CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Socioeconomic predictors of alcohol consumption in the Czech adult population

Master's thesis

Author: Jan Hanzal
Study program: Economics and Finance
Supervisor: doc. Zuzana Havránková, Ph.D.
External advisor: RNDr. Ladislav Kážmer, Ph.D.
Year of defense: 2022

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, August 2, 2022

Jan Hanzal

Abstract

This empirical study focuses on the relation between individual social and economic variables and patterns of alcohol consumption in the Czech Republic. The work is divided into two parts. The first one concentrates on an exploratory analysis of a cross-sectional dataset. The results of this part reveal that several variables are significantly correlated with alcohol consumption, namely education, marital status and household income. The second part attempts to get closer to the actual causal effects of unemployment and household income on alcohol consumption by employing the Arellano-Bond estimator on a separate panel dataset. The results somewhat differ from the first part, with household income having a noticeably higher point estimate. The aim of this thesis is to bring more current and, most importantly, more robust results to the research on the topic.

JEL Classification	I10
Keywords	Alcohol consumption, Arellano-Bond estimator, Economic predictors of health
Title	Socioeconomic predictors of alcohol consumption in the Czech adult population

Abstrakt

Tato práce empiricky zkoumá závislost vzorců konzumace alkoholu v České republice na socioekonomických proměnných, a to ve dvou částech. První z nich je zaměřena na explorativní analýzu průřezových dat. Výsledky této části ukazují, že několik proměnných signifikantně koreluje s konzumací alkoholu, především pak úroveň vzdělání, rodinný stav a příjem domácnosti. Cílem druhé části je přiblížit se ke kauzálním efektům pro dvě, v čase se měnící, proměnné, příjem domácnosti a nezaměstnanost, za použití metody Arellano-Bond na panelových datech. Výsledky této části se poněkud liší, jelikož bodový odhad pro příjem domácnosti je znatelně vyšší. Cílem práce je především přinést aktualnější a ekonometricky robustnější výsledky do současného výzkumu na toto téma.

Klasifikace JEL	I10
Klicova slova	Konzumace alkoholu, Arellano-Bond estimátor, Ekonomické determinanty zdraví
Nazev prace	Socioekonomické prediktory konzumace alkoholu v české dospělé populaci

Acknowledgments

I am deeply grateful to my supervisor doc. Zuzana Havránková, Ph.D. for her invaluable methodological advice and overall help in constructing the story of this thesis. Equally, I would like to express my deepest thanks to my external advisor RNDr. Ladislav Kážmer, Ph.D. of the National Institute of Mental Health for providing deep insights into the sociological aspects of alcohol consumption and their statistical modelling, without which this endeavour would have been extremely difficult.
Furthermore, I would like to thank PhDr. Ladislav Csémy for his advice on the measurement of alcohol intake and for helping me obtain the data for the first part of this thesis, along with the National Institute of Mental Health and the National Institute of Public Health. Finally, I would also like to thank Mgr. Milan Ščasný, Ph.D. for his advice on two-part models.

Typeset in LaTeX using the IES Thesis Template.

Bibliographic Record
Hanzal, Jan: *Socioeconomic predictors of alcohol consumption in the Czech adult population.* Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2022, pages 72. Advisors: doc. Zuzana Havránková, Ph.D., RNDr. Ladislav Kážmer, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

AB GMM Arellano & Bond (1991) generalized method of moments estimator

ABV Alcohol by volume

ML Maximum likelihood

CDF Cumulative distribution function

SES Socioeconomic status

Master's Thesis Proposal

Institute of Economic Studies
Faculty of Social Sciences
Charles University

Author:	**Jan Hanzal**	Supervisor:	doc. Zuzana Havránková, PhD.
E-mail:	-	E-mail:	-
Phone:	-	Phone:	-
Specialization:	ET&M	Defense Planned:	June 2022

Proposed Topic:

Socioeconomic predictors of alcohol consumption patterns in the Czech adult population

Motivation:

Alcohol consumption is one of the leading causes of poor health and premature death. In terms of the global burden of alcohol-attributable mortality, in 2016, alcohol consumption led to approximately 3 million deaths worldwide, which constituted about 5.3 % of the total number of deaths (World Health Organization, 2018). Alcohol abuse significantly increases the risk of liver disease, oropharynx, larynx, oesophagus, liver, colon, rectum and female breast cancers, cardiovascular disease, infectious disease (through the weakening of the immune system and also risky sexual behaviour in relation to AIDS) and mental illness (ibid.) Furthermore, alcohol increases the risks of aggressive behaviour of an individual, which in certain cases leads to violent criminality (Bushman, 2002). It also increases the risk of a number of external causes of morbidity and mortality, particularly with respect to accidents; e.g. it has been estimated that driving under the influence of alcohol increases the risk of a fatal car accident approximately 17.8 times (Martin, Gadegbeku, Wu, Viallon, & Laumon, 2017).

It is important to note that the amount of alcohol consumed, and the drinking behaviour, vary greatly between individuals. Notably, it has been shown by empirical studies that socioeconomic characteristics of an individual are significantly related to the person's drinking habits as well as their negative consequences. Namely, Grittner, Kuntsche, Graham, & Bloomfield (2012) show in a multilevel country and individual analysis that low educational attainment worsens the harmful effects of alcohol. Similarly, in a multilevel analysis of Czech municipal and individual data, Dzúrová, Spilková, & Pikhart (2010) find that single people, unemployed people and people with low education are more likely to engage in risky alcohol consumption. On the other hand, there also exist studies that indicate very different conclusions. For example, a more recent study on Finnish and Chilean data by Peña et al. (2017) shows that groups with low socioeconomic status are more likely to abstain from alcohol, while middle-aged women of high socioeconomic status are the most at risk.

Given the serious impact of alcohol consumption on public health, it is incredibly important to study the impact of the already mentioned, and similar, socioeconomic variables on alcohol consumption patterns and their consequences to identify groups which are the most at risk. The proposed thesis will study how these associations have evolved over the recent decade across Czech regions.

Hypotheses:

1. Hypothesis #1: Individual level socioeconomic variables predict levels and patterns of alcohol use.
2. Hypothesis #2: There is a causal link between household income and alcohol consumption.
3. Hypothesis #3: There is a causal link between marital status and alcohol consumption.

Methodology:

The thesis will use a series of data collected within the National Survey of Alcohol and Tobacco Use in the Czech Republic for the years 2012, 2014, 2016, 2018, 2019 and 2020 (e.g. Csémy et al., 2021). The data provides

detailed individual level information on the amount of alcohol consumed, including categorization by type (beer, wine, liquor), and indicators of risky drinking behaviour. It also includes socioeconomic and demographic variables, such as attained education, gender, age, marital status, or estimated net household income. The datasets also provide information on the region where the individual lives.

The data will be analysed through the lens of a two-part logit and log-regression model for participation and consumption equations. Furthermore, the thesis will make use of an ordinal logit or a multinomial logit model to analyse patterns of risky consumption behaviour, including binge drinking. To at least partially control for unobserved effects in the regression equations, such as local social deprivation, dummies for years and regions, and potentially also regional time-trends, will be used.

Given that the consumption of alcohol is a complex social phenomenon, there might still be unforeseen omitted variables affecting the results, as well as reverse causality. Nevertheless, it might still be possible to state which social groups are more likely to be affected by problematic drinking behaviour on the basis of the above models.

A possible solution to endogeneity could be the use of a separate panel dataset, collected by Sociologický ústav AV ČR (2019). As opposed to the previously mentioned dataset, this one is not based on pooled cross-sections and has detailed information on numerous social and economic variables – therefore, these variables and the dynamic structure of the data could be used to tackle the potential endogeneity. More concretely, the thesis will make use of the model proposed by Arellano & Bond (1991) to approach the causal estimation of the effect of household income and marital status on alcohol consumption, two potentially important time-varying predictors. The Arellano & Bond (1991) model uses the generalized method of moments to estimate coefficients of lagged predetermined variables and includes fixed effects. This will allow the author to:

1. Control for unobserved heterogeneity between individuals that might be correlated with observed time-varying regressors
2. Use a lag of the dependent variable to control for potential reverse causality. Given that alcohol consumption potentially influences income and even marital status, it might bias the (reverse) estimate of interest. If we assume that this effect only acts through a lag, the proposed model will allow us to control for this reverse effect.

Unfortunately, the panel data has only limited information on alcohol consumption and measures it quite imprecisely. Thus it is probably not suitable for the main analysis.

Expected Contribution:

The thesis will provide a detailed overview of the impact of socioeconomic predictors on the patterns of alcohol consumption among the Czech adult population. First, it will begin with a thorough descriptive analysis of Czech alcohol consumption patterns from a regional and chronological perspective, extending current descriptive statistics studies such as Sovinová, Csémy, & Kernová (2014) and Csémy et al. (2021). Second, the thesis will analyse the impact of socioeconomic predictors on these patterns through regression modelling. It will therefore build upon studies such as Dzúrová et al. (2010) by using data on alcohol consumption patterns and socioeconomic status for the recent decade. This data is also more detailed with respect to alcohol consumption and will possibly allow for richer model specifications and more precise conclusions.
Lastly, the thesis will attempt to estimate the causal effect of two time-varying regressors through the use of GMM modelling on a separate panel dataset.

Outline:

Introduction
- motivation

Literature Review
- overview of the literature on the theoretical underpinnings of the relationship between socioeconomic variables and alcohol consumption
- summary of the empirical methods used in the literature to estimate these relationships
- overview of the results of the analyses

Descriptive analysis
- presentation of the data
- summary statistics, maps and charts outlining the topic and the data

Econometric analysis

- hypotheses
- methodology and data transformations
- results
- discussion

Conclusion
- summary and policy implications

Core Bibliography:

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*(2), 277--297.

Bushman, B. J. (2002). Effects of alcohol on human aggression. *Recent Developments in Alcoholism*, 227–243.

Csémy, L., Dvořáková, Z., Fialová, A., Kodl, M., Malý, M., & Skývová, M. (2021). Národní výzkum užívání tabáku a alkoholu v České republice 2020. *Státní Zdravotní Ústav*.

Dzúrová, D., Spilková, J., & Pikhart, H. (2010). Social inequalities in alcohol consumption in the Czech Republic: A multilevel analysis. *Health & Place*, *16*(3), 590–597. https://doi.org/https://doi.org/10.1016/j.healthplace.2010.01.004

Grittner, U., Kuntsche, S., Graham, K., & Bloomfield, K. (2012). Social Inequalities and Gender Differences in the Experience of Alcohol-Related Problems. *Alcohol and Alcoholism*, *47*(5), 597–605. https://doi.org/10.1093/alcalc/ags040

Martin, Gadegbeku, B. J.-L., Wu, D., Viallon, V., & Laumon, B. (2017). Cannabis, alcohol and fatal road accidents. *PLoS One*, *12*(11).

Peña, S., Mäkelä, P., Valdivia, G., Helakorpi, S., Markkula, N., Margozzini, P., & Koskinen, S. (2017). Socioeconomic inequalities in alcohol consumption in Chile and Finland. *Drug and Alcohol Dependence*, *173*, 24–30. https://doi.org/https://doi.org/10.1016/j.drugalcdep.2016.12.014

Sociologický ústav AV ČR. (2019). *Proměny české společnosti*.

Sovinová, H., Csémy, L., & Kernová, V. (2014). *Užívání tabáku a alkoholu v České republice: Zpráva o situaci za období posledních deseti let*. Státní zdravotní ústav.

World Health Organization. (2019). *Global status report on alcohol and health 2018.*

Author **Supervisor**

Chapter 1

Introduction

The dangers of alcohol consumption are clear. Alcohol abuse significantly increases the risk of liver disease, oropharynx, larynx, oesophagus, liver, colon, rectum and female breast cancers, cardiovascular disease, infectious disease and mental illness (World Health Organization 2019). The risk is especially elevated for higher doses of ethanol, but even relatively small amounts of alcohol lead to an increased risk of certain types of cancer (Pelucchi *et al.* 2011). Overall, around 5.3% of the total number of deaths in 2016 were attributable to alcohol consumption worldwide (World Health Organization 2019). Alcohol consumption also acutely increases the aggressiveness of its consumer by its impact on brain functioning, and it does so in a worse manner than any other psychoactive substance (Heinz *et al.* 2011). This alcohol related aggression is then strongly correlated with crime (Bushman 2002). Quite evidently, alcohol also increases the risk of accidents. In particular, Martin *et al.* (2017) have estimated that alcohol increases the risk of a fatal car accident by approximately 17.8 times.

Drinking behaviour shows heterogeneity across countries - there tend to be important differences between "drinking cultures" (Gordon *et al.* 2012). In the Czech Republic, alcohol consumption is very high. Although heavy episodic drinking is not as prevalent as in other countries in central and eastern Europe, frequent, steady consumption of alcohol makes the country's inhabitants one of the most intensive consumers of ethanol in the region (Popova *et al.* 2007). To quote more recent figures, according to 2019 Eurostat statistics, over 41% of Czechs drink weekly or more often and over 20% engage in heavy episodic drinking at least once per month (European Statistical Office 2019).

At the same time, alcohol consumption may be arguably very heteroge-

nous in social characteristics. In targeting campaigns and restrictions in order to limit the harmful effects of alcohol consumption, it is necessary for policy makers to know which individuals tend to show more risky drinking patterns and which ones are less likely to suffer from high alcohol consumption. While econometric modelling does not provide a definite and certain answer - be it because of sampling variation or unavoidable methodological imperfections - it might lead to more informed and qualified decisions.

This leads naturally to the question investigated in this thesis: What are the social and economic predictors of alcohol consumption in the Czech Republic? Compared to earlier studies investigating this question, namely Dzúrová *et al.* (2010) who looked at the inequalities in alcohol consumption from a hierarchical model perspective, this thesis approaches it from a slightly different, econometric viewpoint. Two separate, cross-sectional and panel datasets from the past decade are employed, each with its advantages and drawbacks. In the cross-sectional part, the focus is on an exploratory analysis with several models describing individual drinking behaviour, employing regional and time fixed effects for an increased level of robustness of the results. The multilevel (regional level random effects) approach of Dzúrová *et al.* (2010) is therefore abandoned. In the panel data part, the heart of the empirical investigation lies in modelling the time-varying predictors found in the first part using the Arellano-Bond estimator to get the estimates closer to actual causal effects. Furthermore, the panel analysis is followed by an extending analyses and robustness checks subsection, which employs further techniques from recent research to analyse the potential issues connected to the Arellano-Bond estimator. The main contribution of this thesis is therefore a more robust and thorough estimation of the effects of socio-economic predictors on alcohol consumption in the Czech Republic using recent data and methodologies which, to our knowledge, have not been used in this context before.

The results of the cross-sectional analysis show that in general, higher education, particularly amongst men, is related to lower alcohol consumption. On the other hand, unmarried individuals display a higher propensity to consume more alcohol. While income elasticity is low in the base model measuring the amount of alcohol consumed, it still is statistically significant, as opposed to the effect of unemployment. By contrast, for a model studying truly pathological drinking, the results imply that unemployment is a risk factor, while higher household income seems to be protective. The panel model then confirms the sign of the income elasticity of the base model, but with a larger point estimate.

As will be seen in the discussion of the results, while the panel model should be more robust to a specific type of omitted variable bias, there are other issues that arise in its estimation.

The thesis is structured as follows. First, the literature review briefly introduces the topic from the perspective of economic theory. Next, it proceeds by summarizing the most important empirical findings both in Europe and beyond. The aforementioned cross-sectional and panel analyses follow. Each of these is accompanied by a brief overview of the respective datasets. The panel analysis is followed by the aforementioned robustness check section. Lastly, the conclusion summarizes the main findings and proposes possible further research avenues in this topic.

Chapter 2

Literature review

Before presenting empirical analyses of the effect of socioeconomic variables on alcohol consumption, it is perhaps best to first briefly present the view of current economic theory on the consumption of addictive substances, to which alcohol undoubtedly belongs.

Traditionally, addiction has been seen in economics through the lens of the rational addiction theory (Becker & Murphy 1988). This theory treats addictive goods basically as any other goods, with the utility maximizing agent perfectly planning their consumption, taking into account the actual effects of the addictive substance with perfect foresight. This theory, while prevalent, has been subject to sharp criticism, given that its assumptions are arguably not realistic (see e.g. Rogeberg 2004). Nevertheless, additional extensions of the model, and new models entirely, have appeared in the years following the publication of Becker & Murphy (1988), to better reflect the reality of addiction.

Smith & Tasnádi (2007) attempt to reconcile economic theory with actual biological mechanisms underlying addiction. The authors introduce the concept of positive cues influencing the structure of an individual's consumption of addictive goods. The cues are biological responses of an individual when consuming a given good. On the basis of these cues, the individual forms beliefs about the nutrients contained in the good. Formally, the authors present the problem in the following way:

$$\begin{aligned} \max_{x,a} \quad & P\left(C_x x + C_a a \geq k\right) \\ \text{subject to} \quad & x + pa \leq m, \\ & x \geq 0, a \geq 0 \end{aligned} \tag{2.1}$$

Here a denotes the amount of the addictive good consumed, and x denotes

an alternative. C_i then represents the nutrients contained in the good i, m denotes the budget and p represents the normalized price of good a. If the dot product between the vector of the amount of goods consumed and the vector of their nutrient composition surpasses the threshold k, the individual survives. Thus, the objective is to maximise the probability of survival given a budget constraint with normalized prices.

