

Forecasting Box Office Revenues: An application of SLR analytics/assessment

In the movie industry, weekend #1 means everything:

*In the past, you could start a movie off and it would do OK, and it was good. Week on week, you could pick up business or stay there. Now, if you're not number one or two -- and people are lying about what's number one and two half the time -- you're pretty much dead meat. In two weeks, you're not even in the theater.*¹ Bill Mechanic, former Chairman and CEO of Twentieth Century Fox ... quoted in *Frontline :The Monster that Ate Hollywood - Open Wide, Open Big*²

*It's well-known in the business that Friday nights during the opening weekend are nervous times for the marketing people at movie studios. It is on the strength of the opening weekend of general release that all major decisions pertaining to a film's ultimate financial destiny are made. Since competition for movie screens is fierce, movie theater owners do not want to spend more than the contractually obligatory two weeks on a film that doesn't have "legs." ... Movie theater owners often make the decision to keep a film running based on the strength of its opening weekend. But is it really true that first weekend grosses are predictive for the ultimate total domestic gross?*³ Jeffrey Simonoff, NYU

The weekend number\$ can impress (Top 10 opening weekend domestic revs as of Aug 2019):⁴

Rank	Title (click to view)	Studio	Opening*	% of Total	Theaters	Avg.	Total Gross^	Date**
1	Avengers: Endgame	BV	\$357,115,007	41.6%	4,662	\$76,601	\$858,373,000	4/26/2019
2	Avengers: Infinity War	BV	\$257,698,183	38.0%	4,474	\$57,599	\$678,815,482	4/27/2018
3	Star Wars: The Force Awakens	BV	\$247,966,675	26.5%	4,134	\$59,982	\$936,662,225	12/18/2015
4	Star Wars: The Last Jedi	BV	\$220,009,584	35.5%	4,232	\$51,987	\$620,181,382	12/15/2017
5	Jurassic World	Uni.	\$208,806,270	32.0%	4,274	\$48,855	\$652,270,625	6/12/2015
6	Marvel's The Avengers	BV	\$207,438,708	33.3%	4,349	\$47,698	\$623,357,910	5/4/2012
7	Black Panther	BV	\$202,003,951	28.9%	4,020	\$50,250	\$700,059,566	2/16/2018
8	The Lion King (2019)	BV	\$191,770,759	35.9%	4,725	\$40,586	\$533,992,775	7/19/2019
9	Avengers: Age of Ultron	BV	\$191,271,109	41.7%	4,276	\$44,731	\$459,005,868	5/1/2015
10	Incredibles 2	BV	\$182,687,905	30.0%	4,410	\$41,426	\$608,581,744	6/15/2018

¹ <https://www.pbs.org/wgbh/pages/frontline/shows/hollywood/interviews/mechanic.html>


² <https://www.pbs.org/wgbh/pages/frontline/shows/hollywood/picture/openbig.html>

³ Jeffrey Simonoff, *Predicting total movie grosses after one week*, <http://people.stern.nyu.edu/jsimonof/classes/2301/pdf/movies.pdf> ... also, see Jeffrey S. Simonoff and Ilana R. Sparrow, *Predicting movie grosses: Winners and losers, blockbusters and sleepers*, *Chance*, 13(3), 15-24 (Summer 2000), <http://pages.stern.nyu.edu/~jsimonof/movies/movies.pdf> .

⁴ <https://www.boxofficemojo.com/alltime/weekends/> (unadjusted for inflation)

Forecasting Box Office Revenues v.6

Given the importance of the weekend #1 performance, it's no surprise to see it highlighted on Box Office Mojo's summary pages. Here's an example for [Avengers: Endgame](#):



Avengers: Endgame

Domestic Total Gross: \$858,373,000	
Distributor: Buena Vista	Release Date: April 26, 2019
Genre: Action / Adventure	Runtime: 3 hrs. 1 min.
MPAA Rating: PG-13	Production Budget: \$356 million

