Economics 312
Emmanuel Enemchukwu
Final Project

Does lottery money boost education spending?

A case study of the state of Connecticut

Introduction

Gambling, as a generalization has been in this region of the world as far back as the 15th Century, at which point the United States of America was a cluster of British colonies. The acceptance of gambling in the colonies, however, was short-lived because it was seen as a vice, and a sign of laziness. Though chastised, the revenue to some colonial governments from gambling was simply irresistible, and was instrumental in running these colonies. The Jamestown colony is a case in point of where gambling money was significant in keeping the colony afloat. The debate on the harm and benefit of state sanctioned gambling wasn't left behind in the colonial era. States across the US still grapple on the issue, and currently, six states have not approved state lottery. These states are Alabama, Alaska, Hawaii, Mississippi, Nevada, and Utah.[1] The proponents of state lottery have made arguments about how states have been able to finance themselves with lottery revenue. And its critics argue that it breeds laziness and thus, undermines productivity. Proponents interested in canceling out the argument that gambling undermines productivity argue that even if this assertion were remotely true, it is offset by the role lottery revenue plays in funding education, a factor that significantly boosts productivity. In a bid to put some empirical evidence into this debate, this paper uses Connecticut as a case study to investigate the effect of lottery revenue on education spending.

Literature Review/ Theory

Statistics establishes that the state and the local governments provide 93% of education expenditure[2]. The implication of the above information is that the quality of education in any state depends on the ability of the state to fund its educational program. On another hand, across the US and in areas under US

[1] Wikimedia Foundation. "Lotteries in the United States." Wikipedia. http://en.wikipedia.org/wiki/Lotteries_in_the_United_States (accessed May 8, 2014).
[2] Woodruff, Judy. "How Do We Fund Our Schools?." PBS. http://www.pbs.org/wnet/wherewestand/reports/finance/how-do-we-fund-our-schools/197/ (accessed May 8, 2014).

jurisdiction, not only do states set the lottery laws, they also have very lucrative state lottery schemes[3]. Thus, lottery potentially plays an important role in generating the income used in funding education, a claim this paper seeks to explore and assess.

An assumption made from intuition is that the education variable *educexp* denotes the total education expenditure in the state. Literature already tells us that the total education spending in any US state is made up of Federal Government contribution, which is about 7%, and state and local government contributions, which make up the remaining 93%. We begin to get the idea that the usefulness of lottery revenue only goes as far as giving us an idea of its contribution to state spending on education. The implication of this clarification is that we shouldn't expect to get a very high R2 in our regression as a good chunk education spending, federal and local government spending is beyond our predicting power.

The established understanding on the issue of education funding are that *sales* and *income taxes* (both *corporate* and *personal*), and property tax are the main sources of funds for running education programs. This is more than just an understanding; it is the way the program was designed. It is noteworthy to point out that Property tax, though significant, only becomes necessary when we look at a local level[4]. Thus, considering that we are only interested in the state level, we will not be working with property tax.

Except for a few instances and sectors, government expenditure and revenue, over time, have been increasing in real dollar values. Education spending also should prove to follow this rule of increasing over time. This means that most of the variables in this paper, including education, will be nonstationary.

Methods and Data

To conduct this analysis, this project uses a couple of time series and panel datasets collected and supplied by Prof. Jon Rork and Prof. Jeff Parker, both from the Reed College department of Economics.

---

[3] Wikimedia Foundation. "Lotteries in the United States." Wikipedia. http://en.wikipedia.org/wiki/Lotteries_in_the_United_States (accessed May 8, 2014).

[4] Woodruff, Judy. "How Do We Fund Our Schools?." PBS. http://www.pbs.org/wnet/wherewestand/reports/finance/how-do-we-fund-our-schools/197/ (accessed May 8, 2014).

These datasets include, *lottery.dta, income.dta, expenditure.dta,* and *agebystate.dta.* These datasets span from 1967 to 2012, but this paper only uses information from 1967 to 2002 because of the completeness of data within the range.

