Analysis of Fresh Fish Demand in France and Estimation Methods for Heteroskedastic Systems of Zero Inflated Negative Binomial Models

Arnar Mar Buason

Norwegian University of Life Sciences
UMB School of Economics and Business
Master Thesis 30 credits 2013

Acknowledgments

The author wishes to thank Kyrre Rickertsen and Dadi Kristofersson for many helpful comments on earlier drafts, as well as their unwavering support throughout the writing of this master's thesis.

Arnar Mar Buason

Ås, Norway, May 2013

Contents

List of Figures

List of Tables

Abstract

This study analyzes consumer demand for the French fresh fish market, to provide a better understanding of the determining factors of fresh fish consumption. This is accomplished by estimating a system of demand equations with French household scanner data. In the data set, a large number of zero observations were generated from the low frequency of fresh fish purchases, which poses a problem for conventional methods used for such estimation. In order to analyze the data, an extension of the zero inflated negative binomial model is derived, which is referred to as the group heteroskedastic zero inflated negative binomial model. The model allows an estimation of a system of equations in the framework of zero inflated models, and therefore accounts for excess zeros in the data set, as well as overdispersion of the non zero observations. The results are consistent with previous studies and show that income and age are systematically positively related to the demand for all species of fresh fish.

1 Introduction

France is one of the largest consumer markets for fish products in Europe (INRA, 2007). It is therefore of vital importance to fish suppliers, such as Norway, to understand the consumer side of the market in detail; i.e. what effects would a change in the relative price of cod and salmon or increased income have on demand for fish, and do changes in demographics have an effect on the demand for fish? This study sheds light on the determining factors of French demand for the most important species of fish, such as salmon, cod, saithe, trout and whiting.

Previous studies have found that the typical French consumer of fresh fish is from an upper income bracket, he or she is relatively old and is from a household of two people,where the connection to income is assumed to stem from the fact that fresh fish is an expensive type of food (Girard and Paquotte, 2003). One possible explanation for why larger households consume less fish, might be because of childrens' preferences, as fish is rarely amongst childrens' favorite food. Given the recent demographic changes in Europe it is of interest to note that demand for fish increases with age. Therefore, demand for fish is likely to increase as Europe's populations grow older. This observation could however, also be due to generational differences in attitude towards fish (Girard and Paquotte, 2003).

In the last 20 years, demand for fish in France has gradually changed. The market has undergone structural changes, e.g. in the form of decrease in the price to income ratio, and been hit by intermediate shocks, such as the onset of mad cow crises in 1996 and 2000 (Allais and Nichele, 2007). The two most popular species of fish in France, in terms of value, are cod and salmon. In terms of quantity, salmon was the most bought fish type in France in 2008 and cod ranked seventh. Almost 2,5 kg of

salmon and 1,5 kg of cod were purchased per household. Even though cod ranked seventh in terms of quantity sold, there are still 31% of households who purchase cod, which is the second highest market participation rate after salmon, which 45% of the households purchased (ViaAqua, 2010). The demand for cod and salmon has undergone structural changes in the past few decades, both in the form of decreased landings and changed price to income ratio. For example, in the year 2000 cod landings in France were 12.000 tonnes, but had been 50.000 tonnes in 1987 and 160.000 tonnes in 1968. As a result, France began importing cod in large quantities, but since there was an overall decline in cod landings in the world, the imports did not completely compensate for the decreased landings and total French consumption of cod decreased (Girard and Paquotte, 2003). French consumers have slowly moved towards salmon, because of changes in relative prices between the species, and France has become the largest consumer of salmon in Europe. Furthermore, French consumer demand has moved from salted and dried fish to fresh fish. In the years 2000-2003, French households bought 51% of their cod fresh, 33% frozen and 16% salted or dried. Over the same period, French households bought 52% of their salmon smoked, 38% fresh and 10% frozen (Girard and Paquotte, 2003). The development towards fresh fish has continued and according to INRA (2007) fresh fish was the most frequently purchased product form of fish by households and commercial restaurants in 2006.

The objectives of this study are two. The main objective is to analyze the consumer demand for fresh fish in France, and therefore to better understand what factors influence fresh fish demand. For that purpose, the relationships between consumer demand for fish and prices as well as several socioeconomic variables; such as family size, age and income will be analyzed. Understanding such relationships could help

to predict future demand patterns for fresh fish, which could be important for large exporters of fish, such as Norway. The data set used for the analysis is a scanner data set provided by INRA Worldpanel. Each household in the sample is a consumer of fish, and register their purchases through the use of bar codes. One of the benefits of using scanner data, is that it provides a very detailed information on the properties of each household and their purchases of fish. Other food studies which have used scanner data are for example; Allais et al. (2010) and Bertail and Chaillavet (2008). One characteristic of this data set that complicates the econometric analysis to be performed, is the large number of zeros encountered. To be able to make the most of the data and in order to analyze any differences there might be in the consumption of various fish species, i.e. to perform hypothesis testing, it is desirable to estimate a system of equations. Therefore, the second objective of this paper is to derive an extension of the zero inflated negative binomial model.[1] The extension provides a simple way of estimating a system of demand equations in the framework of zero inflated negative binomial models. The reasons why a count data model such as this one is desirable are; the large share of zero observations which are accounted for by the zero inflated part of the model, overdispersion[2] of the non zero observations are taken into account by the use of the negative binomial distribution, and finally it allows the estimation of several equations which is necessary to conduct cross equation hypothesis as the different species of fish could be substitutes.

Heretofore, the estimation of demand for fish or other types of food has been con-

[1]The zero inflated negative binomial model is a count data model based on a mixture of the negative binomial model and the logistic model, which takes account of a large number of zero observations and overdispersion (Ridout et al., 2001).

[2]Dispersion is measured by dividing the sample variance by the sample mean, and when the variance is larger than the mean this will generate dispersion larger than one which is referred to as overdispersion (Green, 2008).

ducted by the use of almost ideal demand systems, estimated with Zellner's seemingly unrelated regression (SUR) (Zellner, 1962). Such an analysis has for example been conducted by Bertail and Chaillavet (2008) and Allais et al. (2010). To solve the problem created by the large number of zeros in the data set, which is generated by the low frequency of purchase of many food items, Allais et al. (2010) split the sample households into cohorts, based on geographic and socioeconomic classification, and then aggregated the data at the cohort level. In order to prevent the loss of information from such aggregation, this sort of method is not used in this analysis. In this analysis however, the fish consumption was split into four sub categories and a frequency variable created for each one. One variable for the frequency of salmon purchases, one for cod, another for other important species, such as trout and saithe, and finally one variable for all other fish. Since each household did not purchase from all four categories each year a large number of zeros was generated. It would be possible to aggregate the data by creating larger groups of fish for each category, but since salmon and cod were of interest, and not as a group of other fish, the aggregation approach was deemed unfeasible.

The rest of the paper is organized as follows. In section two a literature review of count data models and French food studies will be provided. In section three a theoretical microeconomic model is presented, the general Poisson and negative binomial count data models are presented, as well as the zero inflated negative binomial model and finally the group heteroskedastic zero inflated negative binomial model is derived. A detailed description of the French scanner data is presented in section 4. Section 5 contains the results of the estimation as well as hypothesis testing. Finally, section 6 concludes.

2 Review of the literature

In this chapter a brief literature review of consumer demand for fish in France is provided in order to provide the reader with an overview of the development of consumer demand for fish in France. A review of other French food studies is provided as well in order to compare conventional techniques of demand estimation with the count data estimation used in this analysis. A detailed review of the count data literature in economics is then presented. Everything from the basic Poisson and negative binomial models to the advanced systems of mixed count data models.

