State University of New York at Albany

Omar Villalba

Forecasting Local Sales Taxes

Introduction

Forecasting is a requirement of any government which attempts to prepare its current and future budgetary agreements. For most government, some version of a sales tax forms a large part of their revenues, for example, in the State of New York, it makes up 18% of their income(Cuomo & Mujica, 2017), but this is increasingly becoming more important for local governments, who have historically been more dependant on property taxes rather than sales tax(Afonso, 2013). However, these governments have not usually forecasted their future sales tax using econometric models but instead have used naive judgemental methods to project their expected future earnings (Kong, 2007;Beckett-Camarata,2006). Given the increasing share of revenue that comes from Sales Taxes, and the future base erosions that could bring substantial fiscal stress to the local governments that depend on them (Russo, 2010), proper forecasting of this revenue source if of great importance to local governments . For this paper, utilizing panel data from 57 New York state counties, a comparison is made for several Panel Data models to use in forecasting the Sales Tax revenue of these counties. Regressions are fitted up to a 2013 and an in sample forecast is produced. Lastly, the Mean Absolute Percentage Error(MAPE) is taken for the in sample forecasted years and utilized to compare the different models and determine their quality.

New York Sales & Use Tax

The New York Sales Tax affects most sales of tangible personal property except those that are specifically exempt as well as several specific services. The Use tax is a complement to the state sales tax and it is imposed on sales brought into the state. Currently, the Sales Tax affects(Cuomo & Mujica, 2017; DiNapoli et al., 2015):

- Tangible personal property
- Specific services, including gas, electricity, refrigeration and steam and telephone
- Food and beverages sold at taverns, caterers and restaurants
- Hotel occupancy
- Some admission charges and dues

The Use tax complements the Sales Tax by requiring the purchaser of a good or service to pay a tax if the company selling the product has not done so. This is due to the supreme court ruling in *National Bellas Hess v. Department of Revenue* and *Quill Corp. v. North Dakota.* In these rulings the Supreme Court determined that states could not ask corporations without a "nexus" in the state to pay sales tax. Because of this, the Use tax is collected to make up for lost revenue.

For counties, New York State Law allows counties to impose local sales tax of at most 3%, although they can ask, and have consistently gotten, the state government to allow a tax rate higher than that, with the most common rate being 4% (DiNapoli et al., 2015). Cities are also allowed to impose a sales tax, but for the current analysis, the focus will be on counties.

Data Analysis

As found by previous research(Kong, 2007), obtaining data for each county has proven to be a serious difficulty. Although compared to Kong, a larger pool of data that has been obtained, the data for each specific county is still too small to do any form of time series analysis that will not be seriously biased or inconsistent. While ultimately, the goal of any research into forecasting local sales tax is for local governments to move away from their current guess estimation into a more rigorous and effective forecasting methodology, the lack of data makes this much more difficult than it would otherwise. In order to increase forecasting accuracy, either better data must be made available for counties to work with, or other statistical methods, such as panel data, must be utilized in order to work with the data that does exist. As mentioned before, panel data analysis will be alternative pursued, given the data that was obtainable. The dependent and independent variables can be seen in the table below, as well as a summary of their basic characteristics. It is important to note that per capita personal income is in tens of thousands and the others are in millions, excluding tax rates. Furthermore, in the regressions that follow, differenced, lagged and log of the variables are also utilized. They're identified by a d, L^x or log in front respectively.

	Sales Tax	*Employment*	*Population*	*Personal Income*	*Per Capita Personal Income*	*Tax Rates*
Mean	96.80436	.092288	.1945844	7.914471	3.20547	3.659357
Std. Deviation	191.9157	.1465418	.3002541	15.98181	1.04842	.5424421
Minimum	0	.002185	.004706	.108675	1.6191	0
Maximum	1297.661	.759899	1.502342	105.8607	9.3229	5.25