Deciding the best way to adjust the monetary variables posed some challenges. One way to think of it is to adjust the monetary variables in these datasets for inflation to the 2014-dollar values and then divide by the population to present our values in per capita terms. Presenting these values in the same year dollar values and also converting them to the per capita values ensure that the values are comparable between time and states. These variables that have been standardized for inflation and per capita terms come with the suffix *'iapc'* to distinguish them from the original variables. In this specification, the population variable is irrelevant to the regression because adjusting the variables to per capita terms already accounts for them. Alternatively, we can choose not to present the monetary variables in per capita basis, but instead, we can choose to make the population of any year the base year (let's say, the 1967 population as the base year just for convenience), and then make the population of subsequent years to be ratios of the 1967 values. Because the population is increasing, the ratios will be increasing too. The population ratio can be included in the regression, as a way of accounting for the change in population, and its impact in the regression. Both processes appear plausible, but this paper will work with the former just for the sake of convenience.

The data definition is produced below, the units of the observations were not explicitly stated, and discretion was applied in choosing the units

Obs: **********************

state	state
iapcgsp	inflation adjusted per capita gross state product
lage04	log of sum of ages 0 to 4
lage517	log of sum of ages 5 to 17
lage1824	log of sum of ages 18 to 24
lage2564	log of sum of ages 25 to 64
lage65	log of sum of ages above 65
lpop	log of total population
iapceducexp	inflation adjusted per capita total expenditure on education
iapclottrev	inflation adjusted per capita total lottery revenue
iapcpersoninctx	inflation adjusted per capita total personal income tax
iapccorpinctx	inflation adjusted per capita total corporate income tax
iapcsalesrevenue	inflation adjusted per capita total sales revenue

Summary of variables from across the US:

Variable	Obs	Mean	Std. Dev.	Min	Max
iapceducexp	1548	.3351744	.326578	.0084958	1.791527
iapclottrev	1548	.0000301	.0000641	0	.0007869
iapcpersinctx	1548	.0001593	.0002025	0	.0011619
iapccorpictx	1548	.0000319	.0000394	0	.0006152
iapcttaxrev	1548	.0005149	.0004835	.0000134	.0028455
iapcpci	1548	.0034785	.0055585	.0000289	.0487865
iapcgsp	1548	.0094946	.0085771	.0003415	.0571153

Summary of variables for the state of Connecticut

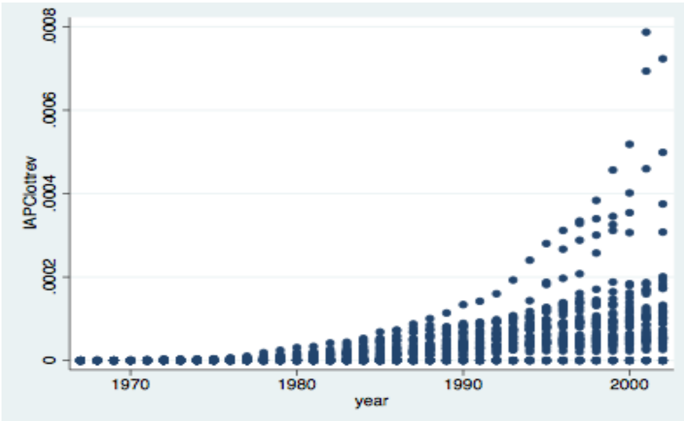
```
. summarize iapcpersinctx iapccorpinctx iapcttaxrev iapcgsp iapclottrev lage517
> lage1824
```

Variable	Obs	Mean	Std. Dev.	Min	Max
iapcpersin~x	36	.0002094	.0003047	0	.0009689
iapccorpin~x	36	.0000614	.0000483	3.89e-06	.0001384
iapcttaxrev	36	.0007334	.0007399	.0000222	.0022984
iapcgsp	36	.0128809	.0119534	.0006599	.0372295
iapclottrev	36	.0000632	.0000666	0	.0002007
lage517	36	11.04657	3.327187	6.270988	13.55313
lage1824	33	10.83702	3.035997	5.916202	12.86119

