Econometrics with YouTube Data
A Time Series Teaching Case

Sebastian Wai
Sebastian.Wai@usd.edu; https://www.youtube.com/@sebastianwaiecon

UNIVERSITY OF SOUTH DAKOTA

Motivation

- Time series is a difficult subject to teach in econometrics courses. It is conceptually more difficult than cross-sectional analysis, causality is difficult to establish, and there can be a lot of math on the theory side. Yet, students are likely to encounter time series data in the business world.
- We can either just scratch the surface or go deep into the math – there is no real middle ground to take. My approach is to go on the lighter side, dedicating two weeks of the course. Hansen (2017)[1] provides a good framework for what to cover.
- Existing time series exercises typically use constructed datasets (Prince 2018 [2]) or macroeconomic data (Wooldridge 2020 [3]), which are not reflective of the data students will see out in the world.

Background of the Case

- My contribution is to use data from my own YouTube channel for a summative case in class. Since 2016, I have posted tutorials and lectures, mostly from my online courses. The Stata tutorials are my most popular videos. The case uses datapoints from August 2016 to March 2022.
- The main research question: does frequent uploading lead to more views? This is the conventional wisdom on YouTube, but does that hold up in this case? Students should already be familiar with YouTube as a platform and the results are potentially actionable. Advanced undergraduates are the target audience, but I also include some adjustments for introductory or graduate-level students.

Regression Table

Example of a formatted table of results:

	Dependent variable:	
	Views	
	(1)	(2)
Last.Video	1.263^{***}	-0.400^{***}
	(0.141)	(0.065)
t		1.048^{***}
		(0.030)
t2		-0.0002^{***}
		(0.00001)
Month Dummies	No	Yes
Observations	2,010	2,010
Adjusted R^2	0.038	0.839
Note:	*p<0.1; **p<0.05; ***p<0.01	

References

[1] Bruce E. Hansen. Time series econometrics for the 21st century. *The Journal of Economic Education*, 48(3):137–145, 2017.

[2] Jeffrey T Prince. *Predictive Analytics for Business Strategy*. McGraw Hill, 1st ed. edition, 2018.

[3] Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage, 7th ed. edition, 2020.

Acknowledgements

Thank you to my students, who tested this case in class and experienced my previous attempts to teach time series. I thank former students in Business Econometrics at Indiana University and Applied Operational Analytics students at the University of South Dakota.

Case Setup and Variables

Your client is the owner of a YouTube channel featuring educational videos on a variety of academic topics. The client initially created the channel for the benefit of his students, but the channel grew in popularity over time. Between the start in August 2016 and March 2022, he has published 186 videos. The owner of the channel wants to know the importance of regular video uploads to his channel. Specifically, do views drop if it has been a long time since the last video was published?

Variable	Meaning
Views	Daily video views
Subscribers	Daily net change in subscribers
Published	Number of videos published that day
Last.Video	Days since the last video was published

Questions and Teaching Notes

(a) Generate a line graph of views over time. Are any trends or seasonality apparent?

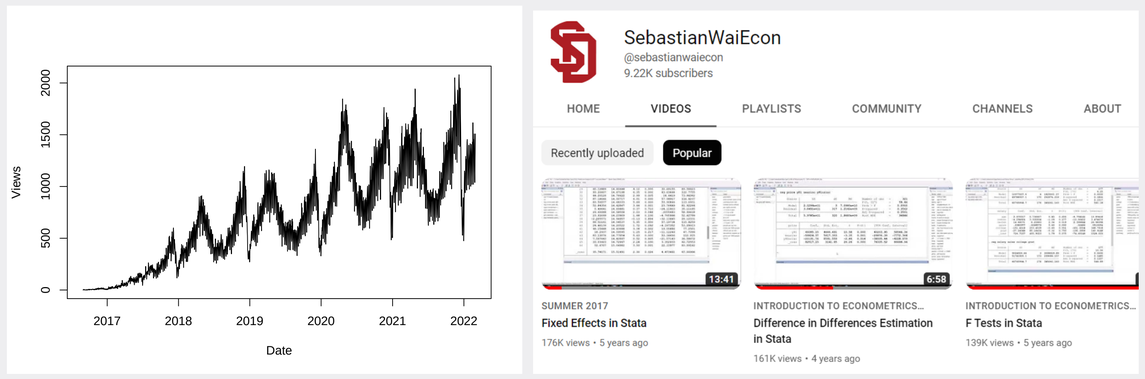

Figure 1: Line graph of views over time (left); A screenshot of the YouTube channel page (right)

To build the graph, students need to create a time variable. More advanced students could convert the date into a usable format in R or Stata. There is clear seasonality consistent with the educational content of the YouTube channel. Views rise through the Spring semester and drop off over the summer, before rising in the fall and dropping at the winter break.

(b) Calculate the total views and subscribers to date.

Total views: 1,420,750. Total subscribers: 6801. This questions gives students some practice with basic data manipulation and calculations.

(c) Run a simple regression to answer the channel owner's main question. Is this result surprising?

The estimated equation (1) shows a positive relationship between $Last.Video$ and $Views$, suggesting views increase when it has been a long time since the last upload. This may be surprising because it contradicts the conventional wisdom on YouTube. I have included the $Last.Video$ variable in the dataset, but graduate instructors could delete it and have students create it on their own using the $Published$ variable.

(d) Add a trend and monthly seasonality to your regression. Describe the seasonality present in the data and give an intuitive explanation.

Students will now have to generate a time variable and, if desired, convert the month variable to numbers. More advanced students might implement a quadratic trend. Introductory statistics instructors may want to end here.

(e) Based on your last regression, how does your answer to the main question change? Why do you think this happened?

The results have flipped from part (c). This aligns with the idea that infrequent uploads decreases views, but the effect is small. This question offers an opportunity to discuss possible omitted variables in the original model. Students can think through the omitted variable bias. A plausible explanation is that upload frequency is also seasonal and exhibits trends over time. We can also see that in August, when views are very low, $Last.Video$ is also low, again leading to positive bias. It could be that August sees more uploads in preparation for the start of the academic year. Students may also notice the R-squared greatly increases from the previous regression. This is an opportunity to explain that adding trends and seasonality often inflate the R-squared in a regression.

(f) Test for the presence of unit roots in the final regression.

This is a more advanced question, suitable for an upper-level undergraduate course or graduate-level course. Students will need to work with the appropriate time series functions in R or Stata to get this done. In an undergraduate course, I recommend using the informal test outlined by Wooldridge (2020) [3]. More advanced students may want to use the Dickey-Fuller test and could also run a regression in first differences.

(g) Generate a professionally formatted table showing your regressions.

There are many ways to get this done. If using R, the *stargazer* package is useful and user friendly. With Stata, the *estout* package works well. See the table at left for an example.