Capstone Project 1: Data Wrangling

1. What kind of cleaning steps did you perform?

Data

In this project, I use country-level aggregate time series-monthly data, including the 2005:1-2015:2 periods for all the variables in the estimation models. As the quantity of demand, I use the total cigarette consumption as packages. As prices, I employ average price per package. Tax data consist of excise taxes for cigarettes, but not other or general tobacco products. Consumption, price, and tax data are monthly obtained from TAPDK (Tobacco and Alcohol Market Regulatory Authority). For real prices, price data are deflated by the Consumer Price Index (CPI). In order to control changes in income, the Total Industry Product Index (TIPI) that is highly correlated with the Gross Domestic Product (GDP) is used[1]. Data regarding TIPI and CPI are taken from the Turkish Statistical Institute (TUIK). All data are used in logarithmic form to interpret coefficients in the estimation models as long-term demand elasticities.

Dummy variables

Additionally, I include dummy variables representing the changes in anti-smoking policies such as taxes and regulations into the. However, I only include dummies for major anti-smoking policies into the model, but not all the changes in taxes and regulations. This is because all taxes and regulations do not affect significantly consumption. As a matter of fact, when I regress dummies for May 2008, May 2010, October 2010, July 2012, and June 2013 along with price, tax, and income together or separately in all the models, coefficients for those dummy variables were economically and statistically insignificant for all the scenarios, even when they were regressed only along with income to consumption. For those reasons, I do not include those dummy variables that represent the changes in May 2008, May 2010, October 2010, July 2012, and May 2013 into the model. However, although the tax increase in January 2013 alone is not a significant increase to raise prices and thus the demand for cigarettes, it becomes

[1] In Turkey, there is no monthly data for GDP.

meaningful, when I evaluate this tax increase along with major anti-alcohol regulations initiated in January 2013. This hypothesis is acceptable, because alcohol and cigarette in Turkey are complementary goods and cigarette consumption in Turkey declines to its lowest level in February 2013. I define these policies as mix strategy, because government simultaneously started to use both the tax increase on cigarettes and the most extensive advertisement and sale bans on alcohol in Turkey as of January 2013.

Accordingly, I employ four different dummies only representing the aforementioned major tax and regulation measures on smoking in Turkey. The first dummy variable is *regulation dummy (July 2009)* used to proxy the effect of extensive smoking bans in July 2009. The second one is *tax dummy (January 2010)* that represents the effect of tax increase in January 2010. *Tax dummy (October 2011)* that is the third dummy is used to proxy the effect of tax increase in October 2011. Lastly, I use a *mix strategy dummy (January 2013)*, which proxies the effects of strict anti-alcohol policies and tax increase in January 2013.

2. **How did you deal with missing values, if any?**

There were missing data.

3. **Were there outliers, and how did you handle them?**

As depicted in the figures in Capstone Project: Data Story 1 (https://github.com/tamercetin/Capstone-Project-1--Data-Story/blob/master/Untitled2.ipynb), there are no outliers in data.

Convert the final document to a .pdf and add it to your GitHub repository for this project. This document will eventually become part of your milestone report.