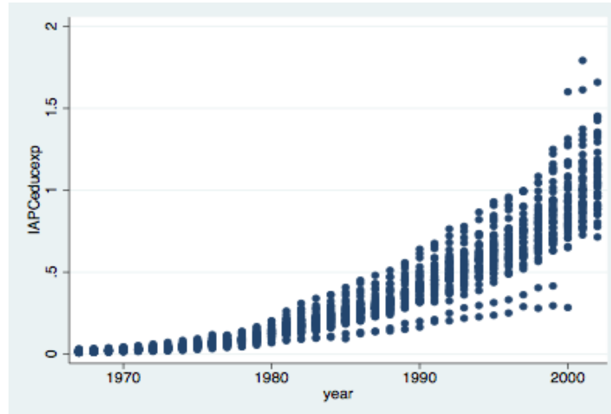
To the extent that we can speak by just looking at these numbers, though some variables are small, these values appear unproblematic. Looking at these tables and their corresponding graphs helped us visual a problem of inconsistent scaling with the *population* variable, which was consequently corrected. Admittedly, though it is easy to detect an inconsistent scaling within a particular variable because of the outliers such inconsistency would generate. Without the keys to the scale, we cannot detect when a variable is scaled consistently but with a wrong scale. Because the means and standard deviations appear really small It wouldn't make much sense to calculate the effect on education of values that are zeros up till the third or fourth decimal place, but the interpretation would make more sense if we start looking at the variables to the 100^{th} or 1000^{th} units.

4 Sample scatterplots of the variables across the US:

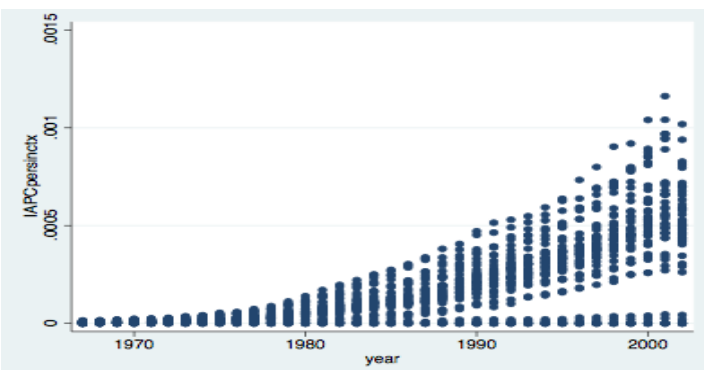
a.) twoway (scatter iapclottrev year)



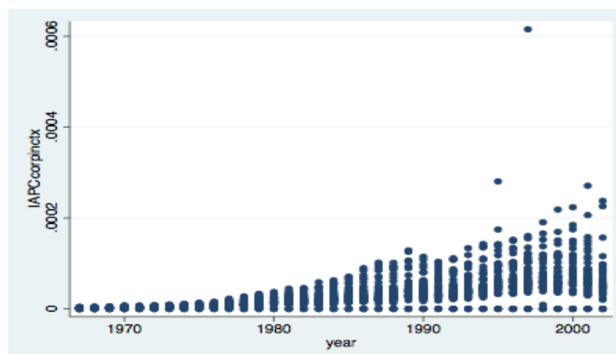
b.) twoway (scatter iapceducexp year)



c.) twoway (scatter iapcpersinctx year)



d.) twoway (scatter iapccorpinctx year)



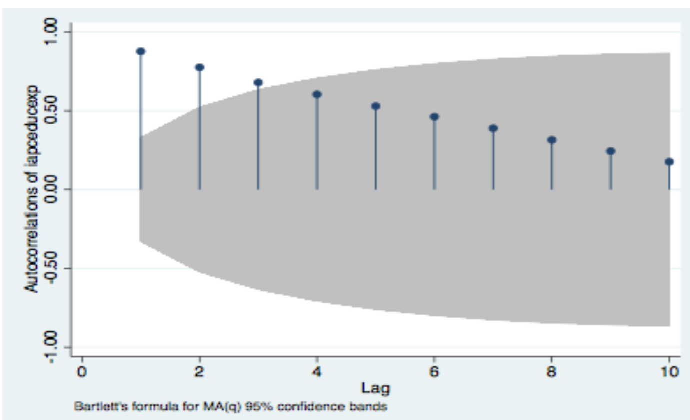
Because of the tediousness in presenting the scatterplots of all the variables, a sample of only 4 scatterplots that have proved to be representative of the whole is presented. In line with the theory, spending and tax revenues, for the most part, have been increasing across the years. There are a few outlying values, but they are unlikely to have large effects on our analyses.

