyraźnie dopasowania modelu. Kryterium informacji Akaike prezentuje więc kompromis pomiędzy dopasowaniem modelu a możliwością przetrenowania.
Do wyboru końcowego modelu użyty może zostać test zgodności G, którego wyniki zostaną zaprezentowane przy poziomie istotności 5%, lub test χ^2. Są to tylko dwie z wielu metod a wybór najlepszego modelu może się różnić w zależności od przyjętego kryterium.

[6] W tym równaniu wektor $\hat{\boldsymbol{\beta}}$ należy rozumieć jako wektor minimalizujący funkcję ℓ uwzględniającą dodatkowe parametry poprzez rozszerzenie macierzy $\boldsymbol{X}$ o dodatkowe informacje.

Przykłady

Dane dotyczące nieregularnych imigrantów w Holandii

Opis zbioru

Zbiór danych pochodzi z pracy P. G. v. d. Heijden et al. 2003 oraz Böhning and P. G. M. v. d. Heijden 2009 zawiera on dane o nieregularnych imigrantach na terenie kilku miast Holandii (Amsterdam, Rotterdam, Utrecht i Haga) z roku 1995, otrzymane z informacji holenderskiej policji, zawiera on zmienne:

- capture - ilość razy dana osoba została zatrzymana zmienna ilościowa,
- gender - płeć osoby zatrzymanej, zakodowana jako 0 dla kobiet i 1 dla mężczyzn,
- age - wiek osoby zatrzymanej zakodowany jako 0 dla osób starszych niż 40 lat i 1 dla osób młodszych niż 40 lat
- nation - region z którego pochodzi zatrzymana osoba
- reason - powód zatrzymania 1 dla nielegalnego przebywania na terenie Holandii 0 w przypadku innego powodu.

W celu przeprowadzenia regresji dla zmiennej nation jedna wartość, konkretnie “American and Australia“, została wybrana jako wartość bazowa, a reszta została przekształcona w 5 kolumn wskazujących, z którego regionu pochodzi dana osoba przykładowo $0, 0, 0, 0, 0$ dla osoby z Australii i $0, 0, 0, 0, 1$ dla osoby z Turcji.

Niestety model ujemny dwumianowy[7] nie mógł zostać dopasowany, ponieważ algorytm nie mógł znaleźć stabilnego minimum dla logarytmu wiarygodności. Dla każdego modelu wybrane zostały zmienne, które minimalizują kryterium informacji, w przypadku modeli Poissona i Geometrycznego odrzucona została zmienna “reason“ co zaskutkowało obniżeniem AIC o odpowiednio 1.995 i 1.976, w obu przypadkach zmienna “reason“ nie była istotna statystycznie (przy 5% poziomie istotności), p-wartości testu t-studenta to odpowiednio 0.95 i 0.88. Dla modeli Chao i Zeltermana odrzucone zostały zmienne “reason“ i “age“, co zaskutkowało obniżeniem AIC o 1.306, przy p-wartościach odpowied-

[7] Podobny problem pojawił się także w jednej z prac w, których korzystano z tego zbioru danych.

nio 0.39 i 0.19. Wyniki dla każdego dopasowanego modelu zostały przedstawione na ilustacji, przedstawionej na końcu sekcji.

Opis estymacji

Poniższa tabela zawiera marginalne liczebności estymowane za pomocą każdego modelu oraz zaobserwowane w danych.

	$\boldsymbol{f}_0$	$\boldsymbol{f}_1$	$\boldsymbol{f}_2$	$\boldsymbol{f}_3$	$\boldsymbol{f}_4$	$\boldsymbol{f}_5$	$\boldsymbol{f}_6$
Zaobserwowane	0	1645	183	37	13	1	1
Model Poissona	10810.4	1612.6	233.7	30.1	3.2	0.3	0.0
Model Chao	13732.53	1672.4	186.6	19.2	1.7	0.1	0.0
Model Zeltermana	13936.1	1672.4	186.6	19.2	1.7	0.1	0.0
Model Geometryczny	22783.8	1627.9	209.4	34.8	6.4	1.2	0.2

