Economics 312: Final Project
John Mills
Daniel Hope
Chris Weber

Demographics and Consumption Tax Revenue

Introduction

In this project we examine the effect of state-level demographics for the 48 contiguous states the percentage of tax revenue generated through various consumption taxes. We estimated the relationship using fixed effects and random effects models. We find that age distribution has a non-negligible effect on the proportional distribution of consumption tax revenues, with respect to alcohol, cigarette, and gasoline tax. Younger demographic groups have a statistically significant, positive effect on proportional tax revenues in all three categories. Additionally, we examine the assumptions of our regressions and find them to be valid.

Data

This project uses a customized data set called finaldata.dta. It was obtained via Professor Jon Rork's data collection. It contains observations of demographic data from 48 states, excluding Alaska and Hawaii due to incongruous market forces at work in those two states in comparison to the contiguous 48 states. The dataset includes the variables representing unemployment rate, per capita income, educational attainment, age distribution and categorical proportions of tax revenues. The finaldata.dta is a panel dataset because it contains both cross-sectional observations for the above variables for a single year, and observations over time for 41 years. We chose to cut the raw panel data from 1960-2011 to 1970-2011 in order to obtain a balanced panel.

The summary statistics for the raw data are as follows:

Variable	Obs	Mean	Std. Dev.	Min	Max
state_name	0				
stfips	2016	24.5	13.85684	1	48
year	2016	1990.5	12.12393	1970	2011
urate	2016	5.955655	2.094736	2.1	18
pci	2016	19917.48	11887.3	2628	57902
Pct__High_~l	2016	75.43418	11.33495	37.8	93
Pct__College	2016	20.03376	6.567956	2.5	40.4
age04	2016	.0728256	.0103275	.0385301	.1321004
age517	2016	.2010248	.0289316	.1387294	.3048576
age1824	2016	.1124094	.0163658	.0779133	.1537592
age2564	2016	.4980857	.0497667	.3901975	.7490116
age65	2016	.1202162	.0199993	.01567	.1854977
gastaxpct	2016	.0859371	.0370386	.0077924	.2594719
alchtaxpct	2016	.0145462	.0126751	0	.0844416
cigtaxpct	2016	.0262404	.0174244	.0011381	.1726619

The variables in the dataset are defined as follows:

state_name: name of state stored as a string
stfips: the Federal Information Processing Standard (FIPS) code for each state
year: the year of observation, ranging from 1970-2011
urate: the unemployment rate of a state in a given year
pci: per capita income for a state, not adjusted for inflation
Pct_High_School: percentage of state population that are high school graduates in a given year (abbreviated Pct_High_~1 above)
Pct_College: percentage of state population that are college graduates in a given year
age04: The percentage of state population between the ages of 0 and 4 in a given year
age517: The percentage of state population between the ages of 5 and 17 in a given year
age1824: The percentage of the population between the ages of 18 and 24 in a given year
age2564: The percentage of the population between the ages of 25 and 64 in a given year
age65: The percentage of the population over the age of 65 in a given year
gastaxpct: The percentage of tax revenue consisting of gasoline tax in a state in a given year
alchtaxpct: The percentage of tax revenue consisting of alcohol tax in a state in a given year
cigtaxpct: The percentage of tax revenue consisting of cigarette tax in a state in a given year

For this paper, we estimate regressions of the proportion at which the consumption taxes of alcohol, gas, and cigarettes are affected by the age distribution of a state population, the state unemployment rate, educational attainment, and per capita income.

Theory and Expectations

Theory tells us that tax revenues are driven by consumption of the taxed good, which of course is balanced via elasticity measures. The more inelastic a good, the greater the tax burden can be placed on the consumer via higher prices. Our analysis is driven by the question of whether the characteristics of the states, notably, their age distributions, affects the proportion of their tax revenues that come from our target categories: alcohol, cigarettes, and gasoline. Our expectations are that having an age distribution such that there are larger amounts of the population in a bracket that would have greater (and perhaps more inelastic) demand for the good in question leads to a higher proportion of tax revenues. Put more succinctly, we expect those states where there are relatively higher numbers in the age brackets who consume relatively more of the good in the face of a tax to have higher proportional revenues. For gasoline, we may expect states with larger numbers of working age citizens to have higher proportions of tax revenues. For cigarettes, we may expect older populations to have higher proportional revenues, given that smoking has decreased over time. For alcohol, we may expect populations with a large number of college-age citizens to have higher proportional revenues.

As for other variables, we might expect unemployment to have some sort of an effect on consumption-based taxes; we might expect proportional revenues to vary inversely with unemployment; however, it is not clear that this would be a simple inverse relationship.