Case study: Connecticut

An interesting way to go about this inquiry is to zoom in to a state and see what interpretations we can get by studying it independently. The dataset for just the state of Connecticut is taken up and scrutinized as a time series dataset. Just like in the national dataset, though not presented, the dataset from the state of Connecticut appear normal, and graphical features such as nonstationarity are still present.

Disregarding the hint from the graph that the data may be nonstationary, the data from the state of Connecticut is first approached without accounting for nonstationarity. One could see this as a test, in its own right, of the consistency of the regression, or its lack. This approach will confirm or allay the intuition that the data is nonstationary. The data is graphically tested for autocorrelation.

Graph for autocorrelation:

	The first few lags are beyond the interval estimates, suggesting that the variables are indeed autocorrelated. The subsequent regression will be run with a lag in an attempt to compensate for the autocorrelation

Theory of the regression (what variables should be included?)

Given that the variables are autocorrelated, the lag of the variables that are affected by autocorrelation will be included.

Below are the explanatory variables and the intuition behind their inclusion:

1.) educexp: It is fair to assume that the education spending in one year won't be very far away from the education spending the previous year. Thus the lag of the dependent variable will be included as an explanatory variable.

2a.) persinctx, corpinctx, salestrev: Literature posits that personal, corporate income, and sales tax are very important sources of fund for educational spending. 2b.) taxrev: Alternatively, we could substitute these specific taxes with the aggregate tax revenue and analyze its effect. However, both should not be used together, because (2a.) constitutes (2b.) and thus, we might get the issue of collinearity by including both.

3.) age517, age1824: The age demography of a state should determine educational spending. States with more people between the ages of 5 and 24, will probably spend more on education than states with more people above 65.

4.) gsp: The wealth or income of a state should potentially also have an effect on its spending on education. We can quantify this by looking at the state's "gross state production".

5.)lottrev: Finally, the lottery revenue variable should be included because it is the variable that inspired this paper. Going about to test its effect is our question.

Outreg

	(a)	(b)	(c)	(d)
	iapceducexp	iapceducexp	iapceducexp	iapceducexp
L.iapceducexp	1.422	1.508	1.571	1.629
	(5.91)**	(5.93)**	(6.98)**	(8.62)**
L.iapccorpinctx	-429.986	-249.416		
	(2.15)*	(1.00)		
L.iapcpersinctx	51.286	1.945		
	(0.71)	(0.02)		
L.iapcsalestrev	66.758	39.814		
	(0.32)	(0.19)		
L.lage517	-0.103	0.117	-0.043	0.313
	(2.12)*	(0.64)	(1.01)	(2.98)**
L.lage1824	0.102	-0.116	0.041	-0.311
	(2.06)	(0.64)	(0.94)	(2.97)**
L.iapcgsp	-17.427	-17.901	-22.009	-16.700
	(2.28)*	(2.27)*	(2.33)*	(2.11)*
L.iapclottrev	1,682.933	1,688.780	1,739.208	1,930.505
	(1.90)	(1.89)	(2.19)*	(2.95)**
lage517		-0.215		-0.393
		(1.22)		(3.47)**
lage1824		0.214		0.391
		(1.22)		(3.44)**
L.iapcttaxrev			27.637	-80.291
			(0.35)	(1.14)
_cons	0.101	0.074	0.071	0.045
	(2.99)**	(1.84)	(2.34)*	(1.73)
R2	1.00	1.00	1.00	1.00
N	32	32	32	32

* $p<0.05$; ** $p<0.01$

Personal income, corporate income, and sales taxes, are mostly insignificant in outreg models (a) and (b), when one of them appeared significant (corporate income), the coefficient had a negative relation with educational spending. This means that more corporate tax results in less education spending. However, when these different taxes are replaced with the variable ttaxrev in models (C) and (d), the negative relationship still persists with the new “total tax revenue variable” and it also has a weak statistical significance. outreg (d) seems to have the most statistically and economically significant variables, but the signs of their coefficients still don’t abide by the theory. The R-squared of the regression is nearly 100%, which is, practically speaking, unattainable. It is very likely that this regression is spurious, and this is probably because the variables are nonstationary.
Note: the “iapc” in front of the variables mean that they are inflation adjusted and presented in per capita terms.

