CHARLES UNIVERSITY IN PRAGUE

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

BACHELOR THESIS

Forecasting Tobacco Consumption in the Czech Republic

Author: **Martin Štrobl**

Supervisor: **prof. Ing. Karel Janda M.A., Dr., Ph.D.**

Year of defence: **2015**

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, May 12, 2015 _____

Signature

Acknowledgment

I would like to express my gratitude to prof. Ing. Karel Janda M.A., Dr., Ph.D. for his guidance and leadership of my thesis. Furthermore, I would like to thank PhDr. Jakub Mikolášek for valuable hints and advice, Mgr. Pavla Chomynová for providing me the access to the data and Bc. Jan Žáček for this neat LaTeX template.

Bibliography Reference

ŠTROBL, Martin. *Forecasting Tobacco Consumption in the Czech Republic.* Prague, 2015. 44 p. Bachelor thesis (Bc.) Charles University, Faculty of Social Sciences, Institute of Economic Studies.
Supervisor: prof. Ing. Karel Janda M.A., Dr., Ph.D.

Extent of the Thesis

61,188 characters (with spaces)

Abstract

This thesis aims to build a theoretical framework to model the future tobacco consumption, size of smoking population and governmental tax revenues in the Czech republic. The main assumption of the model states that smokers determine their future tobacco consumption behavior as adolescents. This strong statement is backed by empirical evidence. Further assumptions are introduced to make the model applicable to the data by the Czech National Monitoring Centre for Drugs and Drug Addiction. The resulting model is simplified, however, is still able to reflect the future trends induced by upcoming demographic changes to the Czech population and provide forecasts. Future teenage smoking rates and average consumption are the inputs to the model; consumption growth coefficients for each age category are estimated using zero-inflated negative binomial regression. Several scenarios are built to model possible developments, including extreme cases. All scenarios showed that all model outcomes are going to grow until 2028 in a very similar pattern. In particular, the projected number of smokers in 2028 is by 4-8% higher than in 2013, the total daily tobacco consumption and tax revenue by 7-26%. This increase is induced by aging of large birth cohorts. Later on, the projected scenarios differ substantially. If the teenagers were to behave as in 2013, the projected number of smokers (after the initial growth) would steadily fluctuate around 2.6 millions, compared to 2.4 in 2013; total daily tobacco consumption around 33-34 millions, compared to 31.8 in 2013; and tax revenues around 50-52 billions of CZK compared to 46.8 in 2013. An appropriate policy reaction to the upcoming growth in the next decade might consist of the anti-tobacco law, currently proposed by the Czech government, and higher taxation on tobacco.

JEL Classificiation	D12
Keywords	smoking, tobacco, cigarettes, consumption, forecasting
Author's email	martin@martinstrobl.cz
Supervisor's email	Karel-Janda@seznam.cz

Abstrakt

Cílem této práce je vytvořit teoretický model schopný odhadnout budoucí vývoj spotřeby tabáku, počtu kuřáků a výnosů ze spotřební daně v České republice. Výsledný model je založen na předpokladu, že kuřáci určují svou budoucí spotřebu už jako mladiství. Tento silný předpoklad je empiricky podložen. Další předpoklady jsou nutné k tomu, aby bylo možné použít data z Národního monitorovacího střediska pro drogy a drogové závislosti. To vede ke zjednodušení modelu, přesto je model schopný zachytit vliv nastávajících demografických změn v české populaci a poskytnout odhady budoucího vývoje. Vstupními parametry modelu jsou podíl kuřáků mezi budoucími teenagery a jejich průměrná denní spotřeba cigaret. K určení koeficientů růstu spotřeby cigaret s věkem je využita negativní binomická regrese (zero-inflated). Pro odhad budoucího vývoje je sestaveno několik scénářů, včetně těch extrémních. Podle všech scénářů budou všechny odhadované veličiny až do roku 2028 růst, a to velmi podobně. Počet kuřáků bude v roce 2028 o 4-8% větší než v roce 2013, celková denní spotřeba a výnosy ze spotřební daně o 7-26%. Tento nárůst je způsoben stárnutím silných ročníků. Následně se ale modelované scénáře velmi rozcházejí. Pokud by budoucí mladiství kouřili ve stejné míře jako v roce 2013, bude se počet kuřáků v ČR v budoucnu (po úvodním nárůstu) pohybovat kolem 2,6 milionu oproti 2,4 v roce 2013. Podobně by se denní spotřeba pohybovala kolem 33-34 milionů cigaret oproti 31,8 v roce 2013 a výnosy z daně kolem 50-52 miliard korun v porovnání s 46,68 miliardy z roku 2013. Odpovídající reakcí na nadcházející růst by mohl být protikuřácký zákon, který právě chystá česká vláda, a případné zvýšení spotřební daně z cigaret.

JEL klasifikace	D12
Klíčová slova	kouření, tabák, cigarety, spotřeba, vývoj
Email autora	martin@martinstrobl.cz
Email vedoucího práce	Karel-Janda@seznam.cz

Bachelor Thesis Proposal

Author	Martin Štrobl
Supervisor	prof. Ing. Karel Janda M.A., Dr., Ph. D.
Proposed topic	Forecasting Tobacco Consumption in the Czech Republic

Předběžná náplň práce Cílem této práce je vytvořit jednoduchý model, který bude simulovat vývoj populace kuřáků v ČR a jejich spotřebu v nadcházejících letech. Model bude využívat demografickou projekci vývoje obyvatelstva EUROPOP 2013 a bude založen na předpokladu, že kuřáci formují své chování během dospívání. Předpovědi modelu tedy budou závislé na budocím vývoji počtu kuřáků mezi náctiletými a průměrné spotřebě cigaret této věkové kategorie. Ke konkrétním předpovědím pro Českou republiku budou využity data z Národního monitorovacího střediska pro drogy a drogové závislosti spolu s negativní binomickou regresí (zero-inflated varianta). Výstupy modelu, včetně simulace příjmů státu ze spotřební daně z cigaret, budou zachycovat nejrůznější scénáře a mohou tak být užiteční pro simulaci reakce na změnu zdanění, omezení kouření či jiné změny a šoky.

Topic characteristics This thesis aims to introduce a simple model which would simulate the changes in the population size of smokers in the Czech republic as well as the tobacco consumption in the forthcoming years. The model will make use of the demographic projection EUROPOP 2013 and will be based on the assumption that smokers form their behavior during their teenage years. The model forecasts will depend on future smoking prevalence and average consumption among adolescents. The data by the Czech National Monitoring Centre for Drugs and Drug Addiction and a zero-inflated negative binomial regression will be used to form forecasts for the Czech republic. Model outcomes including simulations of tax revenue changes will be scenario-based and can be useful for simulations of responses to policy changes, taxation changes, tobacco bans or other shocks and changes.

Outline

1. Introduction
2. Literature
3. Theoretical Model
4. Data
5. Coefficients Estimation
6. Empirical Model
7. Discussion and Conclusion

Core bibliography

[1] Coxe, S., West, S. G., and Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of personality assessment*, 91(2), 121-136.

[2] Hiscock, R., Bauld, L., Amos, A., Fidler, J. A., and Munafo, M. (2012). Socioeconomic status and smoking: a review. *Annals of the New York Academy of Sciences*, 1248(1), 107-123.

[3] Kvaček, J. (2011) Poptávka po cigaretách v České republice a výnos spotřební daně z cigaret. Diplomová práce (Mgr.) Univerzita Karlova, Fakulta sociálních věd, Institut ekonomických studií. Vedoucí diplomové práce PhDr. Martin Gregor, PhD.

[4] Reed, M. B., R. Wang, A. M. Shillington, J. D. Clapp, and J. E. Lange (2007). The relationship between alcohol use and cigarette smoking in a sample of undergraduate college students. *Addictive Behaviors 32* (3), 449-464.

[5] Panday, S., Reddy, S. P., Ruiter, R. A., Bergström, E., and De Vries, H. (2007). Determinants of smoking among adolescents in the Southern Cape-Karoo region, South Africa. *Health Promotion International*, 22(3), 207-217.

Author

Supervisor

Contents

Acronyms

EUROPOP2013	European Population Projections, base year 2013
IRR	incidence rate ratio
MLE	maximum likelihood estimation
NB1	negative binomial distribution where $Var(y_i\|\mathbf{x_i}) = \mu + \alpha\mu$
NB2	negative binomial distribution where $Var(y_i\|\mathbf{x_i}) = \mu + \alpha\mu^2$
NUTS	Nomenclature of Units for Territorial Statistics
OECD	Organisation for Economic Co-operation and Development
OLS	ordinary least squares
PMF	probability mass function
SES	socioeconomic status
UN	United Nations
VOŠ	vyšší odborná škola - follow-up studies
WHO	World Health Organization
WLS	weighted least squares
ZINB	zero-inflated negative binomial regression

List of Figures

List of Tables

1. Introduction

Though the harmful effects of smoking on human's body are nowadays commonly known, smoking is still popular in our society and there has not been a significant decline in smoking rate over the past few years. Currently, more than 21% of Czech population smokes (WHO, 2014). According to the OECD (2014), the Czech Republic is its only member whose percentage change in smoking rate over the period of 2000-2011 was positive (4.7%) while the OECD average was -20.7%. The Czech government proposes a smoking-ban which should regulate the sale of tobacco and ban smoking in restaurants and bars. Such a measure is common in many EU countries. Just like in case of other risky behavior (obesity, alcohol, etc.) this ban intends to discourage desired choices of consumers which may (or may not) be based on imperfect information or myopic behavior and consequently lead to their addiction. Surprisingly, the individual's health may not necessarily be its main reason because tobacco consumption also produces many externalities, e.g., decreased productivity of work (due to more frequent breaks), increased health costs, passive smoking, pregnancy smoking, fires, etc. - and those may lead to economic inefficiency. On the other hand, the tax levied on tobacco is a considerable source of state budget income.

The basic questions regarding smoking that are relevant to every policymaker (and thus to academia as well) are: What is the relationship of price of tobacco to its demand (its price elasticity)? How does the socioeconomic status (SES) influence tobacco consumption? How influential is advertising? Are the anti-tobacco campaigns successful? What are the trends and prospects of tobacco consumption? This thesis deals with the last one, as it aims to explore the future role of tobacco in the Czech Republic. Application of a proposed simple theoretical model on empirical data and various development scenarios provide predictions for smoker population sizes, total daily tobacco consumption and governmental tobacco tax revenues for the next 75 years. The resulting outlook based on demographic projections can be used to

understand the mechanisms of possible transformation in smoking behavior as well as to see the prospects of tobacco industry and governmental tax revenues in the Czech Republic. The main assumption of the model is that smokers determine their future smoking status and consumption levels as adolescents and are unable to quit later on throughout their life. That implies that the model scenarios and outcomes are always based on future smoking rates and average consumption among teenagers.

The model further assumes that tobacco consumption changes with age in a way that is determined as a product of individual's teenage tobacco consumption and some coefficient for his current age category. These parameters are estimated using empirical econometric analysis, namely zero-inflated negative binomial model. The analyzed data set comes from a survey (from Q4/2012) by the Czech National Monitoring Centre for Drugs and Drug Addiction. The data structure allows to focus on age categories of size of 5 years from the age of 15 up to 65. Unfortunately, there were no respondents aged 65 or more and this imposes complications on my latter empirical model, resulting in a necessary simplification and model adjustments. As the coefficients estimation identifies the determinants of consumption and smoking status, its results might be interesting also from a health economist's point of view as well.

This paper is structured in seven parts. In the next section (Literature) I sum up previous research on tobacco consumption and models predicting consumption. The theoretical model, its assumptions, limitations and possible improvements are located in the subsequent section (Theoretical Model). The origin and structure of the population projections as well as of the data set used for coefficients estimation and model adjustments are described in section four (Data). Coefficients estimation, including theoretical background of the estimation methods, follows in section five (Coefficients Estimation). The empirical model is built and analyzed in chapter six (Empirical Model) and its outcomes are summarized and discussed in the final section (Discussion and Conclusion).

