ECONOMETRICS WITH YOUTUBE DATA: A TIME SERIES TEACHING CASE

Sebastian Wai[1]

Abstract

This paper explains a teaching case I have designed for econometrics and business analytics courses. Time series remains a difficult topic to teach, in part because existing textbooks lack examples using real data and applicable to business problems students may encounter. This case aims to fill that gap using data from my own YouTube channel to give students an opportunity to practice time series analysis. Issues include endogeneity, seasonality, and unit roots. I also include discussion on how the case can be adapted for different course levels.

Keywords: time series, econometrics, teaching case, YouTube

JEL Classification: A22, A23

Introduction

One key principle that has made itself clear over years teaching econometrics and business analytics is that students need hands-on practice in the classroom. For an instructor, this means finding appropriate datasets that are simple to understand while demonstrating a particular concept. Widely available business analytics texts, such as Prince (2018) and Camm et al. (2017) tend to rely on constructed datasets for a clean demonstration. Recent articles on econometrics pedagogy (Kassens, 2019; Leamer, 2019) emphasize the importance of tying the classroom experience to problems students will face with real data. Neumann et al. (2013) use student feedback to show real-life data improves student engagement, interest, and motivation. Popular econometrics books, such as that of Wooldridge (2020) and both by Angrist and Pischke (2009, 2015) refer to real datasets, leaning heavily on prior studies in labor economics. While estimating Mincer equations is a useful exercise, students (particularly those in business analytics courses) may desire problems geared more toward what they may encounter in the business world. Such data can be a challenge for an instructor to find.

The case presented here aims to help instructors fill this gap with a regression exercise that gives students an opportunity to work with real data that feels relevant to them. The data come from my own YouTube channel,[2] which houses educational videos for my students and the general public. YouTube provides detailed data for channel owners, including daily views on specific videos, sources, and geography. It is a natural fit to teach basic time series analysis. The main research question is one that can be acted upon: do frequent uploads lead to more views? YouTube is itself a familiar setting for students. Prior research has shown that younger generations of students frequently view YouTube videos and view their incorporation into learning positively (Almobarraz, 2018; Buzzetto-More, 2015). Wessels and Steenkamp (2009) also showed economics and management courses specifically benefit from web and multimedia

[1] Assistant Professor of Economics, Beacom School of Business, University of South Dakota, 414 E. Clark St., Vermillion, SD 57069 (Sebastian.Wai@usd.edu).

[2] https://www.youtube.com/@sebastianwaiecon

content. If students have used my channel to learn Stata, they might even be part of the dataset themselves. Brown (2017) indicates the advantages of familiarity and immediacy of data, writing that large survey datasets used in many statistics courses "seem alien and remote," a problem this case avoids.

Time series can be a particularly tricky topic to teach at the undergraduate level, as the relevant issues are more technical in nature than those present in cross-sectional regressions. However, time series datasets are very common in the business world. It is easy to imagine a data analyst being given a time series of a company's sales and expenses in various areas. My approach, exemplified by this case, focuses on practical work that students are likely to encounter rather than formal tests. Hansen's (2017) article on this topic provides a useful framework on the time series topics undergraduates should be exposed to. Hansen suggests data such as GDP growth, stock returns, and gasoline prices. These are all useful examples but are not immediately actionable from a managerial perspective.

I tested this case as an in-class exercise in the Spring 2022 and 2023 sections of a capstone undergraduate business analytics course. The student response was generally positive. Students worked in small groups to analyze the case. Time series analysis comprises two weeks of the 16-week semester in this course. In the first week, I cover the basic models of static regression, autoregression, and distributed lag, along with seasonality, trends, and implementation in R. In the second week, I cover weakly and strongly dependent processes, unit roots, and first differencing, ending with this case. The case is designed as a summary exercise, covering numerous elements of time series analysis.

The Case

Background

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?

