Capstone Project 1: In-depth Analysis (Machine Learning)

TAMER CETIN

In this capstone project, I empirically investigate the effect of anti-smoking policies such as regulation and taxation on cigarette smoking using country-level aggregate time-series data, which is appropriate for linear regression analysis. Using the relationships in the following figures, I develop hypothesis and build my predictive models. The figures and correlation coefficients show that there is a positive relationship between tax and price, a negative relationship between consumption/demand and tax/price. Also, I use dummy variables proxying regulations to analyze the effect of regulatory policies on consumption/demand.

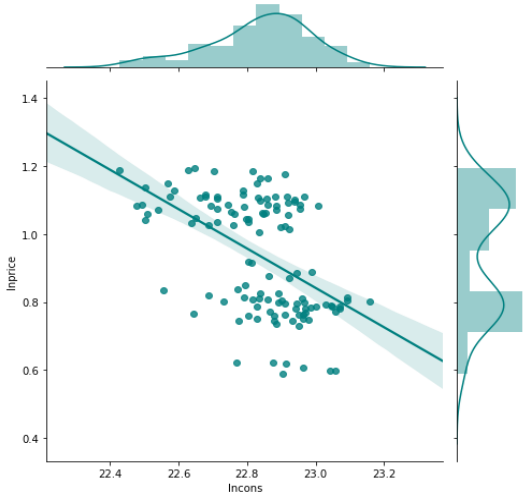

Fig. 1. Relationship between Consumption and Price
(Pearson Corr = -0.50)

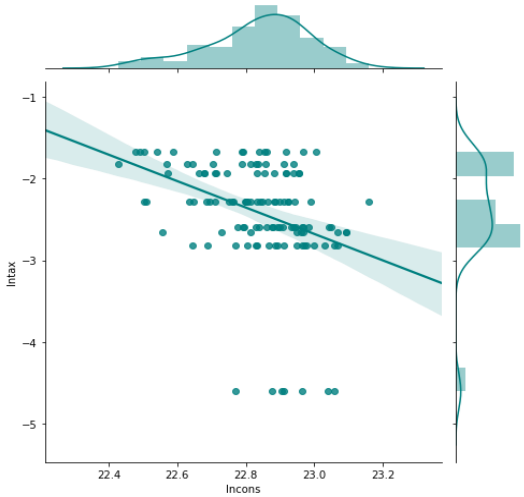

Fig. 2. Relationship between Consumption and Tax
(Pearson Corr = -0.36)

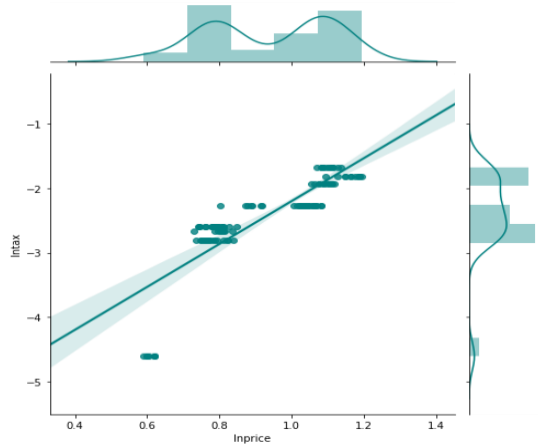

Fig. 3. Relationship between Tax and Price
(Pearson Corr = 0.85)

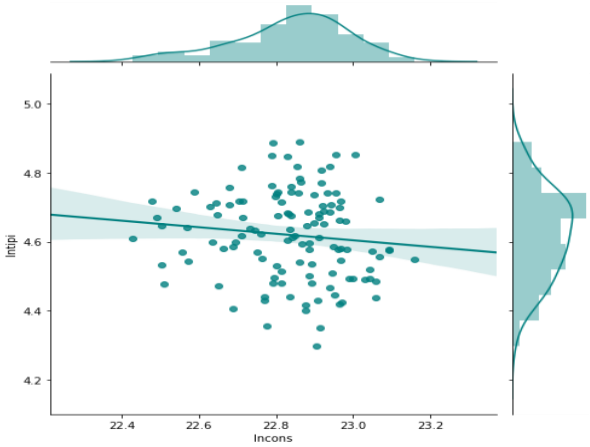

Fig. 4. Relationship between Consumption and Income
(Pearson Corr = -0.11)

