LUND UNIVERSITY
School of Economics and Management

Cannabis on a legalized market

- An evaluation of Washington and Colorado’s cannabis retail markets

Bachelor thesis in economics April 2016

Author: Alexander Morling
Supervisor: Andreas Bergh

Abstract

The aim is to specify a demand function that describes the demand for cannabis on county level in the states that have legalized retail sales in order to test hypotheses regarding what factors influence demand on county and potentially on individual level, as well as to gain some insight in whether legalization is a good strategy for combatting black markets of cannabis. Multiple regression analysis is used to estimate the specification with use of 48 observations. 9 hypotheses are tried in order to see what explanatory variables could help explain sales revenue. The main conclusions is that retail sales is positively significantly explained by the county characteristics of tourism, education, size of the population, share of smokers, share of population between 18-65, and negatively by the share of the population that consumes alcohol to a large extent. Sales are highly explained by these characteristics on county level but it is harder to tell if the effects on individual level are representative. However, individual level research supports the outcome of the results on county level. In the short run legalization supplies a small segment of the cannabis market, and time will tell if and by how much retail prices decrease in order to be competitive with the black and grey markets. The specification has some clear issues that should be dealt with before drawing scientific conclusions of the results.

Table of Contents

1. Introduction

1.1 Background

At present date 23 US states have a system of legalized cannabis for medical purposes. Five states and the district of Colombia have so far legislated retail sales of cannabis. The states Washington and Colorado have since 2014 had retail stores supplying cannabis (Cournoyer, 2016). There are still many states that are debating the advantages and disadvantages of legalizing marijuana and an important measure in the debate is the possibility of evading black market suppliers to prevent revenue for organized crime and to restrict access for minors (Subritzky, Pettigrew & Lenton, 2016, p. 7). The most equivalent example of a similar policy measure is Uruguayôs legalization of cannabis, where they erased the taxes on cannabis in order for the legal market to be competitive with the black market after realizing that the markup gets too high with taxation (Castaldi 2014). Previous studies on the subject have mainly focused on number of users and the quantity of the product demanded and little on the characteristics of the specific segments of customers of retail cannabis stores. Although extensive research have been conducted on the subject on users of cannabis and other drugs at large under normal circumstances of the substance being illegal.

The retail and medical cannabis industry in the United States had revenues of 1.9 billion dollars 2015 and the industry has an expected growth rate of 31.4% annually the next 5 years (Diment, 2015, p.4). An evaluation of the county characteristics of successful counties regarding cannabis retail sales is therefore of great interest for investors who are looking to make informed decisions before entering the industry. For any industry it is of interest to learn more about the individual consumers and what characteristics they possess for aiming proper marketing efforts.

The states Colorado and Washington have passed law reforms allowing production and sales of cannabis, and both states have since 2014 had retail stores that have sold the product (Huddleston, 2016). Because of the newly opened industry there are now available figures for sales at county level in Colorado, as long as one actor is not controlling 80% of the market or there are less than 3 actors, according to Colorado department of revenue (2015). In Washington there are retail and medicinal sales figures at county level for every county that have retail cannabis stores.

1.2 Purpose

The purpose of this thesis is to specify an equation that can describe demand of cannabis on county level in the states of Colorado and Washington. The first year sales of cannabis per county in the states will be used as the dependent variable explained by theoretically motivated independent variables.

A demand function can be useful for budgetary departments of states debating legalization, investors, and enabling marketers to better find the markets containing main segments. The work also consists of testing hypotheses regarding the county- as well as individual consumer characteristics, and the implications for a state to legalize cannabis. There are many stakeholders that can benefit from having a function that can forecast sales in a particular area as well as an evaluation of the societal characteristics of the users of legal cannabis.

- What are the characteristics of consumers of retail cannabis?
- What are the characteristics of well performing retail cannabis counties?
- Is legalization a way of combatting black market supply of cannabis?

1.3 Delimitation

Since data is only available for two states and all counties have not been included, mainly due to local tax laws in Colorado, there are only 48 observations in the regression. The data that are used for this work is sales figures that is available after the states legalized the entire supply chain following a change in the general perception of what cannabis is, the result should for that reason not be interpreted as general results for the complete population of cannabis users, or users in states where cannabis is illegal but rather help describe the characteristics of societies that allows retail and medicinal cannabis as a mainstream product. The main delimitation is therefore that the paper is examining the factors that drive the demand for the retail cannabis market and not the factors for cannabis demand in general. The market for medical and retail cannabis is brand new and is clearly not in equilibrium. Due to the time restraints, the model is not as improved as it could have been.

1.4 Previous studies

Previous studies covering the aim of evaluating how well legalization combats the black market for cannabis are for obvious reasons slim considering full legalization such as in the states covered in this thesis are very few. I am not aware of any published study that evaluates the characteristics of typical consumers of retail cannabis. There have been works published but they assume policy outcomes based on opinions rather than on empirical studies (Subritzky, Pettigrew & Lenton, 2016, p. 1). Uruguay is a recent real world example in which a state has legalized cannabis under similar circumstances as these, where they changed the policy to completely eradicate taxation of cannabis, in order for the legal market to be competitive with the criminal black market (Castaldi, 2014).

Similar are the questions regarding the characteristics of retail cannabis consumers which have not been evaluated under circumstances equivalent to these and it is a topic that is hard to compare between states and continents, such as The Netherlands and Uruguay, which have similar but not equivalent policies. This is a specific topic that has not either been covered that extensively for the same reasons. Previous studies on the characteristics of consumers of cannabis rather focus on cannabis users at large and not on the specific segment that would not consume cannabis when it is illegal but does not mind consumption when it is legal.

