ESTIMATION OF POLICY EFFECTS AND COUNTERFACTUAL DISTRIBUTIONS: AN APPLICATION TO FOOD SECURITY ANALYSIS IN MEXICO

A Dissertation

by

PAUL ESTEBAN NAVAS ALBAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	David Bessler
Committee Members,	Ximing Wu
	David Leatham
	Keli Xu
Head of Department,	Parr Rosson

August 2015

Major Subject: Agricultural Economics

ABSTRACT

The analysis of the effects of a policy and the factors that mediated those effects are among the core objectives of the economic science. The aim of this investigation is to present a recent methodological breakthrough in this field named distribution regression and to use it conjointly with causal analysis, to estimate the effects of a policy. The novelty relies in the estimation not only of the expectation of the difference between observed and counterfactual outcome, the usual Average Treatment Effect, but of the entire distribution of this counterfactual random variable and its distribution as a stochastic process. This estimation uses tools stemming from the field of Empirical Processes, to allow for the estimation of these counterfactual distributions, which will be supported with the methods of machine learning and causal analysis to tackle the problem of the causality structure among a set of variables and to attempt to justify the causal interpretation of the counterfactual distributions obtained.

These new tools of policy analysis are applied to the 2012 avian flu outbreak in Mexico in July. The variable of interest in the analysis is the food security of the population, with the research question driving this application being whether there was a negative effect in the nutritional status of the population, in particular that of rural areas, once other factors such as household expenditure, educational status and others have been controlled for; estimating counterfactual distributions of what would have happened to the whole and to specific portions of the population had the outbreak not occurred finds that the average treatment effect is significant, both for the before and after comparison and for the rural and urban comparison in the

expected way, with the outbreak reducing the food security of the population, with factors as the number of household members, whether the family experienced a setback and the gender of the provider being the driving forces. Another set of findings shows that it is not possible to conclude that the effects are heterogeneous on the population, since the quantile treatment effects are not statistically significant using uniform confidence intervals for any of the different setups investigated.

DEDICATION

To my family

ACKNOWLEDGEMENTS

I want to gratefully acknowledge the support of my committee members, for their help, time and patience they dedicated to me throughout my doctoral studies. They were not only great sources to address the theorical inquiries I had, but showed a predisposition to support me through adverse personal situations, for which I am forever indebted to them. I am indebted to Dr. David Magana, who offered many insightful discussions on this topic and shared many useful insights. I want to acknowledge the support of the Fulbright Scholarship, which supported me financially during a part of my doctoral studies. I want to thank the Department of Agricultural Economics for providing a great atmosphere to perform research and the TAMU Libraries, which honored every request for literature I made, no matter how difficult; they did a superb job. Of course, this enterprise would not have been possible without the relentless support of my family, in particular my wife, Elena, who endured much to get me through this process, the constant bliss that my children bring to me every day, and the guiding light of my family in the distance, always in my thoughts.

NOMENCLATURE

ChFM	Chernozhukov, Fernandez-Val and Melly
DAG	Directed Acyclic Graph
ENGASTO	Mexican National Expenditure Survey
SENASICA	Mexican National Service of Agricultural Food Sanitation, Safety and Quality
vdVW	van der Vaart and Wellner
GC	Glivenko-Cantelli
D	Donsker

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

The analysis of the effects of a policy and the factors that mediated those effects are among the core objectives of the economic science. The last few decades have seen an enormous development of the models and tools employed in this realm. The field of quasi-experimental economics is probably the field that has produced some of the most fruitful innovations; its main approach is to adapt the methodology from experimental science and test the effect of a treatment on a given population by an event analogous to an experiment, where a segment of the population is exposed to the treatment and the results on a variable of interest are compared to those in the control group.

The aforementioned quasi-experimental approach and in particular the analysis undertaken using the tools of linear regression models have achieved a degree of maturity reflected in the possibility of identifying and estimating consistently, under certain circumstances and assumptions, the effect of a policy. These particular circumstances and assumptions will be addressed further in the theoretical section of the dissertation, but they have to do mainly with the degree to which the observational situation resembles that of an experimental setup, at least in its main distinguishing features. The policy effects that be ca identified with this linear approach are the so-called Average Treatment Effects (ATE) of a policy, which are usually the expected differences between the value of the outcome of interest and the values that what would have been, had the policy not been in effect.

This type of what-if analysis is termed counterfactual, and it lies at the heart of

experimental and quasi-experimental economics. As mentioned, the interest of the investigator is in the effect of the policy and the ability to determine the ATE represents a quality leap in this perspective. The aim of this investigation is to present a methodological extension in this direction, which consists in the estimation, not only of the expectation of the difference between observed and counterfactual outcome, but of the entire distribution of this counterfactual random variable and its distribution as a stochastic process. This a recent breakthrough in the field of Econometrics, which uses statistical tools stemming from the field of Empirical Processes, to allow for the estimation of these counterfactual distributions; these distributions have been estimated elsewhere, but the results of the work of Chernozhukov et al (2013), on which this dissertation is based, is innovative and far-reaching, in that it performs inference on the counterfactual distributions, which allows to speak about the precision of the estimation of the whole distribution.

Apart from analyzing the effect of a policy, one is interested in the causes underlying that effect. This is the second theme of this dissertation, which extends the estimation of counterfactual distributions of the variables of interest and asks for the possibility of a causal interpretation of the results. It is customary to simply assume that condition - known as conditional exogeneity-allowing for a causal analysis to hold; it asserts that the variables used as controls in the estimation are sufficient to explain all the ways in which the effect may have come to stand. This is a very strong assumption and has many ramifications; the second aim of this dissertation is to utilize the methods developed in the field of machine learning and causal analysis, mostly by Pearl (2009) and the Carnegie Mellon school, to tackle this problem and to attempt to justify the causal interpretation of the counterfactual distributions obtained.

The third and final objective of this dissertation is to present an application of the aforementioned estimation techniques to a recent outbreak of avian flu that happened in Mexico in July 2012. The variable of interest in the analysis is the food security of the population, which is measured by an index constructed from the answers to 12 questions regarding the food accessibility of the population. These questions are part of the database for the Mexican National Expenditure Survey (ENGASTO, for its name in Spanish), which is a yearly nation-wide survey conducted by the Mexican government. This database is publicly available for the year in mention. The research question driving this application is whether there was a negative effect in the nutritional status of the population, in particular that of rural areas, once other factors such as household expenditure, educational status and others have been controlled for; estimating counterfactual distributions of what would have happened to the whole and to this portion of the population had the outbreak not occurred is the method chosen to attempt an answer to this inquiry.

The dissertation is divided in two main chapters: one theoretical and one empirical. The first one is divided in sections, the first one devoted to an introductory explanation of the counterfactual model of causal inference and elaborating on the difference between estimating an average treatment effect and a counterfactual distribution. With this setup in place, the second section will present the theoretical framework developed by Chernozhukov et al (2013) to estimate the distributions, together with an explanation of the mathematical foundation upon which it is based, namely the theory of empirical processes. The methods explained in this section have a very elaborate mathematical basis, with many theorems and claims needed to establish their validity; in order to maintain the readability, the general setup will be

explained and references will be provided for the proofs and the theorems of relevance.

Next, we will proceed to present the work of Pearl on machine learning and causal analysis, whereby the same admonition regarding mathematical content applies. The final chapter presents the application of these tools to the aforementioned analysis of food security in the face of an avian flu epidemic.

2. THEORETICAL SECTION

2.1 The counterfactual setup for policy analysis

The counterfactual approach to the estimation of the effect of a policy has its roots in the experimental method, originally developed to test the effects treatments and drugs. The principle is simple: the effect of a treatment on a population can be inferred by an experiment where individuals of similar characteristics are assigned randomly to a group that will receive the treatment (T) or to a control group (C) that will not. Given the similarity of the subjects, it is assumed that they will respond to the treatment identically, so any effect observable in a variable of interest can be attributed to the treatment.

In this setting, the experimenter plays a transcendental role, controlling that the mechanism assigning subjects to the groups is indeed random and individuals are not being selected into the treatment or the control group based on some (objective or subjective) criteria. A way to guarantee this lack of selection is to blind (or double-blind) the participants (and the experimenters), so that they are not aware of the subjects membership to a group, having the members of the control take a placebo to keep this lack of knowledge intact. The controlling role of the experimenter is probably the most relevant issue to underline: any manipulation of the entire experiment is under his or her control, so that the behavior of the subjects is not influenced but by their membership to T or C.

An economist interested in estimating the effect of a policy may perform an imaginary experiment and think of an ideal situation enabling the unequivocal de-

termination of the consequences of a policy. In some circumstances, undertaking such an experiment in reality may be possible, but in some others, practical complications and ethical concerns prevent the investigator from doing so. It may be nonetheless possible to perform policy analysis in a non-experimental setup; this is the paradigm underlying observational studies: under some circumstances and assumptions, using data and information that may not have been obtained under an experimental setup may be still render valuable information regarding the effect of a policy. This may occur, for example, if the selection of the individuals being affected (positively or negatively) by a policy happened, probably without intent, in a random fashion.

One can then argue that it is possible to encounter such quasi-experimental situations, where it is plausible to answer the same type of questions as in an experimental study. This approach was pioneered by Donald Rubin (1974), and it is known as the potential outcomes or counterfactual approach, whereby Rubin himself credits the origins and much of the insight to Neyman and Fisher in Rubin (1990). The reason for the name is that, for any individual i, once a treatment status has been assigned and fulfilled, it will only be possible to observe his outcome variable of interest, call it Y_i. But before this status is fulfilled, it is possible to think of the potential outcomes for this variable, $Y_{i,t}$ or $Y_{i,c}$, each of them a legitimate random variable corresponding to the value of the outcome if treated and not treated, respectively. As a random variable, it is possible to calculate its expectation or variance or any other statistical quantity that may be of interest.

Nothing prevents us from thinking that way, but the issue is whether -and how- these quantities can be estimated from the data from an observational study. Indeed, for each individual i, in the best scenario, we *observe* either $Y_{i,t}$ or $Y_{i,c}$, never both,

because they are mutually exclusive, so even then, based on a sample that is at most of size 1, no inference is possible. To overcome this situation, it is customary to assume that individual i is the $i-th$ member of a population of similar characteristics. In the experimental setting, the subjects can be chosen to be as similar as possible, so that any effects can be attributed to the application of the treatment; in observational studies, this is a salient complication, as one is required to argue that the subjects can be regarded as members of a homogeneous population that will respond to the policy under review in a comparable fashion.

This implies, in particular, that there are not characteristics of the subjects that influence their response to the treatment and also, that the subjects in the treatment or the control group were not assigned to either one based on these characteristics, so that there is not a dependence of treatment assignment or response on any particularities of the subjects. If this was the case, the effects on any outcome of interest cannot be traced back to the treatment, but to the interaction of treatment and characteristics of the individuals, and one could at best compare the effect of the treatment on subjects that share that same characteristic. There are many other complications that can arise, even in an experimental study, which could compromise the accuracy of the inference of the treatment effect; for instance, it could happen that a subject receiving or not a treatment may interfere with the reaction of another subject to the treatment or that the intensity of the treatment may not be comparable among subjects because they respond differently to the same situation.

Because of all of these potential complications, the researcher working in an observational setup starts with a set of assumptions under which the estimation of the treatment effect is possible by use of a particular method. The most common

approach to policy analysis is to calculate the expectation of $[Y_{i,t} - Y_{i,c}]$, which is deemed the average treatment effect (ATE). The workhorse of economic policy analysis is the linear regression model, so the usual approach to calculating the ATE is to construct a linear model and to argue that the covariates included in the model are predetermined regressors. With the final assumption of conditional independence, which asserts that, conditional on the covariates included in the model, the outcome variable of interest is independent of the treatment status, it is possible to identify ATE unambiguously and the analysis can be given a causal interpretation. This setup can be extended for treatments with different intensity levels, and many other that will not be of relevance in this work.

2.2 Estimating counterfactual distributions

One of limitations of the linear approach to estimation of effects serves as the motivation for the extension presented here, namely the fact that the linear model allows for estimation and inference of the ATE, given by the expectation of the counterfactual quantities as defined above, which assumes tacitly that the effect is homogeneous on the whole population. Recent techniques permit the estimation of not only the ATE but of the distribution of the counterfactual random variables involved, namely Y_t and Y_c, without additional assumptions as those needed to estimate and attribute a causal meaning to the ATE.