2.1 Brief review of French food studies

Girard and Paquotte (2003) analyze the French market for fresh fish and discuss the opportunities for farmed cod. They analyze the consumer demand for fresh fish in the years 1987 - 2002. The main changes in consumer demand for fresh fish over this period, were the changes of household cod purchases from 41% to 31% and the steady increase of fresh salmon purchases. Over this period French consumer have also moved from salted and dried fish towards fresh and frozen fish. The attitude towards farmed fish has slowly become more positive over the years, but wild fish is still preferred by most consumers, which they show by higher willingness to pay. The authors also claim that the typical profile of the fresh fish consumer is an older (senior), upper income person from a household of two people. Allais and Nichele (2007) analyze a similar period as Girard and Paquotte (2003), but with the aim of estimating structural changes in consumer demand for meat and fish in France in the years 1991 to 2001. Where a Marcov Switching AIDS model is used to estimate the changes to consumer demand for this period. In this period two mad cow crises

occurred in France which resulted in a sudden reduction of beef consumption. The MS-AIDS model was found to be able to accurately capture sudden changes in consumer demand. The paper also emphasizes the importance of accounting for structural changes when analyzing the evolution of demand.

Paquotte and Lem (2008) analyze the seafood markets of the mediterranian countires of the European Union; France, Greece, Italy, Portugal and Spain. France is a large importer of fish products and in 2007 the consumption of fishery and aquaculture products were around 2.2 million tonns, which is more than double the national production. In the 15 years prior to the paper being written, per household consumption of fish products had increased on average by 2% per year. It was also found that fresh and chilled fish products are most popular among French households.

Bertail and Chaillavet (2008) use a finite mixture AIDS model to estimate French consumer demand for fruit and vegetables and its implications to public nutrition policy. For the analysis a scanner data set, similar to the one used in this analysis, is used, where each household records all of their food purchases through the survey period. Many food items are not purchased frequently and therefore resulted in a large share of zero observations, therfore food items were aggregated into large categories to reduce the frequency of zero observations. This aggregation method enables the use of conventional methods for estimation, but at the cost of loss of information.

Allais et al. (2010) estimate the effects of a fat tax on French households purchases of food items, and the taxes impact on nutrition. For the analysis an AIDS model is used, where the estimation method is Zellner's seemingly unrelated regression (SUR). The study uses French scanner data provided by INRA Worldpanel, similar

to the data set used in this analysis. Since a large number of food items have a low frequency of purchase, this will generate a fair share of zero observations in the data set. In order to solve this problem the sample is split into cohorts, based on geographic and socioeconomic classification, and then aggregating the data at the cohort level.

2.2 Review of count data models

Count data models have been applied to a variety of problems, ranging from estimation of demand for recreation and doctor visits, to frequency of infant deaths. The most common use of count data models in the economic literature has been the estimation of recreational demand, which is both vast and has a long history in the econometric literature. Burt and Brewer (1971) proposed Zellner (1962) seemingly unrelated regression (SUR) model in order to estimate the demand for recreation and discuss the problems of doing so. The main problem mentioned stems from the great number of zeros likely to be encountered, which might introduce heteroskedacticity across observations. Burt and Brewer (1971) suggest tricks to circumvent these problems and provide an application in the form of estimating demand for different lakes in Missouri using the SUR model.

Even though Burt and Brewer (1971) recognized some of the difficulties of estimating recreational demand there were still numerous other elements to be considered. Two of them are the "count" nature of the observations and their zero truncation. Creel and Loomis (1990) show the importance of accounting for truncation of the dependent variable and suggest count distributions for the estimation. When collecting data on recreational demand, there will be no zero trip users, therefore the

sample is trunctated at zero. Since the data collected is often the number of trips taken to a specific site, the number of trips is given by a positive integer. The data generating process (DGP) underlying the observed data is therefore given by some probability distribution defined over positive integers. With the zero truncated data generated from a count data process Creel and Loomis (1990) compare estimations of different models. These models are the following; the Poisson, truncated Poisson, negative binomial, ordinary least squares, nonlinear normal and the truncated non-linear normal. The model which performed best in terms of explainatory power, was the Truncated Negative Binomial, which indicated the importance of using a count data model and taking the zero truncation into account.

Hellerstein (1991), as Creel and Loomis (1990), emphasises the importance of using statistical models that recognize the "count" nature of recreation demand data. For example, if OLS with a semilog transformation is used instead of a Poisson model, this will open up for the possibility of obtaining negative fitted values, which in turn will lead to a biased estimate. Even though the Poisson model is convenient, Hellerstein (1991) recognizes its limitations and stringent assumtions, such as the assumption of equidispersion[3] of the Poisson distribution and proposes a solution. Using an extension of the Poisson model, in the form of e.g. the negative binomial, will give a distribution with two parameters and accounts for odverdispersion. When distribution assumptions deviate from the underlying distribution, two estimation methods are applicable; the pseudo- (PML) and quasi generalized pseudo (QGPML) maximum likelihood estimation. Maximum likelihood estimation in this framework does not demand the correct specification of the distribution, which might be a dersirable property in many applications.

[3]The case were the mean is equal to the variance is refered to as equidispersion.

When collecting data on recreation for different sites it is obvious that one does not encounter any zero trip users, but recreation users might visit one site a few times in a given time interval and another site zero times, which might generate a large number of zeros in the data set, which in turn can create overdispersion. Lambert (1992) introduces an extension to the general Poisson model to deal with overdispersion caused by a large number of zeros in the response variable. This method is refered to as the zero inflated Poisson model and is a mixture of a logistic- and a Poisson model. The dependent variable takes the value zero with pobability p and with probability $1-p$ the dependent variable takes a value larger than one and is assumed to be drawn from the Poisson distribution. This method therefore solves the problem of overdispersion by introducing two separate processes. Ridout et al. (2001) developed a score test for determining wheter to use a zero inflated Poisson model or a zero inflated negative binomial model. Even though the inflation part of the Poisson model has taken account of the large number of zeros, and therefore in most cases solved the problem of overdispersion, it is still possible for the non zero observations to be overdispersed. When the sample variance of the non zero observations is significantly larger than the sample mean, the negative binomial distribution might be in order, to take account of the overdispersion.

To extend the count data models to a system of equations King (1989) proposed a joint Poisson regression estimator for estimating a system of two equations in order to improve on the equation-by-equation estimator and Zellner (1962) SUR estimator. The paper combines the Poisson model and the seemingly unrelated regression, into a seemingly unrelated Poisson regressions (SUPREM). The SUR estimator assumes a normally distributed error term, which is not applicable in a "counting" environment, thus using the SUR estimator with no changes will lead to biased, inefficient and

inconsistent estimates. Using an equation-by-equation Poisson model will not allow for cross equation hypothesis testing and assumes that the covariance between the parameter estimates is know *a priori*. The SUPREM estimator is consistent and asymptotically more efficient than the other two estimators applied to count data. The efficiency gain from SUPREM increases with increased correlation between the dependent variables, which also holds for the SUR estimator. One advantage that the SUPREM estimator has over its SUR counterpart, is apparent in the case when both equations contain the same independent variables, as in that case the SUR model collapses into a single equation OLS model. King (1989) also provides an application to presedential vetos, where different estimators are compared. At the onset, linear and log linear models are estimated equation-by-equation as well as jointly, and then the SUPREM model and an equation-by-equation Poisson are estimated. From the results the SUPREM performed best, followed by the equation-by-equation Poisson model. Those results are confirmed by Ozuna and Gomez (1994) who use King (1989) model for the estimation of demand for recreational boating. Three models are compared; SUR, SUPREM and Poisson. The estimation gave similar results for SUPREM and single equation Poisson, but SUR gave substantially different results. The paper mentions two limitations to the SUPREM. If the data presents overdispersion the estimates will be inefficient, and the assumption of an underlying Poisson distribution might not hold which will lead to biased standar errors and inefficient parameter estimates.

Yau et al. (2003) extend the zero inflated negative binomial model to the framework of mixed models. The zero inflated negative binomial model is then applied to pancreas disorder lenght of stay data. Where the newly introduced random effects term is used to account for inter-hospital variations and the dependancy of clustered

observations.