Summary
Daily
Weekend
Weekly
Foreign
Similar Movies

<p>Total Lifetime Grosses</p> <p>Domestic: \$858,373,000 30.7% + Foreign: \$1,937,901,401 69.3%</p> <p>= Worldwide: \$2,796,274,401</p> <p>Domestic Summary</p> <p>Opening Weekend: \$357,115,007 (#1 rank, 4,662 theaters, \$76,601 average) % of Total Gross: 41.6% View All 20 Weekends</p> <p>Widest Release: 4,662 theaters Close Date: September 12, 2019 In Release: 140 days / 20 weeks</p>	<p>The Players</p> <p>Directors: Joe Russo Anthony Russo</p> <p>Writers: Christopher Markus Stephen McFeely</p> <p>Actors: Robert Downey, Jr. Chris Hemsworth Mark Ruffalo Chris Evans Scarlett Johansson</p>	<p>Related Stories</p> <p>7/21 'The Lion King' Debuts with Record \$185M & 'Endgame' Becomes Global #1</p> <p>6/27 'Toy Story 4' Looks to Repeat at Weekend Box Office Over 'Annabelle' & 'Yesterday'</p> <p>6/9 'Pets 2' and 'Dark Phoenix' Top Weekend Box Office, Though Both Underperform</p> <p>6/6 'Secret Life of Pets 2' and 'Dark Phoenix' Both Struggle at the Box Office</p> <p>6/2 'Godzilla' Sequel is Box Office King; 'Rocketman' and 'Ma' Top Expectations</p>
---	--	---

- Lifetime Gross Revenue: \$2.8B Worldwide as of 9/12/2019 - Domestic: \$858M (30.7%); Foreign: \$1.94B (69.3%)
- Opening Weekend (Domestic): \$357M (#1 rank, 4,662 theaters, \$76K average); % of Total Gross: 41.6%

The Challenge: Forecast lifetime box office revenues as a function of weekly revenues. Which week's revenues best predict lifetime film revenues? wk #1?, wk #2?, wk #3?, wk #4, ...
Hint: It is not week #1! *Surprise!*

Bring on the data - *Box Office Mojo*⁵ US Gross Revenues:⁶ 1982 – January, 2017

- Weekly domestic revenue data:⁷ 118,134 observations; 13,186 titles
- Fields: \$Revenue Rank (this wk and last wk) , Title, Studio, Weekly Gross Revenue, Theatre Count, Average Gross Revenue, Cumulative Gross Revenue, Budget, #Weeks
- All revenue figures brought forward to current \$US using the CPI

⁵ <http://www.boxofficemojo.com>

⁶ Foreign revenues for about 60 countries also available, for about the past 15 years.
<http://www.boxofficemojo.com/intl/>

⁷ The data used in this handout are in *movierevs v6.dta*, which includes data through January, 2017.

Forecasting Box Office Revenues v.6

An Example:

The following Figure gives you a sense of what the weekly data look like for *Avengers: Endgame*:

2019

Date (click to view chart)	Rank	Weekly Gross	% Change	Theaters / Change	Avg.	Gross-to-Date	Week #	
Apr 26–May 2	1	\$473,894,638	-	4,662	-	\$101,651	\$473,894,638	1
May 3–9	1	\$186,551,101	-60.6%	4,662	-	\$40,015	\$660,445,739	2
May 10–16	1	\$80,949,131	-56.6%	4,662	-	\$17,364	\$741,394,870	3
May 17–23	2	\$39,936,866	-50.7%	4,220	-442	\$9,464	\$781,331,736	4
May 24–30	3	\$26,357,048	-34.0%	3,810	-410	\$6,918	\$807,688,784	5
May 31–Jun 6	6	\$11,877,156	-54.9%	3,105	-705	\$3,825	\$819,565,940	6
Jun 7–13	8	\$7,408,419	-37.6%	2,121	-984	\$3,493	\$826,974,359	7
Jun 14–20	11	\$5,634,607	-23.9%	1,450	-671	\$3,886	\$832,608,966	8
Jun 21–27	14	\$3,172,195	-43.7%	985	-465	\$3,221	\$835,781,161	9
Jun 28–Jul 4	8	\$8,981,672	+183%	1,985	+1,000	\$4,525	\$844,762,833	10