There have been many efforts aiming at this kind of extension, mainly because of the interest in situations in which the treatment effect may be heterogeneous along some characteristic of the subjects. The ATE, which averages over this heterogeneity by definition, is limited in this respect. In this context, a way to tackle the hetero-

geneity of the treatment is to estimate quantile treatment effects (QTE), which are defined as the horizontal distance between the actual distribution of the outcome variable and the counterfactual distribution. Another measure of interest regarding the heterogeneity of the treatment is the Gini coefficient, which measures the inequality in the distribution of the outcome variable. Again, one can compare the Gini coefficient obtained for the actual distribution with that of the counterfactual distribution and assess the degree to which the probability is not evenly distributed, as is usually done with income and the Gini coefficient as a measure of inequality. We will concentrate in the first approach here.

Both examples require the estimation of the distribution function of the variable and its counterfactual, in order to calculate the QTE or the Gini index. The statistical literature is not short of approaches to estimate distribution functions, whereby parametric approaches impose assumptions in the functional form of the distributions of interest and usually employ maximization techniques; other approaches deemed semiparametric, for instance the probit and logit models, assume no parametric form on the distribution itself but on the parameters of it distribution, so they allow for some flexibility with respect to parametric approaches. Finally, there are nonparametric approaches which further weaken the assumptions made on the model and relax to certain regularity or differentiability assumptions, as is the case with the estimation of distributions by the kernel method. Some salient examples of estimation of counterfactuals using these techniques include the work of DiNardo, Fortin and Lemieux (1996) and Firpo (2007).

The methodology used in these approaches differs from the one presented in this dissertation, which is based in the work of Chernozhukov et al (2013), which will be

referred to as ChFM from this point on. While the results obtained in the aforementioned works are highly relevant and include very compelling policy analyses, the methodology based on empirical processes constitutes a qualitative leap in terms of the type of inference that can be made on the estimated distributions. Indeed, the asymptotic theory of kernel estimators of counterfactual distributions, for instance, is based on pointwise convergence of the estimators, which allows to prove their consistency, with the pointwise- inference based, however, only on specific points of the distribution; the results stemming from empirical processes allow for inference of the entire distribution, which enables inference results for the whole distributions seen as stochastic processes, since the convergence is uniform.

Moreover, once the convergence results for this type of estimation are obtained, the ChFM shows that, in the first of its two most relevant contributions, it is possible to perform inference by means of bootstrapping, which enriches the quality of the results obtained. The second contribution, which is deemed distribution regression, shows that the inference results obtained are model free, in that, independently of the semiparametric link function used, the inference coming from the bootstrap is asymptotically equivalent and overcomes the issue of non-pivotality common to this type of estimation. The next section will elaborate on the theory of empirical processes and the estimation of counterfactual distributions using distribution regression.

2.2.1 Counterfactual distributions by distribution regression

Given two sets of random variables Y and X with a joint probability distribution F_{YX}, the marginal distribution of the variables Y can be computed from integrating

them with respect to their conditional distribution with respect to X, $F_{Y|X}$, over the support of the distribution. For the purposes of presentation, this can be simplified to the case of Y being a single random variable.

Recalling the context of estimation of policy effects presented previously, the variable of interest in this case is the outcome, Y. In the counterfactual setup, given a treatment or source of variation, one has the variable Y_i for the subject i, with its potential outcomes $Y_{t,i}$ and $Y_{c,i}$, which a priori constitute two random variables; a posteriori, only one of the variables will be observable for each subject, but assuming subjects are similar enough, one can consider the population as being divided in two homogeneous groups, T and C. This setup can be extended to allow, for instance, treatments that have different intensities or characteristics as mentioned earlier, but we will maintain only these two populations for ease of exposition.

The distribution of the observed variable Y_t given characteristics X, denoted by $F_{Yt|Xt}$, is obtained from the integration of the conditional distribution over its support, as expressed in the following equation, which is a restatement of an obvious fact. It will become more interesting once reference is made to the populations T and C involved:

$$F_{Yt|Xt}(y) = \int F_{Yt|Xt}(y|x) dFX_t(x) \tag{2.1}$$

To implement this procedure, it is necessary to obtain an estimator of the conditional distribution by a statistical method of choice, and then proceed to the integration; the methods to estimate the conditional distribution range from parametric, over semiparametric and non-parametric. The method chosen in this dissertation is

a semiparametric one, called distribution regression; the reason for this choice has to do with the nature of the data in the empirical section, which is discrete, and for which the method chosen performs well. The method will be explained in the following paragraphs.

The counterfactual distribution of the variable Y_t, the outcome variable under treatment, had the conditions of control population C been given, differs from the distribution mentioned in the last paragraph in that it is not readily estimable from the data, as it is not observed. It is obtained from the integration of the conditional distribution of the variable $F_{Yc|Xc}$ over the support of X_t, which renders the distribution that would have prevailed for Y_t had the subjects faced the conditions of population C. This approach to obtaining a counterfactual distribution has been presented, among others, in the work of DiNardo et al (1996), whereby the conditional distribution is estimated by means of a kernel regression estimator and weights are used to correct for the propensity of treatment assignment given the characteristics of the population.

In order for the estimator to be meaningful, it is necessary that the supports of the treatment and control populations overlap, which is a common assumption made in the context of policy effect analysis. In case this assumption is not fulfilled, the supports of both populations ought to be trimmed, to make the estimation viable. It will also be assumed here that conditional exogeneity with respect to the treatment holds, so that it is not necessary to correct with the propensity weight.

ChFM presents different alternatives to estimate the counterfactual distribution, mainly depending on the choice of estimator of the conditional distribution. One

possible choice, apart from the kernel estimator already mentioned, is estimation by quantile regression, which we will not pursue here. The choice that will be presented here has been deemed distribution regression by their developers. It has its roots on the work of Foresi and Perachi (1995), but it differs from the latter in that the method of distribution regression is shown to be a uniformly consistent estimator and allows to perform inference on the estimated distributions, as will be presented later. It proceeds by the following steps:

- Divide the range of the outcome variables into the numbers of quantiles of the distribution that will be estimated, according to the unconditional distribution of the observed variable.

- For each of the ranges from step 1, create a binary variable that will be zero (or one) for the values above the threshold; this process renders a number of binary variables equal to the number of quantiles chosen in the first step.

- For each binary variable from step two, run a binary regression with the link function of choice fitting the binary variable to the explicative variables of choice. Each of these regressions renders a set of parameters of the dimension of the set of regressors.

- Construct the conditional distribution by obtaining the predicted values from the coefficients estimated in step 3.

This procedure is appealing for a number of reasons, but mainly because of the

flexibility allowed in the specification, as the link function can be chosen from a large set: logistic (logit), probit, cauchy, log-log and other specifications. ChFM argue that many well-known semiparametric transformations of the data correspond to the choice of one of the link functions mentioned above. The choice of link function is not of importance, since the conditional distribution can be approximated to any degree of precision by a rich enough specification of the parameters of the link; given this flexibility, asymptotically, the results obtained from any rich-enough specification ought to be equivalent, which make the estimation less dependent on the model. Once the conditional distribution has been estimated, the covariate distribution F_{Xk} can be estimated by the empirical distribution function. These are all the elements needed to estimate the counterfactual distribution.

2.3 Inference for distribution regression and empirical processes

As mentioned earlier, the asymptotic theory derived by ChFM for the estimators of counterfactual distributions presented in the previous sections is for the entire distribution regarded as a process, that is uniform convergence; the estimators derived are uniformly consistent, under certain assumptions, which will be presented in this section. This is achieved by employing the theory of empirical processes, a field which has brought a number of breakthroughs to long-standing problems which enable this type of results. The distributions estimated are not only uniformly consistent but, by regarding the estimation as an operator and showing that it is Hadamard-differentiable, a concept that will also be explained shortly, it is possible to perform inference based on the bootstrap, to overcome potential problems of non-pivotality and nuisance parameters in the asymptotic distributions. The following paragraphs present a short introduction to the theory of empirical processes used in

ChFM, in order to make the full strength of their results patent.

2.3.1 Empirical processes

The theory of convergence of (probability) measures has been in continuous development ever since Kolmogorov set the mathematical basis for probability based on measure theory, in the 1930s. Based on this foundation, the theory of weak convergence of probability measures was developed by the Russian school of probabilists represented by Skohorod and the American school, with Dudley as its main standard-bearer. The study of empirical processes employing the theory of weak convergence has been undertaken by many researchers, producing many fruitful results. A complete reference for this theory is Billingsley (1968).

We will present the most basic framework required, to state the derivations of ChFM for the case of distribution regression once the former is in place. Given a random sample from a distribution function F on $\mathbb{R}$, the empirical distribution function is defined as $F_n(t) = \frac{1}{n}\sum 1_{(X_i \leq t)}$.

From this perspective, F_n is defined as a real valued function, taking values for each $t \in [0, 1]$. In this setup, since $nF_n(t)$ is binomially distributed with mean nF, the empirical distribution is a pointwise (that is, for each t) unbiased estimator of $F(t)$. By the law of large numbers, it is convergent almost surely, so it is also a pointwise consistent estimator. Finally, by the central limit theorem, it is pointwise asymptotically normal, so that $\sqrt{n}(F_n(t) - F(t)) \rightsquigarrow N(0, F(t)(1 - F(t))$.

The first extension of the law of large numbers to a uniform result is known as the

Glivenko-Cantelli theorem, which refers to the behavior of the Kolmogorov-Smirnov statistic defined as $\sup_t |F_n(t) - F(t)|$. The theorem states that given iid random variables $X_1, \ldots, X_n$, $\sup_t |F_n(t) - F(t)|$ converges as to 0. Likewise, it is possible to think of the extension of the central limit theorem to an uniform version, which is known as Donskers theorem; it states that, given iid random variables as before, with distribution function F, the sequence of empirical processes $\sqrt{n}(F_n - F)$ converges in distribution in the space D, known as Skohorod space, to a tight random element G_F, whose marginal distributions are zero-mean normal with covariance function $E[G_F(t_i)G_F(t_j)] = F(t_i \wedge t_j) - F(t_i)F(t_j)$. The process G_F is known as a F-Brownian Bridge. It is important to realize that this transition to an uniform limit implies the transition from F just as a function, to be seen now as a random function, as a realization of a stochastic process.

Within this very simple example, a major complication in the development of this uniform asymptotic theory was brought to attention by Chibisov, as explained in Ch.18 in Billingsley(1968), regarding the measurability of the random variables whose convergence is the subject of analysis; his counterexample, in which he showed that the empirical distribution function of a group of independently uniformly distributed random variables was not measurable if the space of functions is endowed with the supremum (or uniform convergence) norm as here, so that neither the Glivenko-Cantelli theorem nor the Donsker theorem can be applied to this setup. This complication arises because of the lack of separability of the space on which the functions are defined, which in turn implies that the measures involved in the definition of the empirical process are not tight, even though it is possible to make sense of the limiting process X as a Brownian bridge as the theory requires.

There have been many purported solutions to these measurability complications, for instance, those suggested by Dudley, as explained in Ch.2 of van der Vaart and Wellner (1996), which will from now on be referred to as vdVW, whereby the sigma-fields on which the probabilities are defined are generated by smaller sets of functions, but these generalizations still presented their own problems, as they were suited to handle some issues, but not others; we will not go into this aspect, as it is highly technical and it does not contribute to our main goal.

A more general approach which has proven extremely fruitful is that afforded by the theory of outer convergence, which weakens the requirement of measurability and defines convergence in terms of suprema (and infima) of majorants (and minorants) of measurable functions, not requiring that the functions themselves be measurable. With this generalization, almost all the theory of weak convergence transfers, whereby there are a few places at which one ought to be careful with certain theorems which do not extend, as the Lebesgue theorem of product measures.

The use of outer probability enables to drop the requirement that the random variables and the distributions they determine be measurable in the development of the asymptotic theory. The theory based on this flexible convergence is deemed outer weak convergence of measures, of which vdVW is an extensive account. The outer convergence analogs of the Glivenko-Cantelli theorem and the Donsker theorem read exactly as stated previously, with the only distinction being that measurability is only required of the limiting quantities as opposed to all the variables and distributions involved. This is referred to as asymptotic measurability.

So far, no explicit mention of weak convergence has been made when presenting

the results for the empirical distribution, which was used as the leading example to introduce the GC and Donsker theorems. However, weak convergence permeates the theory, and we will recast the previous results in this more general setup, whereby it will become apparent that the theory developed so far for the empirical distribution, which corresponds to the class of indicator functions as indices of the stochastic process, extends to any class of functions possessing certain properties, which turn it into a GC or Donsker, according to the type of convergence holding.