Egan and Herriges (2006) estimate demand for recreation at Clear Lake in north central Iowa using both observed behavior (OB) data (i.e. number of trips to specific sites) and continget behavior (CB) data (i.e. anticipated trips). Data of this sort contains at least two problems; truncation (exluding non-users) and endogenous stratification (oversampling frequent users), which are controlled for in the estimation. Since the OB and CB data is correlated, the recreational demand is estimated as a system of two equations. The estimation procedures used are; the multivariated Poisson-log normal (MPLN) and seemingly unrelated negative binomial (SUNB). The paper also extends the model for the estimation of recreational demand for multiple sites. The results indicated that, failure to account for on-site sampling procedures will lead to a substantial bias and also that observed trips and anticipated trips do not follow the same demand structure.

In this study a group heteroskedastic zero inflated negative binomial model is derived. The model uses the method of stacking to estimate a system of equations for count data with a large share of zeros. The model accounts for the share of zeros by introducing a zero inflated process, as well as accounting for overdispersion of the non zero observations, by assuming the negative binomial distribution for the non zero observations.

3 Theory

In this chapter a theoretical model for infrequent purchases is presented as well as an empirical model. The theoretical microeconomic model presented is an infrequency model (Meghir and Rogin, 1992), which accounts for the different frequency

of purchases between goods. Then a general overview of the most commonly used count data models will be presented; the Poisson, the negative binomial and the zero inflated negative binomial (ZINB) models. An extension of the ZINB model, the Group Heteroskedastic Zero Inflated Negative Binomial (GHZINB) model, will also be derived. This model provides a simple, and perhaps a less restrictive, way of estimating a system of count data model equations than the seemingly unrelated negative binomial regression (SUNB) from Winkelmann (2000) or the seemingly unrelated Poisson (SUP) regression of King (1989). The GHZINB is a system of two or more independent variables following the ZINB distribution. Each dependent variable has its own variance, which results in a group heteroskedastic covariance structure. The GHZINB model, unlike the SUNB and SUP models, is not restricted to any number of equations, since it is based on independently distributed random variables and not a bivariate Poisson distribution like the SUP and SUNB models, where the latter is derived from the bivariate Poisson distribution. The following sections will be organized as follows: The first section presents the theoretical model for infrequent purchases, the second provides an overview of the Poisson and negative binomial model, the third section provides an overview of the ZINB model, and in the fourth an final section the GHZINB model is derived.

3.1 Theoretical model

Each and every household has to make decisions regarding the purchases of a large number of goods each year. Some of these goods are purchased every week, other only a few times a year. For example, the average frequency of purchases of fresh

fish in France per household is only around 17 times.[4] Other goods such as clothes, plain tickets and recreational activities might fall into a similar category. In standard utility models where the consumer can choose from two different bundles of goods, such as Bockstael et al. (1987), the difference between the frequency of purchases is not modeled specifically. Since this is a study of the demand for fresh fish, it was deemed relevant to account for the difference in frequency of purchases of fresh fish and other goods in the theoretical model.

To model this behaviour in a microeconomic framework, the infrequency model of Meghir and Rogin (1992) is used. Assuming that a consumer purchases a vector of goods, X, in each interval of time, T, this set will contain goods which are purchased every week, some only a few times a year and then there is everything in between. Therefore there will exist an $x_1 \subseteq X$ which contains goods with a very low frequency of purchases, such as fresh fish. The corresponding vector of the frequency of purchases is given by N_1 and the corresponding price vector is given by P_1. All other goods are stored in the set $x_2 = X \setminus x_1$, which has the corresponding average frequency of purchases given by the vector N_2 and a price vector P_2. Each consumer also has some income denoted by M. The time interval T, is split into; hours working h, hours spent purchasing goods $L = L(N_1, N_2)$ and hours spent in other nonmarket activities l. The function $L(N_1, N_2)$ is generally assumed to depend on the location of the household, as well as household demographics, it is also assumed that there is a positive relationship between L and N_i, $\frac{\partial L}{\partial N_i} > 0$, for $i = 1, 2$. The period utility function is then assumed to have the following separable form $U = U\left(u_1(x_1, N_1) + u_2(x_2, N_2)\right)$. The consumer therefore does not only decide how much he wants to purchase, but also how often. Utility is assumed to be concave

[4]According to data provided by INRA Worldpanel, for years 2005 and 2006.

in x_1 and x_2.

$$\frac{\partial u_i}{\partial x_i} > 0 \text{ and } \frac{\partial^2 u_i}{\partial x_i^2} < 0, \text{ for } i = 1, 2 \tag{1}$$

It is also assumed that utility is increasing with frequency of purchases N_1 and N_2, $\frac{\partial u_i}{\partial N_i} > 0$, where this relationship reflects the benefits of keeping a smaller stock of goods in the household, and where freshness is important, for instance for food items such as fish, the frequency of purchases might improve the quality of the goods available in the household. The consumer's utility maximization problem is then given by the following expression:

$$\begin{aligned} \max_{x_1, x_2, N_1, N_2} \quad & U\left(u_1(x_1, N_1) + u_2(x_2, N_2)\right) \\ \text{s.t.} \quad & M - P_1' x_1 - P_2' x_2 = 0 \\ & T - h - l - L(N_1, N_2) = 0 \end{aligned} \tag{2}$$

Solving the maximization problem for the vector x_1, gives a system of linear demand equations $x_1^\star = f(P_1, M, N_1)$, which is a function of prices, income and the frequency of purchases. In the case of fresh fish, $x_1^\star$ would contain the consumer demand for goods, such as; salmon, cod, saithe and trout. All these goods are possible substitutes for each other and therefore a price change for one of these species can have an effect on the demand for another. Thus when examining the demand for fresh fish, it is important to include all the different species. Solving the maximization problem for N_1, gives the condition $\frac{\partial u_1}{\partial N_1} = \frac{\partial L}{\partial N_1}$, thus when the consumer is on an optimal path, the marginal utility from a higher frequency of purchases is equal to the cost of purchasing. The cost of purchasing is in the form of foregone leisure time, thus the opportunity cost of a higher frequency of purchases is leisure time.

A system of demand equations have been derived from a utility maximization problem, were it is assumed that the consumer does not only decide on how much he or she demands in a given interval of time, but also how often to purchase. The bundle of goods of interest are those with low frequency of purchases, such as fresh fish. In the following section a statistical model is derived for a system of such demand equations, where the model assumes low frequency of purchases and splits the decision making process into two discrete decisions, the first is, whether to purchase and the second is, given that one purchases how often does he purchase.

3.2 The Poisson and negative binomial distributions

The usual starting point of a count data analysis is the well known Poisson distribution, which represents the probability of a number of events taking place in a given interval of time. The Poisson density is given by the following expression:[5]

$$P(Y = y) = \frac{\exp(-\lambda)\lambda^y}{\Gamma(y+1)} \qquad y = 0, 1, 2, \ldots \tag{3}$$

where $\Gamma(y+1)$ denotes the gamma function. The random variable y is then said to be Poisson distributed, with the statistical property that $E(y) = \lambda$ and $V(y) = \lambda$. The distribution has the downside of being fairly restrictive, as it only possesses one parameter, λ. This property of the Poisson distribution is some times called equidispersion and for many data sets, this assumption might not hold, even though other aspects of the distribution fit quite well. Possible solutions to this problem and extensions to the general Poisson model will be explored in the following sections.

[5] A formulation of the Poisson distribution can be found in statistics text books, such as Cameron and Trivedi (1998).

In the application in this paper, the random variable y is a count variable and will therefore always be non-negative. To ensure that this attribute of the data still holds, the conditional mean is often given an exponential form, i.e. $E[y \mid x] = \lambda = \exp(x\beta)$. As the exponent is non-negative, this ensures that the conditional mean only takes on values that are non-negative for all values of x and β, where x is a matrix of covariates and β is a vector of unknown parameters.