2. Literature

2.1 Tobacco Consumption in the Czech Republic

Tobacco is a common research topic in the Czech Republic, however, large portion of these publications consists of medical papers. Thus, there are only a few relevant studies which actually use models to predict causalities, future development, or the structure of the smoking population.

Sovinová et al. (2008) used the method of smoking-attributable fractions to estimate (ex-post) that six years earlier, in 2002, smoking could induce 19% (20,550 deaths) of overall Czech mortality. Interestingly, women represented only around 30% out of these deaths. Levy et al. (1997) also studied the Czech mortality, but looked even farther - into the future. Their study models future mortality rates for several tobacco policy scenarios using SimSmoke simulations.

Spilková et al. (2011) analyzed smoking data from 2003 using multilevel modeling which showed that education is negatively related to smoking, unemployment and divorced status positively. Moreover, men proved to be more probable to become smokers.

Because the majority of tobacco consumers initiates smoking as teenagers, some research papers focus particularly on this age group. A good example is the work of Pertold (2009) who compared the past (primary school) and current smoking status of secondary school freshmen in order to estimate the peer effects of their current classmates. The results show that boys are liable to peer pressure but girls aren't. However, the peer group in this case is limited to the classroom; in reality it could differ or be much larger. Tesař (2011) also found the peer effects to be an important factor for youngsters, particularly, the influence of friends and family. According to

a survey by Kralikova et al. (2013) many adolescent initiators are unable to quit their tobacco addiction. Later on, they form the non-teenager percentage of the smoking population as they age.

Oh et al. (2010) analyzed determinants of smoking initiation among women from 5 EU countries (Czech Republic, France, Ireland, Italy, and Sweden). The results indicate that 13,7% of Czech female smokers developed the habit at age 14-15; but that is still the best result compared to other studied countries. The average initiation age for the Czech Republic was 19.6. The logit model estimated peer effects of friends and family and higher age as important determinants of smoking.

Kvaček (2011) applied the theory of rational addiction by Becker and Murphy (1988) to study the demand for tobacco and estimated its price elasticity as -0.2. This indicates that Czech tobacco consumers are quite inelastic.

2.2 Predicting Future Consumption

Mendez et al. (1998) uses similar approach to mine to predict prospective tobacco prevalence. Their model is more detailed as they account for smoking cessation and differ death rates for smokers and nonsmokers. Their approach is specific to tobacco as its addiction is an important factor to reflect in the model.

When it comes to non-addictive goods, many researchers use age-(period-)cohort analysis, which is more advanced than my model, to predict future consumption. This method estimates the effects of age, birth cohort and time period, however, as have Mason et al. (1973) pointed out, in this case can the confounding effect of these variables cause trouble in the regression analysis. This issue has been addressed by Rentz and Reynolds (1991) and further by Carstensen (2007) who proposed incorporating spline functions to estimate age, period and cohort as an continuous variables. However, these approaches are beyond the scope of this thesis and my simple model is based primarily on behavior specific to tobacco consumption, own empirical re-

sults and intuition.

Some examples of age-cohort model applications are by Mori et al. (2004), who used the Bayesian model modification to predict future food consumption in Japan, and by Kerr et al. (2004), who studied the age and cohort effects on alcohol consumption in the United States.

3. Theoretical Model

3.1 Model Description, Assumptions and Limitations

Any model aiming to forecast future smoking prevalence and tobacco consumption through population statistics has to cover several aspects. Especially, a good population size and structure projection is essential. This can be part of the model itself or some external model projections can be directly applied (as in my model). Further, the model has to be able to utilize as much information about the population of smokers as possible in order to provide accurate forecasts. The model that I propose is adjusted to be applied in the Czech Republic where thorough periodical and consistent analyses of smoking prevalence, cessation and death rates of smokers within age categories are not available. Then a simplification is required and that imposes some rather strong assumptions and restrictions which limit the accuracy of the model. Though, for an approximate estimation this model is applicable. The model is based on an assumption that smokers initiate smoking as adolescents and remain smokers throughout their whole life, i.e. the smoking prevalence and some base level of consumption (adolescent/initial consumption) is cohort specific, while coefficients determining the current consumption are age specific. Therefore, any model predictions are scenario-based - depending on the characteristics of future adolescents. The model can be mathematically expressed as:

$$|t_j - t_{j+1}| = |a_i| \tag{3.1}$$

$$s_{a_i,t_j} = P_{a_i,t_j} \cdot r_{t_{j-i+1}} \tag{3.2}$$

$$S_{t_j} = s_{a_1,t_j} + s_{a_2,t_j} + \ldots + s_{a_n,t_j} = \sum_{i=1}^{n} s_{a_i,t_j} \tag{3.3}$$

$$c_{a_i,t_j} = b_{a_i} \cdot \alpha_{t_{j-i+1}} \cdot s_{a_i,t_j} \tag{3.4}$$

$$C_{t_j} = c_{a_1,t_j} + c_{a_2,t_j} + \ldots + c_{a_n,t_j} = \sum_{i=1}^{n} c_{a_i,t_j} \tag{3.5}$$

where:

t_j	=	time; $j \in \mathbb{N} \cup \{0\}$
a_i	=	age category; $i \in [1, \ldots, n]$ with a_1 as the youngest
P_{a_i,t_j}	=	size of population of age category a_i at time t_j
r_{t_j}	=	percentage of smokers within the youngest age category at time t_j
s_{a_i,t_j}	=	population of smokers within age category a_i at time t_j
S_{t_j}	=	total population of smokers at time t_j
α_{t_j}	=	average tobacco consumption of a_1 at time t_j
b_{a_i}	=	tobacco consumption development coefficient for a_i; $b_{a_1} = 1$
c_{a_i,t_j}	=	tobacco consumption of a_i at time t_j
C_{t_j}	=	total tobacco consumption at time t_j.

These are the assumptions of the model:

- a_1 is an age category covering teenage years up to 20 years of age.
- Each age category covers the same number of consequent years of age. Each year of age belongs to only one age category.
- The time between t_j and t_{j+1} in years equals the size of one age category, such that between t_j and t_{j+1} the whole (surviving) population of P_{a_i,t_j} moves to $P_{a_{i+1},t_{j+1}}$ (given by equation (3.1)). This dynamics is assumed, however, not modeled as $P_{a_{i+1},t_{j+1}}$ is taken as an exogenous variable (external population projection model outcomes are plugged in).
- Mortality and migration is accounted for through P_{a_i,t_j}.
- The probability of death is the same for smokers as for nonsmokers.
- All smokers develop(ed) their smoking habit in their teenage years (when in a_1) and are unable to quit later on.
- There is no smoking cessation and initiation after 20 years of age.

- Aging after 20 years of age does not influence smoking status (smoker/nonsmoker) but consumption volume.

- Smokers' consumption in a_i is determined as the product of a corresponding predetermined coefficient b_{a_i} and their past consumption in a_1 (α_{t_j}).

Further, the model can be modified to distinguish sex groups within age categories, which is a desirable property as smoking prevalence and consumption differs significantly between these two subgroups. The upper M or F index denotes the male or female subgroup. Then the model has the following form (assumptions equivalent to those above apply):

$$|t_j - t_{j+1}| = |a_i| \tag{3.6}$$

$$s^M_{a_i,t_j} = P^M_{a_i,t_j} \cdot r^M_{t_{j-i+1}} \tag{3.7}$$

$$s^F_{a_i,t_j} = P^F_{a_i,t_j} \cdot r^F_{t_{j-i+1}} \tag{3.8}$$

$$s_{a_i,t_j} = s^M_{a_i,t_j} + s^F_{a_i,t_j} \tag{3.9}$$

$$S_{t_j} = s_{a_1,t_j} + s_{a_2,t_j} + \ldots + s_{a_n,t_j} = \sum_{i=1}^{n} s_{a_i,t_j} \tag{3.10}$$

$$c^M_{a_i,t_j} = b_{a_i} \cdot \alpha^M_{t_{j-i+1}} \cdot s^M_{a_i,t_j} \tag{3.11}$$

$$c^F_{a_i,t_j} = b_{a_i} \cdot \alpha^F_{t_{j-i+1}} \cdot s^F_{a_i,t_j} \tag{3.12}$$

$$c_{a_i,t_j} = c^M_{a_i,t_j} + c^F_{a_i,t_j} \tag{3.13}$$

$$C_{t_j} = c_{a_1,t_j} + c_{a_2,t_j} + \ldots + c_{a_n,t_j} = \sum_{i=1}^{n} c_{a_i,t_j} \tag{3.14}$$

When it comes to modeling the population size of smokers, the model follows some basic principles of a model introduced by Mendez et al. (1998), however, the original model is more complex as it in addition accounts for smoking cessation and differentiates smokers' and nonsmokers' death rates. On the contrary, my model takes sex differences and migration into account. The assumption that future smoking status of an individual is shaped before the age of twenty is strong, but empirically proven by Kralikova et al. (2013) and numerous surveys in the Czech Republic.

The approach to consumption modeling is based on my own empirical research (see chapter 5), it assumes that the initial level of tobacco consumption (before twenty) is the main determinant of future consumption. Then the future levels are given as factors of the initial consumption of each age group. My analysis shows that these factors can be easily estimated using count variable regression models with suitable data.

Altogether, the model follows these assumptions in this way: It takes population projection P_{a_i,t_j} for each age category; for every birth cohort, currently in some age category, it finds its smoking rate $r_{t_{j-i+1}}$ and average tobacco consumption $\alpha_{t_{j-i+1}}$, both at the period t_{j-i+1} when this birth cohort was located in adolescent age category. Then, it multiplies the current age category size P_{a_i,t_j} by the smoking rate $r_{t_{j-i+1}}$ (smoking rate is assumed to be cohort specific and constant) to get the number of smokers s_{a_i,t_j} in every age category and the sum for whole population S_{t_j} afterwards.

The model assumes that average consumption of a birth cohort changes through life, but it is dependent on its average adolescent tobacco consumption. In particular, it is determined by the birth cohort - by its average adolescent tobacco consumption $\alpha_{t_{j-i+1}}$ - and by the age category this cohort is currently in - by its corresponding coefficient b_{a_i}. These coefficients $b_{a_1}, \ldots, b_{a_n}$ specify the change in consumption with aging, are specific to every age category, predetermined by empirical analysis and constant in time. Hence, for each age category the model multiplies the number of smokers s_{a_i,t_j} by $\alpha_{t_{j-i+1}}$ and b_{a_1} to obtain the tobacco consumption c_{a_i,t_j} within the corresponding category and total tobacco consumption C_{t_j} as their sum afterwards.

With the change to the next period t_{j+1}, each birth cohort moves one age category up. Then, its size changes to $P_{a_{i+1},t_{j+1}}$, and the respective coefficient is now $b_{a_{i+1}}$, the rest stays constant ($r_{t_{j-i+1}} = r_{t_{(j+1)-(i+1)+1}}$;$\alpha_{t_{j-i+1}} = \alpha_{t_{(j+1)-(i+1)+1}}$). This means that the changes of consumption of a cohort are triggered by the change of

respective coefficients, which captures the aging effect, and demographics, which captures the effect of population size change. In this upcoming period, the model needs to be supplied with data for the adolescent population (the youngest age category) and its smoking habits as the model is unable to retrieve this information from the previous period. The population size $P_{a_1,t_{j+1}}$ is provided by the external population projection. The corresponding smoking rate among teenagers and their average consumption for next periods are unknown. In my empirical analysis, I vary these variables according to some scenarios in order to see how much the model output changes. The above described mechanisms of the model work successively for each period, such that the outcomes of the model are only dependent on the characteristics of smoking behavior of future teenagers and their cohort size.

The main limitations of the model are direct consequences of its assumptions. Smoking cessation and initiation rates are not plugged in, but as have been already mentioned this simplification does have at least some empirical justification by Kralikova et al. (2013). Unfortunately, the lack of data creates artificial restrictions that simplify the model without any empirical or theoretical rationale.