However, the authors argue that for the addictive goods, the individual does not necessarily observe C_a, but is influenced by biological cues. The estimate under cues is $\hat{C}_a$, a random variable. Denote $v_t(x, a)$ the solution to the balanced diet problem at time t under the true C_a, and $\tilde{v}_t(x, a)$ the solution to a weighted average of the problem with $\hat{C}_a$ and of the problem with the true C_a. The weights are given by the individual's belief distribution between C_a and $\hat{C}_a$.

If there is a particular kind of mismatch between $v_t(x, a)$ and $\tilde{v}_t(x, a)$, a harmful addiction arises. Specifically, if it holds that

$$\frac{\partial}{\partial a}\tilde{v}_t(x, a) < \frac{\partial}{\partial a}\tilde{v}_{t+1}(x, a)$$

and for the consumed $(x_1, a_1), (x_2, a_2), \ldots$, it holds that

$$v_t(x_t, a_t) > v_{t+1}(x_{t+1}, a_{t+1})$$

there is a harmful addiction.

The authors note that this framework can be readily applied to alcohol.[1] Alcohol can be found naturally in ripe fruits, which were scarce and nutrient rich components of the diet of our ancestors. Biological cues therefore incentivized humans to consume alcohol when possible. In the current world however, alcohol is readily available and is purposefully manufactured. This makes it possible for humans to consume harmful amounts of alcohol, which could not have been found in foraged ripe fruit. At the same time, the genetic makeup of humans is very similar to that of their hunter-gatherer ancestors. Thus, the aforementioned mismatch arises, and, subsequently, alcohol addiction arises. Note that, although the theory is developed with nutritional cues in mind, the authors point out that social and other environmental cues also influence the composition of an individual's consumption. Furthermore, the belief distribution (basically the weight the individual gives to the cues) can

[1] The authors connect their theory primarily to opiates and opioids.

also be strongly influenced by social circumstances. Finding significant associations between socioeconomic predictors and alcohol consumption is therefore in line with this theory as well.

It is also interesting to note that a very similar theory is developed in a widely cited paper by Redish (2004), outside the realm of economics. The author uses a temporal difference reinforcement learning framework with dopamine signals from drug use, which is essentially equivalent to the utility maximisation framework with cues of Smith & Tasnádi (2007) (it is perhaps more narrow, given that other cues than those induced by dopamine are not considered).

Another way to deal with the criticisms of the rational addiction theory is through the concept of bounded rationality. Suranovic *et al.* (1999) present a model modifying Becker & Murphy (1988) by letting the individual choose only current consumption and thus make do without the assumption of perfect foresight. Furthermore, they introduce quitting costs. The model is then able to explain why some people want to quit their addiction, but cannot do so. This would be inconsistent with the rational addiction theory. Suranovic *et al.* (1999) tailor the model to cigarette consumption, but they note that it can be extended to alcohol and other addictive substances with minor adjustments.

The view presented by these theories is important for the interpretation of empirical studies estimating the relation between socioeconomic predictors and alcohol consumption. The theories of Smith & Tasnádi (2007) and Suranovic *et al.* (1999) essentially imply that alcohol should be directly influenced by individual socio-economic circumstances, no matter whether they influence alcohol consumption as cues and beliefs, or through preferences in particular forms of bounded rationality utility maximization problems with individual-specific quitting costs. Furthermore, given these socio-economic circumstances, individuals might consume a higher amount of alcohol than what would be optimal under a lifetime (objective) utility function maximization. In the following sections of the literature review, empirical studies investigating these potential relationships are presented.

2.1 Socioeconomic status and education

This section presents empirical literature studying the impact of either several socioeconomic predictors describing the socioeconomic status (SES) of an individual, or of a proxy variable representing the socioeconomic status, on alcohol consumption. Dzúrová *et al.* (2010) study the inequalities behind alcohol con-

sumption in the context of the Czech Republic. The analysis is done on two levels, individual and municipal. The model used by the authors is that of a logit random effects model at the individual level, with the random intercept allowed to be influenced by municipal characteristics. There are two separate outcome variables - one indicating whether an individual drinks alcohol twice a week or more ("frequent drinking"), and the other indicating whether they drink at least 5 drinks per occasion ("binge drinking"). The authors find that men, single people, unemployed people and people with lower educational attainment are especially at risk of these two types of behaviour in relation to alcohol.

Grittner *et al.* (2012) use a hieararchical model to study the effect of socioecnomic status on the *consequences* of alcohol consumption. This study is cross-national, so compared to the study of Dzúrová *et al.* (2010), the municipal level is exchanged for country level. Importantly, the authors choose the highest attained education level as a proxy for socioeconomic status. While this might be an oversimplification, it highlights the fact that education level, when contained in a regression model measuring alcohol consumption or problems related with it, cannot be directly interpreted as measuring the effect of education itself. Rather, it might be a combination of this effect and the effect of unobserved variables related to the person's SES. The authors conclude that people with lower SES are more likely to experience negative consequences of alcohol consumption, even after controlling for the *level* of consumption.

Peña *et al.* (2017) study the inequalities in alcohol consumption patterns based on the socioeconomic status in two countries with the highest alcohol consumption in their regions, Chile and Finland. Years of completed education are used as a proxy for SES. The authors employ the concentration index, a measure similar to the Gini index, to measure these inequalities.[2] Contrary to Dzúrová *et al.* (2010), they find that higher levels of drinking are more prevalent among people with higher SES, especially women. Finnish men aged 25 to 44 were the only group where lower SES was associated with higher alcohol consumption.

Another, slightly older paper using attained education as a proxy for SES is the study of Bloomfield *et al.* (2006). The authors use multinational survey data, with most included countries being EU members. The Czech Republic forms part of the country pool as well. The authors use separate logistic

[2]The authors define the index as $-2 \cdot Cov(X/\mu_x, 1 - G(Y))$, where X is the variable of interest (a measure of drinking), μ_x its mean, and $G(Y)$ is the CDF of SES.

regressions for each country and gender, which stands in contrast to the random effects (hierarchical) models used by Dzúrová *et al.* (2010) and Grittner *et al.* (2012), In general, quite heterogenous patterns are found across countries and genders. However, a fairly common result across countries is that men with lower SES tend to drink heavily more often than men with higher SES. For several countries, the Czech Republic being among them, heavy episodic drinking is also found to be associated with lower SES.

In what follows, literature investigating the effect of specific individual socioeconomic variables is presented.

2.2 Marital Status

Marital status arguably affects the lifestyle of an individual to a significant degree, and might therefore influence alcohol consumption as well. The observed literature suggests that, in general, marriage has a protective effect in relation to alcohol consumption, while divorce tends to be a risk factor. As we will see, the results come with important caveats.

Prescott & Kendler (2001) perform a latent growth analysis on a panel of female twins to investigate the effects of marital status on alcohol consumption behaviour. To find whether marital status is confounded by other factors in relation to drinking, the authors measured the effect of the marital status of a co-twin on the other co-twin. If there are hidden factors correlated between the twins that influence both marriage and alcohol consumption, there should be an appropriate effect. If not, there should be no effect. The results suggest that marriage and other changes that are related in time with marriage do produce a protective effect separate from that of the common environment and traits of the twins. The authors also find from the latent growth curves that increases in drinking associated with divorce are significant even before the divorce itself occurs. Furthermore, the divorce of a co-twin has a significant association with an increase in drinking. The authors suggest that there are family factors associated with a higher risk of divorce and a higher alcohol consumption.

Kendler *et al.* (2017) perform a survival analysis study on Swedish panel data on alcohol use disorder. They study primarily the effect of divorce, but also first marriage and remarriage. For divorce, the authors find a significant increase in risk, with a hazard ratio of around 6 for men and 7 for women. Similarly to Prescott & Kendler (2001), the authors then perform this analysis for monozygotic twins, thus entirely controlling for genetic differences (monozy-

gotic twins share 100% of their genes) and partially controlling for environmental differences between individuals. The effect is still significant but the hazard ratio falls to about 3.5 for both sexes. The authors also found that both remarriage and first marriage are associated with a decline of risk of alcohol use disorder, suggesting that the marriage itself or the related time-varying variables have a causal effect (otherwise remarriage would have no effect), although the effect associated with remarriage is lower than that of the first marriage.

In a cross-sectional study on American individual data, Ellison *et al.* (2008) point out that the differences in drinking behaviour between married and unmarried people might be related more to the religiosity of the individuals, rather than their actual marital status. Because of various norms and ethical rules, religious people tend to drink less. The authors point out that religious people also tend to get married earlier and more often than non-religious individuals, thus being more prevalent among the married couples than as single people in an adult-only data sample on alcohol consumption. Secondly, they show evidence that couples where both members are religiously conservative drink less - possibly due to the fact that they reinforce and monitor each other in following the norms - while homogamous non-religious couples do not show a reduction in drinking in comparison to single people. In the Czech or even European context, these relationships might not be as strong, given the smaller share (relative to the US) of what Ellison *et al.* (2008) call proscriptive denominations, such as evangelical Protestants, Mormons and Jehovah's witnesses, who have ethical rules against alcohol drinking. Nevertheless, the study shows the importance of controlling for religion as a confounder.

Lastly, Tamers *et al.* (2014) study the impact of stressful life events on excessive alcohol use in a cohort study of French individuals. They estimate trajectories of heavy alcohol use around the time of these events. Surprisingly, they find that alcohol use decreases both before marriage and divorce, in anticipation of these events, and increases *after* marriage and divorce. The authors do not find comparative effects for widowhood.

2.3 Income and unemployment

A large part of the literature on the effect of income and unemployment on alcohol consumption focuses on aggregate economic conditions (e.g. recessions), rather than household/individual income and unemployment status. The interest often lies mainly in discovering whether alcohol consumption is pro- or

counter-cyclical, rather than in individual behaviour. This is not the case of this thesis, nevertheless, even such studies might serve as useful inspiration.

Dávalos *et al.* (2012) use a two-wave panel dataset collected in the USA. They model the dependence of risky alcohol-related behaviour on the unemployment rate. Even though the independent variable of interest is measured at the level of states, thanks to the panel dataset, the authors have access to individual information on alcohol consumption and related behaviour. This allows them to use conditional fixed-effects models to control for unobserved heterogeneity which might potentially be correlated with the predictors. The authors find that state unemployment rates increase the odds of binge drinking, driving under the influence, and alcohol abuse. Therefore, alcohol consumption, or at least the related behaviours, are counter-cyclical.

A very different conclusion is reached in Cotti *et al.* (2015) who use a US time-aggregated high-frequency household-level panel dataset on alcohol purchases in shops (i.e. excluding purchases in bars and similar premises). The authors regress these purchases on state-level unemployment and income using a dynamic panel model. They find that alcohol purchases negatively depend on unemployment, and thus are pro-cyclical. Furthermore, the authors note that failing to account for persistence in alcohol consumption might lead to biased estimates. The authors estimate the effect of the Great Recession to be a decrease of around 220 grams[3] in annual alcohol consumption per household.

By contrast, Popovici & French (2013) study the effects of *individual* unemployment on alcohol consumption. This comes closer to the approach used in this thesis, which concentrates on individual behaviour rather than on aggregate economic conditions. The authors employ the same data on alcohol consumption as Dávalos *et al.* (2012), however they use the unemployment status of the respondents, rather than the state unemployment rate, as the independent variable of interest. The authors perform a fixed-effects regression to conclude that unemployment is a significant risk factor in relation to alcohol consumption. The magnitude of this effect is lower than in the case of a pooled regression on the same data. This points to the possibility that controlling for individual heterogeneity might be key to minimizing bias in similar regression estimates. The authors conclude that high unemployment might generate further societal costs in the form of poor health.

Henkel (2011) summarizes a multitude of older studies relating to substance use and unemployment. In the section talking about the effects of unemploy-

[3]Equivalent of 7.8 ounces actually reported in the American study.

ment on alcohol consumption, the author puts forward two possible hypotheses about the direction of this effect. First, the psychosocial stress related to unemployment can increase alcohol consumption. The author names "financial strain, depression, identity crises, monotony, sleep disorders, and loss of social support" as examples of the manifestations of this type of stress. Second, without work, there is an absence of work-related stress and a decrease in income. These might result in a decrease in alcohol consumption. Out of fourteen studies concentrating on the topic and presented in the paper, nine found an increase in alcohol use related to unemployment.

The literature landscape for income and alcohol consumption is similar. Again, most studies concentrate on estimating income elasticity of alcohol from macro-level data. The papers are indeed numerous, and Nelson (2013) provides a thorough meta-analysis summarising them. After correcting for publication bias and outliers, the author estimates an income elasticity for general alcohol consumption of around 0.6. This would therefore suggest that alcohol is a normal good with relatively low elasticity. That does not mean there are no studies pointing to alcohol being an inferior good, however. For example, Volland (2012) employs a gradual switching model on German macroeconomic data to find that while beer used to be a normal good before 1965, it became inferior by 2004 with a coefficient of -0.59. This result is obviously beverage-specific and beer could be easily substituted by wine or spirits, but it shows that the income elasticity of alcohol might be significantly heterogenous and time-varying. Furthermore, it has been suggested in the literature that women are more sensitive to changes in prices and income in relation to alcohol consumption - possibly because men tend to be more "comitted" drinkers, while women tend to be more "casual" ones (Decker & Schwartz 2000).

In the Czech context, Grosová *et al.* (2017) estimate a model of price and income elasticities, separately for keg and bottled/canned beer consumption. They find a low and statistically insignificant income elasticity of 0.08 for tap beer and a negative income elasticity of around -0.8 for bottled beer, pointing to the possibility of bottled beer being an inferior good in the Czech market. An earlier study by Janda *et al.* (2010) uses a Czech individual-level panel dataset to estimate price and income elasticities for beer, wine and spirits through the use of the Almost Ideal Demand System. They estimate significant coefficients of 0.98, 0.56, 0.35, for beer, wine and spirits, respectively. The authors explain the counterintuitive results (with wine being less income elastic than beer) by stating that while wine is mostly consumed at home, beer tends

to be consumed in restaurants. The data support that the income elasticity for restaurant drinks is higher, in line with Grosová *et al.* (2017). It is important to note that while the authors do use individual-level data, they do not include any individual socio-economic variables other than income that would at least partially control for confounding.

A recent international study by Rousselière *et al.* (2021) also uses individual-level alcohol consumption data for beer, wine and spirits, and relates it to prices and income. The data is taken from a survey of individuals from 21 European countries. The authors however do not have access to individual-level income and use GDP per capita of each country as a replacement. Using a multi-equation generalized Heckman model, the authors conclude that the income elasticity is approximately equal to 0.58 for beer, 0.64 for wine, and 0.1 for spirits. An estimate for total alcohol consumption is not presented.

To conclude this chapter, a multitude of studies show that social, as well as purely economic predictors, do play a role in influencing an individual's consumption. The results often point in different directions. Some suggest that lower income, lower education, unemployment and lack of marriage imply a higher level of alcohol consumption, consistently with the common sense idea that social deprivation is associated with higher levels of alcohol consumption as well as higher prevalence of risky alcohol use. Yet other studies, such as Peña *et al.* (2017) or Cotti *et al.* (2015) suggest opposite effects. It can be concluded that these effects depend heavily on the social context and the methodology of the researchers. This thesis will attempt to reveal how these relationships have looked like in the Czech context during the past decade.

Chapter 3

Cross-sectional analysis

3.1 Empirical methodology

Three empirical models are employed to study three types of alcohol consumption patterns. The first model - the consumption and participation model focuses on the amount of alcohol consumed in grams per year. The second model studies excess/binge drinking as a separate phenomenon. Lastly, the problem drinking model concentrates on modelling risky behaviour in drinking, taking into account both the total amount consumed and the binge drinking patterns. In all cases, although the dependent variable and the modelling method changes, the linear predictor is composed of the same variables. That is to say, for each model $\mathbb{E}(y_i|\mathbf{x_i}) = g^{-1}(\mathbf{x_i}^\intercal\boldsymbol{\beta})$, y_i and its assumed distribution (if any), the link function $g(\cdot)$ and obviously the parameters $\boldsymbol{\beta}$ might be different, but the predictors $\mathbf{x_i}$ will be identical.

The predictors of interest are mainly education level, marital status, household income, unemployment status and religiosity. Control variables then include sex, age (including a squared and a cubic term), number of people living in the household, number of children living in the household, size of the municipality where the individual lives, region and year fixed effects, and regional time trends (interactions between region and year).

