Financial Incentives and Behavior

Four Essays in Applied Health and Labor Economics

Inaugural-Dissertation
zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2011

vorgelegt von
Helmut Farbmacher

Referent:	Prof. Dr. Joachim Winter
Korreferent:	Prof. Dr. Florian Heiss
Promotionsabschlussberatung:	16. Mai 2012

Acknowledgements

I am very grateful to my supervisor, Joachim Winter. His advices and suggestions have been invaluable for the completion of my thesis. Above all, his ability to see the big picture has greatly enriched my dissertation. I also like to thank my co-supervisor, Florian Heiss, who has formed my understanding of computational econometrics. Further, I want to thank Helmut Rainer for agreeing to serve on my thesis committee and Rainer Winkelmann for constructive comments on parts of this dissertation. Amelie Wuppermann, who has co-authored the second chapter of my dissertation, also deserves special thanks.

Additionally, I would like to thank my current and former colleagues at the Munich Center for the Economics of Aging (MEA) and the Seminar for Empirical Research at the University of Munich. Specifically, I would like to mention Tabea Bucher-Koenen, Lucia Maier, Bettina Siflinger, Martin Spindler, Gregor Tannhof and Stefan Vetter.

Parts of the analyses have been done during visits to the PMV forschungsgruppe at the University Hospital of Cologne. I am really grateful to Ingrid Schubert and Peter Ihle for their hospitality and their support. I would like to thank the AOK Hesse and KV Hesse for allowing me to analyze their claims data set. Throughout my dissertation I got financial support from the Munich Center of Health Sciences, which is under the supervision of Reiner Leidl. I am also grateful for his constant support.

To Lisa, Luis and Carolin

Contents

List of Tables

List of Figures

Preface

Individuals face a variety of financial incentives, which are valuable instruments to allocate resources and to steer behavior. Governments use them in a variety of contexts like, for instance, tobacco taxes or subsidies for new technologies. Another example is the prospect of higher earnings due to higher education, which provides an important incentive in our society to advance the long-term investments in education. Empirical evaluations are an important way to assess the extent to which individuals react to incentives.

This dissertation consists of four self-contained chapters. The first two chapters analyze the 2004 health care reform in Germany. An important aim of the reform was to strengthen cost consciousness and personal responsibility of the insured. The focus in the first two chapters is on a particular element of the reform, namely a per-quarter fee for doctor visits, and the question how this treatment affects individuals' decisions to visit a doctor. The time dimension of the fee implies that individuals sometimes do not have to take the fee into account when making decisions. While the treatment status is usually based on characteristics that can easily be observed (like age and gender), in this case, it follows implicitly from the design of the treatment. In this application, an individual's treatment status actually depends on previous and future demand for health care, and this complicates the evaluation of the fee. In the first chapter, I exploit the fact that the treatment status depends on previous health care demand, to form a unique identification strategy. In the second chapter Amelie Wuppermann and I develop an econometric model which takes into account that the perceived treatment status depends on future health care demand. The results suggest that certain groups are ex-ante or ex-post unaffected by the fee. A narrower definition of the group that

is actually affected makes it possible to reveal the true effect of the fee. The focus of the last two chapters is on the enhancement of econometric methods. Examples from health and labor economics are given for illustrative purposes.

In chapter 1, I use the German Socio-Economic Panel to estimate the effects of the 2004 health care reform. Among other things, the reform imposes a fee of €10 for the first visit to a doctor in each quarter of the year. Patients who have already paid the fee are therefore exempt for the rest of the calendar quarter implying that the treatment status depends on previous health care demand. Exploiting random variation of the treatment level over different interview days, I find a substantial effect of the new fee on the probability of visiting a physician. In addition, the identification strategy makes it possible to disentangle this effect from the influences of the contemporaneous increase of co-payments for prescription drugs. I verify the crucial assumptions of my approach using a claims data set from the largest German sickness fund. Overall, the probability of visiting a physician decreased by around 5 percentage points. Due to my identification strategy, I can attribute at least half of this effect to the per-quarter fee for doctor visits.

In chapter 2, Amelie Wuppermann and I revisit the analysis of the reform in spirit of the literature about nonlinear price schedules. We provide empirical evidence of heterogeneous reactions that are in line with theoretical considerations. Using insurance claims data from the largest German sickness fund we find that some individuals strongly react to the new price schedule while there is a group of individuals that does not react at all. This is the group with the worst health in which some individuals may know ex-ante that they cannot avoid the fee. Following van Kleef *et al.* (2009) we suggest a further reform of the system that may help to also increase cost consciousness among individuals in bad health while possibly even decreasing the financial burden for these individuals.

In chapter 3, I extend the literature on hurdle models, which are frequently used to model count data. Recent developments in the count data literature make it possible to relax commonly imposed assumptions of these models. Based on these findings, I develop two extensions of hurdle models which make popular specifications more flexible. Both extensions nest the models that have been estimated previously and

they can thus be tested by appropriate parametric restrictions. An example from health economics illustrates the relevance of both model extensions.

In chapter 4, I employ the new variance estimator for generalized empirical likelihood that has recently been proposed by Newey and Windmeijer (2009) to address the problem that the usual variance estimator understates the true variance. In Monte Carlo examples they show that t-statistics based on the new variance estimator have nearly correct size. I replicate their Monte Carlo simulations and additionally report results for a wider range of the simulation parameters. Moreover, my simulation results suggest that two-stage least squares estimates are poor starting values for the continuous updating estimator, especially when the sample size is small and/or the identification is weak. Finally, I use the continuous updating estimator to assess the private returns to education using a well-known data set, and additionally report the many weak instruments standard errors of Newey and Windmeijer (2009).

Chapter 1

Quarterly co-payments, demand for health care and response behavior - Evidence from survey and claims data

1.1 Introduction

Insurance firms try to implement incentives to avoid excessive claims. This is particularly important in health insurance markets because some therapies depend on patient choice. The first visit to a doctor for a new illness, for instance, is solely a patient's decision. Here co-payments could be an appropriate instrument to reduce moral hazard. The introduction and increase of co-payments have been important instruments of past health care reforms in the German statutory health insurance. There are, for instance, co-payments for drugs, hospitalization and doctor visits. These instruments have a direct fiscal effect because the insurer covers a lower amount. In addition, there might be a reduction in the demand for health care services because the insured avoid excessive use. Such an inhibiting effect on utilization was also a professed goal of the co-payment for doctor visits which was introduced in Germany in 2004. This study exploits random variation in the day of the interview of a survey to reveal the causal effect of the new fee. Accounting for the structure of the data, there is a significant decline in the probability of visiting a doctor. To verify the essential

assumptions of my approach, I imitate a survey with two randomly assigned interview days using claims data from the "Allgemeine Ortskrankenkasse" (AOK), which is the largest sickness fund in Germany.

According to the OECD (2008), around 90% of the German population are covered by statutory health insurance (SHI). The regulation of SHI is heavily influenced by governmental decisions. One example is the implementation of a broad health care reform in 2004 which tried to strengthen cost consciousness and personal responsibility by increasing co-payments. An important part of this reform was the introduction of co-payments for doctor visits. Since 2004, most SHI-insured adults have had to pay €10 for the first visit to a doctor in a calendar quarter. Children and teenagers up to the age of 18 are exempt from co-payments. Moreover, there are also exemption rules for adults. They can apply for an exemption by paying one or two percent of their income in advance. Alternatively, they can choose a gate-keeping model. In this case they often have to pay only €10 a year but must visit a general practitioner (GP) first. When more specialised care is required, the patients receive a referral from this GP.

The €10 fee also covers additional doctor visits within a calendar quarter. So it is a "per-quarter" fee, which is independent of the volume of services rendered in connection with this or later visits within a quarter. This characteristic distinguishes the co-payment from "per-visit" fees. The effects of a per-visit co-payment have been analyzed in several studies (Roemer *et al.*, 1975; Jung, 1998; van de Voorde *et al.*, 2001). For instance, Jung (1998) investigated the effects of implementing such a fee in Korea. He found an remarkable decrease in the number of doctor visits and in the probability of seeking medical care. The effects of a per-quarter co-payment, however, should be different because this fee is not intended to affect all parts of the distribution. It creates a new incentive to avoid the first visit to a doctor in a quarter. However, in contrast to a per-visit fee, it generates no incentives to reduce the number of doctor visits within a quarter once the fee is paid.

Additionally, co-payments for prescription drugs have been increased at the same time as the introduction of the €10 fee and this complicates the evaluation of the fee. Prior to the reform, patients had to pay €4 for small, €4.50 for medium and €5 for large quantities of drugs. Since 2004 it has been a function of the retail price and the

patient has had to bear 10% of the drug price. The co-payment amounts at least to €5 and at most to €10. The effects of increasing co-payments for prescription drugs on the demand for doctor visits were extensively investigated by Winkelmann (2004a, 2004b, 2006). He analyzed the influence of an earlier health care reform implemented in 1997. The most radical element of this reform was the increase of co-payments for prescription drugs (Winkelmann 2004a). All three studies found a link between the propensity to visit a doctor and co-payments for prescription drugs. Therefore, the health care reform of 2004 could affect the behavior of health care consumers through both the increased prescription fees and the introduction of co-payments for doctor visits. This study, however, introduces a method to disentangle these two effects and to uncover the impact of the co-payment for doctor visits.

There are two studies dealing with the introduction of the €10 fee. Both are based on the German Socio-Economic Panel (GSOEP). Augurzky *et al.* (2006) tried to assess the effect of the reform on the probability of seeing a physician using a differences-in-differences approach. They compared statutory health insured participants with privately insured persons, and youths, because the latter two groups are exempt from the fee. Schreyögg and Grabka (2010) applied a similar estimation strategy. Furthermore, they used a zero-inflated negative binomial regression and a negative binomial hurdle model to directly model the number of doctor visits. Both studies concluded that the co-payment for doctor visits had failed to reduce the demand for doctor visits and argued that this ineffectiveness stems from the fact that it is a *per-quarter* fee. The present study, however, reveals that this characteristic does not make the fee ineffective; rather, it is the reason why the effect cannot be observed in the GSOEP using simple differences-in-differences approaches. In addition to a simple comparison of physician visits over time between privately and statutorily insured individuals, this study uses a second natural experiment that exploits exogenous variation in the day of the interview. This allows me to disentangle the impact of the per-quarter fee from the effects of other parts of the reform. Using this approach, I show that the reform as a whole decreases the probability of visiting a physician by 5 percentage points. The per-quarter fee causes at least half of this effect. To put things in perspective, Winkelmann (2004a, 2004b, 2006) has already shown that an increase in prescription

fees indirectly affects the demand for doctor visits, so it would come as a surprise if fees for doctor visits had no direct effect on it.

This chapter is organized as follows. The next section describes the second natural experiment which identifies the causal effect of the new fee. Section 3 explains the data sets used in this analysis and the estimation strategies. Section 4 shows that the co-payment alters the observable behavior in the survey data in a special manner. The effect of the new fee can only be observed once the model accounts for the structure of the data. Section 5 concludes.

1.2 Identification strategy

The GSOEP is an annual survey started in 1984 which, among other things, includes a question about the number of visits to a doctor in the last three months before the interview.[1] Thus the observed three-month period depends on the day of the interview. The interviews are conducted every day from January to October. This variation can be used to identify the causal effect of the new fee if, depending on the day of the interview, the participants are differentially affected by the fee. As already mentioned, a special characteristic of the fee is that it must only be paid at the first visit in a quarter. This characteristic makes it possible to identify random samples of the SHI-insured population that are differently affected by the per-quarter fee. The following example is to show that the probability of having to pay the €10 fee and thus the treatment level depends on the day of the interview.

By way of illustration, Figure 1.1 shows the reporting period for an interview conducted at August 15th. The reporting period can be separated into two equal periods - one period before and one period after the end of the calendar quarter (p_2 and p_3 in Figure 1.1). Period 1, on the other hand, is the unobserved part of the previous calendar quarter. Since period 3 starts at the beginning of a new calendar quarter, all respondents are affected by the per-quarter fee in period 3. However, the treatment status in period 2 is less clear cut because participants do not have to pay the fee in

[1] The question reads as follows: Have you gone to a doctor within the last three months? If yes, please state how often.

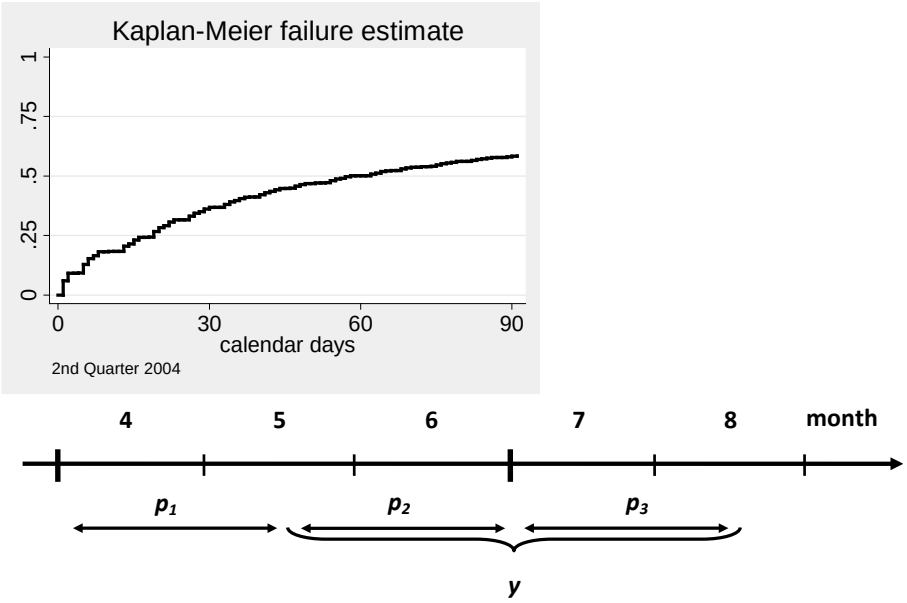


Figure 1.1: Decomposition of the reporting period and degree of misclassification according to the AOK sample

the second period if they have already paid it in the first period. According to the 2004 claims data set, 56% of the population had already paid the fee in period 1 (see also Figure 1.1). Hence, a large fraction of the population was indeed unaffected by the new fee in period 2 which was a part of the reporting period. Previous research results (e. g. Schreyögg and Grabka, 2010) were, however, based on a clear-cut treatment status. They assumed that all participants in the GSOEP were equally affected by the reform independent of the day of the interview. Hence, there was a misclassification in the treatment level, which generally leads to an attenuation bias (Aigner, 1973). This explains why previous studies did not find significant effects. In the next paragraph I explain the underlying problem more formally and provide a solution to overcome it.

The probability of at least one doctor visit within the reporting period can easily be obtained by

$$\begin{aligned} Pr(y > 0) &= 1 - Pr(y = 0) \\ &= 1 - Pr(s_2 = 0, s_3 = 0) \\ &= 1 - Pr(s_2 = 0)Pr(s_3 = 0) \\ &= 1 - [Pr(s_1 = 0, s_2 = 0) + Pr(s_1 > 0, s_2 = 0)]\, Pr(s_3 = 0) \end{aligned} \tag{1.1}$$

where the number of doctor visits in period p_k is s_k for $k = 1, 2, 3$ and the number of visits in the reporting period is $y = s_2 + s_3$. For illustration purposes, I assume that the doctor visits follow a Poisson process. This justifies the third equality because all periods are disjoint time intervals. The Law of Total Probability then gives the fourth equality. It separates the individuals into two groups. Firstly, the group of individuals that had not visited a doctor in the first period and therefore had to pay the fee in the second period. Secondly, the group of individuals that had visited a doctor in the first period and thus had access to free visits in the second period. Compared to the years before the reform, the out-of-pocket costs during period 2 were unchanged in the latter group. This is the variation in the treatment level that I want to exploit in this study.

Whenever the reporting period differs from a calendar quarter, like in a survey, there is a misclassification of the treatment status in a simple before-after comparison in which all observations are considered as treated after the reform. Actually, $Pr(s_1 > 0)$ is the probability of a false-positive treatment status in the second period. Since the reporting period consists of period 2 and 3, I can observe the true reform effect only if $Pr(s_1 > 0) = 0$. The group of participants who were interviewed at the end of a calendar quarter is the only group where I know for sure that this condition is true. I therefore hypothesize that the true reform effect and in particular the causal effect of the per-quarter fee can only be observed in the group of participants who were interviewed at the end of a calendar quarter. To get rid of the misclassification problem, I use different models that account for the day of the interview. The details of these models are explained in the next section.

1.3 Data and estimation

I use two separate data sets to verify my identification strategy. The primary source of data is the GSOEP, which is an annual survey started in 1984. The second data source is a claims data set from the largest German sickness fund. I have used this data set to imitate a survey with two randomly assigned interview days. This enables me to verify essential assumptions of my identification strategy that are untestable with survey data. In the following, I finalize my identification strategy and state my hypotheses. Then, I explain how the claims data set can be used to investigate the validity of the assumptions.

I created a data set using the GSOEP and a data set using claims data from the AOK. I selected a period of four years centered around the health care reform of 2004 and used the years 2002/03 to observe the behavior before the reform and 2005/06 as post-reform years.[2] The sample includes men and women aged 20 to 60. The basic estimation strategy is to pool the data over the four years and evaluate the effect of the fee on the probability of at least one visit to a doctor in the observed three months.[3] I use linear probability models (LPM) to determine the effect of the reform. The conditional probability of at least one doctor visit is $Pr(y > 0|\boldsymbol{x}_k, \boldsymbol{w}) = \boldsymbol{x}_k'\boldsymbol{\beta}_k + \boldsymbol{w}'\boldsymbol{\gamma}$ where y is the number of doctor visits. The index k refers to different parameterizations of the linear index $\boldsymbol{x}'\boldsymbol{\beta}$ which have been estimated to evaluate the effect of the reform. They are explained in more detail in the following paragraph. The vector $\boldsymbol{w}$ stands for other characteristics controlled for in the regressions. It contains a second-order polynomial in age, two indicators for self-reported health status, three indicators for interview season and employment status. Furthermore, I include the variables *female, years of education, married, household size, welfare recipient and household income.*[4]

The LPM have been estimated using different parameterizations. One current

[2] The year 2004 has to be ignored because many interviews in the GSOEP take place in the first three months and thus the observed three-month period lies partly in the pre- and post-reform time.

[3] Generally, it is possible to analyze the effect on the number of visits using a count data model. In this study, however, I am primarily interested in the binary decision whether an individual visits a doctor or not because not visiting a doctor is the only way to avoid the fee.

[4] In the claims data set I can observe only individuals' age and gender.

method to evaluate health care reforms in Germany is to compare privately and statutorily insured persons with a differences-in-differences approach because privately insured persons are unaffected by these changes. Under the assumption of a common trend between privately and statutorily insured persons, this approach can identify the effect of the entire reform only if this effect is independent of when the interview took place. Here $\boldsymbol{x}_k'\boldsymbol{\beta}_k$ is

$$\boldsymbol{x}_1'\boldsymbol{\beta}_1 = \beta_{1,1}\text{after} + \beta_{1,2}\text{SHI} + \beta_{1,3}\text{after} * \text{SHI} \tag{1.2}$$

where the variable *after* indicates the post-reform years and the variable *SHI* is an indicator of whether a person is SHI-insured. The interaction between *after* and *SHI* denotes a statutorily insured observation after the reform.

As hypothesized in section 1.2, SHI-insured participants in the GSOEP are, depending on the day of their interview, differently affected by the new fee. The estimation strategy in equation (1.2), which has also been used in previous studies, ignores the variation of the day of the interview which may lead to a misclassification of the treatment status. The models discussed in the following use the information about the day of the interview to assess the reform effect and in particular the causal effect of the per-quarter fee:

$$\boldsymbol{x}_2'\boldsymbol{\beta}_2 = \beta_{2,1}\text{after} + \beta_{2,2}\text{SHI} + \beta_{2,3}\text{after} * \text{SHI} * \text{q} + \beta_{2,4}\text{after} * \text{SHI} * (1\text{-q}) \tag{1.3}$$

where q measures the degree of misclassification which rises with decreasing overlap between reporting period and calendar quarter. I use a dichotomous and a continuous measure of the misclassification. In the latter case q is the distance of the day of the interview to the nearest end of a calendar quarter and $q = 0$ indicates individuals who were interviewed at the end of a quarter where the misclassification is zero. $\beta_{2,4}$ therefore reveals the true reform effect. The assumption that the reform effect is independent of the day of the interview can be rejected once $\beta_{2,4}$ is significantly different from $\beta_{2,3}$. Additionally, it is possible to identify the reform effect by a dichotomous variable that splits the participants into two groups - similar to the example discussed in section 1.2. In group A the interview took place at the end of a quarter (plus or minus

10 days).[5] Group B contains the remaining sample.[6] The results from the dichotomous measure are very similar to the results from the continuous measure, indicating that the misclassification in group A is close to zero. I therefore rely on the dichotomous measure in the following analysis.

The final estimation strategy makes it possible to disentangle the influence of the per-quarter fee from the effect of the contemporaneous increase of co-payments for drugs. This is because group A and B are equally affected by the increase in prescription fees but they are differently exposed to the new fee for doctor visits. The difference between both groups is caused by some members of group B who had access to free visits in the second period but who would have been induced to participate completely if they had been interviewed at the end of a calendar quarter. The probability of having to pay the fee is thus different between both groups and if the fee works, this will affect each group's demand for medical care differently. Here the analysis is very similar to the estimation of a local average treatment effect (Imbens and Angrist, 1994). Using only SHI-insured observations, the different trends can be estimated by

$$\boldsymbol{x}_3'\boldsymbol{\beta}_3 = \beta_{3,1}\text{after} + \beta_{3,2}\text{A} + \beta_{3,3}\text{after} * \text{A} \tag{1.4}$$

where $\beta_{3,2}$ is expected to be zero since both groups are untreated before the reform. The parameter $\beta_{3,3}$ identifies the post-reform difference between both groups, which is caused by the lower treatment level in group B. Furthermore, $\beta_{3,3}$ should be larger in magnitude in the sicker population because they visit a doctor on average more often than healthy people. Hence, more of them have access to free visits and the misclassification of the treatment status in the second period is higher, implying that

[5] The group A indicator must contain some days around the end of a calendar quarter since too few participants were interviewed exactly at the end of a quarter.

[6] Schreyögg and Grabka (2010) apply a similar approach but do not use the variation of the day of the interview to identify the causal effect of the new fee. They restrict their sample to those respondents who gave their interview within 15 days *before* the end of a quarter. This classification, however, incorrectly assigns persons to their group B that were interviewed close to the end of a calendar quarter where the misclassification is close to zero - namely, those participants who were interviewed at the beginning of a calendar quarter. This classification, thus, decreases the exogenous variation in the degree of misclassification and it is not surprising that they only found slightly larger effects for their group A.

a larger fraction of individuals contributes to the identification.

There are two essential assumptions of my identification strategy. The first one is that the distance to the end of a calendar quarter was assigned to each survey participant in a way that can be considered as random. The evidence in the GSOEP data strongly suggests a random assignment. Nevertheless, I additionally verify this assumption using the claims data set. Here I can randomly split the sample to simulate certain interview or reporting periods and compare the results of equation (1.4) with the corresponding results from the GSOEP. The first "interview period" starts on July 1st and ends on September 30th of each year, i.e. it covers a full calendar quarter. This is group A in the claims data set. The "interview period" of group B is from May 16th to August 15th of each year. I used the claims data set to calculate the number of doctor visits in both "interview periods". The second key assumption is that the probability of visiting a doctor in the different interview periods would have been the same in the absence of the new fee. Here seasonal fluctuations are a potential concern since both "interview periods" are not completely overlapping. I used the 16 to 17 year olds to investigate this assumption. Given a common trend, this group makes it possible to separate seasonal effects, since people younger than 18 do not have to co-pay at all. Hence, in the absence of seasonal effects there should be no difference between both "interview periods" in the group of 16 to 17 year olds. The random assignment and the absence of seasonal effects would suggest that there are also no differences between both "interview periods" in the adult population – apart from the variation in the probability of having to pay the fee.

1.4 Results

Table 1.1 shows that the per-quarter fee alters the observable behavior of the SHI-insured persons in the GSOEP in a special manner. It displays the sample means for the years before and after the reform grouped by whether or not the respondents were interviewed at the end of a quarter. Interestingly, after the reform the share of respondents with at least one doctor visit is significantly lower when participants were interviewed at the end of a quarter (group A) compared to the second group of

interviews which took place sometime in the middle of a quarter (group B). This is, however, not the case before the reform. In both groups 64% visit their doctor at least once in three months before the reform. The unconditional probability decreases to 61.6% in group A after the reform, whereas it stays unchanged at around 64% in group B. Apart from the stronger decline in group B, the results are very similar in the claims data set from the AOK Hesse. Here I can also see a difference between group A and B after the reform but no difference before the reform.

Table 1.1 also gives evidence that the distance to the end of a calendar quarter is quasi-randomly assigned. There are namely no significant differences between group A and B in important predictors of need for medical care. For instance, the average age in the GSOEP is 40 in both groups and around 54% of the respondents are female. Self-reported health (SRHS) is also very similar in both groups. In contrast to the AOK Hesse data set, the assignment to both groups was not by definition random in the GSOEP. Therefore it is here particularly important to see that there is neither a difference in the outcome before the reform nor any differences in important explanatory variables.