A good example is the fact that the model assesses the same probability of death to smokers as to nonsmokers while the nonsmokers' death rate should be certainly higher. Doll et al. (2004) provide evidence from the United Kingdom: at any age higher than 35 they found the probability of dying to be two to three times higher for smokers versus nonsmokers. To illustrate the situation in the Czech Republic, let's focus on year 2002. The smoking rate in 2002 was 24.1% according to OECD database, Sovinová et al. (2008) found out that up to 30% of Czech mortality in 2002 was smoking-attributable. If the probabilities of death were the same, these numbers would be close as smokers would have the same share among dead as among the living population. The assumption of equal probabilities originates from the use of external population projection and is avoidable only though implementation of this projection model into my model along with the differentiation of death rates. However, there are currently no studies or data specifying these death rates of smok-

ers in the Czech Republic.

The model coefficients b_{a_i} are assumed to be constant in time, however, they actually may vary with time. Rather than a cross-sectional data a time series data would be more appropriate for their estimation. Time series analysis would also make sure that the coefficients are not biased by corresponding age cohorts, nevertheless, suitable and consistent data are not available at this point. As a result of that the predictive value of the model is at its highest in the short-run and decreases further on as the coefficients and behavioral patterns of smokers might change over time.

3.2 Possible Model Extensions

As have been discussed before, the model could be more comprehensive and complex if certain data data were available. For now, I ignore these limitations and introduce theoretical model enhancements that improve the model in some ways.

If the model were to include population model (without migration), it would resemble the model of Mendez et al. (1998). This equation would then supplement the model:

$$P_{a_i,t_j} = P_{a_{i-1},t_{j-1}} \cdot (1 - \sigma_{a_{i-1},t_{j-1}}) \tag{3.15}$$

where $\sigma_{a_{i-1},t_{j-1}}$ is the death rate of population aged in a_{i-1} at time t_{j-1}.

If the smoking cessation and initiation were to be included, the equation (3.2) of the model would look as follows:

$$s_{a_i,t_j} = s_{a_{i-1},t_{j-1}} \cdot (1 - \gamma_{a_{i-1},t_{j-1}}) \cdot (1 - \delta_{a_{i-1},t_{j-1}}) + P_{a_{i-1},t_{j-1}} \cdot \mu_{a_{i-1},t_{j-1}} \tag{3.16}$$

where $\gamma_{a_{i-1},t_{j-1}}$, $\delta_{a_{i-1},t_{j-1}}$, $\mu_{a_{i-1},t_{j-1}}$ are the death rate, smoking cessation rate and initiation rate for population aged in a_{i-1} at time t_{j-1}. Unfortunately, this approach would be problematic when modeling resulting tobacco consumption as the new initiators would have no α_{t_j} and one would have to assume they follow the age cohort average.

A significant improvement in accuracy would be achieved if the coefficients b_{a_i} were taken as functions of time $b_{a_i}(t)$ and, thus, allowed to change over time. Then, the modified equation (3.4) would have this form:

$$c_{a_i,t_j} = b_{a_i}(t) \cdot \alpha_{t_{j-i+1}} \cdot s_{a_i,t_j}. \tag{3.17}$$

A better alternative would be the transformation of the consumption part of the model to an age-cohort approach. Then the cohort effects would be clearly separated from age effects, nevertheless, a proper model estimation following the approach of Carstensen (2007) is beyond the scope of this bachelor thesis.

3.3 Forecasting Taxation Revenues

If the model predicts tobacco consumption and we assume that taxation revenue changes proportionally, then the model outcomes can be expressed in terms of future taxation revenues. However, predicting future revenues from the tax on tobacco is a troublesome task as the level of taxation might change in the future and the model is unable to capture these changes effectively. It models all predictions assuming the present price level and taxation policy. Hence, the outcome should be considered only as a qualified approximation of the future state. Here is the key equation showing very simple relation of tax revenues to tobacco consumption:

$$T_{t_j} = T_{t_0} \cdot \frac{C_{t_j}}{C_{t_0}} \tag{3.18}$$

where T_{t_0} is the present tobacco tax revenue. In my empirical analysis, the input for T_{t_0} is the 2013 tax revenue of 46.82 billions CZK.

4. Data

4.1 Population Projections

In my analysis I use EUROPOP2013 (European Population Projections, base year 2013) population projections from 2013 by EUROSTAT. It provides detailed projections specifically for the Czech Republic, for both sexes separately and for every single year of age. EUROSTAT computed several scenarios - the main scenario, low fertility, higher life expectancy, reduced migration and zero migration variants. I apply only the main scenario; its horizon is the year of 2080.

Another option would be to choose the UN Probabilistic Population Projections based on the World Population Prospects: The 2012 Revision, which incorporates Bayesian probabilistic modeling in a way described by Raftery et al. (2012), however, these projections at this point do not provide a sufficient detail - the data specific for the Czech Republic are structured into age categories, not by a single year of age.

4.2 Coefficients Estimation Data

The dataset resulted from a 2012 (Q4) survey collected by SC&C (Czech market research company) for the Czech National Monitoring Centre for Drugs and Drug Addiction. This survey is run periodically every 4 years. Unfortunately, the questions differ substantially from those in previous version from 2008 and the sample is not the same as well. Thus, I could not apply statistics from past years and form a time series data. Hence, I use only the data from 2012 for cross-sectional analysis. The purpose of this survey is to analyze the consumption of drugs (not only tobacco and alcohol but illegal ones as well) and its motivation on an individual level. All of the data is self reported. This dataset was collected by a professional market

research agency, SC&C, using the method of stratified sampling in all parts of the Czech Republic. Then, weighting has been applied to the sample with the following criteria (in the given order):

1. Age
2. Sex
3. Education
4. Region (NUTS2)
5. Municipality size
6. Economic status

The dataset consists of 2,135 observations, however, I am able to use only 1,512 of them. The reason behind this is the fact that many respondents left some parts of the questionnaire blank or their responses simply did not make any sense (mistakes, nonsense). In the regression model I use 38 variables. Their list with descriptions can be found in Table 4.1.

The summary statistics of all variables is located in the Appendices of this paper (Table 7.1) along with tables 7.2-7.9 specifying frequency and relative share of the zero/one values of all dummy variables. Apparently, more than 28% of the respondents were smokers (Table 7.9).

Most of the listed variables are sociodemographic - describing the position of the individual in the society, his SES (through education, income, etc.), or his personal characteristics (age, sex, etc.). Besides that, there is one variable, which allows to observe respondent's health behavior other than smoking - *alcabuse*. It is a dummy that takes the value of one if the respondent is an alcoholic. *Alcabuse* was generated from the original statistics (which specified the frequency of drinking) using the following rule: Abusive drinkers (alcoholics) are determined by consuming five or more drinks during one occasion in a single month (or more frequently). The definition

Variable	Type	Description
cigsdaily	count	number of cigarettes smoked daily
smokes	dummy	=1 if respondent smokes
alcabuse	dummy	=1 if respondent is an alcoholic
male	dummy	=1 if male
age	count	the age of respondent
b1-b6	dummy	Age categories - see Table 4.2
unemp	dummy	=1 if unemployed
student	dummy	=1 if student
maternity	dummy	=1 if on maternity leave
retired	dummy	=1 if retired
disabled	dummy	=1 if disabled
msize1-msize6	dummy	Municipality size - see Table 7.11
edu1-edu5	dummy	Highest earned education - see Table 7.12
inc1-inc7	dummy	Income group - see Table 7.13

Table 4.1: Variables

might seem broad but the responses are possibly underrated and that should offset it.

Variables $b1 - b10$ are dummies corresponding to age categories from the theoretical model. Their estimated incidence rate ratios will play the role of model coefficients.

Dummy variable *smokes* is the binary dependent variable for the binary part of the zero-inflated negative binomial (ZINB) model while *cigsdaily* is the dependent variable for the negative binomial part of the ZINB model. Their types corresponds to the approach to analyzing the tobacco consumption - the first one as a dummy represents the extensive margin (smoke or not to smoke) while the other one the intensive margin (cigarettes per day conditional on being a smoker). The figure above illustrates the distribution of *cigsdaily* among smokers (zero values omitted) in the sample. Apparently, the most common answers were ten and twenty cigarettes daily, probably because of rounding.

	Frequency	Percent	Age
b1	96	6.35	15-19
b2	195	12.90	20-24
b3	181	11.97	25-29
b4	237	15.67	30-34
b5	167	11.04	35-39
b6	125	8.27	40-44
b7	107	7.08	45-49
b8	111	7.34	50-54
b9	126	8.33	55-59
b10	167	11.04	60-64

Table 4.2: Age categories

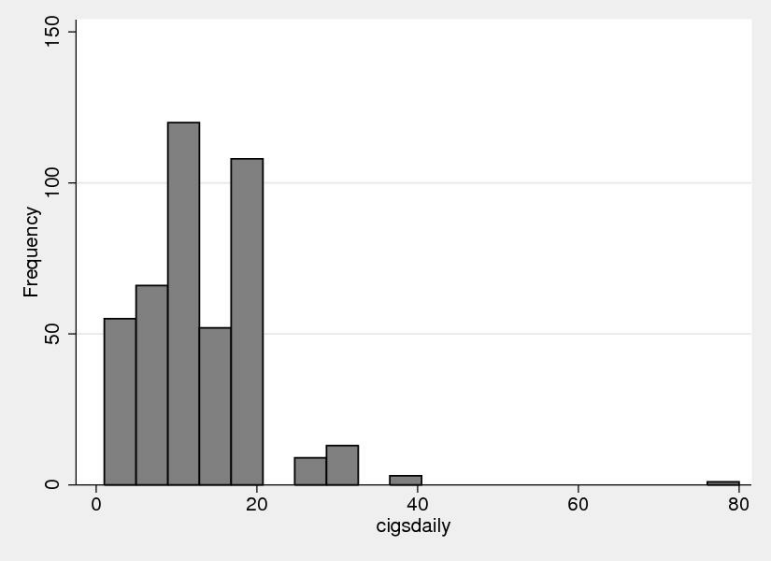

Figure 4.1: Frequency of cigarettes/day in the subsample of smokers

5. Coefficients Estimation

5.1 Methodology

5.1.1 Maximum Likelihood Estimation

For my ZINB model with count dependent variable, using standard linear regression methods like OLS or WLS would be inappropriate. The reason is the non-linearity of $E(y_i \mid \mathbf{x})$. Despite the existence of nonlinear variants of the mentioned methods, using the maximum likelihood estimation (MLE) is a neater option. In short, MLE is an estimation method maximizing the likelihood function. A common definition of likelihood function for discrete variables has the following form:

Definition 1. Let X be a random variable with a discrete probability distribution p with a parameter of θ. Then the function

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x), \tag{5.1}$$

as a function of θ, is called the likelihood function (of θ given the outcome x of X).

To be accurate, I should point out that MLE usually maximizes the log-likelihood function ℓ for the sample of n. One can get this function simply by taking the logarithm of the likelihood function. The definition says that the likelihood function actually represents the odds that certain outcome y is realized.

MLE is based on distribution of y conditional on $\mathbf{x}$, thus it automatically accounts for heteroskedasticity in $Var(y \mid \mathbf{x})$.

5.1.2 Deriving the Zero-Inflated Negative Binomial Model

Count Variable in a Regression

Count variable is a variable whose value can be only discrete, in particular, zero or any positive integer (0,1,2,3,...). Such a variable usually represents count of occurrences in some time frame, that is the reason behind the limitation to positive integers and zero values. Using a count variable in a standard OLS linear regression is not appropriate because assumptions like homoscedasticity and normality of errors might be violated.

Possible alternatives suitable for this kind of data are the Poisson and negative binomial regression models. A common issue when using the Poisson regression is that equidispersion, the equality of conditional mean and variance $Var(y_i|\mathbf{x_i}) = E(y_i|\mathbf{x}_i)$, is assumed in Poisson distribution and the data may (and very often they do) actually violate this rather strong assumption and exhibit overdispersion: $Var(y_i|\mathbf{x_i}) > E(y_i|\mathbf{x}_i)$. This can underestimate the standard errors. Overdispersion can result from some omitted heterogeneity or due to an aspect called state dependence. State dependence occurs when the observed events are not independent (e.g. smoking the second cigarette is not an independent event because the individual had to smoke the first before, thus the likelihood of smoking the first cigarette and the likelihood of smoking two cigarettes are not independent). Though there exists a Poisson model modification for overdispersed data, using negative binomial model which accounts for overdispersion is nowadays a more common practice. Because overdispersion is also an issue relevant to my data I prefer using negative binomial model as well.