The notion of rational addiction does however provide some insight into the matter regarding who may be a consumer if a good has been legalized. The notion of rational addiction assumes that addictive goods can be utility maximized just as any other good, which is provided support for by empirical evidence for a decrease in demand for cigarettes when a future tax increase of tobacco have been announced. This result is showing that consumers behave rationally when making a consumption decision even for addictive goods (Gruber & Köszegi, 2000, p.37). Social control has a part of the consumption decision for different segments of cannabis users. The lower income users who does not see a point in concealing there use of cannabis from employers and family will turn to any dealer, while higher income users with a reputation they have to uphold, will not turn to an illegal market and risking his job, police encounters or being threatened by criminals, for the sake of cannabis (Muscat et al, 2009 p.22).

A study by *Light M. et al. (2014)* was conducted for the state of Colorado after the first six months of legalization 2014, and it reflects on the pricing structure of the market, alternative suppliers and visitor demand in the state of Colorado. Demand of cannabis is estimated for retail vendors as well as the black and grey market demand. There has also been done a similar study by *Kilmer et al (2013)*, for the state of Washington before legalization. The paper prepared for the state of Washington was done before they opened sales, and contains no reflections concerning the particular segments of users of specifically retail cannabis, but rather quantitatively estimates the total market in Washington, not particularly estimating the retail demand.

The main difference between this paper and the above mentioned is that in the articles, the method has been based on using survey data on user prevalence with the main goal on finding quantity demanded, as well as descriptive statistics of characteristics of cannabis users. While this paper examines demand using multiple regression analysis, using demographic and surveys. The work for Colorado has been conducted only six months after sales started, with a brand new market that has not yet settled.

1.5 Disposition

The first chapter introduces the background and the previous studies on this subject. The second chapter will present the theories underlying the hypotheses. The following chapter deals with how the specification was estimated, and the last chapter discusses the results and concludes the questions aimed at with this work.

2. Theory

This chapter will present the theories used in this paper to help explain the sales of retail cannabis, and some insight about how the Colorado market is organized, which is assumed to be representative for the Washington market as well.

2.1 Education

In the heavy user segment of users at large, close to 20% have a college degree or higher educational attainment. Slightly more than 30% of past month users have the same level of education (Kilmer et al. 2013, p.32). Or this could be expressed as 80 % of heavy users at large are estimated to have less than three years college education.

2.2 Smokers of tobacco

There is a strong correlation between tobacco usage and marijuana use. 80% of the heavy users in their surveys also reported tobacco use. This is in the article compared to the general population user rate of tobacco which is 25-28% (Kilmer et al. 2013, p.37).

In addition to the descriptive data in the article by Kilmer et al, a report by the National bureau of economic research finds that an increase in the price of cigarettes decreases the level of cannabis consumed by current users, indicating that the goods are complementary. The research has used data for high school students between the years 1992-1994 (Chaloupka et al,1999, p.16-17).

2.3 Age distribution

Around 80% of heavy users of the national age distribution are between the ages 18-50 years old of all segments of cannabis users (Kilmer et al. 2013, p.33). In Washington and Colorado the state laws does not allow people under the age of 21 to purchase marijuana in retail stores.

2.4 Tourism

When a state allows a product that is highly demanded yet illegal in neighboring states, intuition says there will be a lot of tourism wanting to test out the product or even travelling with the main purpose of taking part of the legalized good. This notion is backed by an article stating that descriptive data from hotels.com and hopper.com supports this notion (Briggs, 2014).

The authors behind the Market size and demand study for the government of Colorado suggest that close to 50% of sales stem from visitors of the state, according to point-of-sale data for the first six months of 2014. The estimate is derived from tax-receipts, point of sale statistics and data from tourist offices (Light et al. 2014, p.3).

An article reports that at major travel and room agencies, searches for rooms and trips to Washington and Colorado increased after the states opened the sales of cannabis. During the 4/20 cannabis festival in Denver it was a 73% increase in searches for rooms, and a 17% increase in room searches the first six month in all of Colorado, compared to the same period the previous year. In Washington, there was a 68% increase in room searches for Seattle the first legal month, compared to the same period the previous year (Briggs 2014).

2.5 Alcohol consumption

In a report, it is concluded that alcohol and cannabis should be seen as substitute products. In the study estimated the effects of an increase of the legal drinking age with the result that it decreased alcohol consumption at the expense of increased marijuana consumption, the two goods are therefore argued to behave like substitute goods (DiNardo & Lemieux 1992, p.43).

A similar result holds for a similar study, where it is concluded that marijuana usage decreases at the age of 21, which is the legal drinking age in many states (Crost & Guerrero 2012, p.1).

2.6 Prices

As with any demand function of a good, the price of the product likely influences the demand. In Colorado there are competing suppliers to the retail shops. In addition to medical and retail dispensaries for marijuana in Colorado, there are two types of suppliers of marijuana that supplies the market. Mainly, any adult in can grow up to 6 plants for own use. In addition, the caregiver program allows adult residents to be a caregiver for up to 5 people that gives the person the right to produce 6 medical marijuana plants per person that the individual is a caregiver for at a 2.9% tax rate. The caregiver can also apply for a waiver that gives the caregiver the right to be a caregiver for more than 5 patients. The size of this portion of the market was in May 2014 estimated to be 43.9%, imputed by the January 2014 sales of Denver and Jefferson counties. The sales tax on medical marijuana was 2.9% compared to 25% plus any local taxes for retail stores (Light et al. 2014, p.26-27).