The aim of including the control variables is in minimizing the omitted variable bias. It is clear that men drink more alcohol than women (World Health Organization 2019) or that consumption might vary with age, however, age and sex/gender are also almost certainly correlated with the predictors of interest. Regional fixed effects and population size of municipality control for several possible confounding factors. Firstly, there might be locally concen-

trated social deprivation in the area where the individual lives, affecting both socioeconomic variables and alcohol consumption patterns. Secondly, cultural norms and habits might again affect both the left-hand and the right-hand side variables of the regression equations. This variation might be expected even in such a small country as the Czech Republic. One might easily expect that the socioeconomic conditions are very different between, for example, cosmopolitan Prague and traditional villages in rural southeastern Moravia. At the same time, the drinking habits might also differ - to continue with our example, bar drinking in Prague and drinking of homemade spirits in the Carpathian mountains can result in very different patterns of alcohol consumption. Indeed, the literature review of Castro *et al.* (2014) points out that the link between culture and alcohol consumption has been shown to be significant in numerous studies, one of the dominant cultural factors being the level of modernization of a given society.

Year dummies and their interactions with regional dummies are included mainly to capture common shocks that might affect both the alcohol consumption patterns and the socioeconomic environment. For example, current macroeconomic theory asserts that in the short-term, inflation affects not only the prices of consumer goods (including alcohol) but also individual economic status (unemployment status). From a microeconomic perspective, the prices of substitutes might change over time vis-a-vis the prices of alcohol, thus possibly changing the income elasticity of alcohol. Given the size of the country (and thus ease of arbitrage), prices in the Czech Republic should not vary much if we control for time, regional, and big town-small town differences. Clearly, having prices of the alcoholic products consumed by each individual would be even better, since the prices of different varieties of alcohol (for example, cheap vodka vs. high quality whisky) might vary independently of each other, and each individual might have a different "alcohol consumer basket". Unfortunately, this kind of data is close to being infeasible to obtain at the individual level.

The interest of including the number of people and children in a household is mainly for ease of interpretation of the coefficient related to household income. The effect of household income might be very different if a single person or a six-member family live on it.

As has been stated in the literature review with regards to similar studies, given the observational and cross-sectional nature of the data, the effects of the variables of interest must be taken with a grain of salt, despite the above

controls. The controls cannot for example contain hidden characteristics, such as intelligence, ability, and others. However, it is not the ambition of this chapter to uncover any causal effects. The interest rather lies in identifying characteristics which predict alcohol use. That is to say, we cannot expect that manipulating any of the variables exogenously would cause alcohol consumption and its patterns to change according to the estimated effect. However, thanks to the models, groups of individuals with a higher alcohol consumption or risky patterns of alcohol use can be identified. Moreover, the robustness of the effects can be tested in further studies. This thesis will attempt to get closer to estimating causal effects for two variables in the next chapter.

3.1.1 Consumption and participation model

The baseline model focuses on the effects of socioeconomic predictors on the amount of alcohol consumed. Consumption of almost any good might contain an important number of zeroes. There are essentially two primary ways of solving this: either use a standard linear regression model that will serve as a crude approximation to the conditional expected value $\mathbb{E}(y_i|\boldsymbol{x_i})$, or use some form of a two-part model approach. The thesis will use the latter. There are two reasons for this.

Firstly, we want to accurately model the distribution of the dependent variable. Given that it contains exact zeroes, the distribution cannot be continuous, since that would imply $P(y_i = 0) = 0$. Rather, the distribution is a mixture of a continuous distribution for positive values, and a discrete distribution for zeroes.

Secondly, we assume that different social and economic relationships might influence the decision on consuming alcohol at all - whether to abstain or not - and the decision on how much to consume. In a review of surveys on alcohol abstinence, Rosansky & Rosenberg (2020) show that the most common reasons for abstaining are a lack of interest in alcohol consumption at all, not liking the effects of alcohol (mostly lifelong abstainers) and health issues (mostly former heavy drinkers). These probably differ at least partially from the reasons determining how much people drink. Most importantly, the group of abstainers is very heterogenous. It includes both people who do not like drinking, these might indeed share characteristics with people who drink only little, and former heavy drinkers. These share characteristics with current heavy drinkers whose alcohol dependence has not yet led to abstinence (or to death). This also means

that models like Tobit which assume that there is an underlying latent variable, which, if it falls below a certain threshold, produces a zero in the data, cannot be used, since this latent variable cannot model the aforementioned heterogeneity of the abstainers group.

We are therefore interested in modelling separately: 1. the probability of participation, 2. the amount of consumption given that consumption is positive. The problem at hand naturally leads to the following conditional density (Cameron & Trivedi 2005, p. 545):

$$f(y_i|\boldsymbol{x_i}) = \begin{cases} P(d_i = 0|\boldsymbol{x_i}) & \text{if } y_i = 0 \\ P(d_i = 1|\boldsymbol{x_i}) f(y_i|d_i = 1, \boldsymbol{x_i}) & \text{if } y_i > 0 \end{cases} \tag{3.1}$$

Here $d_i = 0$ indicates non-participation in alcohol consumption (i.e. abstinence) and $d_i = 1$ indicates participation. Given our data, it seems that the best fit for $f(y_i|d_i = 1, \boldsymbol{x_i}) = f(y_i|y_i > 0, \boldsymbol{x_i})$ is the lognormal distribution (see figure 3.2). To model the probability of participation, the logit model will be used for ease of interpretation of odds ratios. On that note, recall that the logit model represents $log\left(\frac{p}{1-p}\right) = \boldsymbol{x_i}^\intercal \boldsymbol{\gamma} \iff \frac{p}{1-p} = e^{\boldsymbol{x_i}^\intercal \gamma}$ where p is the probability of participation. Therefore interpretation is straightforward and no more calculations are necessary after exponentiating the coefficients.

In this case, the log-likelihood can be written as follows (Hsu & Liu 2008, probit exchanged for logit):

$$\begin{aligned} ll(\boldsymbol{\theta}) = &\sum_i^N \mathbb{I}[y_i = 0] \log(1 - \Lambda(\boldsymbol{x_i}^\intercal \boldsymbol{\gamma})) + \\ &\sum_i^N \mathbb{I}[y_i > 0] \left\{ \log(\Lambda(\boldsymbol{x_i}^\intercal \boldsymbol{\gamma})) + \log(f_{\mathcal{N}}(log(y_i)|\boldsymbol{x_i}^\intercal \boldsymbol{\beta}, \sigma_i^2)) \right\} \end{aligned} \tag{3.2}$$

Here $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma)^T$, $\Lambda(x) = \frac{1}{1+e^{-x}}$ is the CDF of the logistic distribution with the location parameter equal to 0 and the scale parameter equal to 1, $f_{\mathcal{N}}(\cdot|\mu, \sigma)$ is the density of the normal distribution with mean μ and variance σ^2, and $\mathbb{I}[\cdot]$ is the indicator function. It is easily seen that the log-likelihood can be maximised separately on coordinates $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. This gives a standard logit model for participation and a log-normal model for consumption. The latter can be estimated by OLS, with $\log(y_i)$ as the dependent variable. On a similar note,

$$\frac{\partial^2}{\partial \beta_i \partial \gamma_j} ll(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \gamma_j \partial \beta_i} ll(\boldsymbol{\theta}) = 0 \tag{3.3}$$

for any i, j and thus the asymptotic variance of the estimates of the parameters $\boldsymbol{\beta}$ does not depend on $\boldsymbol{\gamma}$ (and vice versa).[1] In any case, robust standard errors are used for the consumption model.

It is important to note that the separated nature of the model implies that the parameters $\boldsymbol{\gamma}$ only describe the conditional probability of participation, and the parameters $\boldsymbol{\beta}$ only describe the conditional expected value $\mathbb{E}(y_i|y_i > 0, \mathbf{x_i})$, *not* the general $\mathbb{E}(y_i|\mathbf{x_i})$. As explained above, this is a feature, not a bug.

Lastly, note that more complex models could also be used (such as sample selection models) which allow for correlation between the errors of the two modelled decisions. Following Madden (2008), it seems however more appropriate to estimate the two-part model, since the sample selection models require exclusion restrictions (we would need certain variables to influence only the participation decision and not the consumption decision) which we cannot *a priori* identify in our data. Even if such restrictions could be found, the two types of models often perform similarly (Smutna & Scasny 2017) or the sample selection model could even perform worse in certain cases (Madden 2008).

3.1.2 Binge drinking & Problem drinking models

In the binge drinking model, the interest is in capturing the predictors of heavy episodic drinking. Following the descriptive study of Csémy *et al.* (2021), four categories of average binge drinking frequency are assigned to each individual: never, one to eleven times a year, one to three times a month, and weekly and more often.

For the problem drinking model, information on average alcohol intake is combined with binge drinking categories to form a comprehensive measure of the riskiness of an individual's drinking habits, which is proposed by Csémy *et al.* (2021) and which the authors call the alcohol consumer pyramid. The pyramid consists of moderate drinkers (and abstainers - which are excluded in this part of the analysis, since they are already treated in the participation model), medium risk drinkers, high risk drinkers, and problem drinkers. The categories are assigned based on the rules in table 3.1.

[1] The Hessian is block-diagonal and thus its inverse can be computed by inverting its blocks.

Table 3.1: Alcohol consumer pyramid

Alcohol consumption	Binge drinking	Pyramid
moderate	never	moderate drinker
moderate	yearly or monthly	medium risk drinker
moderate	weekly or more	high risk drinker
risky	never	medium risk drinker
risky	yearly or more	high risk drinker
harmful	never or yearly	high risk drinker
harmful	monthly or more	problem drinker

Note: The table displays the rules for classification of drinkers into overall risk categories based on their alcohol consumption and binge drinking habits. Note that according to Csémy *et al.* (2021), moderate alcohol consumption of less than 20 grams for women and 40 grams for men per day on average is considered to be moderate, more than 20 g and less than 40 g for women and more than 40 g and less than 60 g for men is considered to be risky, and more than that is considered harmful.

For both models, the dependent variables can be naturally ordered. While a multinomial logistic model could potentially be used, mere pairwise logits of the base category vs. the other categories would throw away the valuable structure of the dependent variable. Another option is the ordinal logit model. Cameron & Trivedi (2005) define this model by means of a latent variable:

$$y_i^* = \boldsymbol{x_i}^\intercal \boldsymbol{\beta} + u_i \tag{3.4}$$

for $i = 1, \ldots, N$. This latent variable part of the model does not include an intercept. Then, if the observed dependent variable $y_i = j,\ j = 1, \ldots, K$ describes that individual i belongs to the j-th category, the model can be expressed as follows:

$$P(y_i = j) = P(\alpha_{j-1} \leq y_i^* \leq \alpha_j) \tag{3.5}$$

That is to say, the individual belonging to the j-th category is equivalent to the value of the latent variable being between some thresholds α_{j-1} and α_j, where $\alpha_0 = -\infty$ and $\alpha_K = \infty$. The parameters $\boldsymbol{\alpha}$ are, along with the regression parameters $\boldsymbol{\beta}$, estimated using maximum likelihood.

The ordinal logit model is based on an important assumption: the so-called proportional odds assumption (Agresti 2010). This means that a single set of parameters is used to model decisions between each of the adjacent categories,

i.e. we assume:

$$P(y_i = j) = P(\alpha_{j-1} \leq \boldsymbol{x_i}^\intercal \boldsymbol{\beta} + u_i \leq \alpha_j) \tag{3.6}$$

instead of:

$$P(y_i = j) = P(\alpha_{j-1} \leq \boldsymbol{x_i}^\intercal \boldsymbol{\beta_j} + u_i \leq \alpha_j) \tag{3.7}$$

This assumption is however rejected for both of our models by the Brant (1990) test which is standardly used for testing this assumption. Therefore another model is needed, one that would preserve the ordinal information and the same time would not necessitate the proportional odds assumption. The model 3.7 is problematic, because the latent curves can cross for different categories (Agresti 2010). Instead, Agresti (2010) defines the so-called adjacent categories with non-proportional odds model:

$$P(y_i = j | y_i = j \vee y_i = j - 1) = \Lambda(a_j + \boldsymbol{x_i}^\intercal \boldsymbol{\beta_j}),\ j = 2, \ldots, K \tag{3.8}$$

where $\Lambda(x) = \frac{1}{1+e^{-x}}$. Or, in another form:

$$log\left(\frac{P(y_i = j)}{P(y_i = j - 1)}\right) = a_j + \boldsymbol{x_i}^\intercal \boldsymbol{\beta_j} \tag{3.9}$$

Simply put, the model describes the probability of switching between adjacent categories. This model can also be estimated by maximum likelihood.

3.2 Data

The data used in this chapter are taken from a series of surveys on alcohol and tobacco use in the Czech Republic for years 2012, 2014, 2016, 2018, 2019, and 2020 (Sovinová & Csémy (2013), Sovinová & Csémy (2015), Váňová *et al.* (2017), Csémy *et al.* (2019), Csémy *et al.* (2020), Csémy *et al.* (2021)). These are cross-sectional studies performed each year on a different random sample. They collect information on the average frequency of alcohol consumption and the average amount (in glasses) consumed by the sampled individuals in a given year. The frequencies and amounts are beverage-specific, i.e. they are recorded separately for beer, wine and spirits. On top of that, the sampled individuals were asked to record how often they consumed alcohol excessively, which was defined as consuming more than three standard glasses of any of the three types of beverages.

In all models, the sample is restricted to the working age population between 25 and 65 years old. The lower bound is set mostly so that only people with a finished education are taken into account. The interest is not in modelling the change of alcohol consumption patterns amongst students and people with finished education, but rather in studying the effect associated with a static, lifelong education status. Unemployment status and income are also more meaningful if most of the sample is part of the labour force. This is also the reason why seniors are excluded as well.

The detailed structure of the survey questions allows to calculate the average amount of alcohol consumed per year fairly precisely. In calculating this amount, we follow the beverage specific quantity-frequency method of Moskalewicz & Sierosławski (2010), similarly to Csémy *et al.* (2021). Specifically, the following formula is used for each individual:

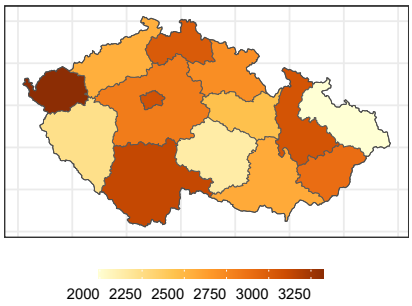
$$consumption = \sum_{k \in \{b,w,l\}} f_k \cdot a_k \cdot v_k \cdot abv_k \cdot 0.8 \quad g \cdot day^{-1} \tag{3.10}$$

where b, w, l are indices for beer, wine and spirits, f_k is the yearly frequency of consuming beverage k, a_k is the average number of glasses of beverage k consumed on one occasion, v_k is the volume of a standard glass of the beverage, abv_k is the standard alcohol by volume percentage for the beverage, and 0.8 is the density of ethanol in grams per milliliter. The volume of a standard glass was taken to be: 500 ml for beer, 200 ml for wine and 50 ml for spirits (if the survey participants drank smaller or larger glasses, they were asked to convert their consumption into glasses of the specified volumes). The ABV for beer, wine and spirits was taken to be 4%, 12% and 40%, respectively.

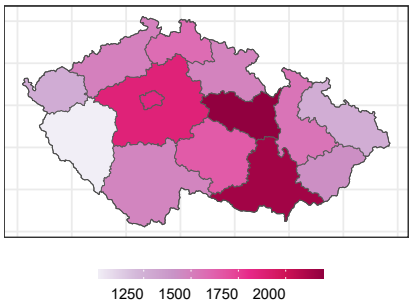
The importance of the beverage-specific approach is illustrated in figure 3.1. It can be seen that the distribution of the consumption of various alcohol drinks is very heterogenous over the Czech territory. While wine consumption is most prominent in the Pardubice region and in Southern Moravia, spirits is consumed the most in the Olomouc and Zlin regions. Beer consumption varies over the whole territory. All of these heterogeneities are statistically significant ($F(13, 7411) = 2.06, p = 0.014$ for beer, $F(13, 7411) = 3.12, p = 0.00012$ for wine, $F(13, 7411) = 2.94, p = 0.00027$ for spirits).[2] While we control for regional fixed effects in the models presented in the previous section, the distributions can also be heterogenous across societal groups. Under- or over-estimating consumption based on a uniform approach across beverages could therefore lead to bias.

Figure 3.1: Average alcohol consumption per region in $g \cdot year^{-1}$: beer, wine, spirits respectively

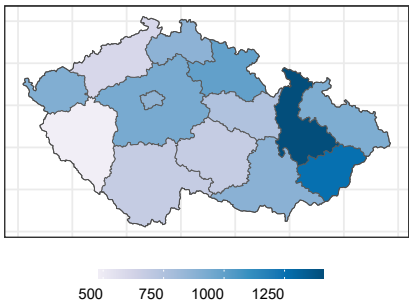
(a) Beer consumption



(b) Wine consumption



(c) Spirits consumption



Note: The maps show beverage-specific distributions of alcohol consumption in grams per year over 14 Czech regions (kraje).