Regression Preparation

For all models, the independent variables have been chosen due to their relationship with the economic health of the counties in question, and naturally with the effect that a good economic status should have in sales tax revenue. For variable selection, a stepwise regression method is utilized. Regressions are fitted from 1995 to 2013, then forecasts are made for 2014 and 2015. Finally, the MAPE is taken for these last 2 years. Variables are added one by one, starting from a regression with just 1 variable, checking every variable and its different forms, and selecting based on which variable has the best effect on the MAPE, until some form of each variable is included in the regression, either its difference, log or regular form. This is done even if it adversely affects the MAPE, due to two reasons. First, because economic theory justifies their inclusion, and secondly, to avoid a case of overfitting a regression. The process is done to sales tax, its log and its difference and the regression with the best MAPE is selected for each of

the models that have been prepared. Lastly, the Mean Square Error(MSE) is also added for reference for all the models.

Regression Analysis

Pooled OLS

The first model to be utilized is pooled ordinary least squares(POLS). A POLS model is equivalent to running Ordinary Least Squares(OLS) for all observations running in both indexes, i and t. It can be written as:

$$y_{i,t} = x_{i,t}\beta + e_{i,t}$$

where $x_{i,t}$ is a 1xK vector including the independent variables and possibly a constant and β is a Kx1 vector including the coefficient. The results of the POLS model can be seen below:

$$d(sales) = \alpha + \beta_1 d(personalincome) + \beta_2 d(population) + \beta_3 employment + \beta_4 taxrates + \beta_5 d(percapitapersonalincome) + \beta_6 log(usretailsales)$$

	Constant	*d(personalincome)*	*d(population)*	*employment*	*taxrates*	*d(percapitapersonalincome)*	*log(usretailsales)*
Coefficient	-5.21866** (2.342172)	4.036255*** (.668039)	-60.011 (171.7189)	26.03006*** (3.193025)	1.10162* (.6204688)	-1.92556 (4.14781)	.9735571 (1.681858)
MAPE	3.709055 (.3026703)						
MSE	134.12 (95.22669)						

*Standard Error in parenthesis, * p<0.10, ** p<0.05, *** p <0.01*

Random Effect

The second model is a Random Effects model. In panel data analysis, models are written in the form $y_{i,t} = x_{i,t}\beta + v_{i,t}$. In this model(and the ones that follow) just as in POLS, $x_{i,t}$ is a vector including the independent variables and possibly a constant, β is a vector including the coefficients and $v_{i,t} = c_i + u_{i,t}$. This is a specific composite error where $u_{i,t}$ are residuals and c_i is the individual heterogeneity. This last random variable captures the differences between individuals and as long as it is in the equation, variables that change across individuals only cannot be included. This unobserved heterogeneity in Random Effects models is assumed to be uncorrelated with $x_{i,t}$ and due to this special form of residuals in this equation the variance covariance matrix has a particular form. In order to take this into account, Random Effects uses a Generalized Least Squares regression utilizing the variance covariance matrix previously mentioned. The results for this model are as follows:

$$d(sales) = \alpha + \beta_1 log(personalincome) + \beta_2 d(population) + \beta_3 employment$$
$$+ \beta_4 d(taxrates) + \beta_5 percapitapersonalincome + \beta_6 d(usretailsales)$$

	Constant	*log(personalincome)*	*d(population)*	*employment*	*d(taxrates)*	*percapitapersonalincome*	*d(usretailsales)*
Coefficient	-.6896145 (1.179003)	-.4425409 (.4879875)	-155.3609 (174.5297)	45.99958*** (3.787259)	.6078865 (1.552325)	-.1954317 (.4121095)	6.877133** (2.145535)
MAPE	3.901288 (.3817678)						
MSE	133.9736 (89.69441)						

*Standard Error in parenthesis, * p<0.10, ** p<0.05, *** p <0.01*

Fixed Effect

The third model is a Fixed Effects regression. This regression takes the mean equation of each individual and subtracts it from each observations, therefore:

$$(y_{i,t} - \overline{y}_i) = (x_{i,t} - \overline{x}_i)\beta + c_i - c_i + (u_{i,t} - \overline{u}_i)$$

$$\widehat{y}_{i,t} = \widehat{x}_{i,t}\beta + \widehat{u}_{i,t}$$

As can be noticed from the equation above, the individual heterogeneity is removed from the new equation. Finally, a pooled OLS regression is run over the new equations. The results of this model can be seen as follows:

$$d(sales) = \alpha + \beta_1 d(personalincome) + \beta_2 population + \beta_3 employment + \beta_4 taxrates + \beta_5 percapitapersonalincome + \beta_6 d(usretailsales)$$