Methods

When dealing with panel data, the major modeling consideration is whether or not to use a fixed effects or a random effects specification. The test that determines whether a fixed effects or a random effects specification is appropropriate is a Hausman test. Implementing a Hausman test for the dependent variable gastaxpct, we get the following output from STATA:

```
              ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed          .          Difference          S.E.
-------------+------------------------------------------------------------------
       urate |   -.0010525    -.0011348        .0000823        .0000232
         pci |   -3.16e-07    -3.31e-07        1.47e-08        2.04e-08
Pct__High_~l |   -.0011544    -.0010674        -.000087        .0000303
Pct__College |   -.0001076    -.0001401        .0000325         .000064
      age517 |    .6605156     .6736995        -.013184         .008597
     age1824 |   -.2924114    -.2799185       -.0124929        .0084849
     age2564 |    .0254131     .0234602         .001953        .0009191
       age65 |    .3324128     .3327613       -.0003485        .0231451
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

              chi2(7) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =       29.83
                 Prob>chi2 =      0.0001
```

At a significance level of 5%, we reject the null hypothesis that gastaxpct should be modeled using a random effects model. We accept the alternative hypothesis of a fixed effect specification, meaning that there is correlation between the regressors and the error term. Implementing a Hausman test for the dependent variable cigtaxpct, we get the following output from STATA:

```
              ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed          .          Difference          S.E.
-------------+------------------------------------------------------------------
       urate |    .0004143     .0004066        7.73e-06        .0000105
         pci |    7.24e-07     6.95e-07        2.87e-08        9.24e-09
Pct__High_~l |   -.0003559    -.0003587        2.82e-06        .0000137
Pct__College |   -.0007007    -.0006278       -.0000729         .000029
      age517 |    .3947889     .3973503       -.0025614        .0039065
     age1824 |    .1373072     .1376396       -.0003324        .0038529
     age2564 |   -.0094821    -.0095589        .0000768        .0004186
       age65 |    .2809003     .2864397       -.0055394        .0106018
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg
```

```
    Test:  Ho:  difference in coefficients not systematic

                  chi2(7) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       24.58
                Prob>chi2 =      0.0009
```

At a significance level of 5%, we reject the null hypothesis that cigtaxpct should be modeled using a random effects model. We accept the alternative hypothesis of a fixed effect specification, again implying that there is correlation between the regressors and the error term. Lastly, implementing a Hausman test for the dependent variable alchtaxpct, we get the following output from STATA:

```
                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed          .          Difference          S.E.
-------------+----------------------------------------------------------------
       urate |   -.0000296    -.0000336        4.00e-06        5.00e-06
         pci |    9.01e-08     8.75e-08        2.60e-09        4.43e-09
Pct__High_~l |   -.0004525    -.0004609        8.40e-06        6.58e-06
Pct__College |    -.000042    -.0000339       -8.12e-06        .0000139
      age517 |    .1374355     .1375996       -.0001642        .0018735
     age1824 |    .0891591     .0899171        -.000758         .001847
     age2564 |    .0013561     .0015855       -.0002294        .0002011
       age65 |    .0070446      .014449       -.0074044        .0051127
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(7) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        8.41
                Prob>chi2 =      0.2980
```

At a significance level of 5%, we fail to reject the null hypothesis that cigtaxpct should be modeled using a random effects model. We therefore do not accept the alternative hypothesis of a fixed effect specification, and instead use a random effects specification to model the dependent variable alchtaxpct.

In summary, we select a fixed effect specification to model gastaxpct and cigtaxpct, and we select a random effects specification to model alchtaxpct.

Results

The following are the outreg tables for our chosen regressions of the percentage of tax revenue due gas, alcohol, and cigarettes.

Fixed Effect Specifications:

	gastaxpct	cigtaxpct
urate	-0.001	0.000
	(2.08)*	(2.03)*

pci	-0.000	0.000
	(1.84)	(3.88)**
Pct__High_Shool	-0.001	-0.000
	(4.02)**	(1.73)
Pct__College	-0.000	-0.001
	(0.15)	(1.54)
age517	0.661	0.395
	(6.98)**	(6.19)**
age1824	-0.292	0.137
	(2.00)	(1.97)
age2564	0.025	-0.009
	(2.53)*	(2.02)*
age65	0.332	0.281
	(1.76)	(1.78)
_cons	0.035	-0.074
	(0.60)	(1.90)
R2	0.72	0.54
N	2,016	2,016

* p<0.05; ** p<0.01

As we can see from the outreg table above, whose constituent regressions use Huber/White robust standard errors, the only regressors which have statistically significant effects on gastaxpct are urate, pct_High_shool, age517, and age2564, with pci and age 517 at the 0.01% significance level, and pct_High_shool and age2564 at the 0.05% significance level. Of these, urate and pct_High_shool have coefficients which are negligible to the point that STATA reports them as being 0. The variables age517 and age2564 are the only regressors that have both statistically significant and numerically significant coefficents. A 1% increase in the proportion of the population between ages 5-17 leads to a 0.661% increase in the percent of tax revenue generated by gasoline taxation; a 1% increase in the portion of the population between ages 25-64 leads to a 0.025% increase in the percent of tax revenue generated by gasoline taxation. In terms of the real world implications, with respect to the coefficient on age517, we speculate that this increase is due to the fact that parents with children in this age range have to drive their offspring most everywhere, whether to school or various activities. This speculation implies increased gas consumption and thus higher tax revenue from sales of gasoline. As for the population aged 25-64, this is the portion of the population that comprises most of the work force, and thus they have to travel to and from work each day, most likely by car. Furthermore, this is the portion of the population most likely to have children within the 5-17 age range, which would increase gas tax revenue for the same reasons discussed above.