The distribution of alcohol consumption of non-abstainers, at least in the present sample, is extremely positively skewed, non-symmetric and contains natural extreme values. This is readily apparent from table 3.2. The value in levels is therefore unsuitable for direct use in the consumption model, which is estimated by OLS. Accordingly, it is more suitable to transform the data using the natural logarithm. The histogram in figure 3.2 shows the transformed data. While the distribution might still be non-normal, given that the histogram is showing slight negative skew, it is certainly more well-behaved, with fewer apparent extreme values. Consequently, the mean, and therefore also most likely the modelled conditional mean, is more representative of the whole distribution.

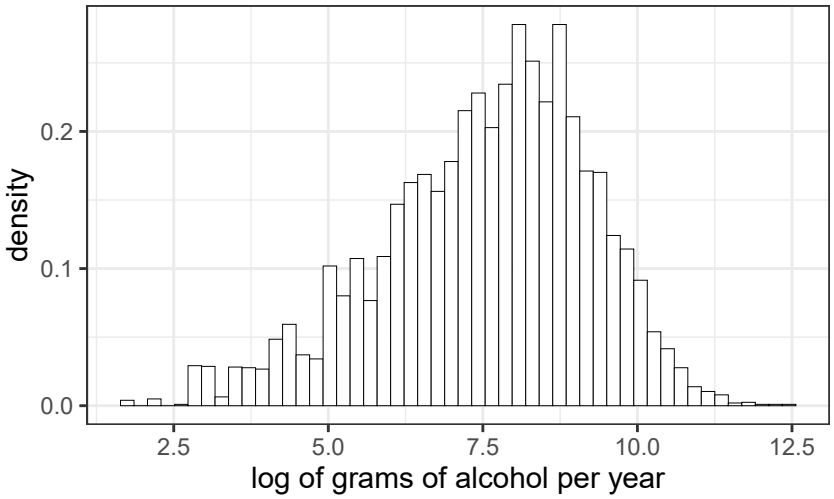
Table 3.2: Descriptive statistics for alcohol consumption of non-abstainers

Min	1st quartile	Median	Mean	3rd quartile	Max	Skewness
6	646	2400	5731	6406	278224	8

Note: The descriptive statistics show extreme positive skew for the untransformed data.

[2]The heterogeneity was tested using the following ANOVA model: $y_{ij} = \alpha + \beta_j + \varepsilon_{ij}$ where y_{ij} is the consumption of individual i in region j, α is the overall mean, β_j is the regional deviation from the mean, and ε_{ij} is the individual error term.

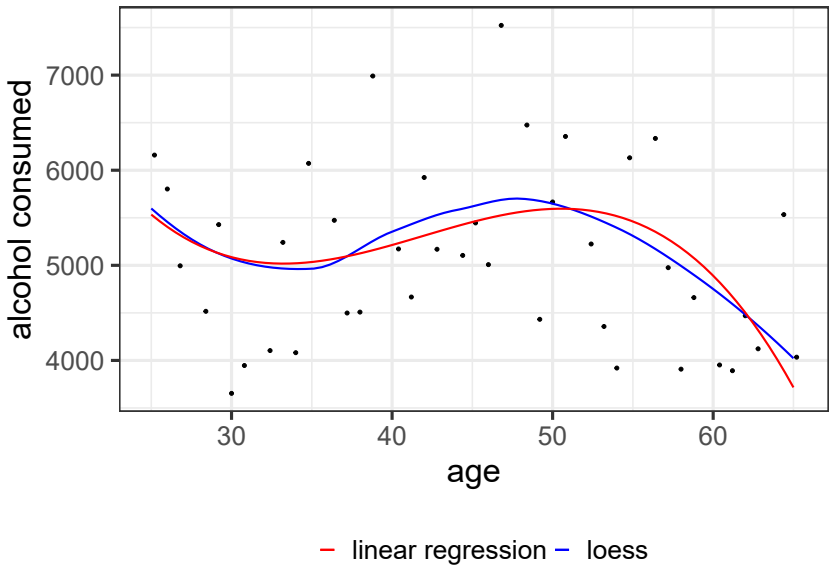
Figure 3.2: Histogram of log of alcohol consumption in $g \cdot year^{-1}$



Note: The logarithmized data display a more well-behaved distribution than the untransformed data. Histogram bandwidth was calibrated using the Freedman & Diaconis (1981) rule.

Some of the independent variables are transformed as well. For the ordinal logit models to converge, it has been necessary to standardize the age variable (subtract the mean and divide by the standard deviation) otherwise the square and the cube of the variable would be too large. A polynomial of the third degree has been chosen to model age since it seems to fit the data quite well if we regress only the amount of alcohol consumed on age. A comparison of a nonparametric local regression model (loess) with a third degree polynomial regression model is presented in figure 3.3. The results are indeed quite close. Furthermore, a cubic polynomial still allows for a concise model without a high risk of overfitting.

Figure 3.3: Dependence of average alcohol consumed $g \cdot year^{-1}$ on age



Note: The plot shows the dependence of the amount of alcohol consumed on age. A comparison of a nonparametric local regression and third degree polynomial regression is provided. For good legibility, the dots represent the means for 50 bins of data based on age. The curves are fitted on original (non-aggregated) data.

The household income variable has been transformed from a categorical one into a continuous one. Even though this introduces a degree of measurement error, in the raw data, there are eight income categories - keeping them untransformed would be problematic for model interpretation. The middle value of each category is used instead. Since income tends to be strongly skewed, and this is the case in the data as well, the income variable is log transformed. This has the added benefit of our being able to interpret the relevant regression coefficient in the estimated consumption equation as an elasticity. All other variables remain untransformed.

3.3 Results

Let us begin with the results for the main model of interest, the consumption model, i.e. the model $\mathbb{E}(y_i|y_i > 0, \mathbf{x_i})$ where y_i is the amount of alcohol consumed per year, on average. The table 3.3 shows the regression coefficients, standard errors, exponentiated coefficients and their 95% confidence intervals, and the p-value. In all models, for factor variables indicating marital status, education and religiosity, the reference levels are married, primary education and non-religious, respectively. Household size and number of children in the household, size of municipality and regional and time fixed effects and their interactions are not shown for conciseness.

Table 3.3: Consumption model results

	Estimate	SE	$e^{estimate}$	95% CI	p value
male	0.98	0.04	2.68	(2.48, 2.89)	0.00
age	0.25	0.08	1.28	(1.1, 1.48)	0.00
age^2	-0.31	0.06	0.73	(0.65, 0.82)	0.00
age^3	-0.31	0.08	0.73	(0.62, 0.86)	0.00
unemployed	-0.07	0.17	0.93	(0.66, 1.31)	0.68
log(income)	0.10	0.04	-	(0.02, 0.18)*	0.02
single	0.37	0.07	1.45	(1.27, 1.65)	0.00
in a relationship	0.35	0.10	1.42	(1.17, 1.73)	0.00
divorced	0.22	0.06	1.24	(1.1, 1.4)	0.00
widowed	-0.03	0.11	0.97	(0.79, 1.2)	0.77
secondary education	-0.10	0.05	0.90	(0.82, 0.99)	0.03
tertiary education	-0.23	0.06	0.80	(0.71, 0.89)	0.00
religious	-0.05	0.05	0.95	(0.86, 1.04)	0.26
member of a church	-0.15	0.08	0.86	(0.74, 0.99)	0.04
N = 6476					

The table shows the point estimate, standard error, exponentiated point estimate, its 95 % confidence interval (*for the log-transformed income the CI is untransformed) and the p-value for the consumption model. The dependent variable is the logarithm of the amount of consumed alcohol. For factor variables indicating marital status, education and religiosity, the base levels are married, primary education and non-religious, respectively. Household size, number of children, size of municipality and regional and time fixed effects and their interactions are not shown for conciseness.

The demographic variables of gender and age display a strong and significant influence on the dependent variable. It is therefore reassuring that these have been included as the primary control variables. As expected, the alcohol consumption of men is much higher than that of women, keeping other variables equal. Values of 2.48 to 2.89 times more all lie in the 95% confidence

interval. Age also plays a significant role and seems to have the assumed cubic relationship presented above (recall that the variable is standardized when looking at the magnitude of the coefficients).

Economic variables, that is unemployment and income, display varying results. The effect of unemployment is estimated very imprecisely, with a large confidence interval, and the sign of the effect cannot be reliably determined, even though the point estimate is negative. On the other hand, on average, a one percent higher income should correspond to a 0.1 % raise in alcohol consumption. While the effect shows statistical significance, the lower bound of the confidence interval approaches zero and the economic effect is modest at best. From these results, it could be judged that alcohol displays income inelasticity. At the same time, it seems to be a normal good, not an inferior one.

All marital statuses except widowhood seem to be risk factors as compared to being married. Most notably, being in a relationship is correlated with alcohol consumption almost as strongly as being single. The model therefore indicates that marriage in itself is a protective factor in relation to alcohol consumption. It is to be hypothesized whether this is due to the institution of marriage itself, or whether a more stable relationship (which more often ends in marriage) induces more modest drinking. However, it seems that the findings of Ellison *et al.* (2008) that the effect of marriage on alcohol consumption is mostly due to married couples being more likely to be religious (see literature review) does not hold in the Czech context, since here a distinct effect of marriage can be observed even after controlling for religiosity. Religiosity displays a modest protective effect, however the sign of the effect can be reliably determined only if the individual is also a member of a church. This is notably only a small portion of the Czech population (Furstova *et al.* 2021).

Furthermore, our data confirm that higher educational attainment is related with a significantly lower alcohol consumption. University education seems to have an even greater effect in absolute value than only high school education, although the difference between the point estimates is not itself statistically significant.

The model has also been applied separately for men and women, given that women and men might have very different attitudes towards alcohol consumption (this is also done in a large portion of the studies presented in the literature review). The results can be seen in table 3.4. Interesting differences appear between men and women, although most are not statistically significant. The sign of the point estimate of the effect of unemployment is reversed for women

Table 3.4: Consumption model results: male and female comparison

	Estimate	SE	$e^{estimate}$	95% CI	p value
			Female		
age	0.12	0.11	1.13	(0.91, 1.41)	0.27
age^2	-0.44	0.09	0.65	(0.54, 0.77)	0.00
age^3	-0.30	0.12	0.74	(0.58, 0.94)	0.01
unemployed	0.06	0.27	1.06	(0.62, 1.8)	0.83
log(income)	0.18	0.06	-	(0.06,0.3)*	0.00
single	0.62	0.10	1.86	(1.53, 2.25)	0.00
in a relationship	0.33	0.16	1.40	(1.03, 1.89)	0.03
divorced	0.33	0.09	1.39	(1.18, 1.65)	0.00
widowed	0.10	0.14	1.11	(0.84, 1.47)	0.46
secondary education	0.06	0.07	1.06	(0.92, 1.23)	0.40
tertiary education	-0.09	0.09	0.92	(0.77, 1.09)	0.33
religious	0.00	0.07	1.00	(0.87, 1.15)	1.00
member of a church	-0.09	0.10	0.91	(0.74, 1.11)	0.36
			Male		
age	0.26	0.10	1.30	(1.07, 1.59)	0.01
age^2	-0.20	0.08	0.82	(0.7, 0.96)	0.01
age^3	-0.20	0.11	0.82	(0.66, 1.01)	0.06
unemployed	-0.19	0.22	0.83	(0.54, 1.27)	0.39
log(income)	0.01	0.06	-	(-0.1,0.13)*	0.91
single	0.22	0.09	1.24	(1.05, 1.47)	0.01
in a relationship	0.38	0.13	1.46	(1.14, 1.88)	0.00
divorced	0.15	0.08	1.16	(0.99, 1.36)	0.07
widowed	-0.08	0.16	0.92	(0.67, 1.26)	0.60
secondary education	-0.22	0.06	0.80	(0.71, 0.9)	0.00
tertiary education	-0.32	0.07	0.73	(0.63, 0.84)	0.00
religious	-0.07	0.06	0.93	(0.82, 1.06)	0.29
member of a church	-0.20	0.11	0.82	(0.66, 1.02)	0.07
$N_{female} = 3141,\ N_{male} = 3335$					

The table shows the point estimate, standard error, the odds ratio ($e^{estimate}$), its 95 % confidence interval (*for the log-transformed income the CI is untransformed) and the p-value for the consumption model with the female and male subsamples. The dependent variable is the logarithm of the amount of consumed alcohol. For factor variables indicating marital status, education and religiosity, the base levels are married, primary education and non-religious, respectively. Household size, number of children, size of municipality and regional and time fixed effects and their interactions are not shown for conciseness.

as compared to men and the whole sample. Again however, the confidence interval is very wide and not much can be reliably stated. The effect of income decreases to practically zero for men, with quite a tight confidence interval around this value. The income elasticity is then greater for women, with a point estimate of 0.18. The difference between the genders is however not

statistically significant. On the other hand, the point estimates of education attainment now approach zero for women, although the confidence intervals are relatively wide and a large array of non-zero values are compatible with the model. The protective effect is now much stronger for men. Men with tertiary education should on average drink only 73% of the amount men with only primary education drink, with the upper endpoint of the 95% confidence interval equaling to 84%, which would still be a large decrease. Religiosity is no longer statistically significant for either of the genders, most likely because of the loss of statistical power due to fewer observations in each subsample.

The higher sensitivity of the alcohol consumption of women to income and the higher effect of education on the alcohol consumption of men is in line with the hypothesis of Decker & Schwartz (2000) that men tend to be more committed drinkers than women. Stable characteristics of individuals (education) have a stronger influence in the case of men, while potentially dynamically changing characteristics (income) have a stronger influence in the case of women. Note that this conclusion has to be made with caution given that the differences between the genders are mostly not statistically significant.

The results of the participation model $P(y_i > 0|\mathbf{x_i})$ can be observed in table 3.5. Notably, men are 1.43 times as likely to drink as women, holding other variables equal. Moreover, a one percent increase in income has an approximately 0.25 percent increase in the odds of the individual drinking non-zero amount of alcohol. (Note that the coefficient does not have to be exponentiated in this case to get the odds ratio, since we receive the model $log(odds) = log(income) + \ldots$ which can be interpreted similarly to a log-log regression model, except that here the dependent variable is the odds.) This is not an economically significant effect, however it points to the possibility that people might not drink because they actually cannot afford it. Furthermore, unemployment has a negative point estimate. An unemployed person should be on average only 0.67 times as likely to drink as an employed person. This effect is not statistically significant, but the confidence interval is mostly negative. It is therefore conceivable that unemployment might possibly actually decrease the odds of drinking alcohol, even after controlling for a decrease in income. This might be because of a higher rate of lifelong abstinence, or also because of a possible higher prevalence of former heavy drinkers amongst unemployed people. For other variables, the confidence intervals are too wide for us to be able to state any conclusions with any sort of confidence. Zero effects are well possible, as are positive and negative ones. This shows the difficulty of

modelling alcohol abstinence. As it has been hypothesized earlier, it is indeed possible that the group of abstainers is so heterogenous that a wide variety of effects of socioeconomic variables might be conceivable. It seems therefore that it is indeed better to model participation in alcohol consumption separately from the amount of alcohol consumed. Separate results for men and women are presented in the Appendix in table 1. Again, only income is significant, with a higher effect for men than for women (difference not statistically significant).

Table 3.5: Participation model results

	Estimate	SE	Odds ratio	95% CI	p value
male	0.36	0.07	1.43	(1.24, 1.65)	0.00
age	-0.08	0.14	0.93	(0.7, 1.22)	0.59
age^2	-0.15	0.10	0.86	(0.71, 1.05)	0.15
age^3	-0.24	0.15	0.78	(0.59, 1.05)	0.10
unemployed	-0.40	0.22	0.67	(0.43, 1.04)	0.07
log(income)	0.25	0.08	-	(0.09,0.41)*	0.00
single	0.10	0.13	1.10	(0.86, 1.41)	0.43
in a relationship	-0.02	0.19	0.98	(0.67, 1.44)	0.93
divorced	-0.05	0.11	0.95	(0.77, 1.17)	0.65
widowed	0.01	0.16	1.01	(0.74, 1.38)	0.96
secondary education	0.10	0.09	1.10	(0.93, 1.3)	0.27
tertiary education	-0.02	0.11	0.98	(0.79, 1.21)	0.82
religious	0.03	0.09	1.03	(0.87, 1.24)	0.71
member of a church	-0.16	0.13	0.85	(0.66, 1.1)	0.22
N = 7423					

The table shows the point estimate, standard error, the odds ratio ($e^{estimate}$), its 95 % confidence interval (*for the log-transformed income the CI is untransformed) and the p-value for the participation model. The dependent variable indicates whether the individual has non-zero alcohol consumption. For factor variables indicating marital status, education and religiosity, the base levels are married, primary education and non-religious, respectively. Household size, number of children, size of municipality and regional and time fixed effects and their interactions are not shown for conciseness.