Znajomość liczebności marginalnych daje nam możliwość wykonania testu zgodności G. Wyniki zostały przedstawione w tabeli poniżej zawierają wartość statystyki testowej[8] ilość stopni swobody oraz p-wartość testu, z dodatkiem wartości kryterium informacyjnego Akaike dla modelu:

	G	df	P	AIC
Model Poissona	34.31	2	$3.6 \cdot 10^{-8}$	1712.901
Model Chao	50.61	2	$1 \cdot 10^{-11}$	1131.723
Model Zeltermana	50.61	2	$1 \cdot 10^{-11}$	1131.723
Model Geometryczny	9.28	3	0.026	1684.904

Niestety wszystkie modele zostały odrzucone przez test zgodności G. Możliwe jest, że wynika to ze złamania założeń modelu. Statystyka testu G (jak również statystyka testu χ^2) mają najniższe wartości dla modelu geometrycznego. Dodatkowo w pracy P. G. v. d. Heijden et al. 2003 autorzy sugerują występowanie niezaobserwowanej niejednorodności, z którą najlepiej powinien poradzić sobie model geometryczny. Z tych powodów model geometryczny wydaje się być najsensowniejszym wyborem dla naszych danych[9] zostaną przedstawione zatem wyniki dla tego modelu, z tym zastrzeżeniem, że najprawdopodobniej nie jest on poprawny jest tylko najlepszą możliwą aproksymacją rzeczywistości

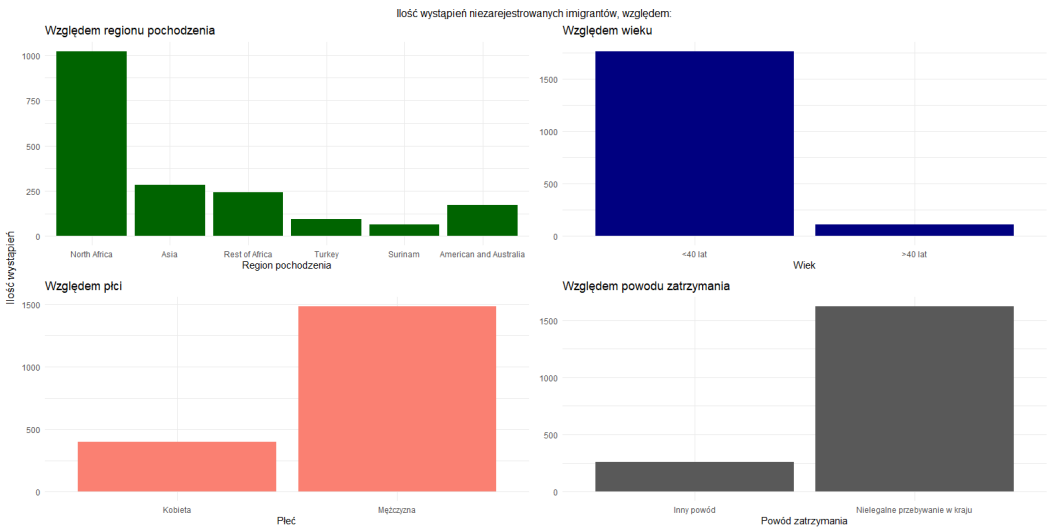
[8] Wielkości $\boldsymbol{f}_0$ zostały zignorowane, a elementy o liczebności mniejszej niż 5 zostały pogrupowane w najbliższą komórkę o liczebności większej niż 5. Nie została zastosowana poprawka na liczbę stopni swobody. Dla modelu geometrycznego wykorzystano wartości $\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \boldsymbol{f}_{4\leqslant}$ oraz $\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_{3\leqslant}$ dla innych modeli.