When the data set to be analysed is over-dispersed[6] the usual extension to the Poisson model relaxes the assumption of equidispersion by introducing unobserved heteroskedasticity in the conditional mean, which is assumed to follow a gamma distribution. The new conditional mean then becomes $E[y \mid x, \varepsilon] = \exp(x\beta + \varepsilon) = h\lambda$, where ε follows the gamma distribution with mean 1 and variance $\frac{1}{\theta}$ and $h = \exp(\varepsilon)$. Adding this new random variable will produce a joint distribution for y and h. Thus integrating over h is needed to get the marginal distribution for y, which is given by the following density;[7]

$$P(Y = y) = \frac{\Gamma(\theta + y) r^{\theta} (1 - r)^{y}}{\Gamma(1 + y)\Gamma(\theta)}, \; r = \frac{\theta}{\theta + \lambda}, \; y = 0, 1, 2, \ldots \tag{4}$$

This marginal distribution of y (after allowing for unobserved heteroskedasticity) is the negative binomial distribution with conditional mean $E[y \mid x] = \lambda$ and conditional variance $V[y \mid x] = \lambda(1 + \frac{1}{\theta}\lambda) = \lambda[1 + \kappa\lambda] = \lambda + \kappa\lambda^2$, where $\kappa = V[h]$. The parameter κ is referred to as a dispersion or scale parameter. Cameron and Trivedi (1998) use the termenology NB2 model, because of its quadratic variance structure.

[6] Dispersion is measured by dividing the sample variance by the sample mean, and when the variance is larger than the mean this will generate dispersion larger than one which is referred to as overdispersion (Green, 2008).

[7] A formulation of the negative binomial can be found in statistics text books, such as Cameron and Trivedi (1998).

The authors also suggested replacing κ with $\lambda\kappa$ which results in the NB1 varinace function $V[y \mid x] = \lambda + \kappa\lambda$. In count data analysis the standard extension to the general Poisson model is the NB2 model. However most researchers do not give any statistical argument for their choice of variance function.

3.3 The ZINB model

The zero inflated models have been used for more than twenty years to extend the general count data model to deal with a large number of zeros. The first zero inflated model was introduced by Lambert (1992), where he extended the Poisson model to the zero inflated Poisson model, which is a mixture of the Poisson model and the logistic model. When a data set contains a large number of zeros as well as having the non-zero observations being over-dispersed, there is a need for another extension, and that is the negative binomial distribution, and then the model becomes a zero inflated negative binomial (ZINB) model. Formulations of the ZINB model can for example be found in Ridout et al. (2001) and in Yau et al. (2003). The zero inflated negative binomial distribution is defined as follows:

$$P(Y = y) = \begin{cases} p + (1-p)(1+\frac{\lambda}{\theta})^{-\theta}, & y = 0 \\ (1-p)\frac{\Gamma(y+\theta)}{\Gamma(1+y)\Gamma(\theta)}(1+\frac{\lambda}{\theta})^{-\theta}(1+\frac{\theta}{\lambda})^{-y}, & y > 0 \end{cases} \tag{5}$$

The distribution is a mixture of the negative binomial distribution and the logistic distribution, and it is split in two, one part for $y = 0$ and another part for $y > 0$. The distribution contains a parameter θ, which is often refered to as a dispersion parameter or scale parameter, which originates from the negative binomial part of the distribution. The Γ symbol represents the gamma function, which is defined

as $\Gamma(\theta) = \int_0^\infty y^{\theta-1} \exp(-\theta) dy$. The distribution contains two more parameters and they are λ and p. λ is the expectation from the negative binomial distribution and p is the probability of observing a $y = 0$. They parameters are defined as follows; $\lambda = \exp(x'\beta)$ and $p = \frac{\exp(z'\gamma)}{1+\exp(z'\gamma)}$.

The expectation and variance of the ZINB distribution are defined as follows:

$$E(y) = (1 - p)\lambda \tag{6}$$

$$V(y) = (1 - p)\lambda(1 + p\lambda + \lambda\theta) \tag{7}$$

From λ and p two linear predictors are created. They are defined as follows:

$$\log\left(\frac{p}{1-p}\right) = z'\gamma \tag{8}$$

$$\log\lambda = x'\beta \tag{9}$$

In equation (8) and (9) z and x are matrices of covariates, and γ and β are vectors of parameters to be estimated. The left hand side of equation (8) is the logarithm of the odds of $y = 0$, and the left hand side of equation (9) is the logarithm of the expectation of the negative binomial distribution.

3.4 The GHZINB model

The group heteroskedastc zero inflated negative binomial (GHZINB) model is an extension to the ZINB count data model, which provides a simple way of estimating a system of equations in the count data model framework, when the data is over-dispersed and when it contains a large share of zero observations. After providing the

brief overview of the ZINB model in the last section, the model can be extended to the GHZINB model, where each dependent variable has its own two linear predictors and variance function. The first step of the GHZINB estimation is stacking the data set, to get the following form:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix} \quad X = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & X_q \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix}$$

$$Z = \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & Z_q \end{pmatrix} \quad \gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_q \end{pmatrix}$$

Where each $Y_i = (y_1, \ldots, y_n)^t$, for $i = 1, 2, ..., q$, where q is the number of equations, $X_i = (x_1, \ldots x_r)$,where x_c is a $n \times 1$ vector, for $c = 1, 2, ..., r$, where r is the number of covariates, and $Z_i = (z_1, \ldots, z_k)$, where z_v is a $n \times 1$ vector, for $v = 1, 2, ..., k$, where k is the number of covariates, andβ_i is a $r \times 1$ dimensional vector of parameters, and γ_i is a $k \times 1$ vector of parameters. It is assumed that each Y_i is independently ZINB distributed. To give each Y_i its own variance, q dispersion parameters are estimated, one for each Y_i.

$$\Theta = \begin{pmatrix} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \theta_q \end{pmatrix} \otimes 1_n$$

where θ_i is the corresponding dispersion parameter for each Y_i and 1_n is an $n \times 1$ dimensional vector of ones. Since the variance function is a function of θ there will be q different variance functions which will create group heteroskedastic covariance structure between the Y_i. The variance covariance matrix is therefore defined as:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_q^2 \end{pmatrix} \otimes I_n$$

where σ_i^2 is the coresponding variance for each Y_i and I_n is an $n \times n$ dimensional identity matrix. This variance covariance structure gives the model its name, group heteroskedastic ZINB model, since the dependent variables are heteroskedastic between groups, but not amongst themselves. The linear predictors can now be expressed on matrix form as well:

$$\Lambda = \begin{pmatrix} \ln(\lambda_1) \\ \ln(\lambda_2) \\ \vdots \\ \ln(\lambda_q) \end{pmatrix} = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & X_q \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix} \tag{10}$$

$$P = \begin{pmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \vdots \\ \text{logit}(p_q) \end{pmatrix} = \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & Z_q \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_q \end{pmatrix} \tag{11}$$

The likelihood function of q independently ZINB distributed random variables Y_i is

given by the following expression:

$$L(\beta, \gamma, \Theta; Y, X, Z) = \prod_{i=1}^{n} \prod_{j=1}^{q} P(Y_{ij} = y_{ij}) \tag{12}$$

Inserting the ZINB distribution from equation (5) into the likelihood function in equation (12) and multiplying with negative one produces the (minus) log likelihood function for the GHZINB model, which is given by the following expression:[8]

$$\begin{aligned} l(\beta, \gamma, \Theta; Y, X, Z) \quad = \quad & - \sum_{j:y=0}^{n} \sum_{i}^{q} \ln \left(\exp(z'_{ij}\gamma_i) + \left(\frac{\exp(x'_{ij}\beta_i) + \theta_i}{\theta_i} \right)^{-\theta_i} \right) \\ & + \sum_{j:y>0}^{n} \sum_{i}^{q} \ln(1 + \exp(z'_{ij}\gamma_i)) \\ & + \sum_{j:y>0}^{n} \sum_{i}^{q} \left(\theta_i \ln \left(\frac{\exp(x'_{ij}\beta) + \theta_i}{\theta_i} \right) + y_{ij} \ln(1 + \exp(-x'_{ij}\beta_i)\theta_i) \right) \\ & + \sum_{j:y>0}^{n} \sum_{i}^{q} (\ln \Gamma(\theta_i) + \ln \Gamma(1 + y_{ij}) - \ln \Gamma(\theta_i + y_{ij})) \end{aligned} \tag{13}$$

The (minus) log likelihood function is then numerically minimized with respect to β, γ and θ, with an optimization algorithm, which is refered to as the Dual Quasy-Newton method (Dennis and More, 1997). The alorithm calculates the gradiants and then approximates the Hessian matrix, to find what values of β, γ and θ, minimize the (minus) log likelihood function.

