Can we predict how good a movie will do ?

Alexandre Chau and Arthur Thibaud

ENAC

April 22, 2018

The Blair Witch project

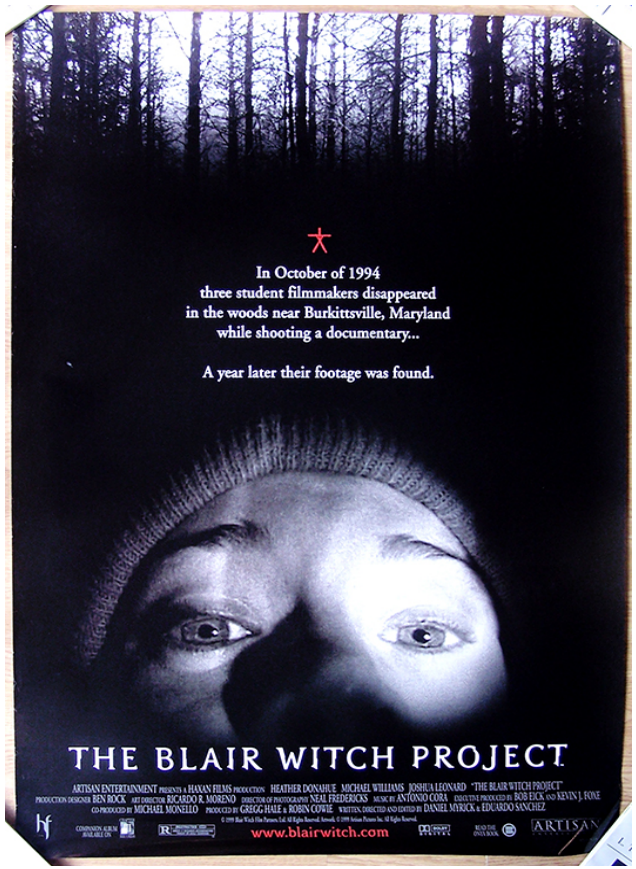

Table: A movie which made a huge profit

Our sample

Purpose

Can we predict if a movie will make money ?

Number of observations

n = 108 movies and k = 26 variables

Non-dummy variables

Variables	Description
box	Box office after a month in the US (in million dollars)
budget	Budget (in million dollars)
critics_meta	Rating of Metacritic (in %)
critics_tomato	Rating of RottenTomatoes (in %)
director	Number of nominations at the Academy Awards for Best Director
higher_budgets	Number of movies released the same month with a higher budget
length	Length of the movie (in minutes)
star	Number of Academy Award nominations for the star actor
supporting	Number of Academy Award nominations for the supporting actor
trailers	Number of trailers
year	Year of release of the movie

Dummy variables

Variables	Description
action	The movie is an action movie
adventure	The movie is an adventure movie
comedy	The movie is a comedy movie
documentary	The movie is a documentary movie
drama	The movie is a drama movie
europe	The movie is an europea movie
fantastic	The movie is a fantastic movie
horror	The movie is an horror movie
new_story	The movie is a new story
pg	The movie is a pg movie
pg_13	The movie is a pg 13 movie
r	The movie is a rated R movie
saga	The movie is a saga
thriller	The movie is a thriller movie
us	The movie is an american movie

