



UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2015:9

On Statistical Methods for Zero-Inflated Models

Julia Eggers

Examensarbete i matematik, 15 hp
Handledare och examinator: Silvelyn Zwanzig
Juni 2015

A large, faint watermark of the Uppsala University seal is visible in the bottom right corner of the page. The seal features a sun with rays, the Latin motto 'VERITAS LIBERABIT VOS', and the text 'UNIVERSITAS UPSALENSIS' around the perimeter.

Department of Mathematics
Uppsala University

Abstract

Data with excess zeros arise in many contexts. Conventional probability distributions often cannot explain large proportions of zero observations. In this paper we shall study statistical models which take large proportions of zero observations into account. We will consider both discrete and continuous distributions.

Contents

1	Introduction	2
2	Models for Zero-Inflated Data	7
3	Models for Semicontinuous Data with Excess Zeros	9
3.1	Tobit Models	9
3.2	Sample Selection Models	10
3.3	Double Hurdle Models	11
3.4	Two-Part Models	12
4	Inference in Models for Zero-Inflated Data	13
4.1	The Likelihood Function	14
4.2	Maximum Likelihood Estimators	14
4.3	Moment Estimators	15
4.4	Cold Spells in Uppsala	16
4.5	Exponential Family	18
5	Inference in Two-Part Models	19
5.1	The Likelihood Function	20
5.2	Maximum Likelihood Estimators	20
5.3	Moment Estimators	21
5.4	Exponential Family	23
5.5	Hypothesis Testing	24
	References	26

1. Introduction

In this paper we will study models for data with a large proportion of zeros. For this we will introduce a few terms.

Definition 1.0.1. *Discrete probability distributions with a large probability mass at zero are said to be zero-inflated.*

Conventional distributions usually cannot explain the large proportion of zeros in zero-inflated data. For this reason different models which can account for a large proportion of zero observations must be applied instead.

Definition 1.0.2. *Probability distributions which are continuous on the entire sample space with the exception of one value at which there is a positive point mass are said to be semicontinuous.*

In this paper we will study models for zero-inflated distributions as well as for semicontinuous distributions with a positive probability mass at zero. We will only consider distributions with non-negative support.

Remark 1.0.1. *Unlike in the case of left-censored data, zeros in semicontinuous data correspond to actual observations and do not represent negative or missing values which have been coded as zero.*

Data with excess zeros may arise in many different contexts. We will start by giving a few examples.

Examples of zero-inflated data

- Cold spells

In his paper on trends for warm and cold spells in Uppsala, Jesper Rydén studied the yearly number of cold spells in Uppsala, Sweden for the period from 1840 to 2012 [07]. He defined a cold spell as a period of at least six consecutive days during which the daily minimum temperature was less

than -13.4°C . The threshold of -13.4°C was chosen as it corresponds to the 5%-quantile of daily minimum temperatures for the reference period 1961 - 1990.

The yearly number of cold spells in Uppsala appears to be zero-inflated as can be seen from the data below. There is a large proportion of zero observations, i.e. years during which there were no cold spells.

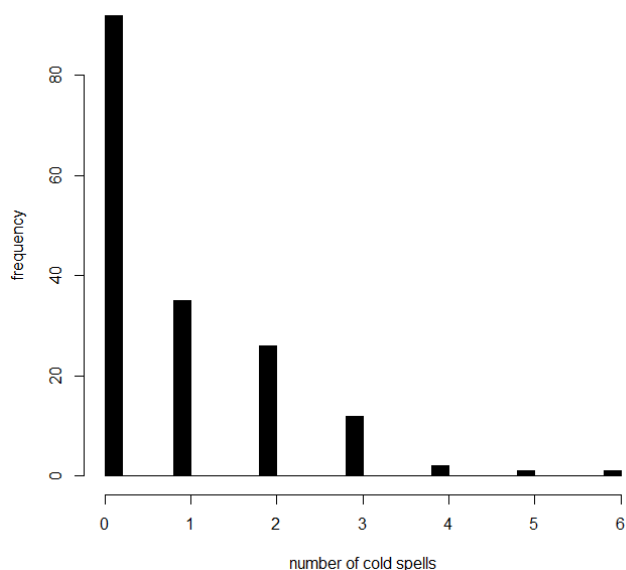


Figure 1.1: Yearly number of cold spells in Uppsala, 1840 - 2012

- Defects in manufacturing

In manufacturing processes defects usually only occur when manufacturing equipment is not properly aligned. If the equipment is misaligned, defects can be found to occur according to a Poisson distribution [08]. This implies that defects in manufacturing occur according to a Poisson distribution with inflation at zero.

Examples of semicontinuous data with excess zeros

- Household expenditures on durable goods

The amount of money a household spends monthly on certain durable goods such as cars or appliances like washing machines or refrigerators is distributed according to a semicontinuous distribution. During most months no such goods are purchased and the expense is zero. If durable goods are purchased, the household expenditure on durable goods for that

month amounts to some positive value, namely the price of the purchased items.

- Alcohol consumption

Consider the alcohol consumption of a population during a certain period of study. Some people belonging to the population may not drink any alcohol at all, thus consuming zero liters of alcohol. These people account for a point mass at zero. People who do consume alcohol may consume arbitrarily large, but positive, amounts. Thus we have a continuous distribution for positive values.

Similarly, the tobacco consumption or consumption of drugs in general is semicontinuously distributed.

- Insurance benefits

The Swedish Social Insurance Agency 'Forsäkringskassan' publishes annual reports on its expenditures. The publication 'Social Insurance in Figures 2014' states that in 2013 a total amount of approximately 24.1 million SEK were paid out as sickness benefits. These sickness benefits are meant to compensate insured for the inability to work due to illness. In 2013, 532 450 people in Sweden received sickness benefits. This corresponds to around 9% of all insured between the ages of 16 and 64.

The amounts of sickness benefits paid out to insured during the year 2013 are semicontinuously distributed. 91% of all insured received no such benefits. We thus have a positive probability mass at zero. Those people who did receive sickness benefits got positive amounts which varied according to factors like income and time spent on sick leave. Therefore, we have a continuous distribution for positive values.

Table 1.1 below gives an account of the average amounts of sickness benefits paid out to insured depending on gender and age group.

Age	Number of recipients		Average number of days		Average amount (SEK per day)	
	Women	Men	Women	Men	Women	Men
16-24	14,070	10,531	59	57	416	499
25-29	28,019	13,269	62	67	496	545
30-34	36,808	15,554	70	77	514	536
35-39	39,283	17,814	81	80	525	560
40-44	40,994	21,007	91	83	527	569
45-49	46,110	25,675	92	87	523	572
50-54	43,431	26,747	92	90	519	563
55-59	43,491	29,325	89	92	519	563
60-	44,721	35,601	90	96	513	561
Total	336,927	195,523	84	85	515	559

Table 1.1: Sickness benefits, 2013

We know from the data that 532 450 insured received sickness benefits in 2013 while around 5 383 661 insured received no such benefits. Since 24.1 million SEK were paid out in total, the average positive amount that was paid out per person that year amounted to approximately 44 822 SEK. Assuming that the paid out benefits are exponentially distributed with parameter λ given that they are positive, we may estimate $\hat{\lambda} = 1/44822$. Generating a sample from this distribution in R, we may illustrate how the sickness benefits may have been distributed.