Accordingly, I develop the following hypotheses:

H1: Tax increases increasing cigarette prices (See Fig. 3)

H2: There will be a negative relative relationship between tax rates, prices, and the demand for cigarettes (See Fig. 2 and Fig. 3).

H3: If anti-measure smoking policies are strong enough there will be a negative relationship between income and the demand for cigarettes (see Fig. 4).

H4: Anti-smoking regulation policies reduce the demand for cigarettes.

Accordingly, I use a classical demand estimation methodology to empirically investigate the relationship between anti-smoking policies and consumption/demand. The traditional models of demand estimation for cigarettes are as follows:

$$Qd_t = f(P_t,\ Y_t,\ R_t) \quad (1)$$

In Eq. (1), Qd_t is cigarette consumption in period *t*, P_t is price in period *t*, Y_t is a vector of shift variables including income, related prices, advertising, and R_t is a vector of regulation and tax variables. Accordingly, my model specifies the log of the demand for cigarettes as a function of the log of the own price, income, and dummy variables representing taxation and regulation. I estimate:

$$lnQd_t^{cig} = \beta_0 + \beta_1 lnP_t^{cig} + \beta_2 lnY_t^{income} + \beta_3 lnD_t^{tax,reg} + \mathcal{E}_t \quad (2)$$

where Qd_t^{cig} is the demand for cigarettes in the country in period *t*, P_t^{cig} is the price of cigarettes in period *t*, Y_t^{income} is income in the country in period *t*, $D_t^{tax,reg}$ dummies for tax and regulation, and $\mathcal{E}_t$ is the unobservable random disturbance term. In Eq. (2), parameters β_0, β_1, β_2, and β_3 are estimated as long-term elasticities, because the variables are used in logarithmic form. I do not include other price variables into the model because there are no close substitute or complementary goods for cigarettes. However, I proxy changes in tastes and preferences by a vector of dummy variables *D*.

According to the hypotheses, data structure, and estimation model functions introduced in Eqs. (1) and (2), the best machine learning technique is linear regression. Thus, in my estimation models, I analyze different relationships between the variables in Eq. (2) through different models under various scenarios. However, I do not include tax and price into the model at the same time, because the inclusion of tax and price that are highly correlated variables can lead to collinearity problem. I estimate the models using log *price* in place of log *tax* or dummies for tax and regulation along with price. In all the scenarios and models, I employ income as explanatory variable for controlling other non-price effects on smoking. I use monthly time-series data, including the 2005:1-2015:2 periods. I develop three different strategies to reveal the most robust results. First, I estimate the long-term dynamics of demand for cigarettes in Turkey under full sample with monthly data. By this strategy, I test nine different demand estimation models under three various scenarios. Second, I examine the pre- and post- taxation and regulation periods with sub samples presenting the different periods for an intertemporal comparison to better understand the effect of tax and regulations. Third, I analyze demand equations estimated in the models developed as per the first-second strategies through quarterly data. The aim is to reveal the probable estimation problems and to find the most robust results.

My findings include the OLS results from 9 different models under 3 various scenarios. Each equation includes the variables used to estimate the related long-term relationships for each scenario. Dummy variables are used to proxy the effect of taxation and regulations on smoking. Under the first scenario, I estimate the relationships between price, income and consumption with and without dummies. Then, I drop price variable from the first-scenario models and estimate relationships between tax, income, consumption and regulation dummies under Scenario 2. Lastly, I observe the relationships between income, dummies, and consumption under Scenario 3. Using those different scenarios and models, I aim to reveal the effect of tax and regulation policies on consumption in the most robust way.