Histogram of the box-office

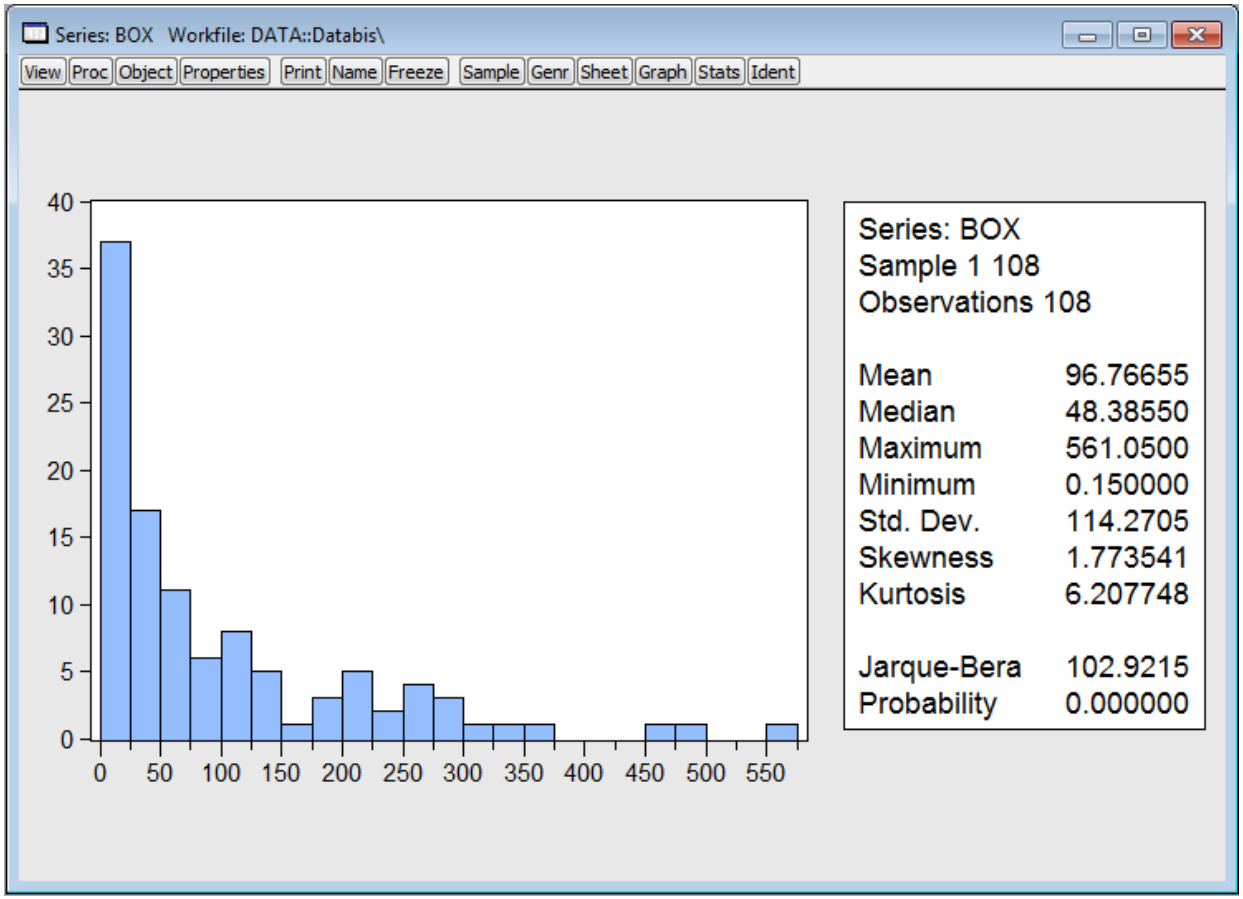

Figure: Histogram and stats for box office.