Top Films ... in the dataset (total gross revenues (real \$)):

rank	lastyr	lastwk	title	studio_hm	studio	budget	total gross revenues		nwks
							Nominal	Real \$2017	
1	2016	22	Star Wars: The Force Awakens	buenavista.htm	BV	245	\$ 936,662,225	\$ 938,617,664	24
2	1998	39	Titanic	paramount.htm	Par.	200	\$ 600,683,057	\$ 884,441,280	41
3	1983	22	E.T.: The Extra-Terrestrial		Uni.	10.5	\$ 353,343,189	\$ 851,429,376	52
4	2010	32	Avatar	fox.htm	Fox		\$ 749,766,139	\$ 825,052,096	34
5	2015	46	Jurassic World	universal.htm	Uni.	150	\$ 652,198,011	\$ 660,453,696	23
6	2012	39	Marvel's The Avengers	buenavista.htm	BV	220	\$ 623,357,910	\$ 651,593,600	22
7	1983	52	Return of the Jedi	fox.htm	Fox	32.5	\$ 249,608,768	\$ 601,466,944	32
8	2000	4	Star Wars: Episode I - The Phantom Menace	fox.htm	Fox	115	\$ 431,088,295	\$ 600,819,904	37
9	2009	9	The Dark Knight	warnerbros.htm	WB	185	\$ 533,345,358	\$ 596,750,080	33
10	1994	41	Jurassic Park	universal.htm	Uni.	63	\$ 356,763,175	\$ 577,754,112	71
11	2004	47	Shrek 2	dreamworks.htm	DW	150	\$ 441,226,247	\$ 560,583,936	21
12	2002	32	Spider-Man	sony.htm	Sony	139	\$ 403,638,985	\$ 538,484,480	15
13	2017	4	Rogue One: A Star Wars Story	buenavista.htm	BV	200	\$ 521,709,512	\$ 521,709,504	7
14	1995	16	Forrest Gump	paramount.htm	Par.	55	\$ 327,838,708	\$ 516,281,440	42
15	1984	52	Ghostbusters	columbia.htm	Col.	30	\$ 221,072,172	\$ 510,657,568	30
16	1985	26	Beverly Hills Cop	paramount.htm	Par.		\$ 226,832,681	\$ 505,946,496	30
17	2006	48	Pirates of the Caribbean: Dead Man's Chest	buenavista.htm	BV	225	\$ 423,315,812	\$ 503,947,392	22
18	1995	7	The Lion King	buenavista.htm	BV	45	\$ 311,455,496	\$ 490,481,120	36
19	1991	25	Home Alone	fox.htm	Fox	18	\$ 277,905,624	\$ 489,701,536	32
20	1982	11	Raiders of the Lost Ark	paramount.htm	Par.	18	\$ 196,614,672	\$ 488,989,856	23
21	2016	49	Finding Dory	buenavista.htm	BV		\$ 486,295,561	\$ 487,310,784	25
22	1989	49	Batman	warnerbros.htm	WB	35	\$ 251,188,924	\$ 486,172,096	25
23	2004	22	The Lord of the Rings: The Return of the King	newline.htm	NL	94	\$ 377,027,325	\$ 479,018,336	24
24	2004	45	Spider-Man 2	sony.htm	Sony	200	\$ 373,286,218	\$ 474,265,184	19
25	2004	30	The Passion of the Christ	newmarket.htm	NM	30	\$ 370,274,604	\$ 470,438,912	22

In the beginning: Hey you! Look at your data! Remember: Data Integrity is #1!

Every econometric analysis should begin with a careful review of the data. You cannot spend enough time looking at summary statistics, histograms, cross-tabs, correlations and covariances, etc.... to get an understanding of your data, and data issues which might need to be addressed. Keep asking yourself: Does what you are seeing make sense? Anything *kafooe*y with the numbers? Anything that you should be investigating? ... and don't just ask... follow up!

Forecasting Box Office Revenues v.6

It is so tempting and so easy to just run regressions... but if you don't understand your data, you have no idea whether your regressions are just garbage, or not (no matter how *oh so technically sophisticated* your analysis).

Put differently: ***Beware data GIGO: Garbage In; Garbage Out!***

So let's look at your data!