We will follow the exposition of Ch.1 and 2 of vdVW and Ch.20 and 21 in van der Vart (1998); Kosorok (2008) is also a useful reference. In broad lines, given a metric space (D, d) and a family of probability measures P_n defined on $(D, \mathbb{D})$, where $\mathbb{D}$ is the Borel sigma algebra generated by the open sets of d, the family P_n of measures is said to converge weakly to the measure P (also defined on $(D, \mathbb{D})$), which is written as $P_n / leadsto P$, if and only if, for every bounded continuous function on D (which is the set $C_b(D)$), $\int f dP_n \to \int f dP$

It is required the each of the P_n be well-defined on the Borel algebra $\mathbb{D}$, which can be equivalently stated as requiring that any random variable X_n on D with distribution P_n be measurable; as mentioned earlier, this is a strong requirement, and weaking it to demand for asymptotic measurability will be the choice to overcome the difficulties posted by it.

Given a collection of random variables on the measurable space $(D, \mathbb{D})$ and a collection of real valued functions F defined on this space, the empirical measure P_n induces a map from F to $\mathbb{R}$ given by $f \mapsto G_n f$, where $G_n f$ is the integral of f with respect to G_n, which is defined as $G_n = n^{-\frac{1}{2}} \sum(\delta_(X_i) - P)$, the empirical process.

From the law of large numbers and the central limit stated above, one obtains that for a given f, $P_n f \to Pf$ almost surely and $G_n f \rightsquigarrow N(0, P(f - Pf)^2)$ respectively. In our example using the empirical measure, where D was the real line, the processes is indexed by the space R^∞, and the class F is the class of indicators of half lines, which is the extension to the random function approach of the indicator for each point in time. The uniform version of the law of large numbers for this class of functions F is $||P_n - P||_F$, where the norm is the supremum norm over the class F, and the central limit theorem for the empirical process can be seen as a map into $l^\infty(F)$ if the functions f are uniformly bounded.

The class F is GC if uniform convergence of $||P_n - P||_F$ holds and is Donsker if $G_n = \sqrt{n}(P_n - P) \rightsquigarrow P$ holds. A more precise statement would be to deem the classes PGC and PD respectively, since the weak convergence is with respect to a probability P that is common to all the random variables on which the empirical measure is defined. The adverb universally GC or D is added when the class is of this type for any probability measure P in the probability space.

The tools to determine whether a given class of functions is GC or Donsker are an extension from the methods derived by Vapnik and Chervonenkis in machine learning theory, which use the entropy or integrability conditions to determine the size of the class of functions according to the number of balls required to cover the class. The interested reader is referred to the sources mentioned at the beginning of the section for the particularities.

Another aspect of the theory of empirical processes the extension of the delta method to this setup. This will be of use, as is with the usual delta method, to

obtain the asymptotic distribution of operators and estimators that possess certain differentiability and local approximation properties. The extension of the delta method to this setup is based on the intuitive idea of von Mises calculus, which asserts that a first order approximation of the empirical distribution or functionals thereof, is possible, once the appropriate concept of differentiation is put forth. In the ordinary case of asymptotically linear estimators, once it has been established that the estimator can be approximated by a Taylor or Edgeworth expansion, the linear term will be the leading term asymptotically, and this so-called influence function will be the one determining the inferential behavior of the estimator, as the other terms will vanish in probability at a speed greater than the rate of convergence.

Von Mises Calculus extends this idea to the setup where the estimators are not the usual point estimators, but any estimator allowing for this type of approximation, at least locally. In the case of the empirical process, the foundations of this approach were laid in the pioneering work of Reeds (1976), who showed that the appropriate concept in this setup is called the Hadamard or compact differentiability of the estimator, which in this case is the estimator of a random function, as we mentioned earlier. Once the asymptotic tightness of the limit process has been assumed, the aforementioned work shows that it is both necessary and sufficient to require the estimator to be differentiable on compacts in a certain direction or subset, for the approximation to be possible.

The analogous results as for estimators hold, namely, the delta method is applicable and the limit distribution can be obtained by analyzing the leading linear term and adjusting the distribution of the process. A thorough exposition of the subject can be found in Fernholz(1983). This is the basis for the use of the bootstrap in this

setup, which is shown to be deliver the right asymptotics for Hadamard-differentiable operators, once it has been ascertained that the transformations done to obtain bootstrap samples maintain the uniform convergence of the family of functions on which the estimator is based; in our case, it is the (universal) Donskerness of the class that needs to be preserved. The main reference for this topic is the work of Gine and Zinn (1990), whereby a succinct explanation can be found in vdVW.

2.3.2 Inference theory for counterfactual distributions

We will state and intuitively explain the main results proved by ChFM, following their nomenclature, to facilitate going back to the original reference. There are three main conditions set forth by the authors, which they name condition S, condition SM and condition DR, which in turn imply condition D, which entails the uniform convergence of the process. In the original reference, the conditions are written in a generality that allows them to be applied to different types of estimator, but we will rephrase them in reference to the case of distribution regression that is relevant here.

Condition S requires that the support of the counterfactual population be included in the support of the actual population with respect to which the integration is done. This is a requirement that the estimator be well-defined. When undertaking policy analysis with counterfactuals, actually the condition of balance of the sample is required, which is related to condition S, and can be modified to furnish the required support inclusion. Indeed, the balance condition serves as a means to validate the randomness of the assignment to the treatment; in its absence, a weighting scheme should be employed, for instances propensity scores, in order to restore the

balance of the sample. In the case of a balanced sample where the variable of interest has some finite support, condition S follows.

Condition SM is a sampling condition that guarantees the permanence of the Donskerness of a class of functions. Donskerness, that is, the functional central limit theorem holding for a class of functions indexing a process, is preserved under many operations, for instance, Lipschitzian maps or maps that have a small enough module of continuity. In this case, the authors require that the sampling of each population under observation be the result of a transformation that preserves the universal Donskerness of the class. Under this type of sampling, the limit processes enjoy the necessary tightness properties required with the asymptotic measurability. This condition will be relevant when obtaining the actual limit distributions.

Condition DR is subdivided into three regularity condition. The first one states that the conditional distributions against which the distribution of the covariates will be integrated, comes from the link model described above, or from any other model with a differentiable link. The second one requires that the regions into which the range of the outcome variable of interest was divided be compact or finite subsets of R. And the last one is a condition on the rank of the score of the link of choice, which ensures that the score is invertible.

Under these three conditions, the authors prove condition D follows, that is, that the class of F counterfactuals is a universally Donsker class, that is, estimating the counterfactual distribution with this process provides estimators of the counterfactual distributions that satisfy the functional central limit theorem for empirical processes. The limit processes to which the distributions converge are non-pivotal,

meaning that they depend on unknown nuisance parameters. As such, the limit processes are proved to exist, but are not of much use, since their exact form is not known.

To address this issue, the authors resort to the theory of Hadamard-differentiability, as stated previously. Their results show that the counterfactual operator or estimator as defined by the process of distribution regression is a Hadamard-differentiable and that the process of empirical bootstrap under condition SM above, maintains the Donskerness of the class of counterfactual distributions, as required. In this fashion, consistent estimators of the uniform confidence bands at a level of significance determined can be obtained by the empirical bootstrap. These estimators overcome the issue of non-pivotality of the initial estimators mentioned above. The construction of the confidence bands is explained in detail in the article, and the construction presented there will be followed to obtain the confidence bands in the empirical application to follow. The calculations will be based on Stata code provided by the authors accompanying the publication and duly modified for the purposes of this dissertation.

3. ESTIMATION OF POLICY EFFECTS AND COUNTERFACTUAL DISTRIBUTIONS

In this section, the theoretical tools previously presented will be applied to address the effect of an avian influenza outbreak on the nutritional status of the Mexican population. We will first describe the generalities of the outbreak, its location, duration and the measures taken to control it. We will then describe the information source, namely the database for the Mexican National Expenditure Survey (ENGASTO, for its Spanish acronym) 2012, and finally we will employ all the tools presented in the theoretical chapter to gain some insight into the effects of the outbreak on the food security of the population. This will entail an analysis of the causal structure of the variables deemed relevant, which will be followed by an estimation of the effects using linear regression analysis, comparing the results using all the explanatory variables and only a d-separating set, and finally performing an estimation of the counterfactual distributions of the food security level for different populations of interest, once with all the explanatory variables, and once a d-separating set, as explanatory variables in the estimation.

3.1 General description of the outbreak

On June 13 2012, the Mexican National Service of Agricultural Food Sanitation, Safety and Quality (SENASICA, for its Spanish acronym) issued a statement announcing that an outbreak of avian influenza had been detected in 7 of 111 farms analyzed in the northern side of the state of Jalisco, in the north of Mexico, the main poultry producing zone in the country. The outbreak had been determined to pertain to the H7N3 strain of the virus and had deemed to be highly pathogenic given the

morbidity rates it presented in the animals. For this reason, a quarantine zone of 40 km (16 miles) of diameter had been established according to the guidelines set by the same organization. It was also decided that all the chicken in the affected farms be slaughtered and the buried or incinerated, also according to established procedures.

In follow-up controls, it was determined that the outbreak had spread to four more farms within the quarantined zone, and later to other farms 20 km outside of it, which made it necessary to extend the diameter of the restricted zone. During the next 10 weeks, new contagion episodes were detected within the zone, but none outside of it, and after 12 weeks of the original report, in the last days of August 2012, the outbreak was deemed controlled, after no more cases had been detected neither inside nor outside the quarantined zone for twenty days. A.1 and A.2, the first from the World Organization for Animal Health database on avian influenza outbreaks, shows the location of the outbreak, and the second presents the second quarantined zone, as determined by SENESICA.

In the final report issued by the authority then, it was mentioned that a total of over 22 million birds were slaughtered. It was also ordered that all the surrounding states with any type of domestic chicken production, be it meat or egg, be vaccinated, and until the day of the last report, over 95 million birds had been inocculated throughout the country. To put this numbers in context, we will present a short table in A.3 with statistics of the production of meat and egg in the country, so that the magnitude of the slaughtering in the national production can be better grasped.

It is also worth mentioning that, right after the outbreak was detected, the quotas on poultry meat and egg import from the United States were lifted, in order to

prevent a spike in the prices. Despite these measures, there was a significant increase in prices, which was attributed by authorities mostly to speculation. Once the outbreak was deemed controlled, there were no more outbreaks in the year 2012, but there were many different ones in 2013 and 2014, which, even though it was the same strain of the virus, were different than the one analyzed in this paper because they were not localized within a small radius and they were not controlled with the speed that this outbreak was, despite the fact that the same control measure were applied.

3.2 Data

The information on which the analysis will be based comes from the 2012 Mexican National Expenditure Survey (ENGASTO, for its Spanish acronym). The database consists of the responses to a multilevel survey, whereby some of the information was collected at the individual level and other was collected at the household level. The information-gathering process was handled by Mexican National Institute of Statistics and Geography (INEGI, for its Spanish acronym), which has been heading this process for the last decades. The whole database with the tabulation of the results of the survey is publicly available at the webpage of the aforementioned institute. Many details about how logistics and validation of the survey are available on the same site from which the dataset was downloaded: www.inegi.gob.mx.

The database contains observations for 57756 households throughout Mexico. Given its extent, it has been divided into 7 subsets, according to the main subject of the questions asked. Four of those datasets will be used in this analysis, namely the one regarding the characteristics of the living space of the household (Vivienda Dataset), which in particular contains information regarding the size of

the population where the household lives. The Vivienda Dataset also contains variable representing a socioeconomic index that was constructed on the basis of all the information gathered in the survey.

The second dataset is called Hogar, which contains other variables related to the household. In particular, this dataset contains the responses based upon which the index of food security is constructed for Mexico; this corresponds to a group of 12 questions with binary answers. The index goes from 0 to 12 and records positive answers to the previous questions as equally representing a degree of food insecurity. This dataset also contains information regarding economic setbacks that the household may have experienced, like job loss or loss of a crop or another business setback; these have been condensed in a variable registering the occurrence of a setback. Finally, this dataset contains information to the number of members of the household, and the number of members of the household that are less than 6 years old.

The dataset Persona contains information at the individual household member level. We will use a variable averaging the indicator of whether a household member has a job or not over the members of the household, since our analysis will be undertaken at that level. Finally the fourth dataset called Gasto Ajustado contains information regarding the expenditure of the household; we will use a variable that expresses the annual average expenditure, and two others that indicate the gender of the main provider and his or her educational level.

The survey was conducted over the whole of the 2012 year, which allows for the distinction between a period before and one after the outbreak. In particular, the survey was divided into 26 biweekly periods, which correspond to the 52 weeks of the

year; a variable is available that records the 2 week period on which the household was interviewed (catorcena, for the Spanish translation). This variable will also be used to determine the effect over time.