Histogram of the budget

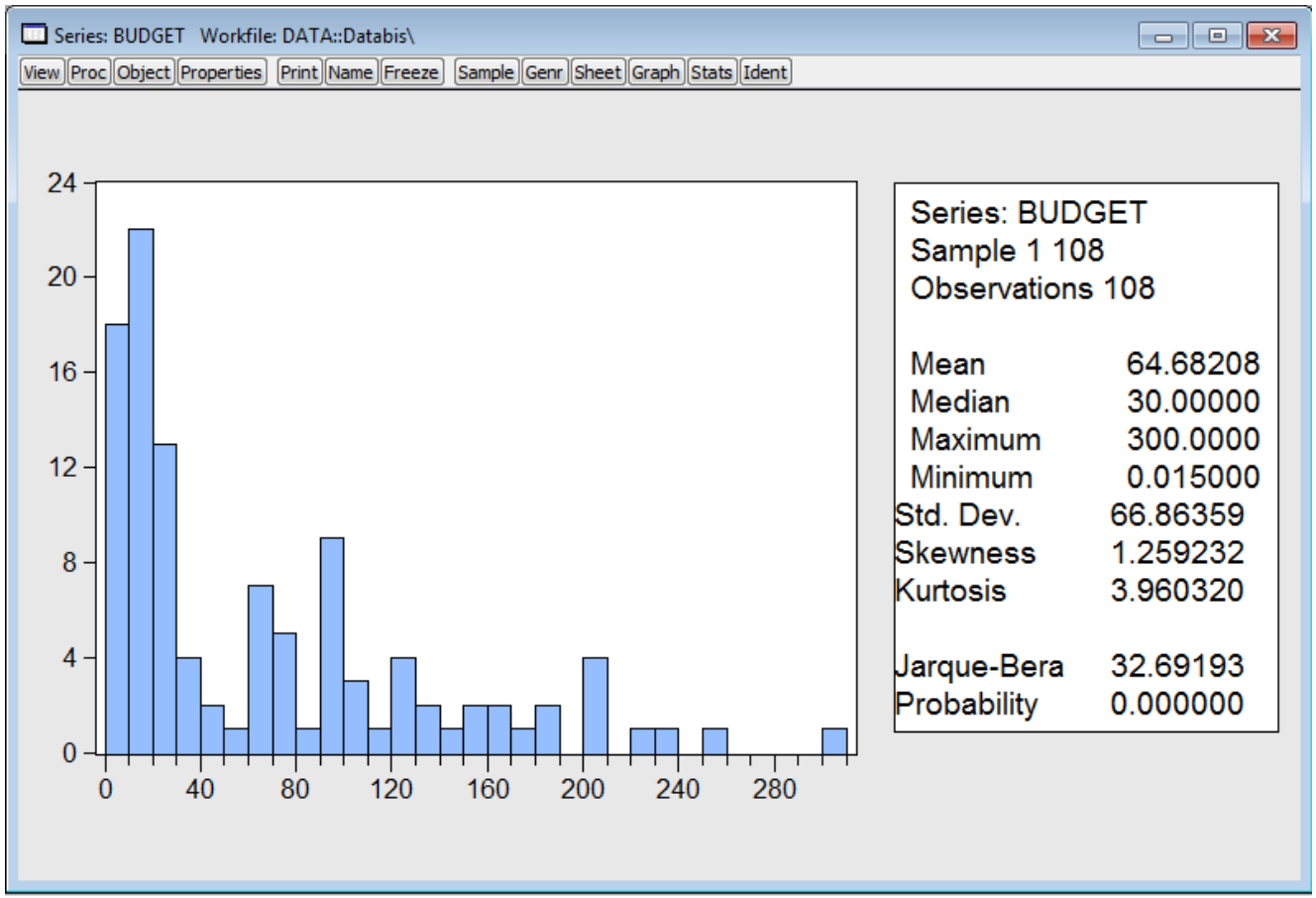

Figure: Histogram and stats for budget.

Correlation matrix

Europe and US correlation

	DRAMA	EUROPE	FANTASTIC
TRAILERS	-0.1803339269334865	0.02354408046740055	-0.05737106000964394
US	0.0530247241230202	-0.8953507092541913	0.1075715527677476
YEAR	-0.2623942516462283	-0.09010983254961958	0.09196512877685935

Critics correlation

	CRITICS_META	CRITICS_TOMATO
COMEDY	-0.2208313573413011	-0.2010949382298678
CRITICS_MET	1	0.8853466398053215
CRITICS_TOM	0.8853466398053215	1
DIRECTOR	0.3144909042241849	0.2669979811134405

First model

We started by looking how the variables were affecting the box-office

Computed equation

Dependent Variable: BOX
Method: Least Squares
Date: 04/09/18 Time: 18:03
Sample: 1 108
Included observations: 108

	Coefficient	Std. Error	t-Statistic	Prob.
C	-777.6090	1925.969	-0.403750	0.6874
ACTION	-15.73112	49.23521	-0.319509	0.7501
ADVENTURE	-38.78276	49.46251	-0.784084	0.4352
BUDGET	0.729520	0.181861	4.011422	0.0001
COMEDY	-0.134679	51.29874	-0.002625	0.9979
CRITICS_TOMATO	1.184037	0.324158	3.652650	0.0004
DIRECTOR	3.262674	4.025879	0.810425	0.4200
DOCUMENTARY	-58.25710	59.57826	-0.977825	0.3309
DRAMA	-41.94050	52.10582	-0.804910	0.4231
EUROPE	-55.06173	23.29586	-2.363584	0.0204
FANTASTIC	-1.900549	51.94742	-0.036586	0.9709
HIGHER_BUDGET	1.043033	1.789588	0.582834	0.5615
HORROR	13.44956	54.23158	0.248002	0.8047
LENGTH	0.620870	0.315778	1.966158	0.0525
NEW_STORY	-2.126899	14.93863	-0.142376	0.8871
PG_13	10.03271	21.74121	0.461461	0.6456
R	-35.75757	22.10094	-1.617921	0.1094
SAGA	62.67925	18.41238	3.404191	0.0010
STAR	0.687561	3.073105	0.223735	0.8235
SUPORTING	-7.251015	4.141676	-1.750744	0.0836
THRILLER	13.15274	53.85596	0.244221	0.8076
TRAILERS	56.09402	15.29043	3.668572	0.0004
YEAR	0.309255	0.965441	0.320325	0.7495

R-squared	0.788930	Mean dependent var	96.76655
Adjusted R-squared	0.734301	S.D. dependent var	114.2705
S.E. of regression	58.90195	Akaike info criterion	11.17607
Sum squared resid	294902.3	Schwarz criterion	11.74727
Log likelihood	-580.5079	Hannan-Quinn criter.	11.40767
F-statistic	14.44139	Durbin-Watson stat	1.848031
Prob(F-statistic)	0.000000		

1/2

Variables	Predicted effect	Actual effect	Significant
action	/	-	no
adventure	/	-	no
budget	+	+	yes
comedy	/	-	no
critics_tomato	+	+	yes
director	+	+	no
documentary	/	-	no
drama	/	-	no
europe	-	-	yes
fantastic	/	-	no
...			