At this point that the specification that was not adjusted for nonstationarity has proven inconsistent,

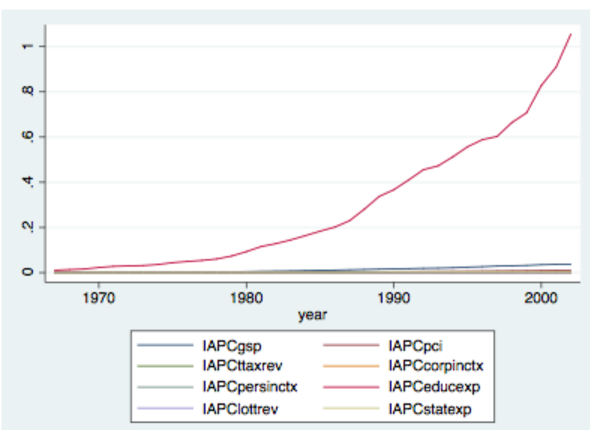
testing for stationary of outreg (d) is in order.

```
. dfuller iapceducexp

Dickey-Fuller test for unit root                   Number of obs   =        35

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             8.474            -3.682            -2.972            -2.618
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 1.0000
```

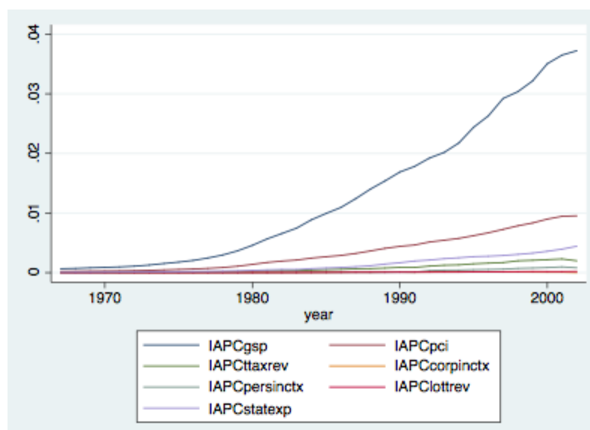
The approximate p-value for this test is 100% and the nonstationary null hypothesis is not rejected at that level. The above regression (d) is definitely nonstationary.

This position that the data is not stationary is reiterated graphically using 'tsline':

a.) tsline iapcgsp iapcpci iapcttaxrev iapccorpinctx iapcpersinctx iapceducexp iapclottrev iapcstatexp



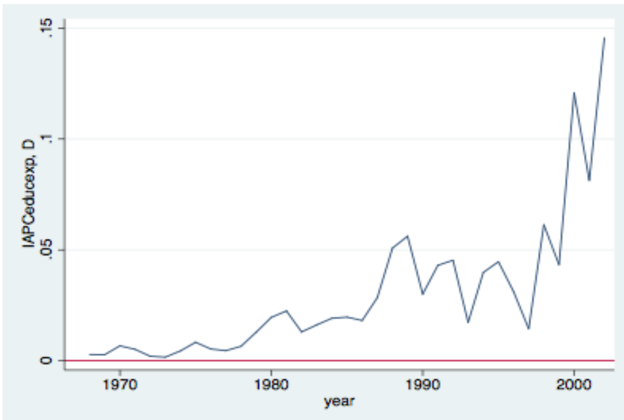
b.) tsline iapcgsp iapcpci iapcttaxrev iapccorpinctx iapcpersinctx iapclottrev iapc statexp