3.3 Analysis

As mentioned in the introduction, the aim of this work is to analyze the effect of the avian influenza outbreak described in the beginning of this section on the food safety of the Mexican population, as measure by the food security index constructed on the basis of the survey from which our data comes. The analysis will proceed in the following steps: first, we will check the plausibility of performing a quasi-experimental analysis by a simple graphic procedure comparing the distributions of the covariates we deem of interest in explaining the effect. Then, we will run the causal discovery algorithms that will provide us with the description of the conditional independence structure of the variables. Based on the latter, we will construct a d-separating set of variables; recalling the theoretical section, this set of variables contains enough information as to explain the effects on our variable of interest, without the need to run the estimation using the other explanatory variables.

We will test the strength of this assumption, first by running an analysis of the effect of the outbreak on the food security variable using a linear regression model. The effect will be calculated by constructing a binary variable representing the before and after the outbreak; whether the outbreak had an effect on the food security will be determined by whether the coefficient of this variable is statistically significant, as corresponds to the classical treatment effect analysis, where our treatment group is the after-population and the control group is the before-population. We will run this

analysis, once using the set of d-separating variables and once with all the covariates and comparing the results, as mentioned.

We will then proceed with the next hypothesis of this dissertation, namely whether the effect of the outbreak was different once it is analyzed at a deeper tier of the population. This deeper stratum of the population in our case will be whether the household was part of the urban population or rural. Again, the analysis will be done at first through linear regression and also comparing the results using a d-separating set of covariates and once with all the covariates, by seeing whether a urban indicator variable is statistically significant.

One final stage of the analysis will be undertaken to address the question whether the effects are heterogeneous over the population, namely whether the effect of the outbreak was the same for the different quantiles of the population as given by the food security variable. Or working hypothesis is that the outbreak may have been had a harsher effect on the nutrition of the population that was already vulnerable, that is, that which exhibits a higher level of food insecurity, which we consider to be those with a value of 8 or higher (over 12). We will, for this purpose, estimate the counterfactual distributions between different setups and test the aforementioned hypotheses graphically, by means of the uniform confidence bands provided by the methods described in the theoretical section.

3.3.1 Causal graphs

As mentioned in the theoretical chapter, we will run the PC-Lingam algorithm on the variables that were described in the data portion of this chapter. We have set up

the algorithm in such a way that the food security variable is the sink receiving the influence of the other variables. We have also incorporated the previous knowledge that variables such as setback, which denotes the occurrence of a setback in the family, cannot be influenced by any other of the covariates; this is also the case for the numbers of members of the household. We have made one further division, making the education level of the provider of the house and his or her gender precede the expenditure of the household, the economic status and whether the member of the household has a job. It is common to incorporate this a priori knowledge, with seems very plausible to argue, in order to limit the number of DAGs output by the algorithm. In fact, in our case, with these restrictions in place, DAG output is unique, and it takes the rather intricate form of A.4

It is clear that there are many interactions among the variables, but we will concentrate of the direction of the causes that appear more relevant, and try to discern the information in the graph. Apart from the sink given by the food insecurity measure, which was chosen to be that way, almost all the links in this graph have a direction that coincides with the output of the PC algorithm. This is reassuring, since it shows that the dependence information encoded in the first two moments does not conflict with the analysis that incorporates entropy as a more robust measure, as done by the ICA embodied in Lingam. The only exception to this is the link between the number of household members and the gender of the provider of the household, which are dependent on each other, but the direction of the dependence cannot be determined.

We have chosen both variables to be in a higher tier of causation, so that both variables will play an important role when deciding on the separating sets. The

next step in the causal analysis is to check for colliders, that is, variables on which arrows from other variable converge. There are six colliders in the graph, namely the socioeconomic status, which has arrows from many variables pointing to it, but does not have any descendants and does not point to food insecurity anyway. The average expenditure, on the other hand, has four incoming arrows and points into the socioeconomic status, the latter seeming very logic. As the theory of d-separation says, neither this variable nor the former can be a part of a d-separating set. The other four colliders are health, the education level of the provider, the indicator of household members under six, and the job indicator. This leaves three variables that have arrows only emanating from them, which are the setback indicator, the number of household members and the gender of the provider, which form a d-separating set.

As a clarification, the above analysis does not mean that the colliding variables are not relevant or do not have an effect on the food insecurity measure; rather, it says that the influence of that variable will already by taken into consideration, say in a linear regression model, once the variables on the d-separating set are taken as covariates. The effect of the variables not in the d-separating set in then mediated by those in it.

3.3.2 Estimation of average effects

We start by presenting the descriptive statistics of the variables in the analysis in B.1, which presents the mean, the standard deviation of each of the variables. We then perform a graphic analysis of the balance of the data, comparing the distribution of the variable before and after the outbreak. A word of caution regarding the outcome variable is in order: the measurement of food insecurity, given that it comes from answer to a survey, incorporates many sources of imprecision, as detailed by

Gundersen et al (2009). Despite the fact that validation analysis were conducted for this national survey, this remark is nonetheless relevant here, and one should exercise care in the use of the results to guide policy. It is also the case that many observations to construct the analysis are missing, many households interviewed did not respond to seven of the questions used to construct the index, so these observations are not included.

We can observe in A.5 and A.6 containing the histograms of each of the variables that the distributions are so similar that it is even difficult to tell them apart, which provides legitimacy to performing an analysis of the effect of the outbreak from a potential outcomes perspective.

As a first step, we will attempt to validate the results from the causal graphic analysis by running linear regressions with the food security indicator as dependent variable; the first regression will incorporate all covariates available, and the second one will only regress on the d-separating set of variable given by the setback indicator, the number of household members and the gender of the provider, as identified previously. The results of the regression are in B.2 and B.3.

For ease of comparison, the results of both models are presented together with a test of the causal results by another regression in B.4. We can see in the table that the economic status and the job indicator are not statistically significant as explanatory variables for food insecurity. We can also observe that all the variables in the d-separated model are significant, and their coefficients do not change a lot when switching between models. As a final test of the causal analysis results, we run a regression incorporating a variable that was deemed a collider in the analysis. It

was argued that the variable would induce a dependence that would not otherwise be present between the variables that collide on it.

We can observe that, though there are some changes in the coefficients of the model, they are not drastic. The greatest change is observed in the coefficient of the number of household members, which goes from 0.12 to 0.17 when the colliders are incorporated, which points to the fact that the indicator of small children is highly related to it. Attempting an interpretation of the results in general, we can see that the signs of the coefficients are very intuitive; for instance, the coefficient of the health status is negative on the food insecurity, as is that of the education level. Factors that increase the food insecurity of the household are the number of household members, the indicator for a setback and whether the gender of the provider of the household is a female and the existence of small children; these general results do not change with the changes of specification above.

It is now the turn of one of the main questions of the dissertation: what was the effect of the outbreak on the food security of the population? To answer this question, we will calculate the coefficient of an indicator variable marking the before and after the outbreak, as given by the trimester on which the information was gathered. This variable is called wave1 in the analysis, and is zero for the after period. From the results of B.5, we can see that the effect of the outbreak is statistically significant at a 5 % level for both models, once with all covariates and once with the d-separated set. The outbreak increased the index by 0.11 in one model and by 0.13 in the other.

In this same line, we will perform one more estimation, which will incorporate a variable that measure the number of weeks after the outbreak had taken place, to

see how the effect developed over time. This estimation is only sensible for the after-outbreak period, so we will cut the sample accordingly. The results of B.6 shows us that the variable biweek is also statistically significant at the 5 % level, and has a negative coefficient, which means that, as time went by and the number of two-week periods after the outbreak increased, the food security index of the population improved.

Finally, we will perform a linear regression to motivate the next section, which will investigate whether the effects change if the analysis is taken into a deeper stratum of the population, in this case based on the distinction between urban and rural. This distinction is embodied in the variable urban, constructed to indicate whether the household surveyed lived in an area of more than 2500 inhabitants. Regressing the last specification including the indicator for the number of two-week periods, the results of B.7 show one of the few disagreements between the two models. In the first specification, with all the covariates, the variable urban is not statistically significant, whereas in the second, it shows a positive effect on the food security of the urban population after the outbreak.

3.3.3 Estimation of counterfactual distributions

In this last section, we will deal with the question whether the effect that we have found using the standard average treatment effect techniques is heterogeneous over the population. To address this, we will estimate the horizontal distance between the actual or observed distribution of the variable of interest in a certain population and the counterfactual distribution of that variable had the conditions pertaining to the comparison population prevailed. This horizontal distance between distribu-

tions is known as quantile treatment effect. As presented in the theoretical section, the method of distribution regression allows us to estimate the counterfactual distribution and the quantile treatment effect together with uniform confidence bands obtained by bootstrapping.

Following the same structure as with the estimation of average effects, we will first compare the situation before and after the outbreak, to test whether the effect that we identified in the previous section is indeed heterogeneous over the whole population. Following the work of ChFM, we will use the logit, probit, cauchit and loglog link for the calculation of the counterfactual distributions and the quantile treatment effects. For each setup, we will display the observed distribution and the estimated counterfactual, the counterfactual distribution with the inferred confidence bands and the quantile treatment effects with their confidence bands, for each of the specified links.

The first setup is the estimation of the counterfactual distribution of the food security had the population before the outbreak faced the conditions prevailing after it. The second setup refers to the stratification between rural and urban, and the counterfactual estimated is the distribution of the food security after the outbreak had the rural population faced the conditions prevailing for the urban one. As we did with the calculation of the average treatment effects, we will perform the calculations once with the full set of covariates and once with the d-separating set. The results are presented in A.7, A.8, A.9, A.10, A.11, A.12, A.13, A.14, A.15, A.16, A.17, A.18. For ease of visualization, figures for the histograms of both situations have been included too; since the histograms coincide for all the links and using all covariates and only the d-separating, only two figures comparing the actual and coun-

terfactual distributions for both setups have been included. These are A.19 and A.20.

As can be observed from all the figures, the uniform confidence bands include the region around zero, which means that the null hypothesis of there not being an effect on the food security of the population cannot be rejected. Graphically, one can observe that, in the first setup comparing the situation before and after the outbreak, even though the actual distribution and the counterfactual distribution are visually different, statistically this difference is not significant at the level of confidence chosen. This is true for both the estimation using all the covariates and the d-separating set. It is worth mentioning that this lack of significance is not by a large margin, so taking into consideration the remarks about the data made earlier and these other considerations, there may be room to interpret these results beyond the strict statistical realm.

The null hypothesis of there not being an effect cannot be rejected in the second specification comparing the urban and rural either, whereby the same admonition as in the previous case applies. In general, one can observe graphically that the effects present do indeed present a degree of heterogeneity by comparing the estimated and actual distributions and that there are only distinct spikes in the uniform confidence bands that make the statistical rejection necessary.

4. CONCLUSIONS

The aim of this dissertation was to undertake the analysis of the effect of an outbreak of avian influenza that took place in 2012 on the food security of the population of Mexico as measured by the index constructed from the Mexican National Expenditure Survey (ENGASTO). The proposed methods to undertake this analysis were the usual analysis of policy effects by regressions, which allowed us to estimate the average effects of the policy by seeing the outbreak as a treatment or exogenous source of variation.

The analysis by linear regression was complemented by the estimation of counterfactual distributions to check for a heterogeneous reaction to the outbreak over the population. This was implemented first comparing the situation before and after the outbreak, and estimating the counterfactual distribution of the food security index of the population before the outbreak had they faced the conditions prevalent after the outbreak. Another specification compared the response of the rural population with respect to that of the urban population, by again estimating the counterfactual distributions and obtaining the quantile treatment effects.

The estimation of the counterfactual distributions mentioned above was undertaken using the novel method of distribution regression, which permitted the construction of uniform confidence intervals around the estimated counterfactual distributions, as well as for the estimated quantile treatment effects.

It was shown that there were many significant factors affecting the food security

of the population, for instance, the level of education of the provider, which affects the level of food security positively, and the gender of the provider, which showed that households with a female as provider show a lower level of food security. Interestingly enough, whether the members of the household has a job, on average, does not have a significant effect on the index.

The results from the calculation of average effects showed a significant effect of the outbreak, which increased the food insecurity by a margin of 0.11 to 0.13 (on a scale of 12), depending on the specification chosen. The analysis also showed that the effect was diminishing over time, which was determined by a statistically significant coefficient for the variable measuring the number of biweekly periods after the outbreak. The analysis comparing the urban and rural population showed a significant difference, whereby there is a considerable difference even in the sign of the effect for both models employed.