Let us now focus on the binge drinking model. The results are displayed in table 3.6. Similarly to the previous alcohol consumption measure, men have higher odds of stronger binge drinking across all of the three comparisons. Unemployment still has wide confidence intervals and it is hard to comment on its effect, however, surprisingly, the sign of the point estimate is reversed to a plus for the most extreme comparison between monthly and weekly binge drinkers. In the first and third comparisons, higher levels of education play a protective role, similarly to previous results. Note especially the large protective effect of university education when comparing yearly binge drinkers to people who

never binge drink and when comparing the monthly and weekly frequency. In both cases, even the conservative right part of the confidence interval points to about 20% lower odds of stronger binge drinking habits, with a point estimate of around 37%. This is truly an immense reduction in odds of more frequent binge drinking. On the other hand, in the second comparison, no conclusion can be reliably reached. This goes also for most other variables, perhaps because yearly and monthly binge drinkers are still casual binge drinkers, possibly sharing many similarities between them. In terms of marital status, the only consistent pattern can be detected for single people, who have higher odds of being more frequent binge drinkers across all three comparisons, although the first one is not statistically significant. The effect of religion is not statistically significant.

Let us now inspect the results for the last model - the problem drinking model - in table 3.7. First, gender: holding other variables equal, while men have clearly higher odds of being medium risk drinkers than moderate drinkers and of being high risk drinkers rather than medium risk drinkers, the distinction disappears for the last comparison. The point estimate is now practically zero, with the confidence interval spanning from 30% lower to 30% higher odds than women. Therefore, although it cannot be stated that the effect for men is zero, it is interesting to see the consistent pattern of men being stronger drinkers disappear.

A crucial result is that while income significantly increases the odds of medium and high risk drinking, this effect is reversed for problem drinking, when compared to high risk drinking. A 1% increase in income equates on average to a 0.33% decrease in the odds of problem drinking. On the other hand, the otherwise always insignificant and mostly negative (i.e. protective) coefficient related to unemployment is now significant and positive. In fact, unemployment should be related to 2.5 times higher odds of problem drinking, holding other variables constant. Therefore, for truly pathological and extreme drinking behaviour (recall that this means consuming on average more than 40g or 60g of ethanol per day for women and men respectively, and monthly or more frequent binge drinking), unemployment can be observed significantly more. This would be in line with the hypothesis that unemployment is related to increased alcohol use through psychosocial stress (e.g. Henkel 2011), since coping with stress through alcohol might result in flat-out pathological alcohol dependence, rather than a normal increase in alcohol consumption. This might also be why a positive coefficient related to unemployment could not be

observed for example in the consumption model above. Another explanation is that of reverse causality - pathological drinkers might be at an especially elevated risk of losing their employment or earnings. This creates an issue in the model formulation which will be addressed further on, even though only for the continuous consumption model.

Education on the whole has mostly a protective effect, although the point estimate is somewhat smaller (and not statistically significant) for secondary education in the first comparison and for tertiary education in the last comparison. Marital statuses, as compared to marriage, seem to be generally risk factors, although the statistical significance is not always present. Widowhood even has a negative point estimate for the problem vs. high risk drinking comparison. The effect of religion is rather ambiguous.

Overall, it can be concluded that income is mostly related to higher alcohol consumption, except for problem drinking, where it shows the opposite effect. Unemployment is for the most part not statistically significant, with a negative point estimate. This is again reversed for problem drinking. Here unemployment seems to be an economically and statistically significant risk factor. Therefore, the results point to a stylised conclusion that while worse economic conditions in general decrease the amount of alcohol consumed, they increase the risk of pathological drinking and alcohol dependence. Note that in case of both variables, there might exist the problem of reverse causality, i.e. changes in alcohol consumption might result in changes of income and employment status. Studies as far as Mullahy & Sindelar (1996) have been in fact interested in this opposite effect. This part of the thesis therefore has to be viewed as an exploratory analysis showing interesting patterns that have to be confirmed by further analyses. This is what will be the aim of the second part of this thesis.

In relation to education and marital status other than marriage, it has been shown that these variables are protective and risk factors, respectively. While unobserved heterogeneity has been controlled for, in the extent allowed by the limitations of the cross-sectional dataset, it is still possible that the effects are related to hidden characteristics of the individuals that are correlated with these variables. Nevertheless, the analysis succeeds in finding groups of individuals that might be at a higher risk of higher or problematic alcohol consumption than others - again showing its exploratory nature. In the second part of this thesis, the effect of hidden static variables will be controlled for, but only in relation to dynamically changing variables of interest.

Table 3.6: Binge drinking model results

	Estimate	Std. Error	Odds ratio	95% CI	p value
Never vs. Yearly					
male	0.61	0.08	1.85	(1.57, 2.17)	0.00
unemployed	-0.45	0.27	0.64	(0.37, 1.09)	0.10
log(income)	0.13	0.08	-	(-0.03, 0.29)*	0.12
single	0.19	0.13	1.21	(0.93, 1.57)	0.15
in a relationship	0.03	0.20	1.03	(0.7, 1.52)	0.88
divorced	0.29	0.12	1.34	(1.06, 1.69)	0.01
widowed	0.00	0.19	1.00	(0.69, 1.44)	1.00
secondary education	-0.06	0.10	0.94	(0.78, 1.13)	0.52
tertiary education	-0.46	0.11	0.63	(0.51, 0.79)	0.00
religious	-0.03	0.10	0.97	(0.81, 1.17)	0.77
member of a church	-0.03	0.14	0.97	(0.74, 1.27)	0.81
Yearly vs. Monthly					
male	0.54	0.06	1.72	(1.52, 1.95)	0.00
unemployed	-0.02	0.26	0.98	(0.59, 1.62)	0.92
log(income)	0.11	0.07	-	(0.03, 0.25)*	0.11
single	0.23	0.11	1.25	(1.02, 1.54)	0.03
in a relationship	-0.00	0.16	1.00	(0.72, 1.37)	0.98
divorced	-0.01	0.10	0.99	(0.82, 1.2)	0.93
widowed	0.20	0.17	1.22	(0.87, 1.72)	0.25
secondary education	-0.09	0.08	0.92	(0.79, 1.07)	0.27
tertiary education	-0.09	0.10	0.91	(0.76, 1.1)	0.35
religious	0.06	0.08	1.06	(0.91, 1.24)	0.45
member of a church	-0.17	0.12	0.84	(0.66, 1.07)	0.16
Monthly vs. Weekly					
male	0.68	0.08	1.97	(1.67, 2.32)	0.00
unemployed	0.26	0.29	1.29	(0.73, 2.28)	0.37
log(income)	-0.02	0.09	-	(-0.2, 0.16)*	0.82
single	0.32	0.13	1.38	(1.08, 1.77)	0.01
in a relationship	0.33	0.19	1.40	(0.96, 2.02)	0.08
divorced	0.29	0.12	1.34	(1.06, 1.68)	0.01
widowed	-0.07	0.22	0.93	(0.6, 1.43)	0.74
secondary education	-0.28	0.09	0.76	(0.63, 0.9)	0.00
tertiary education	-0.45	0.12	0.64	(0.51, 0.8)	0.00
religious	-0.15	0.10	0.86	(0.71, 1.05)	0.13
member of a church	0.04	0.16	1.04	(0.76, 1.43)	0.80
N = 6476					

The table shows the point estimate, standard error, the odds ratio ($e^{estimate}$), its 95 % confidence interval (*for the log-transformed income the CI is untransformed) and the p-value for the binge drinking model model. The dependent variable is an ordinal variable denoting intensity of binge drinking. For factor variables indicating marital status, education and religiosity, the base levels are married, primary education and non-religious, respectively. Age (up to the cube power), household size, number of children, size of municipality and regional and time fixed effects and their interactions are not shown for conciseness.

Table 3.7: Problem drinking model results

	Estimate	Std. Error	Odds ratio	95% CI	p value
Moderate vs. Medium risk drinkers					
male	0.89	0.08	2.43	(2.08, 2.83)	0.00
unemployed	-0.46	0.26	0.63	(0.38, 1.05)	0.08
log(income)	0.16	0.08	-	(0.003, 0.32)*	0.04
single	0.29	0.13	1.34	(1.04, 1.72)	0.02
in a relationship	0.01	0.19	1.01	(0.69, 1.47)	0.96
divorced	0.29	0.11	1.34	(1.07, 1.68)	0.01
widowed	0.03	0.18	1.03	(0.73, 1.45)	0.86
secondary education	-0.11	0.09	0.89	(0.74, 1.07)	0.22
tertiary education	-0.52	0.11	0.60	(0.48, 0.74)	0.00
religious	-0.03	0.09	0.97	(0.81, 1.16)	0.74
member of a church	-0.14	0.13	0.87	(0.67, 1.12)	0.28
Medium risk vs. High risk drinkers					
male	0.58	0.07	1.79	(1.55, 2.06)	0.00
unemployed	-0.18	0.30	0.83	(0.46, 1.5)	0.54
log(income)	0.19	0.08	-	(0.03, 0.35)*	0.01
single	0.35	0.11	1.41	(1.13, 1.77)	0.00
in a relationship	0.45	0.17	1.57	(1.14, 2.17)	0.01
divorced	0.18	0.11	1.20	(0.97, 1.47)	0.09
widowed	0.23	0.19	1.26	(0.87, 1.82)	0.22
secondary education	-0.23	0.08	0.80	(0.68, 0.94)	0.01
tertiary education	-0.40	0.11	0.67	(0.54, 0.82)	0.00
religious	-0.11	0.09	0.90	(0.75, 1.07)	0.24
member of a church	0.08	0.14	1.08	(0.83, 1.42)	0.55
High risk vs. Problem drinkers					
male	0.01	0.13	1.01	(0.78, 1.31)	0.93
unemployed	0.93	0.39	2.53	(1.19, 5.39)	0.02
log(income)	-0.33	0.13	-	(-0.58, -0.08)*	0.01
single	0.38	0.20	1.46	(0.99, 2.16)	0.06
in a relationship	0.17	0.29	1.18	(0.67, 2.09)	0.57
divorced	0.32	0.18	1.37	(0.97, 1.95)	0.07
widowed	-0.52	0.37	0.59	(0.29, 1.24)	0.16
secondary education	-0.35	0.14	0.71	(0.53, 0.94)	0.02
tertiary education	-0.26	0.19	0.77	(0.53, 1.12)	0.17
religious	0.09	0.16	1.10	(0.8, 1.49)	0.57
member of a church	-0.38	0.28	0.68	(0.4, 1.17)	0.17
N = 6476					

The table shows the point estimate, standard error, the odds ratio ($e^{estimate}$), its 95 % confidence interval (*for the log-transformed income the CI is untransformed) and the p-value for the problem drinking model model. The dependent variable is an ordinal variable denoting the level of problem drinking. For factor variables indicating marital status, education and religiosity, the base levels are married, primary education and non-religious, respectively. Age (up to the cube power), household size, number of children, size of municipality and regional and time fixed effects and their interactions are not shown for conciseness.

Chapter 4

Panel analysis

In this chapter, an alternative estimation of several previous results will be presented. The aim is to use a three-wave panel dataset to get potentially more robust estimates using a GMM estimator.

4.1 Empirical methodology

When estimating the consumption equation for alcohol, there is an issue that cannot be effectively solved by observational cross-sectional data. Variables that change in time, such as employment and income, not only potentially influence alcohol consumption, but can themselves be influenced by previous alcohol consumption. Given that alcohol is an addictive good, past consumption should logically have a positive effect on present consumption (this follows from the theoretical models presented in the first chapter). Therefore, by omitting past consumption in the estimated equation, omitted variable bias might be introduced. Panel data allow to control for this bias.

Furthermore, it would be ideal to control for as many possible hidden factors as is feasible. Since experimental data or data where a natural experiment could be found are not available, it is not possible to ensure exogeneity. At the same time, introducing fixed effects allows us to at least control for hidden factors which do not vary over time. This fortunately entails many of the most problematic variables that could bias our estimation, such as intelligence or social background. Panel data allow us to control for these fixed effects.

Therefore we would like to estimate the following equation,

$$y_{it} = \gamma y_{it-1} + \boldsymbol{x}_{\boldsymbol{it}}^{\intercal}\boldsymbol{\beta} + FE_i + \varepsilon_{it}, \tag{4.1}$$

where y_{it} represents alcohol consumption of individual i at time t, the vector $\mathbf{x_{it}}$ contains the variables of interest and time-varying control variables, FE_i denotes individual fixed-effects, and ε_{it} is the idiosyncratic error.

A standard fixed-effects estimation of the equation would however be biased and inconsistent for $N \longrightarrow +\infty$. Denote by $\mathbf{v_{it}} = (y_{it-1}, \mathbf{x_{it}})^\intercal$ (i.e. all of the independent variables) and recall the strict exogeneity assumption (Wooldridge 2010):

$$\mathbb{E}(\varepsilon_{it}|\boldsymbol{v_{i1}}, \ldots, \boldsymbol{v_{iT}}, FE_i) = 0, t = 1, \ldots, T \tag{4.2}$$

which implies

$$\mathbb{E}(\boldsymbol{v_{is}}\varepsilon_{it}) = \mathbf{0}\ \forall s, t. \tag{4.3}$$

But even if $\mathbb{E}(y_{it-1}\varepsilon_{it}) = \mathbb{E}(FE_i\varepsilon_{it}) = 0$ and $\mathbb{E}(\mathbf{x_{it}}\varepsilon_{it}) = \mathbf{0}$, the strict exogeneity assumption is neccesarily broken, because

$$\mathbb{E}(y_{it}\varepsilon_{it}) = \gamma\mathbb{E}(y_{it-1}\varepsilon_{it}) + \boldsymbol{\beta}^\intercal\mathbb{E}(\mathbf{x_{it}}\varepsilon_{it}) + \mathbb{E}(FE_i\varepsilon_{it}) + \mathbb{E}(\varepsilon_{it}^2) = \mathbb{E}(\varepsilon_{it}^2) > 0. \tag{4.4}$$

Instead, let us relax the assumption of strict exogeneity and apply the so-called Arellano-Bond estimator (Arellano & Bond 1991). This estimator employs the Generalized Method of Moments to estimate the regression coefficients using a particular set of instruments. Its use in modelling alcohol consumption is nothing new, to mention one of the papers listed in the literature review, it has been used for example in Cotti *et al.* (2015). To employ this model, first assume sequential exogeneity instead of strict exogeneity, that is, assume the idiosyncratic error's conditional expected value is zero given past and present values of regressors, and not necessarily given future values of regressors (Wooldridge 2010):

$$\mathbb{E}(\varepsilon_{it}|\boldsymbol{v_{i1}}, \ldots, \boldsymbol{v_{it}}, FE_i) = 0\ \forall t. \tag{4.5}$$

This relaxation is useful not only because of the inclusion of the lag of the dependent variable as a regressor (now $\mathbb{E}(y_{it}\varepsilon_{it}) > 0$ does not break any fundamental assumptions), but also in relation to the other regressors. If for example a shock to the idiosyncratic error of alcohol consumption influences future income, which is easily conceivable, this no longer invalidates the model.

An additional assumption that we need is the absence of serial correlation between the idiosyncratic errors, i.e. $\mathbb{E}(\varepsilon_{it}\varepsilon_{is}) = 0$ for $s \neq t$. Plainly speaking, it is assumed that the lag of the dependent variable contains all of its dynamics.

Testing of the validity of the model and thus these assumptions is further described below.

Following Harris *et al.* (2008), to build the model, start by taking the first difference of (4.1) to eliminate the fixed effects.

$$y_{it} - y_{it-1} = (\mathbf{x_{it}} - \mathbf{x_{it-1}})^\intercal \boldsymbol{\beta} + \gamma(y_{it-1} - y_{it-2}) + \varepsilon_{it} - \varepsilon_{it-1} \tag{4.6}$$

From the model assumptions, a valid instrument for $\Delta y_{it-1} = y_{it-1} - y_{it-2}$ is simply y_{it-2}, since it is correlated with $y_{it-1} - y_{it-2}$ and uncorrelated with $\varepsilon_{it} - \varepsilon_{it-1}$. Similarly for $s = 1, \ldots, t-3$, since y_{it} follows an $AR(1)$ model and therefore the terms are correlated. For $\Delta\mathbf{x_{it}}$, a similar logic applies (the instruments are $\mathbf{x_{i1}}, \ldots, \mathbf{x_{it-1}}$). The instruments can then be stacked in the following matrix for each individual i:

$$\mathbf{Z_i} = \begin{pmatrix} y_{i1} & 0 & 0 & \ldots & 0 & \mathbf{x_{i1}^\intercal} & \mathbf{x_{i2}^\intercal} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ y_{i1} & y_{i2} & 0 & \ldots & 0 & \mathbf{x_{i1}^\intercal} & \mathbf{x_{i2}^\intercal} & \mathbf{x_{i3}^\intercal} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & y_{i3} & \ldots & y_{iT-2} & \mathbf{x_{i1}^\intercal} & \mathbf{x_{i2}^\intercal} & \mathbf{x_{i3}^\intercal} & \mathbf{x_{i4}^\intercal} & \ldots & \mathbf{x_{iT-1}^\intercal} \end{pmatrix} \tag{4.7}$$

For our three-period panel, this means that the equation can be estimated only for $t = 3$, because for $t = 2$ and $t = 1$, there would be no instruments. Therefore only the following equation is estimated.

$$y_{i3} - y_{i2} = (\mathbf{x_{i3}} - \mathbf{x_{i2}})^\intercal \boldsymbol{\beta} + \gamma(y_{i2} - y_{i1}) + \varepsilon_{i3} - \varepsilon_{i2} \tag{4.8}$$

With following matrix (row vector) of instruments for individual i.

$$\mathbf{Z_i} = \begin{pmatrix} y_{i1} & \mathbf{x_{i1}^\intercal} & \mathbf{x_{i2}^\intercal} \end{pmatrix} \tag{4.9}$$

The moment conditions of the GMM estimation are:

$$\mathbb{E}(\mathbf{x_{it-s}} \Delta\varepsilon_{it}) = \mathbf{0}, \; s = 1, \ldots, t-1 \tag{4.10}$$

$$\mathbb{E}(y_{it-l} \Delta\varepsilon_{it}) = 0, \; l = 2, \ldots, t-1 \tag{4.11}$$

If the values of the dependent variable Δy_{it} are vertically stacked into a vector $\mathbf{y}$, the observations of the independent variables (that is the first differenced ones in 4.6, including lags of the dependent variable) into a standard regression matrix $\mathbf{X}$, the parameters $\boldsymbol{\beta}, \gamma$ into a vector $\boldsymbol{\theta}$ and the matrices $\mathbf{Z_i}$ into a

larger matrix $\mathbf{Z}$, the following empirical equivalent of (4.11) can be obtained (Roodman 2009).

$$M(\boldsymbol{\theta}) = \mathbf{Z}^\intercal(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}) \tag{4.12}$$

Then the GMM estimator minimizes the following norm.

$$\min_{\boldsymbol{\theta}} \; M(\boldsymbol{\theta})^\intercal \mathbf{W} M(\boldsymbol{\theta}) \tag{4.13}$$

where $\mathbf{W}$ is some positive definite weighting matrix. The solution to this minimization problem is then the following.