[9] Wartym odnotowania jest fakt, że modele Zeltermana i Chao, miałyby niższą wartość statystyk G i χ^2 niż model geometryczny gdyby ograniczyć sumę przy obliczaniu statystyk do $\boldsymbol{f}_1$ i $\boldsymbol{f}_2$

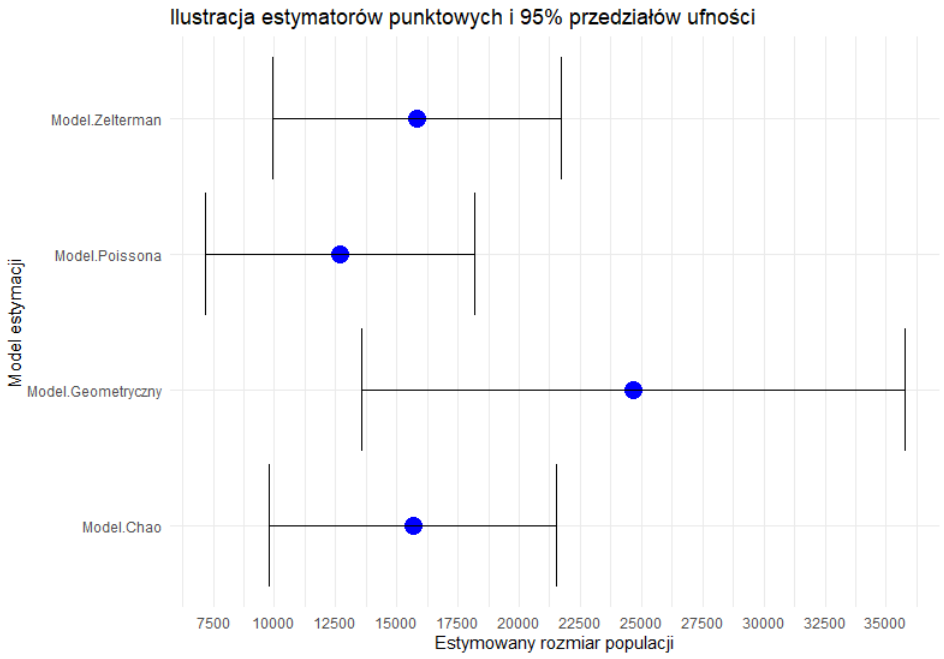
dostępną w kontekście zaprezentowanych modeli.
Model geometryczny sugeruje, że tylko część populacji została zaobserwowana $\approx 7.62\%$ z przedziałem ufności $(5.26\%, 13.686\%)$. Poniżej zaprezentowany został kod języka R prezentujący detale modelu geometrycznego:

```
summary(
estimate_popsize(
 formula = capture ~ . - reason,
 model = "ztgeom",
 data = netherlandsimmigrant,
 method = "mle",
 pop.var = "analytic"
 )
)
```

```
Response Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7470  0.7470  0.9066  1.0002  0.9553  5.7470

Coefficients:
                    Estimate Std. Error z value P(>|z|)
(Intercept)           -2.991      0.476   -6.28 3.4e-10 ***
gender                 0.406      0.179    2.27 2.3e-02   *
age                    0.996      0.429    2.32 2.0e-02   *
nationAsia            -1.114      0.322   -3.46 5.5e-04 ***
nationNorth Africa     0.214      0.217    0.99 3.2e-01
nationRest of Africa  -0.934      0.323   -2.89 3.8e-03  **
nationSurinam         -2.366      1.028   -2.30 2.1e-02   *
nationTurkey          -1.713      0.622   -2.75 5.9e-03  **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
                0.05 '*' 0.1 '.' 1 ' '

AIC: 1684.904
BIC: 1729.216
Deviance: 0

Log-likelihood: -834.4521 on 1872 Degrees of freedom
Number of iterations: 142
---
Population size estimation results:
Point estimate 24663.81
Variance 32092025
Std. Error 5664.982
95% CI:
             lowerBound upperBound
Studentized     13560.65   35766.98
Logtransform    15977.72   38701.70
```



Rysunek 5.1. Wykresy opisujące rozkład zmiennej “capture“ względem różnych wartości zmiennych opisujących.



Rysunek 5.2. Wykres przedstawiający estymatory przedziałowe i punktowe dla dopasowanych modeli modeli.
Niebieski punkt symbolizuje estymator punktowy.