In the first scenario, I find that price has a negative and statistically significant effect on smoking at 1% and 5% significance levels for all equations in Scenario 1, including price variable. The price elasticity of demand varies from -0.50 to -0.63. According to the results from the monthly data analysis, I calculate that the long-term price elasticity of demand for cigarettes in Turkey is -0.56 on average. The findings confirm that the long run demand elasticities in the post-2004 anti-smoking policies period are higher than the previous literature. I estimate that this is because of the consistent taxation and regulation policies in the last decade. Similarly, I find that income has a positive and statistically significant impact

on smoking at 1% and 5% significance levels for all equations in both Scenario 1 and Scenario 2 and Scenario 3. The income elasticity of demand for cigarettes ranges from 0.28 to 0.45. According to the results from the monthly data analysis, I calculate that the long-term income elasticity of demand for cigarettes in Turkey is 0.39 on average. The statistically and economically significant and relatively higher demand and income elasticities confirm that governments led to a remarkable decline in consumption through tax and regulation policies in the last decade in Turkey.

Regarding dummies, whereas the results regarding dummy variables economically, but not statistically, corroborate this inference for the first-scenario models, the second- and third-scenario demand estimation models present more robust evidence affirming that taxes and regulations have been influential in reducing consumption, because coefficients for dummy and tax variables are highly significant in those models both economically and statistically. The first-scenario dummy findings suggest that even though regulations in July 2009, the mix strategy in January 2013, and the tax increases in January 2010 and October 2011 mostly have a negative effect on consumption, they do not have a meaningful effect on smoking, because these coefficients do not have statistically significant values. I estimate that this is due to close correlations between price and dummy variables including the changes in taxes.

Because of concerns about endogeneity between price, taxation, and regulation in the first scenario, I drop the price variable from Eq. (2) in Scenario 2 and directly estimate the relationships between tax and consumption to reveal the effect of taxation and regulation by means of the models with and without regulation dummies. I find that an increase in excise taxes has a negative and statistically significant effect on smoking at 1% and 5% significance levels for all equations in Scenario 2. The long-term tax elasticity of demand for cigarettes varies from -0.05 to -0.11. This finding affirms that a 10% increase in excise taxes in the long run brings bout a decline between 0.5% and 1.8% in the demand for cigarettes in Turkey. This finding is consistent with the results from Scenario 1. Tax elasticities corroborate the findings regarding price elasticities in Scenario 1. Also, tax and price elasticities together suggest that governments in Turkey have generated higher tax revenues through excise taxes on smoking in the last decade.

As expressed before, for the equations in Scenario 2, income also has a positive and statistically significant impact on smoking. Lastly, regulation dummies in the tax models suggest that regulations have a negative impact on smoking. *Regulation dummy (July 2009)* is significant at a 1% significance level and a negative effect under this scenario. As can be understood from the first-two scenarios, the relationships between consumption and dummies are generally as expected, but not statistically significant in some models. One can claim that this is because there are the tax changes and regulations in close intervals to each other and this leads to endogeneity in the analysis with monthly data. For that reason, under Scenario 3, I only investigate the relationship between consumption and dummies using income to control other explanatory variables in these regressions. Table 2 shows the results. Under this scenario, dummies that represent the changes in taxes and regulations are mostly statistically significant and as expected. Clearly, as different from the previous estimations, these findings confirm more strongly the presence of relationship between dummies and consumption. Also, this evidence is consistent with the previous findings and corroborates that the increase in taxes and restrictive regulations on smoking considerably reduce consumption. Shortly, when I investigate the long-term dynamics of consumption only along with income as explanatory variable, dummy variables become statistically more significant. I estimate that this is because there is no any correlation relationship between income and dummy variables.