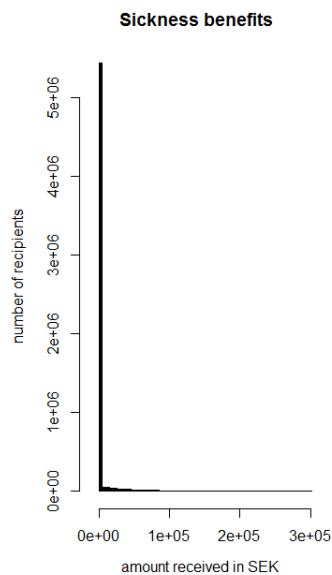


Figure 1.2: A possible distribution of the amount of sickness benefits paid out to insured during the year 2013

- Healthcare expenditures

Healthcare expenditures in general can be found to be semicontinuously distributed. Individuals may or may not choose to seek medical treatment during a certain period of study. There are no costs arising for people who do not seek medical treatment. If medical treatment is sought, however, the cost for the treatment amounts to some positive value. Thus health-care expenditures are continuously distributed for positive values and have a positive probability mass at zero.

2. Models for Zero-Inflated Data

The models for zero-inflated data which we will present here are variations of the following mixture model

$$Y = \Delta Z_1 + (1 - \Delta)Z_2$$

with $\Delta \sim Ber(p)$, $Z_1 \sim P^{Z_1}$ and $Z_2 \sim P^{Z_2}$.

If we let $P^{Z_2} = \delta_{\{0\}}$ and assume Z_1 to be discrete, we obtain a model for zero-inflated data. When modeling count data we have the additional assumption that $P(Z_1 \geq 0) = 1$.

For the above model we have that $Y \sim \delta_{\{0\}}$ with probability $1 - p$ and $Y \sim P^{Z_1}$ with probability p . Letting p_{Z_1} denote the probability mass function of the random variable Z_1 we obtain

$$P(Y = y) = \begin{cases} 1 - p + pp_{Z_1}(0) & , y = 0 \\ pp_{Z_1}(y) & , y > 0 \end{cases}$$

When modeling count data, the negative binomial and the Poisson distribution are common distributions for Z_1 .

If $Z_1 \sim Po(\lambda)$ the above model is referred to as the zero-inflated Poisson model, abbreviated ZIP.

Zero-Inflated Poisson Regression

Zero-inflated Poisson regression is an extension of the zero-inflated Poisson model which was proposed by Diane Lambert in 1992 [08].

The model assumes $Y = (Y_1, \dots, Y_n)$ to be a sample of independent, but not necessarily identically distributed random variables Y_i . In this model we assume $Y_i \sim Po(\lambda_i)$ with probability p_i .

$$\text{Thus } P(Y_i = y_i) = \begin{cases} 1 - p_i + p_i \exp(-\lambda_i) & , y_i = 0 \\ p_i \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!} & , y_i > 0 \end{cases}$$

The parameters $p = (p_1, \dots, p_n)$ and $\lambda = (\lambda_1, \dots, \lambda_n)$ are assumed to satisfy

$\log(\lambda) = B\beta$ and $\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = G\gamma$ with B and G denoting matrices with explanatory variables. β and γ are matrices with coefficients to adequately describe the linear dependency of $\log(\lambda)$ and $\text{logit}(p)$ on B and G respectively.

If p and λ depend on the same explanatory variables, the number of model parameters may be reduced by expressing p as a function of λ . Lambert proposes the relation $\text{logit}(p) = -\tau \log(\lambda)$ for some $\tau \in \mathbb{R}$. This implies that $p_i = \frac{1}{1+\lambda_i^\tau}$. The resulting model is denoted by $ZIP(\tau)$.

3. Models for Semicontinuous Data with Excess Zeros

There are a number of different models which can be applied to semicontinuous data with excess zeros. We will present a few of the most common ones.

In all of these models we will let Y denote the observed variable.

The models we will present are all special kinds of two-component mixture models. A mixture model with two components has the form

$$Y = \Delta Z_1 + (1 - \Delta)Z_2$$

with $\Delta \sim Ber(p)$, $Z_1 \sim P^{Z_1}$ and $Z_2 \sim P^{Z_2}$.

In the models we will present, we have that $P^{Z_2} = \delta_{\{0\}}$.

3.1 Tobit Models

The Tobit model which was proposed by James Tobin in 1958 [03] assumes that Y can be expressed in terms of a latent variable Y^* which can only be observed for values greater than zero.

The random variable Y is defined as follows.

$$Y = \begin{cases} Y^* & , Y^* > 0 \\ 0 & , Y^* \leq 0 \end{cases}$$

The latent variable Y^* is assumed to be linearly dependent on a number of explanatory (and observable) variables and can be expressed as a linear combination of these, i.e.

$$Y^* = X\beta + \epsilon$$

where X is a row vector containing the explanatory variables and β is a column vector with the corresponding coefficients describing the linear dependency of Y^* on X .

The error terms ϵ are assumed to be independently and identically distributed

according to $N(0, \sigma^2)$. Thus the Tobit model assumes an underlying normal distribution.

The probability that Y takes the value zero is given by

$$\begin{aligned} P(Y = 0) &= P(Y^* \leq 0) = P(X\beta + \epsilon \leq 0) = P(\epsilon \leq -X\beta) = \\ &= P\left(\frac{\epsilon}{\sigma} \leq -\frac{X\beta}{\sigma}\right) = \Phi\left(-\frac{X\beta}{\sigma}\right) = 1 - \Phi\left(\frac{X\beta}{\sigma}\right). \end{aligned}$$

This part of the Tobit model corresponds to the so-called Probit model. The name Tobit alludes to the Tobit model having been proposed by Tobin and being based on the Probit model.

The likelihood function L of the uncensored positive values of Y is given by the probability density function of the latent variable Y^* given that it is positive, i.e.

$$L(y|y > 0) = L(y^*|y^* > 0) = \frac{1}{\sigma} \phi\left(\frac{y - X\beta}{\sigma}\right).$$

In Tobit models the probability of a zero observation depends on the same random variable that determines the magnitude of the observation given that it is positive.

Note that in the Tobit model zeros do not represent actual responses. The Tobit model is therefore not appropriate for semicontinuous data. It is, however, often applied to such data in spite of this.

Remark 3.1.1. *There are many variations of the Tobit model. Censoring can for instance be performed at values other than zero. There are also models where censoring is done from above instead of below or from both above and below.*

Remark 3.1.2. *The mixture model above corresponds to the Tobit model if $Z_1 = Y^*$ and $p = P(Y^* > 0) = \Phi\left(\frac{X\beta}{\sigma}\right)$.*

3.2 Sample Selection Models

The sample selection model was first proposed by J. Heckman in 1979 as an extension of the Tobit model.

Sample selection models are based on two latent variables Y_1^* and Y_2^* .

The first latent variable Y_1^* is assumed to be of the form $Y_1^* = X_1\beta_1 + \epsilon_1$ with X_1 being a row vector of explanatory variables and β_1 being the corresponding vector of coefficients describing the linear dependency of Y_1^* on X_1 .

Similarly, the second latent variable Y_2^* is assumed to be of the form $Y_2^* = X_2\beta_2 + \epsilon_2$, again with X_2 being a row vector of explanatory variables and β_2 being the corresponding vector of coefficients.

Sample selection models thus allow for the latent variables to depend on different covariates.

The error terms (ϵ_1, ϵ_2) are assumed to be independently and identically distributed according to a bivariate normal distribution. They may thus be correlated.

The observed variable is defined as $Y = \begin{cases} Y_2^* & , Y_1^* > 0 \\ 0 & , Y_1^* \leq 0 \end{cases}$.