In this paragraph I verify my identification strategy using the claims data set. Table 1.2 compares and contrasts the estimation results from the GSEOP data with the results from the AOK claims data. The corresponding estimation strategy is described in equation (1.4). The first two columns are based on the survey data, the third column shows the corresponding results from the claims data set and the last column shows the results for the 16 to 17 year olds which allows me to separate potential seasonal effects. There are some differences between the survey data and the claims data set. Firstly, while I can observe many potential covariates in the survey, I only observe individuals' age and gender in the claims data set. For comparison reasons, the covariates in Table 1.2 are thus restricted to a second-order polynomial in age and a gender indicator. Secondly, while the individuals in the claims data set are insured with AOK, the survey participants can be insured in all existing statutory sickness funds. This is particularly important because if someone is SHI-insured, he can choose between all statutory sickness funds. As a result, the risk pool of AOK may differ from the other sickness funds. According to official figures, AOK insurees are slightly

Table 1.1: Group means before and after the reform

	GSOEP		AOK Hesse	
	2002 & 2003	2005 & 2006	2002 & 2003	2005 & 2006
At least one doctor visit	**0.640**	**0.616**	**0.658**	**0.581**
	0.640	0.642	0.659	0.610
Age	**39.77**	**40.36**	**40.62**	**40.86**
	39.94	40.65	40.69	40.96
Female	**0.532**	**0.548**	**0.471**	**0.480**
	0.534	0.543	0.477	0.485
SRHS (1: very good, ..., 5: very bad)	**2.475**	**2.529**		
	2.495	2.525		
Education in years	**11.75**	**12.01**		
	11.82	11.90		
Married	**0.626**	**0.616**		
	0.620	0.594		
Household size	**3.063**	**3.024**		
	3.035	2.955		
Welfare recipient	**0.037**	**0.053**		
	0.038	0.069		
Ln(income)	**7.719**	**7.767**		
	7.695	7.695		
Observations	**3,680**	**3,430**	**152,086**	**147,923**
	19,664	16,770	152,091	147,563

Only SHI-insured observations are used in the GSOEP sample (**Group A** / Group B).

Note: The lower fraction of women in the AOK Hesse sample is in accordance with official figures. See e.g. "GKV-Versicherte nach Alter und Wohnort GKV-Statistik KM6 zum 1. Juli 2005", Federal Ministry of Health.

older than the entire population.[7] Therefore I also provide the estimation results for the group of survey participants who are insured with AOK (see column 2). Finally, there is a regional difference. While all individuals in the claims data set live in Hesse, the GSEOP is a German-wide survey. However, I do not believe that this affects the comparability of both samples since Hesse is a large federal state and certainly representative of Germany.

The results are striking. Although I randomly split the claims data set into two groups, there is a significant difference in the probability of visiting a physician between these two groups after the reform which was not the case before the reform (see column 3 of Table 1.2). On the other hand, splitting the 16 to 17 year olds into two groups with different "interview periods" does not lead to a significant difference (see column 4), indicating that the effect on the adult population is not due to seasonal

[7] Source, available only in German: "GKV-Versicherte nach Alter und Wohnort GKV-Statistik KM6 zum 1. Juli 2005", Federal Ministry of Health.

Table 1.2: Estimation results from the different data sets

	GSOEP	GSOEP	AOK Hesse	
		AOK only	20-60	16-17
Age/10	-0.0861	-0.0475	-0.1473	0.3875
	(0.0185)	(0.0326)	(0.0051)	(0.0542)
$Age^2/100$	0.0161	0.0135	0.0234	
	(0.0023)	(0.0040)	(0.0000)	
Female	0.1495	0.1583	0.1719	0.1654
	(0.0058)	(0.0102)	(0.0016)	(0.0061)
After	-0.0024	-0.0064	-0.0523	-0.0130
	(0.0047)	(0.0085)	(0.0016)	(0.0086)
A	0.0015	0.0143	0.0011	-0.0105
	(0.0085)	(0.0147)	(0.0020)	(0.0088)
After x A	-0.0260	-0.0382	-0.0279	-0.0076
	(0.0123)	(0.0217)	(0.0023)	(0.0122)
Observations	43,544	13,760	599,663	27,763

Dependent variable: at least one doctor visit. Parameter estimates after separate linear regressions using only SHI-insured observations. Cluster-robust standard errors in parentheses.

fluctuations. While the post-reform difference in the adult population is significant, the pre-reform difference between both groups is not. This is the expected result when the group assignment is random and when there are no seasonal differences between both "interview periods" in the years before the reform. Given the random assignment in the claims data set and the likely absence of seasonal influences, I therefore conclude that the post-reform difference stems from the variation in the probability of having to pay the new fee as hypothesized in section 1.2.

These results are very similar in the survey data set indicating that my identification strategy also works in the GSOEP. But while the decline in group B, which can be interpreted as the general effect of the reform, is about 5.2% in the claims data set, it is insignificant in the survey data set. Given the accuracy of the claims data set, this may point to a survey effect in the response behavior of the participants. The estimates from the differences-in-differences regression, which are discussed in more detail later in this study, strengthen this suggestion. They reveal an overall effect of 4.2-5.4% in the GSOEP sample (see Table 1.3). This is distinctly larger than the overall effect of 2.8% in the GSOEP sample reported in Table 1.2, and also closer to the overall effect of 8.0% in the AOK sample. The following part of the results is based on the GSOEP

Table 1.3: Estimation results from the GSOEP data set

	full sample			if $A = 1$
Age / 10	-0.0982	-0.0984	-0.0981	-0.0529
	(0.0179)	(0.0179)	(0.0179)	(0.0392)
Age^2/100	0.0123	0.0123	0.0123	0.0073
	(0.0022)	(0.0022)	(0.0022)	(0.0047)
Female	0.1350	0.1351	0.1350	0.1413
	(0.0058)	(0.0058)	(0.0058)	(0.0120)
Education / 10	0.0791	0.0795	0.0792	0.0910
	(0.0110)	(0.0110)	(0.0110)	(0.0230)
Married	0.0293	0.0292	0.0293	0.0311
	(0.0065)	(0.0065)	(0.0065)	(0.0141)
Household size	-0.0247	-0.0247	-0.0248	-0.0232
	(0.0024)	(0.0024)	(0.0024)	(0.0052)
Good health	-0.1691	-0.1693	-0.1692	-0.1780
	(0.0053)	(0.0053)	(0.0053)	(0.0121)
Bad health	0.1629	0.1628	0.1628	0.1520
	(0.0061)	(0.0061)	(0.0061)	(0.0144)
Welfare recipient	-0.0114	-0.0117	-0.0116	-0.0282
	(0.0134)	(0.0134)	(0.0134)	(0.0340)
Ln(income)	0.0375	0.0376	0.0377	0.0313
	(0.0058)	(0.0058)	(0.0058)	(0.0129)
After	0.0146	0.0135	0.0141	0.0128
	(0.0118)	(0.0118)	(0.0118)	(0.0296)
SHI	0.0312	0.0316	0.0314	0.0474
	(0.0104)	(0.0104)	(0.0104)	(0.0238)
After x SHI	-0.0226			-0.0537
	(0.0125)			(0.0314)
q is		continuous	dichotomous	
After x SHI x q		-0.0082*	-0.0184*	
		(0.0136)	(0.0126)	
After x SHI x (1-q)		-0.0417*	-0.0430*	
		(0.0143)	(0.0146)	
Observations	49,326	49,326	49,326	8,084

Dependent variable: at least one doctor visit in the reporting period.
Models also account for seasonal effects and employment status.
Cluster-robust standard errors in parentheses.
* The parameter estimates are significantly different at the 1%-level.

data because in this data set I can take advantage of the richer set of covariates and moreover can observe the privately insured as an additional contemporaneous control group.

Table 1.3 displays the average marginal effects of the probit regressions that compare privately and statutorily insured individuals. Most effects are very similar to those found in Winkelmann (2004a). The probability of visiting a doctor is u-shaped in age and women are more likely to see a physician than men. The effects of education and household size are larger in the present study and married persons are somewhat more likely to visit a physician in Winkelmann's sample. The estimation strategy in the first column is a simple differences-in-differences approach conditional on covariates (see equation (1.2)). According to these estimates, the reform leads to a slight decrease in the probability of visiting a physician in the group of SHI-insured persons. It is only weakly significant at the 10%-level despite the large sample size. This result is in line with previous research which concluded that the per-quarter fee had failed to reduce the demand for doctor visits (Augurzky *et al.*, 2006; Schreyögg and Grabka, 2010). However, this conclusion changes once the reform effect can vary with the degree of misclassification as in equation (1.3) (compare columns 2 and 3 of Table 1.3). Now, there is a strong and highly significant reform effect in both models given the misclassification is close to zero, i.e. $q = 0$. Figure 1.2 displays the reform effect over the entire

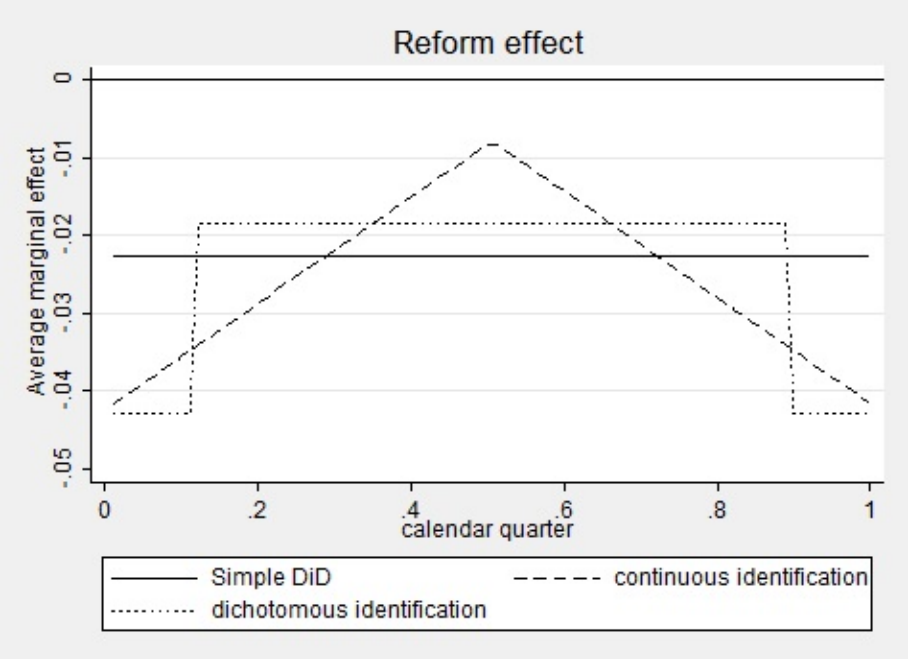


Figure 1.2: Average marginal effects for each day of the interview (GSOEP)

range of q. The average marginal effect of the reform is significantly stronger at the end of a calendar quarter, while I wrongly assume that the reform effect is constant in the simple differences-in-differences regression in equation (1.2). Comparing the reform effect at the end of a calendar quarter with the effect in the middle of a quarter, allows me to assess the effect of the per-quarter fee. According to Figure 1.2, at least half of the reform effect is caused by the per-quarter co-payment for doctor visits. The underlying estimates are significantly different at the 1%-level (see Table 1.3).

The model in the first column of Table 1.3 is inappropriate to evaluate the new co-payment for doctor visits. It assumes that the reform effect is independent of when the interview took place although GSEOP participants are differently affected by the new fee. Column 4 shows the estimation results for the group of participants who were interviewed around the end of a quarter. The misclassification is almost zero in this group. Here there is a significantly stronger decline in the probability of visiting a physician in the group of statutorily health insured individuals than in the group of privately insured. The average reform effect is -0.054 which is very similar to the results from the second and third column (-0.042 and -0.043). The reform effect in column 4 is, however, less significant, which is probably due to the distinctly smaller sample size.

An important result of this study is that the true reform effect can only be found if the reporting period is a full calendar quarter. Table 1.4 reports the estimation results for equation (1.4) which compares the different trends between participants interviewed around the end of a calendar quarter with the remaining sample. The estimation strategy in Table 1.4 is the same as in Table 1.2 but in the former table I only use the GSOEP data set and can thus take advantage of the richer set of covariates. According to the last row of Table 1.4, there is a significantly stronger decline in the probability of visiting a doctor in group A than in group B, similar to the results in Table 1.2. The post-reform difference between group A and B should be larger in the group of sick people, because they are more likely to visit a doctor in the unobserved period than healthy people. Thus, a larger fraction of them has had access to free visits and contributes to the identification. The upper panel in Table 1.4 shows the estimation results of equation (1.4) and the sample means of the outcome variable grouped by

Table 1.4: Comparison of trends for different subgroups of the population (GSOEP)

				Pre-reform		
Regressions	Parameter estimates			Number of	Probability	
conditional on SRHS	After	A	After x A	doctor visits	of any use	Obs.
very good	0.0277	0.0009	0.0239	0.99	0.43	4,075
	(0.0169)	(0.0289)	(0.0429)			
good	-0.0018	0.0013	-0.0392	1.35	0.56	20,138
	(0.0076)	(0.0129)	(0.0186)			
satisfactory	-0.0156	0.0156	-0.0253	2.49	0.72	13,525
	(0.0084)	(0.0152)	(0.0210)			
poor	-0.0021	0.0341	-0.0675	5.02	0.88	4,769
	(0.0112)	(0.0179)	(0.0275)			
bad	0.0122	-0.0014	-0.0717	9.35	0.94	1,030
	(0.0171)	(0.0295)	(0.0466)			
Entire sample	-0.0028	0.0084	-0.0314	2.24	0.64	43,544
	(0.0047)	(0.0085)	(0.0120)			

Dependent variable: at least one doctor visit in the reporting period.

Parameter estimates after separate linear regressions using only SHI-insured observations.

Covariates are the same as in Table 1.3. Cluster-robust standard errors in parentheses.

self-reported health status. The rise in the probability of visiting a doctor at least once and in the number of doctor visits indicate that the group of individuals with access to free visits increases with decreasing health status. As expected, the point estimate for the post-reform difference between group A and B is largest in the sick population. However, it is not significant (p-value=0.124) which might be caused by the distinctly smaller sample size. Apart from the group which reports a satisfactory health status, the point estimate rises in magnitude with decreasing health status. This indicates that the sicker population contributes more to the identification, which might be due to a true reduction in demand for medical care. However, since sick people have a high need for medical care, it is unrealistic that they permanently reduce their visits to zero in order to avoid paying €10 per quarter. The difference between both groups in the sicker population may therefore also be caused by a second effect of the fee. It may have generated an incentive to cluster a given level of care in as few as possible calendar quarters. So once people have access to free visits, they may be tempted to group their visits into this quarter. This would affect the observable behavior in group B stronger because some members of group B are exempt from the fee already at the beginning of their reporting period. The post-reform difference between both groups

could therefore also be caused by an incentive to cluster visits. Such a behavior would also lead to a larger variance in the number of doctor visits. In the claims data set there is indeed an increase in the variance after the reform. While the sample variance was 26.7 before the reform, it rises to 30.5 after the reform. I will investigate this issue in more detail in a follow-up analysis.

1.5 Conclusion

This study exploits exogenous variation in the day of the interview to assess the effect of a per-quarter fee for doctor visits on utilization. This approach is appealing because it compares random samples of the SHI-insured population that are differentially affected by the new fee. Therefore, a differences-in-differences regression makes it possible to disentangle its influence from potential macro effects. In particular, it separates the influence of the fee from the effect of the contemporaneous increase of co-payments for drugs.

The key contribution of this study is to show the necessity of comparing full quarters before and after the reform to assess the effect of the 2004 health care reform. Otherwise the treatment status is not clear-cut since some statutorily insured individuals have had access to free visits after the reform. Ignoring this leads to an underestimation of the reform effect due to a misclassification of the treatment status. The attenuation bias increases with decreasing overlap between reporting period and calendar quarter. The majority of participants in the GSOEP, however, has not been interviewed at the end of a calendar quarter and their treatment status is thus subject to misclassification. The true effect of the fee is therefore diluted in a simple before-after comparison.

The present study overcomes this problem by accounting for the misclassification. The probability of visiting a physician is significantly influenced by the health care reform of 2004. It decreased by around 5 percentage points. Due to my identification strategy, I can attribute at least half of this effect to the per-quarter fee for doctor visits.

Chapter 2

Heterogeneous effects of a nonlinear price schedule for outpatient care[†]

2.1 Introduction

Nonlinear price schedules are a common feature of many health insurance systems. Nonlinearities often arise due to deductibles or combinations of co-payments and maximum out-of-pocket amounts. In order to increase cost consciousness the insured have to bear part of their health care costs. But once the sum of out-of-pocket expenditures exceeds a certain amount, co-payments for further health care use decrease or even drop to zero. Economic theory predicts that not all insured react to co-payments in the same way if the latter are combined with maximum out-of-pocket amounts. Instead, individuals' price sensitivity is predicted to depend on expected future health care use, which naturally varies between individuals. For example, in a price schedule where costs drop to zero once out-of-pocket expenditures exceed a certain amount, individuals who expect that their out-of-pocket expenditures will exceed the maximum amount have little incentive to reduce care today. They will likely have to pay the same overall amount independent of their health care use today (Keeler *et al.*, 1977; Ellis, 1986).

[†] This chapter is joint work with Amelie Wuppermann. Peter Ihle, Ingrid Schubert and Joachim Winter also participate in the project but are not co-authors of this chapter.

In this study, we provide an empirical example for these theoretical considerations. One of the challenges in this type of analysis is that individuals' expectations on future health care use are unobserved. In earlier studies this problem has been solved by predicting the missing information based on observable characteristics or prior health care use (see Ellis, 1986; Contoyannis *et al.*, 2005; Meyerhoefer and Zuvekas, 2010). In this study, we present results that are in line with the theoretical predictions without constructing expectations. Instead, we allow for heterogeneous effects of the introduction of a nonlinear price schedule in a finite mixture model. In this model, we can estimate reactions for different classes of individuals without having to specify a priori which individual belongs to which class. We thus do not need to observe expected health care use a priori.

For our analysis, we use exogenous variation in the price schedule introduced by a recent reform of the German statutory health insurance system. The statutory health insurance is the public health insurance system in Germany that is mandatory for most employees and covers around 90% of the German population. In 2004, a nonlinear price schedule for doctor visits was introduced in this system. Before 2004 the publicly insured did not have to co-pay for doctor visits. Since 2004, they have to pay a fee of €10 for the first visit to a doctor in each calendar quarter. Additional visits in the same quarter are free of charge. The consumer price thus drops from €10 to €0 after the first doctor visit in a quarter. This per-quarter fee should mainly affect the decision of a first visit in a quarter, because it is the first visit that determines whether the fee has to be paid. Additional visits within one quarter do not change the overall costs.[1] We therefore focus on the question whether the reform affected the probability of at least one visit in a quarter. We call this access to outpatient care.

Due to the nonlinearity in the price for doctor visits introduced by this reform, we expect that whether individuals change their behavior following the reform depends on individuals' expectations of health care use which in turn depend on their health status. For individuals who expect that they will likely have to visit a doctor within the next three months, access to outpatient care might not change. Healthy individuals,

[1] Of course, individuals could try to fit as many visits as possible into one quarter once the first visit has taken place in order to avoid paying the fee in later quarters. We focus on this possible heaping of visits in a follow-up analysis.

however, might expect that they can avoid paying the fee and the probability of no doctor visits might increase.

The literature that focuses on the effect of the specific reform of the German statutory health insurance delivers mixed results. Augurzky *et al.* (2006) and Schreyögg and Grabka (2010) find that the reform had essentially no effect on the health care use of the statutorily insured in the German Socio-Economic Panel (GSOEP). Using the same data set Farbmacher (2009), on the contrary, presents evidence according to which the statutorily insured on average reduced their propensity to visit a doctor. Farbmacher's results are in line with Rückert *et al.* (2008) who find that individuals surveyed in the Bertelsmann Healthcare Monitor report avoiding and delaying doctor visits. We add to this literature in two ways. First, we use a new data set for our analysis. Our data is based on health insurance claims from the largest German sickness fund. The main advantage compared to survey data is that we reliably observe doctor visits. Second, we are the first to take into account that the newly introduced per-quarter fee has an implicit deductible structure. We therefore focus on heterogeneous effects in our analysis.

Our results indicate that the average probability of no doctor visit significantly increases after the reform by about 4 percentage points. This result is in line with Rückert *et al.* (2008) and Farbmacher (2009) and indicates that the reform affected access to health care on average. Furthermore, we find evidence for heterogeneous effects that are in line with the theoretical considerations by Keeler *et al.* (1977): The results of our finite mixture model indicate that for about 36% of individuals access to outpatient care is not changed by the reform. Among the remaining individuals, on the contrary, access decreases significantly after the reform. Post-estimation analyses further indicate that the individuals who do not react to the reform are sicker than the others. They might not react to the new co-payment because they expect that they cannot avoid paying the fee due to their health status, i.e. they assume to have at least one visit in a quarter anyway.

While our results indicate that the per-quarter fee is successful in influencing the healthier individuals' behavior, sicker individuals do not react to the fee. Following an idea proposed by van Kleef *et al.* (2009), we suggest to change the timing of the fee for

sick individuals: Instead of a fee that is due for the first visit in a quarter, which sick individuals cannot avoid due to their health status, individuals should get a certain number of free visits before the fee applies. The number of free visits should ideally be individual specific and depend on the unavoidable number of visits. As this is difficult to reliably observe, characteristics that are not easily influenced by an individual, such as age and sex, could serve as criteria. Our analysis suggest that women up to the age of 50 should for example get one visit for free, while men in the same age group should continue to pay at the first visit.

This chapter is structured as follows: The next section describes the health care reform in more detail. Section 3 introduces the data set and Section 4 explains our estimation strategy. In Section 5 the results are presented. Section 6 contains a discussion of the results and Section 7 concludes.

2.2 Incentive effects of the reform

The health care reform that we analyze became effective at the beginning of the year 2004. With this reform various financial incentives have been implemented in the German statutory health insurance with the intend to increase patients' cost consciousness which may help to reduce moral hazard. The most radical element of the reform has been the introduction of a per-quarter fee for doctor visits. While patients did not have to co-pay for doctor visits before the reform, they have to pay €10 for the first visit to a physician in each quarter of the year since the reform in 2004. Further doctor visits to the same doctor within this quarter are free of charge. Visits in the same quarter to other doctors are also exempt from the fee if the patient gets a referral by the doctor whom he visited at first. Alternatively, patients can visit other doctors without referral and pay the fee again.

Additional parts of the reform have been an increase in prescription fees and the abolishment of the possibility to prescribe over the counter medications. Since 2004 the patients have to copay at least €5 and at most €10 for their drugs - depending on the drug price. The pre-reform prescription fees were between €4 and €5. Thus in the best case there has been no increase in prescription fees while the increase in

fees could have been up to 150% in the worst case. Furthermore, the sickness funds no longer pay for eyeglasses and visual aids. Figure 2.1 shows the changes in Germany's consumer price index for medical care. The reform has permanently increased the prices for medical care. According to this index, it has been the largest health care reform in Germany for more than a decade.

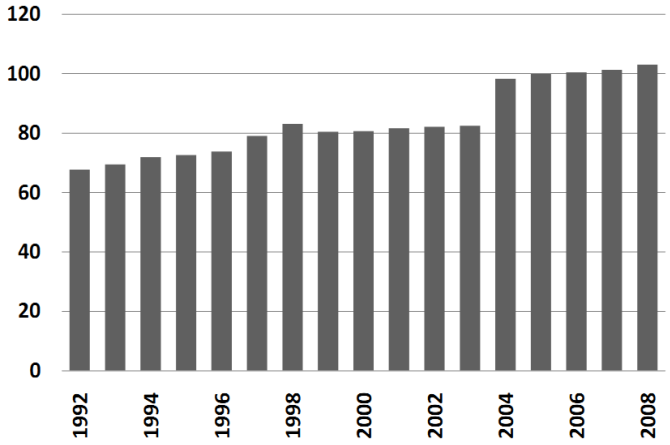


Figure 2.1: Germany's consumer price index for medical care
Source: German Federal Statistical Office, own visualization

The per-quarter fee for doctor visits was the central element of the reform and attracted a lot of attention in the media. We are mainly interested in its effect on the probability of visiting a physician. As paying the fee can only be avoided by not visiting any physician within a quarter, it should mainly affect access to outpatient care where access is measured as the probability of at least one visit to any type of physician.

The per-quarter fee has introduced a nonlinearity in the price schedule. The reason for this is its implicit deductible. It has to be paid only at the first visit in a calendar quarter. Hence, given a referral the patient's price for doctor visits drops to zero after the first visit. This nonlinearity generates varying incentives depending on the individual's health status. Keeler *et al.* (1977), for instance, show that under uncertainty a rational individual facing a deductible will not base decisions on nominal prices. The authors instead argue that "the greater the chance that future expenditures will exceed the deductible, the cheaper today's visit to the doctor". A rational individual will thus anticipate that the price drops to zero at a certain consumption level. In the German

case the intertemporal effect on prices is relatively easy to assess because it only depends on individuals' knowledge about their probability of visiting a physician in the next three months. In the extreme case their behavior is unaffected by the per-quarter fee once they know that they have to visit a doctor in a certain calendar quarter (e. g. to get a new prescription of a medicine for a chronic disease). The effective price for doctor visits is thus lower for individuals with chronic conditions. Hence, if demand for outpatient care depends on effective prices, we expect a weaker decrease in demand for individuals with high risks of doctor visits than among low risk individuals. In this argument we assume that individuals still visit doctors in case of major conditions and take medically indicated drugs (e.g. to treat chronic diseases) despite the increase in co-payments. This assumption can be justified by the low co-payment level. Generally, individuals in Germany will not get into severe financial troubles due to out-of-pocket expenditures.