YouTube was founded in 2005. In 2006, Google purchased the company for $1.65 billion. By 2007, YouTube introduced the partner program, allowing users to run ads on their videos. As of 2022, applying for partner status requires 1000 subscribers and 4000 viewing hours over a 12-month period.

This case uses the dataset *Youtube.csv*.[3] The data were collected daily from your client's YouTube channel. The variables include:

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

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

[3] The dataset is available at https://doi.org/10.6084/m9.figshare.21382104.

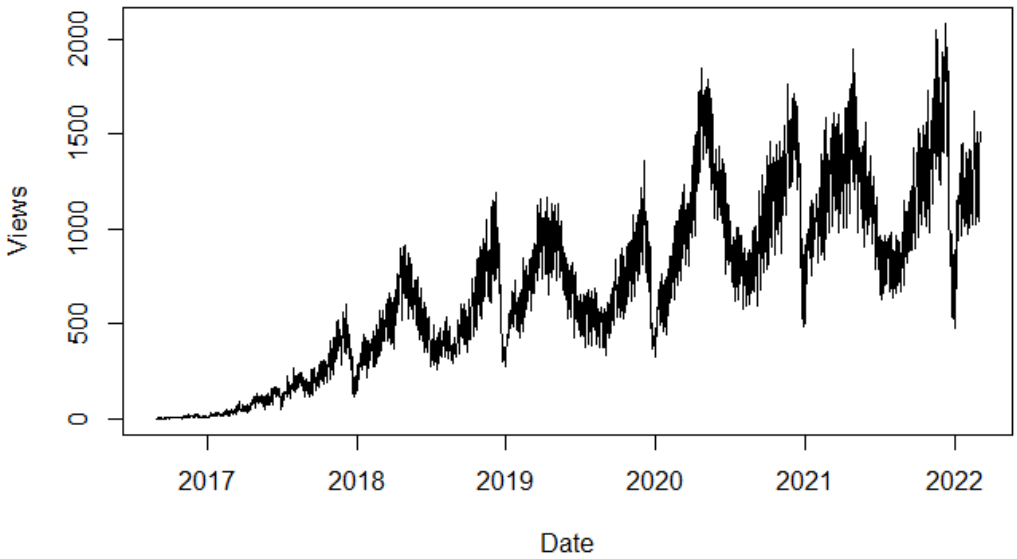
(b) Calculate the total views and subscribers to date.
(c) Run a simple regression to answer the channel owner's main question. Is this result surprising?
(d) Add a trend and monthly seasonality to your regression. Describe the seasonality present in the data and give an intuitive explanation.
(e) Based on your last regression, how does your answer to the main question change? Why do you think this happened?
(f) Test for the presence of unit roots in the final regression.
(g) Generate a professionally formatted table showing your regressions.

Teaching Notes

In this section, I will outline example solutions to the case and provide additional commentary. While this case was designed for an upper-level business analytics or econometrics course, instructors can adapt it to introductory statistics or graduate-level courses. Throughout the example solutions, I will note places where such adjustments can be made.

(a) To build the graph, students need to create a time variable. A basic solution would be to make a simple $t = 1,2,\ldots$ variable. More advanced students could convert the date into a usable format in R or Stata. Figure 1 shows an example graph generated with R. The graph shown here is quite basic, but this can also be an opportunity for an instructor to demonstrate more advanced graphing techniques. We can see a general upward trend which flattens out as time goes on. 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. This provides an opportunity to discuss the value of visual inspection of data, particularly as a tool to present to a less quantitative audience.

Figure 1: Channel Views Over Time.



(b) This question gives students some practice with basic functions. In R, the sum function can be used. Total views: 1,420,750. Total subscribers: 6801.

(c) This part gives students the chance to practice running and interpreting a regression. The estimated equation 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. Note, however, that the effect is relatively small (1.26 views per day). 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) 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. With a quadratic trend, the results show a positive trend that diminishes over time. For seasonality, students could either generate dummy variables or use factor methods within R or Stata. The seasonality results are consistent with the intuition in part (a) and comparison with the graph is a useful exercise. A regression-based introductory statistics class may want to stop the case here before we get into more complex econometric issues.