The sample selection model coincides with the Tobit model if $X_1 = X_2$ and $\beta_1 = \beta_2$ (i.e. $Y_1^* = Y_2^*$).

Remark 3.2.1. *The mixture model above corresponds to the sample selection model if $Z_1 = Y_2^*$ and $p = P(Y_1^* > 0)$.*

3.3 Double Hurdle Models

Similarly to sample selection models, double hurdle models are based on two latent variables Y_1^* and Y_2^* .

These latent variables are again assumed to be of the form $Y_1^* = X_1\beta_1 + \epsilon_1$ and $Y_2^* = X_2\beta_2 + \epsilon_2$ with X_1 and X_2 denoting row vectors with observed values of explanatory variables and β_1 and β_2 denoting column vectors that contain the corresponding coefficients describing the linear dependency of Y_1^* and Y_2^* on X_1 and X_2 respectively.

(ϵ_1, ϵ_2) are again assumed to be independently and identically distributed according to a bivariate normal distribution.

In double hurdle models, the observed variable is defined as

$$Y = \begin{cases} Y_2^* & , Y_1^* > 0 \text{ and } Y_2^* > 0 \\ 0 & , \text{otherwise} \end{cases}.$$

To illustrate the idea behind double hurdle models, we will apply it to the example of tobacco consumption. We thus let Y denote the amount of tobacco consumed by an individual during a certain period of time.

The first latent variable may Y_1^* determine whether an individual is a smoker or non-smoker. This may depend on certain socioeconomic factors which can be accounted for by the dependency of Y_1^* on X_1 .

The second latent variable Y_2^* may thereafter be used to determine how much tobacco is consumed by an individual given that the individual is a smoker. This quantity may depend on other covariates than the ones that affected the probability of the individual being a smoker in the first place.

Note that it is possible for a smoker not to consume any tobacco during the period of the study, in other words we may have $Y_2^* \leq 0 | Y_1^* > 0$.

We see that in order to observe positive values of Y two hurdles need to be overcome. The individual must be a smoker and smoke during the period of the study. Hence the name double hurdle model.

Remark 3.3.1. *The mixture model above corresponds to the double hurdle model if $Z_1 = Y_2^*$ and $p = P(Y_1^* > 0, Y_2^* > 0)$.*

3.4 Two-Part Models

As the name suggests, two-part models consist of two parts. In the first part of the model a random variable Δ determines whether the observation is zero or positive. In the second part another random variable Z determines the magnitude of the observation given that it is positive. The value of the random variable Z is not observed if Δ has taken the value zero. The random variables Δ and Z are assumed to be independent. Moreover, we assume that $P(Z > 0) = 1$.

In other words we have the following model for the random variable Y

$$Y = \mathbf{1}_{\{1\}}(\Delta)Z = \Delta Z, \quad (3.1)$$

with $\Delta \sim \text{Ber}(\theta_1)$ and $Z \sim P^Z \in \{P_{\theta_2}\}$ being independent and $P(Z > 0) = 1$. Thus $Y \sim P^Y \in \{P_{\theta}, \theta = (\theta_1, \theta_2)\}$.

Two-part models do not assume an underlying normal distribution and can therefore be applied to a wider range of data than for instance Tobit models.

Note that in two-part models we do not have a latent variable. Zeros correspond to actual observations, and are not the result of censoring as in the previously presented models. Consequently, two-part models are more appropriate for modeling semicontinuous data than the other models we have presented. In the following, we will therefore restrict ourselves to the study of two-part models.

Remark 3.4.1. *The mixture model above corresponds to the two-part model if $Z_1 = Z$.*

Remark 3.4.2. *Note that in all the models for semicontinuous data with excess zeros we have that $P(Z_1 = 0) = 0$ and $P(Z_2 = 0) = 1$. We can therefore distinguish between observations from Z_1 and Z_2 . For zero-inflated count data, however, we have that $P(Z_2 = 0) = 1$ and $P(Z_1 = 0) > 0$. Here we are unable to distinguish between zero observations from Z_1 and zero observations from Z_2 .*

4. Inference in Models for Zero-Inflated Data

Let Y denote the observed variable. We will assume the following model $P^Y \in \{P_\theta, \theta = (p, \lambda)\}$ for Y .

$$Y = \Delta Z_1 + (1 - \Delta)Z_2$$

with $\Delta \sim Ber(p)$, $Z_1 \sim P^{Z_1} \in \{P_\lambda\}$ and $Z_2 \sim \delta_{\{0\}}$ being independent. Moreover, we assume that Z_1 is discrete and that $P(Z_1 \geq 0) = 1$.

Note that the observed variable Y has non-negative support. Thus the above model can be applied to, for instance, count data.

$$\text{For this model we have } P(Y = y) = \begin{cases} 1 - p + pp_{Z_1}(0) & , y = 0 \\ pp_{Z_1}(y) & , y > 0 \end{cases} .$$

with p_{Z_1} denoting the probability mass function of Z_1 .

In the case that $Z_1 \sim Po(\lambda)$ we obtain a zero-inflated Poisson model with

$$P(Y = y) = \begin{cases} 1 - p + p \exp(-\lambda) & , y = 0 \\ p \frac{\lambda^y \exp(-\lambda)}{y!} & , y > 0 \end{cases} .$$

Theorem 4.0.1. *The expected value $E[Y]$ and variance $Var[Y]$ of Y are given by $E[Y] = pE[Z_1]$ and $Var[Y] = pVar[Z_1] + (1 - p)pE[Z_1]^2$.*

Proof. The expected value of Y is given by

$$E[Y] = E[\Delta Z_1 + (1 - \Delta)Z_2] = E[\Delta]E[Z_1] + E[Z_2] - E[\Delta]E[Z_2] = pE[Z_1].$$

The variance of Y is given by

$$\begin{aligned} Var[Y] &= Var[\Delta Z_1] + Var[Z_2] + Var[\Delta Z_2] = Var[\Delta Z_1] + Var[\Delta Z_2] = \\ &= E[\Delta^2]E[Z_1^2] - E[\Delta]^2E[Z_1]^2 + E[\Delta^2]E[Z_2^2] - E[\Delta]^2E[Z_2]^2 = \\ &= E[\Delta^2]E[Z_1^2] - E[\Delta]^2E[Z_1]^2 = \\ &= (Var[\Delta] + E[\Delta]^2)(Var[Z_1] + E[Z_1]^2) - E[\Delta]^2E[Z_1]^2 = \\ &= (p(1 - p) + p^2)(Var[Z_1] + E[Z_1]^2) - p^2E[Z_1]^2 = \\ &= pVar[Z_1] + p(1 - p)E[Z_1]^2 \end{aligned}$$

□

Corollary 4.0.1. *In the zero-inflated Poisson model the expected value $E[Y]$ and variance $Var[Y]$ of Y are given by $E[Y] = p\lambda$ and $Var[Y] = p\lambda(1 + \lambda - p\lambda)$.*

Proof. In the zero-inflated Poisson model $Z_1 \sim Po(\lambda)$ and $E[Z_1] = Var[Z_1] = \lambda$. Plugging these values in to the expressions for $E[Y]$ and $Var[Y]$ yields the above result. \square

4.1 The Likelihood Function

Definition 4.1.1. *The likelihood function $L(\theta, y) : \Theta \rightarrow \mathbb{R}_+$ for an observation y of a random variable Y with probability function $p(\theta, y)$ is given by*

$$L(\theta, y) := p(\theta, y). \quad (4.1)$$

For a sample $Y = (Y_1, \dots, Y_n)$ of independent and identically distributed random variables the likelihood function is given by

$$L(\theta, y) := \prod_{i=1}^n p(\theta, y_i). \quad (4.2)$$

We will now consider a sample $y = (y_1, \dots, y_n)$ of independent and identically distributed random variables $Y_i \sim P^Y$. Let r denote the number of zero observations in the sample y .