2.7 Hypotheses for this paper

The characteristics of the legalized retail market for cannabis are likely to shape its customer base. The prices are likely to influence the consumer behavior to a large extent and it is for that reason needed to consider what sort of price structure currently exists when assuming hypotheses for the different variables. Because of the higher costs of having production facilities, stores and logistic costs, the markup will be higher than on the alternative markets where the customers do not expect much more than being handed the product they want to purchase. Since it currently exist a large market that are considered as black or grey, the customers of retail cannabis will be individuals who can afford to pay the higher price.

2.7.1 Expected signs of the coefficients

Education is for that reason assumed to have a positive sign even though educated cannabis consumers are relatively scares considering the estimated portion of heavy users who are educated on the national level. Tobacco is expected to have a positive sign as well, considering the high degree of cannabis users who also reported tobacco use. According to studies tobacco use and cannabis are substitutes and should for that reason too be positively related to sales of cannabis. The variables for age that are included are assumed to be negatively related to sales since the largest part of users are between the age 18 and 65. Tourism and the size of the population are assumed to be positively related as well. It could be that typical tourists are not that interested in consuming cannabis, but a lot speaks for tourists being a large market for retail cannabis. Alcohol will be expected to have a negative coefficient since the products behaves like substitutes. The portion of a county that consumes alcohol to a large extent is assumed to prefer that over cannabis. Even though the data to be used are estimates from earlier years than actual legalization it is likely that it reflects a portion of the population that previously has preferred cannabis in general.

2.7.2 The specification of the model

Retail cannabis is expected to be demanded by the capital intensive portion of the market. The size of the population and the number of tourists are expected to have positive coefficients. The proportion of the population that is under 18, over 65 and that are heavy users of alcohol are expected to have negative signs. The proportion of the population that are tobacco users and that are educated, are expected to have positive slope coefficients.

$$RCB = ɓ0 + ɓ1pop + ɓ2tour - ɓ3<18 - ɓ4drinking - ɓ5>65 + ɓ6smoke + ɓ7edu + ʿ$$

Where

RCB = Revenue of retail cannabis

ɓ0 = Intercept

Pop = Size of population

Tour = Number of tourists

<18 = Share of population under 18

Drinking = Share of population who drinks

>65 = Share of population above 65

Smoke = Share of population who are smokers

Edu = Share of population who has a college degree or higher

ʿ = Stochastic error term

3. Method

The data for the variables have been collected at various sources, all of which are highly trustable and reference governmental bodies of the United States on both state and national level. This chapter will describe what the data used measures, as well as how raw data have been remodeled to be comparable, and to measure the variables hypothesized in this paper. I will present the method used to find the best regressions.

3.1 Data per variable

3.1.1 Revenue of cannabis sales, RCB

The dependent variable of this model is the revenue generated through retail and medical marijuana stores per county in the states Colorado and Washington the 12 month period July 2014 ï end of June 2015. The Colorado data has been reported on monthly basis, and for months where a county have not reported sales, the value have been imputed by the average of all reported months for that county. The data have been collected from Colorado department of revenue, and Washington state liquor and cannabis control board. The data for the counties in Colorado have been reported as tax revenue for the state and not as retail sales. These numbers have for that reason been divided by the appropriate tax rate to get the revenue incurred.

3.1.2 Population

The variable population per county is an estimate of the population in 2014. It is a numeric number giving the estimate of the number of people, collected from the US census bureau.

3.1.3 Population under 18

The population under the age of eighteen is a percentage estimate from US census bureau for 2014.

3.1.4 Population above 65

This variable is an estimate of the portion of the population that is sixty five years old or older 2014. The data is taken from the US census bureau.

3.1.5 Population who has a college degree or higher

The variable describes the percentage of the county population 25 or older, who has a bachelorôs degree or higher education. The estimate is for the years 2009-2013, and the data is collected from the US census bureau.

3.1.6 Population who drinks

This variable is a percentage of the adult population per county who reported heavy or binge drinking, defined as drinking five or more units of alcohol on one occasion the last 30 days. The data is for the years 2006-2012 from the county health rankings organization.

3.1.7 Population who are smokers

Measure of the percentage of the adult population per county who is a smoker of tobacco between the years 2006-2012 collected from the county health rankings organization.

3.1.8 Tourism

The estimation for number of overnight tourists to each county in Colorado is found by travel expenditures per county 2014 divided by the average spending among visitors. The estimate is provided in the report *Colorado travel impacts 2009-2014p,* prepared for the state of Colorado (Dean Runyan Associates 2015a). The data from Washington is provided in the report *Washington state county travel impacts & visitor volume 1991-2014p* (Dean Runyan Associates 2015b). This data has been processed the same way as that for Colorado.

3.1.9 Per capita income

Per capita income is a measure of the income the past 12 month between the years 2009-2013, from us census bureau.

3.1.10 Arrests

The data is published by Puzzanchera, C. & Kang, W, 2014. The responsible authors represent a governmental organization that publishes FBI statistics. The data available is total drug related arrests per county, which has been discounted by the nationwide proportion of marijuana related arrests.

3.2 Estimation method

This subchapter will describe the model used and how the best equation was specified, which is determined by adjusted R2, how significant the variables are, how robust the equations are to changes in specification and statistical invalidity have been excluded as a source of coefficient results. I will present the tools and the theories underlying the analysis.

3.2.1 Ordinary Least Squares

Ordinary least squares will be used to estimate the best specification presented in this paper. The statistical model is minimizing the squared sum of the residuals (Studenmund, 2013, p.37). In other words it estimates the coefficients to minimize the distance between the regression line and each of the sample observations, it is therefore the most accurate model as long as the classical assumptions are met (Studenmund, 2013, p.97).