2.3 Data

The analysis is based on insurance claims data from the largest German sickness fund in the years 2002 to 2005. The data contains information on a 18.75% random subsample of all individuals in the German state of Hesse who are insured with this sickness fund. At the beginning of each year a sample refreshment is taken in order to keep the sample representative for the insured population.[2]

The data contains information on doctor visits, the type of doctor visited, diagnoses made at each visit measured in ICD-10 codes and prescribed medications. As we are interested in the reactions to the introduction of the per-quarter fee, which can only be avoided by not visiting any physician within a quarter, we aggregate the information in the claims data to the quarterly level. Furthermore, we group information on the different doctor visits into visits to general practitioners (GPs) and visits to specialists. The data then contains information on the number of GP visits per quarter and the number of specialist visits per quarter for each individual.

[2] See *http://www.pmvforschungsgruppe.de/content/02_forschung/02_b_sekundaerd_1.htm* for a short description of the data in German.

The main advantage of using claims data compared to survey data is that doctor visits are reliably observed. However, the only information on individuals' health that is contained in the data comes from the diagnosis codes and prescription drugs. This information is only available for individuals who have seen a doctor. Independent of doctor visits only information on age and sex is available. A disadvantage of the data is thus that it only contains few observables that do not depend on whether an individual has visited a doctor.

An additional drawback is that the data set consists only of publicly insured individuals and therefore includes no adults for whom nothing has changed due to the reform and who could thus serve as a control group in our analysis. Only individuals younger than 18 are generally exempt from paying the per-quarter fee. They, however, may not be suitable as a control group for the entire adult population.[3] We thus revert to before-after comparisons to identify the effects of the reform in the adult population. Our results therefore rely on the assumption that in the absence of the reform no changes in health care use would have occurred or that if there were changes they were not considerably large.

Our sample is further restricted to all individual-quarter pairs for which we observe the use of outpatient services within the entire quarter. Individual-quarter pairs are excluded, for example, if the individual switches from or to a different insurer within the quarter. This ensures that the length of the period at risk is the same for each observation.

Table 2.1 shows descriptive statistics for the third quarter of each year.[4] The average age is almost unchanged over time reflecting the conducted refreshments of the sample. The average number of doctor visits in our sample is around 4.5 per quarter. On average individuals visit a GP a little less than once a month and a specialist 1.7 times per quarter. This in international comparison relatively high use of physician services is in line with information from other data on doctor visits in Germany (see Grobe *et al.*, 2010).

While the average number of doctor visits per quarter does not show a clear change

[3] This feature of the reform suggests a natural division in treatment and control group among teenagers. We conduct a difference-in-difference analysis for teenagers in a follow-up study.

[4] The descriptive statistics are very similar for all quarters of the year.

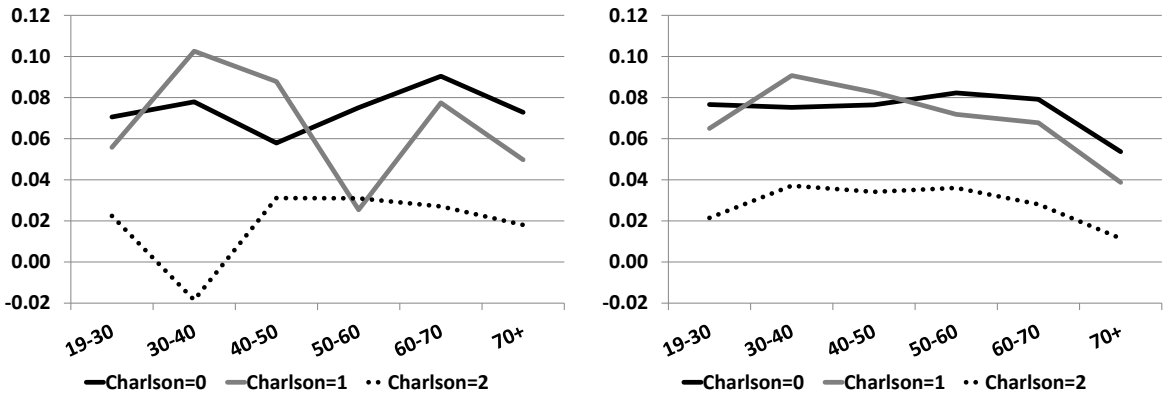
Table 2.1: Descriptive statistics

	3Q 2002		3Q 2003		3Q 2004		3Q 2005	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	51.87	18.71	52.16	18.75	52.31	18.81	52.17	18.88
19-39	0.31	0.46	0.30	0.46	0.29	0.46	0.29	0.46
40-59	0.31	0.46	0.31	0.46	0.32	0.47	0.33	0.47
60-79	0.31	0.46	0.31	0.46	0.31	0.46	0.30	0.46
≥ 80	0.07	0.26	0.08	0.27	0.08	0.27	0.08	0.27
Female	0.52	0.50	0.52	0.50	0.52	0.50	0.52	0.50
# GP visits	2.77	4.12	2.75	4.11	2.77	4.16	2.85	4.33
GP>0	0.63	0.48	0.63	0.48	0.60	0.49	0.59	0.49
# GP visits truncated at 0	4.43	4.45	4.40	4.45	4.65	4.51	4.83	4.72
# Specialist visits	1.75	3.65	1.70	3.49	1.63	3.57	1.71	3.73
Specialist>0	0.46	0.50	0.46	0.50	0.42	0.49	0.42	0.49
# Specialist visits truncated at 0	3.83	4.60	3.70	4.37	3.89	4.64	4.05	4.85
GP= 0 & Specialist = 0	0.27	0.44	0.27	0.44	0.33	0.47	0.33	0.47
N	256,071		249,851		246,379		248,328	

after the reform, two possible effects of the increased co-payments become evident in Table 2.1. Between 2003 and 2004, the fraction of individuals with at least one GP visit and the fraction with at least one specialist visit in the third quarter both decline, from 63% to 60% for GPs and from 46% to 42% for specialists. Individuals thus seem to avoid to contact either type of doctor after the reform.

As the co-payment can only be avoided by seeing neither type of physician, we are particularly interested in how the probability of no doctor visit within a quarter changed after the reform. Information on this is contained at the bottom of Table 2.1. While in the third quarter of 2002 and 2003, roughly 27% of the sample visit neither a GP nor a specialist, this is the case for 33% of individuals in the years after the reform. The average probability of no doctor visit per quarter thus increases by about 6 percentage points after the reform.

The change in the probability of no doctor visit in 2004 compared to 2003 is also depicted in Figure 2.2. This figure shows the changes in the third quarter of 2004 compared to 2003 separately for men and women in different age groups and in different health status. The health status is captured by the Charlson Index (Charlson *et al.*, 1987). This index is based on 17 diseases identified from the diagnosis codes available in our data set. Each disease is assigned a weight between 1 and 6 depending on disease



(a) Men (b) Women

Figure 2.2: Changes in the probability of no doctor visit by age and Charlson index

severity.[5] The Charlson Index is the sum of these weights, truncated at 2. A value of 0 thus indicates that an individual had no diagnosis of any of the Charlson conditions and a value of 1 or 2 indicates the presence of more severe co-morbidities.

As the Charlson Index is based on diagnosis codes and those codes are only available if an individual has seen a doctor, the Charlson Index is endogenous. In order to mitigate this problem, we construct the Charlson Index for the observations in the third quarter of each year based on their diagnoses in the two prior quarters. For example, the "Charlson 0" group in 2003 contains all individuals who had no diagnoses of Charlson conditions in the first two quarters of 2003.

Figure 2.2 presents evidence for heterogeneous effects across the different groups. The probability of no doctor visit generally increases after the reform with similar magnitude for both genders. These increases are much smaller for individuals with a Charlson Index of 2 than for the other groups. Sicker individuals thus seem to react less to the reform than healthier ones. This holds true for all age groups and both genders. Furthermore, there is some evidence that women who are older than 70 react less to the reform than younger women conditional on the Charlson group. For men there is no clear age pattern.

Overall, Figure 2.2 indicates that the change in the probability of no doctor visit is stronger for healthy than for less healthy individuals. These descriptive results,

[5] A list of the diseases and corresponding weights is displayed in Appendix A.1.

however, rely on an endogenous measure of health. In order to test our hypothesis of heterogeneous results without having to rely on the health information in the data, we use a finite mixture model. This model allows us to estimate different effects for separate groups in the population without having to explicitly stratify the data by observable characteristics a priori.

2.4 Econometric framework

Our data consists of a panel of individuals across time and across different physician types. Individuals can seek care from GPs (y_{1it}) and/or specialists (y_{2it}). The panel is unbalanced over time, and each individual i is observed in T_i quarters. Over time and physician types, individual i is thus observed $2 \cdot T_i$.

Suppose that individual i belongs to a latent class j for the entire observational period. The probability of belonging to class j is π_j. Within a latent class, we use bivariate probits to jointly model the decision to visit a GP and/or a specialist. Although we are mostly interested in whether individuals visit any doctor within a quarter, independent of the type of doctor visited, we use the separate information on GPs and specialists in order to gain potentially relevant information. This additional information might allow a more accurate classification of individuals into latent classes. The joint probability of the dependent variables over the observed period is the product of T_i independent probabilities, given fixed class membership, i.e.,

$$Pr(y_{1i}, y_{2i}|x_i, \theta_j) = \prod_{t=1}^{T_i} \Phi_2[(2y_{1it} - 1)x_{it}\beta_j, (2y_{2it} - 1)x_{it}\gamma_j, (2y_{1it} - 1)(2y_{2it} - 1)\rho_j] \quad (2.1)$$

where $\Phi_2()$ stands for the cumulative bivariate normal distribution function and x_i denotes the vector of covariates that includes age, sex, seasonal fixed effects and year fixed effects. θ_j contains the vector of parameters for GP visits (β_j), the vector of parameters for specialist visits (γ_j), and the parameter ρ_j. The latter indicates the extent to which the errors in the underlying structural model covary.

The log-likelihood function is given by

$$ln(L) = \sum_{i=1}^{I} ln \left(\sum_{j=1}^{J} \pi_j Pr(y_{1i}, y_{2i}|x_i, \theta_j) \right) \tag{2.2}$$

where I is the number of individuals in the dataset and J is the number of latent classes. The likelihood function is maximized directly using the Newton-Raphson algorithm. First and second derivatives are calculated numerically in Stata's optimization package. In order to get a manageable data set for this estimation, we use a 3% random subsample for this part of our analysis. This gives us on average a little more than 7,500 individuals per quarter.

For the interpretation of our results we calculate average marginal effects of the variables in x_i on $Pr(y_{1i} = 1|x_i, \theta_j)$ and $Pr(y_{2i} = 1|x_i, \theta_j)$. Standard errors of the marginal effects are calculated using the delta method. As the reform effect is captured by changes in the probability of no doctor visit, we also calculate marginal effects on the probability of no doctor visit, i.e. on $Pr(y_{1i} = 0, y_{2i} = 0|x_i, \theta_j)$.[6]

Furthermore, we calculate posterior probabilities of membership in the different latent classes for each individual i as

$$Pr(y_{1i}, y_{2i} \in k|x_i, \widehat{\theta}) = \frac{\widehat{\pi}_k Pr(y_{1i}, y_{2i}|x_i, \widehat{\theta}_k)}{\sum_{j=1}^{J} \widehat{\pi}_j Pr(y_{1i}, y_{2i}|x_i, \widehat{\theta}_j)} \tag{2.3}$$

where $Pr(y_{1i}, y_{2i}|x_i, \widehat{\theta}_k)$ is defined as in equation (2.1).

These posterior probabilities on the one hand help to see how well the different latent classes are separated by the estimation. On the other hand, one can assign individuals to a specific latent class based on their posterior probabilities and then characterize each latent class using observable characteristics. In addition to the variables age and sex that are included in x_i we use the health information contained in the claims data in this part of the analysis. As this information is only available conditional on doctor visits we do not include it in the vector of control variables and it is thus external to the estimation.

[6] See Appendix A.2 for details.

2.5 Results

Table 2.2 reports AIC and BIC information criteria for the finite mixture bivariate probit model described in the last section with different number of latent classes. In addition to finite mixture models with 2, 3, and 4 latent classes we estimated a standard one-component bivariate probit model. Estimations with more than 4 latent classes failed to converge and are likely overparameterized. The information criteria displayed in Table 2.2 indicate that the model with 4 latent classes fits the data best. Furthermore, the posterior probabilities for the four latent classes are well separated as the figure in Appendix A.3 shows. We therefore focus on the results of this model.

Marginal effects and their standard errors based on the results of our finite mixture bivariate probit model with 4 latent classes are reported in Table 2.3. In addition to the marginal effects in the different latent classes, Table 2.3 displays the overall effects that are derived as weighted averages of the effects in the different latent classes. As a comparison, the last column of Table 2.3 reports marginal effects and standard errors based on the standard bivariate probit model.

The overall marginal effects and the marginal effects of the standard bivariate probit model have the same signs and are similar in magnitude. Women have a higher probability to visit a GP and to visit a specialist at least once in a quarter than men. The probability to visit a GP at least once increases with age, particularly so for individuals aged 40 to 60. The probability to visit a specialist at least once only increases for individuals in this age group. The probabilities to visit either type of doctor at least once are lower in the summer months (quarter 2 and 3) and higher in the last quarter compared to the first quarter of a year. Overall, the marginal effects thus show expected signs.

Table 2.2: Model selection

Model	N	LogL	df	AIC	BIC
Bivariate Probit	120,521	-148,116.6	25	296,283	296,526
2 LC FMM Bivariate Probit	120,521	-130,342.6	51	260,787	261,282
3 LC FMM Bivariate Probit	120,521	-122,594.1	77	245,342	246,089
4 LC FMM Bivariate Probit	120,521	-118,899.9	103	**238,006**	**239,005**

Table 2.3: FMM bivariate probit – marginal effects

	LC 1	LC 2	LC 3	LC 4	Overall	BiProbit
GP						
Female	0.028**	0.016	0.041**	0.082*	0.037***	0.067***
	(0.010)	(0.016)	(0.013)	(0.035)	(0.009)	(0.007)
Age Splines						
Age 19-40	0.003***	0.002	0.003**	0.001	0.003***	0.002***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Age 40-60	0.002	0.002	0.007**	0.006	0.004**	0.007***
	(0.002)	(0.003)	(0.002)	(0.005)	(0.001)	(0.001)
Age 60-80	-0.001	-0.001	-0.005	-0.009	-0.003	0.000
	(0.003)	(0.008)	(0.004)	(0.013)	(0.003)	(0.001)
Age $>$ 80	-0.003	0.004	0.008	0.013	0.004	-0.008*
	(0.007)	(0.028)	(0.013)	(0.034)	(0.008)	(0.003)
Quarter Dummies						
Q2	-0.005	-0.012	-0.007	-0.031***	-0.011***	-0.011***
	(0.005)	(0.006)	(0.006)	(0.009)	(0.003)	(0.003)
Q3	-0.011*	-0.021**	-0.016**	-0.046***	-0.020***	-0.020***
	(0.005)	(0.007)	(0.006)	(0.011)	(0.003)	(0.003)
Q4	0.017***	0.008	0.022***	-0.014	0.012***	0.010***
	(0.005)	(0.008)	(0.006)	(0.011)	(0.003)	(0.003)
Year Dummies						
2003	0.007	-0.012	0.023***	-0.008	0.006	0.001
	(0.006)	(0.008)	(0.006)	(0.015)	(0.003)	(0.003)
2004	0.031***	-0.055***	-0.052***	0.017	-0.022***	-0.027***
	(0.008)	(0.009)	(0.009)	(0.020)	(0.004)	(0.004)
2005	0.028**	-0.033**	-0.065***	0.041	-0.018***	-0.027***
	(0.009)	(0.011)	(0.010)	(0.023)	(0.004)	(0.004)
Specialist						
Female	0.083***	0.081***	0.110***	0.197***	0.107***	0.164***
	(0.015)	(0.011)	(0.013)	(0.028)	(0.009)	(0.007)
Age Splines						
Age 19-40	0.002	-0.001	0.002	-0.006**	0.000	-0.001*
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Age 40-60	0.004	0.002	0.002	0.014***	0.004**	0.008***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.001)	(0.001)
Age 60-80	-0.009***	0.000	-0.009***	0.006	-0.005*	-0.006***
	(0.002)	(0.003)	(0.002)	(0.007)	(0.002)	(0.001)
Age $>$ 80	-0.005	-0.009	-0.007	-0.045	-0.012	-0.014***
	(0.004)	(0.011)	(0.005)	(0.024)	(0.007)	(0.003)
Quarter Dummies						
Q2	-0.002	-0.003	-0.002	-0.010	-0.003	-0.004
	(0.007)	(0.005)	(0.006)	(0.010)	(0.003)	(0.003)
Q3	-0.004	-0.002	-0.009	-0.034**	-0.009**	-0.010**
	(0.006)	(0.006)	(0.006)	(0.010)	(0.003)	(0.003)
Q4	0.005	0.000	-0.006	-0.009	-0.002	-0.003
	(0.006)	(0.006)	(0.006)	(0.010)	(0.003)	(0.003)
Year Dummies						
2003	0.041***	-0.014*	-0.009	0.006	0.005	0.002
	(0.007)	(0.007)	(0.007)	(0.011)	(0.004)	(0.004)
2004	0.004	-0.047***	-0.031***	-0.090***	-0.033***	-0.037***
	(0.009)	(0.007)	(0.008)	(0.015)	(0.004)	(0.004)
2005	-0.004	-0.028***	-0.017	-0.107***	-0.028***	-0.033***
	(0.010)	(0.008)	(0.009)	(0.016)	(0.004)	(0.004)
ρ	0.367***	0.559***	0.410***	0.236***		0.489***
	(0.021)	(0.017)	(0.014)	(0.030)		(0.008)
π_j	0.265	0.248	0.355	0.132		

Notes: π_j is the probability of class membership in latent class j. Standard errors for the marginal effects in parentheses. They are calculated using the delta method. Standard errors of the underlying coefficients are clustered at the individual level. *p<0.05; **p<0.01; ***p<0.001

Of particular interest for our analysis are the changes in the probabilities of doctor visits over the years. The effects of the year dummies capture these changes compared to the reference year 2002. The overall results of the finite mixture bivariate probit model and the standard bivariate probit concordantly show significant reductions in the probabilities in the years 2004 and 2005 compared to 2002. In 2003, however, there is no significant difference to 2002. As 2003 is a pre-reform year, the absence of an effect in 2003 supports the assumption that there would have been no changes in the outcome variable in the absence of the reform.

These overall effects, however, are composed of different effects in the latent classes. While there are reductions in the probability to visit a GP and in the probability to visit a specialist in latent classes 2 and 3 after the reform, in latent class 4 only the probability of a specialist visit is reduced. In latent class 1, there is even an increase in the probability to visit a GP after the reform while the probability to consult a specialist does not change significantly.

The results in Table 2.3 thus suggest that the reform has an effect on access to outpatient care and that this effect might be heterogeneous across individuals. However, as the per-quarter fee has to be paid at the first visit to a doctor in a quarter independent of the type of doctor visited, while all additional visits to other doctors are free of charge, it is more informative to analyze the change in the probability of not visiting any doctor. The marginal effects on $Pr(y_{1i} = 0, y_{2i} = 0|x_i, \theta_j)$ are displayed in Figure 2.3.

The overall effect displayed in Figure 2.3 is not significant in the pre-reform year 2003 compared to 2002, but highly significant in the post-reform years 2004 and 2005. It indicates that individuals went to see a doctor at least once with an almost 4 percentage points lower probability after the reform. Furthermore, Figure 2.3 presents evidence for heterogeneous reform effects: While latent class 1 does not react to the reform, there are strong reactions in latent classes 2, 3 and 4.

In latent class 1, the probability of no doctor visit is reduced by about 1 percentage point in the post-reform years 2004 and 2005 compared to 2002. However, the same change already occurs in the pre-reform year 2003. These changes therefore cannot be interpreted as reform effects.

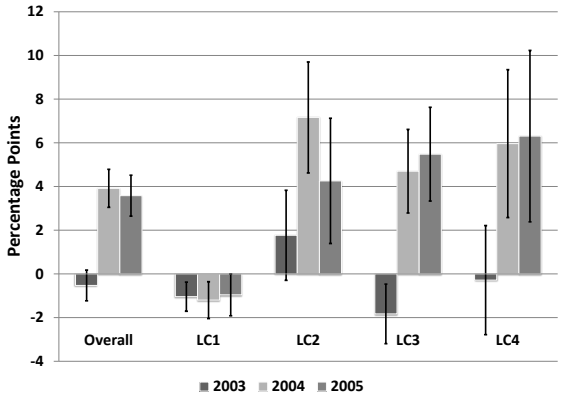


Figure 2.3: Changes in the probability of no doctor visit
Note: Error bars indicate 99%-confidence intervals.

In latent classes 2, 3 and 4 to the contrary, we see strong changes in the probability of no doctor visit in the post-reform years compared to 2002. The changes in 2003 are only significant in latent class 3 and much smaller in magnitude than the post-reform changes. Individuals in latent classes 2, 3 and 4 thus react to the reform by going to see a doctor with a lower probability.

Naturally, the question arises, what distinguishes the individuals belonging to the different latent classes? Given the parameter estimates, we derive the posterior probabilities for each individual i to belong to the four different latent classes as described in the previous section. Each individual is then assigned to the latent class with the highest posterior probability. Table 2.4 reports averages of observed characteristics for the different latent classes. Beside age and sex, Table 2.4 includes information on diagnoses in form of the Charlson Index and on prescription drugs that individuals got from GPs and specialists. The latter are measured in defined daily doses (DDDs).[7] These give a rough measure of drug consumption within the different classes and adjust for the fact that different drugs can be of different potency. The DDDs do not take into account, however, which amount of the drugs was actually prescribed.

Additionally, Table 2.4 shows the fraction of observations in each latent class that is exempt from co-payments. Before and after the reform in 2004, individuals could apply for an exemption from co-payments if the amount of their co-payments exceeded 2% of

[7] DDDs are defined by the World Health Organization, see *http://www.whocc.no/ddd/definition_and_general_considera/* for a definition.

Table 2.4: Comparison of latent classes

	LC 1	LC 2	LC 3	LC 4
Age	57.43	44.03	56.72	44.64
Female	0.60	0.40	0.51	0.62
Charlson Index	0.98	0.17	0.85	0.47
Fraction Charlson=0	45.19	83.54	51.32	69.45
Fraction Charlson=1	9.57	6.17	9.80	8.25
Fraction Charlson=2	43.52	5.21	36.72	18.38
DDD	201.26	13.07	156.22	79.73
Exempt from co-payments	0.18	0.02	0.12	0.08
% of individuals	36.24	29.91	26.67	7.18

Notes: Fractions of Charlson Index do not add up to 1 because of missing observations. DDD stands for defined daily dose.

their gross yearly income. Welfare recipients and chronically sick individuals could be entirely exempt from co-payments before the reform. Since 2004, however, the 2% rule also applies to welfare recipients and the chronically sick can only be exempt from further payments within a year when they have already co-paid 1% of their gross yearly income. Individuals can apply for these exemptions already at the beginning of each year. In order to do so, they have to pay either 1% or 2% of their gross yearly income up front to their insurer. The information on whether individuals are exempt from co-payments is only available after the reform. The results in Table 2.4 thus only include observations in the post-reform period.[8]

The last row of Table 2.4 shows that about 36% of individuals were assigned to latent class 1, 30% to latent class 2, 27% to latent class 3 and 7% to latent class 4. Comparing the different latent classes, it becomes evident that latent class 1 does not seem to differ generally from the other latent classes in terms of age or sex. Observations assigned to latent class 1 are on average of the same age as observations in latent class 3, and the sex composition in latent class 1 is similar to the one in latent class 4. There are differences, however, concerning the Charlson Index, the DDDs and exemptions from co-payments. Observations in latent class 1 have a higher average Charlson Index and an average DDD that is markedly higher than in the other latent classes. Observations in latent class 1 thus have more severe diagnoses and take more potent

[8] For all variables that are available before the reform the results are almost identical for the pre-reform period.

drugs on average, indicating that these observations are sicker than observations in the other latent classes. This confirms the hypothesis that among the individuals who do not react to the reform are the relatively sick.

Given the exemption rules described above, one could argue that individuals who know ex ante that they will be exempt from all co-payments do not react to the introduction of the per-quarter fee. In line with this argument, Table 2.4 shows that latent class 1 has the largest fraction of observations that are exempt from co-payments. However, there are still 82% of observations in latent class 1 that do not react to the introduction of the per-quarter fee, even though they are not exempt from co-payments in general. The nonlinearity of the per-quarter fee thus results in heterogeneous reactions not only through the general exemption rules.

Overall, the results from the finite mixture model confirm that the introduction of the per-quarter fee in the German statutory health insurance had heterogeneous effects: Around 36% of individuals are ex post assigned to the class that does not react to the reform, while there are strong reactions among the rest. Consistent with the theory of Keeler *et al.* (1977), the results indicate that among the individuals who do not react to the introduction of the per-quarter fee are the ones who are relatively sick.