The likelihood function $L(p, \lambda, y)$ of the sample y is given by

$$\begin{aligned} L(p, \lambda, y) &= \prod_{i=1}^n P(Y_i = y_i) = \prod_{y_i=0} (1 - p + pp_{Z_1}(0)) \prod_{y_i>0} pp_{Z_1}(y_i) = \\ &= (1 - p + pp_{Z_1}(0))^r p^{n-r} \prod_{y_i>0} p_{Z_1}(y_i). \end{aligned}$$

If $Z_1 \sim Po(\lambda)$ we have

$$L(p, \lambda, y) = (1 - p + p \exp(-\lambda))^r p^{n-r} \frac{\lambda^{\sum_{i=1}^n y_i} \exp(-\lambda(n-r))}{\prod_{i=1}^n y_i!}.$$

4.2 Maximum Likelihood Estimators

Definition 4.2.1. *The maximum likelihood estimator $\hat{\theta}_{MLE}$ of a variable θ is a value of θ which maximizes the likelihood function, i.e.*

$$\hat{\theta}_{MLE} \in \max_{\theta \in \Theta} L(\theta, y) \quad \forall y \in \chi \quad (4.3)$$

with χ denoting the sample space of Y .

Theorem 4.2.1. *Let $p \in (0, 1)$ and $Z_1 \sim Po(\lambda)$, i.e. assume a zero-inflated Poisson model for the sample y . The maximum likelihood estimators $\hat{p}_{MLE}(Y)$ and $\hat{\lambda}_{MLE}(Y)$ are given by*

$$\hat{p}_{MLE}(Y) = \frac{n-r}{n(1 - e^{-\hat{\lambda}_{MLE}})}.$$

and

$$(1 - e^{-\hat{\lambda}_{MLE}}) \sum_{i=1}^n y_i = \hat{\lambda}_{MLE}(n - r).$$

Proof. The likelihood function $L(p, \lambda, y)$ of the sample y is given by

$$L(p, \lambda, y) = (1 - p + p \exp(-\lambda))^r p^{n-r} \frac{\lambda^{\sum_{i=1}^n y_i} \exp(-\lambda(n-r))}{\prod_{i=1}^n y_i!}$$

The values of p for which $L(p, \lambda, y)$ is maximized satisfy

$$\begin{aligned} \frac{\partial}{\partial p} L(p, \lambda, y) &= 0 \\ \Leftrightarrow \frac{\partial}{\partial p} \ln(L(p, \lambda, y)) &= 0 \\ \Leftrightarrow \frac{\partial}{\partial p} (r \ln(1-p+p \exp(-\lambda)) + (n-r) \ln(p) + \sum_{i=1}^n y_i \ln(\lambda) - \lambda(n-r) - \ln(\prod_{i=1}^n y_i!)) &= 0 \\ \Leftrightarrow \frac{r(-1+e^{-\lambda})}{1-p+pe^{-\lambda}} + \frac{n-r}{p} &= 0 \\ \Leftrightarrow p &= \frac{n-r}{n(1-e^{-\lambda})} \end{aligned}$$

The values of λ which maximize $L(p, \lambda, y)$ satisfy

$$\begin{aligned} \frac{\partial}{\partial \lambda} L(p, \lambda, y) &= 0 \\ \Leftrightarrow \frac{\partial}{\partial \lambda} (r \ln(1-p+p \exp(-\lambda)) + (n-r) \ln(p) + \sum_{i=1}^n y_i \ln(\lambda) - \lambda(n-r) - \ln(\prod_{i=1}^n y_i!)) &= 0 \end{aligned}$$

Inserting $p = \frac{n-r}{n(1-e^{-\lambda})}$ gives

$$(1 - e^{-\lambda}) \sum_{i=1}^n y_i = \lambda(n - r)$$

□

Remark 4.2.1. *Numerical methods must be applied to solve the equation*

$$(1 - e^{-\lambda}) \sum_{i=1}^n y_i = \lambda(n - r)$$

above.

4.3 Moment Estimators

Definition 4.3.1. *Let $Y = (Y_1, Y_2, \dots, Y_n)$ be a sample from independent and identically distributed random variables with distributions depending on a parameter θ . The moment estimator of order k for θ is given by the value of θ for which*

$$E[Y^k] = g(\theta) = \frac{1}{n} \sum_{i=1}^n Y_i^k$$

where g is some function specifying the expected value.

Theorem 4.3.1. Let $Z_1 \sim Po(\lambda)$, i.e. assume a zero-inflated Poisson model for the sample y . The moment estimators $\hat{p}_{MME}(Y)$ and $\hat{\lambda}_{MME}(Y)$ are given by

$$\hat{p}_{MME}(Y) = \frac{\left(\frac{1}{n} \sum_{i=1}^n Y_i\right)^2}{\frac{1}{n} \sum_{i=1}^n Y_i^2 - \frac{1}{n} \sum_{i=1}^n Y_i}$$

and

$$\hat{\lambda}_{MME}(Y) = \frac{\sum_{i=1}^n Y_i^2}{\sum_{i=1}^n Y_i} - 1.$$

Proof. The moment estimators $\hat{p}_{MME}(Y)$ and $\hat{\lambda}_{MME}(Y)$ are given by values of p and λ which satisfy

$$\begin{cases} E[Y] = \frac{1}{n} \sum_{i=1}^n Y_i = p\lambda \\ E[Y^2] = \frac{1}{n} \sum_{i=1}^n Y_i^2 = Var[Y] + E[Y]^2 = p\lambda(1 + \lambda - p\lambda) + p^2\lambda^2 = p\lambda(1 + \lambda) \end{cases}$$

$$\Rightarrow (1 + \lambda) = \frac{\frac{1}{n} \sum_{i=1}^n Y_i^2}{\frac{1}{n} \sum_{i=1}^n Y_i} = \frac{\sum_{i=1}^n Y_i^2}{\sum_{i=1}^n Y_i}$$

$$\Rightarrow \lambda = \frac{\sum_{i=1}^n Y_i^2}{\sum_{i=1}^n Y_i} - 1$$

$$\Rightarrow p = \frac{\frac{1}{n} \sum_{i=1}^n Y_i}{\lambda} = \frac{\frac{1}{n} \sum_{i=1}^n Y_i}{\frac{\sum_{i=1}^n Y_i^2 - \sum_{i=1}^n Y_i}{\sum_{i=1}^n Y_i}} = \frac{\left(\frac{1}{n} \sum_{i=1}^n Y_i\right)^2}{\frac{1}{n} \sum_{i=1}^n Y_i^2 - \frac{1}{n} \sum_{i=1}^n Y_i}$$

□

4.4 Cold Spells in Uppsala

We will now assume a zero-inflated Poisson model for the data regarding cold spells in Uppsala (see chapter 1, figure 1.1) [07]. We let $y = (y_1, \dots, y_{169})$ denote the corresponding sample and assume that the observations are independent and identically distributed. Note that, in reality, the number of cold spells that occur during two consecutive years may not actually be independent so this assumption may not hold.

For this sample we have $\sum_{i=1}^{169} y_i = 142$ and $\sum_{i=1}^{169} y_i^2 = 340$.