For cigtaxpct, the only regressors which have statistically significant effects are urate, pci, age517 and age2564. Of these, urate and pci both have coefficients that are negligible to the point that STATA reports them as being 0. This leaves age517 and age2564, which have coefficients of 0.395 and -0.009 respectively. Determining the real world reasoning behind these results however, is more difficult, and we cannot think of a reason why age517 in particular would have such a noticeable effect on the percentage of tax revenue made up by cigarette sales, given that 18 years old is the legal age to buy cigarettes in the United States, with some

states requiring an individual be even older than 18. The possible reasoning behind why age2564 has a negative effect on cigarette tax revenue is that most people in this age range, grew up during the time period when the health risks of cigarettes were coming to light and the product as a whole was being attacked more, decreasing the interest in smoking. It should be noted that the upper age bracket does not have a significant effect; we might have expected there to be an effect here. However, with all respect to good taste, it is feasible that the smokers drop out of the dataset earlier, and so there is not a significant subset of the oldest age bracket in our data that are smokers such that there is a statistically significant effect.

Random Effects Specification:

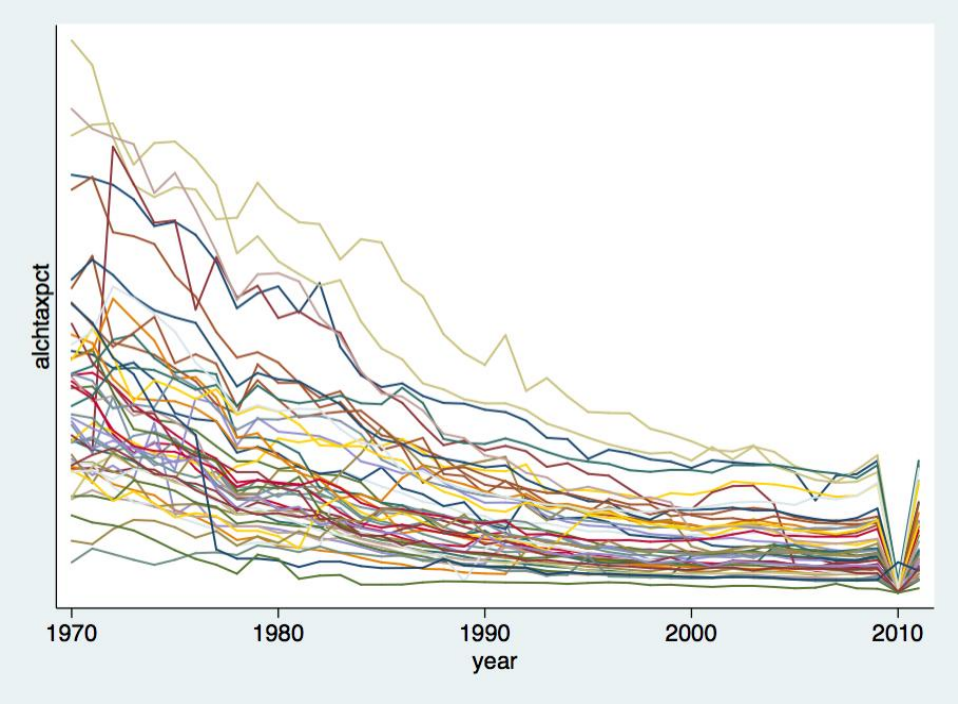
	alchtaxpct
urate	-0.000
	(0.22)
pci	0.000
	(0.87)
Pct__High_Shool	-0.000
	(3.15)**
Pct__College	-0.000
	(0.14)
age517	0.138
	(4.02)**
age1824	0.090
	(3.05)**
age2564	0.002
	(0.51)
age65	0.014
	(0.19)
_cons	0.008
	(0.47)
N	2,016