Now that the GHZINB model has been derived, as an extension to the ZINB model, it can be taken to the data in order to estimate a system of equations. This, for example, enables the researcher to conduct cross equation hypothesis testing, in the environment of the ZINB models. Even though this model is an extension to

[8] A similar formulation of the ZINB log likelihood function can be found in Mwalili et al. (2007).

the ZINB model, it can just as easily be formulated for the standard zero inflated Poisson (ZIP) regression model. In the following chapters the GHZINB model is put to use, by estimating a system of demand equations for fresh fish in France.

4 Data

The scanner data used in this analysis is provided by INRA Worldpanel. The data set contains information on weekly purchases of fish for around 6,000 households in France for the years 2005 and 2006. Each household is a buyer of fish, however not necessarily fresh fish. Each household registers its purchases through the use of bar codes. The main objective of this study is to estimate the demand for fresh fish in France, and only purchases of fresh fish are included. Due to the large number of different species in the sample, they were divided into four categories. Some criteria was needed to determine the categories and it was decided to distinctly model the most important species in France, which can also be caught in Norwegian territorial waters. The two most important types are salmon and cod, which were each alloted their own categories. The third category contains other important fish types in France; whiting, trout, saithe and perch, which will be referred to as the benchmark category. The fourth category contains of all other fresh fish types. General statistics of the four categories are presented in table 1.

Table 1: General statistics

	Mean	Variance	S.E.	Frequency of Zeros	Index of Dispersion
Fish	16.54	774.09	27.82	16.95%	46.80
Salmon	2.29	25.23	5.02	52.03%	11.02
Benchmark	1.96	27.83	5.28	60.78%	14.20
Cod	1.95	25.77	5.08	59.21%	13.22
Other	10.35	360.77	18.99	25.23%	34.86

The sample average frequency of purchases of fresh fish is 16.54. Thus, on average, each household in the sample purchases fresh fish about 17 times each year. In table 1 the index of dispersion shows that fish purchases are overdispersed, since the variance is almost 47 times as large as the mean. This great overdispersion stems from the large number of small counts compared to the small number of large counts in the data set, this great difference is partly generated from the large number of zero observations. In table 2 general statistics of the non zero observations are presented.

Table 2: General statistics (Non zero observations)

	Mean	Variance	S.E.	Index of Dispersion
Fish	19.92	864.83	29.41	43.41
Salmon	4.77	40.77	6.39	8.55
Benchmark	5.01	55.75	7.47	11.13
Cod	4.77	49.72	7.05	10.42
Other	13.84	434.18	20.84	31.37

The average frequency of purchases of fresh fish is 19.92 for non zero observations. Thus, households which are consumers of fresh fish purchase on average about 20 times each year, which is slightly higher than for the whole sample, that is consumers who purchase fresh and non-fresh fish. The index of dispersion for the non zero observations is quite lower for salmon, cod and the benchmark category, than for the whole sample, but all categories are still overdispersed. This is a justification of the use of the negative binomial distribution in the GHZINB model, since the negative binomial distribution assumes overdispersion. The frequency distribution of the whole sample is given in figure 1.

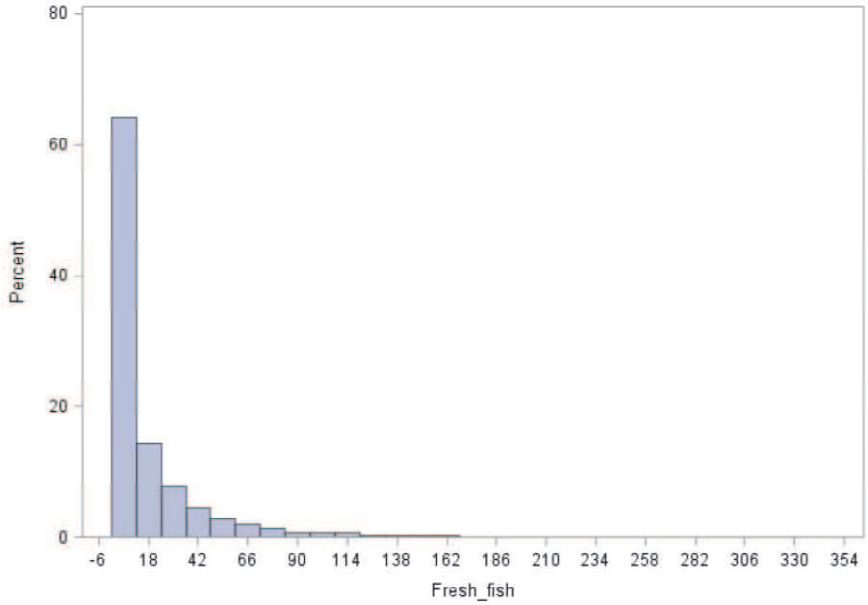

Figure 1: Frequency distribution of fresh fish

Turning to the different types of fish, the average frequency of fresh salmon purchased is around 2 each year. The index of dispersion is 11.02 which means that the variance is about 11 times as large as the mean, and the sample is clearly overdispersed. The frequency of zeros in the data set for fresh salmon is 5,929 out of 11,396 observations, thus around 48% of households which purchase fresh fish purchase fresh salmon. This percentage is consistent with the 45% purchase rate found for the year 2008 (ViaAqua, 2010). The frequency distribution for fresh salmon is presented in figure 2.

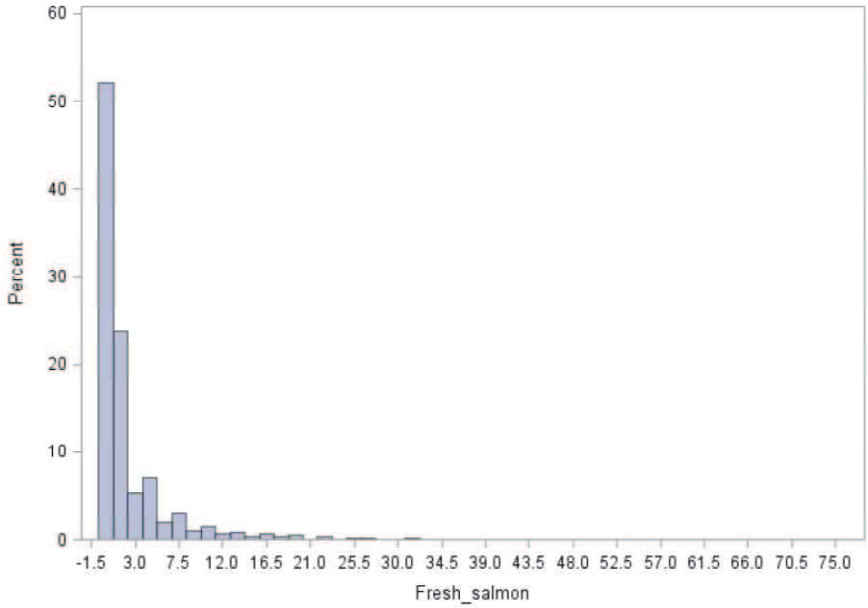

Figure 2: Frequency distribution of fresh salmon

The average frequency of fresh cod purchases is also around 2 each year, the same as for fresh salmon. The index of dispersion is 13.22 which means that the variance is about 13 times as large as the mean, and the sample is clearly overdispersed. The frequency of zeros in the data set for fresh cod is 6,748 out of 11,396 observations, thus around 41% of households who purchased fresh fish purchased fresh cod. The frequency distribution for cod is given in figure 3.