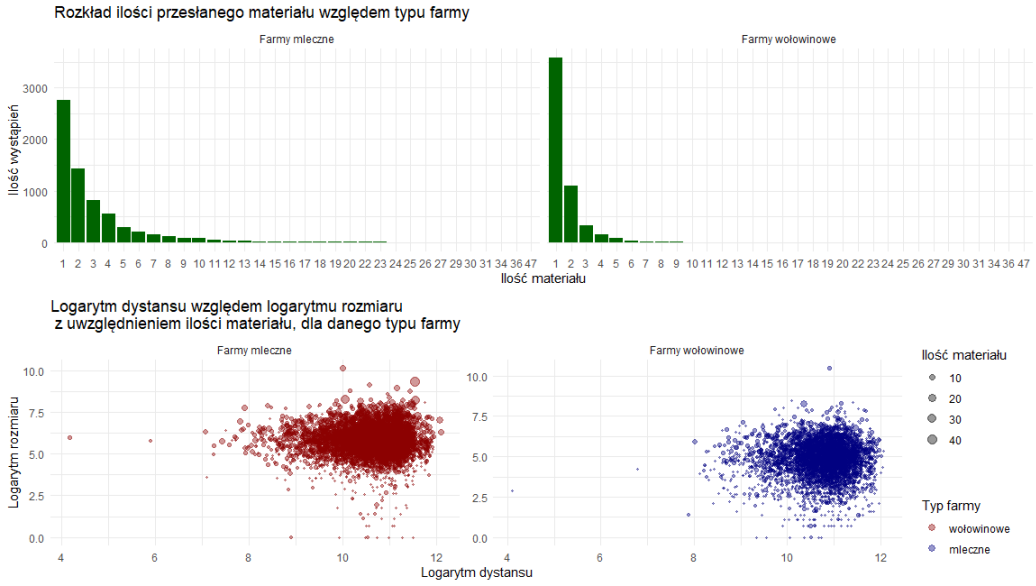
Dane dotyczące nadesłanego materiału z farm

Opis zbioru

Zbiór danych pochodzi z publikacji Böhning, Vidal-Diez, et al. 2013. Zawiera dane o przesłanym materiale z brytyjskich farm do AHVLA (Animal Health and Veterinary Laboratories Agency). W przypadku gdy prywatny weterynarz nie jest w stanie zidentyfikować powodu śmierci zwierzęcia brytyjskie farmy mogą wysłać próbki lub padlinę pochodzącą ze zwierzęcia do AHVLA, w przypadku gdy nie ma podejrzenia groźnej choroby decyzja przypada do weterynarza, w przeciwnych wypadku jest to wymagane. Zbiór danych zawiera informacje o takich farmach, z następującymi zmiennymi:

- TOTAL_SUB - Liczba przesłanych materiałów
- log_size - logarytm rozmiaru farmy
- log_distance - logarytm dystansu od najbliższego centrum AHVLA
- C_TYPE - Typ farmy zmienna zakodowana jako 1 dla farmy produkującej nabiał i 0 w innym przypadku

Rozkłady zmiennych zostały przedstawione na poniższej ilustracji.



Rysunek 5.3. Wykresy opisujące rozkład zmiennych.

Opis estymacji

Podobnie jak w poprzednim przypadku dokonano selekcji zmiennych dodatkowych, odrzucona została tylko zmienna log_distance w modelach Zelteramana i Chao, w tym wypadku 0.92 - p-wartość testu walda, obniżka AIC o 1.99. Tabela porównująca pierwsze sześć dopasowanych częstości marginalnych dla pięciu dopasowanych modeli oraz wartości empiryczne przybiera następującą postać