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\intercal \boldsymbol{Z} \boldsymbol{W} \boldsymbol{Z}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{Z} \boldsymbol{W} \boldsymbol{Z}^\intercal \boldsymbol{y} \tag{4.14}$$

This is almost the answer to how to estimate equation 4.1. The question remains how to choose the matrix $\mathbf{W}$ and how to estimate the standard errors of the coefficients $\hat{\boldsymbol{\theta}}$. To answer it, we follow Roodman (2009) and Cameron & Trivedi (2005). It turns out that to obtain an efficient estimator, it is optimal to set

$$\mathbf{W} = Var(\mathbf{Z_i}^\intercal(\mathbf{Y_i} - \mathbf{X_i}\boldsymbol{\theta}))^{-1}. \tag{4.15}$$

Setting $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}} = (\mathbf{Y_i} - \mathbf{X_i}\hat{\boldsymbol{\theta}})$, the variance in (4.15) can be consistently estimated by $\hat{\mathbf{S}} = \sum_i^N \mathbf{Z_i}^\intercal \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}} \hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}}{}^\intercal \mathbf{Z_i}$. Note however that to receive this estimate, an estimate $\hat{\boldsymbol{\theta}}$ is already needed, while to receive an efficient estimate $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{S}}$ is needed in turn. This can be solved by first getting any consistent estimate for the inverse of the weighting matrix (4.15), for example $\hat{\mathbf{S}}_{\mathbf{naive}} = \sum_i^N \mathbf{Z_i}^\intercal \mathbf{Z_i}$, calculating $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}}$, and then finally obtaining $\hat{\mathbf{S}}$ and the final efficient estimate $\hat{\boldsymbol{\theta}}$. This then allows us to compute the asymptotic variance of the estimates, and thus the standard errors:

$$AVar(\hat{\boldsymbol{\theta}}) = (\mathbf{X}^\intercal\mathbf{Z}\mathbf{W}\mathbf{Z}^\intercal\mathbf{X})^{-1}\,\mathbf{X}^\intercal\mathbf{Z}\mathbf{W}\hat{\mathbf{S}}\mathbf{W}^\intercal\mathbf{Z}^\intercal\mathbf{X}\,(\mathbf{X}^\intercal\mathbf{Z}\mathbf{W}\mathbf{Z}^\intercal\mathbf{X})^{-1}\,. \tag{4.16}$$

This estimate might be downward biased in small samples and thus the Windmeijer (2005) correction which addresses this problem is applied.

Lastly, let us look at how to test the model assumptions. If there is more instruments than sequentially exogenous variables, it is possible to apply a test of overidentifying restrictions. The test statistic is the following (Cameron & Trivedi 2005),

$$1/N \left(\sum_i^N \hat{\boldsymbol{\varepsilon}}_i^\intercal \mathbf{Z}_i \right) \hat{\mathbf{S}}^{-1} \left(\sum_i^N \mathbf{Z}_i^\intercal \hat{\boldsymbol{\varepsilon}}_i \right) \tag{4.17}$$

which is asymptotically distributed as χ^2_{k-l} under the null hypothesis where k is the number of instruments and l is the number of sequentially exogenous variables. The null hypothesis is the validity of the instruments. Simply put, the test checks for whether it has been possible to minimize the objective function sufficiently. This means that the test indirectly checks for the assumption of the lack of autocorrelation of the idiosyncratic errors mentioned above. There also exist direct tests for this assumption, but they necessitate $T > 4$.

4.2 Data

The dataset used in this part of the thesis is the Czech Household Panel Survey (Sociologický ústav AV ČR 2018). This is a four-wave longitudinal survey performed each year from 2015 to 2018 on a constant random sample of Czech households and their members. In this study, only the last three waves can be used, since the first wave does not contain data on household income. The survey contains a multitude of topics and it is not concentrated on a single issue (such as alcohol). It contains data on demographics, family life, health (including alcohol consumption), education, employment, social stratification, housing, political opinions and many other topics. This generality implies both benefits and disadvantages for the purposes of this study. On the one hand, it allows us to include a variable controlling for the health of the individual. As noted below, this is a potentially important control variable which is missing from the cross-sectional dataset.

On the other hand, since alcohol consumption is only one of the many variables in the survey, its measurement is not very precise. Notably, the beverage-specific approach of the previous chapter is not viable, given that the data measures only the intake of alcoholic beverages in general, and not specific types. Furthermore, the frequency of consuming alcoholic beverages is given in quite vague intervals. To be able to calculate approximate average yearly alcohol consumption, crude numeric approximations have to be assigned to these frequency descriptions according to table 4.1. This is then multiplied by the average number of alcoholic beverages consumed on one occasion to obtain the average number of drinks consumed.[1] It is clear why this dataset is not used for the main analysis, but rather only for a confirmatory analysis.

The aforementioned health control variable is represented by a subjective

[1]The amount is not transformed into grams, since the kind of the drinks is not known. The results of the modelling would not change anyway, except for their scale.

Table 4.1: Alcohol frequency in the panel dataset

Survey frequency	Assigned frequency
several times a day	365
daily	365
several times a week	182
once a week	52
several times a month	36
less frequently	12
never	0

This table shows the assignment of a numeric frequency (in times per year) to the panel survey answers on frequency of alcohol consumption.

measure of personal health expressed on the Likert scale. It is important to include this variable, given that it might easily be correlated with the alcohol consumption of the individual - health issues might be a cause as well as an effect of heightened alcohol consumption - as well as the variables of interest - income and unemployment could also be conceivably influenced by health. Links between subjective health and alcohol consumption were found for example in Poikotainen *et al.* (1996). Importantly, health changes in time and therefore it cannot be captured by fixed effects. The subjective measure concisely summarises the health of the individual and gives us possibly more usable information than detailed data on every health issue an individual has. Moreover, the variable incorporates the individual's perception of their health status and thus reflects the individual's health related quality of life better than objective measures (Albrecht 1994).

Compared to the dependent variable, household income in the panel dataset has more precise values than in the cross-sectional dataset. There are 18 tight income intervals as opposed to 8. This might reduce attenuation bias normally arising from measurement errors in the independent variables. As in the cross-sectional dataset, the logarithm of the middle value of the interval is used in the final model. Again, this reduces the skew naturally present in income, improves model fit, and allows for an elasticity interpretation of the related coefficient. Also similarly to the previous analysis, the number of adults and the number of children in the household are added as further control variables to account for how many people live off the household income. Age has been included along with its squared value. The results are robust to including the cubed value of age, and thus it is not included in the estimations. Lastly, given that the panel

dataset is a household survey, the clustering of standard errors is performed on the level of households. Even if there is no cross-sectional dependence (we will see further on why this might be an even more severe issue) Abadie *et al.* (2017) recommend to cluster on the sampled units - which are the households in this case, rather than the individuals.

4.3 Results

The main results of this chapter are shown in table 4.2. The table displays the point estimate, standard error, exponentiated point estimate, 95% confidence interval and exponentiated 95% confidence interval, and p-value for the model outlined above. Firstly, let us concentrate on the result of the overidentifying restrictions test. The p-value of 0.8 is in line with the instruments being valid. At the same time, remember that the power of the test is not available and thus we cannot actually control for Type II error. Still, it is reassuring to see that the test does not indicate that the assumptions are violated.

The lag of the dependent variable is not statistically significant. At the same time, it is not our primary aim to interpret its coefficient, but rather to use it basically as a control variable. Nevertheless, it is reassuring that the point estimate is positive and that most of the 95 % confidence interval is positive as well. This is in line with alcohol being an addictive good. The more a person consumes one year, the more they consume the next year, on average. Specifically, a one percent change in last year's consumption should correspond to about -0.01 % to 0.24 % change in this year's consumption.

Income has a significant and positive coefficient of 0.6. As compared to the cross-sectional results, where the effect has been estimated to be 0.1, the effect is much more economically significant. A 1% increase in income should lead to an approximate average increase of 0.6% in alcohol consumption. Alcohol therefore displays a much higher household income elasticity, even though the confidence interval is fairly wide and would still contain even the original estimate. Going back to the literature, the estimate exactly corresponds to the meta-analysis estimate of Nelson (2013), somewhat extending the validity of the findings of this study to Czech data. It is also similar to the estimates of income elasticity of beer and wine of Rousselière *et al.* (2021) which are 0.58 and 0.64, respectively. In the Czech context, the income elasticity of wine of 0.56, estimated by Janda *et al.* (2010), is also very close, while the income elasticity of beer is higher, at 0.98. Beer is by far the most consumed beverage in Czechia

Table 4.2: Consumption model: GMM estimation

	Estimate	SE	$e^{estimate}$	95% CI	95% CI exp.	p value
$\log(y_{it-1})$	0.11	0.07	-	(-0.01, 0.24)	-	0.08
$\log(income)$	0.60	0.29	-	(0.03, 1.17)	-	0.04
unemployed	0.30	0.28	1.35	(-0.26, 0.85)	(0.77, 2.34)	0.30
health	0.14	0.09	1.15	(-0.03, 0.31)	(0.97, 1.36)	0.10
age	-0.27	0.65	0.76	(-1.54, 1.01)	(0.21, 2.75)	0.68
age^2	0.16	0.35	1.17	(-0.53,0 .86)	(0.59, 2.36)	0.64
Test of overidentifying restrictions: $\chi^2_3 = 1.02,\ p = 0.8$						
N = 2050						

The table shows the point estimate, standard error, exponentiated point estimate, 95% confidence interval and exponentiated 95% confidence interval, and p-value for the GMM model. The dependent variable is the logarithm of the amount of consumed drinks per year. The health variable is on a Likert scale, with larger values indicating worse health. Coefficients for the number of adults and number of children in the household are not shown for conciseness.

(see for example figure 3.1), therefore a closer estimate of the overall alcohol elasticity to this estimate could be expected. On the other hand, the current estimate does not agree with the very small or negative estimates of Grosová *et al.* (2017). When making these comparisons however, keep in mind that our analysis models $\mathbb{E}(y_i|y_i > 0, \boldsymbol{x_i})$ and not the general $\mathbb{E}(y_i|\boldsymbol{x_i})$ found in some of the presented studies.

There might be several reasons behind the differences in the cross-sectional and panel estimates. Firstly, this might simply be brought about by sampling variability. Secondly, the structural differences of the data (e.g. the different measurement of alcohol consumption or household income) might be at play. Lastly, the difference in the modelling approach might be the ultimate cause, showing bias in one of the methodologies. Table 4.3 shows the results of a simple fixed-effects regression performed on the panel dataset with standard errors clustered at the household level. The estimated coefficient corresponding to income is equal to 0.13 which is certainly closer to the value estimated on the cross-sectional dataset than to the GMM estimate. This possibly means that the difference is caused by the model and it is in line with the hypothesis that the cross-sectional estimates of the effect of income are biased because the equations do not contain the lag of the dependent variable. The sign of the bias is indeed quite logical: it is to be expected that long-term alcohol consumption is negatively correlated with income, while income is positively correlated with current alcohol consumption (income effect), thus resulting in

a negative bias of the coefficient. On the other hand, let us not discard the cross-sectional estimate yet. As it will be discussed in the robustness checks section, the Arellano & Bond (1991) GMM estimator has a high finite-sample variance and a degree of finite-sample bias. This could also be the cause of the disparity between the results.

Unemployment has a positive point estimate, but also a very wide confidence interval, even more so than in the previous chapter. Therefore, no evidence has been found to support the hypothesis that unemployment is related to the amount of alcohol consumed. A more clear effect can be spotted in the case of the effect of subjective health. On average, a one point increase on the Likert scale (with higher values meaning worse subjective health) means the individual consumes 1.15 times more alcohol. This estimate is not statistically significant, but positive values are more compatible with the model and the data than negative ones. No conclusions can be safely drawn from this result, but it shows that including this control variable is probably more than reasonable.

Table 4.3: Consumption model: Fixed effects estimation

	Estimate	SE	$e^{estimate}$	95% CI	95% CI exp.	p value
log(*income*)	0.13	0.07	-	(-0.01, 0.27)	-	0.07
unemployed	0.03	0.12	1.03	(-0.21, 0.27)	(0.81, 1.30)	0.78
health	0.02	0.03	1.02	(-0.04, 0.08)	(0.96, 1.08)	0.57
age	-0.91	1.83	0.4	(-4.50, 2.68)	(0.01, 14.54)	0.62
age^2	-0.14	0.22	0.87	(-0.57, 0.29)	(0.56, 1.34)	0.52
N = 2050						

The table shows the point estimate, standard error, exponentiated point estimate, 95% confidence interval and exponentiated 95% confidence interval, and p-value for the fixed effects model. The dependent variable is the logarithm of the amount of consumed drinks per year. The health variable is on a Likert scale, with larger values indicating worse health. Coefficients for the number of adults and number of children in the household are not shown for conciseness.

Let us also analyse the differences in the results on male and female subsamples of the data, similarly to the previous chapter. The results can be inspected in table 4.4. Interesting differences arise between the genders, although none of them are statistically significant. The test of overidentifying restrictions does not reject the validity of either of the two models. The point estimate of income elasticity is higher for women than for men, as in the previous chapter. Also notice that the lagged consumption elasticity is significant in the case of the female subsample, with an estimate of 0.19. On the other hand, the male

Table 4.4: Consumption model: male and female comparison - GMM estimation

	Estimate	SE	$e^{estimate}$	95% CI	95% CI exp.	p value
Female						
$\log(y_{it-1})$	0.19	0.08	-	(0.03, 0.35)	-	0.02
$\log(income)$	0.69	0.28	-	(0.14, 1.23)	-	0.01
unemployed	0.55	0.37	1.73	(-0.18, 1.27)	(0.84, 3.56)	0.14
health	0.05	0.10	1.05	(-0.15, 0.26)	(0.86, 1.30)	0.66
age	-0.06	0.71	0.94	(-1.46, 1.34)	(0.23, 3.82)	0.93
age^2	-0.16	0.51	0.85	(-1.16, 0.86)	(0.31, 2.36)	0.76
Test of overidentifying restrictions: $\chi^2_3 = 1.91,\ p = 0.59$						
N = 1135						
Male						
$\log(y_{it-1})$	-0.05	0.08	-	(-0.20, 0.10)	-	0.52
$\log(income)$	0.50	0.51	-	(-0.50, 1.51)	-	0.33
unemployed	-0.05	0.39	0.95	(-0.82, 0.72)	(0.44, 2.05)	0.90
health	0.25	0.13	1.28	(-0.003, 0.51)	(0.997, 1.67)	0.05
age	-0.48	1.13	0.62	(-2.68, 1.73)	(0.07, 5.64)	0.67
age^2	0.60	0.45	1.82	(-0.29, 1.48)	(0.75, 4.39)	0.19
Test of overidentifying restrictions: $\chi^2_3 = 2.32,\ p = 0.51$						
N = 915						

The table shows the point estimate, standard error, exponentiated point estimate, 95% confidence interval and exponentiated 95% confidence interval, and p-value for the GMM model with female and male subsamples. The dependent variable is the logarithm of the amount of consumed drinks per year. The health variable is on a Likert scale, with larger values indicating worse health. Coefficients for the number of adults and number of children in the household are not shown for conciseness.

subsample shows a point estimate close to zero and actually slightly negative. Although the estimated income elasticities are in line with the hypothesis of Decker & Schwartz (2000) - that men tend to be more committed drinkers than women (reacting more readily to economic changes than men) - mentioned in the previous chapter, the elasticity of the dependence on the lagged consumption does not actually support this conclusion. In the case of unemployment, the confidence intervals are very wide and not much can be drawn from the results. Subjective health, on the other hand, is on the verge of statistical significance for the male subsample, allowing us to state that worse health is likely related to an increase in alcohol consumption in men, possibly around 1.28 times for each increase on the Likert scale.