	Constant	*d(personalincome)*	*population*	*employment*	*taxrates*	*percapitapersonalincome*	*d(usretailsales)*
Coefficient	-13.08556 (8.263169)	3.99658*** (.5328281)	-103.4627** (47.56493)	312.1884*** (68.46644)	2.027063** (.9515128)	-.4582961 .0627424	4.343267 ** (2.144649)
MAPE	17.48199 (2.772942)						
MSE	276.3461 (187.9503)						

*Standard Error in parenthesis, * p<0.10, ** p<0.05, *** p <0.01*

First Difference

The fourth model is the first difference estimator. In this model the lagged equation is subtracted from the original one, therefore:

$$(y_{i,t} - y_{i,t-1}) = (x_{i,t} - x_{i,t-1})\beta + (c_i - c_i) + (u_{i,t} - u_{i,t-1})$$

$$\Delta y_{i,t} = \Delta x_{i,t}\beta + \Delta u_{i,t}$$

And then pooled OLS is run over the new equations. When choosing between Fixed effects and first difference, the key issue is the error term. If the error term is serially correlated, then fixed effects is more efficient, whereas if it follows a random walk, then the first difference is more efficient. The results of this model are as follows:

$$d(sales) = \alpha + \beta_1 d(personalincome) + \beta_2 d(population) + \beta_3 d(employment) + \beta_4 d(taxrates) + \beta_5 d(percapitapersonalincome) + \beta_6 d(usretailsales)$$

	Constant	*d(personalincome)*	*d(population)*	*d(employment)*	*d(taxrates)*	*d(percapitapersonalincome)*	*d(usretailsales)*
Coefficient	2.809563*** (.5303705)	7.179588*** (.500391)	421.5022*** (162.465)	797.8241*** (116.574)	.1613082 (1.53828)	-15.6673*** (3.958218)	.4507405 (2.212306)
MAPE	5.045257 (1.231544)						
MSE	145.9662 (100.333)						

*Standard Error in parenthesis, * p<0.10, ** p<0.05, *** p <0.01*

Between Estimator

The fifth model is the between estimator, which is simply OLS over the averaged equation:

$$\overline{y}_i = \overline{x}_i\beta + \overline{u}_i$$

Since we are using averages, variables that change across time but not individuals cannot be included(since that would just be a constant), which is why the retail sales variable is not included in this regression. The results of this regression can be seen below.

$$d(sales) = \alpha + \beta_1 d(personalincome) + \beta_2 population + \beta_3 d(employment)$$
$$+ \beta_4 taxrates + \beta_5 log(percapitapersonalincome)$$

	Constant	*d(personalincome)*	*population*	*d(employment)*	*taxrates*	*log(percapitapersonalincome)*
Coefficient	.1201725*** (2.103616)	-3.164156*** (.9527898)	24.30687*** (1.656343)	1974.061*** (263.5832)	.2929853 (.422849)	-1.596652 (1.165718)
MAPE	4.018521 (.6461808)					
MSE	183.1162 (122.9396)					

*Standard Error in parenthesis, * p<0.10, ** p<0.05, *** p <0.01*

Arellano-Bond/Blundell estimator

And the last model is System Generalized Method of Moments Arellano-Bond/Blundell estimator. This model is different from the others in that it includes lags of the dependent variable as a predictor. Because this is included, there is likely correlation between that predictor & the residual, which violates an assumption required for other models. In order to deal with this, the model takes first difference of the of the variables, including the lag. Hence:

$$(y_{i,t} - y_{i,t-1}) = (x_{i,t} - x_{i,t-1})\beta + (y_{i,t-1} - y_{i,t-2})\alpha + c_i - c_i + (u_{i,t} - u_{i,t-1})$$

$$\Delta y_{i,t} = \Delta x_{i,t}\beta + \Delta y_{i,t-1}\alpha + \Delta u_{i,t}$$

