Reproducing Angelucci & Cagé's 'Newspapers in Times of Low Advertising Revenues'

Yuyan Xu

22 December 2020

Abstract

The present paper aims to replicate the results from Angelucci & Cagé's paper, and provide more validity to the finding that alternate advertising platforms tend to decrease the newspaper producers'incentives to advertise, whcih leads to a decrease in journalistics-intensive content in the newspapers.

Keywords

Advertising revenues; French newspaper; Journalistic-intensive content; Readership; Pricing strategies;

Github Repository

Code for this analysis is available at:

Introduction

The advancement in communication technology today provides general readers with more ways to access information. People can easily fetch news from TV, smartphones, tablets, etc. It is of interest that how newspaper industry change their method of delivering news, and how their advertising revenues changes corresponding with the newspaper content and pricing. In Angelucci & Cagé's "Newspapers in Times of Low Advertising Revenues", they explored the shock brought to the advertising side of French newspaper industry with the introduction of television advertising. Also, they used empirical analysis to discover the relationship between advertising revenues and the incentive of newspaper producers put journalistic-intensive contents in their publications (Angelucci & Cagé, 2019). They found that there was a decline in newspaper producers' willingness to pay for advertising for attracting readers' attention with the introduction of alternative advertising platforms; also, a decrease in advertising avenue lead to a corresponding decrease in journalistic-intensive content in newspaper (Angelucci & Cagé, 2019). The present paper aims to reproduce these results using simple linear model with ordinary least squares estimators. The data would be the data used in Angelucci & Cagé's paper: the information about newspaper prices, revenues, and circulation from the French Ministry of Information's non-publicly available records in the National archives, which includes 68 local newspapers and 12 national newspaper; the number of journalist from the non-publicly available paper records of the "Commission de la carte d'identité des journalistes professionnels" (CCIJP) , which includes 63 out of the 68 local newspapers for which we have revenue data, and 11 out of the 12 national newspapers; information about advertising prices and quantity of advertising from "Tarif Media," and French National Library respectively.

Data

Table 1: Data summary

Name	dta
Number of rows	1196
Number of columns	18
Column type frequency:	
factor	4
numeric	14
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
id_news	0	1.00	FALSE	80	1: 15, 3: 15, 6: 15, 7: 15
after_national	1100	0.08	FALSE	1	1: 96
local	180	0.85	FALSE	1	1: 1016
national	1016	0.15	FALSE	1	1: 180

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	
year	0	1.00	1967.02	4.32	1960.00	1963.00	
ra_cst	144	0.88	91531796.91	137207312.44	549717.25	13551599.00	359
qtotal	126	0.89	130817.51	172954.33	1480.00	27822.94	
pages	150	0.87	15.75	6.50	2.00	12.00	
news_hole	150	0.87	12.46	3.96	1.86	10.21	
nb_journ	131	0.89	62.79	65.71	1.00	14.00	
ps_cst	152	0.87	2.78	0.74	0.68	2.23	
po_cst	133	0.89	3.23	0.88	0.82	2.59	
qs_s	124	0.90	27.19	22.66	0.70	8.49	
rs_cst	150	0.87	97666503.64	125257120.29	255760.08	18541837.50	407
R_sh_edu_primaire_ipo	783	0.35	63.85	15.21	11.53	58.35	
R_sh_edu_secondaire_ipo	783	0.35	13.49	6.21	2.16	9.01	
R_sh_edu_no_ipo	841	0.30	-0.11	4.17	-30.40	0.12	
R_sh_edu_sup_prof_ipo	783	0.35	22.82	10.80	4.67	15.98	
Figure 1 Overview o	f dataset wi	th variables of	interest				

Model

The models we are interested in are based on the Ordinary Least Sqaures (OLS) method, which is a method commonly used to examine linear regression models :

$ln(Y_{n,t}) = \beta_0 + \beta_1 + ... + \beta_n + \epsilon$

where

- $\hat{\beta}_0$ is the OLS estimator of the intercept coefficient β_0
- $\hat{\beta}_1 ... \hat{\beta}_n$ are the OLS estimators of the slope coefficients $\beta_1 ... \beta_n$

- ϵ is a random error term

The rationale of the estimated models in present paper are based on the OLS simple regression function

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_n$

By using the OLS estimation, we can minimize the sum of the squared residuals, in other words, minimize the squared prediction errors of the model. This means that we can obtain a more accurate model using OLS. In this case, Y represents the predictor variable, which is advertising avenue, X represents the response varaibles, which were selected based on the findings in Angelucci & Cagé, 2015. By doing this, we aimed to reproduce their results by examining the significance of the relationship between these response variables X and the predictor variable Y. We built two models, one is from the aspect of newspaper production, another is from the aspect of readership. Therefore, the response variables were selected respectively in terms of these two aspects, while the predictor variable, advertising avenue, stayed the same in each model.