We thus obtain the following moment estimates for the model parameters p and λ .

$$\hat{p}_{MME}(y) = \frac{\left(\frac{1}{169} \sum_{i=1}^{169} y_i\right)^2}{\frac{1}{169} \sum_{i=1}^n y_i^2 - \frac{1}{169} \sum_{i=1}^{169} y_i} = \frac{142^2}{169(340 - 142)} \approx 0.6026$$

$$\hat{\lambda}_{MME}(y) = \frac{\sum_{i=1}^{169} y_i^2}{\sum_{i=1}^{169} y_i} - 1 = \frac{340}{142} - 1 = \frac{99}{71} \approx 1.3944$$

Note that $\hat{p}_{MME}(y) \approx 0.6 < 1$ so the yearly number of cold spells in Uppsala does indeed appear to be zero-inflated.

Theorem 4.4.1. *An approximate level α test for the testing problem*

$$H_0 : p = 1, \lambda = 1.39$$

$$H_1 : 0 < p < 1, \lambda = 1.39$$

is given by

$$\phi(y) = \begin{cases} 1 & , -2 \ln(\Lambda(y)) \geq \chi_\alpha^2(1) \\ 0 & , -2 \ln(\Lambda(y)) < \chi_\alpha^2(1) \end{cases}$$

where

$$\Lambda(y) = \left(1 - \frac{n-r}{n(1-e^{-1.39})} + \frac{n-r}{n(1-e^{-1.39})} e^{-1.39}\right)^{-r} \left(\frac{n-r}{n(1-e^{-1.39})}\right)^{-(n-r)} e^{-1.39r}.$$

Proof. The likelihood ratio $\Lambda(Y)$ is given by

$$\Lambda(Y) = \frac{\max\{p_0(y) : p = 1, \lambda = 1.39\}}{\max\{p(y) : 0 < p \leq 1, \lambda = 1.39\}} = \frac{1}{\max_{0 < p \leq 1} (1-p + pe^{-1.39})^r p^{n-r} e^{-1.39r}}$$

$$= \left(1 - \frac{n-r}{n(1-e^{-1.39})} + \frac{n-r}{n(1-e^{-1.39})} e^{-1.39}\right)^{-r} \left(\frac{n-r}{n(1-e^{-1.39})}\right)^{-(n-r)} e^{-1.39r}$$

According to Wilk's Theorem $-2 \ln(\Lambda(Y)) \sim \chi^2(1)$ approximately as $n \rightarrow \infty$. We can thus reject H_0 at significance level α if $-2 \ln(\Lambda(Y)) \geq \chi_\alpha^2(1)$. This yields the above test. □

For the sample y we obtain $-2 \ln(\Lambda(y)) = 66.92 \geq \chi_{0.05}^2(1) = 3.84$. It therefore follows that H_0 can be rejected at significance level 0.05.

Moreover, the p-value for the test is given by $p = P_0(-2 \ln(\Lambda(Y)) \geq 66.92) = 1 - P_0(-2 \ln(\Lambda(Y)) < 66.92) = 3.33 * 10^{-15}$, so H_0 can be rejected at any significance level $\alpha > 3.33 * 10^{-15}$.

We conclude that the yearly number of cold spells in Uppsala are zero-inflated.

4.5 Exponential Family

Definition 4.5.1. A class of probability measures $P = \{P_\theta : \theta \in \Theta\}$ is called an exponential family if

$$L(y; \theta) = A(\theta) * \exp\left(\sum_{j=1}^k \zeta_j(\theta) T_j(y)\right) * h(y) \quad (4.4)$$

for some $k \in \mathbb{N}$, real-valued functions ζ_1, \dots, ζ_k on Θ , real-valued statistics T_1, \dots, T_k and a function h on the sample space χ .

Theorem 4.5.1. If the class of probability measures $P = \{P_\theta : \theta \in \Theta\}$ forms an exponential family, then all P_θ are pairwise equivalent, i.e. for any $P, Q \in P$ we have $P(N)=0$ iff $Q(N)=0$ [06].

Theorem 4.5.2. If $p \in [0, 1]$, then P^Y does not form an exponential family.

Proof. Consider the probability measures $P_{(0, \lambda_1)}$ and $P_{(p_2, \lambda_2)}$ with $p_2 \in (0, 1]$. We have that $P_{(0, \lambda_1)}(\mathbb{R} \setminus \{0\}) = 0$ but $P_{(p_2, \lambda_2)}(\mathbb{R} \setminus \{0\}) > 0$. Therefore it follows that P^Y does not form an exponential family. \square

Theorem 4.5.3. If P^{Z_1} does not form an exponential family, then P^Y does not form an exponential family.

Proof. The likelihood $L(p, \lambda, y)$ of y is given by $L(p, \lambda, y) = ((1 - p) + pp_{Z_1}(0))^r p^{n-r} \prod_{y_i > 0} p_{Z_1}(y_i)$.

Since P^{Z_1} does not form an exponential family, $p_{Z_1}(y_i)$ is not of the form (4.4). Thus $L(p, \lambda, y)$ is not of the form (4.4). Consequently, P^Y is not an exponential family. \square

Theorem 4.5.4. If $p \in (0, 1]$ and P^{Z_1} is a k -parameter exponential family with natural parameters $\zeta_j(\lambda)$ and sufficient statistics $T_j(z)$, $j = 1, \dots, k$, then P^Y is a $k+1$ parameter exponential family with natural parameters $\zeta_j(\lambda)$, $j = 1, \dots, k$ and $\zeta_{k+1}(p) = \ln\left(\frac{1-p+pp_{Z_1}(0)}{p}\right)$ and sufficient statistics $T_j(y)$, $j = 1, \dots, k$ and $T_{k+1}(y) = r$.

Proof. We have that $L(p, \lambda, y) = ((1 - p) + pp_{Z_1}(0))^r p^{n-r} \prod_{y_i > 0} p_{Z_1}(y_i) = \exp(r \ln(1 - p + pp_{Z_1}(0)) + (n - r) \ln(p)) \prod_{y_i > 0} p_{Z_1}(y_i) = p^n \exp\left(r \ln\left(\frac{1-p+pp_{Z_1}(0)}{p}\right)\right) \prod_{y_i > 0} p_{Z_1}(y_i) = p^n A(\lambda) \exp\left(r \ln\left(\frac{1-p+pp_{Z_1}(0)}{p}\right) + \sum_{j=1}^k \zeta_j(\lambda) T_j(y)\right) h(y)$

Thus $P^Y \in \{P_{(p, \lambda)}\}$ forms an exponential family with natural parameters $\zeta_j(\lambda)$, $j = 1, \dots, k$ and $\zeta_{k+1}(p) = \ln\left(\frac{1-p+pp_{Z_1}(0)}{p}\right)$ and sufficient statistics $T_j(y)$, $j = 1, \dots, k$ and $T_{k+1}(y) = r$. \square

5. Inference in Two-Part Models

Consider a random variable $Y \sim P^Y \in \{P_\theta, \theta = (\theta_1, \theta_2)\}$ distributed according to the two-part model, i.e. let

$$Y = \mathbf{1}_{\{1\}}(\Delta)Z = \Delta Z,$$

with Δ and Z being independent, $\Delta \sim \text{Ber}(\theta_1)$, $Z \sim P^Z \in \{P_{\theta_2}\}$ and $P(Z > 0) = 1$.