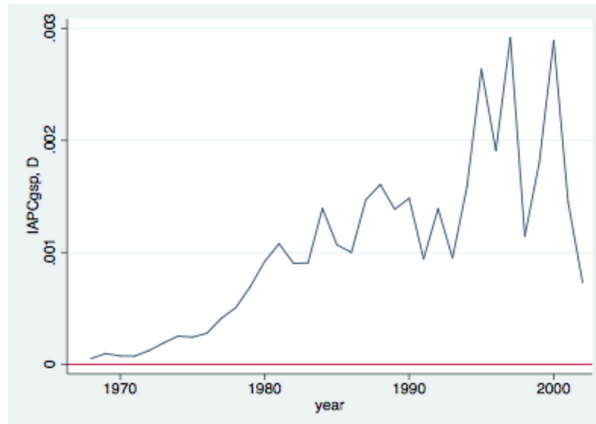
Because of the inclusion of education in graph (a), which is significantly larger than the other variables, the other variables are compressed. This makes their nonstationarity not very graphically visible. In graph (b) however, when education is removed, and the other variables spread out more, their nonstationarity become more apparent. Because the direction of these variables is related, the suspicion that the above is a spurious regression is not unfounded.

Having reached the conclusion that the data is nonstationary, both sides are differenced to get stable means. Sample diagrammatic representation of differenced variables helps determine the level of differencing that would get a stable mean.

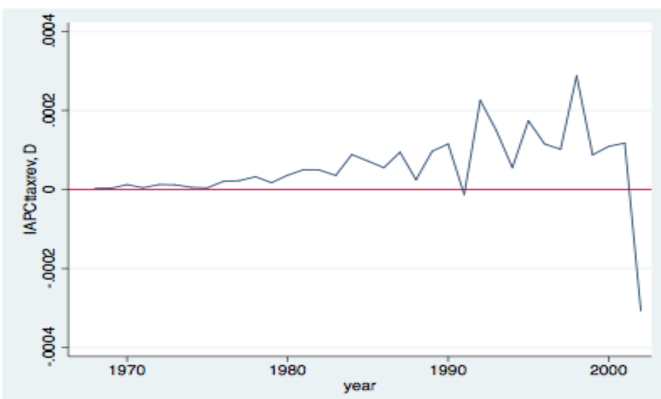
qui tsline D.iapceducexp, name(diapceducexp, replace) yline(0)



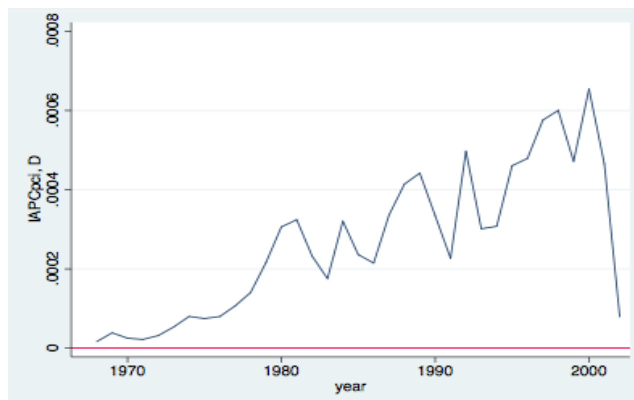
qui tsline D.iapcgsp, name(diapcgsp, replace) yline(0)



qui tsline D.iapcttaxrev, name(diapcttaxrev, replace) yline(0)



qui tsline D.iapcpci, name(diapcpci, replace) yline(0)



To take care of the issue of nonstationarity, we difference once to get a stable mean. We test again for nonstationarity, but this time with the differenced values.

```
. . dfuller D.iapceducexp, noconstant lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =        33

                               ---------- Interpolated Dickey-Fuller ----------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             2.442           -2.647            -1.950            -1.603
```

From the Augmented Dickey-Fuller test for unit root, we reject the null hypothesis of nonstationarity. Thus we dismiss the speculation that the regression from the differenced variables might still be spurious.

We test for autocorrelation with regression, but this time with regression (b), our most preferred regression. Testing for autocorrelation helps us determine if the estimated variance of our coefficient are inflated.