R2 is a measure of how well the variation in Y is explained by the estimated equation. It is a ratio between the variation that can be explained by the model and the variation that is unexplained. R2 lies in the interval of 0 and 1 where 1 indicates a perfect degree of explanation and 0 a complete failure of the model to explain the variation (Studenmund, 2013, p.51). R2 adjusted is the R2 value corrected for the number of explanatory variables included in the equation. R2 never decreases with the inclusion of an additional variable and is therefore adjusted for the degrees of freedom. If the improvement of fit outweighs the loss of degrees of freedom, adjusted R2 will increase (Studenmund, 2013, p.55-56).

The simplified version of the t-statistic formula consists of the variables estimated coefficient in the numerator and the standard error of the variables coefficient in the denominator (Studenmund, 2013, p.135). The p-value can be used instead of the t-test and gives the probability of observing a t-score of given size, or larger under the assumption that the null hypothesis is true. For the level of significance used in this paper, of 5%, the null hypothesis is rejected if the p-value is smaller than 0.05 (Studenmund, 2013, p.141-142).

3.2.2 Specification search

A commonly accepted practice when specifying an equation is called sequential specification search and it basically means that the inclusion or remittance of one variable at the time will be done. The output will then carefully be analyzed, in particular the R2 adjusted, the coefficients and the variance. By doing a sequential specification search, it is assumed that the true equation will also be consistent with the best equation estimated (Studenmund, 2013, p.192-193). This is the method used to specify the equation in this paper and an important part to justify the results would be to rerun the equations on a different data set.

3.2.3 The Gauss-Markov theorem

In order for OLS to be the minimum variance linear unbiased estimator, the classical assumptions must hold. If one of them does not, other techniques may be of greater advantage. The classical assumptions can be summarized as a model that is linear, has the correct functional form, has no omitted variables, homoscedastic residuals, no multicollinearity, has uncorrelated error terms, and have an error term that has a zero mean. The assumption that the mean of the error term must be zero is met by the inclusion of a constant term. Correlated error terms or autocorrelation is mainly an issue for time-series models, and not for cross-sectional data models such as the one specified in this paper (Studenmund, 2013, p.98-104).

When searching for the best specification it is necessary to be aware of these potential problems. Outlined below are the methods and statistical theories used when searching for the best specification and evaluating how reliable the best equation really is.

3.2.4 Multicollinearity

The presence of severe multicollinearity in a model means that one or more of its independent variables are correlated to each other. If the simple correlation coefficient between two variables is larger than .8, it is common practice to view the multicollinearity as severe. In addition to examining the simple correlation coefficients between the variables, it is also common to measure the variance inflation factors in a model. VIF is a measure of the extent to which the explanatory variables in an equation is correlated to each other, while the simple correlation coefficient only measures the extent to which one variable is correlated to one other variable. A VIF higher than 5 is an indicator of severe multicollinearity.

The main consequence of severe multicollinearity is that the standard errors increase, thus reducing the t-scores. When a model suffers from severe multicollinearity, it may be difficult to get significant variables and it is likely that variables will get unexpected signs. Even though there may be no significant variables, the adjusted R2 is usually unaffected even when there is severe multicollinearity (Studenmund 2013, p.266-275).

A typical measure used to detect multicollinearity is to look at the simple correlation coefficients. When doing the testing for this paper, a correlation matrix was used mainly as help when including variables to the regression, which is included as table 3.1. This was a very useful tool to efficiently see if multicollinearity could have infringed any effects in the output.

Table 3.1

	nr.				**Excessive**		**BA or**
	tourists	**pop 2014**	**< 18 years**	**cigarette use**	**drinking**	**>65 years**	**higher**
nr. tourists	1						
pop 2014	0,22	1					
< 18 years	0,02	0,19	1				
cigarette use	-0,01	-0,09	0,24	1			
Excessive drinking	0,22	-0,06	-0,42	-0,27	1		
>65 years	-0,25	-0,32	-0,45	0,27	-0,33	1	
BA or higher	0,23	0,18	-0,52	-0,57	0,62	-0,34	1

3.2.5 Omitted/irrelevant variable

The classical assumption that the error term is independent of the explanatory variables is violated by exclusion of a relevant variable. The bias is a function of the omitted variables coefficient times the correlation between the included and omitted variable. The direction of bias can therefore be estimated by the formula (Studenmund, 2013, p.179-181):

Bias = $\boldsymbol{\beta}$om$f$($r$in,om)

The main things to consider in addition to being aware of the bias are theoretical validity, if the added variable is significant with the hypothesized sign and if the adjusted R2 increase. When including an omitted variable, the variance of the previously included variables coefficients will increase. This is in contrast to including an irrelevant variable which would decrease adjusted R2 and increase the variance, thus lowering the t-scores (Studenmund, 2013, p.186-188).

3.2.6 Residual plots

To meet the classical assumptions of using the correct functional form and the model to be linear, a plot of the residuals can be examined whether they are distributed with a constant variance, which is one of the classical assumptions as well.

The residual is the difference between the actually observed data and the predicted data for each observation. When plotting every observation residual against the X-axis, the pattern of that graph can be examined for constant variance. A normal pattern would be a horizontally distributed clutter around the mean with standardized residuals between -2 and 2.

If the residual plot for instance exhibits clear patterns such as a U-shape, funnel or a data point that is much larger than the rest of the residuals, the fit of the model could be improved. A funnel shaped scatterplot indicate heteroscedasticity, a U-shape indicate that the relationship between the dependent and independent variable are non-linearly related to each other and a data point that has a high residual value indicates that there is an outlying data point in Y, which could mean that it is a data error or that the observation is extreme and may bias the coefficients.