Running the likelihood-ratio test compares both models and checks for overdispersion (more on tests in a separate section later on). Hence, one can easily decide which model suits the data better.

Negative Binomial Model

As mentioned above, the negative binomial model, unlike the Poisson model, accounts for overdispersion. This feature originates from the form of the variance of the NB2 model (notation from Cameron and Trivedi (1986)) which is the standard and the most common form of the negative binomial model. Its variance function has the following form:

$$Var(y_i|\mathbf{x_i}) = \mu + \alpha\mu^2, \tag{5.2}$$

where α is the overdispersion (or heterogeneity) parameter. For example, the other (less used) version of the model, denoted NB1, has variance of $Var(y_i|\mathbf{x_i}) = \mu + \alpha\mu$. Notice, that if the dispersion parameter α in any of the models is zero, the model variance has properties of the Poisson model ($Var(y_i|\mathbf{x_i}) = E(y_i|\mathbf{x}_i)$). Because NB2 is adopted (as default option) by all major statistical programs, I will, from now on, work only with NB2.

A convenient way of deriving the NB2 model is from a Poisson-gamma mixture - a Poisson model with gamma heterogeneity (with the mean of 1) which accounts for correlated and overdispersed outcomes. Hilbe (2011) defines this mixture as

$$f(y;\lambda,u) = \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} \tag{5.3}$$

he further modifies this formula and derives

$$P(y_i = Y) = f(y_i;\lambda,u) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1+\alpha\mu_i}\right)^{y_i} \tag{5.4}$$

which is the probability mass function (PMF) of negative binomial distribution where $\alpha > 0$ is the overdispersion parameter and $\mu > 0$ represents the mean of y. PMF denotes the probability that the count variable y equals some value Y.

Hence, the general forms of the likelihood function and log-likelihood functions are:

$$L(\mu; y, \alpha) = \prod_{i=1}^{n} \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1+\alpha\mu_i}\right)^{y_i} \tag{5.5}$$

$$\ell(\mu; y, \alpha) = \sum_{i=1}^{n} \Bigg[y_i(\alpha + \mu) + \Big(y_i + \frac{1}{2}\Big)\Big(\ln(\Gamma) - \ln(1+\alpha\mu)\Big) - \ln(\Gamma)\Big(y_i + 1 + \frac{1}{\alpha}\Big)\Bigg] \tag{5.6}$$

Now, let's take into account that the standard negative binomial model has this form:

$$\ln\mu = \mathbf{x}\beta \tag{5.7}$$

where $\mathbf{x}$ is the set of explanatory variables with ones in the first column and β is the vector of regression coefficients that we want to estimate. Using 5.7 in 5.5 and 5.6 gives

$$L(y, \alpha, \beta) = \prod_{i=1}^{n} \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha e^{x_i\beta}}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1+\alpha e^{x_i\beta}}\right)^{y_i} \tag{5.8}$$

$$\ell(y, \alpha, \beta) = \sum_{i=1}^{n} \Bigg[y_i(\alpha + e^{x_i\beta}) + \Big(y_i + \frac{1}{2}\Big)\Big(\ln(\Gamma) - \ln(1+\alpha e^{x_i\beta})\Big) - \ln(\Gamma)\Big(y_i + 1 + \frac{1}{\alpha}\Big)\Bigg]. \tag{5.9}$$

Maximizing the log-likelihood function (maximum likelihood estimation) and estimating β and α can be achieved using several methods. The the vast majority of statistical packages uses the Newton-Raphson method.

Zero-Inflated Model

Having a count data in a regression can cause several problems, where many of them are caused by zero counts. The one that is relevant to my analysis is the inflation of zeros, the fact that zeros can be generated by a different process (decision) than the positive values. For example, consider two respondents to a survey, being questioned about the volume of their alcohol consumption. The first one as a never drinker reports zero, while the other one is a potential drinker but reports zero as well (e.g., he can choose to consume zero on a given day purely as an economic choice). Obviously, these zeros are not the same and have to be somehow accounted

for in the model. The key is to distinguish between the structural zeros (the never drinker zero) and zeros of those participating in the process that generates the discrete values of the count variable. This is where the zero-inflated model steps in.

Usually, zero-inflated models use binary regression models (logit or probit) to model the participation in the activity and a count regression models (NB2, Poisson) for estimation within the group of participants. Alternatively, one can also apply hurdle models, however, hurdle models exclude all the zeros from the count process. For example, everyone shopping at a store is then accounted as buyer, though he may choose to buy nothing. Due to this limitation and its lower strictness are zero-inflated models more suitable for my data as it is hard to draw a clear line between smokers and nonsmokers and clearly categorize occasional and relapsing smokers.

Now, let's discuss the zero-inflated negative binomial model (ZINB) as it builds upon the model derived in the previous section. Though it is possible to use probit with ZINB, I choose to use logit as it is a more common practice. Nevertheless, both ways are applicable. Usually, the second stage of the model consists of the NB2 version of the negative binomial model. Notice, that the log-likelihood function must be then different for cases when $y = 0$ and $y > 0$. The log-likelihood functions for ZINB with logit are given by Hilbe (2011) as:

$$\text{if } y = 0: \quad \ell(\alpha, \beta) = \Sigma_{i=1}^{n} \left\{ \ln\left(\frac{1}{1 + \exp(-x_i\beta_1)} \right) + \frac{1}{1 + \exp(x_i\beta_1)} \left(\frac{1}{1 + \alpha\exp(x_i\beta)} \right)^{\frac{1}{\alpha}} \right\}; \tag{5.10}$$

$$\text{if } y > 0: \quad \ell(\alpha, \beta) = \Sigma_{i=1}^{n} \left\{ \ln\left(\frac{1}{1 + \exp(-x_i\beta_1)} \right) + \ln\Gamma\left(\frac{1}{\alpha} + y_i\right) - \ln\Gamma(y_i + 1) - \ln\Gamma\left(\frac{1}{\alpha}\right) + \left(\frac{1}{\alpha}\right)\ln\left(\frac{1}{1 + \alpha\exp(x_i\beta)} \right) + y_i\ln\left(1 - \frac{1}{1 + \alpha\exp(x_i\beta)} \right) \right\} \tag{5.11}$$

where β_1 stands for the regression coefficients for the binary part, whereas β for the count part of the model. Hence, the model can be estimated using MLE.

Likelihood-Ratio and Vuong Tests

Testing the fit of the model is an essential part of any econometric analysis. The likelihood-ratio test compares the values of maximized log-likelihood functions of each model. The test has the following form

$$LR = -2\ell_1 + 2\ell_2 \tag{5.12}$$

where ℓ_1 is the log-likelihood value of the model preferred under null hypothesis and ℓ_2 is of the other one. LR follows chi-squared distribution. After computing the difference in degrees of freedom, one can find the critical value for a given confidence level and decide which model is better.

Vuong test (Vuong, 1989) is a test comparing the ZINB model to the standard NB2 model. It checks for significant difference between the fitted values of the models. The Vuong test forms a z-statistic

$$Z = \frac{\sqrt{n}\mu}{\sigma_i} \tag{5.13}$$

where μ is the mean of u_i, σ_i is the standard deviation of u_i and

$$u_i = \ln\left(\frac{\Sigma_i P_{NB2}(y_i|x_i)}{\Sigma_i P_{ZINB}(y_i|x_i)}\right). \tag{5.14}$$

The test compares the probabilities $P(y_i|x_i)$ of outcome y_i given x_i. Under the null hypothesis, ZINB is the better model. The test incorporates the normal distribution; one can easily check for any confidence level and decide which model suits the data better.

Interpretation

Unfortunately, the interpretation of the ZINB model is not that straightforward. In this model, the NB2 part is of primary interest. Because the ZINB model has a log-form, the beta-parameter estimates represent the differences in expected counts. For a one-unit change in p^{th} predictor x_p (where $1 \leq p \leq n$) this is the difference in predicted counts:

$$\begin{aligned} \log(y \mid \mathbf{x}') - \log(y \mid \mathbf{x}) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p(x_p + 1) + \ldots + \beta_n x_n \\ &\quad -(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \ldots + \beta_n x_n) \qquad (5.15) \\ &= \beta_p \qquad (5.16) \end{aligned}$$

where $\mathbf{x}'$ is the set of explanatory variables after the change in p^{th} variable. After exponentiation we get an incidence rate ratio (IRR):

$$\text{IRR}_p = \frac{y \mid \mathbf{x}'}{y \mid \mathbf{x}} = e^{\beta_p} \qquad (5.17)$$

IRR denotes the ratio of incident rates for different values of a given predictor.

5.2 Consumption Coefficients Estimation

The ZINB regression model was estimated using MLE in STATA with *cigsdaily* as a dependent variable. The estimates of the logit part of the model have only auxiliary function to NB2 estimation, are of no interpretative value and actually vary from a standard logit results as currently a modified log-likelihood function is incorporated. The insignificant variables are omitted from the logit part including the constant term. Logit was estimated with variable *age* instead of $b1 - b10$ dummies for simplicity, *age* is significant at 5% level.

The model estimation with fully specified NB2 part (Table 7.14 in Appendix) shows that income, municipality size, retirement, unemployment, maternity, and disability are not significant predictors of tobacco consumption at the 5% level. Resulting model (Table 5.1) confirms significance for all the other variables (jointly for categorical) as well as for the intercept at the 5% significance level and *student* at 10%. The corresponding calculated IRRs supplement the estimated coefficients in Table 5.1.

The positive coefficient of alpha (the heterogeneity parameter) of 0.2429 proves that some overdispersion is really present. Though, careful testing is still necessary. The LR test between this ZINB model and its Poisson alterative, ZIP (zero-inflated Poisson), results in a p-value of zero, that implies that ZINB suits the data better. The Vuong test compares ZINB to the standard NB2 model; with a z-statistic of 12.67 and corresponding p-value of zero is ZINB again the better of the two. This concludes that choosing ZINB model among other count models was the right choice. Unfortunately, no direct indicators for the interpretative value of the model (like R^2/pseudo R^2) are provided by Stata.

Let's focus on the IRRs and interpret our estimation results as this might be interesting for health economists. Among those who smoke, being an abusive drinker increases tobacco consumption by 21.75% (by a factor of 1.2175), holding all other

Variable	**IRR**	**Coefficient**	**(Std. Err.)**
alcabuse	1.2175	0.197 **	(0.059)
male	1.3749	0.318 **	(0.062)
b2	1.13	0.122	(0.175)
b3	1.3381	0.291	(0.195)
b4	1.2966	0.260	(0.197)
b5	1.2454	0.219	(0.200)
b6	1.2124	0.193	(0.207)
b7	1.5759	0.455 *	(0.208)
b8	1.6906	0.525 **	(0.203)
b9	1.6262	0.486 *	(0.204)
b10	1.4531	0.374 †	(0.205)
student	0.7347	-0.308†	(0.158)
edu2	0.6985	-0.359**	(0.099)
edu3	0.6613	-0.414**	(0.102)
edu4	0.6423	-0.443*	(0.191)
edu5	0.5886	-0.530**	(0.135)
Intercept		2.318 **	(0.174)
Inflate (logit):	Table 7.10 in Appendices		
\lnalpha		-1.415**	(0.100)
alpha		.2429	(0.024)
N (zero/nonzero)		1512 (1085/427)	
Log-likelihood		-2209.77	
$\chi^2_{(8)}$		98.11	

Significance levels : † : 10% * : 5% ** : 1%

Table 5.1: Estimation results : zinb

factors constant. Similarly, being a male by a factor of 1.3749 (37.49%). The effects of indicators of education are very similar. Student's volume consumed is smaller by a factor of 0.7347 (-26.53%) than non-student's. The respective factors for $educ2-5$ are 0.6985 (-30.15%), 0.6613 (-33.87%), 0.6423 (-35.77%) and 0.5886 (-41.14%), compared to $educ1$ level.