Distribution of total revenues... some perspectives

Let's start with some basic *summary statistics*:

```
. summ rtotgross wk1 wk2 wk3 wk4 wk5
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rtotgross	11,419	23.0794	53.06492	.0000357	938.6177
wk1	11,419	8.333653	19.75821	.0000117	395.8036
wk2	9,114	6.252487	12.10332	4.40e-06	264.4164
wk3	7,842	4.551125	8.127558	2.23e-06	118.661
wk4	6,952	3.280906	5.86471	.000019	76.94823
wk5	6,182	2.386494	4.542509	.000012	72.99082.

Not surprisingly, the means are dropping week-by-week. But notice that as well, Obs changes by the week, no doubt due to missing observations. Should we worry about this? Models working with more weeks will have fewer observations due to missing data. That might make models incomparable. We'll get back to this issue.

Want more detail? Try the *detail* option:

```
. sum rtotgross, detail
```

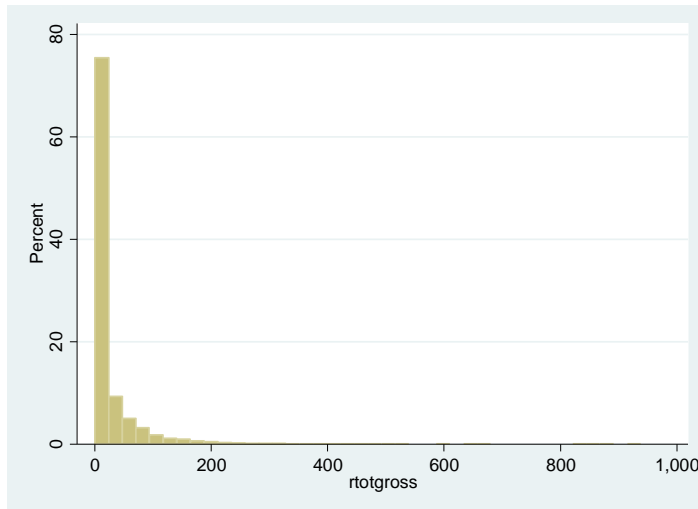
rtotgross					
Percentiles		Smallest			
1%	.0014406	.0000357			
5%	.0055608	.0000701			
10%	.0118573	.0000742	Obs		11,419
25%	.0560516	.0000749	Sum of Wgt.		11,419
50%	<u>.9426789</u>		Mean		<u>23.0794</u>
		Largest	Std. Dev.		53.06492
75%	22.64797	825.0521			
90%	71.14035	851.4294	Variance		2815.886
95%	117.0008	884.4413	Skewness		5.115296
99%	247.5804	938.6177	Kurtosis		47.07394

So while mean movie revenues are \$23.1M, more than 50% of the films have lifetime revenues of under \$1M, and 75% of the films have total revenues of under \$22.65M. This is evidence of a highly skewed distribution, with a relatively small number of *megahits/blockbusters* driving mean revenues well above the median.

Forecasting Box Office Revenues v.6

Histograms

```
histogram rtotgross, percent  
(bin=40, start=.00003571, width=23.465441)
```



So 50% of the films have less than \$1M in box office revenues, and about 62% are under \$10M in revenues.

And what about the correlation between those weekly box office revenues and total film gross revenue? Use Stata's *corr* command to generate the correlations:

```
. corr rtotgross wk1 wk2 wk3 wk4 wk5  
(obs=5,856)
```

	rtotgr~s	wk1	wk2	wk3	wk4	wk5
rtotgross	1.0000					
wk1	0.8707	1.0000				
wk2	0.9388	0.9296	1.0000			
wk3	0.9463	0.8316	0.9430	1.0000		
wk4	0.9261	0.7299	0.8571	0.9443	1.0000	
wk5	0.8918	0.6212	0.7767	0.8722	0.9466	1.0000

Not surprisingly, most of the correlations are above .80, if not .90.

But Wait!

We have 11,419 observations in the dataset... How come the correlations have obs= 5,856?

Forecasting Box Office Revenues v.6

Let's try this again with... *pwcorr*, to generate the *pairwise correlations* (so no dropped records just because one variable happens to be missing; note that the *pw* in *pwcorr* stands for *pairwise*):