First, we aimed to investigate the potential variables in newspaper production (total circulation, number of pages in publication, news-hole, number of journalists) that could be significantly influenced corresponding to the changes in advertising avenues.

```
##
## Call:
## lm(formula = ra_cst ~ qtotal + pages + news_hole + nb_journ,
##     data = dta)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -215025609  -24083481    1410071   19360215  416074583
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.412e+07  7.128e+06   1.980    0.048 *
## qtotal       2.126e+02  2.389e+01   8.896   <2e-16 ***
## pages        1.165e+07  6.425e+05  18.129   <2e-16 ***
## news_hole   -1.672e+07  1.056e+06 -15.840   <2e-16 ***
## nb_journ     1.204e+06  6.596e+04  18.262   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58600000 on 913 degrees of freedom
##   (278 observations deleted due to missingness)
## Multiple R-squared:  0.8337, Adjusted R-squared:  0.8329
## F-statistic:  1144 on 4 and 913 DF,  p-value: < 2.2e-16
```

Figure 1 The summary statistics for model1

We first included all the selected variables in model1 and the summary statistics indicated that they all have significant relationships with the predictor variables, since the p-values were extremely small. However, a best model should contain fewer number of variables if possible, since there is a chance that a variables had small p-value because of its co-effects with another variable contained in the model. In this case, if remove the other variable from the mode, as a result, the former variable would not have a significant relationship or its significant value decreases. Therefore, we removed variable of total pages of publications from the model in order to examine if there is a variable consequently has reduced significant level. Based on the summary statistics of model2, the p values basically remained unchanged. In the same way, we constructed the third model by removing variable, number of pages, from model1. The p-values of all the variables still remained unchanged. However, it is worth mentioning that no matter which variable we remove from model1, the value of adjusted R-sqaured decreases. This suggests that all the variables contained in model1 are significant to the fit of regression model, in other words, by adding each of the variable into the model,

the model became increasingly better. Therefore, for the newspaper production, variables which are total circulation, number of pages in publication, news-hole, number of journalists, can be affected by a change in advertising avenues.

However, the vif value of variable, number of journalists, was 5.18, which is higher than the threshold of 5. It means that this model has problem estimating the coefficients although its predictive ability is not affected. The vif values of model2 were good, which were all below the threshold of 5.

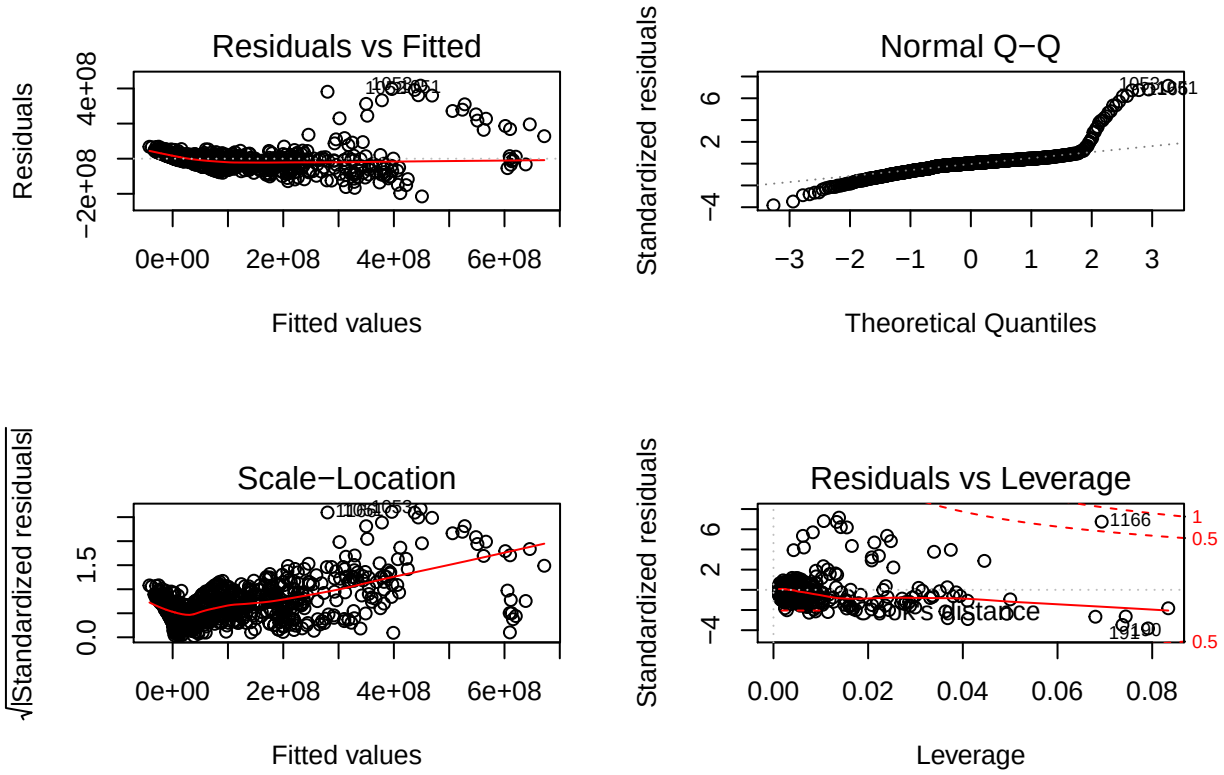

Figure 2 Plotting model2 and examining its validity

We also examined the validity of model2 by plotting the model in residuals vs. fitted graph, QQ plots, Scale-location plot and residual vs leverage plot.From the produced plots above, we can see that although there are some significant leverage points, the overall linear trend of model1 is clear.