Theorem 5.0.5. *The expected value $E[Y]$ and variance $\text{Var}[Y]$ of Y are given by $E[Y] = \theta_1 E[Z]$ and $\text{Var}[Y] = \theta_1 \text{Var}[Z] + (1 - \theta_1)\theta_1 E[Z]^2$.*

Proof. The expected value of Y is given by $E[Y] = E[\Delta Z] = E[\Delta]E[Z] = \theta_1 E[Z]$.

The variance of Y is given by

$$\begin{aligned} \text{Var}[Y] &= \text{Var}[\Delta Z] = E[\Delta^2]E[Z^2] - E[\Delta]^2 E[Z]^2 \\ &\text{since } \Delta \text{ and } Z \text{ are independent. Thus} \\ \text{Var}[Y] &= E[\Delta^2]E[Z^2] - E[\Delta]^2 E[Z]^2 = \\ &= (\text{Var}[\Delta] + E[\Delta]^2)(\text{Var}[Z] + E[Z]^2) - E[\Delta]^2 E[Z]^2 = \\ &= \text{Var}[\Delta]\text{Var}[Z] + \text{Var}[\Delta]E[Z]^2 + E[\Delta]^2 \text{Var}[Z] + E[\Delta]^2 E[Z]^2 - E[\Delta]^2 E[Z]^2 = \\ &= \text{Var}[\Delta]\text{Var}[Z] + \text{Var}[\Delta]E[Z]^2 + E[\Delta]^2 \text{Var}[Z] = \\ &= (\text{Var}[\Delta] + E[\Delta]^2)\text{Var}[Z] + \text{Var}[\Delta]E[Z]^2 = \\ &= ((1 - \theta_1)\theta_1 + \theta_1^2)\text{Var}[Z] + (1 - \theta_1)\theta_1 E[Z]^2 = \\ &= \theta_1 \text{Var}[Z] + (1 - \theta_1)\theta_1 E[Z]^2 \end{aligned}$$

□

It can be shown that if $E[\hat{Z}]$, $E[\hat{Z}]^2$ and $\text{Var}\hat{Z}$ are unbiased estimators of $E[Z]$, $E[Z]^2$ and $\text{Var}[Z]$, then

$$E\hat{Y} = \begin{cases} \frac{n-r}{n} E\hat{Z} & , n-r > 0 \\ 0 & , n-r = 0 \end{cases}$$

and

$$\text{Var}\hat{Y} = \begin{cases} \frac{n-r}{n} \text{Var}\hat{Z} + \frac{n-r}{n} \frac{r}{n-1} E\hat{Z}^2 & , n-r > 0 \\ 0 & , n-r = 0 \end{cases}$$

are unbiased estimators of $E[Y]$ and $\text{Var}[Y]$ respectively, see [10].

5.1 The Likelihood Function

Theorem 5.1.1. *Let $y = (y_1, \dots, y_n)$ be a sample from independent and identically distributed random variables Y_i distributed according to the two-part model. Moreover, let r denote the number of zero observations in the sample y and let $z = (z_1, \dots, z_{n-r})$ be the subsample of positive observations of y .*

The likelihood function $L(\theta_1, \theta_2, y)$ of the sample y is given by

$$L(\theta_1, \theta_2, y) = L_1(\theta_1, y)L_2(\theta_2, y)$$

where $L_1(\theta_1, y) = (1 - \theta_1)^r \theta_1^{n-r}$ and $L_2(\theta_2, y) = L(\theta_2, z)$.

Proof. The likelihood function of the sample y is given by

$$\begin{aligned} L(\theta_1, \theta_2, y) &= \prod_{y_i=0} P(Y_i = 0) \prod_{y_i>0} P(Y_i > 0) f(y_i | y_i > 0) = \\ &= \prod_{y_i=0} (1 - \theta_1) \prod_{y_i>0} \theta_1 f(y_i | y_i > 0) = (1 - \theta_1)^r \theta_1^{n-r} \prod_{y_i>0} f(y_i | y_i > 0) \end{aligned}$$

with $f(y_i | y_i > 0)$ denoting the probability density function of the observations given that they are positive, i.e. the probability density function of Z . Thus $f(y_i | y_i > 0) = f(z_i)$.

It follows that $L(\theta_1, \theta_2, y) = L_1(\theta_1, y)L_2(\theta_2, y)$ with $L_1(\theta_1, y) = (1 - \theta_1)^r \theta_1^{n-r}$ and $L_2(\theta_2, y) = \prod_{y_i>0} f(y_i | y_i > 0) = \prod_{i=1}^{n-r} f(z_i) = L(\theta_2, z)$. □

Remark 5.1.1. *Here $L(\theta_2, z)$ denotes the likelihood of the subsample z .*

5.2 Maximum Likelihood Estimators

Theorem 5.2.1. *Let $\theta_1 \in (0, 1)$. The likelihood estimates of θ_1 and θ_2 are given by*

$$\hat{\theta}_{1,MLE} = \frac{n-r}{n}$$

and

$$\hat{\theta}_{2,MLE} \in \max_{\theta_2 \in \Theta_2} L(\theta_2, z).$$

Proof. Since the likelihood function L of the sample y can be expressed as a product of two functions L_1 and L_2 that each depend on only θ_1 or θ_2 , L may be maximized by maximizing L_1 and L_2 respectively. In other words, we have

$$\max_{\theta \in \Theta} L(\theta_1, \theta_2, y) = \max_{\theta_1 \in \Theta_1} L_1(\theta_1, y) \max_{\theta_2 \in \Theta_2} L_2(\theta_2, y).$$

Maximum likelihood estimation of θ_1

We have that $\hat{\theta}_{1,MLE} \in \max_{\theta_1 \in \Theta_1} L_1(\theta_1, y)$. In other words the maximum likelihood estimate $\hat{\theta}_{1,MLE}$ of θ_1 is a value of θ_1 which maximizes $L_1(\theta_1, y) = (1-\theta_1)^r \theta_1^{n-r}$. The values of θ_1 for which $L_1(\theta_1, y)$ is maximized satisfy the following equation.

$$\begin{aligned} \frac{\partial}{\partial \theta_1} L(\theta_1, y) &= 0 \\ \Leftrightarrow \frac{\partial}{\partial \theta_1} \ln(L(\theta_1, y)) &= 0 \\ \Leftrightarrow \frac{\partial}{\partial \theta_1} r \ln(1 - \theta_1) + (n - r) \ln(\theta_1) &= 0 \\ \Leftrightarrow \theta_1 &= \frac{n-r}{n} \end{aligned}$$

Maximum likelihood estimation of θ_2

The maximum likelihood estimate $\hat{\theta}_{2,MLE}$ of θ_2 is a value of θ_2 which maximizes $L_2(\theta_2, y) = L(\theta_2, z)$, i.e. $\hat{\theta}_{2,MLE} \in \max_{\theta_2 \in \Theta_2} L(\theta_2, z)$. To obtain a maximum likelihood estimate of θ_2 we therefore only need to consider the subsample z of positive observations of y . □

5.3 Moment Estimators

Theorem 5.3.1. *The moment estimators for θ_1 and θ_2 satisfy*

$$\begin{cases} \theta_1 E[Z] = \frac{1}{n} \sum_{i=1}^n Y_i \\ \theta_1 \text{Var}[Z] + \theta_1 E[Z]^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2 \end{cases}$$