This indeterminacy led us to estimate the counterfactual distributions and to test the existence of effects by checking the statistical significance of quantile treatment effects. As mentioned in the presentation of results, the higher reliability of uniform intervals comes usually at the higher price of them being wider, which in this case forced us to conclude that the effects that are visually present in our estimations are not statistically significant. Despite this fact, we are of the opinion that this complementary analysis permitted a valuable insight into the heterogeneity of the effect on the population, both when analyzing the before and after scenario, and the urban and rural setup.

There have been more outbreaks of the same strain of the virus in Mexico and

in many other countries. The analysis undertaken here could be repeated in this situations, keeping track of other variables that could lead the policy maker in his decision-making process. This type of extension of this work in appealing, and will hopefully be pursued, conditioned on the availability of reliable data, with the hope that the new methods presented here allow for a more informed analysis.

REFERENCES

Billingsley, P., *Convergence of Probability Measures*, Wiley, First Edition, 1968.

Chernozhukov,V.,Fernandez-Val, I., Melly, B., *Inference on Counterfactual Distributions*, Econometrica, Vol. 81, No. 6, 2013, p. 2205-2268.

DiNardo, J., Fortin, N., Lemieux, T., *Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach*, Econometrica, Vol. 64, No.5, 1996, p. 1001-1044.

Fernholz, L., *Von Mises Calculus for Statistical Functionals*, Springer, First Edition, 1983.

Firpo, S., *Efficient Semiparametric Estimation of Quantile Treatment Effects*, Econometrica, Vol. 75, No. 4, 2007, p. 259-276.

Foresi, S., Peracchi, F., *Conditional Distribution of Excess Returns: An Empirical Analysis*, Journal of the American Statistical Association, Vol. 90, 1995, p. 451466.

Gine, E., Zinn, J., *Bootstrapping General Empirical Measure*, Annals of Probability, Vol. 18, No. 2, 1990, p. 851-869.

Gundersen, C., Kreider, B., *Bounding the Effects of Food Insecurity on Children Health Outcomes*, Journal of Health Economics, Vol. 28, 2009, p. 971-983.

Kosorok, M., *Introduction to Empirical Processes and Semiparametric Inference*, Springer, First Edition, 2008.

Pearl, J., *Causality*, Cambridge University Press, Second Edition, 2009.

Reeds, J., *On the Definition of von Mises Functional*, published PhD Dissertation, Harvard University, 1976.

Rubin, D., *Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies*, Journal of Educational Psychology, 66, 1974, p. 688-701.

Rubin, D., *On the Application of Probability Theory to Agricultural Experiments, Comment: Neyman and Causal Inference in Experiments and Observational Studies*, Statistical Science, Vol. 5, No. 4, 1990, p. 472-480.

SENASICA, Historical Information on Avian Flu Outbreak (in Spanish), available online under http://www.senasica.gob.mx/?id=5368, last accessed on March 3, 2015.

Spirtes, P., Glymour, C., Scheines, R., *Causation, Prediction and Search*, MIT Press, Second Edition, 2000.

van der Vaart, A., Wellner, J., *Weak Convergence of Empirical Processes: With Applications to Statistics*, Springer, First Edition, 1996.

van der Vaart, A., *Asymptotic Statistics*, CUP, First Edition, 1998.

World Organisation for Animal Health (OIE), World Animal Health Information Database (WAHID) available online under http://www.oie.int/wahis_2/public/wahid.php/Wahidhome/Home, last accesed on March 3, 2015.

APPENDIX A

FIGURES

Figure A.1: World Organization for Animal Health (OIE) map of influenza outbreak within July and November 2012. With the kind authorisation of the World Organisation for Animal Health [OIE] www.oie.int. Figure extracted from OIE WAHID at 2:05 pm on March 3, 2015 on the OIE website.

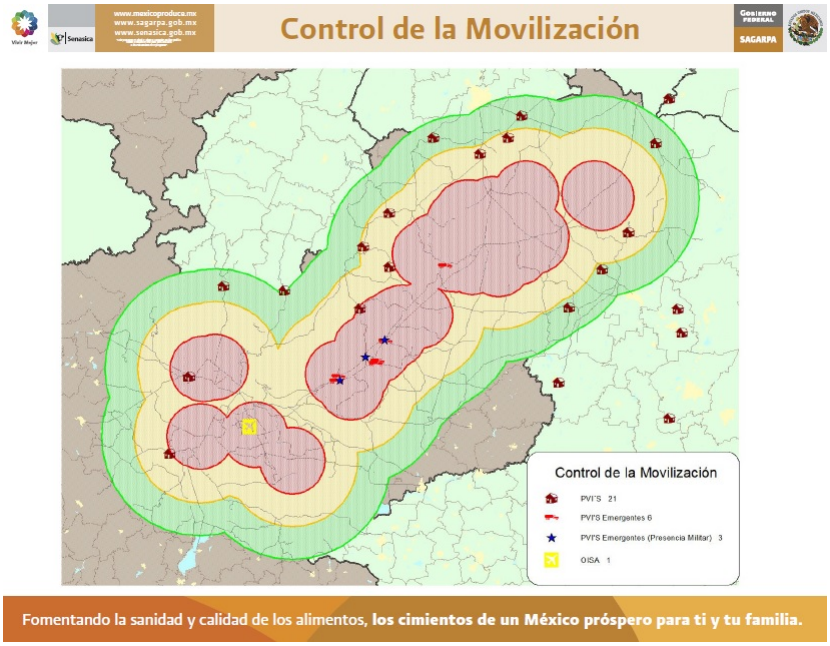

Figure A.2: Close-up map of quarantined zone. Source: www.senasica.gob.mx

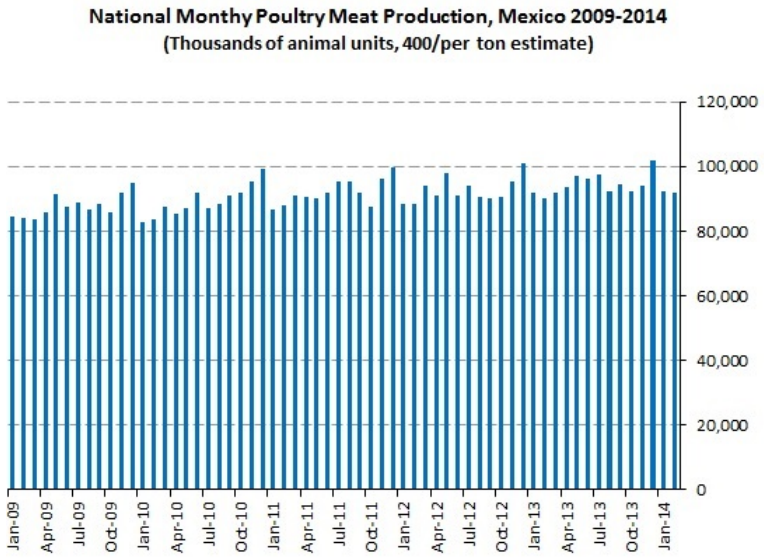

Figure A.3: Mexican National Poultry Production. Source: Author

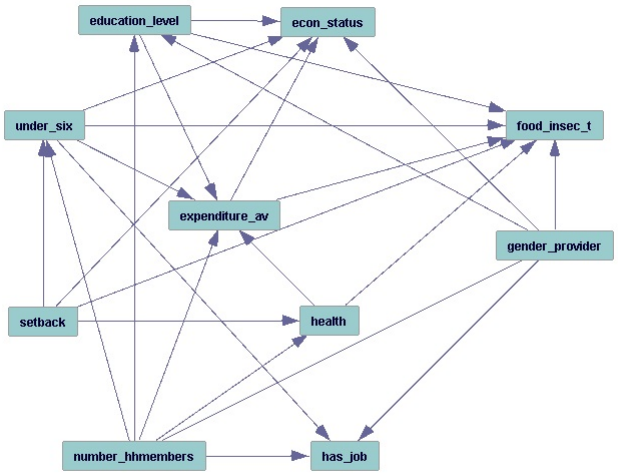

Figure A.4: Directed Acyclic Graph output by the PC-Lingam algorithm by the TETRAD software

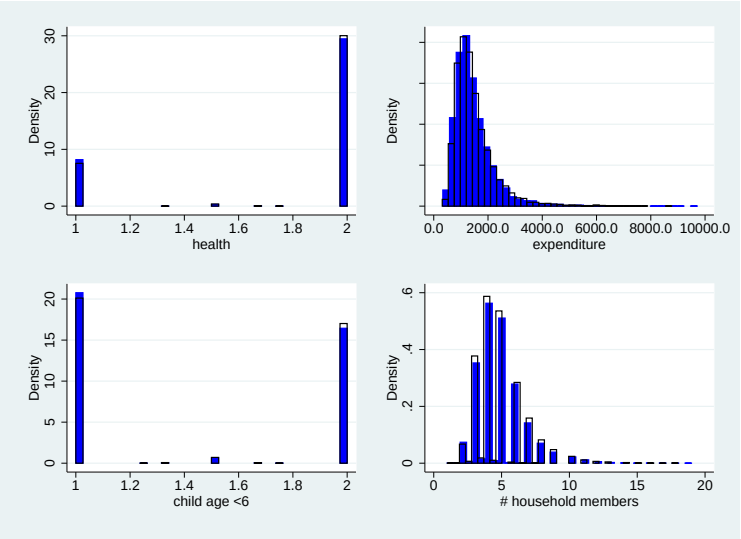

Figure A.5: Graphic balance test (a)

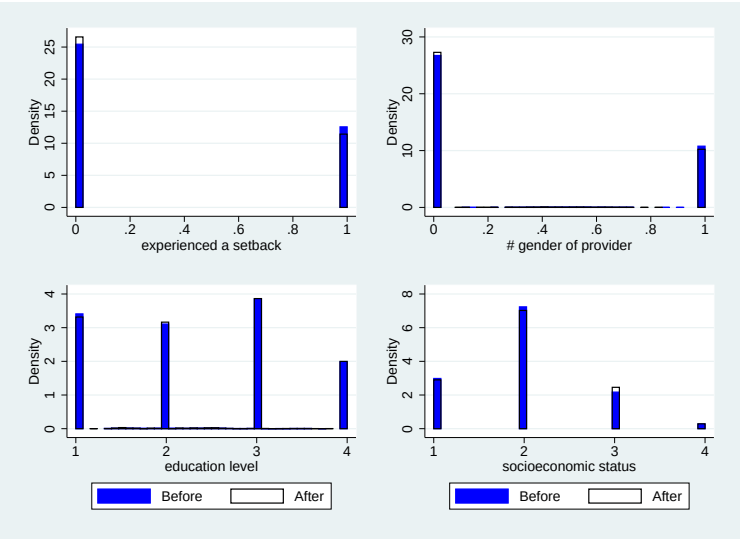

Figure A.6: Graphic balance test (b)

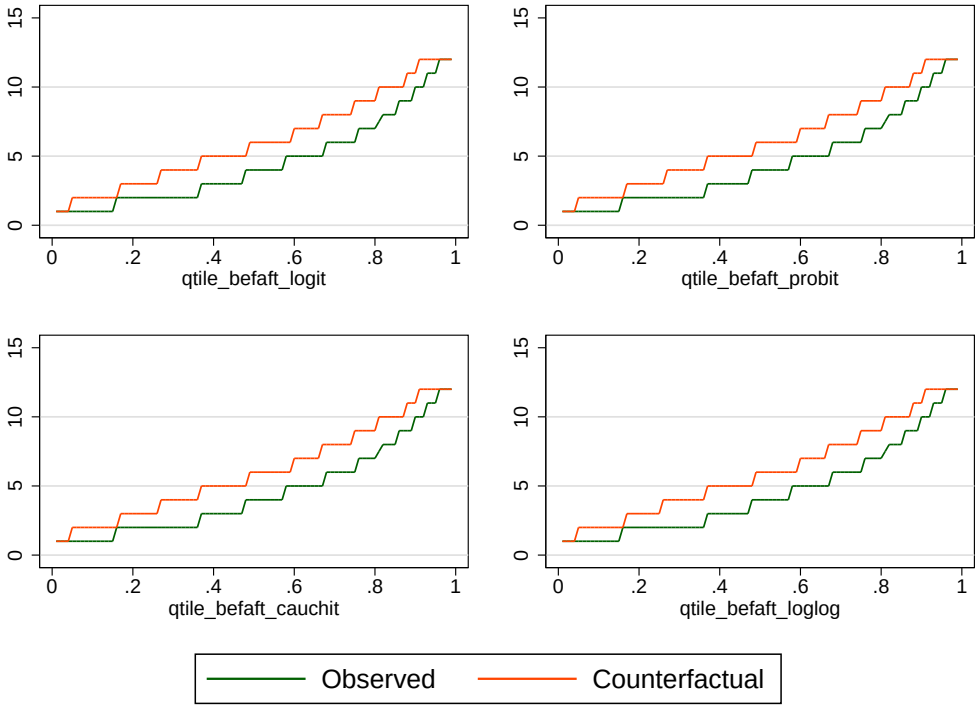

Figure A.7: Observed vs counterfactual distributions with all covariates for before and after outbreak comparison