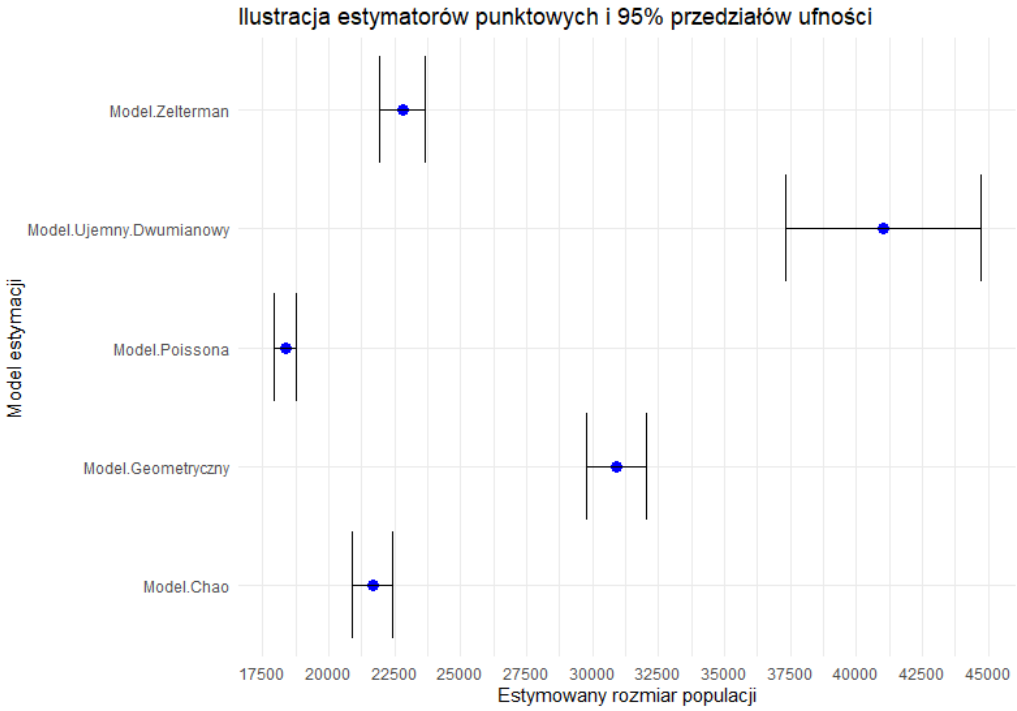
	$\boldsymbol{f}_0$	$\boldsymbol{f}_1$	$\boldsymbol{f}_2$	$\boldsymbol{f}_3$	$\boldsymbol{f}_4$	$\boldsymbol{f}_5$	$\boldsymbol{f}_6$	$\boldsymbol{f}_7$
Zaobserwowane	0	6340	2520	1149	709	380	249	173
Model Poissona	6309.7	4896.8	3057.8	1815.1	1040.7	574.7	307.6	161.5
Model Chao	9621.0	7469.5	3150.2	1033.4	288.9	72.3	16.8	3.8
Model Zeltermana	10754.9	7469.5	3150.2	1033.4	288.9	72.3	16.8	3.8
Model Geometryczny	18871.1	6046.0	2602.7	1319.1	739.1	442.7	277.9	180.8
Model Ujemny Dwumianowy	28984.3	6317.9	2485.3	1226.1	685.4	414.7	265.0	176.3

Wyniki testu zgodności G zostały przedstawione w tabeli poniżej zawierają wartość statystyki testowej[10] ilość stopni swobody oraz p-wartość testu, z dodatkiem wartości kryterium informacyjnego Akaike dla modelu:

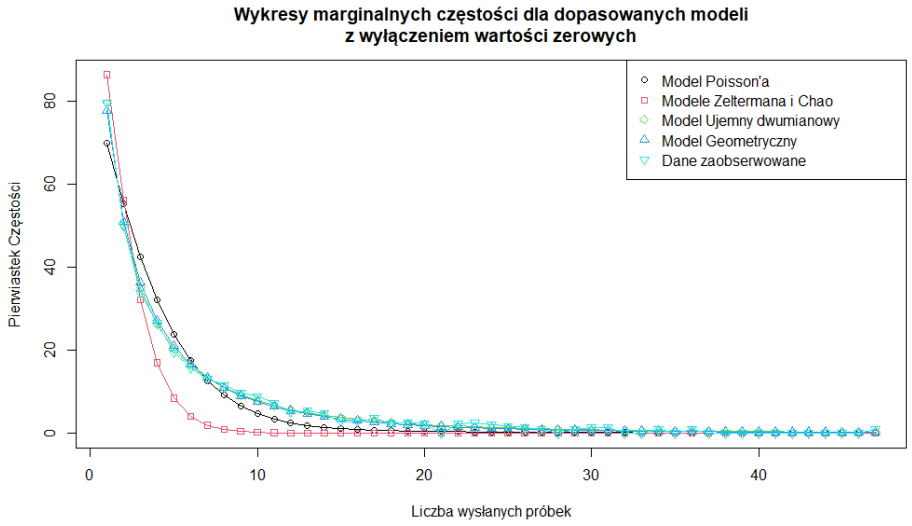
	G	df	P	AIC
Model Poissona	996.10	8	10^{-209}	39886.85
Model Chao	915.81	3	$3.3 \cdot 10^{-198}$	10293.7
Model Zeltermana	915.81	3	$3.3 \cdot 10^{-198}$	10293.7
Model Geometryczny	44.55	14	$4.8 \cdot 10^{-5}$	34621.42
Model Ujemny Dwumianowy	19.14	14	0.16	34538.73