The IRRs of $b1 - b10$ show that the age influence is not linear, these are the coefficients sought for $b_{a_1}, \ldots, b_{a_{10}}$. Though some coefficients are insignificant, $b1 - b10$ are jointly significant at the 10% level and thus applicable in the theoretical model. These IRRs along with corresponding age categories are once again for clarity reported in the following table:

Age category	Age	b_{a_i}
a_1	15-19	1
a_2	20-24	1.13
a_3	25-29	1.3381
a_4	30-34	1.2966
a_5	35-39	1.2454
a_6	40-44	1.2124
a_7	45-49	1.5759
a_8	50-54	1.6906
a_9	55-59	1.6262
a_{10}	60-64	1.4531
a_{11}	65+	1.4531

Table 5.2: Age categories and corresponding coefficients

The analysis fails to predict $b_{a_{11}}$ for elderly aged 65+. Hence, another assumption must be formulated:

- On average, people turning 65 do not change their consumption behavior and sustain this consumption level until their death, i.e., $b_{a_{10}} = b_{a_{11}}$

6. Empirical Model

6.1 Additional Model Adjustments

Because I work only with present data, the values of $r_{t_{j-i+1}}$ and $\alpha_{t_{j-i+1}}$ for age cohorts $a_2, \ldots, a_{10}$ are not defined and have to be computed or approximated. For $r_{t_{j-i+1}}$ is the situation quite clear as the model assumes the age cohort's percentage of smokers stays constant and thus $r_{t_{j-i+1}}$ equals the current percentage at time t_0 in given age category. In case of $\alpha_{t_{-i+1}}$ ($j = 0$ because this is an initial model input) I use the assumption that consumption follows the pattern given by $b_{a_1}, \ldots, b_{a_{10}}$. Then, the following formula applies:

$$\alpha_{t_{-i+1}} = \frac{\hat{c}_{a_i}}{b_{a_i}} \tag{6.1}$$

where $\hat{c}_{a_i}$ is average consumption within a_i at time t_0. The derived input data are summarized in the following table:

a_i	$r^M_{t_{-i+1}}$	$r^F_{t_{-i+1}}$	$\alpha^M_{t_{-i+1}}$	$\alpha^F_{t_{-i+1}}$
a_1	0.3243	0.2373	9.08	9.07
a_2	0.3100	0.2421	10.31	8.93
a_3	0.4444	0.2900	10.59	7.32
a_4	0.3254	0.2432	11.12	7.31
a_5	0.3377	0.1778	13.09	5.87
a_6	0.2778	0.2830	11.59	7.09
a_7	0.3000	0.2807	11.00	7.42
a_8	0.3750	0.2727	11.18	7.18
a_9	0.3065	0.2031	11.94	7.14
a_{10}	0.3117	0.1556	9.98	7.96
a_{11}	0.2110	0.2110	9.24	9.24

Table 6.1: Model inputs at time t_0

Again, the absence of respondents aged 65+ (in a_{11}) in the dataset causes trouble. This time I plug in the results from Sovinová et al. (2014) survey where 21.10% of respondents aged 65+ were smokers. The value of corresponding $\alpha_{t_{-10}}$ is based on one more additional assumption:

- I assume, that

$$\alpha_{t_{-10}} = \alpha_{t_{-9}}. \tag{6.2}$$

The actual numbers with M/F indexes for a_{10} differ as $\alpha_{t_{-10}}$ ignores sex differences in a_{11} and counts the population average. This number is consistent with the findings of Sovinová et al. (2014).

This problem was caused by the structure of the dataset, but the missing coefficients and descriptive statistics are only the consequences of a much larger problem - the disproportional size of age group a_{11} which violates assumption (3.1). The model is able to model behavior from 15 to 64, however, later on it lacks the ability to capture any changes and it has to rely on approximation of some constant rates. This artificial limitation and assumption violation can cause some bias in predictive power of the model. Unfortunately, the data does not allow any other solution than the one introduced above. A more thorough and detailed survey data is required for any future improvements.

6.2 Model Scenarios

The outcomes of my model are dependent on future teenage smoking characteristics. This means that several models with different inputs must be built to capture all possible future developments. Firstly, I analyze the smoking population and daily consumption in 2013 (t_0). Secondly, I focus on possible developments of teenage smoking rates and consumption. In this respect, I consider 4 possible period-to-period changes - by $\pm 5\%$ and $\pm 10\%$. The time interval between two periods is 5 years, starting with 2013. These scenarios have this property: The positive development (+5% and +10%) scenarios' parameters grow with time at an increasing

pace, while negative development (-5% and -10%) scenarios' parameters decline with time at an decreasing pace. This model property is intentional and based on the studies of peer influence among adolescents (e.g., Pertold, 2009). Those proved that peer influence is an important determinant of smoking status, that implies the higher the prevalence the higher peer pressure to smoke, the lower prevalence the lower peer pressure but some compulsive smokers still remain present. This makes perfect sense for modeling smoking rates, however, not for average consumption as intensive smoking doesn't make existing smokers to smoke more. Therefore, the average teenage consumption will be studied through different scenarios based on data from Table 6.1.

6.2.1 The Initial State in 2013

The model estimates the total population of smokers in the Czech Republic in 2013 as 2.41 millions. Table 6.2 provides more detail on the distribution within age categories. The estimated daily consumption of cigarettes is 31.72 millions. See Table 6.3 for more detail. The tax revenue according to the Czech Ministry of Finance was 46.82 billions CZK.

a_i	Age	Smokers	Men	Women
a_1	15-19	143 844	84 828	59 016
a_2	20-24	182 697	104 500	78 198
a_3	25-29	262 915	162 827	100 089
a_4	30-34	232 176	136 222	95 955
a_5	35-39	243 497	162 512	80 985
a_6	40-44	209 065	106 441	102 624
a_7	45-49	206 383	108 865	97 517
a_8	50-54	205 687	119 863	85 825
a_9	55-59	185 554	110 018	75 536
a_{10}	60-64	169 100	109 286	59 815
a_{11}	65+	372 967		
Total		**2 413 887**		

Table 6.2: Smokers in 2013

a_i	Age	Daily consumption	Men	Women
a_1	15-19	1 305 885	770 524	535 361
a_2	20-24	2 005 692	1 216 916	788 776
a_3	25-29	3 286 889	2 306 711	980 178
a_4	30-34	2 873 377	1 963 583	909 794
a_5	35-39	3 242 397	2 650 198	592 199
a_6	40-44	2 378 063	1 495 494	882 569
a_7	45-49	3 026 732	1 886 996	1 139 736
a_8	50-54	3 307 314	2 265 977	1 041 338
a_9	55-59	3 014 048	2 136 670	877 378
a_{10}	60-64	2 276 783	1 584 643	692 140
a_{11}	65+	5 005 614		
Total		**31 722 794**		

Table 6.3: Daily cigarettes consumption in 2013

6.2.2 Status Quo

Suppose the current teenage smoking rate (32.43% for men and 23.73% for women) and average consumption (9.1 cigarettes) are steady-states. Then, there will be no changes in these parameters in future. The projected development of smokers in the Czech population is depicted by Figure 6.1, the projected consumption in Figure 6.2 and taxation revenues in Figure 6.3. For numerical results, see Table 6.4.

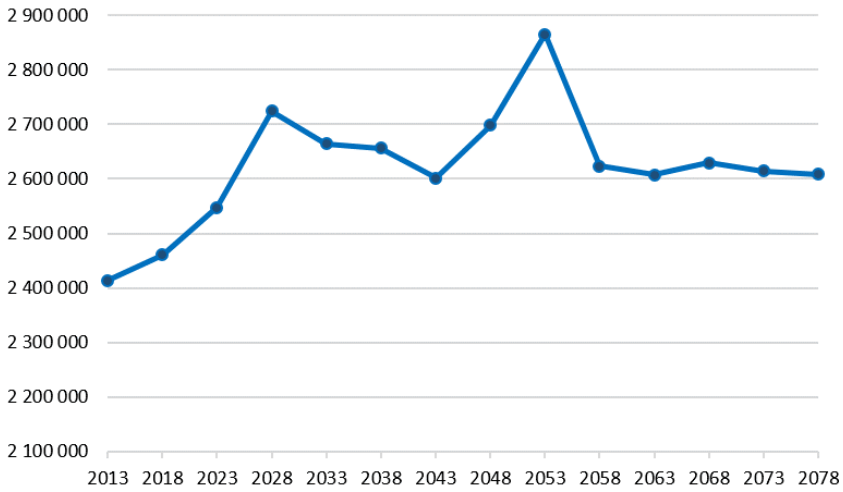

Figure 6.1: Status quo scenario: Smokers

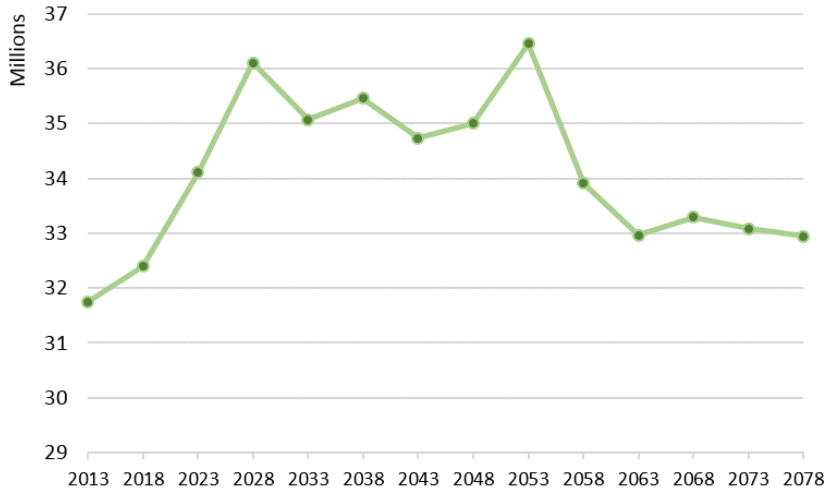

Figure 6.2: Status quo scenario: Daily cigarette consumption

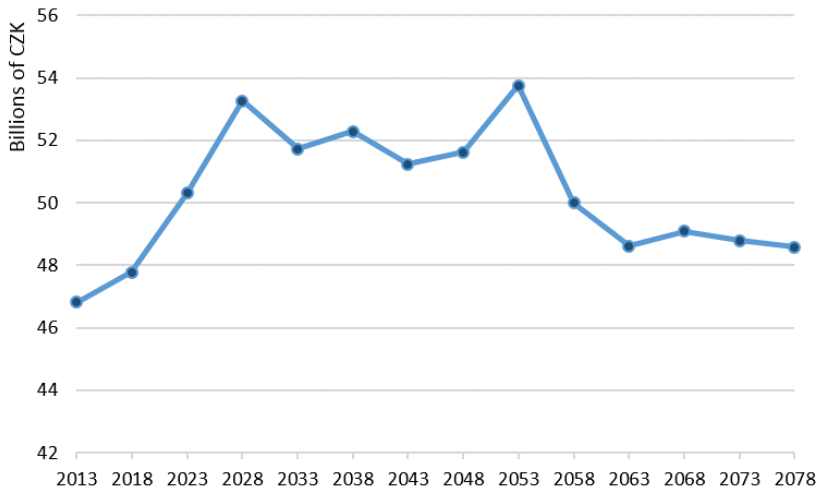

Figure 6.3: Status quo scenario: Tax revenue (in CZK)

All of the charts follow a similar pattern - a boost until 2028, a steady decline, another surge peaking in 2053 followed by a plunge. The relation of consumption to tax revenue is proportional, hence, the graphs look the same. On the other hand, relation of consumption to number of smokers is not so trivial and the plots actually differ. Let's focus on the growth in 2013-2023. The overall daily consumption of cigarettes grew by 7.44%, while the number of smokers only by 5.55%,