2.6 Discussion

Our results indicate that the per-quarter fee fails to influence the health care use of high-risk individuals. One way to affect the behavior of high-risk individuals as well might be to introduce a fee for every single visit to a doctor. However, the disadvantage of a *per-visit* fee would be an increasing financial burden, especially for high-risk individuals. A different solution that would allow to affect high-risk individuals' behavior without increasing their financial burden is related to the idea of "shifted deductibles" by van Kleef et al. (2009). The authors propose a new design of nonlinear price schedules which might even decrease the financial burden for high-risk individuals. They suggest to use a deductible that does not start at zero like a traditional deductible but at an individual specific starting point which depends on risk characteristics of the individuals. This overcomes the problem that high-risk individuals often know

for sure that their expenditures will reach the level of the deductible in which case co-payments will only have an income effect. As a shifted deductible increases the uncertainty about the out-of-pocket expenditures, it may increase the incentive effect for high-risk individuals.

Shifted deductibles in our application translate to allowing different numbers of free doctor visits before the per-quarter fee has to be paid. Depending on their health status some statutorily insured individuals in Germany should receive free care up to a certain threshold. How many visits individuals receive for free before the fee applies should optimally be individual specific. However, as van Kleef *et al.* (2009) note using objective criteria like age and sex might be "practical and understandable to consumers".

Similar to van Kleef *et al.* (2009) we calculate the number of free visits that maximize the uncertainty about the out-of-pocket expenditures based on the observed number of visits for different groups of individuals. Figure 2.4 shows the empirical cumulative distribution function of doctor visits in the third quarter of 2003 - truncated at 10 - and the optimal number of free visits for six different age groups separately for men and women. According to van Kleef *et al.* (2009) the optimal starting point of a deductible maximizes the uncertainty about the out-of-pocket expenditures. We use the variance as a measure of the uncertainty. In our application the variance is maximized at the threshold that splits the empirical cumulative distribution function into two equal parts. This is because people either have to pay the fee or not, which is a dichotomous event. The variance of this event is thus maximized at a probability of 0.5.

According to Figure 2.4, the optimal number of free visits is almost always unequal to zero which is the current number of free visits. While young men should pay the fee at the first visit in each quarter, young women should get one or two free visits per quarter. This gender difference can be justified, for instance, by preventive examinations that are provided regularly to young women. An additional cause might be the contraceptive pill that is available only with prescription in Germany leading to additional doctor visits for women compared to men.

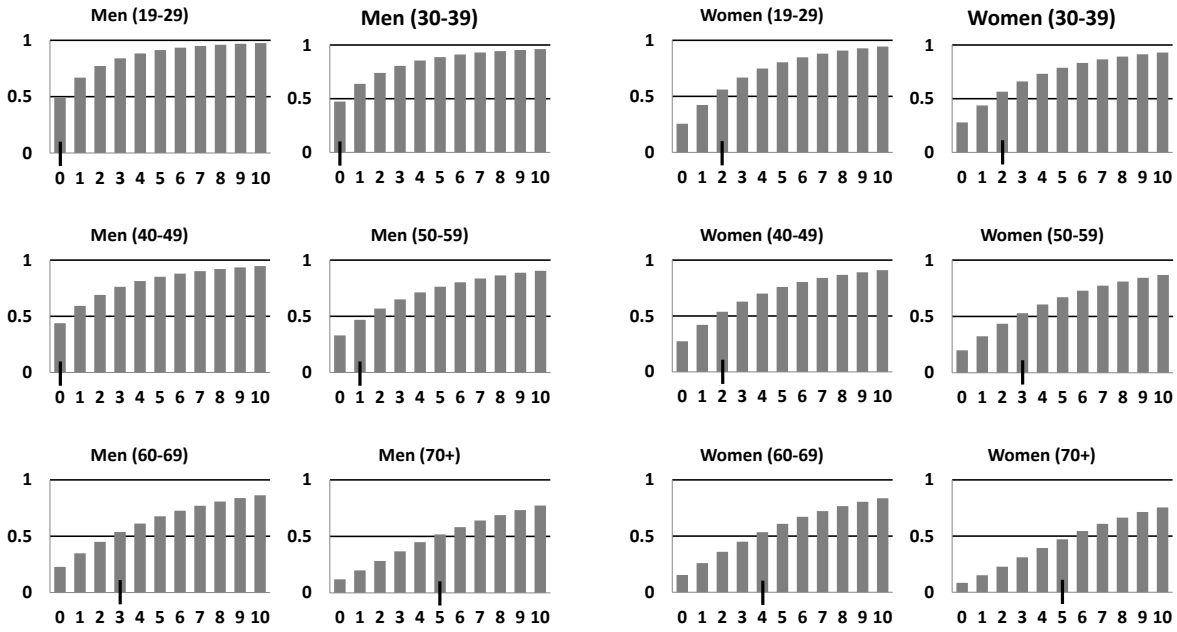


Figure 2.4: Empirical CDF of doctor visits and number of free visits grouped by age and sex

With four free visits per quarter even a 60 year old woman, for instance, might have a real chance to avoid the fee by slightly changing her behavior. Assuming that people within the groups are homogeneous, implies that the entire population now faces similar effective prices. However, it is very likely that the homogeneity assumption does not hold in Figure 2.4 since the classification of the groups is too crude. An additional dimension like the Charlson index as a measure of chronic conditions is certainly helpful to make the groups more homogeneous. On the other hand, this measure is not as objective as age and sex. Table 2.5 contains the optimal number of free visits by age and Charlson groups separately for men and women. Since the level of doctor visits is generally higher, women should get more free visits than men which would then also cover, for instance, the higher level of preventive care. According to Table 2.5, women up to the age of 50 should get one free visit per quarter but only if they have none of the Charlson diseases. Otherwise they should get up to five free visits depending on their risk characteristics.

Table 2.5: Number of free visits separately for age, sex and Charlson index

	Charlson index					
	Men			Women		
Age	0	1	2+	0	1	2+
19-29	0	1	2	1	3	4
30-39	0	1	3	1	3	4
40-49	0	1	3	1	3	5
50-59	0	2	4	2	3	5
60-69	1	2	5	2	4	6
70+	1	3	6	3	5	7

2.7 Conclusion

In this paper, we present empirical evidence for heterogeneous effects of a nonlinear price schedule that was introduced in the German statutory health insurance. The nonlinearity takes the form of a co-payment for doctor visits that only has to be paid for the first visit in each quarter of the year. Prices for doctor visits in the same quarter drop to zero once the fee has been paid.

Following theoretical considerations on health care demand in the presence of nonlinear price schedules, we anticipate that the per-quarter fee changes access to health care differently across individuals. In particular, individuals who expect in the beginning of a quarter that they will have to visit a physician at some point within the quarter have a lower incentive to change behavior than individuals who expect that no visit will be necessary. As the expectations on doctor visits likely depend on individuals' health status we expect that the reform effects vary between individuals with good and bad health.

In a descriptive analysis we find that individuals in worse health react less to the reform than healthier individuals. Our measure of health, however, depends on the outcome variable. Namely, it is only observed if individuals visit a doctor. We therefore allow for unobserved heterogeneity in a finite mixture model. The results of this model show that some individuals react to the reform while others do not. Examining the different groups indicates that those individuals who do not react are in worse health. Our results are thus in line with the theoretical predictions of how nonlinear price schedules affect the demand for medical care.

Our findings allow two conclusions. First, the per-quarter fee seems to be effective to increase cost consciousness for many individuals, in particular the healthier ones. Second, there are individuals – among them the sick – who do not react to the introduction of the per-quarter fee. The reform thus seems to fail to increase cost consciousness and to reduce moral hazard in this group.

Following the idea of van Kleef *et al.* (2009) we suggest a slight refinement to the current system that might be more effective in reducing moral hazard for all individuals. We suggest shifted thresholds as a means to change the behavior of higher-risk individuals as well. With a shifted threshold individuals would get different numbers of doctor visits for free in each quarter before having to pay a fee. Since the number of free visits increases with decreasing health status, even high-risk individuals would have a chance to avoid the new fee by changing their behavior. Furthermore, allowing a certain number of free visits for high-risk individuals has the potential of reducing the financial burden for these individuals.

Appendices

A.1 Definition of Charlson index

Table 2.6: Definition of Charlson Score

	Charlson Comorbidity	Assigned Weights
1	Myocardial infarction	1
2	Congestive heart failure	1
3	Peripheral vascular disease	1
4	Cerebrovascular disease	1
5	Dementia	1
6	Chronic pulmonary disease	1
7	Rheumatic disease	1
8	Peptic ulcer disease	1
9	Mild liver disease	1
10	Diabetes without complications	1
11	Diabetes with chronic complications	2
12	Hemiplegia or paraplegia	2
13	Renal disease	2
14	Cancer	2
15	Moderate or severe liver disease	3
16	Metastatic carcinoma	6
17	AIDS/HIV	6

A.2 Marginal Effects

This appendix describes how the marginal effects on the probabilities of each type of doctor visit and on the joint probability that neither type of visit occurs ($Pr(y_{1i} = 0, y_{2i} = 0|x_i, \theta_j)$) are calculated. Standard errors for the marginal effects are derived using the delta method. Each marginal effect is calculated for each individual i and for each of the J latent classes. We report the average marginal effects of each latent class j as the average over all individual marginal effects in this class. Furthermore, the weighted average of the effects in the different latent classes gives an overall effect. For continuous explanatory variables the marginal effects are calculated using the calculus method. Marginal effects of binary variables are calculated with the finite difference method.

For a continuous variable x the marginal effects on $Pr(y_{ki} = 1|x_i, \widehat{\theta}_j)$ with $k \in \{1, 2\}$ in latent class j for individual i are calculated as

$$ME_{xi}Pr(y_{1i} = 1|x_i, \widehat{\theta}_j)_j = \widehat{\beta}_{x,j}\phi(x_i'\widehat{\beta}_j) \tag{2.4}$$

$$ME_{xi}Pr(y_{2i} = 1|x_i, \widehat{\theta}_j)_j = \widehat{\gamma}_{x,j}\phi(x_i'\widehat{\gamma}_j) \tag{2.5}$$

where $\phi()$ stands for the standard normal density function. We report $\frac{1}{I}\sum_{i=1}^{I} ME_{xi}Pr(y_{ki} = 1|x_i, \widehat{\theta}_j)_j$.

The marginal effects of a continuous variable x on the joint probability for y_{1i} and y_{2i} in latent class j are calculated for individual i as follows:

$$\begin{aligned} ME_{xi}Pr(y_{1i} = 0, y_{2i} = 0|x_i, \widehat{\theta}_j)_j &= \frac{\partial \Phi_2(q_{1i}x_i'\widehat{\beta}_j, q_{2i}x_{it}'\widehat{\gamma}_j, q_{1i}q_{2i}\widehat{\rho}_j)}{\partial x_i} \\ &= q_{1i}\widehat{\beta}_{x,j}\phi(q_{1i}x_{it}'\widehat{\beta}_j)\Phi\left(\frac{q_{2i}x_{it}'\widehat{\gamma}_j - q_{1i}^2 q_{2i}\widehat{\rho}_j x_{it}'\widehat{\beta}_j}{\sqrt{1-\widehat{\rho}_j^2}}\right) \\ &+ q_{2i}\widehat{\gamma}_{x,j}\phi(q_{2i}x_{it}'\widehat{\gamma}_j)\Phi\left(\frac{q_{1i}x_i'\widehat{\beta}_j - q_{1i}q_{2i}^2\widehat{\rho}_j x_i'\widehat{\gamma}_j}{\sqrt{1-\widehat{\rho}_j^2}}\right) \end{aligned} \tag{2.6}$$

where $\Phi_2()$ stands for the cumulative bivariate normal distribution function, $\Phi()$ indicates the standard normal cdf, and $q_{ki} = 2y_{ki} - 1$, with $k \in \{1, 2\}$.

The calculation of marginal effects of discrete variables is illustrated for the year dummies. The marginal effects of the years 2003, 2004 and 2005 with 2002 as reference in latent class j for individual i are calculated as

$$ME_{year,i}Pr(y_{1i} = 1|x_i, \widehat{\theta}_j)_j = \Phi(x_i'\widehat{\beta}_j + \widehat{\beta}_{year,j}) - \Phi(x_i'\widehat{\beta}_j) \tag{2.7}$$

$$ME_{year,i}Pr(y_{2i} = 1|x_i, \widehat{\theta}_j)_j = \Phi(x_i'\widehat{\gamma}_j + \widehat{\gamma}_{year,j}) - \Phi(x_i'\widehat{\gamma}_j) \tag{2.8}$$

for the marginal probabilities, and as

$$\begin{aligned} ME_{year,i}Pr(y_{1i} = 0, y_{2i} = 0|x_i, \widehat{\theta}_j)_j = \Phi_2 \Big(q_{1i}(x_i'\widehat{\beta}_j + \widehat{\beta}_{year,j}), q_{2i}(x_i'\widehat{\gamma}_j + \widehat{\gamma}_{year,j}), q_{1i}q_{2i}\widehat{\rho}_j\Big) \\ - \Phi_2(q_{1i}x_i'\widehat{\beta}_j, q_{2i}x_i'\widehat{\gamma}_j, q_{1i}q_{2i}\widehat{\rho}_j) \end{aligned} \tag{2.9}$$

for the joint probability of y_{1i} and y_{2i}. $\widehat{\beta}_j$ and $\widehat{\gamma}_j$ now stand for the vectors of parameter estimates for all variables excluding the year indicators. Marginal effects for the variable female and the quarter dummies are calculated analogously.

The overall marginal effect, i.e. the marginal effect averaged over the latent classes, for any continuous or discrete variable x is then derived as weighted average of the marginal effects across the different latent classes

$$ME_{xi} = \sum_{j=1}^{J} \pi_j ME_{xi,j} \tag{2.10}$$

Again, averages over all individuals are reported.

Standard errors for the average marginal effects are derived using the delta method that delivers the variance for each average marginal effect as

$$Var(ME) = \nabla_g' Var(\theta) \nabla_g \tag{2.11}$$

where θ is the vector of all parameters that are estimated (β_j,γ_j,ρ_j, and p_j, where

$\pi_j = \frac{exp(p_j)}{1+\sum_{c=1}^{C} exp(p_c)}$ with $C = J - 1$), and ∇_g stands for the gradient of the marginal effect, $ME = g(\theta)$, with respect to θ. In order to calculate the variance on the average marginal effect, we set each element in the gradient to its sample average.

A.3 Posterior probabilities

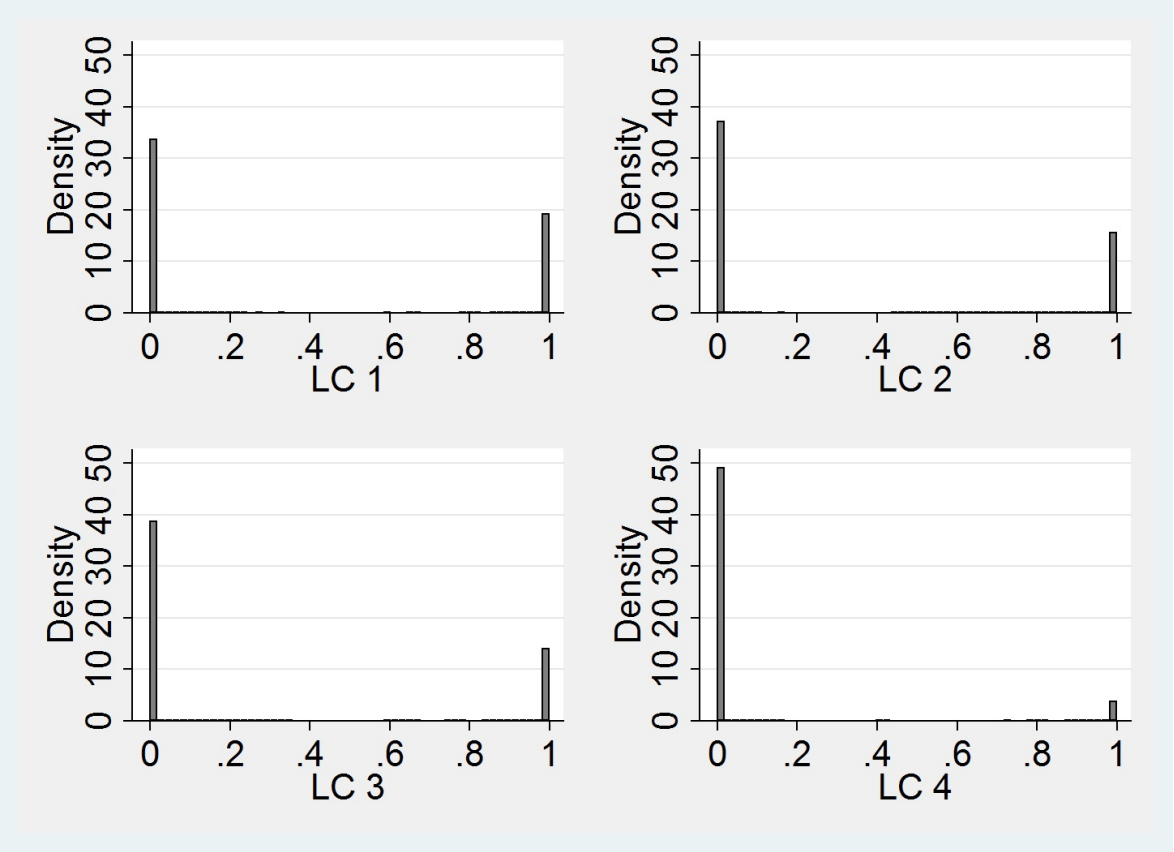


Figure 2.5: Posterior probabilities

Chapter 3

Extensions of hurdle models for overdispersed count data[†]

3.1 Introduction

Proposed by Mullahy (1986), the notion of a hurdle model is still very popular in modeling count data. It can be used in various contexts, such as job changes, fishing, or use of health care. The hurdle model typically combines a binary model to model participation (for example, modeling the patient's decision to visit the doctor) with a zero-truncated count data model to model the extent of participation for those participating (for example, modeling the number of doctor visits). In contrast with a single-index model, the hurdle model permits heterogeneous effects for individuals below or above the hurdle. In many applications the hurdle is set at zero and can therefore also solve the problem of excess zeros, that is, the presence of more zeros in the data than what was predicted by single-index count-data models. There are many possible combinations of binary and truncated count-data models. An often-used model combines a probit or logit model with a zero-truncated negative binomial model (for example, Vesterinen *et al.*, 2010 and Wong *et al.*, 2010).

In health economics, for instance, Pohlmeier and Ulrich (1995) has been one of the

[†] Parts of this chapter have been published in Farbmacher (2011a)

first studies to analyze the number of doctor visits using a hurdle model. The number of doctor visits may serve as a proxy for demand for health care. This measure may be determined by a two-part decision process. At first, it is up to the patient whether to visit a doctor. After the first contact, though, the physician influences the intensity of treatment (Stoddart and Barer, 1981 and Pohlmeier and Ulrich, 1995). Assuming that the error terms of the binary and the truncated models are uncorrelated, the maximization process can be separated. In this case, one can first maximize a binary model with at least one doctor visit as the dependent variable using the full sample. Second, one can estimate a zero-truncated regression separately using only observations with positive counts.

To account for unobserved heterogeneity, Pohlmeier and Ulrich (1995) applied a zero-truncated model based on the negative binomial distribution. While this often improves the fit of the model, some of the underlying assumptions are mainly based on convenience. Recent developments in the count data literature make it possible to relax these assumptions. For instance, Greene (2008) proposed a generalization of the negative binomial model and Dhaene and Santos Silva (2011) showed a general way to increase the flexibility of models in which unobserved heterogeneity has to be integrated out. Based on these findings, I develop extensions of the truncated negative binomial and the truncated Poisson log-normal model, which can be used to make the second part of hurdle models more flexible.

This chapter is structured as follows: The next section describes the basic specifications of truncated count data models and explains the proposed model extensions. These extensions are then applied to the Pohlmeier and Ulrich (1995) data set in section 3. Section 4 concludes.

3.2 Econometric models

To account for unobserved heterogeneity, inference is often based on the marginal distribution $h(y_i|\mathbf{x}_i)$ obtained after integrating out u_i:

$$h(y_i|\mathbf{x}_i) = \int_0^\infty f(y_i|\mathbf{x}_i, u_i)g(u_i|\mathbf{x}_i)\mathrm{d}u_i \tag{3.1}$$

where $g(\cdot)$ is called the mixing distribution. Santos Silva (2003) mentioned that there are two alternative approaches to construct hurdle models if unobserved heterogeneity is present. On the one hand, if $f(y_i|\mathbf{x}_i, u_i)$ is an untruncated count data distribution, the mixing is done in the first step and the truncating follows in the second step. On the other hand, if $f(y_i|\mathbf{x}_i, u_i)$ is already a truncated distribution, the order between mixing and truncating is the other way round. The choice between these two alternatives is not innocuous, and this seemingly slight difference can lead to substantially different results. The reason for this is the assumption of independence, $g(u_i|\mathbf{x}_i) = g(u_i)$, which is required before integration. It can be assumed to hold in the actual population or in the truncated one, but generally not in both populations at the same time (see also footnote 4 in Santos Silva, 2003).[1]

To get a closed-form solution of the integral in (3.1), all truncated models based on the negative binomial distribution belong to the class of models where the mixing is done in the first step. If, for instance, $f(y_i|\mathbf{x}_i, u_i)$ is of the Poisson form and u_i is independent of $\mathbf{x}_i$ and follows a gamma distribution, we obtain the negative binomial (NB) models. Greene (2008) proposed the NB-P model which encompasses the often used NB-1 and NB-2. Its probability function is

$$Pr(y_i = n|\mathbf{x}_i) = \frac{\Gamma(m_i + y_i)}{\Gamma(m_i)\Gamma(y_i + 1)} \left(\frac{m_i}{\lambda_i + m_i}\right)^{m_i} \left(\frac{\lambda_i}{\lambda_i + m_i}\right)^{y_i} \quad \text{for} \quad n \geq 0 \tag{3.2}$$

[1] A simple example illustrates this problem: Assume that a count data variable is generated by a Poisson process where the mean depends on a covariate x and an unobserved variable u. Independence holds in the actual population. For a given value of x, it is now more likely to get truncated if the unobserved individual effect is lower than average. As a consequence, x and u are correlated in the truncated population.

where $\lambda_i = exp(\mathbf{x}_i\boldsymbol{\beta})$, $m_i = \frac{1}{\delta}\lambda_i^{(2-P)} = exp((2-P)\mathbf{x}_i\boldsymbol{\beta} - ln(\delta))$. δ and P are parameters to be estimated in addition to $\boldsymbol{\beta}$. Setting $P = 1$ or $P = 2$ gives the NB-1 or NB-2 model.

I use a truncated version of the NB-P model to analyze strictly positive counts. It can be obtained by dividing the probability function by $1 - Pr(y_i = 0|\mathbf{x}_i)$:

$$Pr(y_i = k|y_i > 0, \mathbf{x}_i) = \frac{Pr(y_i = k|\mathbf{x}_i)}{1 - Pr(y_i = 0|\mathbf{x}_i)} \quad \text{for} \quad k \geq 1 \tag{3.3}$$

Thus the log-likelihood contribution of the zero-truncated NB-P is

$$\begin{aligned} lnL_i^I = {} & ln\Gamma(m_i + y_i) - ln\Gamma(m_i) - ln\Gamma(y_i + 1) \\ & + ln(s_i) + y_i ln\left(\frac{\lambda_i}{\lambda_i + m_i}\right) - ln(1 - s_i) \end{aligned} \tag{3.4}$$

where $s_i = (m_i/(\lambda_i + m_i))^{m_i}$.

While the likelihood of the zero-truncated NB-P model has a closed-form expression, there is no such result for the zero-truncated Poisson log-normal model. Winkelmann (2004a), for instance, used the latter model to analyze demand for health care. In contrast to the NB-P model, $f(y_i|\mathbf{x}_i, u_i)$ is now already a truncated probability function (namely, the zero-truncated Poisson function) and u_i is log-normal distributed, independent of $\mathbf{x}_i$. The likelihood contribution of the zero-truncated Poisson log-normal model is

$$L_i^{II} = \int_{-\infty}^{\infty} \frac{exp(-exp(\mathbf{x}_i\boldsymbol{\beta} + \sigma\epsilon_i))exp(\mathbf{x}_i\boldsymbol{\beta} + \sigma\epsilon_i)^{y_i}}{(1 - exp(-exp(\mathbf{x}_i\boldsymbol{\beta} + \sigma\epsilon_i)))y_i!}\phi(\epsilon_i)\mathrm{d}\epsilon_i \tag{3.5}$$

where $\epsilon_i = ln(u_i)$ is standard normally distributed and $\phi(\cdot)$ denotes the density function of the standard normal distribution. It has to be approximated numerically (e.g. by Gauss-Hermite quadrature). Dhaene and Santos Silva (2011) proposed a general way to increase the flexibility of models in which unobserved heterogeneity has to be integrated out. Their basic idea is "that replacing $g(\cdot)$ with some other density is equivalent to transforming ϵ_i monotonically and keeping $g(\cdot)$". Using this procedure, we can

thus relax the distributional assumptions about ϵ_i and get the following likelihood contribution:

$$L_i^{III} = \int_{-\infty}^{\infty} f(y_i|\mathbf{x}_i\boldsymbol{\beta} + \sigma d(\epsilon_i))\phi(\epsilon_i)\mathrm{d}\epsilon_i \tag{3.6}$$

where $f(\cdot)$ denotes the probability function of the zero-truncated Poisson model as in equation (3.5) and $d(\cdot)$ links the true but unknown density of ϵ_i to the maintained distributional assumptions about $g(\cdot)$. To make the model from (3.5) more flexible, we can replace ϵ_i with some approximation of $d(\epsilon_i)$. Dhaene and Santos Silva (2011) suggested to use a polynomial in ϵ_i or a transformation to normality. For the latter, they use $d(\epsilon_i) = sinh(\theta\epsilon_i)/\theta$. I choose the same transformation and adapt their procedure to the truncated model. After a change of variable, the integrals in (3.5) and (3.6) can be written as

$$L_i^{II,III} = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} f(y_i|\mathbf{x}_i\boldsymbol{\beta} + \sigma d(\sqrt{2}\nu_i))exp(-\nu_i^2)\mathrm{d}\nu_i \tag{3.7}$$

where $d(\cdot)$ is the identity function in model II and $d(\cdot) = sinh(\theta\sqrt{2}\nu_i)/\theta$ in model III. If θ goes to zero, $d(\cdot)$ becomes the identity function and model III converges to model II. Since there is no analytical solution of the integral in (3.7), it has to be approximated numerically. The likelihood contributions of the actually estimated models therefore are

$$L_i^{II,III} \doteq \frac{1}{\sqrt{\pi}} \sum_{r=1}^{R} f(y_i|\mathbf{x}_i\boldsymbol{\beta} + \sigma d(\sqrt{2}\nu_r))\omega_r \tag{3.8}$$

where ν_r and ω_r are the nodes and weights for the quadrature.[2]

[2] Details about the approximation are given in Appendices A.1 and A.2.