```
. pwcorr rtotgross wk1 wk2 wk3 wk4 wk5
```

	rtotgr~s	wk1	wk2	wk3	wk4	wk5
rtotgross	1.0000					
wk1	0.8831	1.0000				
wk2	0.9418	0.9342	1.0000			
wk3	0.9474	0.8393	0.9456	1.0000		
wk4	0.9275	0.7367	0.8609	0.9451	1.0000	
wk5	0.8921	0.6232	0.7777	0.8718	0.9469	1.0000

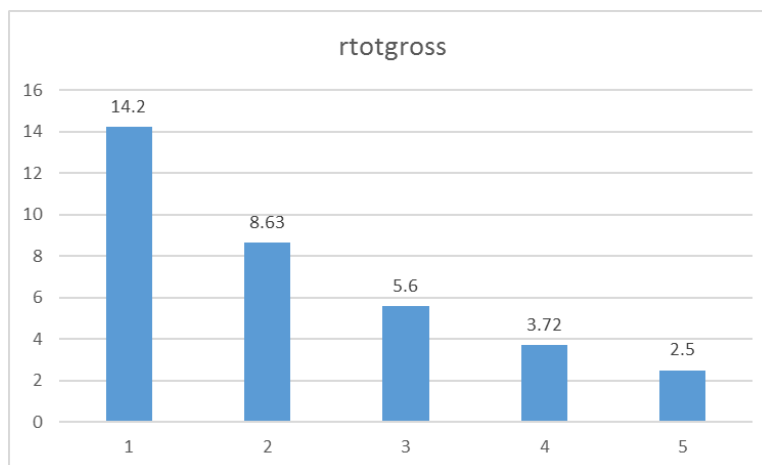
The new correlations are within 1% point (if not 0.5% points) of the prior figures. So turning to pairwise correlations hasn't changed things much. *But it was certainly worth a look!*

Pattern over time?

So let's take another look at those weekly revenues, looking only at films with data for all of the weeks (wk1-wk5). On average, about a third of total revenues are in week 1... and at least in the first several weeks, each week's revenues are on average about 2/3rds of the revenues in the prior week:

```
. summ rtotgross wk1 wk2 wk3 wk4 wk5 if wk2!=. & wk3!=. & wk4!=. & wk5!=.
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rtotgross	5,856	40.84843	68.83194	.0015503	938.6177
wk1	5,856	14.19789	25.88587	.0000304	395.8036
wk2	5,856	8.630448	14.3202	4.40e-06	264.4164
wk3	5,856	5.598127	9.048863	2.23e-06	118.661
wk4	5,856	3.716118	6.240694	.0000391	76.94823
wk5	5,856	2.496333	4.621171	.000012	72.99082



Forecasting Box Office Revenues v.6

Comparing total revenues to week 1 revenues:

We can generate a *multiple* showing *rtotgross* as a multiple of *wk1* revenues. In the simplest of all models, you might be tempted to use that *multiple* to predict total revenues based on week 1 revs (after all, financial wizards do this all the time when they are valuing enterprises). But maybe *multiples* vary significantly by film... in which case you would likely want to shy away from just using *multiples*, or relying too heavily on *multiples analysis*.

```
. gen mult1=rtotgross/wk1
. summ mult1, detail
```

mult1					
Percentiles		Smallest			
1%	1	.37058			
5%	1	.7788957			
10%	1	.8171461	Obs		11,419
25%	1.463347	.9485417	Sum of Wgt.		11,419
50%	2.202446		Mean		10.92286
		Largest	Std. Dev.		122.6946
75%	4.529478	3144.167			
90%	14.26041	4468.3	Variance		15053.95
95%	30.30049	7648.038	Skewness		53.47511
99%	139.2657	8196.829	Kurtosis		3248.459

So a simple model would multiply wk1 revs by the mean multiple of 10.9... or would you use the median multiple of 2.2? It would make a difference, yes? And the inter-quartile range of multiples is 1.5 – 4.5 ... hmmm, maybe not as precise as you might like?

But wait! Some films have *rtotgross* that are more than 1,000X *wk1* revs? ... *I don't think so!* What are those films?

Here's a culprit: *Inglourious Basterds*, of course! Who sees the problem?