Year	Smokers	Consumption	Tax revenue (CZK)
2013	2 413 887	31 746 646	46 820 000 000
2018	2 460 308	32 400 922	47 784 928 117
2023	2 547 749	34 109 318	50 304 471 881
2028	2 723 743	36 111 068	53 256 655 761
2033	2 664 696	35 071 554	51 723 579 196
2038	2 656 084	35 456 392	52 291 139 060
2043	2 601 042	34 737 000	51 230 178 440
2048	2 697 990	34 999 844	51 617 820 895
2053	2 864 609	36 453 267	53 761 331 904
2058	2 623 438	33 905 238	50 003 495 530
2063	2 607 316	32 959 754	48 609 093 293
2068	2 628 833	33 289 395	49 095 248 728
2073	2 614 761	33 081 122	48 788 087 615
2078	2 607 917	32 945 002	48 587 337 699

Table 6.4: Status quo scenario: summary data

6.2.3 Teenage Smoking Rate Scenarios

These scenarios assumes that the percentage of smokers in the group of teenagers varies while average teenage tobacco consumption sustains its initial level from 2013. The model predicts ±5% and ±10% period-to-period changes. These mechanisms can be expressed as:

$$r_{t_{j+1}}=r_{t_j} \cdot (1 \pm 0.05) \text{ or } r_{t_{j+1}}=r_{t_j} \cdot (1 \pm 0.10), \forall j \in \mathbb{N} \cup \{0\}.$$

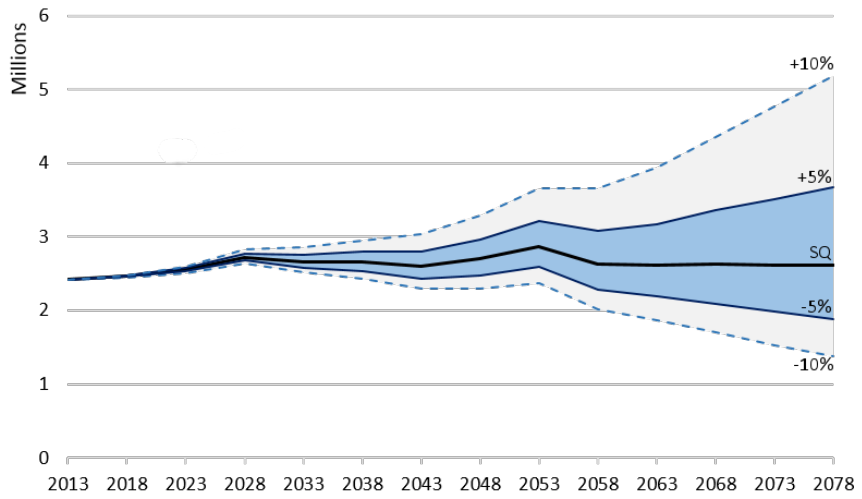

Figure 6.4: Teenage rate scenarios: Smokers

Figure 6.4 shows the predicted smoker population size with the status quo scenario (as SQ) and period-to-period change ±5% and ±10% scenarios. Similarly, the predicted consumption is described in Figure 6.5 and tax revenues in Figure 6.6.

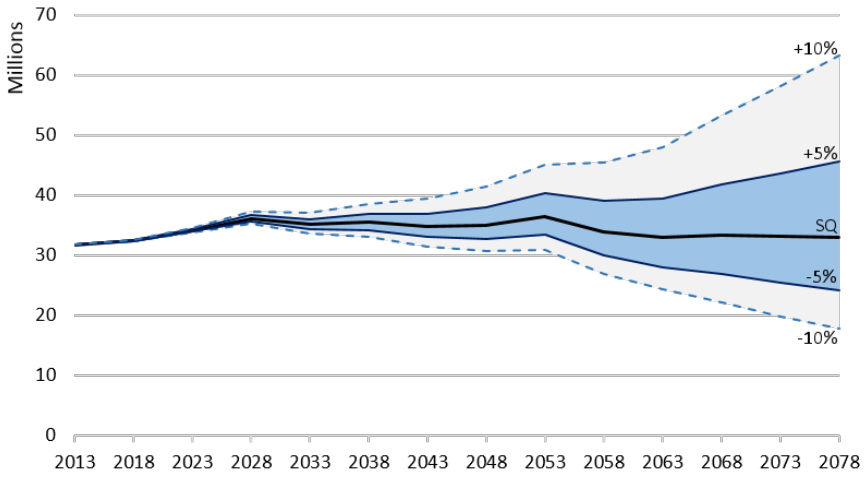

Figure 6.5: Teenage smoking rate scenarios: Total daily consumption

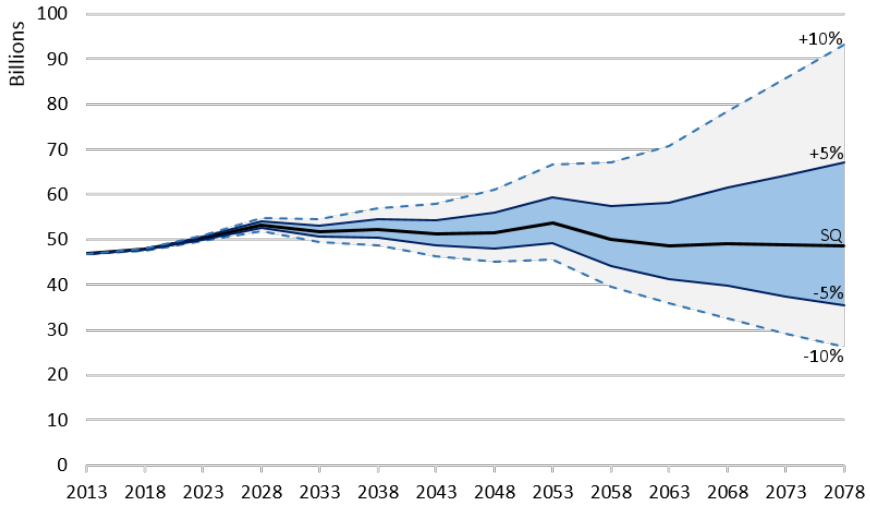

Figure 6.6: Teenage smoking rate scenarios: Tax revenue (in CZK)

Complete model predictions data are summarized in Table 6.5. The numbers indicate that number of smokers, daily consumption as well as tax revenues are going to increase within the next decade in any of these scenarios. These variables will ever reach their initial value from 2013 and further fall only in scenarios -5% and -10%. The drop below the initial state is predicted in period 2038-2043 at the earliest.

		2013	2018	2023	2028	2033	2038	2043	2048	2053	2058	2063	2068	2073	2078
Number of smokers	SQ	2 413 887	2 460 308	2 547 749	2 723 743	2 664 696	2 656 084	2 601 042	2 697 990	2 864 609	2 623 438	2 607 316	2 628 833	2 614 761	2 607 917
	-10%	2 413 887	2 447 212	2 504 882	2 631 581	2 510 865	2 428 932	2 291 753	2 298 745	2 363 974	2 008 649	1 871 699	1 702 366	1 524 379	1 370 939
	-5%	2 413 887	2 453 760	2 525 934	2 676 028	2 583 822	2 534 888	2 433 681	2 478 901	2 585 828	2 275 931	2 185 470	2 095 782	1 980 647	1 878 411
	5%	2 413 887	2 466 856	2 570 327	2 774 855	2 754 105	2 794 320	2 797 873	2 963 861	3 214 480	3 075 185	3 175 113	3 357 591	3 505 873	3 668 302
	10%	2 413 887	2 473 404	2 593 669	2 829 494	2 852 693	2 951 542	3 028 705	3 285 684	3 652 692	3 661 652	3 938 912	4 359 103	4 761 723	5 180 432
Daily consumption	SQ	31 746 646	32 400 922	34 109 318	36 111 068	35 071 554	35 456 392	34 737 000	34 999 844	36 453 267	33 905 238	32 959 754	33 289 395	33 081 122	32 945 002
	-10%	31 746 646	32 286 978	33 739 098	35 217 200	33 514 990	33 037 905	31 438 979	30 637 175	30 843 668	26 822 082	24 314 396	22 136 730	19 798 206	17 775 855
	-5%	31 746 646	32 346 424	33 938 403	35 660 909	34 274 018	34 169 756	32 975 876	32 626 007	33 350 889	29 919 937	28 028 410	26 909 260	25 401 131	24 049 711
	5%	31 746 646	32 465 316	34 357 796	36 641 347	36 030 495	36 915 026	36 883 609	37 916 186	40 331 996	38 966 348	39 423 316	41 750 417	43 556 336	45 504 318
	10%	31 746 646	32 524 762	34 577 884	37 180 430	37 039 733	38 564 568	39 339 478	41 389 835	45 127 200	45 477 389	48 023 923	53 233 196	58 130 567	63 254 834
Tax revenue	SQ	46 820	47 785	50 304	53 257	51 724	52 291	51 230	51 618	53 761	50 003	48 609	49 095	48 788	48 587
	-10%	46 820	47 617	49 758	51 938	49 428	48 724	46 366	45 184	45 488	39 557	35 859	32 647	29 198	26 216
	-5%	46 820	47 705	50 052	52 593	50 547	50 394	48 633	48 117	49 186	44 126	41 336	39 686	37 462	35 469
	5%	46 820	47 880	50 671	54 039	53 138	54 442	54 396	55 919	59 482	57 468	58 142	61 574	64 237	67 110
	10%	46 820	47 968	50 996	54 834	54 626	56 875	58 018	61 042	66 554	67 070	70 826	78 508	85 731	93 288

* Consumption is expressed in number o cigarettes, tax revenues in millions of CZK

Table 6.5: Teenage smoking rate scenarios: Model predictions

6.2.4 Teenage Consumption Scenarios

Now, let's assume the smoking rate among adolescents stays constant at its current rate and only consumption level changes. Interestingly, if its development mechanism were analogical to those in the previous section (5%/10% period-to-period growth/decline), the resulting predictions and charts would be identical to those in the previous section. If current teen smokers increase tobacco consumption by 30%, it is the same as if there were 30% more teen smokers while they all keep their initial consumption level. Nevertheless, as have been discussed before, this approach is not realistic and I prefer to base the scenarios on empirical observations from Table 6.1.

The approach is such that I take the average of all $\alpha^M_{t_{-i+1}}$ and $\alpha^F_{t_{-i+1}}$ (separately) from the table. The resulting $\alpha^{\mathbf{M}} = 10.99$ and $\alpha^{\mathbf{F}} = 7.53$ denote the average adolescent tobacco consumption for each gender during past ten years. I plug these numbers into the model as if they were for whole adolescent population. As it is unlikely that the total teenage consumption average would fall bellow $\alpha^{\mathbf{F}}$ or above $\alpha^{\mathbf{M}}$, the predicted outcomes should very likely be located inside the interval created by these scenarios.

Furthermore, I find the minimum of $\alpha^F_{t_{-i+1}}$ (5.87) and maximum of $\alpha^M_{t_{-i+1}}$ (13.09) and use them as scenarios for extreme cases. In every scenario, I assume the change of consumption takes place in t_1 and there are no changes further on.

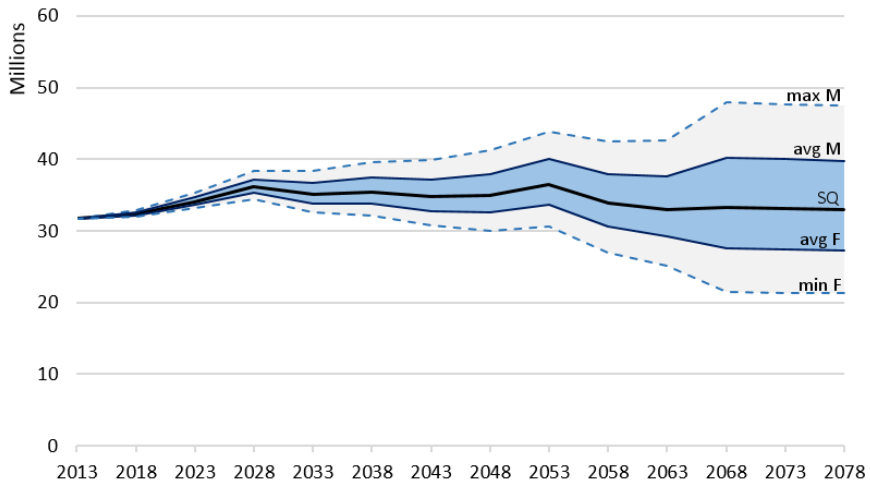

Figure 6.7: Teenage consumption scenarios: Total daily consumption

As the portion of smokers among adolescents is assumed constant at its current rate, there is no point in plotting a chart for total population of smokers because that would equal to Figure 6.1. Predicted consumption is shown in Figure 6.7 and predicted tax revenues in Figure 6.8. The predictions of the model in a numerical form are located in Table 6.6.