To compare the results of the models, I calculate relative marginal effects which in the zero-truncated negative binomial models are

$$\frac{\partial E(\cdot)^+}{\partial x_j}\frac{1}{E(\cdot)^+} = \beta_j - \frac{s_i m_i \left(Q\beta_j ln(\frac{m_i+exp(\mathbf{x}_i\boldsymbol{\beta})}{m_i}) + \frac{m_i Q\beta_j + exp(\mathbf{x}_i\boldsymbol{\beta})\beta_j}{m_i+exp(\mathbf{x}_i\boldsymbol{\beta})} - Q\beta_j\right)}{1-s_i} \tag{3.9}$$

where $Q = (2 - P)$. Generally, marginal effects of models based on integration have to be approximated numerically. The relative marginal effects of the zero-truncated Poisson log-normal models are

$$\frac{\partial E(\cdot)^+}{\partial x_j}\frac{1}{E(\cdot)^+} = \int_{-\infty}^{\infty} \beta_j - \frac{exp(-\mu_i)\mu_i\beta_j}{1-exp(-\mu_i)}\phi(\epsilon_i)\mathrm{d}\epsilon_i \tag{3.10}$$

where $\mu_i = exp(x_i\beta + \sigma\epsilon_i)$ and $\mu_i = exp(x_i\beta + \sigma(sinh(\theta\epsilon_i)/\theta))$ for model II or III, respectively. In the considered application, however, evaluating the marginal effects at $\epsilon_i = 0$ gives almost the same results and eases the calculation of standard errors. Therefore I report the marginal effects evaluated at $\epsilon_i = 0$.[3] The reported marginal effects of binary regressors are the relative mean differences between groups.

3.3 Application

In this section, I use the data studied by Pohlmeier and Ulrich (1995) to illustrate the model extensions described above. The aim of their study was to emphasize that the decision to contact a physician is a two-part decisionmaking process. The importance of this two-part process was strengthened by a former institutional setting of the statutory health insurance in Germany: "[O]nce [the patient] has submitted the sickness voucher to his physician of choice, medical services for the relevant quarter are supplied by this physician only" (Pohlmeier and Ulrich, 1995, my insertions). If the patient wanted to visit another doctor in the relevant quarter, he needed a referral from the doctor whom he visited first. The physicians thus were "gatekeepers" in the statutory health insurance system, which covered more than 90% of those living in Germany.

[3] Table 3.3 in Appendix A.3 shows the marginal effects based on quadrature.

Table 3.1: Descriptive statistics

		$GP \geq 0$		$GP \geq 1$	
		Mean	SD	Mean	SD
Dependent variable					
GP	Number of visits to a general practitioner in the last three months	1.325	3.235	3.178	4.383
Explanatory variables					
Age	Age in years * 10^{-1}	3.989	1.113	4.122	1.148
Age^2	Age^2 * 10^{-3}	1.715	0.898	1.831	0.940
HH-income	Monthly net household income * 10^{-4}	0.336	0.217	0.322	0.221
Education	Years of education beyond the age of 16	1.985	3.128	1.630	2.819
Physician density	Physicians per 100,000 residents	0.260	0.049	0.258	0.044
Unemployment	Duration of unemployment in months in the last year	0.240	1.300	0.171	1.064
Female		0.368	0.482	0.394	0.489
Marital status	1 = single; 0 = other	0.173	0.378	0.147	0.354
Chronic illness	Chronic illness for at least one year	0.230	0.421	0.315	0.465
Private insurance	Private insurance in the last year	0.087	0.282	0.065	0.246
Heavy labor job	Heavy physical work (agree completely=1)	0.189	0.392	0.218	0.413
Stress	Job involves a high level of stress (agree completely=1)	0.260	0.439	0.280	0.449
Job variety	Job offers a lot of variety (agree completely=1)	0.517	0.500	0.488	0.500
Self-determining	Individual can plan his job tasks (agree completely=1)	0.363	0.481	0.340	0.474
Control	Work performance is strictly controlled (agree completely=1)	0.170	0.376	0.192	0.394
Population I	Place of residence has less than 5,000 residents = 1	0.124	0.330	0.144	0.351
Population II	Place of residence has between 5,000-20,000 residents = 1	0.247	0.431	0.283	0.451
Population III	Place of residence has between 20,000-100,000 residents = 1	0.275	0.446	0.269	0.443
Hospitalized	More than 7 days hospitalized in previous year = 1	0.074	0.261	0.109	0.312
Sick leave	Missed more than 14 work days in previous year = 1	0.183	0.387	0.256	0.436
Disability	Degree of disability is greater than 20% = 1	0.058	0.234	0.081	0.274
Number of observations		5,096		2,125	

The dataset is a cross-section of 5,096 individuals from the 1985 German Socio-Economic Panel. The authors separately analyzed the number of visits to a general practitioner (GP) and the number of visits to a specialist in the last three months. In both cases they used a zero-truncated NB-1 in the second part of the hurdle model. Using the same dataset, Santos Silva and Windmeijer (2001) showed that an important assumption of the hurdle model, namely the assumption of a single illness spell, is violated in the case of visits to specialists. Therefore I only use visits to GPs to illustrate the model extensions discussed in the previous section. Table 3.1 shows the variables used in the analysis with descriptive statistics for the full and the truncated sample.

Table 3.2 presents the relative marginal effects for the models discussed in this chapter. The vector of explanatory variables is the same as in Pohlmeier and Ulrich (1995).[4] Relative marginal effects are reported only for those variables that are significant at the 10% level in at least one of the specifications. Within the negative binomial models, the zero-truncated NB-P model has the greatest likelihood. The additionally estimated parameter P is around 1.33 and significantly different from 1 and 2. Thus the zero-truncated NB-1 and NB-2 models are rejected in favor of the NB-P model. The relative marginal effects of the NB-P model tend to be larger in magnitude than the NB-1 results reported by Pohlmeier and Ulrich (1995). In particular the effects of physician density and private insurance increase distinctly in the NB-P model and are both now clearly significant.

The specifications based on the Poisson log-normal distribution fit the data even better than its negative binomial alternatives. In addition, a Vuong (1989) test is performed to select between the truncated NB-P model and the flexible version of the truncated Poisson log-normal model more formally. Since both models are overlapping, I have first tested whether they overlap in this particular application. The additional parameters are, however, significantly different from the values which would reduce them to a standard truncated Poisson model. The Vuong test statistic for the flexible

[4] The variation in some of these variables is probably endogenous. The effect of private insurance, for example, might be partly due to a selection effect. Therefore, in what follows I will not reveal causal mechanisms. The aim of the following discussion is to show how model choice can affect the estimates in the analysis of overdispersed count data.

Table 3.2: Relative marginal effects

	Zero-truncated				
	Negative binomial			Poisson log-normal	
	NB 1	NB 2	NB P	Standard	Flexible
	(1)	(2)	(3)	(4)	(5)
Education	-0.002	-0.016	-0.016	-0.012	-0.012
	(0.010)	(0.009)	(0.011)	(0.007)	(0.008)
Physician density	0.548	1.129	1.003	0.842	0.839
	(0.289)	(0.656)	(0.450)	(0.415)	(0.433)
Female	0.041	0.050	0.060	0.091	0.100
	(0.035)	(0.056)	(0.047)	(0.039)	(0.040)
Chronic illness	0.333	0.476	0.463	0.399	0.403
	(0.141)	(0.099)	(0.110)	(0.069)	(0.074)
Private insurance	-0.073	-0.198	-0.195	-0.111	-0.105
	(0.047)	(0.063)	(0.075)	(0.054)	(0.056)
Heavy labor job	0.070	0.218	0.152	0.104	0.093
	(0.046)	(0.085)	(0.065)	(0.052)	(0.055)
Self-determining	-0.074	-0.099	-0.119	-0.105	-0.109
	(0.034)	(0.050)	(0.042)	(0.033)	(0.033)
Sick leave	0.225	0.408	0.376	0.325	0.328
	(0.101)	(0.106)	(0.096)	(0.071)	(0.071)
δ	4.082	5.490	5.347	—	—
	(0.340)	(1.886)	(0.952)		
P	1.000	2.000	1.327	—	—
	(fixed)	(fixed)	(0.069)		
σ	—	—	—	0.982	0.885
				(0.030)	(0.043)
θ	—	—	—	—	0.380
					(0.051)
Log-likelihood	-3,945.08	-3,931.00	-3,921.16	-3,886.16	-3,881.90

Dependent variable: Number of visits to a GP given any use of GP services.

Models also account for the other covariates displayed in table 3.1.

Standard errors are obtained by the delta rule.

Vuong test statistic for (5) versus (3): 3.72. H_0 : *both models are equivalent.* The test statistic is standard normally distributed.

version of the truncated Poisson log-normal model against the NB-P model is 3.72 and hence the null hypothesis is rejected in favor of the specification based on the Poisson log-normal distribution. The relative marginal effects of the truncated Poisson log-normal models are mostly in between the NB-1 and NB-P results. While the gender difference is insignificant in the negative binomial models, there is a significant difference between men and women according to the Poisson log-normal models.

3.4 Conclusion

Hurdle models based on the zero-truncated Poisson log-normal distribution are rarely used in applied work, although they incorporate some advantages compared with their negative binomial alternatives. These models are appealing from a theoretical point of view and, additionally, perform much better in many applications.

Recent developments in the count data literature make it possible to relax commonly imposed assumptions of hurdle models. I use these techniques to propose two extensions of hurdle models. Both extensions nest the models that have been estimated previously. This allows one to simply test these models by appropriate parametric restrictions. An example from health economics shows that the more flexible models can lead to distinctly different marginal effects.

Appendices

A.1 Gauss-Hermite Quadrature

When there is no analytical solution of an integral, it can be approximated numerically using e.g. Gauss-Hermite quadrature. The likelihood contributions that are actually used in the estimation can be derived by:

$$\begin{aligned} L_i^{II,III} &= \int_{-\infty}^{\infty} f(y_i|\mathbf{x}_i\boldsymbol{\beta} + \sigma d(\epsilon_i))\phi(\epsilon_i)\mathrm{d}\epsilon_i \\ &= \int_{-\infty}^{\infty} f(y_i|\mathbf{x}_i\boldsymbol{\beta} + \sigma d(\epsilon_i))\frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}\epsilon_i^2)\mathrm{d}\epsilon_i \end{aligned}$$

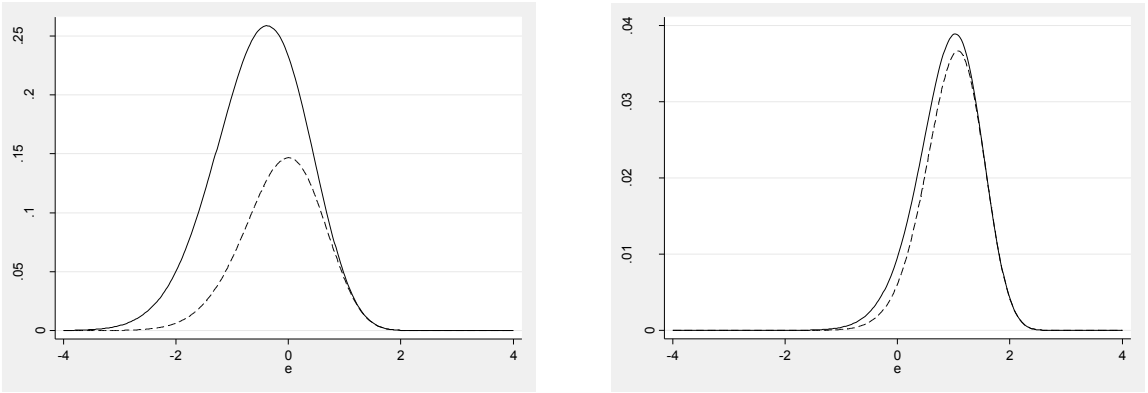
after a change of variable $\epsilon_i = \sqrt{2}\nu_i$,

$$\begin{aligned} L_i^{II,III} &= \int_{-\infty}^{\infty} f(y_i|\mathbf{x}_i\boldsymbol{\beta} + \sigma d(\sqrt{2}\nu_i))\frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}(\sqrt{2}\nu_i)^2)\sqrt{2}\mathrm{d}\nu_i \\ &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} f(y_i|\mathbf{x}_i\boldsymbol{\beta} + \sigma d(\sqrt{2}\nu_i))exp(-\nu_i^2)\mathrm{d}\nu_i \\ &\doteq \frac{1}{\sqrt{\pi}}\sum_{r=1}^{R} f(y_i|\mathbf{x}_i\boldsymbol{\beta} + \sigma d(\sqrt{2}\nu_r))\omega_r \end{aligned}$$

where $\doteq$ indicates the approximation.

A.2 Implementation in Stata

The models discussed in this chapter are estimated in Stata using adaptive Gauss-Hermite quadrature. Adaptive quadrature shifts and scales the quadrature points to place them under the peak of the integrand which most likely improves the approximation (compare section 6.3.2 in Skrondal and Rabe-Hesketh, 2004 for a detailed discussion). Many Stata commands such as `xtpoisson` implement an approach proposed by Liu and Pierce (1994). They argue that the mode of the integrand and the curvature at the mode can be used as shifting and scaling factors. Instead of calculating these factors, I use the corresponding values of the standard (untruncated) Poisson log-normal model to implement the adaptive quadrature. The reason for this is a built-in



(a) y=1, xb=0 (b) y=4, xb=0

Figure 3.1: Integrand of zero-truncated model (solid curve) and standard model (dashed curve)

command in Stata that only gives the corresponding values for the standard Poisson log-normal model. The integrand, however, is often very similar in both models, especially for high values of the dependent variable. This indicates that these values might also be good guesses for the scaling and shifting factors of the zero-truncated models. Figure 3.1 shows the integrands of the standard and zero-truncated Poisson log-normal model. In this example they are almost identical for higher values of the dependent variable.

A.3 Relative marginal effects using quadrature

Table 3.3: Relative marginal effects using quadrature

	Zero-truncated Poisson log-normal	
	Standard	Flexible
Education	-0.012	-0.012
	(0.007)	(0.007)
Physician density	0.843	0.838
	(0.459)	(0.452)
Female	0.091	0.100
	(0.037)	(0.036)
Chronic illness	0.405	0.408
	(0.061)	(0.058)
Private insurance	-0.111	-0.105
	(0.058)	(0.057)
Heavy labor job	0.105	0.093
	(0.057)	(0.057)
Self-determining	-0.105	-0.108
	(0.032)	(0.035)
Sick leave	0.327	0.329
	(0.064)	(0.066)

Models also account for the other covariates displayed in table 3.1.
Bootstrap standard errors based on 100 replications.
Flexible model converged in only 66% of the replications.

Chapter 4

Continuously updated GMM with many weak moment conditions: An application in labor economics[†]

4.1 Introduction

Endogeneity is a common phenomenon in applied econometrics and it generally prevents a causal interpretation of ordinary least squares regressions. The availability of valid instruments can solve this problem. One criterion for a valid instrument is sufficient correlation between the endogenous variable and the instrument. Instruments that do not fulfill this criterion are called weak. Weak instruments are not unusual in applied econometrics. Angrist and Krueger (1991)'s quarter-of-birth instrument is a famous example for a weak instrument. They used it to estimate individuals' returns to education. Yogo (2004) presented an example of weak instruments in macroeconomics. The poor performance of two-stage least squares (2SLS) estimation with weak instruments has been extensively discussed in Staiger and Stock (1997)'s seminal study. Thus, there is a need for an estimator with better properties than 2SLS in case of weak instruments.

The continuous updating estimator, for instance, could be such an estimator.

[†] Parts of this chapter have been published in Farbmacher (2011b)

Hansen *et al.* (1996) showed that the continuous updating estimator is typically less median biased than 2SLS. However, there are still some problems remaining. For instance, Guggenberger (2005, 2008) mentioned that its criterion function is difficult to optimize, which can lead to spurious results. Moreover, the dispersion of the estimator is often tremendously high, which may complicate the interpretation of the results (see e.g. Hansen *et al.*, 1996 and Hausman *et al.*, 2011). Furthermore, the usual formula for the variance estimator seems to understate the true variance especially when identification is weak (see e.g. Han and Phillips, 2005 or Newey and Windmeijer, 2009). For this reason, Newey and Windmeijer (2009) proposed a new variance estimator for generalized empirical likelihood (GEL) estimators. It addresses the problem that usual standard errors are too small when there are many weak instruments. In Monte Carlo simulations they apply the new variance estimator for a member of GEL estimators, namely the continuous updating estimator. t-statistics based on the new variance estimator have nearly correct size in a wide range of cases. I replicate their simulations for the linear model using the continuous updating estimator with usual standard errors and the many weak instruments standard errors of Newey and Windmeijer. Moreover, I report results for a wider range of the parameters involved. For comparison the jackknife instrumental variable estimator (JIVE2) of Angrist *et al.* (1999) and two-stage least squares have also been estimated. Finally, I re-estimate Angrist and Krueger (1991)'s returns to education using these estimators and additionally compare them to the limited information maximum likelihood (LIML) estimator.

An additional finding of my Monte Carlo simulations is that two-stage least squares estimates are particularly poor starting values for the continuous updating estimator, especially when the sample size is small and/or the identification is weak. A potential reason for this is the likely presence of local optima and the fact that the continuous updating estimator often converges to these local optima if they are close to the 2SLS estimates. The nearness of the local optima to the 2SLS estimates then suggest that the performance of the continuous updating estimator is also affected by the properties of 2SLS. Furthermore, this study shows that extreme estimates of the continuous updating estimator, which are often reported in Monte Carlo simulations, are more likely to be a failure of the optimization routine than a property of the continuous

updating estimator.

This chapter is organized as follows. The next section describes the continuous updating estimator and the new variance estimator proposed by Newey and Windmeijer (2009). Section 3 first explains the design of the conducted Monte Carlo simulation and discusses the expected performance of 2SLS under these conditions. Then, it provides the simulation results. Section 4 applies the new variance estimator to the Angrist and Krueger (1991) data set. Section 5 concludes.

4.2 Continuously updated GMM

The model considered here is linear and homoskedastic with

$$y_i = x_i'\beta_0 + u_i \tag{4.1}$$

where y_i is a scalar, x_i is a $l \times 1$ vector of explanatory variables and β_0 is a $l \times 1$ vector of true parameters satisfying the moment conditions

$$E(z_i u_i) = 0 \tag{4.2}$$

where z_i is a $m \times 1$ vector of instruments. Denote

$$g_i(\beta) = z_i(y_i - x_i'\beta), \quad \hat{g}(\beta) = n^{-1}\sum_{i=1}^{n} g_i(\beta), \tag{4.3}$$

$$\hat{\Omega}(\beta) = n^{-1}\sum_{i=1}^{n} z_i z_i'(y_i - x_i'\beta)^2 \tag{4.4}$$

where β is a $l \times 1$ parameter vector to be estimated. Hansen (1982)'s two-step generalized method of moments (GMM) estimator is given by

$$\ddot{\beta} = \arg\min_{\beta} \ddot{Q}(\beta), \quad \ddot{Q}(\beta) = \hat{g}(\beta)'\hat{W}\hat{g}(\beta)/2, \quad \hat{W} = \hat{\Omega}(\dot{\beta})^{-1} \tag{4.5}$$

with $\dot{\beta}$ is obtained using a suboptimal choice of the weighting matrix. Then, $\hat{W}$ minimizes the asymptotic variance of $\ddot{\beta}$. The continuous updating estimator proposed by Hansen *et al.* (1996) simultaneously minimizes over β in the sample analogue of the moment conditions and the weighting matrix, that is

$$\hat{\beta} = \arg\min_{\beta} \hat{Q}(\beta), \quad \hat{Q}(\beta) = \hat{g}(\beta)'\hat{\Omega}(\beta)^{-1}\hat{g}(\beta)/2. \tag{4.6}$$

The jth element of the first derivative of $\hat{Q}(\beta)$ is

$$\frac{\partial \hat{Q}(\beta)}{\partial \beta_j} = \frac{\partial \hat{g}(\beta)'}{\partial \beta_j}\hat{\Omega}(\beta)^{-1}\hat{g}(\beta) - \hat{g}(\beta)'\hat{\Omega}(\beta)^{-1}\hat{\Lambda}_j(\beta)\hat{\Omega}(\beta)^{-1}\hat{g}(\beta) \tag{4.7}$$

and the jkth element of the second derivative of $\hat{Q}(\beta)$ is

$$\begin{aligned}\frac{\partial^2 \hat{Q}(\beta)}{\partial \beta_j \beta_k} = {} & \frac{\partial^2 \hat{g}(\beta)}{\partial \beta_j \partial \beta_k}\hat{\Omega}(\beta)^{-1}\hat{g}(\beta) + \frac{\partial \hat{g}(\beta)'}{\partial \beta_j}\hat{\Omega}(\beta)^{-1}\frac{\partial \hat{g}(\beta)}{\partial \beta_k} \\ & - 2\frac{\partial \hat{g}(\beta)'}{\partial \beta_j}\hat{\Omega}(\beta)^{-1}\hat{\Lambda}_k(\beta)\hat{\Omega}(\beta)^{-1}\hat{g}(\beta) - 2\frac{\partial \hat{g}(\beta)'}{\partial \beta_k}\hat{\Omega}(\beta)^{-1}\hat{\Lambda}_j(\beta)\hat{\Omega}(\beta)^{-1}\hat{g}(\beta) \\ & - \hat{g}(\beta)'\hat{\Omega}(\beta)^{-1}\frac{\partial \hat{\Lambda}_j(\beta)}{\partial \beta_k}'\hat{\Omega}(\beta)^{-1}\hat{g}(\beta) \\ & + 4\hat{g}(\beta)'\hat{\Omega}(\beta)^{-1}\hat{\Lambda}_k(\beta)\hat{\Omega}(\beta)^{-1}\hat{\Lambda}_j(\beta)\hat{\Omega}(\beta)^{-1}\hat{g}(\beta)\end{aligned} \tag{4.8}$$

where in the considered linear model

$$\frac{\partial \hat{g}(\beta)}{\partial \beta_j} = n^{-1}\sum_{i=1}^{n} -z_i x_{ij}, \quad \frac{\partial^2 \hat{g}(\beta)}{\partial \beta_j \partial \beta_k} = 0, \tag{4.9}$$

$$\hat{\Lambda}_j(\beta) = \frac{\partial \hat{\Omega}(\beta)}{\partial \beta_j}/2 = n^{-1}\sum_{i=1}^{n} -z_i z_i' x_{ij}(y_i - x_i'\beta), \tag{4.10}$$

$$\frac{\partial \hat{\Lambda}_j(\beta)}{\partial \beta_k} = n^{-1}\sum_{i=1}^{n} z_i z_i' x_{ij} x_{ik}. \tag{4.11}$$

The usual formula of the variance-covariance matrix is

$$\hat{V}_c = \left(\hat{G}'\hat{\Omega}(\hat{\beta})^{-1}\hat{G}\right)^{-1}, \quad \text{with} \quad \hat{G} = \frac{\partial \hat{g}(\hat{\beta})}{\partial \beta} = \left(\frac{\partial \hat{g}(\hat{\beta})}{\partial \beta_1}, \ldots, \frac{\partial \hat{g}(\hat{\beta})}{\partial \beta_l}\right). \tag{4.12}$$

According to the usual formula, the asymptotic variance of $\hat{\beta}$ is $\hat{V}_c/n$. Standard errors based on this matrix are often too small in applications with many instruments (see e.g. Han and Phillips, 2005). t-statistics based on the new variance estimator of Newey and Windmeijer (2009) have, however, nearly correct size in a wide range of cases. The proposed variance estimator of Newey and Windmeijer (2009) is

$$\hat{V} = \hat{H}^{-1}\hat{D}(\hat{\beta})'\hat{\Omega}(\hat{\beta})^{-1}\hat{D}(\hat{\beta})\hat{H}^{-1} \tag{4.13}$$

where $\hat{H}$ is a Hessian term containing the elements defined in equation (4.8) and $\hat{D}_j(\hat{\beta})$ is the jth column of $\hat{D}(\hat{\beta})$:

$$\hat{D}_j(\hat{\beta}) = \frac{\partial \hat{g}(\hat{\beta})}{\partial \beta_j} - \hat{\Lambda}_j(\hat{\beta})\hat{\Omega}(\hat{\beta})^{-1}\hat{g}(\hat{\beta}). \tag{4.14}$$

The asymptotic variance of $\hat{\beta}$ is $\hat{V}/n$. Newey and Windmeijer (2009, p.692) noted that $\hat{V}$ converges to $\hat{V}_c$ when m is fixed and identification is strong. In this case, $\hat{g}(\hat{\beta})$ converges in probability to zero and the second term in equation (4.14) vanishes. Furthermore, the Hessian term is then equal to the numerator in equation (4.13).