A system instrumental variable regression is then run on this equation. Because there can still be correlation in this equation between the differenced lagged dependent variable and the differenced lagged residual, deeper lags of the level dependent variable and of the differenced dependent variable are used as instruments for it. The results from this model are then as follows:

$$\begin{aligned} d(sales) = {} & \alpha + \beta_1 d(personalincome) + \beta_2 log(population) + \beta_3 log(employment) \\ & + \beta_4 d(taxrates) + \beta_5 d(percapitapersonalincome) + \beta_6 d(usretailsales) \\ & + \delta L^1 d(sales) \end{aligned}$$

	Constant	*d(personalincome)*	*log(population)*	*log(employment)*	*d(taxrates)*	*d(percapitapersonalincome)*	*d(usretailsales)*	*L^1d(sales)*
Coefficient	.262268*** (.0692119)	.0001463 (.0054048)	-.2822155*** (.0743865)	.277266*** (.068145)	-.0036327 (.0137029)	.0532847* (.0319167)	.064955*** (.0138019)	.990696*** (.0090569)
MAPE	3.388062 (.30217)							
MSE	134.9721 (81.75026)							

*Standard Error in parenthesis, * p<0.10, ** p<0.05, *** p <0.01*

Conclusion

As can be seen from the results, the Arellano-Bond/Blundell model is the best of according to the MAPE, while the Random Effect is the best by the MSE. However, the difference between most estimators based on the MAPE measure is relatively small outside of Fixed Effects, and in the case of MSE it is small outside of Fixed Effects and Between Estimator. For this reason, in a case by case basis, it would be reasonable to expect the best forecasting model is not necessarily the one with best MAPE or MSE. Nonetheless, if choosing a model, using the MAPE as a forecast measure is preferable for 2 reasons. First, it is easy to comprehend, which can be very important for people without statistical preparation(like government officials), and secondly, unlikely MSE it does not heavily weight outliers. While the MAPE does have its own issues (Toffalis, 2014), this are either not present for the models presented (cannot have a 0 realized values) or not substantially problematic (bias towards under forecasting) for the presented forecasting case. Lastly, in this case, both measures show relatively similar results when analyzing the quality of the models, which makes the simpler to understand MAPE more attractive.

Ultimately, besides panel data modeling, attempts to obtain good enough data for time series forecasting models should nonetheless continue. This is not just because this models could provide better forecasts, but because ultimately, local governments look for simpler, more easily automated models(Williams & Kavanagh, 2016). While to some extent panel data forecasting could be made accessible to local governments, particularly with the currently available

statistical software(and due to the POLS model being effective, which is a simple model), it would be very unlikely that panel data analysis can be made simpler than time series. Nonetheless, it is certainly possible that for larger counties, the capacity to pay for panel data modeling is worth it given their income from Sales Taxes, for this reason a comparison between time series model and panel data models for forecasting county level Sales Tax is a topic to be pursued for future research.

References

1. Williams, D. W., & Kavanagh, S. C. (2016). LOCAL GOVERNMENT REVENUE FORECASTING METHODS: COMPETITION AND COMPARISON. *Journal of Public Budgeting, Accounting & Financial Management*, *28*(4).
2. Kong, D. (2007). Local Government Revenue Forecasting: The California County Experience. *Journal of Public Budgeting, Accounting & Financial Management*, *19*(2), 178.
3. Afonso, W. B. (2013). Diversification toward stability? The effect of local sales taxes on own source revenue. *Journal of Public Budgeting, Accounting & Financial Management*, *25*(4), 649.
4. Beckett-Camarata, J. (2006). Revenue forecasting accuracy in Ohio local governments. *Journal of Public Budgeting, Accounting & Financial Management*, *18*(1), 77.
5. Russo, B. (2010, 07). Is past prologue? Prospects for state and local sales tax bases. *Applied Economics, 42*(18), 2261-2274. doi:10.1080/00036840701858000
6. Cuomo, Andrew, M., Mujica, Robert, F. (2017). *FY 2018 Economic & Revenue Outlook*. New York: New York State Government
7. DiNapoli, Thomas, P., State Comptroller (2015). *Local Government Sales Taxes in New York State: 2015 Update*. New York: Division of Local Government and School Accountability & Office of the New York State Comptroller
8. Tofallis, C. (2014, 12). Erratum: A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society, 66*(3), 524-524. doi:10.1057/jors.2014.124