If a variable has a residual plot that has the residuals plotted on the lower values of the X-axis and have one or a few residuals that are on the larger part of the axis, this suggests that these extreme observations may be influential. An X-axis with large values is not per definition problematic, but could be an indicator that the coefficients are biased.

When fixing these issues the variable could be transformed by taking the log, making the variable quadratic, or include a dummy variable. When a variable that exhibits non-normal patterns is fixed, the fit and the reliability of the model will improve (Statwing, 2016).

3.2.7 Sensitivity analysis

In order to get a sense of how good a model is sensitivity analysis is a common form of testing how robust an equation is. By running regressions with changes to the specification, variables, subsets and functional forms, the results are determined by how well the results hold up to these changes. When sensitivity analysis is conducted, results that appear statistically significant in some specifications and statistically insignificant in other functional forms tend to be viewed as less significant, even though it looks good in a particular specification (Studenmund, 2013, p.194).

4. Result

In this chapter I will present the outputs from the OLS regressions, as well as explain how the data outputs has been analyzed in accordance with empirically accepted interpretation of statistical output in order to motivate the quality of the best regression model estimated.

Table 4.1

	Regression 1	**Regression 2**	**Regression 3**	**Regression 4**
pop 2014	51,69***	24,39***	25,68***	25,41***
nr. tourists		28,82***	28,79***	29,34***
Persons under 18 years percent 2014			-53000000	-95000000*
Excessive drinking 2006-2012 18+ %				-95000000*
Persons 65 years and over percent 2014				
cigarette use 2009-2013 18+ %				
BA or higher 25+ 2009-2013				
Adjusted R2	0,1096	0,908	0,907	0,9096

Table 4.2

	Regression 5	**Regression 6**	**Regression 7**	**Regression 8**
pop 2014	23,42***	23,70***	21,47***	20,70***
nr. tourists	29,23***	28,98***	28,72***	28,65***
Persons under 18 years percent 2014	-150000000**	-210000000**	-112000000	
Excessive drinking 2006-2012 18+ %	-140000000**	-156000000**	-175000000**	-145000000**
Persons 65 years and over percent 2014	-600000000	-107000000*	-56000000	
cigarette use 2009-2013 18+ %		108000000*	144000000**	141000000**
BA or higher 25+ 2009-2013			46000000*	66000000***
Adjusted R2	0,9094	0,913	0,915	0,917

***** significant on 1% level**

****significant on 5% level**

***significant on 10% level**

4.1 regression results

Regression 1: $RCB = \beta_0 + \beta_1 pop + \epsilon$

The main point of interest in this regression run with solely the population as the independent variable is that the variable is statistically significant at less than 1%. That is; the risk of falsely accepting a false null hypothesis is less than 1%. The adjusted R2 is very low, which is expected as the theoretical arguments to include a number of additional variables are very strong.

Regression 2: $RCB = \beta_0 + \beta_1 pop + \beta_2 tour + \epsilon$

As the adjusted R2 have increased to 90.8 %, the variable is clearly an important variable in the model. The p-value for population is even more significant than in the previous regression, and the p-value for the new variable is also significant at a very high confidence level. The coefficient for population changes from 51.69 to 24.39 which is a major change of the coefficient. All of the points mentioned above are indicators suggesting that the new variable belong in the equation.

An additional point to make is that the negative change of the population coefficient is expected if tourism was an omitted variable in the first equation, since the expected bias would be positive. The population coefficient therefore changed negatively when the omitted variable was included.

Regression 3: $RCB = \beta_0 + \beta_1 pop + \beta_2 tour - \beta_3 {<}18 + \epsilon$

Theoretically the percentage of people under the age of 18 in a county should be valid in explaining the variance in the sales of cannabis. The adjusted R2 decreased, the p-value is very insignificant for the new variable. The coefficients changed in the opposite direction of previous bias due to an omitted variable, though really small changes. The standard errors are practically constant and the t-scores are slightly decreased.

The little change in the previous included variables and the decreased adjusted R2 are typical indicators of an irrelevant variable. The theoretical argument to include the variable is very strong and the sign is in the expected direction and the unadjusted R2 increased, typically it would remain constant if the variable was irrelevant. The variable will for these reasons not be excluded at this point.

A likely reason for the output is current omitted variable bias. The variable persons younger than eighteen is not highly correlated with the other included variables, and the size of the impact from including the variable is therefore not expected to be large. When including more variables that individually or in combination has a higher correlation with <18, the variable is likely to appear better fitted.

Regression 4: $RCB = \beta_0 + \beta_1 pop + \beta_2 tour - \beta_3 {<}18 - \beta_4 drinking + \epsilon$

When including the variable excessive drinking, the adjusted R2 increased. The coefficients for population and tourism remained fairly constant, while the coefficient for <18 decreased a lot. All the coefficients changed in the predicted direction if the previous output was infected by omitted variable bias. The t-scores, standard errors and p-values for the variables population and tourism were constant with the inclusion of the new variable. The effect on <18 was larger, with a t-score that decreased in magnitude, increased standard error and a p-value that went from significant at 19% to significance at 7,8%.

The result implies that the variable belongs in the equation and that bias was incurred in particularly in the <18 variable, which has the strongest simple correlation coefficient with the excessive drinking variable. The new variable is significant at 7% in the hypothesized direction. With that said, both of the variables <18 and excessive drinking now seems to belong in the equation and the previous poor result for <18 variable was caused by omitted variable bias. The p-values do not yet indicate that they definitely belong in the equation but the p-values are likely to improve when including additional variables that currently incur specification bias.