4.3 Monte Carlo simulation

4.3.1 Simulation design

The design of the Monte Carlo experiment is

$$\begin{aligned} y_i &= \alpha + \beta_x x_i + u_i \\ x_i &= z_i'\pi + v_i \\ u_i &= \rho v_i + \sqrt{1-\rho^2} w_i \\ v_i \sim N(0,1),\ \ w_i \sim N(0,1),\ z_i \sim N(0,I), &\quad \pi = \sqrt{\frac{CP}{mn}}\iota_m \end{aligned}$$

where ι_m is a m-vector of ones. x has no causal effect on y (i.e. $\beta_x = 0$) and the constant is set to zero as well. Firstly, I use the same parameters as Newey and Windmeijer (2009) to replicate some of their results. The sample size n is 200 in their simulation; the concentration parameter CP is equal to 10, 20 or 35; the degree of endogeneity ρ is equal to 0.3, 0.5 or 0.9; the number of instruments m is 3 or 15; the number of replications is 10,000. Secondly, I extend their analysis using a wider range of the simulation parameters. Throughout the continuous updating estimator is estimated with Mata's optimize function. The gradient and the Hessian are calculated analytically.

I use Hahn and Hausman (2002)'s expression for the approximate finite sample bias of the 2SLS estimator as a theoretical guideline for the interpretation of the Monte Carlo results:[1]

$$\begin{aligned} E(b_{2SLS}) - \beta &\approx \frac{cov(u,v)}{\frac{n}{m}\pi' R \pi + var(v)} \\ &= \frac{\rho}{\frac{CP}{m^2}\iota_m' R \iota_m + 1} \end{aligned} \tag{4.15}$$

[1] Bun and Windmeijer (2011) recently developed an alternative bias approximation. They compared it with the Hahn and Hausman approximation and showed that the latter may be inaccurate for modest m.

where $R = E[z'z/n]$ and $z = (z_1' \cdots z_n')'$. The equality in the second line follows from the choice of the simulation parameter π. Now, n is no longer in the denominator implying that increasing n does not reduce bias. Thus we cannot expect unbiased estimates from 2SLS under the asymptotics where $\pi = \sqrt{a/n}$, even if the sample size is large as, for instance, in Angrist and Krueger (1991). This has been shown by Staiger and Stock (1997) and more recently by Hahn and Hausman (2002). According to equation (4.15), the approximate bias of 2SLS is increasing in ρ and m and decreasing in CP/m, which is the population F-statistic from the first stage regression.

For comparison purposes, Table 4.5 in the appendix replicates the results from Newey and Windmeijer (2009). It reports median bias and rejection frequencies of Wald tests for the null hypothesis $H_0 : \beta_x = 0$. These are very similar to Newey and Windmeijer's Monte Carlo results. Figure 4.2 to 4.4 report the results graphically for a wider range of the simulation parameters and for different sample sizes ($n = 25, 100, 800$). These results are discussed in section 4.3.3. Only one parameter is varied at a time. The other two values are fixed at a certain level ($\rho = 0.9$, $m = 5$, $CP = 10$). When m is varied, I additionally hold the first stage F-statistic fixed as the number of instruments increases. It implies that the set of instruments is equally strong with increasing m. Setting $\pi = \sqrt{\frac{CP}{n}}\iota_m$ the approximate bias becomes

$$E(b_{2SLS}) - \beta \approx \frac{\rho}{\frac{CP}{m}\iota_m' R \iota_m + 1} \tag{4.16}$$

which does now not depend on m given that the first stage F-statistic is fixed.

4.3.2 Continuous updating estimator and starting values

In the next section, I compare the theoretical considerations about the approximate mean bias of 2SLS to the simulation results. However, instead of the mean bias, I report the median bias from the simulations. The reason for this is the "no-moments problem" of the continuous updating estimator that has been reported extensively in the literature. Guggenberger (2005, 2008), for instance, evaluated the criterion function of the continuous updating estimator using a grid over the parameter space, and his

results do therefore not rely on a minimization routine. He found that the continuous updating estimator takes on large values with a much higher probability than 2SLS. This leads to a substantial dispersion in the estimates. In a recent Monte Carlo simulation Hausman *et al.* (2011) compared the performance of different estimators – among them the continuous updating estimator. They used a derivative based optimization which could potentially not converge. In their results some extreme estimates of β_x completely preclude the interpretation of the first two moments of the continuous updating estimator.

The following simulation suggests that these extreme estimates are, however, from estimations that did not converge. I provide two different sets of estimates from the continuous updating estimator. The first set of estimates is based on just one optimization in which the starting values are obtained from 2SLS, which might be considered as a natural choice of the starting values.[2] On the other hand, for the second set of estimates I use five fixed starting values (FSV) for β_x (-2,-1,0,1,2)[3] and additionally the 2SLS estimates. The results of the continuous updating estimator in the second set are then obtained from the optimization with the lowest criterion function $\hat{Q}(\beta)$ given that the optimization converged. Table 4.1 shows measures of central tendency and dispersion for both sets of estimates and additionally the fraction of simulation replications that converged. It is interesting to see that some extreme estimates of β_x completely preclude the interpretation of the mean and the variance whenever the results contain some estimates from optimizations that did not converge. On the other hand, when I focus on converged optimizations, I can eliminate these extreme estimates (compare last two columns in Table 4.1). This may indicate that extreme estimates of β_x are rather a failure of the optimization routine than a property of the continuous updating estimator. Overall, the probability that the optimization converges increases when I additionally use fixed starting values, and it even reaches 100% in the simulations with 100 or 800 observations. Nevertheless, the variance and in some cases also the mean are still tremendously high, pointing to the "no-moments problem".

[2] For instance, a user-written command in Stata uses the 2SLS estimates as starting values (see Baum *et al.* 2007).

[3] The starting value for the constant is set to zero.

Table 4.1: Continuous updating estimator and starting values

			overall						given convergence		
SV	CP	n	med.	mean	IQR	NDR	var.	conv.	med.	mean	var.
2SLS	1	25	0.637	-7.2E+26	0.808	5.938	5.1E+57	95.40%	0.643	0.296	152.1
FSV & 2SLS	1	25	0.560	-9.4E+08	1.147	7.858	1.6E+22	99.96%	0.560	0.197	3337
2SLS	1	100	0.494	3.5E+16	1.020	8.850	1.6E+37	94.90%	0.494	0.143	694.9
FSV & 2SLS	1	100	0.470	-0.341	1.104	7.753	3,005	100.0%	–	–	–
2SLS	1	800	0.469	-2.3E+24	1.059	8.240	5.3E+52	97.77%	0.472	0.569	577.7
FSV & 2SLS	1	800	0.471	0.260	1.072	7.858	991.8	100.0%	–	–	–
2SLS	10	25	0.073	-2.3E+14	0.565	2.189	3.4E+32	98.80%	0.076	-0.146	10.24
FSV & 2SLS	10	25	0.037	-9.3E+09	0.624	2.727	5.1E+23	99.98%	0.037	-0.195	150.1
2SLS	10	100	-0.001	1.9E+13	0.502	1.693	3.7E+30	99.46%	-0.000	-0.180	5.904
FSV & 2SLS	10	100	-0.001	-0.028	0.501	1.676	72.61	100.0%	–	–	–
2SLS	10	800	0.002	-2.3E+09	0.478	1.504	8.0E+22	99.91%	0.002	-0.180	3.972
FSV & 2SLS	10	800	0.003	-0.170	0.478	1.504	9.544	100.0%	–	–	–

Linear model with constant and under homoskedasticity: $m = 5$; $\rho = 0.9$; Number of replications: 10,000.

If 50 iterations are reached without finding a minimum, then non-convergence is declared.

Abbreviations: FSV: Fixed starting values (-2,-1,0,1,2); SV: starting values; med.: median; var.: variance;

conv.: converged; IQR: interquartile range; NDR: nine decile range, that is, the 0.95 quantile minus the 0.05 quantile.

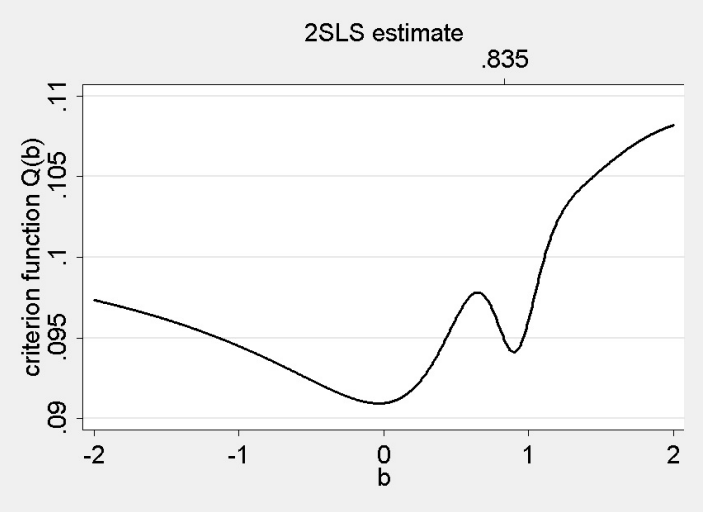


Figure 4.1: Criterion function of the continuous updating estimator with multiple optima ($m = 5$, $\rho = 0.9$, $CP = 1$, n=25)

Adding fixed starting values appears to affect the median bias as well. It tends to be larger in the first set of estimates in which the starting values are obtained from 2SLS. This effect is particularly pronounced in the simulations with 25 observations. But it can also be observed in larger samples given that the identification is extremely weak. For instance, using Hausman *et al.* (2011)'s simulation with $n = 400$, $\rho = 0.3$, $m = 50$ and $CP = 8$, the median bias is 0.136 with 2SLS starting values and 0.091 with fixed and 2SLS starting values. This result is not reported in the table.

Figure 4.1 suggests a potential reason for the difference in the median bias. It displays the criterion function of an example data set which has been evaluated for all values of β_x between -2 and 2 using a step size of 0.01.[4] In Figure 4.1 you can see a local minimum which is very close to the 2SLS estimate marked on the upper axis. In this example the continuous updating estimator ends up in this local minimum when I use the 2SLS estimate as starting value. Such a situation can be observed very often when the sample size is small and/or the identification is weak.

I performed an additional Monte Carlo simulation using the simple example discussed in the previous paragraph to analyze the effect of starting values on the reported

[4] Such a brute force approach has also been suggested by Hansen *et al.* (1996) and Guggenberger (2005, 2008). It finds the true minimizer given that it lies between -2 and 2. However, since this approach is computationally demanding in higher dimensions, I revert to a model without constant.

performance of the continuous updating estimator more systematically. Again, the criterion function has been evaluated using a fine grid with step size of 0.01, but now for all values of β_x in between -5 and 5. Using this range, there was at least one minimum in almost all simulation replications. The minimum of all these optima is denoted as global optimum (β_{global}).[5] All other optima are local (β_{local}). The variables ζ and C_{SV} shall be defined as follows:

$$\zeta = \mathbf{1}\left(|\hat{\beta}_{2SLS} - \beta_{local}| < |\hat{\beta}_{2SLS} - \beta_{global}|\right), \tag{4.17}$$

$$C_{SV} = \mathbf{1}\left(\left|\frac{\hat{\beta}^{SV}_{CUE} - \beta_{local}}{(\hat{\beta}^{SV}_{CUE} + \beta_{local})/2}\right| < 0.05\right) \tag{4.18}$$

where $\hat{\beta}_{2SLS}$ is the 2SLS estimator and $\hat{\beta}^{SV}_{CUE}$ denotes the continuous updating estimator that is supposed to depend on the choice of the starting values (SV). The first variable, ζ, is equal to one when $\hat{\beta}_{2SLS}$ is closer to the local than to the global optimum. The second variable, C_{SV}, indicates the convergence of the continuous updating estimator to a local optimum where convergence is assumed if the relative difference between the continuous updating estimator and the local optimum is smaller than 5%.

Table 4.2 shows the simulation results of the continuous updating estimator for both sets of starting values, and additionally provides the results from the 2SLS estimator and the global optimum. Throughout, the presence of local optima decreases with the sample size and the strength of the identification. Local optima are particularly likely in simulations with both small sample size and low concentration parameter. For instance, 44% of the replications exhibit local optima in the simulation with $n = 25$ and $CP = 1$. Table 4.2 also shows the fraction of replications in which the continuous updating estimator converges to a local optimum, i.e. $Pr(C_{SV} = 1)$. Overall, this probability increases considerably when only 2SLS is used as starting value compared to using fixed starting values as well (0.253 vs. 0.001 in the example with $n = 25$ and $CP = 1$). This effect is particularly pronounced when the 2SLS estimates are closer to the local than to the global optimum, i.e. $\zeta = 1$ (0.540 vs. 0.001 in case of $n = 25$

[5] Of course, it is just the global optimum in between -5 and 5.

and $CP = 1$). On the other hand, the 2SLS estimates work well when they are nearer to the global optimum, i.e. $\zeta = 0$ (0.017 vs. 0.001 in the simulation with $n = 25$ and $CP = 1$). While it is not surprising that the continuous updating estimator converges more likely to a local optimum when the chosen starting values are close to this local optimum, it is particularly problematic in this case because 2SLS is generally more biased than the continuous updating estimator. The nearness of the local optimum to the 2SLS estimate may then suggest that the reported performance of the continuous updating estimator is affected by the properties of 2SLS.

The last three columns in Table 4.2 support this suggestion. They show the median bias and the nine decile range for the continuous updating estimator with 2SLS as starting values and, on the other hand, with fixed starting values and 2SLS. Additionally, I report the global optima from the grid evaluation and the estimates from 2SLS. The median bias and the nine decile range of the continuous updating estimator turn out to be almost equally large for both sets of starting values when the continuous updating estimator with 2SLS starting values does not converge to a local optimum ($C_{2SLS} = 0$). On the other hand, they are distinctly different when the continuous updating estimator with 2SLS starting values converges to a local optimum ($C_{2SLS} = 1$). As expected, choosing 2SLS as starting values increases the reported bias but, on the other hand, also decreases the reported dispersion of the continuous updating estimator. The results from the grid evaluation are always very similar to the results from the continuous updating estimator with fixed starting values. On the contrary, the continuous updating estimator with 2SLS starting values performs as poor as the 2SLS estimator given that it converges to a local optimum. This confirms the importance of trying different starting values not only in real-world data sets but also in Monte Carlo simulations. The impact of this on the unconditional results depends essentially on the frequency of local optima. For instance, while a difference in the median bias can also be observed in larger samples and with stronger identification, this effect does not change the unconditional median bias since the frequency of local optima is almost zero in these examples.[6]

[6] This can also be observed in Table 4.1. Note, however, that the estimations in Table 4.1 are not completely comparable because they include a constant.

Table 4.2: Continuous updating estimator (CUE), starting values and local optima

				Local optima*				$Pr(C_{SV}=1)$ given			Median bias/nine decile range given the presence of a local optimum and given		
$\hat{\beta}$	SV	CP	n	0	1	2+	$Pr(\zeta=1)$	$\zeta=0$	$\zeta=1$	overall	$C_{2SLS}=0$	$C_{2SLS}=1$	overall
2SLS	–	1	25								0.765/0.745	0.816/0.588	0.781/0.707
CUE	2SLS	1	25	0.520	0.407	0.033	0.451	0.017	0.540	0.253	0.650/2.528	0.831/1.065	0.717/2.198
CUE	FSV & 2SLS	1	25					0.001	0.001	0.001	0.644/2.895	0.285/6.943	0.614/4.471
Global	–	1	25								0.650/2.810	0.300/6.830	0.620/4.350
2SLS	–	1	100								0.848/0.423	0.867/0.334	0.854/0.397
CUE	2SLS	1	100	0.832	0.094	0.000	0.479	0.007	0.618	0.300	0.728/1.952	0.883/0.846	0.804/1.638
CUE	FSV & 2SLS	1	100					0.000	0.001	0.001	0.733/1.959	0.227/6.790	0.690/4.080
Global	–	1	100								0.740/1.960	0.240/6.790	0.700/4.080
2SLS	–	1	800								0.895/0.149	0.894/0.115	0.894/0.142
CUE	2SLS	1	800	0.910	0.010	0.000	0.500	0.004	0.665	0.335	0.917/1.435	0.904/0.809	0.911/1.212
CUE	FSV & 2SLS	1	800					0.000	0.000	0.000	0.915/1.401	0.255/7.152	0.858/4.512
Global	–	1	800								0.925/1.400	0.265/7.160	0.870/4.510
2SLS	–	10	25								0.322/0.657	0.421/0.582	0.336/0.662
CUE	2SLS	10	25	0.611	0.352	0.025	0.317	0.008	0.401	0.133	0.103/1.665	0.489/1.247	0.159/1.689
CUE	FSV & 2SLS	10	25					0.000	0.001	0.001	0.097/2.018	-0.503/6.535	0.067/2.939
Global	–	10	25								0.110/1.970	-0.480/6.370	0.080/2.830
2SLS	–	10	100								0.487/0.544	0.653/0.404	0.504/0.548
CUE	2SLS	10	100	0.980	0.010	0.000	0.361	0.003	0.321	0.118	0.107/1.453	0.763/0.785	0.154/1.609
CUE	FSV & 2SLS	10	100					0.000	0.000	0.000	0.107/1.440	-0.779/6.106	0.069/2.243
Global	–	10	100								0.120/1.440	-0.770/6.105	0.080/2.240
CUE	2SLS	10	800	0.995	0.000	0.000	–†						

Linear model without constant and under homoskedasticity: $m=5$; $\rho=0.9$; Number of replications: 50,000.

In the estimation non-convergence is declared if 50 iterations are reached without finding a minimum.

Local and global optima are obtained using a grid over the parameter space (from -5 to 5; step size 0.01). * The difference to 1 is the fraction of replications in which there was no minimum between -5 and 5. FSV: Fixed starting values (-2,-1,0,1,2); SV: starting values. † No observations.

Using the same line of arguments, it follows that all estimates, which are more biased than the continuous updating estimator, are generally poor starting values whenever local optima are present. Han and Phillips (2006), for example, used ordinary least squares estimates as starting values for the continuous updating estimator in their simulations. This, although not reported here, do not work either. A generalization of this approach to a multidimensional parameter space is in principle possible but cumbersome due to the computational burden.

4.3.3 Median bias and rejection frequencies

Figures 4.2, 4.3 and 4.4 show the simulation results for $n = 25$, $n = 100$ and $n = 800$, respectively.[7] The graphs show the relationships predicted by equations (4.15) and (4.16). The 2SLS median bias increases with the degree of endogeneity and decreases with the concentration parameter. Apart from the specifications that are close to the just-identified case, the median bias is also independent of the number of instruments as has been shown in equation (4.16). Moreover a comparison of the graphs shows, as expected, that increasing the sample size does not reduce the median bias of 2SLS under weak instruments asymptotics. The continuous updating estimator appears to perform slightly worse in the tiny sample with just 25 observations (Figure 4.2) than in Figure 4.3, where the sample size is 100, and Figure 4.4, where the sample size is 800. However, it clearly outperforms 2SLS and JIVE2 once the sample size is reasonably large (see Figures 4.3 and 4.4).

According to the simulation with the largest sample size, the continuous updating estimator is almost median unbiased and this result is, interestingly, independent of the degree of endogeneity. Obviously, the bias increases in the neighborhood of an unidentified model where the instruments are completely uninformative. Nevertheless, the continuous updating estimator becomes median unbiased once the concentration parameter is around 5 which implies a population first stage F-statistic of around 1. This is considerably lower than the rule of thumb for 2SLS, which is around 10.

This result is consistent with recent Monte Carlo results of Hansen *et al.* (2008).

[7] The continuous updating estimator has been estimated with different starting values (-2,-1,0,1,2) to prevent convergence problems and local optima.

They found that the use of LIML, which matches the continuous updating estimator in the case of a linear model and homoskedasticity (see e.g. Hausman *et al.*, 2011 and Guggenberger, 2005, 2008), is often adequate in situations where the F-statistic takes on values around one. For instance, in their simulation LIML is almost unbiased with $\rho = 0.8$, $m = 8$ and $CP = 8$.

While the continuous updating estimator is less median biased than 2SLS and JIVE2 in almost all situations, the rejection frequencies for $H_0 : \beta_x = 0$ with the usual standard errors (CUE) are far too high especially in small samples. In contrast to this, the rejection frequencies with the new variance estimator (CUEC) proposed by Newey and Windmeijer (2009) are often very close to the nominal level once the sample size is reasonably large.

According to Figures 4.3 and 4.4, the rejection frequency of 2SLS increases constantly with rising number of instruments although the sets of instruments are equally strong. This has already been shown in Staiger and Stock (1997)'s figure 3. The rejection frequency of the continuous updating estimator based on the new variance estimator is, however, close to the nominal size and independent of the number of instruments once the number of instruments is larger than 4. It performs also better than the continuous updating estimator with usual standard errors especially when the sample size is 100. Furthermore, it depends only slightly on the degree of endogeneity and is nearly level-correct for all values of the concentration parameter apart from situations in which the model is almost unidentified.

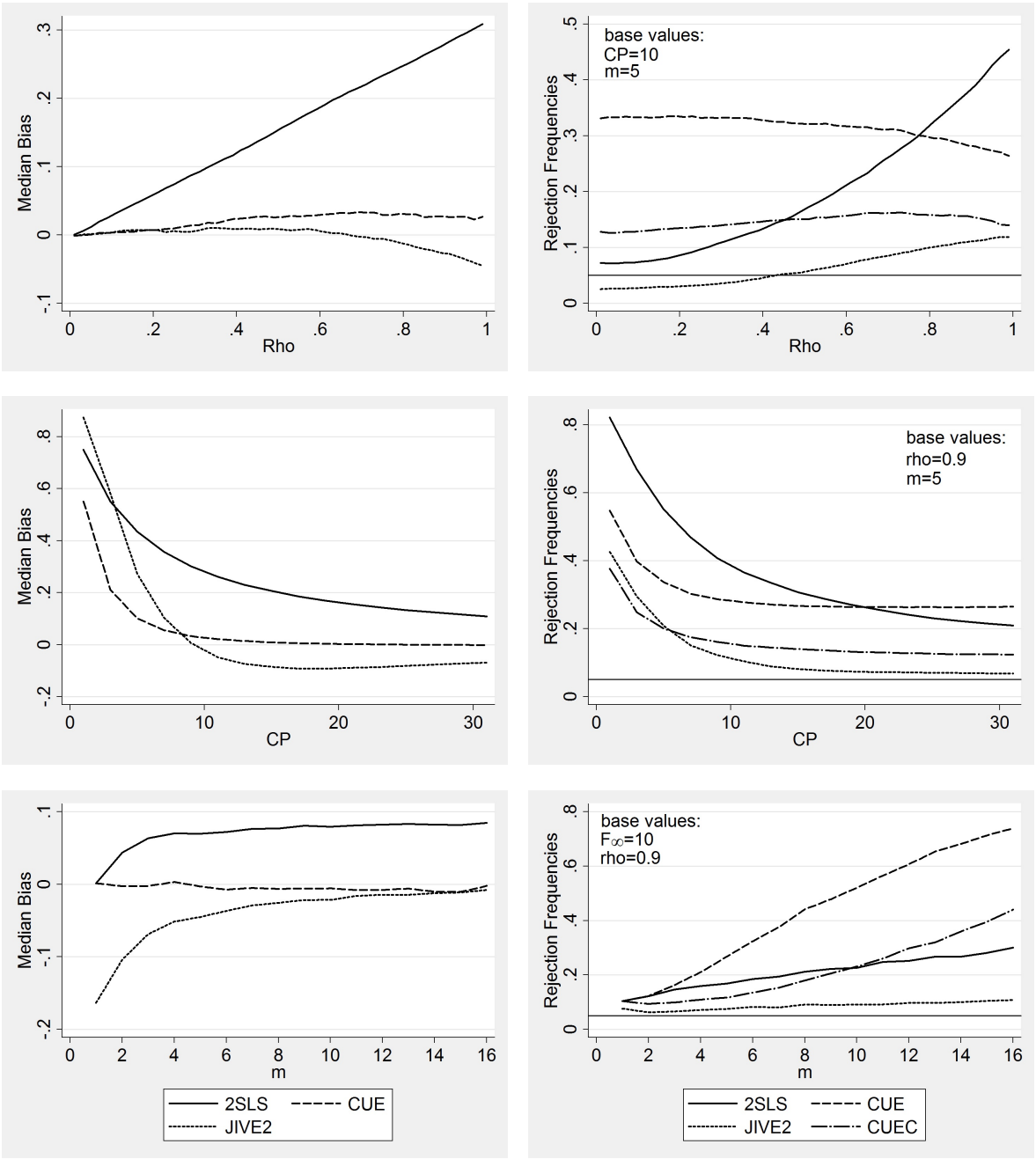


Figure 4.2: Median bias and rejection frequencies when $\boldsymbol{n = 25}$
(Step size: $\rho = 0.02$, $CP = 2$, $m = 1$; 10,000 replications each)