Data Sources

1. **New York County Level Sales Tax :** *New York State Government, Office of Information and Technology Services*

   https://data.ny.gov/Government-Finance/State-and-Local-Sales-Tax-Distributions-Beginning-/5g2s-tnb7/data

2. **New York County Level Population:** *New York State Government, Office of Information and Technology Services*

   https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k/data

3. **New York County Level Employment:** *US Department of Labor, Bureau of Labor Statistics*

   https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k/data

4. **New York County Level Tax Rates:** *New York State Government, Department of Taxation and Finance*

   https://www.tax.ny.gov/pdf/publications/sales/pub718a.pdf

5. **New York County Personal Income and Per Capita Personal Income:***US Department of Commerce, Bureau of Economic Research*

   https://www.bea.gov/iTable/index_nipa.cfm

6. **United States Retail Sales:** *US Federal Government, United States Census Bureau*

   https://www.census.gov/retail/index.html

Stata Code

Fix variables and label county dummy

```
destring employment, replace

label define countylabels 1 "Albany" 2 "Allegany" 3 "Broome" 4 "Cattaraugus" 5 "Cayuga" 6 "Chautauqua" 7 "Chemung" 8 "Chenango" 9 "Clinton" 10 "Columbia" 11 "Cortland" 12 "Delaware" 13 "Dutchess" 14 "Erie"  15 "Essex" 16 "Franklin" 17 "Fulton" 18 "Genesee" 19 "Greene" 20 "Hamilton" 21 "Herkimer" 22 "Jefferson" 23 "Lewis" 24 "Livingston" 25 "Madison" 26 "Monroe" 27 "Montgomery" 28 "Nassau" 29 "Niagara" 30 "Onandaga" 31 "Oneida" 32 "Ontario" 33 "Orange" 34 "Orleans" 35 "Oswego" 36 "Otsego" 37 "Putnam" 38 "Rensselaer" 39 "Rockland" 40 "Saratoga" 41 "Schenectady" 42 "Schohaire" 43 "Schuyler" 44 "Seneca" 45 "St. Lawrence" 46 "Steuben" 47 "Suffolk" 48 "Sullivan" 49 "Tioga" 50 "Tompkins" 51 "Ulster"  52 "Warren" 53 "Washington" 54 "Wayne" 55 "Westchester" 56 "Wyoming" 57 "Yates"

label values counties countylabels
```

Set Panel Data

```
xtset counties year
```

Generating variables

Log Variables

```
generate logsales = log(sales)

generate logpopulation = log(population)

generate logpersonalincome = log(personalincome)

generate logpercapitapersonalincome = log(percapitapersonalincome)

generate logemployment = log(employment)

generate logusretailsales = log(usretailsales)
```

Differences Variables

generate dsales = sales-L.sales

generate dpopulation = population-L.population

generate dpersonalincome =personalincome - L.personalincome

generate dpercapitapersonalincome = percapitapersonalincome - L.percapitapersonalincome

generate demployment = employment - L.employment

generate dusretailsales = usretailsales - L.usretailsales

generate dtaxrates = taxrates - L.taxrates

Scaled variables

generate percapitapersonalincomem = percapitapersonalincome/10000

generate populationm = population/1000000

generate personalincomem = personalincome/1000000

generate employmentm = employment/1000000

generate usretailsalesm = usretailsales/1000000

generate dpercapitapersonalincomem = (percapitapersonalincome -

L.percapitapersonalincome)/10000

generate dpopulationm = (population-L.population)/1000000

generate dpersonalincomem = (personalincome - L.personalincome)/1000000

generate demploymentm = (employment - L.employment)/1000000

generate dusretailsalesm = (usretailsales - L.usretailsales)/1000000

Regressions Analysis
Note: Only the code for the final results are included for reference. This is due to the specific stepwise regression process used not being automatable. In order to replicate this process, remove all variables in the first to line but the first, then repeat the blocks of code adding and removing variables depending on the MAPE result.