Inglourious Basterds

Domestic Total Gross: **\$120,540,719**

Distributor: **Weinstein Company**

Release Date: **August 21, 2009**

Genre: **War**

Runtime: **2 hrs. 32 min.**

MPAA Rating: **R**

Production Budget: **\$70 million**

Summary Daily Weekend **Weekly** Foreign Dvd / Home Video Similar Movies

2009

Date (click to view chart)	Rank	Weekly Gross	% Change	Theaters / Change	Avg.	Gross-to-Date	Week #
Aug 14-20	-	\$15,761	-	28	\$563	\$15,761	0
Aug 21-27	1	\$53,703,427	+340,636%	3,165	+3,137	\$16,968	1
Aug 28-Sep 3	2	\$26,476,420	-50.7%	3,165	-	\$8,365	2
Sep 4-10	2	\$17,567,244	-33.6%	3,358	+193	\$5,231	3
Sep 11-17	3	\$8,535,813	-51.4%	3,215	-143	\$2,655	4
Sep 18-24	7	\$5,438,870	-36.3%	2,519	-696	\$2,159	5

Forecasting Box Office Revenues v.6

And what about those multiples that are less than 1? Something's fishy here!

So ... Look at your data! To repeat what I said earlier: Good econometric practice requires that you spend some time looking at your data before you jump in. It is so easy and tempting to just run regressions. But the truth is that if you don't understand your data, your econometric analysis will suffer accordingly. So put the time into understanding your data!

We've identified a number of data problems/issues. We'll ignore them for now (*lazy! lazy!*)... but we'll definitely want to do something about these issues in the future! ... and certainly before we started bragging about our econometric analysis.

Time for Some Regressions!

The goal here is to determine which week's revenues has the most explanatory power in predicting lifetime movie box office revenues (domestic). So let's run six SLR models, one for each of the first six weeks, and compare results:

```
. esttab, r2 scalar(rmse) compress
```

	(1)	(2)	(3)	(4)	(5)	(6)
	rtotgross	rtotgross	rtotgross	rtotgross	rtotgross	rtotgross
wk1	2.372*** (201.14)					
wk2		4.516*** (267.49)				
wk3			7.174*** (262.21)			
wk4				10.20*** (206.86)		
wk5					13.28*** (155.19)	
wk6						14.98*** (109.54)
_cons	3.313*** (13.10)	0.403 (1.75)	0.0565 (0.22)	2.399*** (7.29)	7.156*** (16.30)	14.14*** (24.22)
N	11419	9114	7842	6952	6182	5606
R-sq	0.780	0.887	0.898	0.860	0.796	0.682
rmse	24.90	19.51	19.69	24.12	30.55	39.68

t statistics in parentheses

* p<0.05, ** p<0.01, *** p<0.001

Forecasting Box Office Revenues v.6

Here are the SRFs from the different SLR models:

- $SRF_1 : \hat{y} = 3.3 + 2.4 \cdot wk1$
- $SRF_2 : \hat{y} = 0.4 + 4.5 \cdot wk2$
- $SRF_3 : \hat{y} = 0.06 + 7.2 \cdot wk3$
- $SRF_4 : \hat{y} = 2.4 + 10.2 \cdot wk4$
- $SRF_5 : \hat{y} = 7.2 + 13.3 \cdot wk5$
- $SRF_6 : \hat{y} = 14.14 + 15.0 \cdot wk6$

A few observations:

1. The estimated slope coefficients are all positive. No surprise there! Explain why.
2. The estimated slope coefficients increase as you move left to right. No surprise with that! This is to be expected. Why?
3. The R^2 are all in the .7-.9 range, so each model has a fairly impressive amount of explanatory power.
4. The R^2 are increasing as you start to move left to right, reach a maximum value of 0.898 with wk3, and then decline thereafter.
5. So the pattern of R^2 's suggest that wk3 (and not wk1!) has the most explanatory power... wk2 has the second most explanatory power... then come wk4, and wk5, and finally wk1. So the R^2 's suggest that of the six weeks considered, only wk6 has less explanatory power than wk1! So much for wk1 telling you all you need to know about total box office revenues!

But wait! ... not so fast! ... something is not right!

6. The pattern of $RMSE$'s tells a different story. $RMSE$'s are minimized with wk2, second best with wk3, third with wk4, and finally we hit wk1 at fourth best.
7. wk1 does not fare well by either measure.... *What's going on here?*
8. Earlier we showed that under certain conditions, the R^2 's and $RMSE$'s would move in opposite directions... so whichever week maximized R^2 would also minimize $RMSE$. But that is clearly not happening here.
9. *Anyone see why? What's the problem?*