(e) The results have flipped from part (c). Using a model with a quadratic trend and dummy variables for months, the estimated effect is now -0.4 views per day. This aligns with the idea that infrequent uploads decrease views. This is again a relatively small effect. This question offers an opportunity to discuss possible omitted variables in the original model. A plausible explanation is that upload frequency is also seasonal and exhibits trends over time. If we run a regression of *Last.Video* on the quadratic trend and month dummies, we can see that *Last.Video* trends upward like *Views*, leading to a positive bias. 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 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. More advanced students could be asked to detrend the data and run the regression again. This would show the same coefficient estimate for *Last.Video*, but a much lower R-squared. As an extension, students could also be asked to use the estimated model to forecast views at a more recent date.

(f) 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). In this test, we run an AR(1) regression for each variable and check if the coefficient estimate is close to 1. For *Last.Video*, the estimate is 0.986 and for *Views*, the estimate is 0.969. As such, these results suggest the presence of unit roots. Graduate students may want to use the Dickey-Fuller test (1979) for a more formal approach. The *tseries* package available for R contains functions to execute this test. Doing so confirms the above results. If desired, students can be asked to estimate the model using first differences as a possible solution to the unit root problem. This would be suitable for more advanced students.

(g) 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. Table 1 shows an example using *stargazer*. This is a good opportunity for the instructor to discuss best practices in displaying tables. For example, I have omitted the constant term and dummy variables for months but have added a line noting their presence in the regression.

Table 1: Example Table Using Stargazer

	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)
Observations	2,010	2,010
Adjusted R^2	0.038	0.839
Note:	*p<0.1; **p<0.05; ***p<0.01	

Discussion

This case presents a teaching opportunity for several important issues in time series analysis and econometrics in general. Students get practice with the special ways in which time series data is handled in whichever statistical package is chosen. There are clear trends and seasonality in the data to discuss as well as a potential unit root to investigate. Finally, there are interesting omitted variables, which provide an avenue for discussing endogeneity and its solutions. As most students should be familiar with YouTube as a platform, the data is easy for them to understand. To maintain the immediacy of the case, the dataset can also be updated as time goes on.

This case could also be used as the basis for a project assignment. Students could find data on popular YouTube channels[4] or other websites and do their own analysis. The project could be expanded to a panel data analysis by including multiple channels. Students could practice presentation skills by giving a brief talk on their results to the class.

Acknowledgements

I thank session participants at the Missouri Valley Economics Association conference and the Council for Economic Education Poster Session at the ASSA Annual Meetings for their helpful feedback. I also thank my students for testing the case in class and inspiring me to develop better teaching materials.

Disclosure Statement

The author reports there are no competing interests to declare.

[4] For example, the website Social Blade provides daily viewing estimates for YouTube channels at https://socialblade.com/.

Data Availability Statement

The data used in this case, along with the sample solutions using both R and Stata, are openly available on Figshare at https://doi.org/10.6084/m9.figshare.21382104.

References

Almobarraz, Abdullah. 2018. "Utilization of YouTube as an information resource to support university courses." *The Electronic Library* 36 (1): 71-81.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

———. 2015. *Mastering 'Metrics*. Princeton, NJ: Princeton University Press.

Brown, Mark. 2017. "Making students part of the dataset: a model for statistical enquiry in social issues." *Teaching Statistics Trust* 39 (3): 79-83.

Buzzetto-More, Nicole. 2015. "Student Attitudes Towards The Integration Of YouTube In Online, Hybrid, And Web-Assisted Courses: An Examination Of The Impact Of Course Modality On Perception." *MERLOT Journal of Online Learning and Teaching* 11 (1): 55-73.