Proof. The first moment of Y is given by

$$E[Y] = \frac{1}{n} \sum_{i=1}^n Y_i = \theta_1 E[Z].$$

The second moment is given by

$$\begin{aligned} E[Y^2] &= \frac{1}{n} \sum_{i=1}^n Y_i^2 = \text{Var}[Y] + E[Y]^2 = \theta_1 \text{Var}[Z] + (1 - \theta_1) \theta_1 E[Z]^2 + \theta_1^2 E[Z]^2 = \\ &= \theta_1 \text{Var}[Z] + \theta_1 E[Z]^2. \end{aligned}$$

This yields the above result. □

Corollary 5.3.1. *If $Z \sim \text{Exp}(\theta_2)$, the moment estimators $\hat{\theta}_{1,MME}(Y)$ and $\hat{\theta}_{2,MME}(Y)$ are given by*

$$\hat{\theta}_{1,MME}(Y) = \frac{2 \left(\sum_{i=1}^n Y_i \right)^2}{n \sum_{i=1}^n Y_i^2}$$

and

$$\hat{\theta}_{2,MME}(Y) = \frac{2 \sum_{i=1}^n Y_i}{\sum_{i=1}^n Y_i^2}.$$

Proof. If $Z \sim Exp(\theta_2)$, $E[Z] = \frac{1}{\theta_2}$ and $Var[Z] = \frac{1}{\theta_2^2}$. Plugging these values into the equations for the first and second moment, we obtain

$$\begin{cases} \frac{\theta_1}{\theta_2} = \frac{1}{n} \sum_{i=1}^n Y_i \\ \frac{2\theta_1}{\theta_2^2} = \frac{1}{n} \sum_{i=1}^n Y_i^2 \end{cases}$$

$$\Rightarrow \frac{2}{\theta_2} \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n Y_i^2$$

$$\Leftrightarrow \theta_2 = \frac{2 \sum_{i=1}^n Y_i}{\sum_{i=1}^n Y_i^2}$$

$$\Rightarrow \theta_1 = \frac{\theta_2}{n} \sum_{i=1}^n Y_i = \frac{2(\sum_{i=1}^n Y_i)^2}{n \sum_{i=1}^n Y_i^2}$$

□

We will now consider the two parts of the two-part model separately.

First we consider the part of the two-part model which determines whether the observation is zero or positive, i.e. the part corresponding to the random variable Δ . Consider the sample $\delta = (\delta_1, \dots, \delta_n)$ defined by $\delta_i := \begin{cases} 0 & \text{,if } y_i = 0 \\ 1 & \text{,if } y_i > 0 \end{cases}$. We have that δ is a sample from i.i.d. random variables $\Delta_i \sim Ber(\theta_1)$.

The moment estimator of order 1 for θ_1 can be determined as follows.

$$E[\Delta] = \theta_1 = \frac{1}{n} \sum_{i=1}^n \Delta_i$$

\Rightarrow The moment estimator $\hat{\theta}_{1,MME}$ for θ_1 is given by

$$\hat{\theta}_{1,MME}(\Delta) = \frac{1}{n} \sum_{i=1}^n \Delta_i.$$

Now consider the second part of the two-part model. Again, letting $z = (z_1, \dots, z_{n-r})$ denote subsample of positive observations of y , we get that $Z \sim Exp(\theta_2)$. The first moment estimator for θ_2 is given by

$$E[Z] = 1/\theta_2 = \frac{1}{n-r} \sum_{i=1}^{n-r} Z_i$$

$$\Leftrightarrow \theta_2 = n - r / \left(\sum_{i=1}^{n-r} Z_i \right)$$

\Rightarrow The moment estimator of order 1 for θ_2 is given by

$$\hat{\theta}_{2,MME}(Z) = \frac{n - r}{\sum_{i=1}^{n-r} Z_i}.$$

Remark 5.3.1. We see that the moment estimators for the separate parts of the models are not the same as the moment estimators for the joint model. When deriving moment estimates for the joint model, the two parts of the two-part model may therefore not be considered separately.

5.4 Exponential Family

Let $Y \sim P^Y \in \{P_\theta, \theta = (\theta_1, \theta_2)\}$ be distributed according to the two-part model, i.e. let

$$Y = \Delta Z$$

with Δ and Z being independent, $\Delta \sim \text{Ber}(\theta_1)$, $Z \sim P^Z \in \{P_{\theta_2}\}$ and $P(Z > 0) = 1$. Let $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$.

Moreover, let y be a random sample of size n from independent and identically distributed random variables Y_i . Let r denote the number of zero observations in the sample.

Theorem 5.4.1. If $\Theta_1 = [0, 1]$, then $P^Y \in \{P_\theta\}$ does not form an exponential family.

Proof. Consider the probability measures $P_{(1, \alpha_1)}$, $P_{(0, \alpha_2)}$ and $P_{(\beta, \alpha_3)}$ with $\alpha_1, \alpha_2, \alpha_3 \in \Theta_2$ and $\beta \in (0; 1)$.

We have that $P_{(1, \alpha_1)}(\mathbb{R}^- \cup \{0\}) = 0$ but $P_{(\beta, \alpha_3)}(\mathbb{R}^- \cup \{0\}) > 0$ since $P_{(\beta, \alpha_3)}(0) = 1 - \beta > 0$.

Moreover, $P_{(0, \alpha_2)}(\mathbb{R} \setminus \{0\}) = 0$ but $P_{(\beta, \alpha_3)}(\mathbb{R} \setminus \{0\}) > 0$.

Therefore it follows that P^Y does not form an exponential family. □

Theorem 5.4.2. If $P^Z \in \{P_{\theta_2}\}$ is not an exponential family, then $P^Y \in \{P_\theta\}$ does not form an exponential family.

Proof. We have that the likelihood $L(\theta_1, \theta_2, y)$ of the sample y is of the form

$$L(\theta_1, \theta_2, y) = L_1(\theta_1, y)L_2(\theta_2, y)$$

with $L_1(\theta_1, y) = (1 - \theta_1)^r \theta_1^{n-r}$ and $L_2(\theta_2, y) = \prod_{y_i > 0} f(y_i | y_i > 0) = \prod_{i=1}^{n-r} f(z_i) =$

$L(\theta_2, z)$. Since $P^Z \in \{P_{\theta_2}\}$ is not an exponential family, $L(\theta_2, z)$ is not of the form (4.4). Thus $L(\theta_1, \theta_2, y)$ is not of the form (4.4) either. This implies that $P^Y \in \{P_\theta\}$ does not form an exponential family. □

Theorem 5.4.3. Let $P^Z \in \{P_{\theta_2}\}$ form a k -parameter exponential family with natural parameters $\zeta_j(\theta_2)$ and sufficient statistics $T_j(z)$, $j = 1, \dots, k$. Moreover, let $\Theta_1 = (0, 1)$. Then $P^Y \in \{P_\theta\}$ with $\theta = (\theta_1, \theta_2)$ forms a $k+1$ parameter exponential family with natural parameters $\zeta_j(\theta_2)$, $j = 1, \dots, k$ and $\zeta_{k+1}(\theta_1) = \ln(\frac{1-\theta_1}{\theta_1})$ and sufficient statistics $T_j(y)$, $j = 1, \dots, k$ and $T_{k+1}(y) = r$.

Proof. We have that $L(\theta_1, \theta_2, y) = L_1(\theta_1, y)L_2(\theta_2, y)$ with $L_1(\theta_1, y) = (1 - \theta_1)^r \theta_1^{n-r}$ and $L_2(\theta_2, y) = \prod_{y_i > 0} f(y_i | y_i > 0) = \prod_{i=1}^{n-r} f(z_i) = L(\theta_2, z)$.