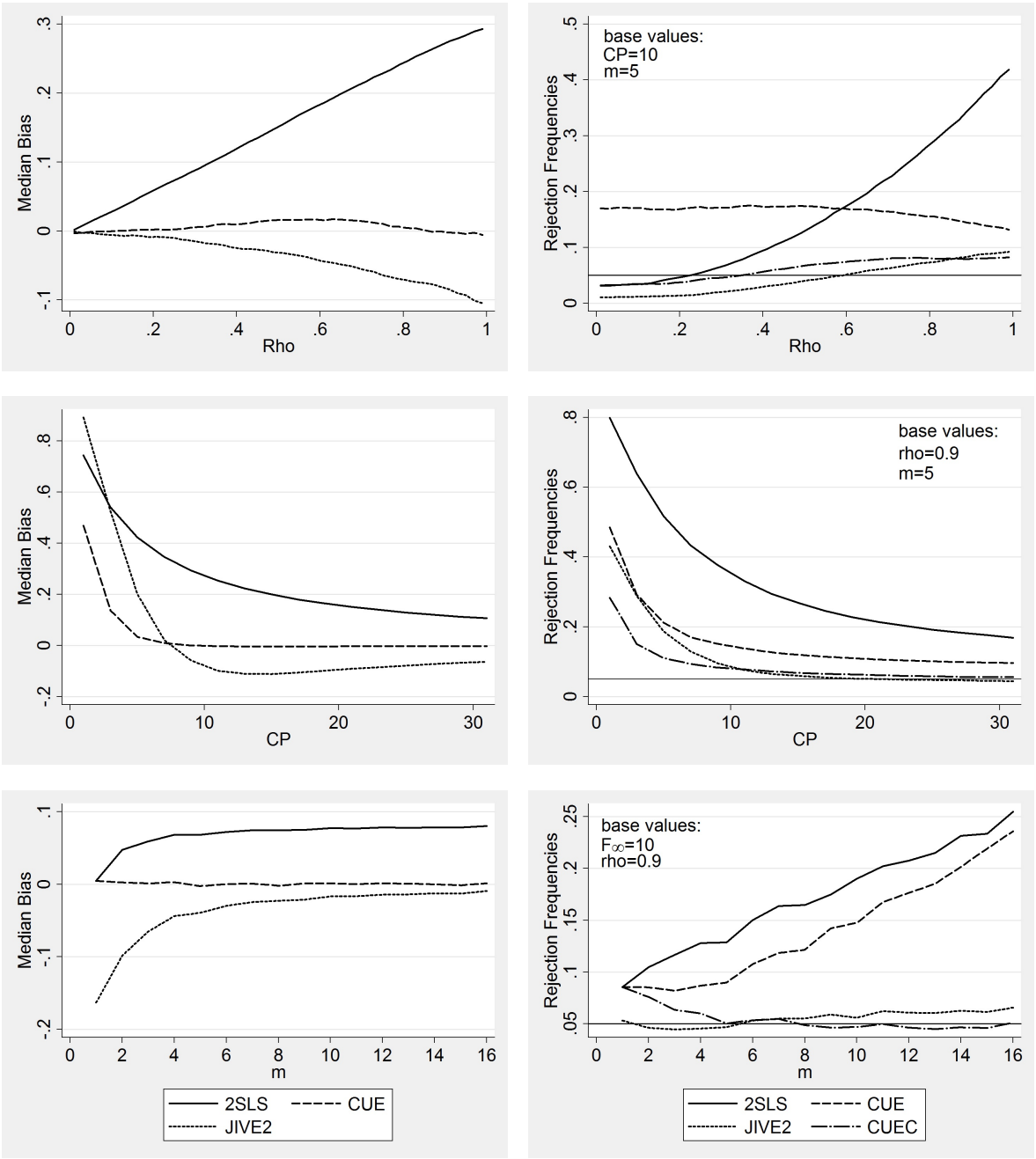


Figure 4.3: Median bias and rejection frequencies when $\boldsymbol{n = 100}$
(Step size: $\rho = 0.02$, $CP = 2$, $m = 1$; 10,000 replications each)

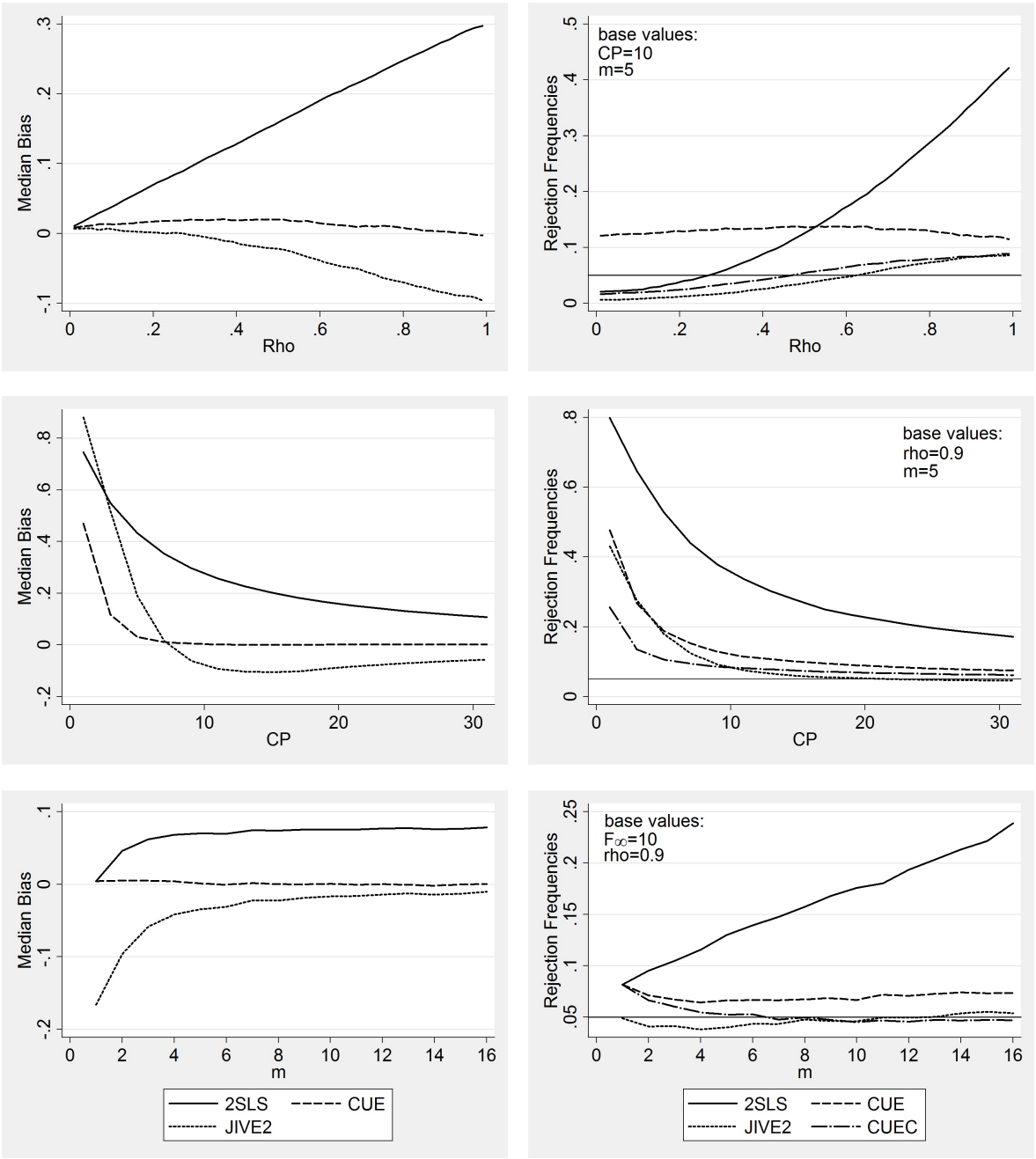


Figure 4.4: Median bias and rejection frequencies when $\boldsymbol{n = 800}$
(Step size: $\rho = 0.02$, $CP = 2$, $m = 1$; 10,000 replications each)

So far, the degree of endogeneity has been fixed at a very high level to evaluate the performance of Newey and Windmeijer's variance estimator under extreme conditions. Based on a literature review, Hansen *et al.* (2008) argue that values of $\rho \geq 0.8$ may not be very relevant for practice. In the following, I set $\rho = 0.3$ which is, according to Hansen *et al.* (2008)'s literature review, a more realistic value. Following Chamberlain and Imbens (2004), I also add an additional element to the Monte Carlo design that makes the simulations more comparable to the situation in Angrist and Krueger (1991)'s data set. Chamberlain and Imbens (2004) divide the set of instruments into a small set of basic instrumental variables and a set of doubtful instrumental variables. Using Angrist and Krueger (1991)'s application, they argue that the basic variation stems from the quarter-of-birth dummies, while the quarter-of-birth interactions are the doubtful set of instruments. In the following, I also split the set of instruments into two subsets. The first set contains a fixed number of relevant instruments (m_1) while the second set of instruments is completely irrelevant (m_2). In Figures 4.5 to 4.7 the number of irrelevant instruments is 0, 2 and 6, respectively, while the concentration parameter has been fixed at six different values. Hence, the population first stage F-statistic decreases from Figures 4.5 to 4.7, which is similar to the application discussed in the next section. Once the doubtful quarter-of-birth interactions are included, the first stage F-statistic decreases considerably.

In this simulation design the performance of 2SLS is extremely poor because not only the bias increases with the number of instruments but also the rejection frequency, given a constant bias, increases with the number of instruments (see e.g. Figure 4.4). Therefore I only discuss the results from the continuous updating estimator in the following. Without including the irrelevant instruments (see Figure 4.5), the divergence between t-statistics based on the usual standard errors and NW's standard errors does not seem to be great. While both densities appear to be non-normal for low values of the concentration parameter, they converge to a normal distribution as the concentration parameter increases. Figures 4.6 and 4.7 show that the inclusion of irrelevant instruments affects especially the t-statistics based on the usual standard errors. The density becomes even bimodal for low values of the concentration parameter. However, the performance of t-statistics based on the new variance estimator stays almost iden-

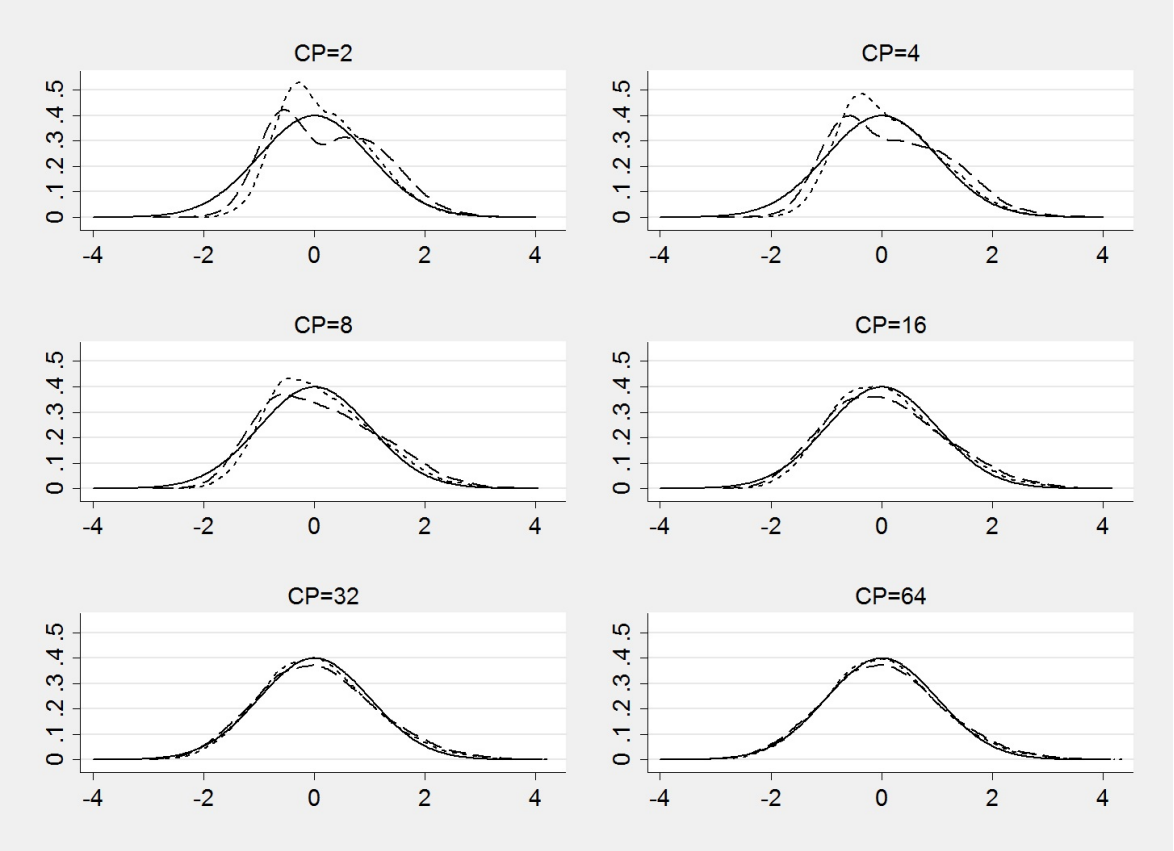


Figure 4.5: Simulation densities of t-statistics when $n = 100$ ($\rho = 0.3$, $m_1 = 2$, $\boldsymbol{m_2 = 0}$; 10,000 replications each). Dashed lines show CUEC; longer dashes show CUE; solid lines show standard normal distributions.

tical as the number of irrelevant instruments increases. For instance, with 6 irrelevant instruments and a concentration parameter of 16, the density of the CUEC is almost normal. Throughout, whenever the performance of both variance estimators is poor, the actual size of the test seems to be slightly higher for positive estimates than for negative ones. This pattern is the other way round when the correlation is negative. The corresponding graph for $m_2 = 6$ is displayed in Appendix A.2.

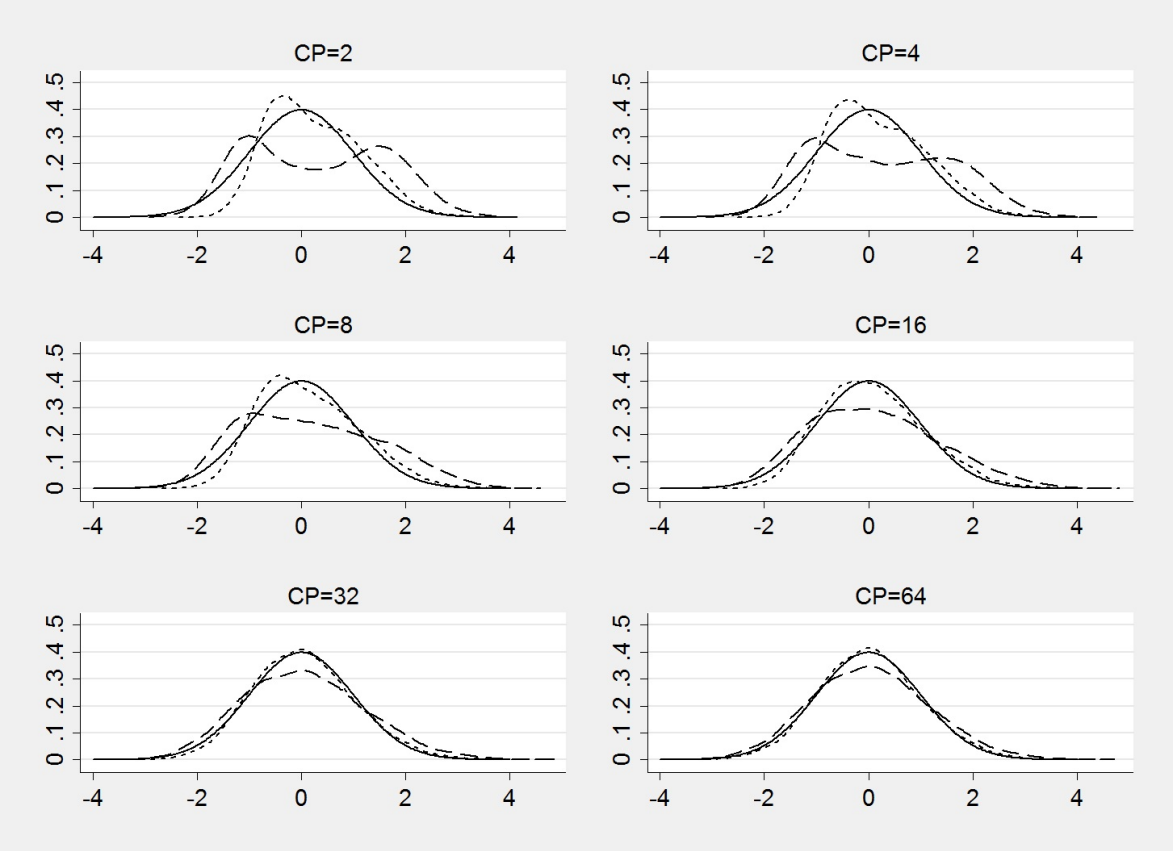


Figure 4.6: Simulation densities of t-statistics when $n = 100$ ($\rho = 0.3$, $m_1 = 2$, $\boldsymbol{m_2 = 2}$; 10,000 replications each). Dashed lines show CUEC; longer dashes show CUE; solid lines show standard normal distributions.

4.4 Application†

Angrist and Krueger (1991) estimated the effect of schooling on income using a sample of 329,500 men born 1930-39 from the 1980 census. This sample has been extensively used as an application in the weak instruments literature. Using randomly generated indicators for quarter of birth, Bound *et al.* (1995), for instance, showed that 2SLS suffer from finite-sample bias even if the sample size is as large as in Angrist and Krueger's sample. Staiger and Stock (1997) suggested that LIML point estimates are

† Following the results from the previous section, I tried different starting values for the education parameter in all regression specifications where the identification was particularly weak (F-statistic<1.5). The 2SLS starting values, however, performed well in this application.

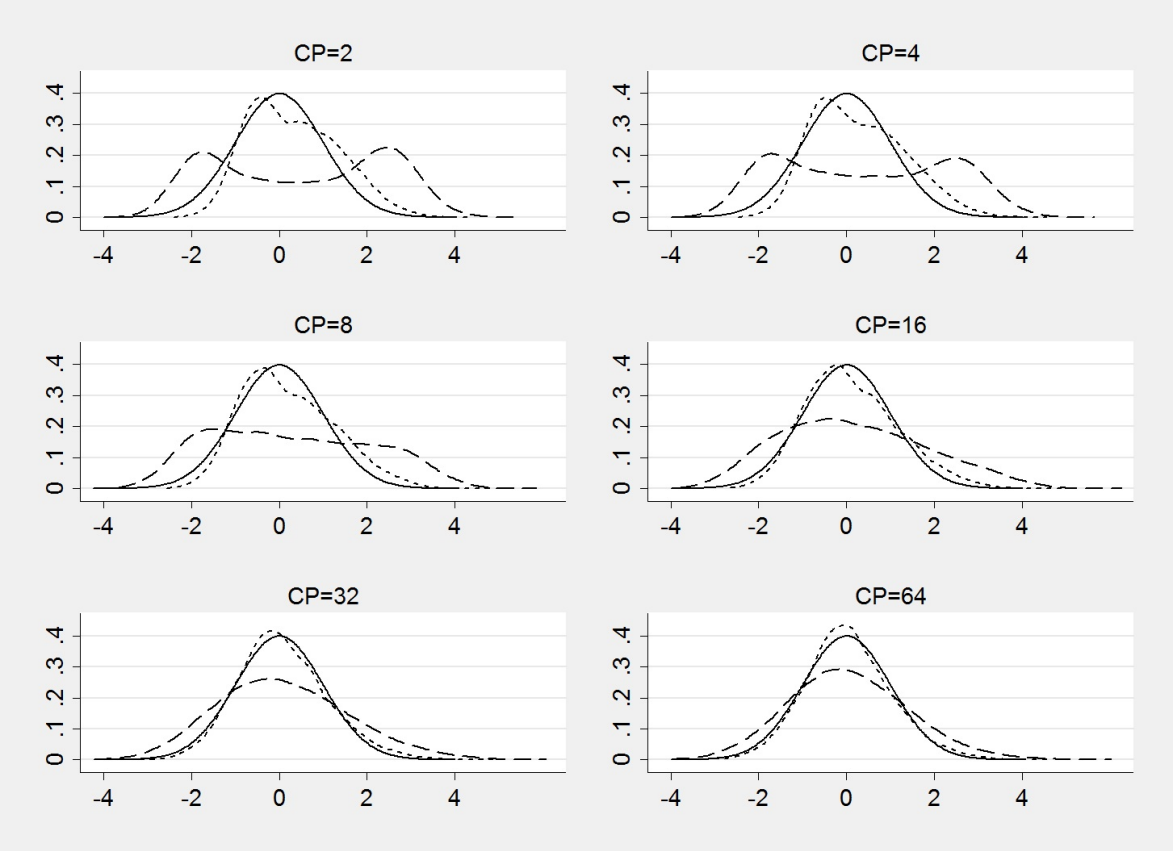


Figure 4.7: Simulation densities of t-statistics when $n = 100$ ($\rho = 0.3$, $m_1 = 2$, $\boldsymbol{m_2 = 6}$; 10,000 replications each). Dashed lines show CUEC; longer dashes show CUE; solid lines show standard normal distributions.

more reliable in situations where the first-stage F-statistic is small. However, they also showed that the coverage rate of LIML using the conventional standard errors might be too low in case of weak instruments. Almost a decade later, Cruz and Moreira (2005) picked up this issue and calculated LIML confidence intervals for this application that are based on a conditional likelihood test. Moreira (2003) showed that confidence regions based on this test have correct coverage probability even when identification is weak. Chamberlain and Imbens (2004) proposed a new estimator which also performs better in terms of coverage rate and employed this estimator to the Angrist and Krueger (1991) data set. In both studies the usual standard errors of LIML turned out to be too small. I therefore report sandwich-type standard errors which lead to the same test decisions as in Cruz and Moreira (2005).

While the LIML estimator is more reliable than 2SLS, it is not robust against certain types of heteroskedasticity (see e.g. Bekker and van der Ploeg, 2005 and Hausman *et al.*, 2007). The contribution of the present study is to analyze the Angrist and Krueger (1991) data set using an estimator that is robust to both heteroskedasticity and weak identification. The presence of heteroskedasticity in the wage equation is not unlikely. Klein and Vella (2009), for instance, discussed some potential sources of heteroskedasticity. However, the extent to which heteroskedasticity affects the point estimates of returns to education is an empirical question.

The aim of Angrist and Krueger (1991) is to estimate the causal effect of compulsory schooling on earnings using the variation that quarter of birth induces in school attendance. An increasing number of studies argues that this variation is not, or at least not completely, exogenous. Such a conclusion would preclude a causal interpretation of the results. For instance, in a very early study, Bound *et al.* (1995) discussed potential channels of a direct effect of quarter of birth on wages. They concluded that the existence of such a direct effect is quite plausible. Bound and Jaeger (2000) renewed their concerns some years later. More recently, Buckles and Hungerman (2008) showed that mothers' socioeconomic characteristics vary depending on season of birth. These characteristics can explain a considerable fraction of the relationship between quarter of birth and schooling which indicates that the variation in quarter of birth is at least not completely exogenous. Buckles and Hungerman (2008) also noted that controlling for family background characteristics might not be enough to produce consistent estimates since there could still be a correlation between season of birth and the unobservables in the model. Following the literature about weak instruments, I nevertheless use this application to illustrate the performance of Newey and Windmeijer (2009)'s new variance estimator.

I re-estimate the model in the second column of Angrist and Krueger (1991)'s table V for 100 random subsamples using the continuous updating estimator. Each subsample contains around 30% of the dataset. Table 4.3 shows the median and the standard deviation of the estimated coefficients and the corresponding Wald statistics for 2SLS, JIVE2 and the continuous updating estimator with the usual and the new variance estimator. The effect of education is slightly stronger when I use the

Table 4.3: Estimates of returns to education from 100 random subsamples

# IVs	State effects	Year effects	Coeff.	2SLS	JIVE2	CUE	CUEC
30	No	Yes	Effect size	0.082	0.100	0.092	0.092
				(0.022)	(0.202)	(0.041)	(0.041)
			Wald statistic	11.90	2.490	15.37	4.865
				(5.814)	(2.467)	(9.762)	(4.320)

Median (Standard deviation); Wald test of H_0 : *no returns to education* (χ^2_1-distributed).

continuous updating estimator compared with 2SLS. It is even somewhat stronger when I use JIVE2. The JIVE2 estimates are much more dispersed than the estimates from the continuous updating estimator. While the usual variance estimator of the continuous updating estimator leads to even larger Wald statistics than 2SLS, the corrected variance estimator gives distinctly lower Wald statistics. JIVE2 is more conservative than Newey and Windmeijer's corrected variance estimator in this regression specification. Nakov (2010) also used this data set from which he drew 100 random subsamples with 50,000 observations in each. He reported the estimates for JIVE1 and JIVE2. While the median of the effect size is very similar to his findings, the standard deviation is distinctly lower in my simulation. The smaller standard deviation is, however, in line with Davidson and MacKinnon (2006). Although they found that JIVE1 is more dispersed than 2SLS, the differences are quite modest in large samples.