Regression 5: *RCB =* **ɓ***0 +* **ɓ***1pop +* **ɓ***2tour -* **ɓ***3<18 -* **ɓ***4drinking +* **ɓ***6smoke +* ‘

With the inclusion of the variable smoking, the adjusted R2 increased. The coefficients changed in the expected direction assuming previous bias due to omission of a significant variable. The coefficients changed the most for the variables with the highest simple correlation coefficient with the smoking variable. The standard errors for the population and tourism variables were kept relatively constant. It was relatively larger changes for the <18 and drinking variables. All p-values decreased except for the p-value for drinking, which increased to significant at roughly 10%.

The coefficient for the new variable is in the expected direction but it is not significant at a very low level. There is however a strong theoretical justification for keeping the variable in the equation. The drinking variable will also be kept since it is still significant at a pretty high level of confidence and its theoretical argument seems strong.

Since the next variable to be included is persons above the age of 65, there is an assumed current positive bias in the drinking coefficient if the new variable to be included is significantly different from zero. The t-statistic for the variable is therefore predicted to increase in magnitude with the inclusion, thus lowering the p-value. Considering that the simple correlation coefficient between the variables is -0.325, the change of the coefficient is likely to be relatively large. The same argument holds for the smoking variable.

Regression 6: $RCB = \beta_0 + \beta_1 pop + \beta_2 tour - \beta_3 {<}18 - \beta_4 drinking + \beta_6 smoke - \beta_5 {>}65 + \epsilon$

The new variable caused the adjusted R2 to increase. The coefficients all changed significantly in the expected direction, which indicates that the variable for persons over 65 was previously an omitted relevant variable. In particular; the variables <18, drinking and cigarette use changed the most dramatically. All coefficients are now significant at 5% except for persons older than 65 which is significant at 6.5% and smoking variable is significant at 5.25%

Regression 7-8: $RCB = \beta_0 + \beta_1 pop + \beta_2 tour - \beta_4 drinking + \beta_6 smoke + \beta_7 edu + \epsilon$

When testing regression 6 with the additional hypothesis that the level of education would be significantly different from zero, the adjusted R2 increased and the variable was significant at the 10 % confidence level, which could be an indicator that the variable is significantly different from zero at an even lower level. At the same time the significance levels of the variables for age rose sharply, which have also been the case when testing the model for general changes in specification, and in particular the variable >65 have been sensitive. The insignificant result for the age variables when including a variable for education likely stems from multicollinearity between the variables. 66% of the variation in percentage of the population with higher education is explained by the percentage of the population who is under 18 or older than 65 within the county.

Since the insignificant results are likely due to multicollinearity, conclusions regarding the effects of the age distribution are still assumed to be valid, as well as the result for education. The specification resulted in really good p-values for the variables included, as well as increased adjusted R2. When testing how sensitive this model was to changes in specification, the results all hold really robust. When including the variables arrests and PCI, the adjusted R2 decreases while the R2 is constant. The coefficients of the variables included in the regression do not change with the new variable and the p-values are constant. The p-value of the included variable is however really high.

4.2 Reliability analysis

4.2.1 Robustness testing

In order to test how well a model performs, it is good to change the specifications to see if the previous results still hold or if the model responds sharply to small changes in specification. I have tested to include the additional variables per capita income and number of marijuana related arrests to see what happens to the result in regression 6.

When testing how sensitive the model is for changes in specification, the variable <65 appear to change quite a bit in the different regression specifications even though the included variables does not infer collinearity. In particular the p-value rises to high levels of insignificance. In general the results hold up fairly well.

4.2.2 Influential points

When excluding the variable Denver which has sales of roughly 350 million $, compared to 70 million $ for the second highest observation, the output changes quite much. This means that this variable is an influential variable. It is however not an incorrectly recorded observation and it could therefore be left in the model, even though it may not be very reliable to let one observation influence the results depending on how much leverage the observation has. When it will be possible to run the regression with more observations having x-values that are of similar size, the point will be less influential and moving the regression line towards the true population.

4.2.3 Multicollinearity

It can be seen in the simple correlation coefficient matrix that the included variables does not have correlation coefficients larger than .8. In addition to checking the simple correlation coefficients, the variables have been measured according to their respective variance inflation factor. For all variables the VIF is below 5.

4.2.4 Residual plots

The residual plots for each variable have been briefly examined in order to see any non-linearites or non-constant variances that could be transformed to improve the model further. The main result is that the variables for smoking and persons younger than eighteen have equal variance over the distribution of residuals.

The variables >65 and heavy drinkers show traits of heteroscedasticity since the distributions of the residuals exhibits a funnel shaped pattern. In both plots, the variance decreases as x increases.

The residuals for the variables tourism and the size of the population have their residuals clustered around the lower x-values and with one observation being much larger horizontally. These are indicators of leverage points biasing the coefficients, which should be examined further.

All variables have one positive and one negative outlier.

5. Discussion

This chapter will present my own thoughts and conclusions regarding the results and the work process. I will discuss and interpret what these results may be used for. I will in this chapter also propose how this work could be used as a foundation for looking at other study areas that is within the concepts that have been evaluated in this paper.

5.1 Discussion of the regression results

The size of the population is significant in every regression, which is not a very surprising result. The result that areas with high tourism have high sales of cannabis could suggest that Colorado and Washington have tourists that visit the states with the main intention of consuming marijuana, or that counties with a lot of tourist visitors is also common for the in-state population to visit. It is however strongly theoretically justified to assume that the measure may dominantly be a measure of tourists buying marijuana considering the estimated 50% demand base by tourists, estimated for the state of Colorado.