Ponieważ tylko model ujemny dwumianowy nie został odrzucony w wyniku testu G, p-wartość jest zdecydowanie wyższa niż wartość progowa 0.05, najprawdopodobniej prezentuje on najlepszy estymator z dopasowanych modeli. Na następnych dwóch stronach znajdują się kod R prezentujący detale modelu ujemnego dwumianowego oraz ilustracje przedstawiające marginalne częstości oraz wyznaczone estymatory.

[10] Wielkości $\boldsymbol{f}_0$ zostały zignorowane, a elementy o liczebności mniejszej niż 5 zostały w tym przypadku zostały odrzucone, ponieważ w tym wypadku nie ma prostego sposobu na połączenie komórek. Poprawka na stopnie swobody z powodu estymacji parametrów została zastosowana



Rysunek 5.4. Wykres przedstawiający estymatory przedziałowe i punktowe dla dopasowanych modeli modeli. Niebieski punkt symbolizuje estymator punktowy.



Rysunek 5.5. Wykres przedstawiający pierwiastek marginalnych częstości dla dopasowanych modeli oraz dla wartości zaobserwowanych.

```
summary(
estimate_popsize(
 formula = TOTAL_SUB ~ .,
 model = "ztnegbin",
 data = farmsubmission,
 method = "mle",
 pop.var = "analytic"
 )
)
```

```
Response Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-13.0209   0.2472   0.7367   1.2950   1.6635  43.7657

Coefficients:
                Estimate Std. Error z value  P(>|z|)
log(dispersion)    0.582      0.073    7.97  1.5e-15  ***
(Intercept)       -3.171      0.259  -12.22  2.4e-34 ****
log_size           0.639      0.018   36.11 1.7e-285 ****
log_distance      -0.080      0.022   -3.58  3.4e-04  ***
C_TYPE             0.660      0.034   19.33  2.9e-83 ****
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**'
                0.05 '*' 0.1 '.' 1 ' '

AIC: 34538.73
BIC: 34575.71
Deviance: 0

Log-likelihood: -17264.37 on 24064 Degrees of freedom
Number of iterations: 1144
---
Population size estimation results:
Point estimate 41020.34
Variance 3563465
Std. Error 1887.714
95% CI:
             lowerBound upperBound
Studentized     37320.49   44720.19
Logtransform    37550.34   44962.27
```

Podsumowanie

Dobrze znanym problemem w zwyczajnej regresji Poissona jest nadmierna ilość jednostek dla których $y_k = 0$, podobnie w przedstawionych modelach wystąpić może problem zbyt wysokiej ilości jednostek spełniających $y_k = 1$. Ponieważ bardzo możliwym jest, że osoba przebywająca nielegalnie w danym kraju zostanie wydalona lub zacznie się ukrywać. Istnieje zatem bardzo realne podejrzenie, że w przypadku zbioru danych z imigrantami w Holandii występuje taki nadmiar. Jest to istotne, ponieważ jednym z założeń modelu Poissona było wykorzystanie rozkładu Poissona[11] w celu opisania zmiennej zależnej Y. Sprawia to, że podane estymatory mogą nie być poprawne[12]. Jest to też jeden z powodów, dla których model geometryczny był przeze mnie preferowany względem modeli Zeltermana i Chao. Jedno z możliwych rozwiązań modelu znajduje się w publikacji Böhning and P. G. M. v. d. Heijden 2019, w której udowodniono, że model uwzględniający nadmierną ilość "jedynek" jest tożsamy z modelem uciętym w zerze i jedynce w tym sensie, że logarytmy funkcji wiarygodności tych modeli różnią się o wartość niezależną od wektora $\boldsymbol{\beta}$. Inną metodą poradzenia sobie z tym problemem może być jawna estymacja parametru "inflacji", w taki sam sposób jak parametr α był estymowany w modelu ujemnym dwumianowym.