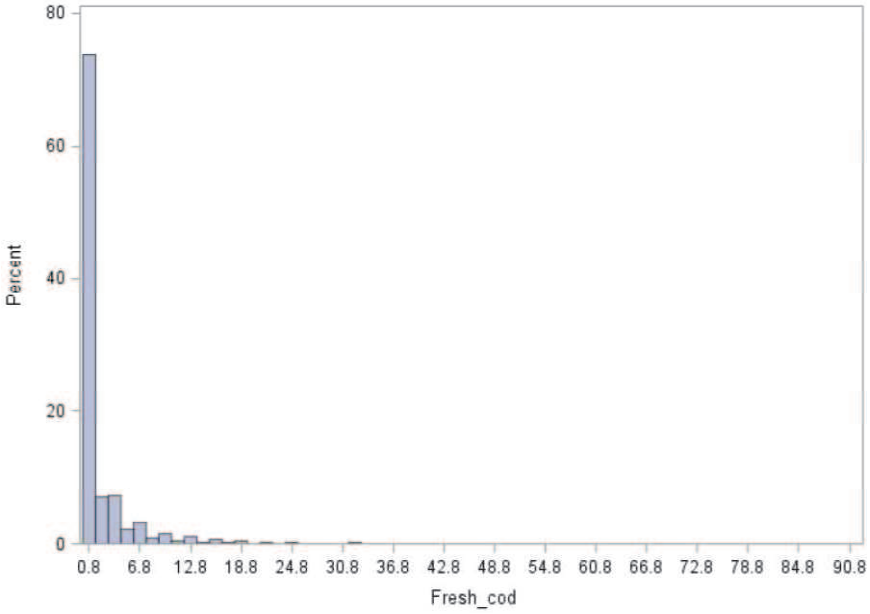

Figure 3: Frequency distribution of fresh cod

Information on a large number of socioeconomic variables, such as income, family size, age, number of cats in household etc. is provided in the data set. The set of variables that are included in the analysis is therefore restricted to those variables that have already been shown to be empirically relevant in other food studies such as Bertail and Chaillavet (2008) and Allais et al. (2010). These variables include; income, region, family size and age. In table 3 a description of the variables used in this analysis is given.

Table 3: Description of variables

Variable	Mean	S.E.	CV
Household income	1013.72	1363.32	1.34
Family size	2.58	1.34	0.52
Age	54.3	15.15	0.28
Coastal region	0.35	0.48	1.37
Unit price of salmon vs other	1.51	0.99	0.66
Unit price of cod vs other	1.44	0.57	0.39
Unit price of benchmark vs other	0.98	0.53	0.54

Even though these variables are included because of the empirical relevance found in other food studies, a general explanation of their relevance is in order. The relevance of price and income is clear from microeconomic theory, while the relevance of the other variables might need more explanation. The variable age has been found to be empirically relevant when estimating the demand for fish, and the reason might stem from the fact that older people are more concerned about their health than young individuals. The variable family size might be directly related to the age variable, since when people get older they are likely to live in a household of one or two individuals, but people of child bearing age are more likely to live in a household of three or more individuals, but young adults are more likely to live in a household of one or two people. Thus family size should be concave with age of the household representative. One more variable used in the analysis is a dummy variable for coastal regions, as individuals living in those regions have historically been consumers of fish.

The price per kg of each fish type is not given in the data set, only the quantities in grams purchased and the amount paid in euros. Then the price per kg faced by each household can be calculated and used in the estimation as a price. There are at least two problems associated with this method and that is the fact that those who did

not purchase any fish will not appear to be faced with any price and will therefore be dropped from the estimation. To solve this, the average price for all households in each region in France is calculated and that price is allocated to each household in that same region. Since there are seven regions; north, south, east, west, center west, center east and the Paris area, this will generate the variation needed to use the new price variable in the estimation. It should aslo be noted that the prices used in the analysis are relative to the unit price of the category of all other fish. The second problem is that unit values are likely to be endogenous since the unit value also depends on the quality of the good that is determined jointly with the quantity (Cox and Wohlgenant, 1986). An interesting topic for future applications would be to correct for this possible bias due to endogeneity.

5 Results

The results from the GHZINB model estimation of a system of four equations is presented in table 4. Since the model has two linear predictors, one for the logistic part of the model and another for the negative binomial (NB) part of the model, there are two sets of estimates for each equation, as well as one scale parameter for each equation in the system. If the scale parameter in the negative binomial model is equal to zero the model collapses into the Poisson model (Cameron and Trivedi, 1998).[9]

[9]For further discussion of the scale parameter see section 3.2 and 3.3.

Table 4: Results from the GHZINB model

		NB		Logistic Part	
	Coefficient (S.E.)	NB	Elasticity	Odds ratio (S.E.)	logistic part
Salmon	Constant	0.9501(0.1583)*		Constant	2.05(13.47)
	Family size	0.03315(0.01936)		Family size	0.06(0.7887)*
	Income	0.00017(0.000019)*	**0.3076**	Income	0.99(0.000193)
	Age	0.01948(0.001785)*		Age	1.02(0.01429)
	Price	-1.1051(0.03419)*	**-1.7485**		
	Coastal	0.04085(0.03902)			
Cod	Constant	-0.03743(0.1895)		Constant	6.46(1.1579)
	Family size	-0.02793(0.02365)		Family size	0.04(0.6164)*
	Income	0.000249(0.000022)*	**0.4506**	Income	1.00(0.000158)
	Age	0.03241(0.01953)*		Age	0.99(0.01292)
	Price	-1.0333(0.04048)*	**-1.5907**		
	Coastal	0.03188(0.0436)			
Benchmark	Constant	-0.9044(0.1891)*		Constant	187.86(0.6146)*
	Family size	-0.07134(0.02302)*		Family size	0.04(0.6351)*
	Income	0.000139(0.00002)*	**0.2515**	Income	1.00(0.000134)*
	Age	0.0396(0.002114)*		Age	0.95(0.008337)*
	Price	-1.0452(0.05197)*	**-0.9963**		
	Coastal	0.4949(0.04347)*			
All other	Constant	-0.00754(0.1106)		Constant	26,14(0,4112)*
	Family size	0.03351(0.01726)*		Family size	0,76(0,09332)*
	Income	0.000148(0.000015)*	**0.2678**	Income	1(0,000132)*
	Age	0.03554(0.001364)*		Age	0,95(0,0056)*
	Price				
	Coastal	0.1114(0.03302)*			
Scales	a1	0.8165(0.02753)*			
	a2	1.0201(0.02878)*			
	a3	1.0419(0.02834)*			
	a4	0.6039(0.02982)*			

Notes: All elasticities are conditional on households being consumers of the corresponding category of fresh fish. The * in the table represent that the corresponding parameter estimate is at least statistically significant at the 5% level.

Four equations were estimated and therefore four scale parameters were estimated, which all turned out to be highly significant, thus justifying the use of the a negative binomial model. The GHZINB model has two parts which represents two discrete decisions of the households. The logistic part represents the decision of whether

or not to purchase the corresponding species. The negative binomial (NB) then represents the decision of how often a household purchases, given that it is a buyer of the corresponding species.

The logistic results for salmon show that the estimate for family size was significant. All estimates from the logistic part of the model are presented as odds ratios,[10] therefore the estimate of 0.06 indicates that family size has a negative effect on the odds and probability of not purchasing salmon, and therefore family size has a positive effect on the odds and probability of purchasing salmon, i.e., an increase in family size increases the probability of a family buying salmon. The NB results for salmon show that the estimates for income, age and relative price were significant. Income and age are positively related to the frequency of purchases of salmon, given that the household is a buyer of salmon. The estimated conditional income elasticity of demand is 0.31, thus a 1% increase in monthly income increases the conditional frequency of purchases by 0.31%. Finally, price of salmon relative to other fish is negatively related to the conditional frequency of purchases of salmon, as was to be expected. The estimated conditional price elasticity of demand is -1.75, thus a 1% increase in relative price reduces the conditional frequency of purchases by -1.75%. Even though the estimates for age and income are highly significant, it is worth noting that the relationship between these variables and the conditional frequency of purchases of salmon is fairly weak.