The drinking variable is negatively correlated with the sales of cannabis on county level. Alcohol and cannabis should be viewed as substitute products (DiNardo & Lemieux 1992, p.43). The percentage of heavy and binge drinkers in a county should therefore reflect on how large the population is that currently prefers alcohol over cannabis. A county that has a large population and a high percentage of a population that prefers alcohol over cannabis could potentially have a large number of convertible consumers of cannabis if given the incentive.

Tobacco has been concluded to be a complementary good to cannabis in previous research and 80% of self-reported cannabis users also reported tobacco use. The result in this paper is that the variable has proved to be significant at a low level and it can therefore be said that county level tobacco use is an indicator of county level cannabis sales. I would not say that this proves that the goods are complements; it may rather be a measure of the size of the population that does not feel obliged to follow health norms, and therefore counties with a high percentage of smokers may have a population that tend to enact in other potentially harmful activities.

Assuming that the variable persons over sixty five is a variable that is equal to zero is a result that could indicate that counties that have a larger portion of persons who are sixty five years or older, does not have a general tendency to have lower sales of marijuana. In some regressions the variable is fairly significant and has the expected sign, which is consistent with theory of age characteristics of cannabis users in general. But it is however a variable that does not have very consistent results and it is for that reason not a result as strong as the other variables.

It is motivated to include a variable for persons under the age of eighteen, since this portion of a county are not allowed to purchase cannabis in retail stores and that most users are over that age. The hypothesis that as the portion of minors increase, the sales revenue will decrease is accepted. The results are more consistent than the variable for older people which indicates that it is more reliable to assume that the portion of younger people is a more certain indicator of negative sales than the portion of persons sixty five years of age or older.

Regarding the education variable, it was not obvious that it would be significant at a very low level and also provide the best fit for the model. Considering that only 20% of users identified as heavy users has a three year college degree or higher education, and the variable being highly significant holding up very well to different specifications. An explanation could be that with the higher price of retail cannabis other market segments choose cheaper priced supply channels of cannabis. Educated people should also be a segment that may have not used cannabis to a large extent when it has been illegal, and is a group with increased prevalence as an effect of legalization. This is in accordance with the notion of rational addiction which is a theory assuming that drug usage is a rational choice. This concept is reflected upon in the book by *Muscat et al (2009),* where it is reasoned that the consumer weighs all the advantages and disadvantages when forming a consumption decision. When a drug is legalized with a change in the perception of the drug and it is becoming accepted by society, a segment of users will find that it may be that using the product maximizes their utility. This can be assumed to explain who chooses to visit retail cannabis shops in Colorado and Washington.

It could also be argued that the result is due to the education variable is acting as a proxy for the age distribution in the county, given that 66% of the variation of education is explained by the age distributions.

Considering that lower income segments typically are overrepresented in drug usage it is interesting that the result is insignificant. This may suggest that the level of income is less of an indicator than education itself. And it seems likely that the insignificant result comes from the fact that per capita income is an average which makes it unrepresentative for how much money most people make.

Arrests was assumed to be an significant indicator and it is possible that the insignificance stems from drug arrests represent the part of a market that is illegal and not the segment that drives demand for retail cannabis.

5.2 Reliability of the model

The model does not have the best fit achievable since it has residual plots that exhibit traits of heteroscedasticity, leverage points and potential non-linearity. The model should be improved further with respect to these issues before relying too much on the results of the coefficients and the level of significance. The variables could be improved by testing different transformations or by plotting the exponent of X towards the correlation between X and Y, a method also known as Box-Cox linearity plot (NIST/SEMATECH, 2013).

Since there is at least one identified influential point, this should probably be excluded from the model to get more accurate coefficient estimates, as well as other influential points that may exist. Other observations of high influence may be recognized by using the measurement Cook´s D, which identifies the level of leverage of each data point. This is a measurement of how much the regression output changes when each observation is deleted from the input, if the value is above 1 it is considered to be too influential (idre, 2016).

With a standard error of fourteen million for the model it is not completely useless for determining in what range the sales outcome of a county could be considering that sales ranges between 78000 and 350 million. But it would be good to improve the fit and reduce the standard error of the estimate through the individual variables.

One issue is that a variable for the price of cannabis in the various counties was not included, which would be an important part in a demand function, and useful for hypothesis testing regarding market characteristics of the good. The dependent variable is not either adjusted for price differences, which is also an issue since two counties may have had similar retail sales but with very different quantity demand.

The variables does not measure the actual time period in which the sales incurred, it is not likely that very dramatic changes occurred in the variables used but it still is not optimal.

6. Conclusion

The optimal location for a cannabis retail store would be a county where a large population of inhabitants is 18-65 years old, there is an established tourism, a large percentage of the population is smokers, a small percentage of the population is heavy or binge drinkers and that have an educated population. It can be argued that counties with a large proportion of heavy alcohol users may be convertible local markets.

It is hard to know exactly what is measured when using county level percentages. If however theoretically motivated assumptions are made, it is possible to interpret the results on an individual level. If the assumptions implied by the theories underlying the hypotheses are assumed to be what the output measures, user segments can be identified by these results. Tobacco smokers is one segment since the proportion of tobacco smokers increase the revenues from cannabis sales. It can be argued that individuals that consume a lot of alcohol may be convertible customers since it is assumed that the goods are perfect complements if given the right incentives. If it is the number of tourists that drive demand, it is still hard to know if that mostly represent cannabis tourists, or if a typical tourist customer visits these shops as an event while visiting the states. It seems likely that it is two different segments of tourist customers. Higher retail prices excludes customers that cannot afford the higher prices, it therefore makes sense that a large proportion of retail customers are educated and are financially better off than the typical cannabis user, which have been adjusted for in Uruguay where taxes have been eradicated. It would be useful to study the characteristics of individual customers to retail cannabis stores to draw more accurate conclusions about the customers.