Table 4.4 shows the results using the full sample of men born 1930-39. I use the same four regression specifications as Angrist and Krueger (1991) in their table V and table VII (column 7 and 8). Additionally, I estimate the specification proposed by Bound *et al.* (1995). They included only three indicator variables for quarter of birth, which is the basic source of exogenous variation. The first-stage F-statistic is 13.5 in this regression specification and thus above the rule of thumb for 2SLS. Not surprisingly, the estimates of the returns to education in the first column of Table 4.4 are very similar for 2SLS, LIML and the continuous updating estimator. The many weak instruments standard errors are also very close to the usual standard errors. The number of excluded instruments is distinctly larger in Angrist and Krueger (1991)'s specification. The instruments in column 2 to 5 of Table 4.4 are obtained by interacting quarter of birth with 9 year of birth indicators. The instruments in the last column are obtained by

interacting quarter of birth with 9 year of birth and 50 state of birth indicators. The first-stage F-statistic reported for these five specifications is now distinctly lower which might bias the 2SLS estimator towards the OLS estimator (see e.g. Bound *et al.*, 1995 and Staiger and Stock, 1997). Staiger and Stock (1997) also show that this bias is less of a problem for LIML than 2SLS.

The point estimates of the continuous updating estimator are always close to the LIML estimates. This is also the case for the many weak instruments standard errors

Table 4.4: Returns to education on men's log weekly earnings (born 1930-1939)

	BJB			AK		
	(1)	(2)	(3)	(4)	(5)	(6)
OLS	0.063	0.071	0.071	0.063	0.063	0.063
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
2SLS	0.142	0.089	0.076	0.081	0.060	0.081
	(0.033)	(0.016)	(0.029)	(0.016)	(0.029)	(0.011)
LIML	0.146	0.093	0.081	0.084	0.057	0.098
	(0.035)	(0.020)	(0.060)	(0.020)	(0.052)	(0.022)
JIVE2	0.202	0.096	0.116	0.090	0.086	0.144
	(0.066)	(0.022)	(0.266)	(0.026)	(0.211)	(0.052)
CUE	0.146	0.095	0.085	0.086	0.061	0.102
usual SE	(0.033)	(0.016)	(0.029)	(0.017)	(0.029)	(0.011)
many weak instruments SE	(0.035)	(0.020)	(0.058)	(0.020)	(0.050)	(0.022)
Excluded instruments:						
Quarter of birth	×	×	×	×	×	×
Quarter of birth × year of birth		×	×	×	×	×
Quarter of birth × state of birth						×
Number of excluded instruments	3	30	28	30	28	178
F(excluded instruments)	13.5	4.80	1.42	4.62	1.40	1.78
Control variables:						
Age, Age^2	×		×		×	×
Race, SMSA, married	×			×	×	×
9 Year-of-birth dummies		×	×	×	×	×
8 Region-of-residence dummies	×			×	×	×
50 State-of-birth dummies						×

Number of observations: 329,509. Robust SE in parentheses.

BJB: Bound *et al.* (1995)'s regression specification; AK: Angrist and Krueger (1991)'s regression specification.

and the standard errors of LIML. Figure 4.4 from the simulation results shows that the difference between the usual and NW's standard errors is inversely related to the strength of the instruments. This pattern can also be observed in the real data set. While the relative difference is around 6% in Bound *et al.* (1995)'s specification, it is up to 100% in the other specifications where the first-stage F-statistic is considerably lower.[8]

4.5 Conclusion

This study analyzes the finite-sample properties of the continuous updating estimator. While it is well-known that trying different starting values is necessary to obtain the global minimum of the criterion function of the continuous updating estimator, it is interesting to see that this can also affect its reported performance in Monte Carlo simulations. To put it the other way around, choosing the two-stage least squares estimates as starting values turns out to be a poor choice, especially when the sample size is small and/or the identification is weak. A potential reason for this is the likely presence of local optima and the fact that the continuous updating estimator often converges to these local optima if they are close to the 2SLS estimates. In these cases the continuous updating estimator seems to be distinctly more biased and less dispersed. This study also shows that extreme estimates of the continuous updating estimator, which are often reported in Monte Carlo simulations, are more likely to be a failure of the optimization routine than a property of the continuous updating estimator.

The continuous updating estimator becomes almost median unbiased in my simulation design once the population first stage F-statistic is around 1. This is considerably lower than the rule of thumb for 2SLS, which is around 10. The median bias appears to be independent of the degree of endogeneity. Throughout, the continuous updating estimator outperforms 2SLS and JIVE2 once the sample size is reasonably large. However, the rejection frequencies with the usual standard errors are far too

[8] This pattern can also be observed in the 1920-29 cohort and in the 1940-49 cohort. The corresponding results are reported in Appendix A.3 and A.4.

high especially in small samples. In contrast to this, the rejection frequencies with the new variance estimator proposed by Newey and Windmeijer (2009) are often very close to the nominal level once the sample size is reasonably large. Additionally, the new variance estimator depends only slightly on the inclusion of irrelevant instruments. This property makes it particularly attractive for the analysis of Angrist and Krueger (1991)'s regression specifications in which a large set of quarter-of-birth interactions weakens the identification.

Appendices

A.1 Narrow replication of NW (2009)

Table 4.5: Median bias and rejection frequencies

	CP=10		CP=20		CP=35	
	Bias	RF	Bias	RF	Bias	RF
			$\rho = 0.3$			
m=3						
2SLS	0.0507	0.0447	0.0267	0.0474	0.0140	0.0503
JIVE2	-0.0366	0.0208	-0.0388	0.0289	-0.0216	0.0388
CUEC	0.0050	0.0376	0.0016	0.0407	0.0020	0.0436
m=15						
2SLS	0.1796	0.1622	0.1291	0.1333	0.0894	0.1112
JIVE2	0.0615	0.0167	-0.0160	0.0236	-0.0161	0.0334
CUEC	0.0328	0.0746	0.0052	0.0642	0.0020	0.0554
			$\rho = 0.5$			
m=3						
2SLS	0.0871	0.0820	0.0444	0.0722	0.0253	0.0636
JIVE2	-0.0695	0.0368	-0.0591	0.0339	-0.0332	0.0389
CUEC	0.0034	0.0533	0.0016	0.0469	0.0017	0.0465
m=15						
2SLS	0.2983	0.4022	0.2125	0.3039	0.1483	0.2254
JIVE2	0.0870	0.0438	-0.0279	0.0376	-0.0270	0.0373
CUEC	0.0475	0.1018	0.0069	0.0737	-0.0005	0.0546
			$\rho = 0.9$			
m=3						
2SLS	0.1589	0.1908	0.0825	0.1323	0.0479	0.0993
JIVE2	-0.1354	0.0679	-0.1007	0.0439	-0.0569	0.0394
CUEC	0.0029	0.0767	0.0022	0.0624	0.0015	0.0548
m=15						
2SLS	0.5356	0.9343	0.3801	0.7988	0.2644	0.6255
JIVE2	0.0861	0.1405	-0.0586	0.0778	-0.0447	0.0615
CUEC	0.0149	0.0955	0.0022	0.0685	0.0006	0.0555

$n = 200$; $\alpha = \beta_x = 0$; 10,000 replications.

Rejection frequencies for $H_0 : \beta_x = 0$ using Wald tests.

The results for JIVE2 are not comparable since Newey and Windmeijer have estimated a generalization of JIVE2.

A.2 Simulation densities of t-statistics when $\rho < 0$

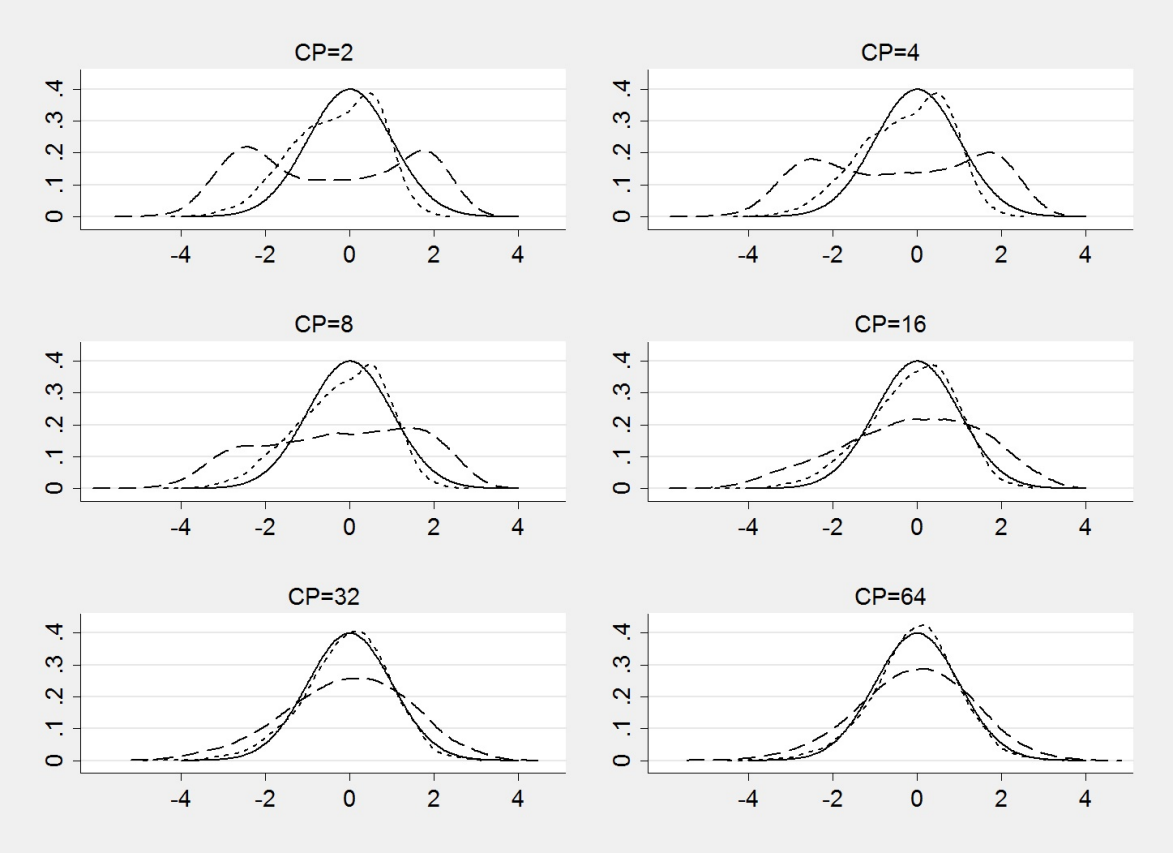


Figure 4.8: Simulation densities of t-statistics when ρ is negative ($\rho = -0.3$, $m_1 = 2$, $\boldsymbol{m_2} = \mathbf{6}$, $n = 100$; 10,000 replications each). Dashed lines show CUEC; longer dashes show CUE; solid lines show standard normal distributions.

A.3 Returns to education (men born 1920-1929)

Table 4.6: Returns to education on men's log weekly earnings (born 1920-1929)

	BJB			AK		
	(1)	(2)	(3)	(4)	(5)	(6)
OLS	0.070	0.080	0.080	0.070	0.070	0.069
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
2SLS	0.056	0.077	0.131	0.067	0.101	0.089
	(0.021)	(0.015)	(0.034)	(0.015)	(0.034)	(0.011)
LIML	0.055	0.076	0.255	0.066	0.282	0.131
	(0.021)	(0.020)	(0.156)	(0.020)	(0.357)	(0.037)
JIVE2	0.051	0.076	-0.081	0.065	0.032	0.481
	(0.027)	(0.021)	(0.137)	(0.024)	(0.042)	(0.638)
CUE	0.056	0.075	0.257	0.065	0.286	–†
usual SE	(0.021)	(0.015)	(0.045)	(0.015)	(0.052)	
many weak instruments SE	(0.021)	(0.021)	(0.124)	(0.020)	(0.242)	
Excluded instruments:						
Quarter of birth	×	×	×	×	×	×
Quarter of birth × year of birth		×	×	×	×	×
Quarter of birth × state of birth						×
Number of excluded instruments	3	30	28	30	28	176
F(excluded instruments)	24.7	4.60	1.17	4.57	1.12	1.51
Control variables:						
Age, Age^2	×		×		×	×
Race, SMSA, married	×			×	×	×
9 Year-of-birth dummies		×	×	×	×	×
8 Region-of-residence dummies	×			×	×	×
50 State-of-birth dummies						×

Number of observations: 247,199. Robust SE in parentheses. † Estimation did not converge.

BJB: Bound *et al.* (1995)'s regression specification; AK: Angrist and Krueger (1991)'s regression specification.

A.4 Returns to education (men born 1940-1949)

Table 4.7: Returns to education on men's log weekly earnings (born 1940-1949)

	BJB			AK		
	(1)	(2)	(3)	(4)	(5)	(6)
OLS	0.052	0.057	0.057	0.052	0.052	0.052
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
2SLS	0.201	0.055	0.095	0.039	0.078	0.067
	(0.059)	(0.014)	(0.022)	(0.015)	(0.024)	(0.011)
LIML	0.293	0.054	0.137	0.029	0.124	0.088
	(0.119)	(0.025)	(0.049)	(0.027)	(0.070)	(0.028)
JIVE2	1.623	0.055	0.124	0.035	0.128	0.117
	(3.888)	(0.017)	(0.042)	(0.019)	(0.077)	(0.048)
CUE	0.293	0.055	0.140	0.030	0.129	0.090
usual SE	(0.073)	(0.014)	(0.023)	(0.015)	(0.025)	(0.011)
many weak instruments SE	(0.105)	(0.025)	(0.049)	(0.027)	(0.067)	(0.029)
Excluded instruments:						
Quarter of birth	×	×	×	×	×	×
Quarter of birth × year of birth		×	×	×	×	×
Quarter of birth × state of birth						×
Number of excluded instruments	3	30	28	30	28	178
F(excluded instruments)	6.25	7.27	3.27	6.54	2.71	1.93
Control variables:						
Age, Age^2	×		×		×	×
Race, SMSA, married	×			×	×	×
9 Year-of-birth dummies		×	×	×	×	×
8 Region-of-residence dummies	×			×	×	×
50 State-of-birth dummies						×

Number of observations: 486,926. Robust SE in parentheses.

BJB: Bound *et al.* (1995)'s regression specification; AK: Angrist and Krueger (1991)'s regression specification.

Bibliography

Aigner DJ, 1973. Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics* **1**(1): 49–59. DOI: 10.1016/0304-4076(73)90005-5

Angrist JD, Imbens GW, Krueger AB, 1999. Jackknife instrumental variables estimation. *Journal of Applied Econometrics* **14**(1): 57–67. DOI: 10.1002/(SICI)1099-1255(199901/02)14:1<57::AID-JAE501>3.0.CO;2-G

Angrist JD, Krueger AB, 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* **106**(4): 979–1014. DOI: 10.2307/2937954

Augurzky B, Bauer TK, Schaffner S, 2006. Copayments in the German health system - Do they work? *RWI Discussion Papers* **43.**

Baum CF, Schaffer ME, Stillman S, 2007. Enhanced routines for instrumental variables / generalized method of moments estimation and testing. *Stata Journal* **7**(4): 465–506.

Bekker PA, van der Ploeg J, 2005. Instrumental variable estimation based on grouped data. *Statistica Neerlandica* **59**(3): 239–267. DOI: 10.1111/j.1467-9574.2005.00296.x

Bound J, Jaeger DA, 2000. Do compulsory school attendance laws alone explain the association between quarter of birth and earnings? *Research in Labor Economics* **19**: 83–108. DOI: 10.1016/S0147-9121(00)19005-3

Bound J, Jaeger DA, Baker RM, 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory

variable is weak. *Journal of the American Statistical Association* **90**: 443–450. URL http://www.jstor.org/stable/2291055

Buckles K, Hungerman DM, 2008. Season of birth and later outcomes: Old questions, new answers. NBER Working Paper 14573.

Bun MJG, Windmeijer F, 2011. A comparison of bias approximations for the two-stage least squares (2SLS) estimator. *Economics Letters* **113**(1): 76–79. DOI: 10.1016/j.econlet.2011.05.047

Chamberlain G, Imbens GW, 2004. Random effects estimators with many instrumental variables. *Econometrica* **72**(1): 295–306. DOI: 10.1111/j.1468-0262.2004.00485.x

Charlson ME, Pompei P, Ales KL, MacKenzie CR, 1987. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases* **40**(5): 373–383. DOI: 10.1016/0021-9681(87)90171-8

Contoyannis P, Hurley J, Grootendorst P, Jeon SH, Tamblyn R, 2005. Estimating the price elasticity of expenditure for prescription drugs in the presence of non-linear price schedules: an illustration from Quebec, Canada. *Health Economics* **14**(9): 909–923. DOI: 10.1002/hec.1041

Cruz LM, Moreira MJ, 2005. On the validity of econometric techniques with weak instruments. *Journal of Human Resources* **40**(2): 393–410.

Davidson R, MacKinnon JG, 2006. The case against JIVE. *Journal of Applied Econometrics* **21**(6): 827–833. DOI: 10.1002/jae.873

Dhaene G, Santos Silva JMC, 2011. Specification and testing of models estimated by quadrature. *Journal of Applied Econometrics,* forthcoming. DOI: 10.1002/jae.1196

Ellis RP, 1986. Rational behavior in the presence of coverage ceilings and deductibles. *The RAND Journal of Economics* **17**(2): 158–175. URL http://www.jstor.org/stable/2555381

Farbmacher H, 2009. Copayments for doctor visits and the probability of visiting a physician - evidence from a natural experiment. Munich Discussion Paper No. 2009-10.

Farbmacher H, 2011a. Estimation of hurdle models for overdispersed count data. *Stata Journal* **11**(1): 82–94.

Farbmacher H, 2011b. GMM with many weak moment conditions: Replication and application of Newey and Windmeijer (2009). *Journal of Applied Econometrics,* forthcoming. DOI: 10.1002/jae.1277

Greene W, 2008. Functional forms for the negative binomial model for count data. *Economics Letters* **99**(3): 585–590. DOI: 10.1016/j.econlet.2007.10.015

Grobe TK, Dörning H, Schartz FW, 2010. BARMER GEK Arztreport. Schriftenreihe zur Gesundheitsanalyse.

Guggenberger P, 2005. Monte-carlo evidence suggesting a no moment problem of the continuous updating estimator. *Economics Bulletin* **3**(13): 1–6.

Guggenberger P, 2008. Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator. *Econometric Reviews* **27**(4-6): 526–541. DOI: 10.1080/07474930801960410

Hahn J, Hausman JA, 2002. Notes on bias in estimators for simultaneous equation models. *Economics Letters* **75**(2): 237–241. DOI: 10.1016/S0165-1765(01)00602-4

Han C, Phillips PC, 2005. GMM with many moment conditions. Cowles Foundation Discussion Paper No. 1515, Yale University.

Han C, Phillips PCB, 2006. GMM with many moment conditions. *Econometrica* **74**(1): 147–192. DOI: 10.1111/j.1468-0262.2006.00652.x

Hansen C, Hauman JA, Newey WK, 2008. Estimation with many instrumental variables. *Journal of Business and Economic Statistics* **26**(4): 398–422. DOI: 10.1198/073500108000000024

Hansen LP, 1982. Large sample properties of generalized method of moments estimators. *Econometrica* **50**(4): 1029–1054. URL http://www.jstor.org/stable/1912775

Hansen LP, Heaton J, Yaron A, 1996. Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics* **14**(3): 262–280. URL http://www.jstor.org/stable/1392442

Hausman JA, Lewis R, Menzel K, Newey WK, 2011. Properties of the CUE estimator and a modification with moments. *Journal of Econometrics* **165**(1): 45–57. DOI: 10.1016/j.jeconom.2011.05.005

Hausman JA, Newey WK, Woutersen T, Chao J, Swanson N, 2007. Instrumental variable estimation with heteroskedasticity and many instruments. URL http://econ-www.mit.edu/files/1544

Imbens GW, Angrist JD, 1994. Identification and estimation of local average treatment effects. *Econometrica* **62**(2): 467–475. URL http://www.jstor.org/stable/2951620

Jung KT, 1998. Influence of the introduction of a per-visit copayment on health care use and expenditures: The Korean experience. *The Journal of Risk and Insurance* **65**(1): 33–56. URL http://www.jstor.org/stable/253490

Keeler EB, Newhouse JP, Phelps CE, 1977. Deductibles and the demand for medical care services: The theory of a consumer facing a variable price schedule under uncertainty. *Econometrica* **45**(3): 641–655. URL http://www.jstor.org/stable/1911679

Klein R, Vella F, 2009. Estimating the return to endogenous schooling decisions for Australian workers via conditional second moments. *Journal of Human Resources* **44**(4): 1047–1065.

Liu Q, Pierce DA, 1994. A note on Gauss-Hermite quadrature. *Biometrika* **81**(3): 624–629. DOI: 10.1093/biomet/81.3.624

Meyerhoefer CD, Zuvekas SH, 2010. New estimates of the demand for physical and mental health treatment. *Health Economics* **19**(3): 297–315. DOI: 10.1002/hec.1476

Moreira MJ, 2003. A conditional likelihood ratio test for structural models. *Econometrica* **71**(4): 1027–1048. DOI: 10.1111/1468-0262.00438

Mullahy J, 1986. Specification and testing of some modified count data models. *Journal of Econometrics* **33**(3): 341–365. DOI: 10.1016/0304-4076(86)90002-3

Nakov A, 2010. Jackknife instrumental variables estimation: replication and extension of Angrist, Imbens and Krueger (1999). *Journal of Applied Econometrics* **25**(6): 1063–1066. DOI: 10.1002/jae.1191

Newey WK, Windmeijer F, 2009. Generalized method of moments with many weak moment conditions. *Econometrica* **77**(3): 687–719. DOI: 10.3982/ECTA6224

OECD, 2008. *OECD Health Data 2008*. Organisation for Economic Co-operation and Development: Paris.

Pohlmeier W, Ulrich V, 1995. An econometric model of the two-part decisionmaking process in the demand for health care. *The Journal of Human Resources* **30**(2): 339–361. URL `http://www.jstor.org/stable/146123`

Rückert I, Böcken J, Mielck A, 2008. Are German patients burdened by the practice charge for physician visits (Praxisgebühr)? A cross sectional analysis of socioeconomic and health related factors. *BMC Health Services Research* **8**(232). DOI: 10.1186/1472-6963-8-232

Roemer MI, Hopkins CE, Carr L, Gartside F, 1975. Copayments for ambulatory care: Penny-wise and pound-foolish. *Medical Care* **13**(6): 457–466. URL `http://www.jstor.org/stable/3763483`

Santos Silva JMC, 2003. A note on the estimation of mixture models under endogenous sampling. *Econometrics Journal* **6**(1): 46–52. DOI: 10.1111/1368-423X.00100

Santos Silva JMC, Windmeijer F, 2001. Two-part multiple spell models for health care demand. *Journal of Econometrics* **104**(1): 67–89. DOI: 10.1016/S0304-4076(01)00059-8

Schreyögg J, Grabka MM, 2010. Copayments for ambulatory care in Germany: a natural experiment using a difference-in-difference approach. *European Journal of Health Economics* **11**(3): 331–341. DOI: 10.1007/s10198-009-0179-9

Skrondal A, Rabe-Hesketh S, 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall/CRC.

Staiger D, Stock JH, 1997. Instrumental variables regression with weak instruments. *Econometrica* **65**(3). URL `http://www.jstor.org/stable/2171753`

Stoddart GL, Barer ML, 1981. Analyses of demand and utilization through episodes of medical service. In van der Gaag J, Perlman M (eds.) *Health, Economics, and Health Economics*, 149–170. North-Holland Publishing Company.

van de Voorde C, van Doorslaer E, Schokkaert E, 2001. Effects of cost sharing on physician utilization under favourable conditions for supplier-induced demand. *Health Economics* **10**(5): 457–471. DOI: 10.1002/hec.631

van Kleef RC, van de Ven WPMM, van Vliet RCJA, 2009. Shifted deductibles for high risks: More effective in reducing moral hazard than traditional deductibles. *Journal of Health Economics* **28**(1): 198–209. DOI: 10.1016/j.jhealeco.2008.09.007

Vesterinen J, Pouta E, Huhtala A, Neuvonen M, 2010. Impacts of changes in water quality of recreation behavior and benefits in Finland. *Journal of Environmental Management* **91**(4): 984–994. DOI: 10.1016/j.jenvman.2009.12.005

Vuong QH, 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**(2): 307–333. URL `http://www.jstor.org/stable/1912557`

Winkelmann R, 2004a. Health care reform and the number of doctor visits - An econometric analysis. *Journal of Applied Econometrics* **19**(4): 455–472. DOI: 10.1002/jae.764

Winkelmann R, 2004b. Co-payments for prescription drugs and the demand for doctor visits - Evidence from a natural experiment. *Health Economics* **13**(11): 1081–1089. DOI: 10.1002/hec.868

Winkelmann R, 2006. Reforming health care: Evidence from quantile regressions for counts. *Journal of Health Economics* **25**(1): 131–145. DOI: 10.1016/j.jhealeco.2005.03.005

Winkelmann R, 2008. *Econometric Analysis of Count Data*. Springer, Heidelberg.

Wong IOL, Lindner MJ, Cowling BJ, Lau EHY, Lo SV, Leung GM, 2010. Measuring moral hazard and adverse selection by propensity scoring in the mixed health care economy of Hong Kong. *Health Policy* **95**(1): 24–35. DOI: 10.1016/j.healthpol.2009.10.006

Yogo M, 2004. Estimating the elasticity of intertemporal substitution when instruments are weak. *Review of Economics and Statistics* **86**(3): 797–810. DOI: 10.1162/0034653041811770