The results for cod were very similar to the ones for salmon. The logistic results for cod showed the estimate for family size to be significant, and positively related

[10] Odds are presented as $\frac{p}{1-p}$, where p is the probability of not purchasing the corresponding fish type, which is a function of covariates x. The odds ratio is then the odds estimated at $x + \Delta x$ divided by the odds estimated at x. Thus providing the ratio $\frac{odds_{x+\Delta x}}{odds_x} = exp(\beta)$. The derivation of the odds ratio can be found in Green (2008).

to the odds and probability of purchasing cod. The NB results for cod showed that the estimates for age, income and relative price were significant. The variables age and income are positively related to the conditional frequency of purchases of cod. The estimated conditional income elasticity of demand for cod is 0.45, and has the same interpretation as the one for salmon. The price of cod relative to the price of other fish is negatively related to the conditional frequency of purchases of cod. The estimated conditional price elasticity of demand for cod is -1.59. Thus the income elasticity for cod is slightly higher than for salmon, but the price elasticity for cod is lower than for salmon. Later in this chapter the results from hypothesis testing are presented, which will determine if these differences are statistically significant or not.

The results for the benchmark category were quite different in some aspects to the result for salmon and cod. The logistic results for the benchmark category showed the estimates for family size, income and age to be significant. Both family size and age are positively related to the odds and probability of purchasing fish from the benchmark category. Even though the estimate for income is significant, it has approximately no effect on the odds and probability of purchasing from the benchmark category. The NB results showed that family size, age, income, relative price and the dummy variable for coastal regions are all significant. The variables age, income and coastal were all positively related to the conditional frequency of purchasing from the benchmark category. The estimated conditional income elasticity of demand is 0.25. The variables family size and relative price were negatively related to the conditional frequency of purchases. It is worth noting that the negative sign of the family size parameter shows the opposite effect to the effect of family size on the probability of purchasing. This might be because a larger household is more likely

to purchase any variety of goods, even though the frequency of purchases might be affected in a different manner. The estimated conditional price elasticity of demand for the benchmark category is -1, which is quite lower than the estimates for salmon and cod.

The results for the category of all other fish was quite similar to the results for the benchmark category. The logistic results for this category showed the estimates for family size, income and age to be significant. Both family size and age are positively related to the odds and probability of purchasing fish from the this category. Even though the estimate for income is significant, it has approximately no effect on the odds and probability of purchasing from the category of all other fish. The results for NB results showed that the estimates for family size, age, income and the dummy variable for coastal regions were significant. There is no estimate for relative price, since the price for each species was relative to the price of this category. The variables family size, age, income and coastal were all positively related to the conditional frequency of purchases from this category. The estimated conditional income elasticity of demand is 0.27.

The GHZINB model was estimated as a system of equations and thereby making it relatively simple to do hypothesis testing. To find out if there are any statistical differences between the parameter estimates across equations in the NB part of the model, a likelihood ratio test was conducted between all parameters of the same sort. In table 5 results from this likelihood ratio test are provided .[11]

[11]The likelihood ratio test statistic $S = 2\left(l(\beta_r) - l(\beta_u)\right)$, where $l(\beta_r)$ is the log likelihood function under the restricted estimationand $l(\beta_u)$ is the log likelihood function under the unrestricted estimation. The likelihood ration test statistic will be asymptotically chi-square distributed with $q - m$ degrees of freedom, when the null hypothesis is true, according to Wilks' theorem, where q is the number of parameters in the model under the assumption that the parameters are not equal and m is the number of parameters under the assumption that the parameters are equal. Then the p-values are computed using the asymptotic chi-square distribution of the chi-square statistic.

Table 5: Results from hypothesis tests

	Estimate (P-value)			Estimate (P-value)	
Family size	fs1-fs2	0.1045(0.001)*	**Income**	inc1-inc2	0.000031(0.266)
	fs1-fs3	0.06108(0.046)*		inc1-inc3	-0.00007(0.011)
	fs1-fs4	-0.00037(0.989)		inc1-inc4	0.000022(0.370)
	fs2-fs3	-0.04341(0.189)		inc2-inc3	-0.0001(0.000)*
	fs2-fs4	-0.1049(0.000)*		inc2-inc4	-0.0000091(0.723)
	fs3-fs4	-0.06145(0.036)*		inc3-inc4	0.000096(0.000)*
Age	age1-age2	-0.02012(0.000)*	**Coastal**	coa1-coa2	-0.454(0.000)*
	age1-age3	-0.01294(0.000)*		coa1-coa3	0.008968(0.878)
	age1-age4	-0.01607(0.000)*		coa1-coa4	-0.07053(0.168)
	age2-age3	0.007189(0.013)*		coa2-coa3	0.463(0.000)*
	age2-age4	0.00406(0.107)		coa2-coa4	0.3835(0.000)*
	age3-age4	-0.00313(0.189)		coa3-coa4	0.3835(0.000)*
Price	p1-p2	-0.05986(0.336)			
	p1-p3	-0.07181(0.175)			
	p2-p3	-0.01195(0.856)			

Note: In the table, 1 denotes the equation for salmon, 2 denotes the equation for the benchmark category, 3 denotes the equation for cod and finally 4 denotes the equation for all other fish.

The hypothesis tests for the parameter family size (fs), indicated that there was statistical difference between all estimates across equations, except for equation one (salmon) and equation four (all other), and equation two (benchmark) and equation three (cod). It is worth noting that the estimate for family size was only significant for the benchmark category and the category of all other fish. Thus the only hypothesis test for which the parameter family size was significant in the model and significantly different between two equations, was between the benchmark category and the category of all other fish.

The hypothesis tests for the parameter income (inc), indicated that there was significant difference between three estimates; equation one (salmon) and three (cod), equation two (benchmark) and three (cod), and equation three (cod) and four (all other). It is worth noting that all income estimates were found positive and significant in the model. The estimated difference between the income parameter indicates

that the effect from increased income is stronger for cod than all three other categories. This is no surprise, since cod showed the highest income elasticity from the model estimates.

The hypothesis tests for the parameter age, indicated that there was statistical difference between all estimates except for equation two (benchmark) and four (all other), and equation three (cod) and four (all other). It is worth noting that all age estimates were positive and significant in the model. The estimated difference between the parameter estimates indicates that the effect of increased age of the head of the household has the weakest impact on the conditional frequency of purchases of all the categories.

The hypothesis tests for the parameter coastal (coa), indicated that there was significant difference between all estimates except for, equation one (salmon) and three (cod), and equation one (salmon) and four (all other). The parameter estimates were however only significant for the benchmark and all other fish. The parameter estimates between those two categories were significantly different. The difference between the parameter estimates indicate that there is a stronger effect from living in a coastal region for the conditional frequency of purchases from the benchmark category, than for the category of other fish.

Finally the hypothesis tests for the parameter price (p), are not able to show any significant difference in the estimates. Thus, one cannot reject that an increase in the relative price of salmon, cod and the benchmark category to the price of other fresh fish will have the same effect on the conditional frequency of purchases.

The estimation results from the GHZINB model produced similar results as other food studies, such as Bertail and Chaillavet (2008) and Allais et al. (2010), where

variables such as; age, family size, region of residence etc. have been shown to be statistically relevant. The results show a systematically positive effect from age and income on demand across all categories, which is consistent with the claims from Girard and Paquotte (2003) that the consumer of fresh fish is an older person with above average income. The only difference between these results and the claims of Girard and Paquotte (2003) is that family size is not systematically negatively related to demand for fresh fish. The only systematic effect from family size is the positive relation between family size and the probability of purchasing from any category of fresh fish.