Najczęściej rozważanymi przedziałami ufności dla opisanych estymatorów w literaturze są studentyzowane przedziały ufności. Istotną ich wadą jest symetryczność względem estymatora punktowego, uznawane jest to za wadę, ponieważ w literaturze można spotkać się z podejrzeniem o dodatniej skośności rozkładu estymatora $\hat{N}$. Inny sposób konstruowania przedziałów ufności, zakładający asymptotyczną normalność statystyki $\ln\left(\hat{N} - N_{obs}\right)$, został wspomniany w pracy Chao 1989 i przybrał on postać:

$$\left(N_{obs} + \frac{\hat{N} - N_{obs}}{G}, N_{obs} + \left(\hat{N} - N_{obs}\right) G\right)$$

[11] Jest to też naturalnie problem dla innych modeli nie uwzględniających nnadmiaru obserwacji $y_k = 1$.

[12] estymatory Zeltermana i Chao zależą najsilniej od obserwacji z $\boldsymbol{f}_1$ i $\boldsymbol{f}_2$ są one najbardziej narażone na wpływ nadmiernej wartości $\boldsymbol{f}_1$ na końcowy wynik.

gdzie $G = \exp\left(z\left(1-\frac{\alpha}{2}\right)\sqrt{\ln\left(1+\frac{\hat{\sigma^2}}{\left(\hat{N}-N_{obs}\right)^2}\right)}\right)$ i $\hat{\sigma^2}$ jest estymatorem wariancji $\hat{N}$. Kolejną metodą jest konstrukcja przedziału ufności za pomocą bootstrapu, jednakże bootstrap w przypadku metody capture-recapture jest bardziej skomplikowany niż zazwyczaj ponieważ $N_{obs} = \sum_{k=1}^{N} I_k$ także jest zmienną losową oraz $\hat{N}$ zależy od N_{obs} poprzez zmienne indykatorowe $I_1, \ldots, I_N$, więc musi to zostać uwzględnione w procedurze bootstrapu. Zagadnienie to zostało omówione w pracach dla jednorodnej populacji Norris and Pollock 1996 oraz Zwane and Van der Heijden 2003 dla niejednorodnej populacji. Wyróżnione zostały 3 rodzaje bootstrapu dla metody capture-recapture. Nieparametryczna polegająca na pobraniu próby bootstrapowej z próby uczącej $\mathcal{L} = \{(y_1, \boldsymbol{x}_1), \ldots, (y_{N_{obs}}, \boldsymbol{x}_{N_{obs}})\}$ i wyliczeniu na nich estymatora $\hat{N}$, metoda ta zaniża wariancje estymatora $\hat{N}$. Semiparametryczna i w pełni parametryczna uwzględnia losowość N_{obs}, różnią się one tym, że metoda semiparametryczna zakłada poprawność estymatora $\hat{N}$ a metoda w pełni parametryczna poprawność rozkładu zakładanego przez model (oraz poprawność estymatora wielkości populacji)[13].

Częściej stosowane niż metody jednoźródłowe są metody estymacji wielkości populacji wykorzystujące wiele źródeł. Przykładowo jedna z takich metod została wykorzystana w publikacji Beręsewicz, Gudaszewski, and Szymkowiak 2019 w celu estymacji ilości nieudokumentowanych emigrantów na terenie Polski. Metody te mają oczywistą przewagę nad modelami uwzględniającymi jedno źródło, ponieważ korzystają z większej liczby informacji o ile tylko źródła są połączone prawidłowo to znaczy można przypisać obserwację jednostki w pierwszym źródle obserwacji tej samej jednostki w drugim źródle i tak dalej. Nie zawsze może być to możliwe, dodatkowo korzysta się wtedy na przykład z probabilistycznych metod łączenia źródeł. W skrajnych przypadkach gdy połączenia są niepoprawne estymacja jednoźródłowa może okazać się dokładniejsza niż estymacja wykorzystująca wiele źródeł. Metody wykorzystujące dane z jednego źródła wymagają też mniejszej ilości danych więc także potencjalnie niższego nakładu finansowego.