Pooled OLS regression

```
regress dsalesm dpersonalincomem dpopulationm employmentm taxrates
dpercapitapersonalincome logusretailsalesm if tin(1995,2013)
predict PredPoolDiff
generate poolSales=L.salesm + PredPoolDiff
generate poolSalesErr=abs(100*(salesm-poolSales)/salesm)
generate poolSalesMSE=(salesm-poolSales)^2
mean(poolSalesMSE) if tin(2014,2015)
mean(poolSalesErr) if tin(2014,2015)
drop PredPoolDiff
drop poolSales
drop poolSalesErr
```

Random Effects regression

```
xtreg dsalesm dtaxrates employmentm logpersonalincomem percapitapersonalincomem
dpopulationm dusretailsalesm if tin(1995,2013), re
predict PredPoolDiff
generate poolSales=L.salesm + PredPoolDiff
generate poolSalesErr=abs(100*(salesm-poolSales)/salesm)
generate poolSalesMSE=(salesm-poolSales)^2
mean(poolSalesMSE) if tin(2014,2015)
mean(poolSalesErr) if tin(2014,2015)
drop PredPoolDiff
drop poolSales
drop poolSalesErr
```

Fixed Effect Regression

```
xtreg dsalesm taxrates percapitapersonalincomem dpersonalincomem populationm
dusretailsalesm employmentm if tin(1995,2013), fe
predict PredPoolDiff
generate poolSales=L.salesm + PredPoolDiff
generate poolSalesErr=abs(100*(salesm-poolSales)/salesm)
generate poolSalesMSE=(salesm-poolSales)^2
```

```
mean(poolSalesMSE) if tin(2014,2015)
mean(poolSalesErr) if tin(2014,2015)
drop PredPoolDiff
drop poolSales
```

First Difference Estimator

```
regress dsalesm dpercapitapersonalincome dusretailsalesm demploymentm dpersonalincomem
dtaxrates dpopulationm if tin(1995,2013)
predict PredPoolDiff
generate poolSales=L.salesm + PredPoolDiff
generate poolSalesErr=abs(100*(salesm-poolSales)/salesm)
generate poolSalesMSE=(salesm-poolSales)^2
mean(poolSalesMSE) if tin(2014,2015)
mean(poolSalesErr) if tin(2014,2015)
drop PredPoolDiff
drop poolSales
drop poolSalesErr
```

Between Effects Regression

```
xtreg dsalesm dpersonalincomem demploymentm populationm taxrates
logpercapitapersonalincome if tin(1995,2013), be
predict PredPoolDiff
generate poolSales=L.salesm + PredPoolDiff
generate poolSalesErr=abs(100*(salesm-poolSales)/salesm)
generate poolSalesMSE=(salesm-poolSales)^2
mean(poolSalesMSE) if tin(2014,2015)
mean(poolSalesErr) if tin(2014,2015)
drop PredPoolDiff
drop poolSales
drop poolSalesErr
```

Arellano-Bond/Blundell System GMM estimator

```
xtdpdsys dsalesm dpersonalincomem logpopulationm logemployment dtaxrates
dpercapitapersonalincomem dusretailsales if tin(1995,2013), lags(1) artests(2)
predict PredPoolDiff
generate poolSales=L.salesm + PredPoolDiff
generate poolSalesErr=abs(100*(salesm-poolSales)/salesm)
generate poolSalesMSE=(salesm-poolSales)^2
mean(poolSalesMSE) if tin(2014,2015)
```

```
mean(poolSalesErr) if tin(2014,2015)
drop PredPoolDiff
drop poolSales
drop poolSalesErr
```