Since $P^Z \in \{P_{\theta_2}\}$ is a k-parameter exponential family,

$$L(\theta_2, z) = L_2(\theta_2, y) = A(\theta_2) \left(\sum_{j=1}^k \zeta_j(\theta_2) T_j(z) \right) h(z).$$

Moreover, we have that $L_1(\theta_1, y) = (1 - \theta_1)^r \theta_1^{n-r} = \theta_1^n \exp(r \ln(\frac{1-\theta_1}{\theta_1}))$.

It follows that

$$L(\theta_1, \theta_2, y) = \theta_1^n A(\theta_2) \exp\left(r \ln\left(\frac{1-\theta_1}{\theta_1}\right) + \sum_{j=1}^k \zeta_j(\theta_2) T_j(y)\right) h(y)$$

which yields the above result. \square

5.5 Hypothesis Testing

We assume the following two-part model for the sample $y = (y_1, \dots, y_n)$ of observations from independent and identically distributed random variables Y_i

$$Y_i = \mathbb{1}_{\{1\}}(\Delta_i) Z_i,$$

where $\Delta_i \sim Ber(\theta_1)$, $\theta_1 \in (0, 1)$ and $Z_i \sim Exp(\theta_2)$, $\theta_2 \in \mathbb{R}^+ \setminus \{0\}$ are independent random variables. Thus $Y_i \sim P_\theta$ where $\theta = (\theta_1, \theta_2)$. Note that P_θ belongs to an exponential family.

Let r denote the number of zero observations in the sample y . Note that r is an observation from the random variable $R \sim Bin(n, 1 - \theta_1)$

Neyman-Pearson tests for simple hypotheses and known θ_2

Theorem 5.5.1. *The Neyman-Pearson test of size α for the testing problem*

$$\begin{aligned} H_0 : Y_i \in P_0 \text{ i.e. } \theta_1 = \alpha_1, \theta_2 = \beta \\ H_1 : Y_i \in P_1 \text{ i.e. } \theta_1 = \alpha_2, \theta_2 = \beta \end{aligned}$$

is given by

$$\phi(y) = \begin{cases} 1 & , R < c \\ \frac{\alpha - P_0(R < c)}{P_0(R \leq c)} & , R = c \\ 0 & , R > c \end{cases} .$$

The value of c is given by the solution to $P_0(R < c) = \alpha$ or, if such c doesn't exist, $P_0(R < c) < \alpha < P_0(R \leq c)$. Note that $R \sim Bin(n, 1 - \alpha_1)$ under H_0 .

Proof. We want to test

$$\begin{aligned} H_0 : Y_i \in P_0 \text{ i.e. } \theta_1 = \alpha_1, \theta_2 = \beta \\ H_1 : Y_i \in P_1 \text{ i.e. } \theta_1 = \alpha_2, \theta_2 = \beta \end{aligned}$$

The Neyman-Pearson tests are of the form

$$\phi(y) = \begin{cases} 1 & , p_0(y) < kp_1(y) \\ \gamma & , p_0(y) = kp_1(y) \\ 0 & , p_0(y) > kp_1(y) \end{cases}$$

We have that

$$p_0(y) = (1 - \alpha_1)^r \alpha_1^{n-r} \beta^{n-r} e^{\beta \sum_{i=1}^n y_i}$$

$$p_1(y) = (1 - \alpha_2)^r \alpha_2^{n-r} \beta^{n-r} e^{\beta \sum_{i=1}^n y_i}$$

To obtain a test of size α k is chosen so that

$$P_0(p_0(Y) < kp_1(Y)) = \alpha$$

or, if such k doesn't exist, so that

$$P_0(p_0(Y) < kp_1(Y)) < \alpha < P_0(p_0(Y) \leq kp_1(Y)).$$

$$\begin{aligned} P_0(p_0(Y) < kp_1(Y)) &= P_0\left(\frac{p_0(Y)}{p_1(Y)} < k\right) = P_0\left(\frac{(1 - \alpha_1)^r \alpha_1^{n-r}}{(1 - \alpha_2)^r \alpha_2^{n-r}} < k\right) = \\ &= P_0\left(\left(\frac{(1 - \alpha_1)\alpha_2}{(1 - \alpha_2)\alpha_1}\right)^r < k\left(\frac{\alpha_2}{\alpha_1}\right)^n\right) = P_0\left(r \ln\left(\frac{(1 - \alpha_1)\alpha_2}{(1 - \alpha_2)\alpha_1}\right) < \ln\left(k\left(\frac{\alpha_2}{\alpha_1}\right)^n\right)\right) = \\ &= P_0\left(r < \ln\left(k\left(\frac{\alpha_2}{\alpha_1}\right)^n\right) / \ln\left(\frac{(1 - \alpha_1)\alpha_2}{(1 - \alpha_2)\alpha_1}\right)\right) \end{aligned}$$

$$\text{Let } c = \ln\left(k\left(\frac{\alpha_2}{\alpha_1}\right)^n\right) / \ln\left(\frac{(1 - \alpha_1)\alpha_2}{(1 - \alpha_2)\alpha_1}\right).$$

Under H_0 , r is an observation of the random variable $R \sim \text{Bin}(n, 1 - \alpha_1)$.

We have that $P_0(p_0(Y) < kp_1(Y)) = P_0(R < c)$

Similarly, $P_0(p_0(Y) \leq kp_1(Y)) = P_0(R \leq c)$.

The value for c such that either

$$P_0(p_0(Y) < kp_1(Y)) = P_0(R < c) = \alpha \text{ or}$$

$$P_0(p_0(Y) < kp_1(Y)) = P_0(R < c) < \alpha < P_0(p_0(Y) \leq kp_1(Y)) = P_0(R \leq c)$$

can easily be computed.

From this we can derive

$$\gamma = \frac{\alpha - P_0(R < c)}{P_0(R \leq c)}.$$

Finally, k can be obtained through $k = \left(\frac{(1 - \alpha_1)\alpha_2}{(1 - \alpha_2)\alpha_1}\right)^{\frac{c}{n}} \left(\frac{\alpha_1}{\alpha_2}\right)^n$. □

Remark 5.5.1. *The Neyman-Pearson test is the most powerful test of size α for the above testing problem [06].*

References

- [01] Social Insurance in Figures 2014 (2014). Försäkringskassan.
- [02] Duan N., Manning W. G., Morris C. N. and J. P. Newhouse (1983). A Comparison of Alternative Models of the Demand for Medical Care. *Journal of Business Economics and Statistics* 1, pp. 115-126.
- [03] Tobin J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica*, Vol. 26, pp. 24-36.
- [04] Maren K. Olsen, Joseph L. Schafer. A Class of Models for Semicontinuous Longitudinal Data.
- [05] Yongyi Min, Agresti, A. (2002). Modeling Nonnegative Data with Clumping at Zero: A Survey. *JIRSS* , Vol. 1, Nos. 1-2, pp. 7-33.
- [06] Liero H., Zwanzig S. (2012). Introduction to the Theory of Statistical Inference.
- [07] Rydén J. (2014). A Statistical Analysis of Trends for Warm and Cold Spells in Uppsala by Means of Counts. *Geografiska Annaler: Series A, Physical Geography*
- [08] Lambert D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics*, Vol. 34, No. 1 , pp. 1-14.
- [09] Jie Gao (2007). Modeling Individual Healthcare Expenditures by Extending the Two-part Model, pp. 7-17.
- [10] Aitchison J. (1955). On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin. *Journal of the American Statistical Association*, Vol.50, No. 271, pp. 901-908.
- [11] Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, Vol. 47, pp. 153-161.