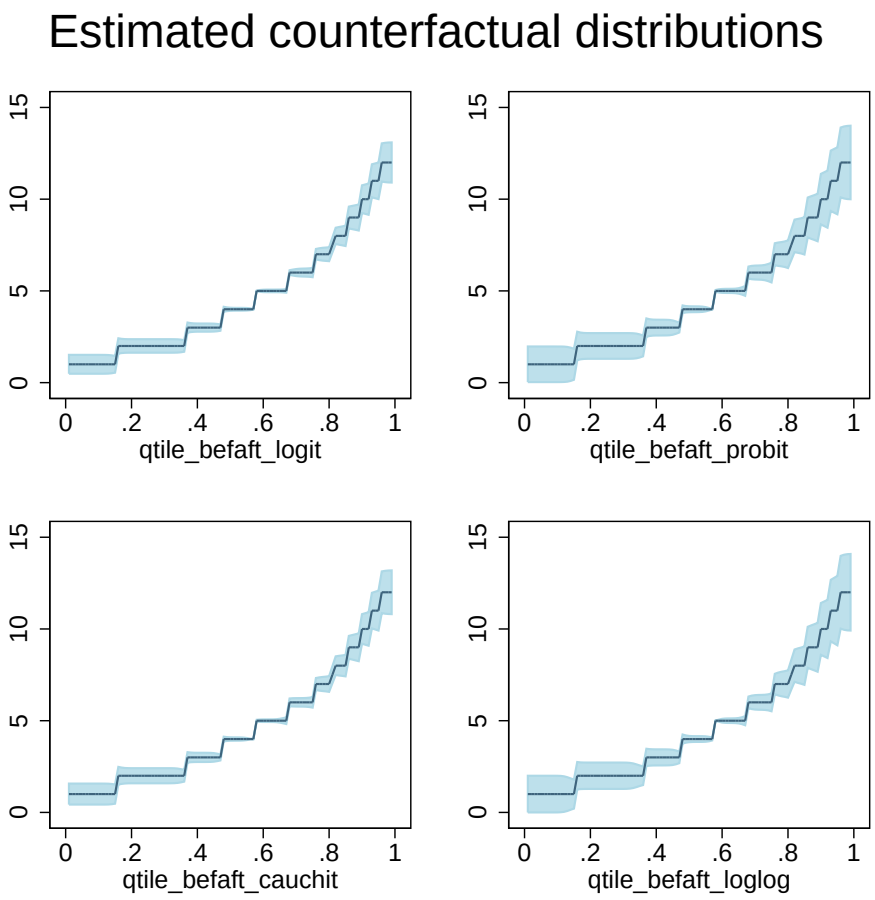

Figure A.8: Counterfactual distributions with inferred confidence bands with all covariates for before and after outbreak comparison

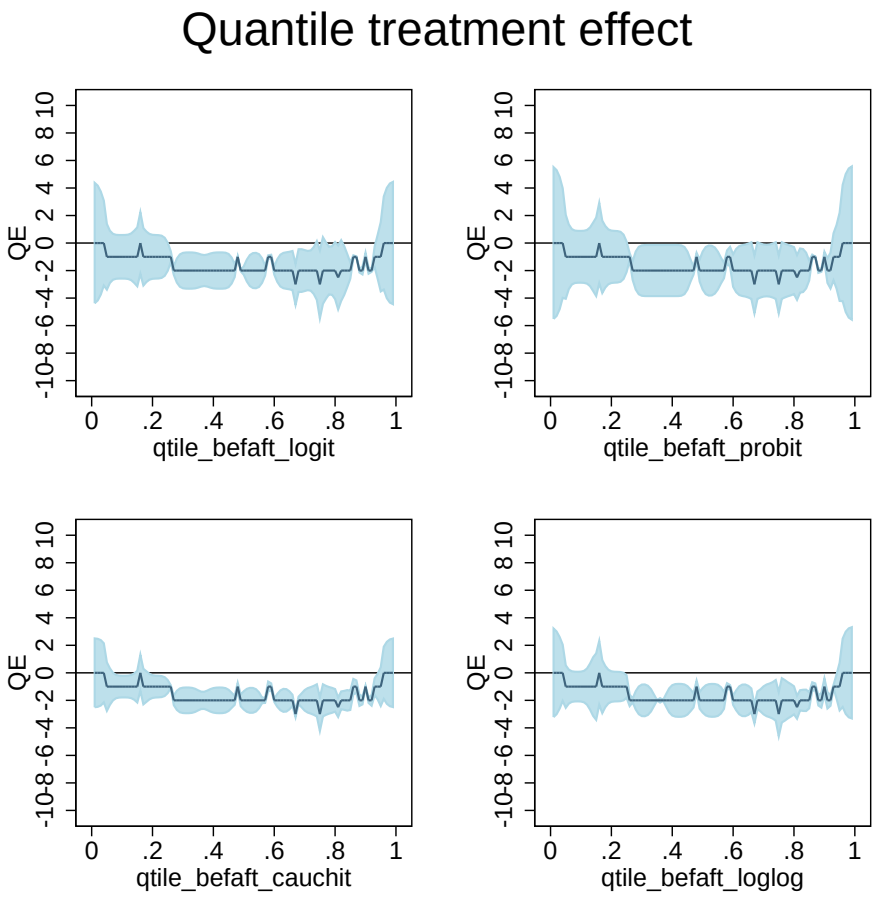

Figure A.9: Quantile treatment effects with confidence bands with all covariates for before and after outbreak comparison

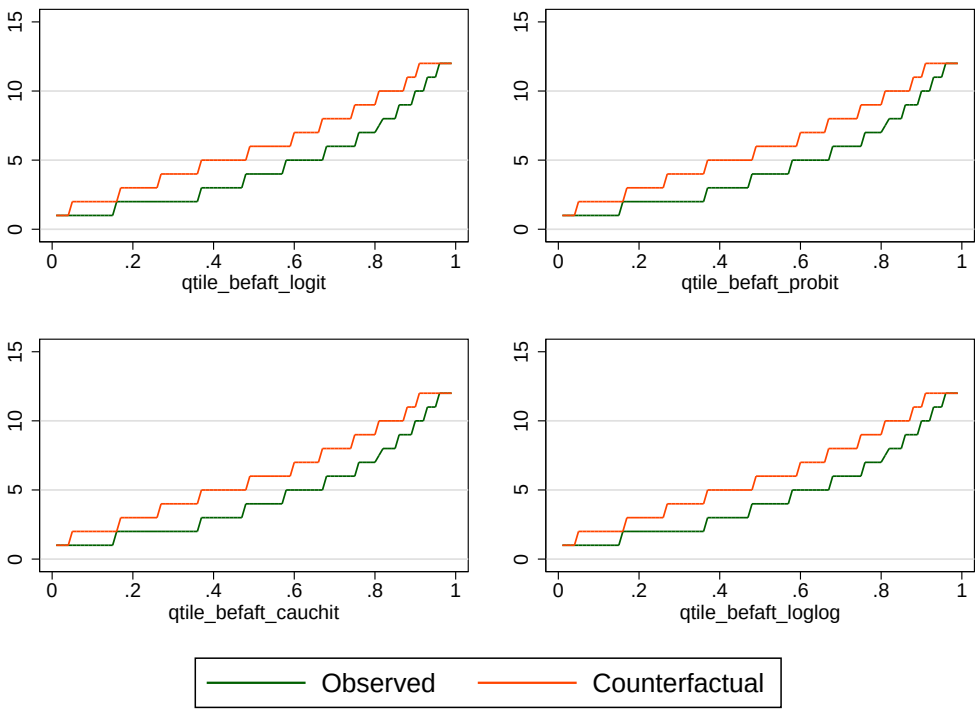

Figure A.10: Observed vs counterfactual distributions with d-separating set for before and after outbreak comparison

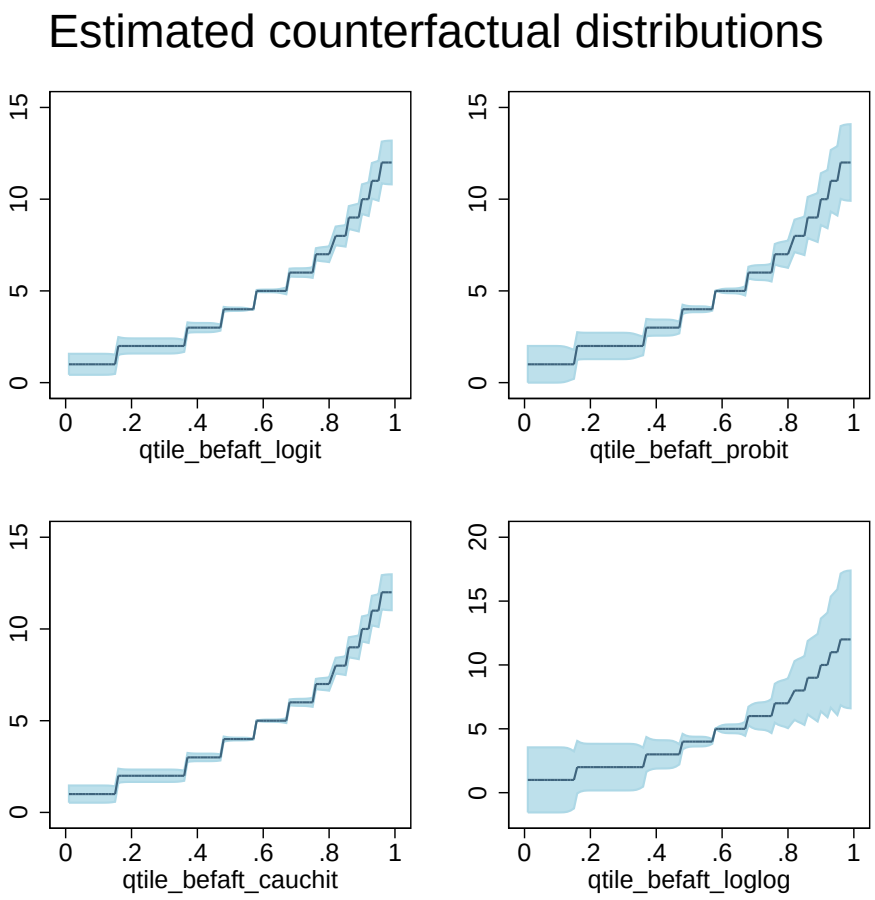

Figure A.11: Counterfactual distributions with inferred confidence bands with d-separating set for before and after outbreak comparison

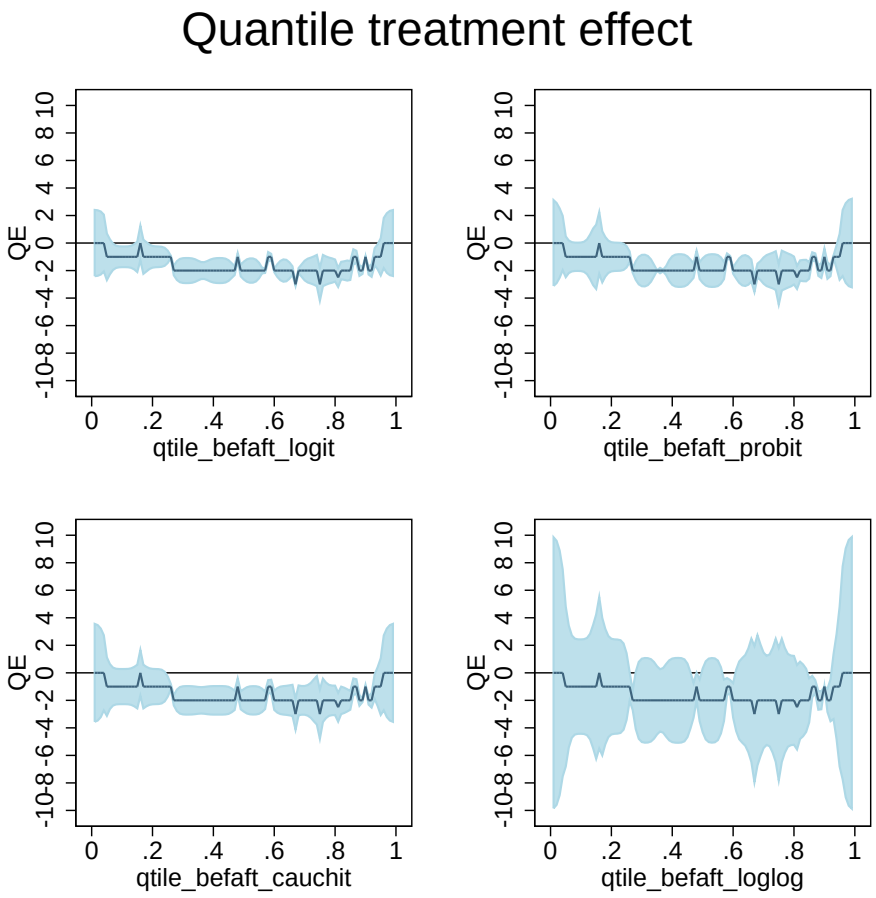

Figure A.12: Quantile treatment effects with confidence bands with d-separating set for before and after outbreak comparison