Therefore, for newspaper production side, we decided to choose model2 as our estimated model, which contains variables: number of pages in publications, news-hole and number of journalists.

Second, we examined the potential variables in the reader side: subscription price per issue, unit buyer price, percentage of share of subscribers, reader's level of education. We removed the variables that showed no significant relationship with advertising revenues from model_1, and obatined model_2 with the rest of the variables: subscription price per issue, unit buyer price, percentage of share of subscribers, reader with primary education.

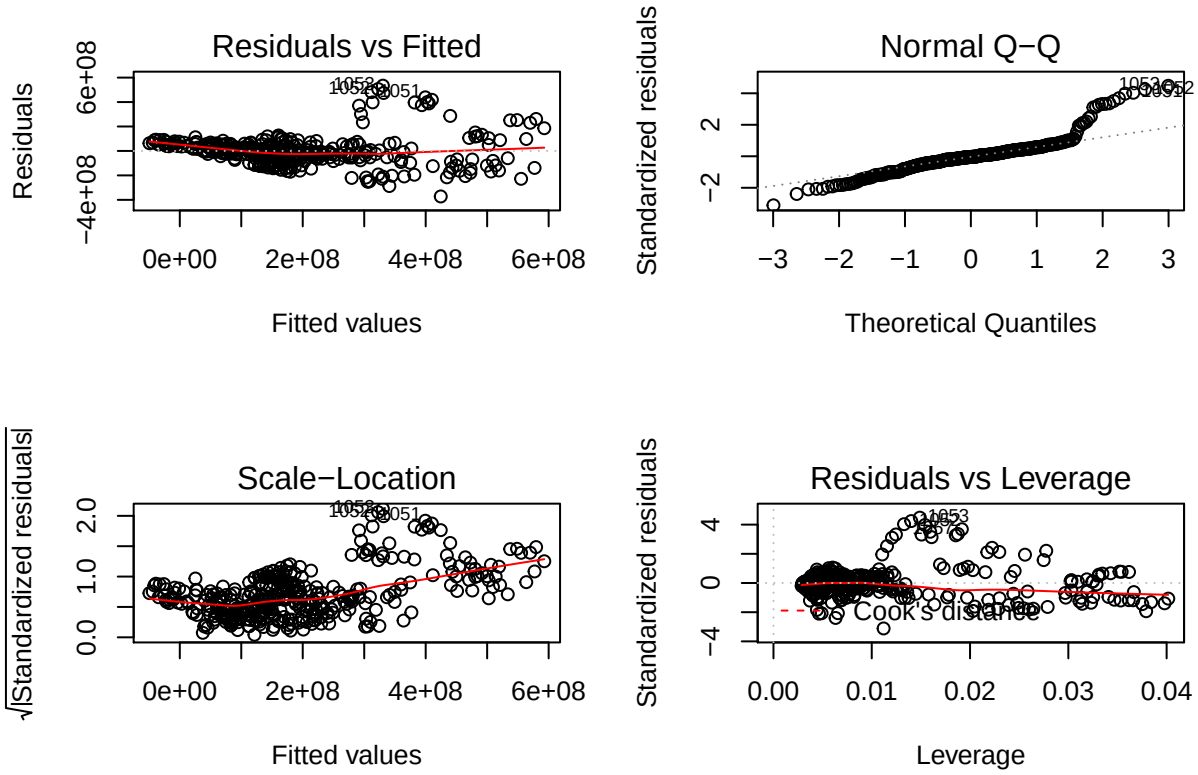

Figure Plotting model_3 and examining its validity

Like what we did for model2 in newspaper production side, we also produced plots for model_3 in reader side. Also, the vif values for model_3 were all around 1, which means that the varaibles in model_3 did a good job on estimating coefficents.

Result

We have several key findings summarized from the model section. First, the number of pages in publications, number of news-holes, and number of journalists were all associated with advertising revenues for newspaper industries; Second, the percentage of subsribers, revenues from sale, reader's education level were all poitively correlated with advertising revenues; Third, which is an interesting finding, the number of news hole had negative correlated relationship with advertising revenues.

Discussion

Reference

Angelucci, Cage. (2015). Newspapers in Times of Low Advertising Revenues. American Economic Journal: Microeconomics, 11(3). DOI: 10.1257/mic.20170306

Luca Braglia (2016). RStata: A Bit of Glue Between R and Stata. R package version 1.1.1. https://CRAN.R-project.org/package=RStata

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Matt Dowle and Arun Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.8. https://CRAN.R-project.org/package=data.table

Martin Elff (2020). memisc: Management of Survey Data and Presentation of Analysis Results. R package version 0.99.27.3. https://CRAN.R-project.org/package=memisc