* p<0.05; ** p<0.01

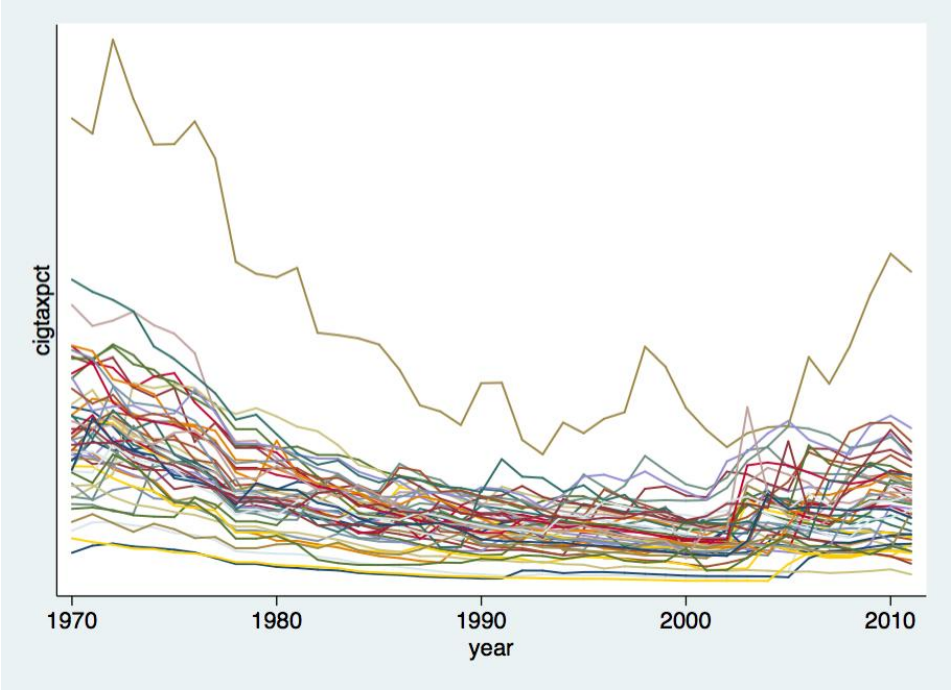
The outreg table for alchtaxpct, shown above, uses the random effects specification. The only regressors which have statistically significant coefficients are Pct_High_Shool, age517 and age1824, though Pct_High_Shool's coefficient is negligible to the point that STATA reports it as being 0. Our results indicate that a 1% increase in the portion of the population between the ages 5-17 leads to a 0.138% increase in the portion of a states' tax revenue generated by taxing alcohol; a 1% increase in the portion of the population between the ages 18-24 leads to a 0.090% increase. The real world reasoning behind age1824's coefficient is easy to grasp, given that this range contains college students, as well as underage drinkers, who are the most likely to binge drink and purchase large amounts of alcohol frequently. Age517's effect is harder to grasp however; we hypothesize that this may be due to parents of children in that age range drinking to take the edge off, or relax after dealing with their offspring. However, this is entirely speculative.

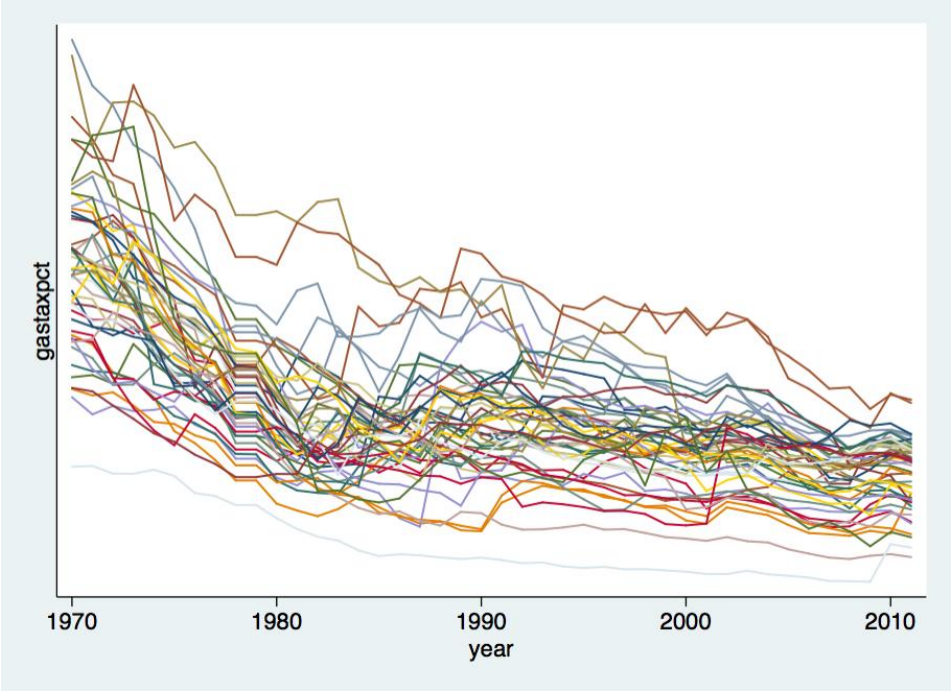
Assessment of Validity

To determine whether our models are biased or inefficient, we need to examine the assumptions made in the fixed effects and random effects model. First, we assume that the relationship dependent variable can be linearly approximated. We examine this cursorily by