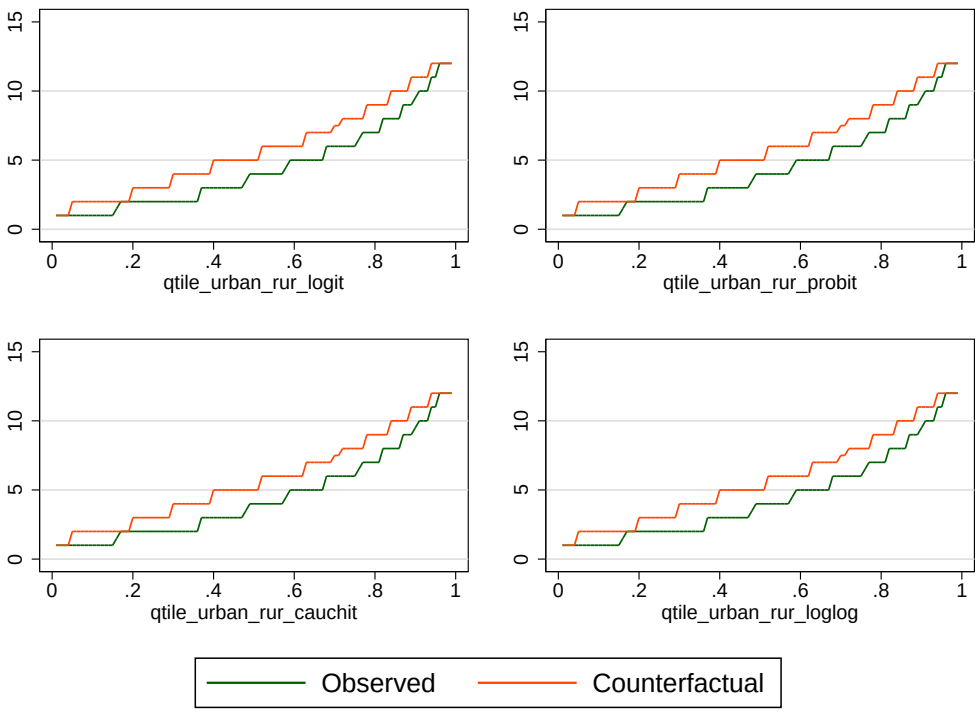

Figure A.13: Observed vs counterfactual distributions with all covariates for urban-rural comparison

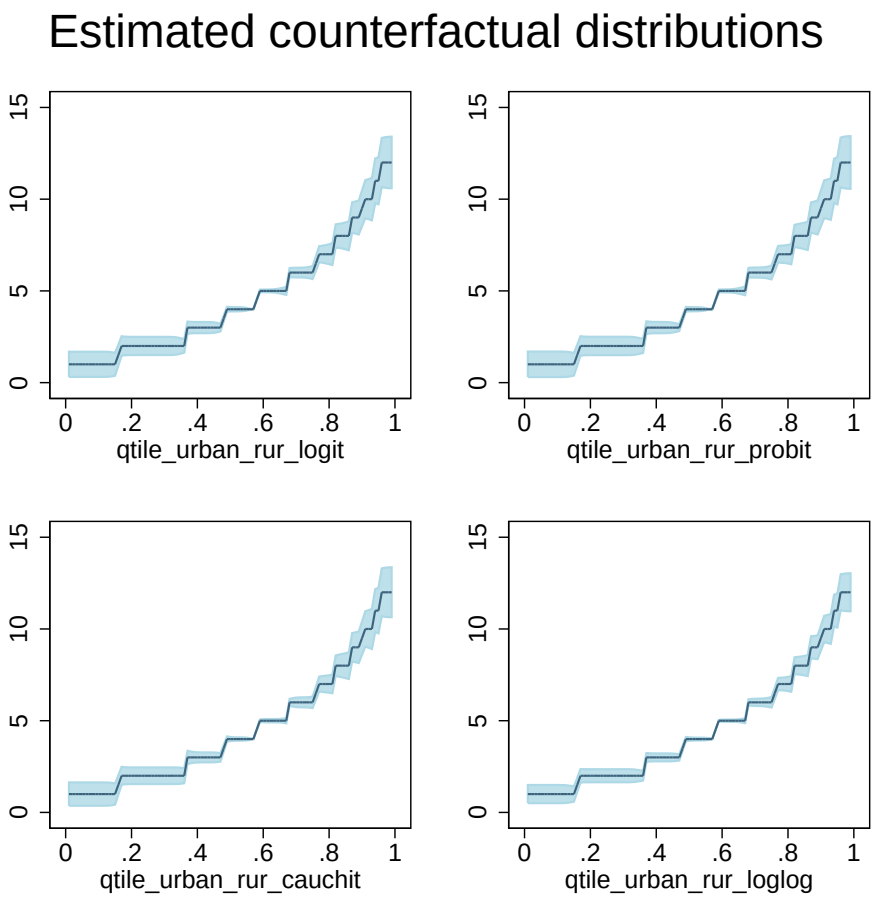

Figure A.14: Counterfactual distributions with inferred confidence bands with all covariates for urban-rural comparison

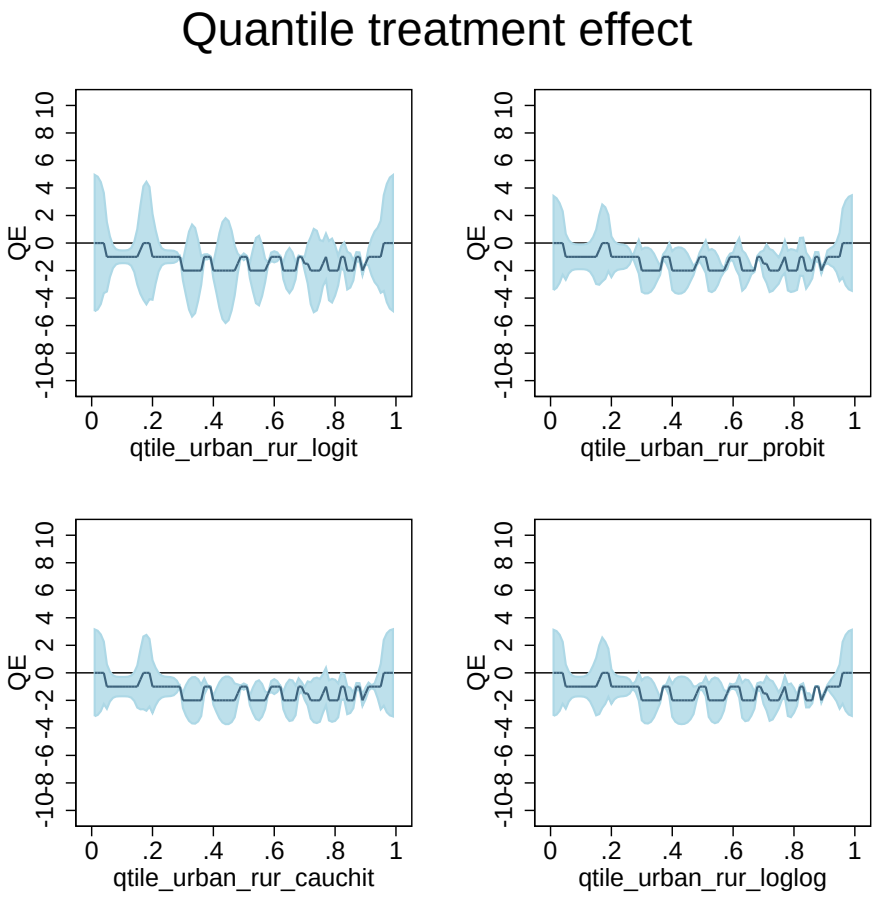

Figure A.15: Quantile treatment effects with confidence bands with all covariates for urban-rural comparison

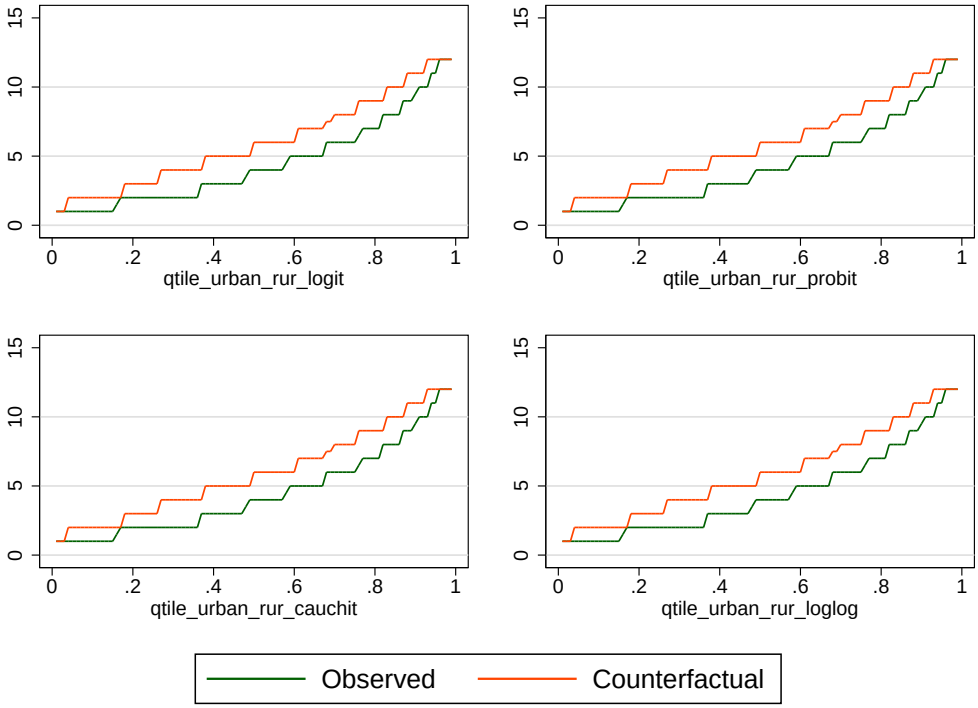

Figure A.16: Observed vs counterfactual distributions with d-separating set for urban-rural comparison