4.4 Further analyses and robustness checks

Extended analyses as well as robustness checks of the baseline panel model are performed in the following subsections to paint a fuller picture of the results and their limitations.

4.4.1 Marital status

Another time-varying variable that has been used in the first part of the thesis is marital status. At the same time, this variable might also be potentially affected by long-term changes in alcohol consumption, thus introducing possible reverse causality. The previous results have shown that marriage is protective in comparison to almost all other marital statuses and for this reason, only the indicator of whether an individual is married has been chosen as the variable of interest, instead of all of the dummy variables of which the marital status variable consists.

Given that the AB GMM estimator captures only the within variation of each individual in the sample, it can be hypothesized that the effect of marriage will probably not be significant (either statistically or economically). Rather than the ceremony of marriage itself, it is probably the factors related to it that affect individual behaviour, such as a stable long-term relationship leading to the marriage. For this reason, a regression capturing the variation between individuals, such as the one in the previous chapter, might be better suited to the task. This is why the variable has not been included in the main panel regression. Indeed, looking at table 4.5, the estimate of the effect of marriage is far from being statistically significant at any reasonable level, even though the point estimate indicates a large protective effect. Looking at the other variables, the estimates stay essentially the same, except for income, the elasticity of which decreases by 0.04 and is barely no longer statistically significant, even though the standard error is essentially identical.[2]

4.4.2 Including abstainers in the analysis

The fact that the log function on the dependent variable has been used to transform the dependent variable means that individuals who drank zero drinks

[2]See e.g. Vasishth *et al.* (2018) why relying on whether $p < 0.05$ might lead to overestimation of effect sizes since the smaller effect sizes are not passing though the "significance filter".

Table 4.5: Consumption model: marital status - GMM estimation

	Estimate	SE	$e^{estimate}$	95% CI	95% CI exp.	p value
married	-0.99	1.26	0.37	(-3.46, 1.49)	(0.03, 4.44)	0.43
$\log(y_{it-1})$	0.10	0.07	-	(-0.03, 0.24)	-	0.14
$\log(income)$	0.56	0.29	-	(-0.02, 1.13)	-	0.06
unemployed	0.31	0.28	1.36	(-0.25, 0.86)	(0.78, 2.36)	0.28
health	0.14	0.08	1.15	(-0.03, 0.30)	(0.97, 1.35)	0.11
age	-0.09	0.67	0.91	(-1.40, 1.22)	(0.25, 3.39)	0.89
age^2	0.07	0.39	1.07	(-0.71, 0.84)	(0.49, 2.32)	0.87
Test of overidentifying restrictions: $\chi^2_4 = 2.82,\ p = 0.59$						
N = 2050						

The table shows the point estimate, standard error, exponentiated point estimate, 95% confidence interval and exponentiated 95% confidence interval, and p-value for the GMM model with the addition of the marriage variable. The dependent variable is the logarithm of the amount of consumed drinks per year. The health variable is on a Likert scale, with larger values indicating worse health. Coefficients for the number of adults and number of children in the household are not shown for conciseness.

in any single year are not used for the estimation (remember that at least three time periods are needed for estimation). At the same time, a linear model has proven to be an extremely bad and unstable fit. A participation model alike to the one estimated in the previous chapter could not be estimated, since linear probability estimation of the model through the Arellano-Bond estimator has caused numerical errors. To estimate whether including abstainers (be it one-year abstainers or lifelong ones) influences the regression results, we use the inverse hypebolic sine transformation (Bellemare & Wichman 2020)

$$\sinh^{-1}(y) = \log\left(y + \sqrt{y^2 + 1}\right). \tag{4.18}$$

This is applied to both the dependent variable and the household income variable. At the same time, zero values of the household income variable have been kept out, since these are clearly errors. The household income variable includes any form of income, including gifts and social security benefits, and it is implausible for an individual to live in a household with no income for a whole year. If both the dependent variable and the indepedent variable of interest are $\sinh^{-1}$-transformed, an elasticity interpretation of the coefficient can be approximately used for large values of both variables, since if (Bellemare & Wichman 2020):

$$y = \sinh(\alpha + \beta \sinh^{-1}(x) + \varepsilon) \tag{4.19}$$

then

$$\frac{\partial y}{\partial x}\frac{x}{y} = \beta\frac{\sqrt{y^2+1}}{y}\frac{x}{\sqrt{x^2+1}} \tag{4.20}$$

and

$$\lim_{(x,y)\to(+\infty,+\infty)} \frac{\partial y}{\partial x}\frac{x}{y} = \beta. \tag{4.21}$$

Therefore for large values of both x and y, the elasticity is approximately equal to the coefficient β. For linear dummy variable predictors, which constitute the rest of our explanatory variables of interest, the semi-elasticity expression and the interpretation are more complicated. We will therefore only look at the sign and the statistical uncertainty for these variables. It is also important to note that while the inverse hyperbolic sine allows for the inclusion of zero-valued variables, the expression (4.21) cannot be used for values at or close to zero, for obvious reasons. Therefore the focus will be only on the robustness of the previous results for reasonably large values when abstainers are included in the analysis as well.

The results can be observed in table 4.6. Both the point estimate of the coefficient of the lagged dependent variable and of household income have slightly increased. They have stayed roughly close to the original results and the slight deviation might be caused by either the inclusion of abstainers, or the inverse hyperbolic sine transform itself. Nevertheless, it seems the estimates are fairly robust to the change. The same goes for the statistical significance of the estimates. The unemployment and health estimates have also barely changed and thus seem to be robust, even though their interpretation would now be more difficult.

4.4.3 Alternative estimation through maximum likelihood

While the Arellano-Bond estimator is consistent, its finite sample properties might not be ideal. Allison *et al.* (2017) point to three problems related to the AB GMM estimator. First, it may suffer from important small sample bias. Second, it does not use all of the moment restrictions that follow from the model assumptions. Third, the wrong choice of the amount of instruments might worsen the small sample bias of the estimates. Moral-Benito (2013) offers an estimator to alleviate these issues. Namely, the author proposes to estimate the non-differenced equation (4.1) with the assumption (4.5) through maximum likelihood. This parametric method assumes multivariate normality of the variables, however, as the author points out, it is consistent and asymp-

Table 4.6: Model with abstainers - GMM estimation with $\sinh^{-1}$

	Estimate	SE	95% CI	p value
$\sinh^{-1}(y_{it-1})$	0.16	0.08	(0.004, 0.32)	0.05
$\sinh^{-1}(income)$	0.79	0.34	(0.13, 1.44)	0.02
unemployed	0.32	0.39	(-0.43, 1.09)	0.40
health	0.14	0.10	(-0.05, 0.33)	0.16
age	-0.43	0.78	(-1.97, 1.11)	0.58
age^2	0.51	0.50	(-0.46, 1.48)	0.30
Test of overidentifying restrictions: $\chi^2_3 = 0.63,\ p = 0.89$				
N = 2366				

The table shows the point estimate, standard error, 95% confidence interval and p-value for the GMM model with the dependent variable and the income variable transformed by the inverse hyperbolic sine. The dependent variable is the inverse hyperbolic sine of the amount of consumed drinks per year. The health variable is on a Likert scale, with larger values indicating worse health. Coefficients for the number of adults and number of children in the household are not shown for conciseness.

totically normally distributed as a pseudo maximum likelihood estimator as well. While the multivariate normality assumption is untenable in most empirical applications and thus we cannot rely on the efficiency gained from it, there is another source of efficiency to be exploited. Harris *et al.* (2008) note that the following also holds given the standard AB GMM assumptions

$$\mathbb{E}(\varepsilon_{iT}\Delta\varepsilon_{it}) = 0,\ t = 2, \ldots, T-1. \tag{4.22}$$

This was originally derived by Ahn & Schmidt (1995) to introduce a more efficient GMM estimator than the Arellano-Bond estimator. This non-linear estimator seems to be non-trivial in terms of implementation. However, the aforementioned maximum likelihood estimator is asymptotically equivalent to it, and thus more asymptotically efficient than the AB GMM estimator (Moral-Benito 2013). Note that there exists yet another widely used GMM estimator which is more asymptotically efficient than the AB GMM estimator, the Blundell & Bond (1998) estimator. The assumptions of this model are however slightly more restrictive and the overidentifying restrictions test has shown that they might not hold for our data and model specification.

Moral-Benito (2013), Allison *et al.* (2017) and Moral-Benito *et al.* (2019) apply the maximum likelihood estimator to empirical examples and compare the results to those of the Arellano-Bond GMM estimator. The point estimates are often quite different and even a different sign of the estimate is not impos-

sible. All three papers also perform simulation experiments and show that in most cases, the maximum likelihood estimator indeed shows less bias. The AB GMM bias tends to be greater especially for the lagged dependent variable parameter and in smaller samples.

As has been pointed out in the previous section of this chapter, the differences between the cross-sectional results and the panel results might be due to the imperfections of the AB GMM estimator. For this reason, let us also estimate equation (4.1) with the maximum likelihood estimator and see whether and how the results differ between the two methods. The results are displayed in table 4.7. The standard errors are only robust and not clustered since this option is not yet implemented in standard econometric software.

Table 4.7: Consumption model: dynamic panel ML estimation

	Estimate	SE	$e^{estimate}$	95% CI	95% CI exp.	p value
$\log(y_{it-1})$	0.10	0.06	-	(-0.02, 0.23)	-	0.09
$\log(income)$	0.18	0.19	-	(-0.19, 0.55)	-	0.33
unemployed	0.29	0.28	1.34	(-0.26, 0.84)	(0.77, 2.32)	0.31
health	0.11	0.08	1.12	(-0.05, 0.29)	(0.95, 1.34)	0.17
age	1.49	2.31	4.44	(-3.03, 6.01)	(0.05, 407.48)	0.52
age^2	0.21	0.35	1.23	(-0.46, 0.90)	(0.63, 2.46)	0.55
N = 2050						

The table shows the point estimate, standard error, exponentiated point estimate, 95% confidence interval and exponentiated 95% confidence interval, and p-value for the Moral-Benito (2013) maximum likelihood model. The dependent variable is the logarithm of the amount of consumed drinks per year. The health variable is on a Likert scale, with larger values indicating worse health. Coefficients for the number of adults and number of children in the household are not shown for conciseness.

None of the coefficients are statistically significant. The coefficients belonging to the lagged dependent variable, unemployment and health are only slightly lower than in the AB GMM results. The biggest difference is in the coefficient of household income which has strikingly decreased by 0.42. This result is therefore much closer to the fixed-effects model ($\beta_{income} = 0.13$) and to the cross-sectional model ($\beta_{income} = 0.10$). This points to the aforementioned possibility that the discrepancy between the cross-sectional results and the AB GMM results might not be caused solely by an underlying omitted variable/reverse causality bias in the cross-sectional part, but also by the variance and finite sample bias of the AB GMM method.

4.4.4 Checking for cross-sectional dependence

Even though the Arellano-Bond estimator is often used for matters such as studying income on country samples (see e.g. Acemoglu *et al.* (2008)) where there might be an arguably large cross-sectional dependence, Sarafidis & Robertson (2009) note that cross-sectional dependence in the panel sample introduces asymptotic bias in the estimator. While it is reasonable to assume that the disturbances in equation (4.1) are uncorrelated across individuals, the sample nevertheless contains members of the same households and therefore the potential for some form of correlation of errors across individuals does exist.

To see how results would change if the sample was constituted of only one individual per household, the so-called within-cluster sampling or multiple outputation procedure is employed. This is a resampling method that is applied to an existing estimator (in our case the AB GMM estimator). The method had been proposed by Hoffman *et al.* (2001) for generalized linear models with binary dependent variables. Follmann *et al.* (2003) then showed that the estimator thus received is consistent and asymptotically normal (under mild technical assumptions) for a general underlying estimator which is consistent and approximately normally distributed. This is the case for the AB GMM estimator. The procedure goes as follows.

Repeat the following for $k = 1, \ldots, K$ where K is the desired number of iterations:

1. Randomly sample one individual from each cluster.
2. Compute the estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ and their covariance matrix $\boldsymbol{Cov}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}})$ for the selected individuals and save the results.

Then the final estimate and its covariance matrix are equal to

$$\hat{\boldsymbol{\theta}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\boldsymbol{\theta}}_{\boldsymbol{k}} \tag{4.23}$$

$$\boldsymbol{Cov}(\hat{\boldsymbol{\theta}}) = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{Cov}(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}) - \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{\boldsymbol{k}} - \hat{\boldsymbol{\theta}})^{\intercal}. \tag{4.24}$$

While this algorithm ensures that there is no cross-sectional dependence on the level of households, it introduces new issues of its own. Firstly, even though a potential source of asymptotic bias has been eliminated, more finite-sample bias has been introduced since the samples in each iteration are smaller.

Recall from the previous section that the AB GMM estimator might suffer from severe bias in smaller sample sizes. Secondly, this method oversamples single individuals whose alcohol-related behaviour might be different to those living in larger households. In this case it is therefore best to use this method as a robustness check rather than as the main analysis.

2500 iterations of the algorithm have been chosen to be sufficient in our case, since after adding further iterations, the results have barely changed. The final results are displayed in table 4.8. Again, none of the coefficients of interest are statistically significant, and all are lower than in the main estimation but by a smaller margin than in the ML estimation case. The signs of the coefficients are still consistent with the earlier results. The confidence interval of the first coefficient points to a lower range of values, indicating a strikingly low, even negative, dependence of alcohol consumption on its previous values that is compatible with the model and the data. The income elasticity is lower by 0.16 compared to the main model, with a wide confidence interval which also includes values higher than the point estimate of the original model. Unemployment continues to be hard to interpret given its wide confidence interval, while the health variable confidence interval stays mostly in the positive part of the real axis.

Table 4.8: Consumption model: within-cluster resampling

	Estimate	SE	$e^{estimate}$	95% CI	95% CI exp.	p value
$\log(y_{it-1})$	0.07	-	-	(-0.06, 0.19)	-	0.30
$\log(income)$	0.44	-	-	(-0.07, 0.96)	-	0.09
unemployed	0.20	0.32	1.23	(-0.42, 0.82)	(0.66, 2.28)	0.52
health	0.10	0.08	1.11	(-0.06, 0.27)	(0.94, 1.30)	0.22
age	0.01	0.62	1.01	(-1.20, 1.22)	(0.30, 3.40)	0.98
age^2	0.07	0.36	1.07	(-0.64, 0.78)	(0.53, 2.18)	0.85
N = 2050						

The table shows the point estimate, standard error, exponentiated point estimate, 95% confidence interval and exponentiated 95% confidence interval, and p-value for the within-cluster resampled AB GMM model. The dependent variable is the logarithm of the amount of consumed drinks per year. The health variable is on a Likert scale, with larger values indicating worse health. Coefficients for the number of adults and number of children in the household are not shown for conciseness.

Since the differences between the main model and the within-clustered model are neither statistically nor practically significant, it can be concluded that the bias arising from a potential cross-sectional dependence of the errors

in the baseline model does not seem to influence the analysis in a major way. Nevertheless, the possibility that the bias exists cannot be ruled out - the same as the possibility that the difference between the results is in fact caused by a larger finite-sample bias and variance of the within-clustered results. In fact, it has been indicated by the simulations performed by Moral-Benito *et al.* (2019) and Allison *et al.* (2017) that the finite-sample bias in the AB GMM estimator tends to exhibit a downward direction.

Chapter 5

Conclusion

In this thesis, the influence of socioeconomic predictors on alcohol consumption in the Czech Republic has been investigated. First, the literature review has provided a brief exploration of the theoretical underpinnings and the empirical studies concentrating on this topic. Second, the cross-sectional analysis has investigated the impact of socioeconomic and demographic characteristics on individual drinking behaviour in an exploratory manner. Third, the panel analysis has attempted to recover the causal effects for two of the variables, income and unemployment. This has then been followed by several robustness checks and extending analyses.

In general, the results of the cross-sectional analysis show that individuals with higher household income and individuals who are not married drink more alcoholic beverages, while higher education levels seem to exhibit a protective relationship towards alcohol consumption. Compared to the previous study of Dzúrová *et al.* (2010), no evidence has been found to state that unemployment increases alcohol consumption in the base model. What is confirmed is that men drink significantly more than women. And while income seems to play a larger role for women, education seems to be more important in the case of men. This points to the possibility that stable variables influence men more than women, while relatively variable characteristics impact women more than men. Modelling alcohol abstinence is particularly difficult, likely because of the heterogeneity of the abstainers group. Furthermore, the coefficients are very different for the model focusing on pathological drinking. In this case, there is no evidence that the probability of being a problem drinker is higher for men. Importantly, unemployment is related to higher odds of being a problem drinker, while higher income shows a decrease in these odds. This shows that

standard and extremely pathological alcohol consumption display very different socioeconomic patterns.