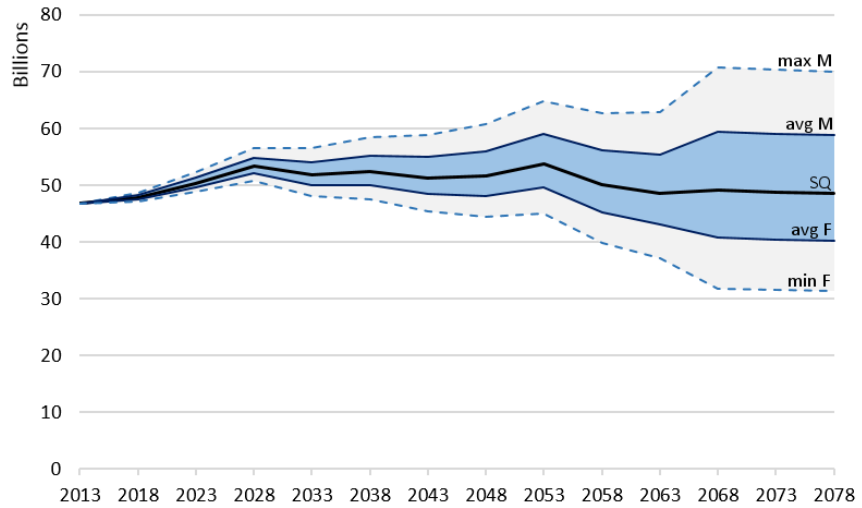

Figure 6.8: Teenage consumption scenarios: Tax revenue (in CZK)

The charts show that in the short run, the change in consumption of adolescents has only minor effect as they account only for a small part of population. However, smokers affected by these changes age and new ones come, so this transforms the population such that the portion of smokers with the new level of consumption grows

steadily with time. Then, in the long run, the scenarios split and their differences grow with time.

		2013	2018	2023	2028	2033	2038	2043	2048	2053	2058	2063	2068	2073	2078
Daily consumption	SQ	31 722 794	32 400 922	34 109 318	36 111 068	35 071 554	35 456 392	34 737 000	34 999 844	36 453 267	33 905 238	32 959 754	33 289 395	33 081 122	32 945 002
	max M	31 722 794	32 931 790	35 386 973	38 351 816	38 314 951	39 632 162	39 874 276	41 230 184	43 868 957	42 506 746	42 647 887	47 957 533	47 650 321	47 448 162
	avg M	31 722 794	32 655 925	34 735 323	37 189 140	36 634 152	37 441 795	37 200 776	37 975 766	39 985 448	37 984 418	37 551 651	40 242 652	39 984 862	39 815 223
	avg F	31 722 794	32 202 995	33 665 408	35 280 194	33 874 524	33 845 529	32 811 274	32 632 482	33 609 287	30 559 408	29 184 367	27 575 933	27 399 284	27 283 041
	min F	31 722 794	31 985 898	33 152 578	34 365 202	32 551 784	32 121 774	30 707 306	30 071 349	30 553 078	27 000 467	25 173 776	21 504 545	21 366 789	21 276 139
Tax revenue	SQ	46 820	47 821	50 342	53 297	51 762	52 330	51 269	51 657	53 802	50 041	48 646	49 132	48 825	48 624
	max M	46 820	48 604	52 228	56 604	56 549	58 494	58 851	60 852	64 747	62 736	62 944	70 781	70 328	70 029
	avg M	46 820	48 197	51 266	54 888	54 069	55 261	54 905	56 049	59 015	56 062	55 423	59 395	59 014	58 764
	avg F	46 820	47 529	49 687	52 070	49 996	49 953	48 426	48 163	49 604	45 103	43 074	40 700	40 439	40 267
	min F	46 820	47 208	48 930	50 720	48 044	47 409	45 321	44 383	45 094	39 850	37 154	31 739	31 535	31 402

* Consumption is expressed in number of cigarettes, tax revenues in millions of CZK

Table 6.6: Teenage consumption scenarios: Model predictions

6.2.5 Combined Scenarios

In the previous forecasts I always focused on one particular parameter of the model. However, it is likely that a decline in the smoking rate of adolescents would be accompanied by a decline in average consumption and vice versa. Now, I will try to combine both approaches and provide a more complex analysis. For simplicity, I will combine the existing scenarios from previous sections (refer there for details) in a manner given by the following table:

Scenario	Smoking rate change	Daily consumption
S1	+10%	13.09
S2	+5%	10.99
S3	-5%	7.53
S4	-10%	5.87

Table 6.7: Combined scenarios

There is no reason to plot the chart for population size of smokers. The scenarios copy those from Section 6.2.3 and thus the chart would look the same as Figure 6.4. Figure 6.9 shows the forecast for consumption and Figure 6.10 for tax revenues. The following holds for both of them: The model predicts a 6.5-25.7% growth in period 2013-2028. A decline towards the initial value is forecast in 2033 at the earliest and

a -10% decline below the initial state as late as in 2048. In an extreme case, the daily consumption can reach the double of 2013 volume in 2063. Complete model output data are reported in Table 6.8.

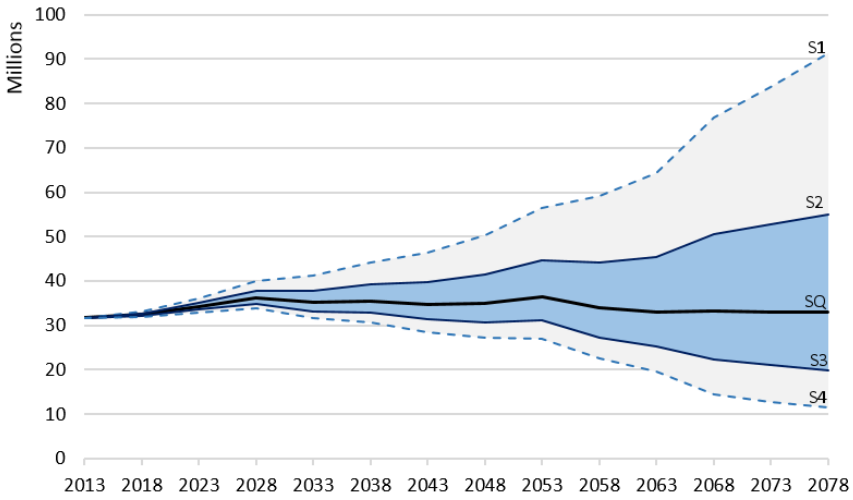

Figure 6.9: Combined scenarios: Total daily consumption

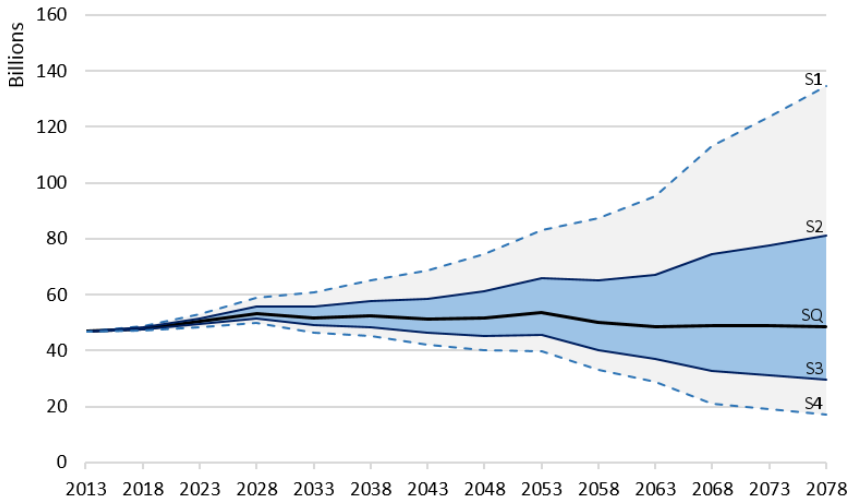

Figure 6.10: Combined scenarios: Tax revenue (in CZK)

Comparison of the graphs to those from previous sections shows that the combination of scenarios increases the pace in which the plotted line grows or falls and that increases the difference between scenarios in the long run.

		2013	2018	2023	2028	2033	2038	2043	2048	2053	2058	2063	2068	2073	2078
Daily consumption	SQ	31 722 794	32 400 922	34 109 318	36 111 068	35 071 554	35 456 392	34 737 000	34 999 844	36 453 267	33 905 238	32 959 754	33 289 395	33 081 122	32 945 002
	S1	31 722 794	33 103 274	36 011 869	39 859 364	41 097 456	44 115 337	46 454 447	50 409 407	56 362 627	59 226 958	64 415 697	76 784 188	83 848 214	91 239 839
	S2	31 722 794	32 727 874	34 993 315	37 801 704	37 747 533	39 207 284	39 750 003	41 474 092	44 665 518	44 134 389	45 408 163	50 533 985	52 719 814	55 077 658
	S3	31 722 794	32 153 693	33 494 365	34 886 807	33 180 699	32 778 485	31 317 174	30 642 192	31 026 380	27 270 844	25 117 434	22 319 086	21 068 203	19 947 331
	S4	31 722 794	31 909 003	32 890 291	33 771 452	31 519 807	30 557 621	28 548 153	27 232 956	26 917 268	22 432 363	19 600 166	14 318 334	12 805 739	11 497 671
Tax revenue	SQ	46 820	47 821	50 342	53 297	51 762	52 330	51 269	51 657	53 802	50 041	48 646	49 132	48 825	48 624
	S1	46 820	48 857	53 150	58 829	60 656	65 110	68 563	74 400	83 186	87 414	95 072	113 327	123 752	134 662
	S2	46 820	48 303	51 647	55 792	55 712	57 866	58 667	61 212	65 922	65 138	67 018	74 584	77 810	81 290
	S3	46 820	47 456	49 435	51 490	48 972	48 378	46 221	45 225	45 792	40 249	37 071	32 941	31 095	29 440
	S4	46 820	47 095	48 543	49 844	46 520	45 100	42 135	40 193	39 727	33 108	28 928	21 133	18 900	16 970

* Consumption is expressed in number of cigarettes, tax revenues in millions of CZK

Table 6.8: Combined scenarios: Model predictions

7. Discussion and Conclusion

This thesis aimed to build a theoretical framework to model the future tobacco consumption, size of smoking population and governmental tax revenues in the Czech Republic. The constituted model had to be adjusted by certain limitations and assumptions (mainly due to the lack of time series data) in order to be applicable to the data by the Czech National Monitoring Centre for Drugs and Drug Addiction. These restrictions did simplify the model, however, the model projections should still be able to capture the future trends induced by upcoming demographic changes to the Czech population and provide approximate forecasts. On the contrary, not all assumptions are artificially imposed because of the lack of data, some are based on empirically justifiable reasoning - a good example is the main assumption of the model, stating that smokers form their tobacco consumption behavior as adolescents.

The model predictions are dependent on the future characteristics of the adolescents - their smoking rate and average daily cigarette consumption. The data from Czech National Monitoring Centre for Drugs and Drug Addiction were used to identify the dependence of tobacco consumption (of each age category) on the prior teenage consumption and provide the resulting coefficients to the model. The same data set was used to supply input for the initial period (2013) of the model. The model inputs (adolescent tobacco consumption characteristics) for the upcoming years have been varied using several scenarios. The resulting forecasts have shown that no matter how large the change in these parameters is, the effect is little in the short run. The reason is that teenagers form only a small portion of the whole population. Nevertheless, in the long run, the effect increases as the portion of population affected by these changes grows and ages.