6 Conclusions

This study analyzed consumer demand of the French fresh fish market, to provide a better understanding of the determining factors of fresh fish consumption. France has for a long time been one of the largest consumer markets of fish products in Europe. Furthermore, the development of fish consumption in France has moved towards fresh products and increased quality (Girard and Paquotte, 2003), where fresh fish was the most frequently purchased product form by households and commercial restaurants in France in 2006 (INRA, 2007). This development has created opportunities for suppliers of fresh fish in France to provide a more expensive high quality product. Being able to predict consumption patterns and to understand the relationship between fresh fish consumption and socioeconomic variables such as; income, family size and age, is of vital importance for suppliers of fresh fish, such as Norwegian fish exporters, in order to predict future demand patterns. The study mainly focused on the consumption of the most popular species in France, salmon

and cod, but also other popular species such as; trout, saith, whiting and perch, which are all caught in Norwegian territorial waters.

The data set analyzed was French scanner data for the years 2005 and 2006, provided by INRA Worldpanel. The data set contains information on 6,000 French households. Each household in the data set purchases fish and registers its purchases through the use of bar codes. The data set also contains a very detailed description of each household's characteristics, everything from income to the number of pets and shoe size. This very detailed information is one of the strengths of scanner data, which have been used in other food studies such as Allais et al. (2010) and Bertail and Chaillavet (2008). With such detailed information the choice of socioeconomic variables to include in the analysis becomes complicated, since the possible combinations are nigh endless. For that reason, the set of variables that were used were restricted to those that had already been shown to be empirically relevant in studies of French food purchases. The included variables are relative prices, income, region, family size and age. Those variables have for example been used in Bertail and Chaillavet (2008).

Other food studies such as Allais et al. (2010) which have estimated similar relationships as the ones in this study have estimated an almost ideal demand system using the seemingly unrelated regression (SUR) of Zellner (1962). In order to get reliable results, using these conventional methods, in the presence of a large share of zero observations, some data manipulation is needed. In Allais et al. (2010) this problem was solved by splitting the sample households into cohorts, based on geographic and socioeconomic classification, and then the data was aggregated at the cohort level. Such aggregation methods were deemed undesirable in this analysis, since the

frequency of salmon and cod purchases were of interest individually, and not as a part of a group of other species. Therefore other methods were needed. Count data methods were used, because of how easily they can account for a large share of zero observations and overdispersion problems. For the estimation an extension to the zero inflated negative binomial model was derived. The extension is the group heteroskedastic zero inflated negative binomial (GHZINB) model, which can be used to estimate a system of equations, as well as deal with the large number of zeros in the data set and overdispersion.

Four equations were estimated as a system using the GHZINB model, where the dependent variables were the frequency of purchases of fresh salmon, fresh cod, other popular fish in France, including; saithe, trout whiting and perch, and the last dependent variable was the frequency of purchases of all other fish. The estimation results showed a systematically strong relationship between the frequency of fresh fish purchases and income, age and relative prices. Even though the results do not account for quantity purchased, it is likely that each household of similar size will purchase roughly the same quantity on each occasion regardless of how frequently they purchase fresh fish, since the product is bought fresh and is therefore unlikely to be stored for a long time. Age of the head of the household was found to be highly positively related to the purchases of fresh fish, households monthly income was also found to be highly positively related to fresh fish purchases, though both age and income showed fairly weak effects on the frequency of purchases. Lastly price was found to be highly negatively related to the frequency of purchases, as was to be expected, but different from the other relationships, the effect of changes in price on the frequency of purchases was fairly strong. These results are consistent with the claims made by other studies such as Girard and Paquotte (2003) that the consumer

of fresh fish is an older person in an upper income bracket. The only difference between the results in this study and the claims in Girard and Paquotte (2003) is that family size is not found to be systematically negative related to the frequency of purchases of fresh fish. For the most popular species, salmon and cod, no relationship was found between family size and the frequency of purchases, but family size was found to be positively related to the probability of being a consumer of fresh fish for all categories. This study therefore provides further evidence of the positive relation between fresh fish purchases and age and income, but not its negative relation to family size. If anything the results indicate a positive relationship between family size and the probability of being a consumer of fresh salmon and fresh cod.

For suppliers of fish in France, the results from this analysis could be used in at least three ways. Firstly, the results could be used in order to sell more fish to those households which are already frequent consumers of fish, by supplying fresh fish the whole year round, or by introducing marketing schemes aimed towards households where the head of the household is an older individual with above average income. Secondly the results could be used in order to sell more fish to those households which are not frequent consumers of fish, which are households of young adults with below average income. Thirdly the strong negative relationship between frequency of purchase and relative prices indicate that suppliers can increase the supply of fresh fish on the French market if they are willing to receive a slightly lower price for their product.

References

Allais, O., Nichele, V., 2007. Capturing structural changes in french meat and fish demand over the period 1991-2002. European Review of Agricultural Economics 34, 517–538.

Allais, O., P.Bertail, Nichele, V., 2010. The effects of a fat tax on french households' purchases: A nutritional approach. American Journal of Agricultural Economics 92, 228–245.

Bertail, P., Chaillavet, F., 2008. Fruit and vegetable consumption patterns: A segmentation approach. American Journal of Agricultural Economics 90, 827–842.

Bockstael, N. E., Strand, I. E., Hanemann, W. M., 1987. Time and recreational demand model. American Journal of Agrecultural Economics 69, 293–302.

Burt, O. R., Brewer, D., 1971. Estimation of net social benefits from outdoor recreation. Econometrica 39, 813–827.

Cameron, A. C., Trivedi, P. K., 1998. Regression Analysis of Count Data. Cambridge University Press.

Cox, T. L., Wohlgenant, M. K., 1986. Price and quality effects in cross-sectional demand analysis. American Journal of Agrecultural Economics 68, 908–919.

Creel, M. D., Loomis, J. B., 1990. Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in california. American Journal of Agricultural Economics 72, 434–441.

Dennis, J. E., More, J. J., 1997. Quasi-newton method, motivation and theory. SIAM Review 19, 46–89.

Egan, K., Herriges, J., 2006. Multivariate count data regression models with individual panel data from an on-site sample. Journal of Environmental Economics and Management 52, 567–581.

Girard, S., Paquotte, P., 2003. The french market for fresh fish: an opportunity for farmed cod? In: XV EAFE Conference Proceedings.

Green, W. H., 2008. Econometric Analysis. Prentice Hall.

Hellerstein, D. M., 1991. Using count data models in travel cost analysis with aggregate data. American Journal of Agricultural Economics 73, 860–866.

INRA, 2007. 5 scenarios for french fish farming to 2021. Tech. rep., INRA Fish Commission.

King, G., 1989. A seemingly unrelated poisson regression model. Sociological Methods and Research 17, 235–255.

Lambert, D., 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.

Meghir, C., Rogin, J. M., 1992. Frequency of purchase and the estimation of demand systems. Journal of Econometrics 53, 53–85.

Mwalili, S. M., Lesaffre, E., Declerck, D., 2007. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Statistical Methods in Medical Research 17, 123–139.

Ozuna, T., Gomez, I. A., 1994. Estimating a system of recreation demand functions using a seemingly unrelated poisson regression approach. The Review of Economics and Statistics 76, 356–360.

Paquotte, P., Lem, A., 2008. Seafood markets and trade: A global perspective and an overview of eu mediterranean countries. Options Mediterraneenes 62, 43–55.

Ridout, M., Hinde, J., Demetrio, C. G. B., 2001. A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57, 219–223.

ViaAqua, 2010. Elements about french cod market. Tech. rep., Via Aqua Seafood and Prospective.

Winkelmann, R., 2000. Seemingly unrelated negative binomial regression. Oxford Bulletin of Economics and Statistics 62, 553–560.

Yau, K. K. W., Wang, K., Lee, A. H., 2003. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical Journal 45, 437–452.

Zellner, A., 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association 57, 348–368.