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
```

lags(p)	chi2	df	Prob > chi2
1	2.181	1	0.1398

H0: no serial correlation

In the Breusch-Godfrey test for autocorrelation, we cannot reject the null that there is no autocorrelation.

We have discovered that the differenced variables are neither nonstationary nor autocorrelated; we will now go ahead to run the time series regression again. This regression, having taking care of nonstationarity, will have more validity to it than the earlier regression. Just before regressing, we construct the covariance matrix in order to avoid collinearity and omitted variable bias in our regressions.

```
. corr D.iapceducexp L.D.iapceducexp L.D.iapclottrev L.D.iapcpersinctx L.D.iapccorpinctx L.D
> apcttaxrev L.D.iapcgsp L.D.iapcsalestrev L.D.lage517 L.D.lage1824
(obs=31)
```

	D. iapced~p	LD. iapced~p	LD. iapclo~v	LD. iapcpe~x	LD. iapcco~x	LD. iapctt~v	LD. iapcgsp	LD. iapcsa~v
iapceducexp								
D1.	**1.0000**							
LD.	**0.7017**	**1.0000**						
iapclottrev								
LD.	**0.2183**	**-0.0585**	**1.0000**					
iapcpersin~x								
LD.	**0.3541**	**0.4874**	**0.1552**	**1.0000**				
iapccorpin~x								
LD.	**-0.3113**	**-0.1508**	**-0.1433**	**0.1393**	**1.0000**			
iapcttaxrev								
LD.	**0.3220**	**0.5129**	**0.4015**	**0.8402**	**0.1989**	**1.0000**		
iapcgsp								
LD.	**0.5795**	**0.6691**	**0.5337**	**0.3768**	**-0.2594**	**0.5392**	**1.0000**	
iapcsalest~v								
LD.	**0.3958**	**0.3646**	**0.4338**	**-0.0567**	**-0.3214**	**0.3928**	**0.4822**	**1.0000**
lage517								
LD.	**0.1238**	**0.0442**	**0.1005**	**0.0695**	**-0.1808**	**0.1296**	**0.0747**	**0.2803**
lage1824								
LD.	**0.1169**	**0.0371**	**0.0900**	**0.0584**	**-0.1787**	**0.1167**	**0.0644**	**0.2757**

	LD. lage517	LD. lage1824
lage517		
LD.	**1.0000**	
lage1824		
LD.	**0.9997**	**1.0000**

From the table we can see just two almost perfect-correlated variables.

ρ(L.D.lage1824)(L.D.lage517) = 0.9997

Therefore, we won't include both in the same regression to avoid collinearity.

Other variables with relative high but not perfect correlation all have to be put into the regression, otherwise we will have a problem of omitted variable bias.

(A) (B) (C)