Camm, Jeffrey D., James J Cochran, Michael J Fry, Jeffrey W Ohlmann, David R Anderson, Dennis J Sweeney, and Thomas A Williams. 2017. *Business Analytics*. 3rd ed. Cengage.

Dickey, David A, and Wayne A Fuller. 1979. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association* 74 (366a): 427–31.

Hansen, Bruce E. 2017. "Time Series Econometrics for the 21st Century." *The Journal of Economic Education* 48 (3): 137–45.

Kassens, Alice L. 2019. "Theory Vs. Practice: Teaching Undergraduate Econometrics." *The Journal of Economic Education* 50 (4): 367–70.

Leamer, Edward. 2019. "Teaching the Art of Pulling Truths from Economic Data: Comment on 'Is Precise Econometrics and Illusion?'" *The Journal of Economic Education* 50 (4): 362–66.

Neumann, David L., Michelle Hood, and Michelle M. Neumann. 2013. "Using Real-Life Data When Teaching Statistics: Student Perceptions of this Strategy in an Introductory Statistics Course." *Statistics Education Research Journal* 12 (2): 59-70.

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

Wessels, P.L. and L.P. Steenkamp. 2009. "Generation Y students: Appropriate learning styles and teaching approaches in the economic and management sciences faculty." *South African Journal of Higher Education* 25 (3): 1039-1058.

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

Appendix

This appendix contains example code in R and Stata which can be used to work through the case. Both scripts assume the working directory contains the dataset Youtube.csv.

R Code

```
#Note Stargazer must be installed.

yt = read.csv('Youtube.csv')

#A
yt$d1 = as.Date(yt$Date, format="%m/%d/%Y")
plot(yt$d1, yt$Views, type='l', ylab="Views", xlab="Date")

#B
sum(yt$Views)
sum(yt$Subscribers)

#C
reg1 = lm(Views ~ Last.Video, yt)
summary(reg1)

#D
yt$mon = match(yt$Month, month.name)
yt$t = seq(1,length(yt$Date))
yt$t2 = yt$t^2
trend = lm(Views ~ t + t2 + factor(mon), yt)
summary(trend)
lines(yt$t, trend$fitted.values, col='red')
reg2 = lm(Views ~ Last.Video + t + t2+ factor(mon), yt)
summary(reg2)

#E
summary(lm(Last.Video ~ t + t2 + factor(mon), yt))

#F
last = ts(yt$Last.Video)
views = ts(yt$Views)
last1 = lag(last, -1)
view1 = lag(views, -1)
t = ts(yt$t)
t2 = ts(yt$t2)
mon = ts(yt$mon)
dlast = diff(last)
dviews = diff(views)
ytts = ts.union(last, last1, dlast, views, view1, dviews, t, t2, mon)
```

```
summary(lm(last ~ last1, ytts))
summary(lm(views ~ view1, ytts))

#FD regression (extension)
reg3 = lm(dviews ~ dlast + t + t2 + factor(mon), ytts)
summary(reg3)

#G
library(stargazer)
stargazer(reg1, reg2, type="text", omit=c("Constant", "mon"), omit.stat=c("f","rsq","ser"))
```

Stata Code

```
*Note that esttab must be installed to make the tables.
import delimited Youtube.csv, clear
eststo clear

*A
gen date2 = date(date,"MDY")
format date2 %d
line views date2, xtitle("Date")

*B
summarize views
display r(sum)
summarize subscribers
display r(sum)

*C
eststo: reg views lastvideo

*D
encode month, gen(mon)
gen t = _n
gen t2 = t^2
eststo: reg views lastvideo t t2 i.mon

*E
reg lastvideo t t2 i.mon

*F
tsset t
reg lastvideo l1.lastvideo
reg views l1.views

*FD regression
gen dviews = views - l1.views
```

```
gen dlast = lastvideo - l1.lastvideo
reg dviews dlast t t2 i.mon

*G
esttab, se
```