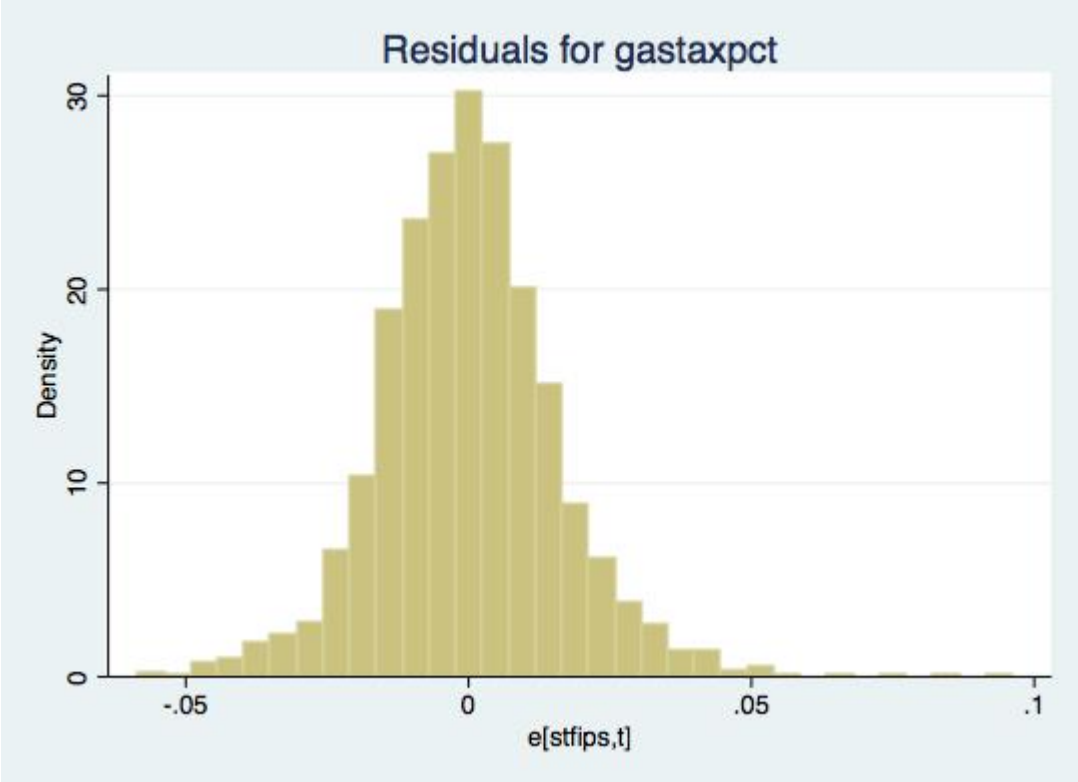
looking at the graphs of each dependent variable over time for each state:

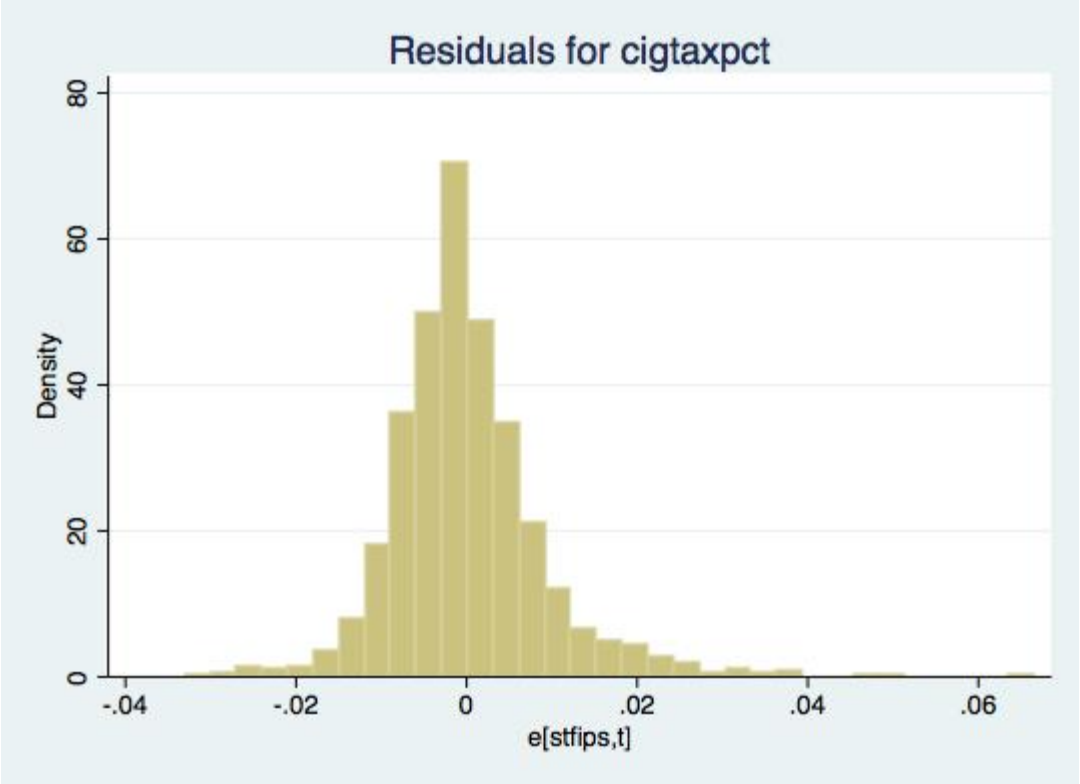

cigtaxpct
1970
1980
1990
2000
2010
year

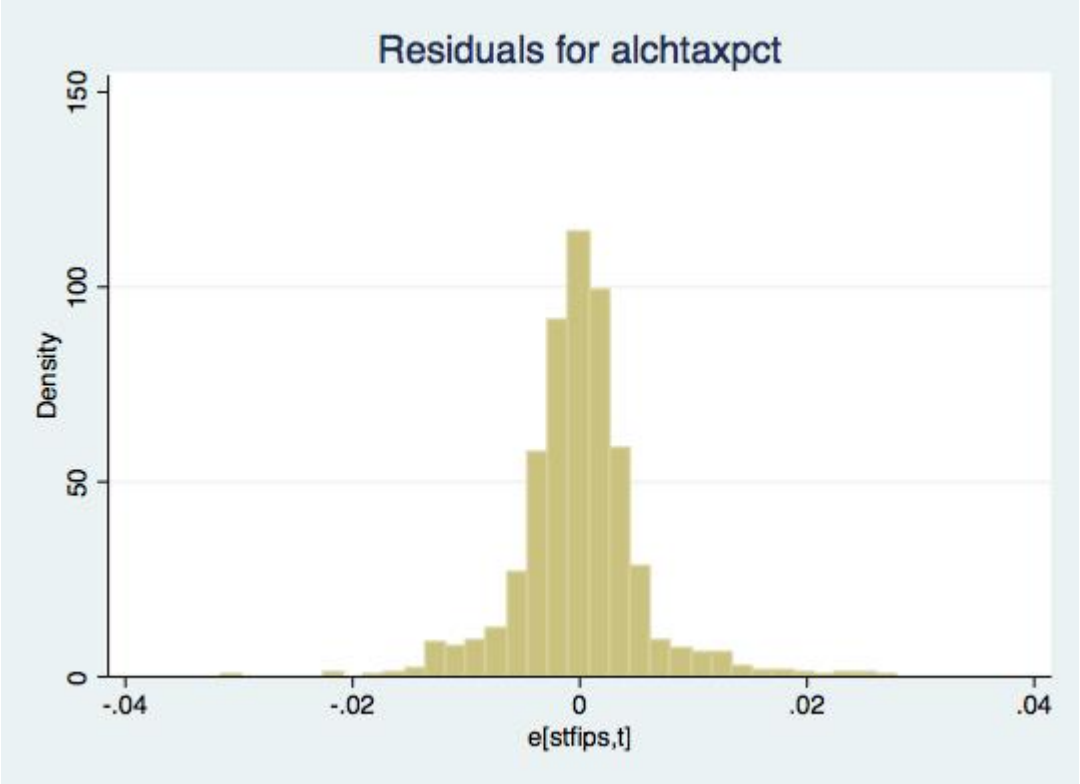

gastaxpct
1970
1980
1990
2000
2010
year

The panel over time appears linear as a whole, but is certainly not perfectly linear. Furthermore, there appear to be some outliers present (New Hampshire gets a lot of their tax revenue from cigarette taxation, for example), which of course detracts from our analysis. However, the functional form certainly doesn't appear to be cubic, or any other sort of rare form.

Second, we assume that the error terms in the regression are normally distributed. To test this assumption, we graph the residuals of the regressions:

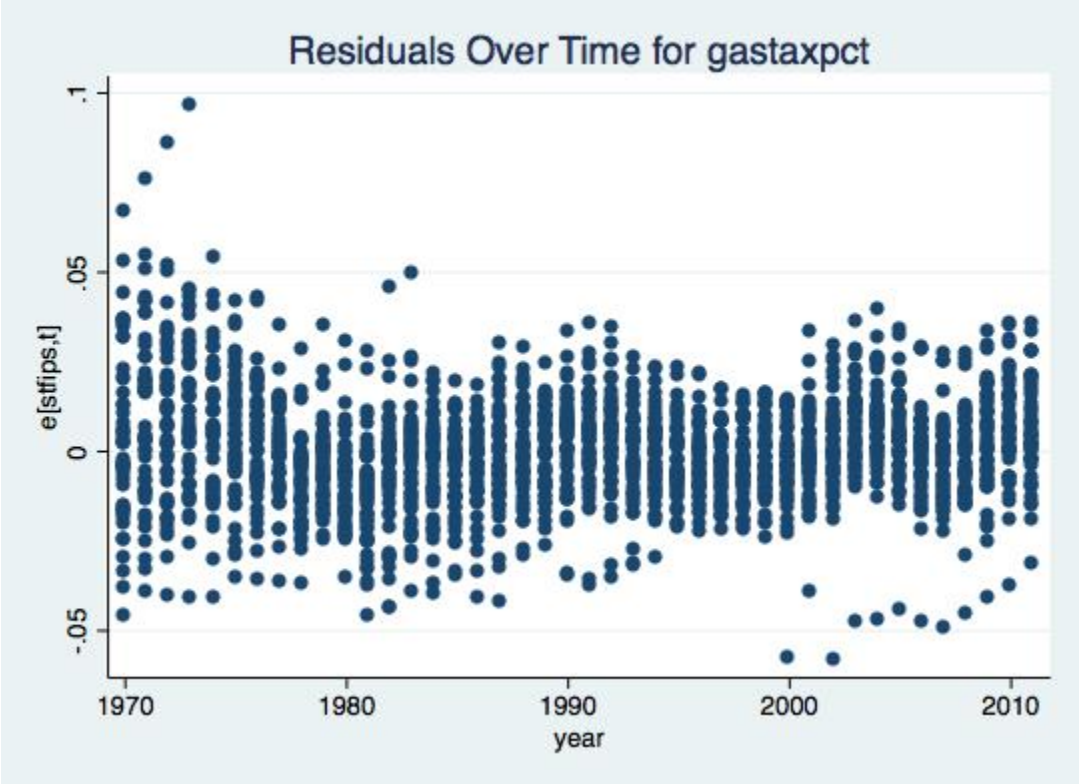


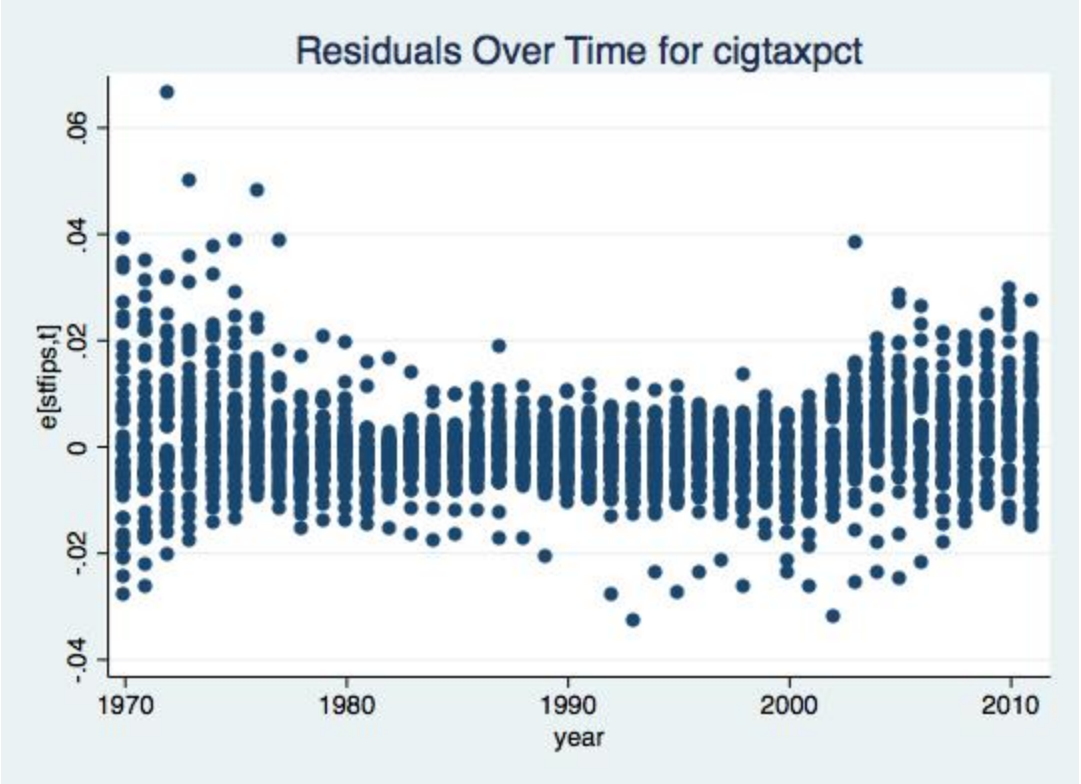

Residuals for cigtaxpct
Density
0
20
40
60
80
-.04
-.02
0
.02
.04
.06
e[stfips,t]