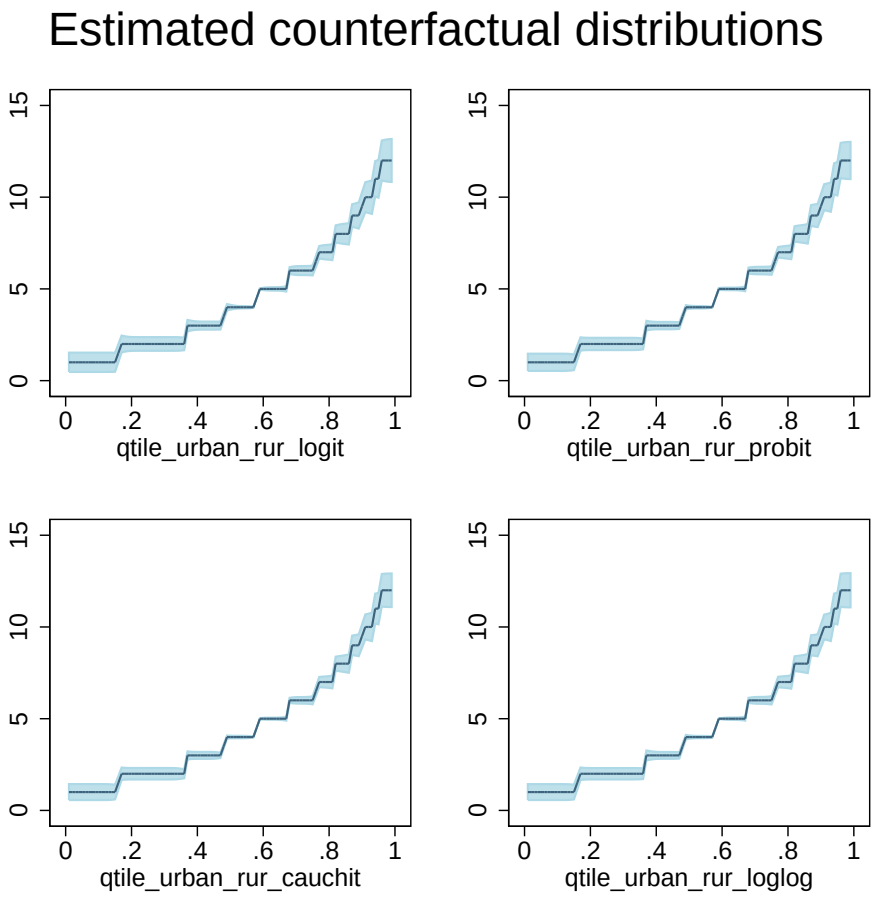

Figure A.17: Counterfactual distributions with inferred confidence bands with d-separating set for urban-rural comparison

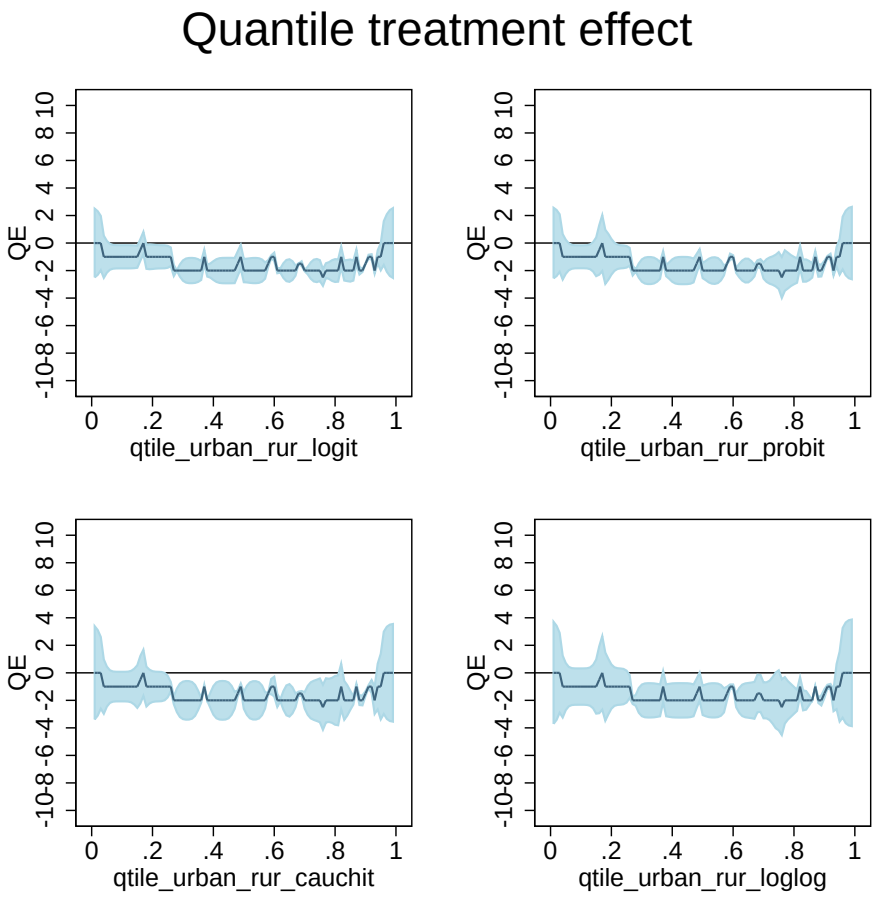

Figure A.18: Quantile treatment effects with confidence bands with d-separating set for urban-rural comparison

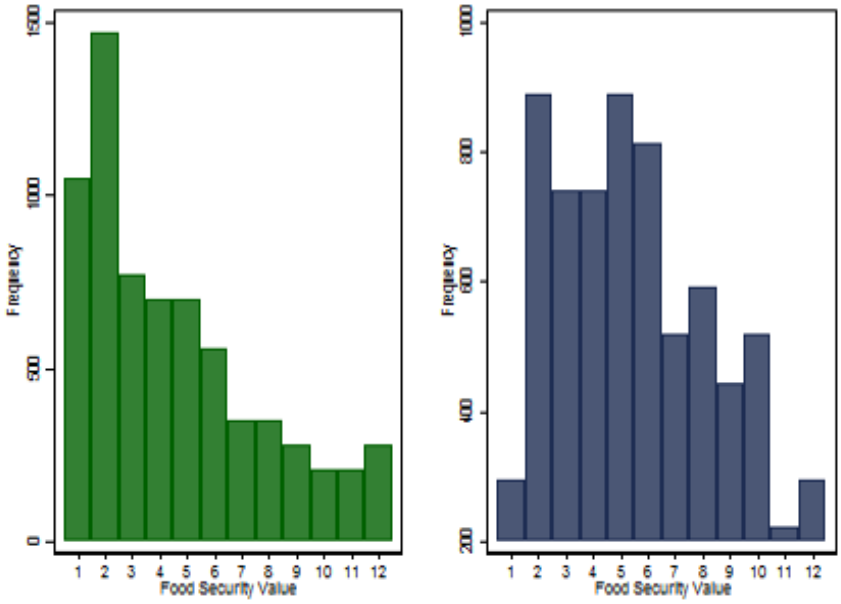

Figure A.19: Histogram for observed vs counterfactual distributions for before and after outbreak comparison

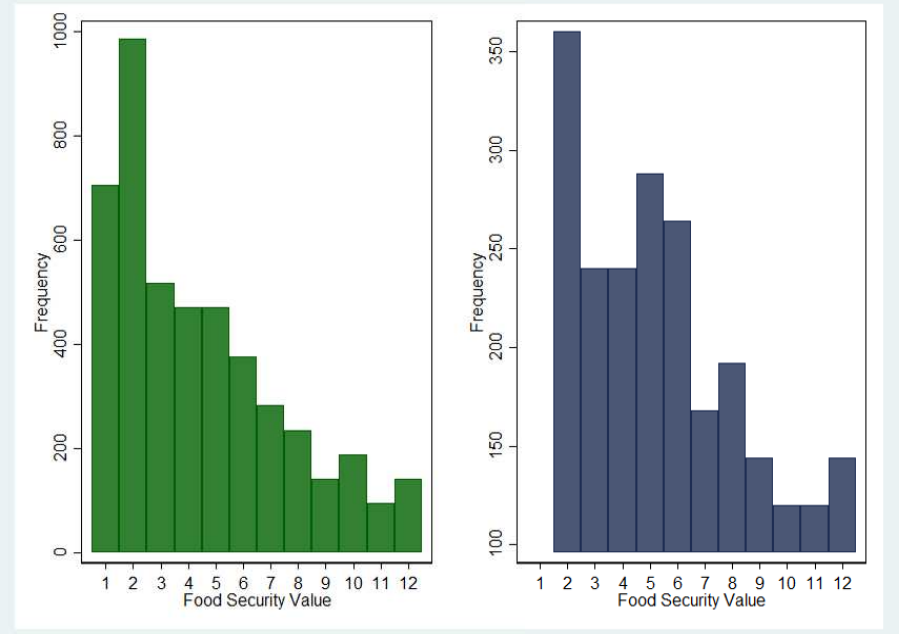

Figure A.20: Histogram for observed vs counterfactual distributions for urban and rural after outbreak comparison

APPENDIX B

TABLES

Table B.1: Descriptive statistics of the variables

variable	**mean**	**sd**
expenditure average (pesos)	1486.96	766.49
under six	1.45	0.49
number members	4.87	1.80
economic status	1.99	0.71
education level	2.36	1.04
gender provider	0.28	0.45
has job	1.40	0.27
setback	0.32	0.46
health	1.79	0.41
food insecurity	4.66	3.25

Table B.2: Regression of food security on all covariates

variable	**Coefficient and Std Error**	**p value**
health	-0.70	0.000
	(0.065)	
number hhmembers	0.12	0.000
	(0.016)	
econ status	-0.05	0.272
	(0.041)	
setback	0.85	0.000
	(0.057)	
expenditure av	-0.00	0.000
	(0.000)	
education level	-0.24	0.000
	(0.028)	
gender provider	0.57	0.000
	(0.060)	
has job	0.14	0.168
	(0.099)	
under six	0.39	0.000
	(0.057)	
cons	5.57	0.000
	(0.240)	

Table B.3: Regression of food security on d-separating set

variable	**Coefficient and Std Error**	**p value**
number hhmembers	0.10	0.000
	(0.015)	
gender provider	0.62	0.000
	(0.060)	
setback	0.88	0.000
	(0.058)	
cons	3.75	0.000
	(0.083)	

Table B.4: Comparison of regression with all covariates, d-separated and colliders

variable	All Covariates		D-separation		Collider	
	b/se	p	b/se	p	b/se	p
health	-0.70	0.000				
	(0.065)					
# hhmembers	0.12	0.000	0.10	0.000	0.17	0.000
	(0.016)		(0.015)		(0.016)	
econ status	-0.05	0.272				
	(0.041)					
setback	0.85	0.000	0.88	0.000	0.89	0.000
	(0.057)		(0.058)		(0.057)	
expenditure av	-0.00	0.000			-0.00	0.000
	(0.000)				(0.000)	
education level	-0.24	0.000				
	(0.028)					
gender provider	0.57	0.000	0.62	0.000	0.62	0.000
	(0.060)		(0.060)		(0.060)	
has job	0.14	0.168				
	(0.099)					
under six	0.39	0.000			0.46	0.000
	(0.057)				(0.056)	
cons	5.57	0.000	3.75	0.000	3.62	0.000
	(0.240)		(0.083)		(0.133)	

Table B.5: Calculation of average effect of outbreak on food security

variable	All Covariates		D-separation	
	b/se	p	b/se	p
health	-0.70	0.000		
	(0.065)			
number hhmembers	0.12	0.000	0.10	0.000
	(0.016)		(0.015)	
setback	0.84	0.000	0.87	0.000
	(0.057)		(0.058)	
expenditure av	-0.00	0.000		
	(0.000)			
education level	-0.24	0.000		
	(0.027)			
gender provider	0.56	0.000	0.61	0.000
	(0.060)		(0.060)	
under six	0.40	0.000		
	(0.056)			
wave1	0.11	0.036	0.13	0.020
	(0.053)		(0.054)	
cons	5.62	0.000	3.68	0.000
	(0.201)		(0.087)	

Table B.6: Calculation of average effect of outbreak on food security over time

variable	All Covariates		D-separation	
	b/se	p	b/se	p
health	-0.73	0.000		
	(0.094)			
number hhmembers	0.07	0.001	0.06	0.006
	(0.023)		(0.021)	
setback	0.81	0.000	0.82	0.000
	(0.082)		(0.083)	
expenditure av	-0.00	0.000		
	(0.000)			
education level	-0.30	0.000		
	(0.039)			
gender provider	0.43	0.000	0.51	0.000
	(0.086)		(0.087)	
under six	0.38	0.000		
	(0.080)			
biweek	-0.02	0.031	-0.02	0.033
	(0.010)		(0.010)	
cons	6.23	0.000	4.07	0.000
	(0.294)		(0.137)	

Table B.7: Calculation of average effect of outbreak on food security over time and urban-rural distinction

variable	All Covariates		D-separation	
	b/se	p	b/se	p
health	-0.73	0.000		
	(0.094)			
number hhmembers	0.08	0.001	0.05	0.010
	(0.023)		(0.021)	
setback	0.80	0.000	0.84	0.000
	(0.082)		(0.084)	
expenditure av	-0.00	0.000		
	(0.000)			
education level	-0.31	0.000		
	(0.039)			
gender provider	0.42	0.000	0.53	0.000
	(0.086)		(0.087)	
under six	0.38	0.000		
	(0.080)			
urban	0.07	0.408	-0.25	0.002
	(0.083)		(0.082)	
cons	6.05	0.000	4.09	0.000
	(0.286)		(0.130)	