```
. outreg, merge
```

	D.iapceducexp	D.iapceducexp	D.iapceducexp
LD.iapceducexp	**0.747**	**0.785**	**0.801**
	(2.94)**	**(5.02)****	**(5.34)****
LD.iapclottrev	**2,130.409**	**2,232.616**	**2,258.421**
	(2.85)**	**(4.21)****	**(4.36)****
LD.iapcpersinctx	**1,283.691**	**1,269.534**	**1,252.206**
	(5.14)**	**(5.41)****	**(5.50)****
LD.iapccorpinctx	**1,239.264**	**1,216.861**	**1,215.583**
	(3.06)**	**(3.18)****	**(3.23)****
LD.iapcttaxrev	**-1,253.026**	**-1,241.230**	**-1,229.216**
	(5.42)**	**(5.67)****	**(5.75)****
LD.iapcgsp	**1.651**		
	(0.20)		
LD.iapcsalestrev	**1,607.460**	**1,587.897**	**1,556.552**
	(4.64)**	**(4.87)****	**(4.98)****
LD.lage517	**-0.001**	**-0.001**	
	(0.40)	**(0.42)**	
_cons	**0.007**	**0.007**	**0.007**
	(1.37)	**(1.51)**	**(1.49)**
R2	**0.83**	**0.83**	**0.83**
N	**34**	**34**	**34**

* p<0.05; ** p<0.01

Model c is the best specification; all its variables are statistically and economically significant at 1% significance level. It has an R2 of 83%. Model A and B have the same R2 as Model C, but they have variables (L.D.lage517 & LD.iapcgsp) that are not statistically and economically significant. The interpretation of our coefficients from Model C is that an increase of one dollar in last years education spending will increase this year's education by $0.801. The variables, which the theory states as important (Income, corporate and sales tax), are also supported by the regression. An increase of one dollar in *LD.iapcpersinctx* will increase education spending by $1252.206, whilst an increase of one dollar for the variables *LD.iapccorpinctx* and *LD.iapcsalestrev* will increase education spending by $1215.583 and $1556.552 respectively. The outcome for LD.iapclottrev (Lottery revenue) is that when it increases by a dollar, education spending will increase by $2,258. It is worthy to note that iapcttaxrev (total tax revenue) has a negative effect on education spending because as it increases by a dollar, education spending falls by about $1229.

Though the correlation matrix tells us that LD.lage517 and LD.lage1824 are almost perfectly collinear and should not be included together. From our regression we realized that neither is actually needed because of the lack of statistical significance. One thing looks out of place, the rate of increase seems too large, and will a dollar increase in sales tax actually lead to $1556 increase education spending? This seems very unlikely. After checking the data thoroughly, our best guess is that these variables were scaled by either a 10,000x or 1000x less. This means that the dependent variable is either 10,000x or 1000x more than the explanatory variables. We use either 10,000x or 1000x because both seem plausible, and given that the dataset has no description/key we are left to guess for ourselves. Except changing where the decimal point should be placed, this scaling problem does not affect our coefficient in other way. The real interpretation of the sales tax variable is that when it increases by a dollar, education spending will increase by 0.15 (assuming the scale is 10,000x), and this transformation applies to all the variables except the constant and the lag of education.

Analysis/ Results

The scaling error does not affect our coefficient or standard error beyond the extent of pushing back the decimal place. This problem is one that we can easily resolve. It is important to pay some attention to the variable L.D.iapcttaxrev (total tax revenue). We began this process by assuming that it is collinear with the other tax variables (personal, corporate, and sales tax), and thus should not be included in the regression. However, the correlation matrix showed that it isn't collinear; rather, omitting it will result to an omitted variable bias, because of the high level of correlation. The discovery extends up to its interpretation, whilst income and sales taxes showed a positive relationship with education spending, total tax showed a negative relationship with education spending. A rationale behind this result is that taxes as a whole has the effect of reducing people and corporate's disposable income, income and sales taxes are unique in the sense that they are transferred to the government who subsequently spend it on education, thus still improving education spending, however, other forms of taxation reduce people's purchasing power to spend on education themselves, and the government do not spend these other taxes on education, thus they have the effect of reducing spending on education (the negative relationship)

Conclusion:

The result from the best model (c) of our regressions, which has its variables differenced to correct for nonstationarity, is that taxation significantly reduces spending in education whilst lottery revenue, income and sales taxes enormously increases spending on education. Quantifying this observation, an increase of one dollar in the variables LD.iapclottrev, *LD.iapcpersinctx, LD.iapccorpinctx* and *LD.iapcsalestrev* will increase education spending by $2,258, $1252.206, $1215.583 and $1556.552 respectively. Total tax revenue- iapcttaxrev -has a negative effect on education spending because as it increases by a dollar, education spending falls by about $1229. These results support the literature that incomes and sales revenues are the bedrock of education spending. It also vindicates the proponents of the lottery scheme, who argue that revenue from lottery can be used to improve education. Whether the above information is enough to sway states to approve a state lottery scheme is a normative question not an econometrics question. Another discovery this paper makes is that reducing other taxes is likely to lead to increase education spending by individuals and corporates.

Next Step

In our bid to expand the frontiers of knowledge, we should not rest on our laurel here. It will be useful to run this regression on a few other randomly selected states, and then using a fixed effect model, run the regression for the entire country across the years that data is available. With this result, we can compare if the regression from these states are consistent with the regression from the country as a whole. This would go a long way to determine, given the present policies if lottery revenue really determines education spending across the US.