All of the scenarios predict a growth in the tobacco industry within the next 13 years (until 2028). In particular, the projected number of smokers in 2028 is by

4-8% higher than in 2013, the total daily tobacco consumption and tax revenue by 7-26%. This increase is induced by aging of large birth cohorts.

Interpreting the predictions in the long run is troublesome as the extreme scenarios differ substantially. For example, the total daily consumption can either double or halve in about 50 years. Focusing on the status quo scenario is probably the best way to understand the future development. Within the next 50 years, the perception, policies and behavioral patterns of smoking are very likely to change and debase any predictions. Nevertheless, it is interesting to see the long run outlook if the current state were to continue unchanged. The status quo scenario predicts an increase in number of smokers from 2.4 in 2013 to 2.7 millions in 2028, later on fluctuating around 2.6 millions with small deviation in 2048-2053 with a peak of 2.8 millions in 2053. Similar pattern is observed for daily tobacco consumption and tax revenues. A boost up to 36.1 millions of cigarettes in 2028 from current 31.8 millions and fluctuation around 33-34 millions with a deviation to 36.5 millions in 2053. Tax revenues are expected to grow towards 53.3 billions of CZK in 2028 and further move around 50-52 billions, compared to 46.8 in 2013.

Clearly, the question emerging from these outcomes is: Does this prospect of an increase in tobacco prevalence and consumption within the next decade call for a policy response or any taxation change? The Czech government currently discusses a new anti-tobacco policy. This law, which is in the works at the Ministry of Health, should ban all smoking in restaurants and bars. This would certainly influence the forthcoming development and the effect of this policy could partially offset the upcoming boom. However, possible trade-off is caused by the decrease in future tax revenues. Kvaček (2011) found the tobacco demand of Czech smokers to be quite inelastic, thus, higher taxation on tobacco may be suitable as an addition to this policy and can provide some funds to cover this opportunity cost.

Possible model improvements and further research are strongly depending on an availability of time series data with consistent structure and necessary detail. An

ideal approach would be to combine the methods of Mendez et al. (1998) and the age-cohort model. Such analysis would be able to estimate the cohort effects, improve the robustness of model coefficients and model the smoking behavior much better (with cessation, higher probability of death for smokers, etc.). Altogether, it would result in a model with higher predictive power.

Bibliography

Becker, G. S. and K. M. Murphy (1988). A Theory of Rational Addiction. *The Journal of Political Economy*, 675–700.

Cameron, A. C. and P. K. Trivedi (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics 1*(1), 29–53.

Carstensen, B. (2007). Age–period–cohort models for the Lexis diagram. *Statistics in Medicine 26*(15), 3018–3045.

Doll, R., R. Peto, J. Boreham, and I. Sutherland (2004). Mortality in relation to smoking: 50 years' observations on male british doctors. *Bmj 328*(7455), 1519.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Kerr, W. C., T. K. Greenfield, J. Bond, Y. Ye, and J. Rehm (2004). Age, period and cohort influences on beer, wine and spirits consumption trends in the US National Alcohol Survey. *Addiction 99*(9), 1111–1120.

Kralikova, E., A. Kmetova, K. Zvolska, M. Blaha, and Z. Bortlicek (2013). Czech adolescent smokers: unhappy to smoke but unable to quit. *The International Journal of Tuberculosis and Lung Disease 17*(6), 842–846.

Kvaček, J. (2011). Poptávka po cigaretách v české republice a výnosnost spotřební daně z cigaret. Master's thesis, IES FSV UK.

Levy, D. T., H. Ross, A. Kmetova, E. Kralikova, M. Stoklosa, K. Blackman, O. A. Diab, E. M. Abdelrahim, M. Esmail, J. F. Golding, et al. (1997). The Czech

Republic SimSmoke: The Effect of Tobacco Control Policies on Smoking Prevalence and Smoking Attributable Deaths in the Czech Republic. *J Prev Med Public Health 30*(4), 697–707.

Mason, K. O., W. M. Mason, H. H. Winsborough, and W. K. Poole (1973). Some Methodological Issues in Cohort Analysis of Archival Data. *American Sociological Review*, 242–258.

Mendez, D., K. E. Warner, and P. N. Courant (1998). Has Smoking Cessation Ceased? Expected Trends in the Prevalence of Smoking in the United States. *American Journal of Epidemiology 148*(3), 249–258.

Mori, H., D. L. Clason, et al. (2004). A Cohort Approach for Predicting Future Eating Habits: The Case of At-Home Consumption of Fresh Fish and Meat in an Aging Japanese Society. *International Food and Agribusiness Management Review 7*(1), 22–41.

OECD (2014). OECD Factbook 2014. *http://dx.doi.org/10.1787/factbook-2014-en*.

Oh, D. L., J. E. Heck, C. Dresler, S. Allwright, M. Haglund, S. S. Del Mazo, E. Kralikova, I. Stucker, E. Tamang, E. R. Gritz, et al. (2010). Determinants of smoking initiation among women in five European countries: a cross-sectional survey. *BMC Public Health 10*(1), 74.

Pertold, F. (2009). Sorting into Secondary Education and Peer Effects in Youth Smoking. *CERGE-EI Working Paper Series* (399).

Raftery, A. E., N. Li, H. Ševčíková, P. Gerland, and G. K. Heilig (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*.

Rentz, J. O. and F. D. Reynolds (1991). Forecasting the Effects of an Aging Population on Product Consumption: An Age-Period-Cohort Framework. *Journal of Marketing Research*, 355–360.

Sovinová, H., L. Csémy, B. Procházka, and S. Kottnauerová (2008). Smoking-attributable mortality in the Czech Republic. *Journal of Public Health 16*(1), 37–42.

Sovinová, H., L. Csémy, and P. Sadílek (2014). Užívání tabáku v české republice 2013. *Státní zdravotní ústav*.

Spilková, J., D. Dzúrová, and H. Pikhart (2011). Inequalities in smoking in the Czech Republic: Societal or individual effects? *Health & Place 17*(1), 215–221.

Tesař, T. (2011). Comparative analysis of factors influencing children's smoking. Bachelor's thesis, IES FSV UK.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica: Journal of the Econometric Society*, 307–333.

WHO (2014). WHO Database. *data.euro.who.int*.

Appendix

	Frequency	Percent	Info
msize1	517	34.19	below 5 000 inhabitants
msize2	290	19.18	5 001-20 000 inhabitants
msize3	161	10.65	20 001-50 000
msize4	154	10.19	50 001-100 000
msize5	171	11.31	over 100 000 except Prague
msize6	219	14.48	Prague

Table 7.11: Municipality size

Variable	Obs	Mean	Std. Dev.	Min	Max
cigsdaily	1512	3.622	7.189	0	80
smokes	1512	.282	.45	0	1
alcabuse	1512	.326	.469	0	1
male	1512	.488	.5	0	1
age	1512	38.595	14.041	15	64
b1	1512	.064	.244	0	1
b2	1512	.129	.335	0	1
b3	1512	.12	.325	0	1
b4	1512	.157	.364	0	1
b5	1512	.11	.314	0	1
b6	1512	.083	.275	0	1
b7	1512	.071	.257	0	1
b8	1512	.073	.261	0	1
b9	1512	.084	.276	0	1
b10	1512	.11	.314	0	1
unemp	1512	.048	.213	0	1
student	1512	.122	.327	0	1
maternity	1512	.051	.22	0	1
retired	1512	.095	.293	0	1
disabled	1512	.024	.153	0	1
inc1	1512	.103	.304	0	1
inc2	1512	.048	.213	0	1
inc3	1512	.164	.37	0	1
inc4	1512	.253	.435	0	1
inc5	1512	.238	.426	0	1
inc6	1512	.138	.345	0	1
inc7	1512	.056	.229	0	1
msize1	1512	.342	.475	0	1
msize2	1512	.192	.394	0	1
msize3	1512	.106	.309	0	1
msize4	1512	.102	.303	0	1
msize5	1512	.113	.317	0	1
msize6	1512	.145	.352	0	1
edu1	1512	.088	.283	0	1
edu2	1512	.3	.459	0	1
edu3	1512	.39	.488	0	1
edu4	1512	.045	.207	0	1
edu5	1512	.177	.382	0	1

Table 7.1: Summary statistics

alcabuse	Frequency	Percent
0	1,019	67.39
1	493	32.61

Table 7.2: Tabulate: alcabuse

male	Frequency	Percent
0	774	51.19
1	738	48.81

Table 7.3: Tabulate: being male

unemp	Frequency	Percent
0	1,440	95.24
1	72	4.76

Table 7.4: Tabulate: being unemployed

student	Frequency	Percent
0	1,328	87.83
1	184	12.17

Table 7.5: Tabulate: being a student

maternity	Frequency	Percent
0	1,435	94.91
1	77	5.09

Table 7.6: Tabulate: being on a maternity leave

retired	Frequency	Percent
0	1,369	90.54
1	143	9.46

Table 7.7: Tabulate: being retired

disabled	Frequency	Percent
0	1,476	97.62
1	36	2.38

Table 7.8: Tabulate: being disabled

smokes	Frequency	Percent
0	1,085	71.76
1	427	28.24

Table 7.9: Tabulate: smokes

Variable	**Coefficient**	**(Std. Err.)**
alcabuse	-0.897**	(0.124)
age	0.017**	(0.004)
unemp	-0.641*	(0.261)
student	1.183**	(0.211)
msize2	-0.222	(0.170)
msize3	-0.279	(0.219)
msize4	-0.564**	(0.213)
msize5	-0.585**	(0.204)
msize6	-0.557**	(0.197)
educ2	0.346†	(0.191)
educ3	0.894**	(0.180)
educ4	1.352**	(0.353)
educ5	1.807**	(0.243)

Significance levels : † : 10% * : 5% ** : 1%

Table 7.10: Inflate (logit)

	Frequency	Percent	Info
educ1	133	8.80	primary education
educ2	454	30.03	secondary without 'maturita'*
educ3	589	38.96	secondary with 'maturita'
educ4	68	4.50	follow-up study (VOŠ)
educ5	268	17.72	university

*Maturita is the high-school exit exam in the Czech Republic

Table 7.12: Education

	Frequency	Percent	Info
inc1	156	10.32	no own income
inc2	72	4.76	below 5 000 Kč
inc3	248	16.40	5 001-10 000 Kč
inc4	383	25.33	10 001-15 000 Kč
inc5	360	23.81	15 001-20 000 Kč
inc6	209	13.822	001-30 000 Kč
inc7	84	5.56	over 30 001 Kč

Table 7.13: Income

Variable	Coefficient	(Std. Err.)
alcabuse	0.197**	(0.059)
male	0.351**	(0.065)
unemp	0.118	(0.121)
student	-0.151	(0.181)
maternity	0.326*	(0.149)
retired	0.237	(0.163)
disabled	0.141	(0.177)
inc2	0.058	(0.165)
inc3	0.058	(0.149)
inc4	0.118	(0.154)
inc5	0.076	(0.155)
inc6	0.066	(0.168)
inc7	0.439*	(0.198)
msize2	-0.066	(0.076)
msize3	-0.145	(0.104)
msize5	0.190*	(0.088)
msize6	0.003	(0.090)
edu2	-0.289**	(0.099)
edu3	-0.362**	(0.104)
edu4	-0.426*	(0.198)
edu5	-0.536**	(0.142)
b2	0.082	(0.178)
b3	0.297	(0.199)
b4	0.231	(0.201)
b5	0.218	(0.204)
b6	0.240	(0.210)
b7	0.446*	(0.212)
b8	0.576**	(0.206)
b9	0.485*	(0.210)
b10	0.261	(0.241)
Intercept	2.104**	(0.215)
Inflate (logit): Table 7.10 in Appendices		
lnalpha	-1.494**	(0.102)

Significance levels : † : 10% * : 5% ** : 1%

Table 7.14: Estimation results : ZINB - all variables