The legalized cannabis markets in Colorado and Washington supplies a portion of the market that does not represent the typical cannabis user. The high tax rates, and the high overhead costs of running production facilities and having stores, are making it hard to compete with suppliers that do not have the same costs of supplying their product. These sellers have likely in many cases also been on the market for quite some time and having a solid customer base. It is likely that the demand will eventually decrease considering more states will legalize the cannabis market and that a very large portion appears to be driven by tourism. It is also likely that the suppliers will continue to increase seeing the profits that can be made. Microeconomic theory tells us that both of these occurrences will cause the market price to decrease. To foresee the actual size of these effects, a different type of study would need to be done measuring the effects on both supply and demand as a function of price. The outcome of such a study could suggest that the retail stores will eventually be competing with black and grey market suppliers, or that the taxes are too high to erase these alternative markets, even when the market is in equilibrium.

7. Bibliography

Briggs, Bill. 2014. Marijuana Tourists: Are More Flocking to Washington and Colorado? *Nbcnews.* August 14. http://www.nbcnews.com/storyline/legal-pot/marijuana-tourists-are-more-flocking-washington-colorado-n176636 (retrieved 10/03-2016)

Castaldi, Malena. 2014. Uruguay to sell marijuana tax-free to undercut drug traffickers. *Reuters.* May 19. http://www.reuters.com/article/us-uruguay-marijuana-idUSKBN0DZ17Z20140519 (Retrieved 26/03-2016)

Caroline Cournoyer. 2016. *State Marijuana Laws Map*. Governing Magazine. http://www.governing.com/gov-data/state-marijuana-laws-map-medical-recreational.html. (Retrieved 16/05-2016)

Chaloupka, Frank J. et al. 1999. *Do higher cigarette prices encourage youth to use marijuana?* Report/National bureau of economic research: 6939. Cambridge: National bureau of economic research

Colorado department of revenue. *Colorado Marijuana Tax Data.* https://www.colorado.gov/pacific/revenue/colorado-marijuana-tax-data. (Retrieved 03/11-2015)

County health rankings organization. *Rankings Data.* http://www.countyhealthrankings.org/rankings/data. (Retrieved 13/12-2015)

Crost, Benjamin. & Guerrero, S. 2012. The effect of alcohol availability on marijuana use: evidence from the minimum legal drinking age. *Journal of health economics* 31(1):112-21. DOI: 10.1016/j.jhealeco.2011.12.005

Dean Runyan Associates. 2015a. *Colorado travel impacts 2009-2014p*. Report/Dean Runyan Associates. Portland: Dean Runyan Associates

Dean Runyan Associates. 2015b. *Washington state county travel impacts & visitor volume 1991-2014p*. Report/Dean Runyan Associates. Portland: Dean Runyan Associates

Diment, Dmitry. 2015. *New Highs: Growing acceptance of medical and recreational marijuana fuels industry expansion*. Report/IBISWorld Industry Report: OD4141. Publisher: IBISWorld Inc.

DiNardo, John. & Lemiuex, Thomas. 1992. *Alcohol, marijuana and American youth: The unintended effects of government regulation.* Report/National bureau of economic research: 4212. Cambridge: National bureau of economic research

Gruber, Jonathan & Köszegi, Botond. 2000. *Is addiction "rational"? theory and evidence.* Report/National bureau of economic research: 7505. Cambridge: National bureau of economic research

Huddleston, Tom Jr. 2016. Legal marijuana sales could hit $6.7 billion in 2016. *Fortune.* February 1. http://fortune.com/2016/02/01/marijuana-sales-legal/ (retrieved 16/05-2016)

idre. 2016. *Stata Data Analysis Examples, Robust Regression*. University of California. http://www.ats.ucla.edu/stat/stata/dae/rreg.htm. (Retrieved 10/03-2016)

Kilmer, Beau. et al. 2013. *Before the grand opening.* Report/RAND Drug policy center. Publisher: RAND Drug policy research center

Light M. et al. 2014. *Market size and demand for marijuana in Colorado.* Report/The Marijuana Policy Group. Denver: Colorado department of revenue

Muscat, Richard et al. 2009. *Signals from drug research.* Strasbourg: Council of Europe publishing

NIST/SEMATECH. 2013. *E-handbook of Statistical Methods*. Publisher: NIST/SEMATECH. http://www.itl.nist.gov/div898/handbook/eda/section3/boxcoxno.htm. (Retrieved 10/03-2016)

Puzzanchera, C. & Kang, W. 2014. *Easy Access to FBI Arrest Statistics*. http://www.ojjdp.gov/ojstatbb/ezaucr/asp/ucr_display.asp. (Retrieved 06/01-2016)

Statwing. 2016. *Interpreting residual plots to improve your regression*. Statwing Inc. http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/#x-unbalanced-header. (Retrieved 10/03-2016)

Studenmund, A.H. 2013. *Using econometrics: a practical guide.* 6. Ed. Boston: Pearson Education Limited.

Subritzky, Todd. Pettigrew, Simone & Lenton, Simon. 2016. Issues in the implementation and evolution of the commercial recreational cannabis market in Colorado. *International journal of drug policy* 27: 1-12. DOI: http://dx.doi.org/10.1016/j.drugpo.2015.12.001

US Census Bureau. *UNITED STATES QuickFacts from the US Census Bereau*. http://www.census.gov/quickfacts/table/PST045215/00. (Retrieved 23/11-2015)

Washington State Liquor and Cannabis control board. *Frequently requested lists.* http://www.liq.wa.gov/records/frequently-requested-lists. (Retrieved 22/11-2015)