While the exploratory analysis does not directly show the causes of higher alcohol consumption, it does provide an overview of which groups are the most likely to drink and which are particularly at risk of problems related to increased ethanol consumption in the Czech republic. This is also relevant for policy makers, for example in targeting the relevant groups with information campaigns or in introducing restrictions on alcohol sales.

The results of the second part seem to confirm the positive relationship between household income and alcohol consumption. What is more, the point estimate is higher than in the cross-sectional part. Nevertheless, the robustness checks have shown that this discrepancy might not be simply due to omitting the lag of the dependent variable in the cross-sectional analysis, but possibly because of the high variance and finite sample bias of the Arellano-Bond GMM estimator.

The main contribution of this thesis is bringing more robust and current results to provide a solid empirical basis for example for the policy decisions mentioned above, as well as creating a further starting point for future research. Further research needs to be performed to answer the questions posed in this thesis with greater certainty. For one, the panel analysis should be replicated if and when more waves of the Czech Household Panel Survey become available. This will allow to directly test the assumption of no autocorrelation of the Arellano-Bond estimator and provide a richer set of instruments. Moreover, other methods should be used to get rid of the inherent endogeneity present in the current problem and to approach it from a different angle to avoid the pitfalls of the Arellano-Bond estimator. Furthermore, it seems equally interesting to study not only the level of alcohol consumption, but also its impact on different socioeconomic groups, similarly to Grittner *et al.* (2012). Undoubtedly, calculating more robust results on new data in the Czech context would be extremely useful in this case as well. In the end, if we only know which socioeconomic groups drink the most, but not *how* they are affected, the picture of the issue is incomplete.

Bibliography

Abadie, A., S. Athey, G. W. Imbens, & J. Wooldridge (2017): "When should you adjust standard errors for clustering?" *Technical report*, National Bureau of Economic Research.

Acemoglu, D., S. Johnson, J. A. Robinson, & P. Yared (2008): "Income and democracy." *American economic review* **98(3)**: pp. 808–42.

Agresti, A. (2010): *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons.

Ahn, S. C. & P. Schmidt (1995): "Efficient estimation of models for dynamic panel data." *Journal of econometrics* **68(1)**: pp. 5–27.

Albrecht, G. L. (1994): "Subjective health assessment." In "Measuring health and medical outcomes," volume 7. UCL Press London.

Allison, P. D., R. Williams, & E. Moral-Benito (2017): "Maximum likelihood for cross-lagged panel models with fixed effects." *Socius* **3**.

Arellano, M. & S. Bond (1991): "Some tests of specification for panel data: Monte carlo evidence and an application to employment equations." *The review of economic studies* **58(2)**: pp. 277–297.

Becker, G. S. & K. M. Murphy (1988): "A theory of rational addiction." *Journal of Political Economy* **96(4)**: pp. 675–700.

Bellemare, M. F. & C. J. Wichman (2020): "Elasticities and the inverse hyperbolic sine transformation." *Oxford Bulletin of Economics and Statistics* **82(1)**: pp. 50–61.

Bloomfield, K., U. Grittner, S. Kramer, & G. Gmel (2006): "Social inequalities in alcohol consumption and alcohol-related problems in the study

countries of the EU concerted action "Gender, culture and alcohol problems: A multi-national study".” *Alcohol & Alcoholism* **41(1)**: pp. i26–i36.

Blundell, R. & S. Bond (1998): “Initial conditions and moment restrictions in dynamic panel data models.” *Journal of econometrics* **87(1)**: pp. 115–143.

Brant, R. (1990): “Assessing proportionality in the proportional odds model for ordinal logistic regression.” *Biometrics* pp. 1171–1178.

Bushman, B. J. (2002): “Effects of alcohol on human aggression.” *Recent developments in alcoholism* pp. 227–243.

Cameron, A. C. & P. K. Trivedi (2005): *Microeconometrics: methods and applications*. Cambridge university press.

Castro, F. G., M. Barrera Jr, L. A. Mena, & K. M. Aguirre (2014): “Culture and alcohol use: Historical and sociocultural themes from 75 years of alcohol research.” *Journal of Studies on Alcohol and Drugs, Supplement* **(s17)**: pp. 36–49.

Cotti, C., R. A. Dunn, & N. Tefft (2015): “The great recession and consumer demand for alcohol: a dynamic panel-data analysis of US households.” *American Journal of Health Economics* **1(3)**: pp. 297–325.

Csémy, L., Z. Dvořáková, A. Fialová, M. Kodl, M. Malý, & M. Skývová (2020): “Užívání tabáku a alkoholu v České republice 2019. In English: Tobacco and alcohol use in the Czech Republic 2019.” *Technical report*.

Csémy, L., Z. Dvořáková, A. Fialová, M. Kodl, M. Malý, & M. Skývová (2021): “Národní výzkum užívání tabáku a alkoholu v České republice 2020. In English: Tobacco and alcohol use in the Czech Republic 2020.” *Technical report*.

Csémy, L., Z. Dvořáková, A. Fialová, M. Kodl, & M. Skývová (2019): “Užívání tabáku a alkoholu v České republice 2018. In English: Tobacco and alcohol use in the Czech Republic 2018.” *Technical report*.

Dávalos, M. E., H. Fang, & M. T. French (2012): “Easing the pain of an economic downturn: macroeconomic conditions and excessive alcohol consumption.” *Health economics* **21(11)**: pp. 1318–1335.

Decker, S. L. & A. E. Schwartz (2000): "Cigarettes and alcohol: Substitutes or complements?" *Working Paper 7535*, National Bureau of Economic Research.

Dzúrová, D., J. Spilková, & H. Pikhart (2010): "Social inequalities in alcohol consumption in the Czech Republic: a multilevel analysis." *Health & place* **16(3)**: pp. 590–597.

Ellison, C. G., J. B. Barrett, & B. E. Moulton (2008): "Gender, marital status, and alcohol behavior: The neglected role of religion." *Journal for the Scientific Study of Religion* **47(4)**: pp. 660–677.

European Statistical Office (2019): "European health interview survey."

Follmann, D., M. Proschan, & E. Leifer (2003): "Multiple outputation: inference for complex clustered data by averaging analyses from independent data." *Biometrics* **59(2)**: pp. 420–429.

Freedman, D. & P. Diaconis (1981): "On the histogram as a density estimator: L2 theory." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57(4)**: pp. 453–476.

Furstova, J., K. Malinakova, D. Sigmundova, & P. Tavel (2021): "Czech out the atheists: A representative study of religiosity in the Czech Republic." *The International Journal for the Psychology of Religion* **31(4)**: pp. 288–306.

Gordon, R., D. Heim, & S. MacAskill (2012): "Rethinking drinking cultures: A review of drinking cultures and a reconstructed dimensional approach." *Public Health* **126(1)**: pp. 3–11.

Grittner, U., S. Kuntsche, K. Graham, & K. Bloomfield (2012): "Social inequalities and gender differences in the experience of alcohol-related problems." *Alcohol and alcoholism* **47(5)**: pp. 597–605.

Grosová, S., M. Masár, O. Kutnohorská, & V. Kubeš (2017): "The demand for beer in Czech Republic: understanding longrun on-and off-trade price elasticities." *Czech Journal of Food Sciences* **35(2)**: pp. 165–170.

Harris, M. N., P. Sevestre *et al.* (2008): "Dynamic models for short panels." In "The econometrics of panel data," pp. 249–278. Springer.

Heinz, A. J., A. Beck, A. Meyer-Lindenberg, P. Sterzer, & A. Heinz (2011): "Cognitive and neurobiological mechanisms of alcohol-related aggression." *Nature Reviews Neuroscience* **12(7)**: pp. 400–413.

Henkel, D. (2011): "Unemployment and substance use: a review of the literature (1990-2010)." *Current drug abuse reviews* **4(1)**: pp. 4–27.

Hoffman, E. B., P. K. Sen, & C. R. Weinberg (2001): "Within-cluster resampling." *Biometrika* **88(4)**: pp. 1121–1134.

Hsu, A.-C. & S.-C. Liu (2008): "The hurdle models choice between truncated normal and lognormal." *Applied Economics* **40(2)**: pp. 201–207.

Janda, K., J. Mikolášek, & M. Netuka (2010): "Complete almost ideal demand system approach to the Czech alcohol demand." *Agricultural Economics* **56(9)**: pp. 421–434.

Kendler, K. S., S. L. Lönn, J. Salvatore, J. Sundquist, & K. Sundquist (2017): "Divorce and the onset of alcohol use disorder: a Swedish population-based longitudinal cohort and co-relative study." *American Journal of Psychiatry* **174(5)**: pp. 451–458.

Madden, D. (2008): "Sample selection versus two-part models revisited: The case of female smoking and drinking." *Journal of health economics* **27(2)**: pp. 300–307.

Martin, J.-L., B. Gadegbeku, D. Wu, V. Viallon, & B. Laumon (2017): "Cannabis, alcohol and fatal road accidents." *PLoS one* **12(11)**: p. e0187320.

Moral-Benito, E. (2013): "Likelihood-based estimation of dynamic panels with predetermined regressors." *Journal of Business & Economic Statistics* **31(4)**: pp. 451–472.

Moral-Benito, E., P. Allison, & R. Williams (2019): "Dynamic panel data modelling using maximum likelihood: an alternative to Arellano-Bond." *Applied Economics* **51(20)**: pp. 2221–2232.

Moskalewicz, J. & J. Sierosławski (2010): "Drinking population surveys: Guidance document for standardised approach: Final report prepared for the project standardizing measurement of alcohol-related troubles - SMART." *Technical report*, Institute of Psychiatry and Neurology.

Mullahy, J. & J. Sindelar (1996): "Employment, unemployment, and problem drinking." *Journal of health economics* **15(4)**: pp. 409–434.

Nelson, J. P. (2013): "Meta-analysis of alcohol price and income elasticities–with corrections for publication bias." *Health economics review* **3(1)**: pp. 1–10.

Pelucchi, C., I. Tramacere, P. Boffetta, E. Negri, & C. L. Vecchia (2011): "Alcohol consumption and cancer risk." *Nutrition and cancer* **63(7)**: pp. 983–990.

Peña, S., P. Mäkelä, G. Valdivia, S. Helakorpi, N. Markkula, P. Margozzini, & S. Koskinen (2017): "Socioeconomic inequalities in alcohol consumption in Chile and Finland." *Drug and alcohol dependence* **173**: pp. 24–30.

Poikotainen, K., E. Vartiainen, & H. J. Korhonen (1996): "Alcohol intake and subjective health." *American Journal of Epidemiology* **144(4)**: pp. 346–350.

Popova, S., J. Rehm, J. Patra, & W. Zatonski (2007): "Comparing alcohol consumption in central and eastern Europe to other European countries." *Alcohol & alcoholism* **42(5)**: pp. 465–473.

Popovici, I. & M. T. French (2013): "Does unemployment lead to greater alcohol consumption?" *Industrial Relations: A Journal of Economy and Society* **52(2)**: pp. 444–466.

Prescott, C. A. & K. S. Kendler (2001): "Associations between marital status and alcohol consumption in a longitudinal study of female twins." *Journal of studies on alcohol* **62(5)**: pp. 589–604.

Redish, A. D. (2004): "Addiction as a computational process gone awry." *Science* **306(5703)**: pp. 1944–1947.

Rogeberg, O. (2004): "Taking absurd theories seriously: economics and the case of rational addiction theories." *Philosophy of science* **71(3)**: pp. 263–285.

Roodman, D. (2009): "How to do xtabond2: An introduction to difference and system GMM in Stata." *The stata journal* **9(1)**: pp. 86–136.

Rosansky, J. A. & H. Rosenberg (2020): "A systematic review of reasons for abstinence from alcohol reported by lifelong abstainers, current abstainers and former problem-drinkers." *Drug and Alcohol Review* **39(7)**: pp. 960–974.

Rousselière, S., G. Petit, T. Coisnon, A. Musson, & D. Rousselière (2021): "A few drinks behind - alcohol price and income elasticities in Europe: A microeconometric note." *Journal of Agricultural Economics* .

Sarafidis, V. & D. Robertson (2009): "On the impact of error cross-sectional dependence in short dynamic panel estimation." *The Econometrics Journal* **12(1)**: pp. 62–81.

Smith, T. G. & A. Tasnádi (2007): "A theory of natural addiction." *Games and Economic Behavior* **59(2)**: pp. 316–344.

Smutna, S. & M. Scasny (2017): "Selectivity problem in demand analysis: Single equation approach." *Technical report*, IES Working Paper.

Sociologický ústav AV ČR (2018): "České panelové šetření domácností. In English: Czech household panel survey." Retrieved from: `http://nesstar.soc.cas.cz/webview/`.

Sovinová, H. & L. Csémy (2013): "Užívání tabáku a alkoholu v České republice 2012. In English: Tobacco and alcohol use in the Czech Republic 2012." *Technical report*.

Sovinová, H. & L. Csémy (2015): "Užívání tabáku a alkoholu v České republice 2014. In English: Tobacco and alcohol use in the Czech Republic 2014." *Technical report*.

Suranovic, S. M., R. S. Goldfarb, & T. C. Leonard (1999): "An economic theory of cigarette addiction." *Journal of health economics* **18(1)**: pp. 1–29.

Tamers, S. L., C. Okechukwu, A. A. Bohl, A. Gueguen, M. Goldberg, & M. Zins (2014): "The impact of stressful life events on excessive alcohol consumption in the French population: Findings from the GAZEL cohort study." *PLOS ONE* **9(1)**: pp. 1–8.

Váňová, A., M. Skývová, & L. Csémy (2017): "Užívání tabáku a alkoholu v České republice 2016. In English: Tobacco and alcohol use in the Czech Republic 2016." *Technical report*.

Vasishth, S., D. Mertzen, L. A. Jäger, & A. Gelman (2018): "The statistical significance filter leads to overoptimistic expectations of replicability." *Journal of Memory and Language* **103**: pp. 151–175.

Volland, B. (2012): "The history of an inferior good: Beer consumption in Germany." *Papers on Economics and Evolution* .

Windmeijer, F. (2005): "A finite sample correction for the variance of linear efficient two-step gmm estimators." *Journal of Econometrics* **126(1)**: pp. 25–51.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*. MIT press.

World Health Organization (2019): *Global status report on alcohol and health 2018*. World Health Organization.

Table 1: Participation model results: male and female comparison

	Estimate	SE	Odds ratio	95% CI	p value
		Female			
age	0.18	0.19	1.20	(0.82, 1.75)	0.34
age^2	-0.31	0.13	0.73	(0.57, 0.95)	0.02
age^3	-0.45	0.20	0.64	(0.43, 0.95)	0.03
unemployed	-0.53	0.32	0.59	(0.31, 1.11)	0.10
log(income)	0.21	0.10	-	(0.01, 0.4)*	0.05
single	0.21	0.17	1.23	(0.88, 1.73)	0.22
in a relationship	0.29	0.29	1.34	(0.75, 2.38)	0.32
divorced	-0.03	0.14	0.97	(0.73, 1.29)	0.84
widowed	0.12	0.21	1.13	(0.75, 1.69)	0.56
secondary education	0.12	0.12	1.13	(0.89, 1.42)	0.32
tertiary education	0.01	0.15	1.01	(0.75, 1.36)	0.95
religious	0.09	0.12	1.09	(0.86, 1.38)	0.46
member of a church	-0.06	0.17	0.95	(0.68, 1.33)	0.75
		Male			
age	-0.43	0.21	0.65	(0.43, 0.99)	0.04
age^2	0.09	0.16	1.10	(0.8, 1.5)	0.56
age^3	0.03	0.23	1.03	(0.66, 1.62)	0.90
unemployed	-0.34	0.32	0.71	(0.38, 1.33)	0.28
log(income)	0.32	0.12	-	(0.08,0.56)*	0.01
single	-0.09	0.19	0.91	(0.63, 1.32)	0.62
in a relationship	-0.39	0.27	0.68	(0.4, 1.15)	0.15
divorced	-0.05	0.16	0.95	(0.69, 1.3)	0.74
widowed	-0.20	0.26	0.82	(0.49, 1.37)	0.44
secondary education	0.02	0.13	1.02	(0.79, 1.31)	0.89
tertiary education	-0.06	0.16	0.94	(0.68, 1.3)	0.71
religious	-0.01	0.14	0.99	(0.75, 1.31)	0.94
member of a church	-0.22	0.21	0.80	(0.53, 1.21)	0.29
	$N_{female} = 3683,\ N_{male} = 3740$				

The table shows the point estimate, standard error, the odds ratio ($e^{estimate}$), its 95 % confidence interval (*for the log-transformed income the CI is untransformed) and the p-value for the participation model with the female and male subsamples. The dependent variable indicates whether the individual has non-zero alcohol consumption. For factor variables indicating marital status, education and religiosity, the base levels are married, primary education and non-religious, respectively. Further control variables are not shown for conciseness.