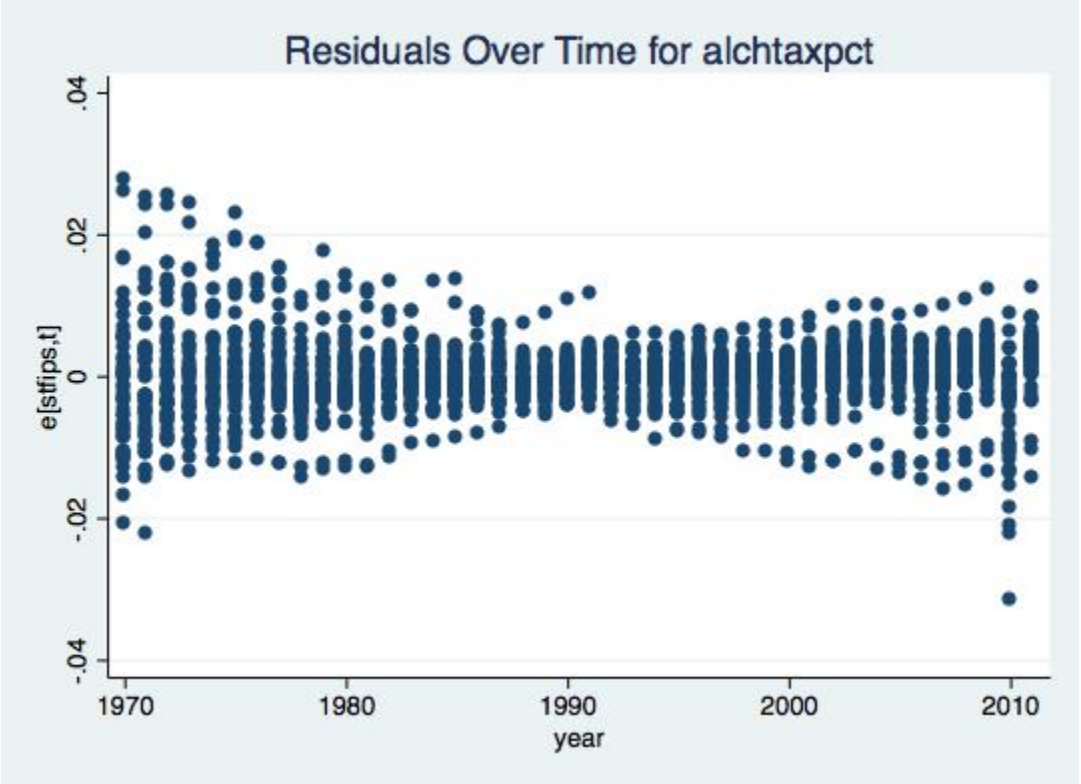

Residuals for alchtaxpct
Density
0
50
100
150
-.04
-.02
0
.02
.04
e[stfips,t]

The residual plots appear to be nearly normal, which is in line with our assumption.

Third, we assume that the error terms of the regression are independent and identically distributed. To check this, we look at the residuals over time:




Residuals Over Time for cigtaxpct
e[stfips,t]
.06
.04
.02
0
-.02
-.04
1970
1980
1990
2000
2010
year


Residuals Over Time for alchtaxpct
e[stfips,t]
.04
.02
0
-.02
-.04
1970
1980
1990
2000
2010
year

The residuals appear to be non-trended over time, and imply our observations are reasonably independent, satisfying this assumption.

Fourth, we assume homoskedasticity of the error terms in our regressions. This is a minor concern and the only effect of heteroskedasticity would be to make our estimate inefficient. Heteroskedasticity would not bias our coefficient estimates. To deal with any potential heteroskedasticity, we employ robust standard errors in our regressions using the option vce(robust).

Fifth, we assume that there is no autocorrelation in our fixed effect regressions. The random effects model uses autocorrelation in its application. For the fixed effects model, given our use of Huber/White robust standard errors, we are not concerned with any within-panel serial correlation.

Sixth, we assume that no measurement error occurred during data collection. Since we have no connection to the data collection process for our data set, we are unable to comment on the nature of any potential measurement error that could potentially affect our regressions.

Seventh, we assume exogeneity in our regressions. The dependent variable does not have an effect on the regressors for our equations. Endogeneity would be a severe problem for our models, however we lack the correct instruments to correct for any endogeneity. The percentage of tax revenue could affect the consumption in an age group in the event that the state in question imposes extremely drastic taxation because of the large number of people in the population that perhaps are especially prone to consumption of a particular good. (i.e. taxing at high, almost discriminatory rates to discourage consumption). This could bias the effect downward over the course of the sample period as consumption is gradually discouraged over time, and perhaps those most inclined to binge leave the state for friendlier tax rates.

Lastly, and almost implicitly, we assume that we are using the right model to generate estimates from our data. This assumption is unproblematic give the nature of our data. The fact that we have panel data dictates that we need to use a pooled model. In the methods section above, we discriminated whether or not we should use fixed effects or random effects to generate our estimates, and chose according to our test statistics.

Conclusion

After analyzing our regressions, we saw gastaxpct and alchtaxpct largely matched up with our theoretical expectations. The results of our gastaxpct regression indicated that states with a larger working age demographic tended to have higher gas tax revenue. Our results for the alchtaxpct regression also supported our theoretical claims where states with a higher college age population generated more revenue from the taxation of alcohol.

The results of our cigtaxpct regression did not match up with our theoretical expectations. We were unable to come up with an adequate explanation as to why age517 would have such a noticeably positive effect on cigarette tax revenue, given that this age group is legally unable to buy cigarettes.

However, on the whole, our results did match up with our theoretical expectations. Given more time and a larger data set, it would be worthwhile to examine more in depth the makeup of these demographics, looking at variables such as gender, race, and political background of a state. It is worth noting that we did not find any sort of dramatic effects here. It is likely that state tax revenue proportions are largely driven by political pressures and timing in the state

themselves. These nuances are not captured in our project; we attempted to stay out of politics and tried to characterize the proportional revenues as driven by the population themselves, rather than their political dynamics. This might be the incorrect way of approaching the topic, but the project provided a useful look into the non-political drivers of how states structure their revenue streams.