Należy pamiętać, że metody dla jednoźródłowego capture-recapture są stosunkowo nowe i z swojej natury trudne pod względem określenia ich dokładności jako, że szacują wielkość która z definicji jest nieobserwowalna. W przytoczonych publikacjach znajdziemy oczywiście testy symulacyjne szacujące pokrycie N przez przedziały ufności estymatorów $\hat{N}$ w przypadku spełnienia oraz braku spełnienia założeń. Jednakże testy symulacyjne to nie do końca to samo co konsensus naukowy dotyczący dokładności estymatora.

[13] Bootstrap może także potencjalnie posłużyć w ocenie, czy rozkład statystyki $\hat{N}$ może zostać przybliżony rozkładem normalnym. Jednakże najczęściej nie jest to weryfikowane.

Bibliografia

Horvitz, Daniel G and Donovan J Thompson (1952). "A generalization of sampling without replacement from a finite universe". In: *Journal of the American statistical Association* 47.260, pp. 663–685.

Zelterman, Daniel (1988). "Robust estimation in truncated discrete distributions with application to capture-recapture experiments". In: *Journal of statistical planning and inference* 18.2, pp. 225–237.

Chao, Anne (1989). "Estimating population size for sparse data in capture - recapture experiments". In: *Biometrics*, pp. 427–438.

Norris, James L and Kenneth H Pollock (1996). "Including model uncertainty in estimating variances in multiple capture studies". In: *Environmental and Ecological Statistics* 3.3, pp. 235–244.

Heijden, Peter GM van der et al. (2003). "Point and interval estimation of the population size using the truncated Poisson regression model". In: *Statistical Modelling* 3.4, pp. 305–322. DOI: 10.1191/1471082X03st057oa. eprint: https://doi.org/10.1191/1471082X03st057oa. URL: https://doi.org/10.1191/1471082X03st057oa.

Van Der Heijden, Peter GM, Maarten Cruyff, and Hans C Van Houwelingen (2003). "Estimating the size of a criminal population from police records using the truncated Poisson regression model". In: *Statistica Neerlandica* 57.3, pp. 289–304.

Zwane, EN and PGM Van der Heijden (2003). "Implementing the parametric bootstrap in capture–recapture models with continuous covariates". In: *Statistics & probability letters* 65.2, pp. 121–125.

Cruyff, Maarten J. L. F. and Peter G. M. van der Heijden (2008). "Point and Interval Estimation of the Population Size Using a Zero-Truncated Negative Binomial Regression Model". In: *Biometrical Journal* 50.6, pp. 1035–1050. DOI: https://doi.org/10.1002/bimj.200810455. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.200810455. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200810455.

Böhning, Dankmar and Peter G. M. van der Heijden (2009). "A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations". In: *The Annals of Applied Statistics* 3.2, pp. 595–

610. DOI: 10.1214/08-AOAS214. URL: https://doi.org/10.1214/08-AOAS214.

Böhning, Dankmar (2010). "Some general comparative points on Chao's and Zelterman's estimators of the population size". In: *Scandinavian Journal of Statistics* 37.2, pp. 221–236.

Böhning, Dankmar, Alberto Vidal-Diez, et al. (2013). "A Generalization of Chao's Estimator for Covariate Information". In: *Biometrics* 69.4, pp. 1033–1042. DOI: https://doi.org/10.1111/biom.12082. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12082. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12082.

Beręsewicz, Maciej, Grzegorz Gudaszewski, and Marcin Szymkowiak (Nov. 2019). "Estymacja liczby cudzoziemców w Polsce z wykorzystaniem metody capture-recapture". In: *Wiadomości statystyczne (Warsaw, Poland: 1956)* 64, pp. 7–35.

Böhning, Dankmar and Peter G. M. van der Heijden (2019). "The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain". In: *The Annals of Applied Statistics* 13.2, pp. 1198–1211. DOI: 10.1214/18-AOAS1232. URL: https://doi.org/10.1214/18-AOAS1232.