2/2

Variables	Predicted effect	Actual effect	Significant
higher_budgets	-	+	no
horror	/	+	no
length	/	+	yes
new_story	+	-	no
pg_13	/	+	no
r	-	-	yes
saga	+	+	yes
star	+	+	no
supporting	+	-	yes
thriller	/	+	no
trailers	+	+	yes
year	/	+	no

Normality of the errors

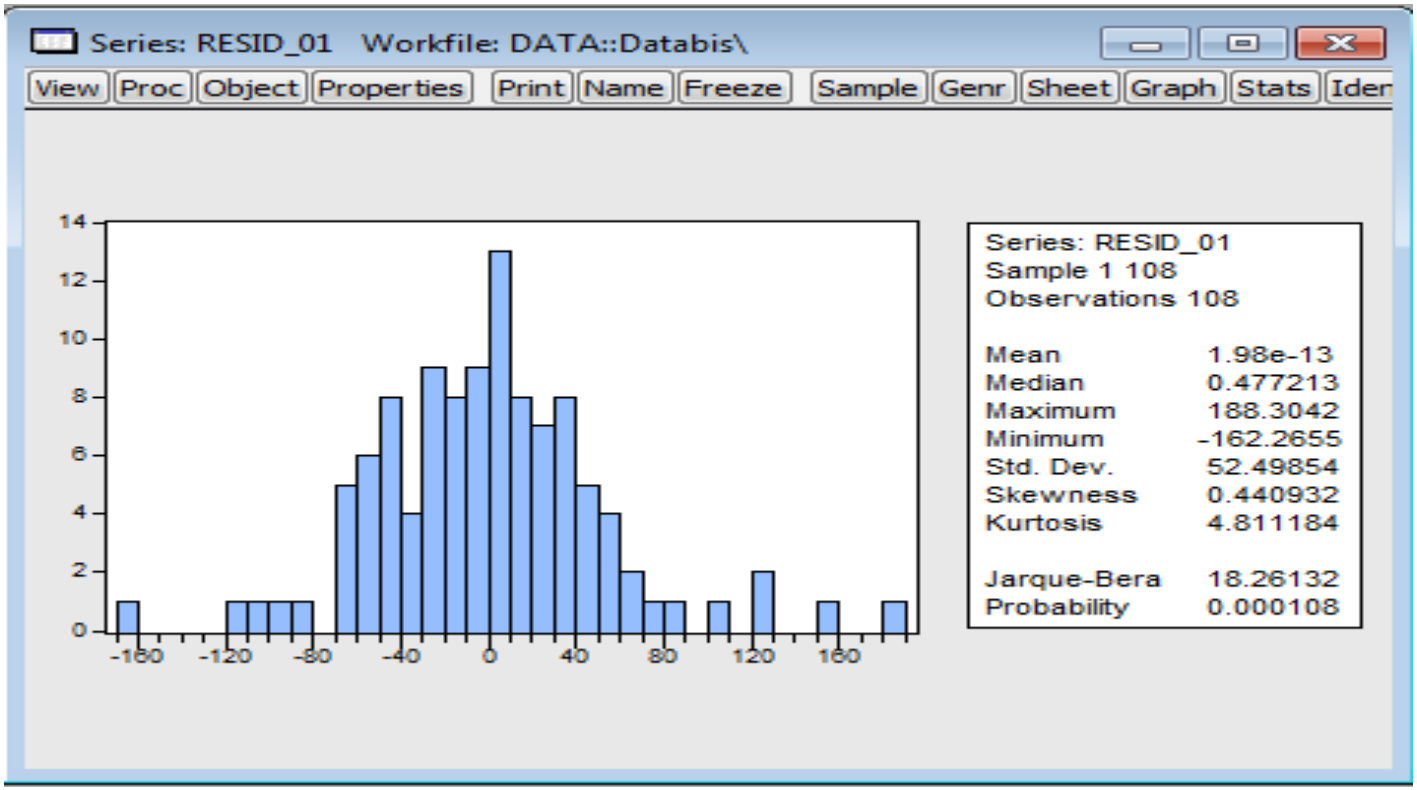

Figure: The errors don't follow a normal law.

Heteroscedasticity

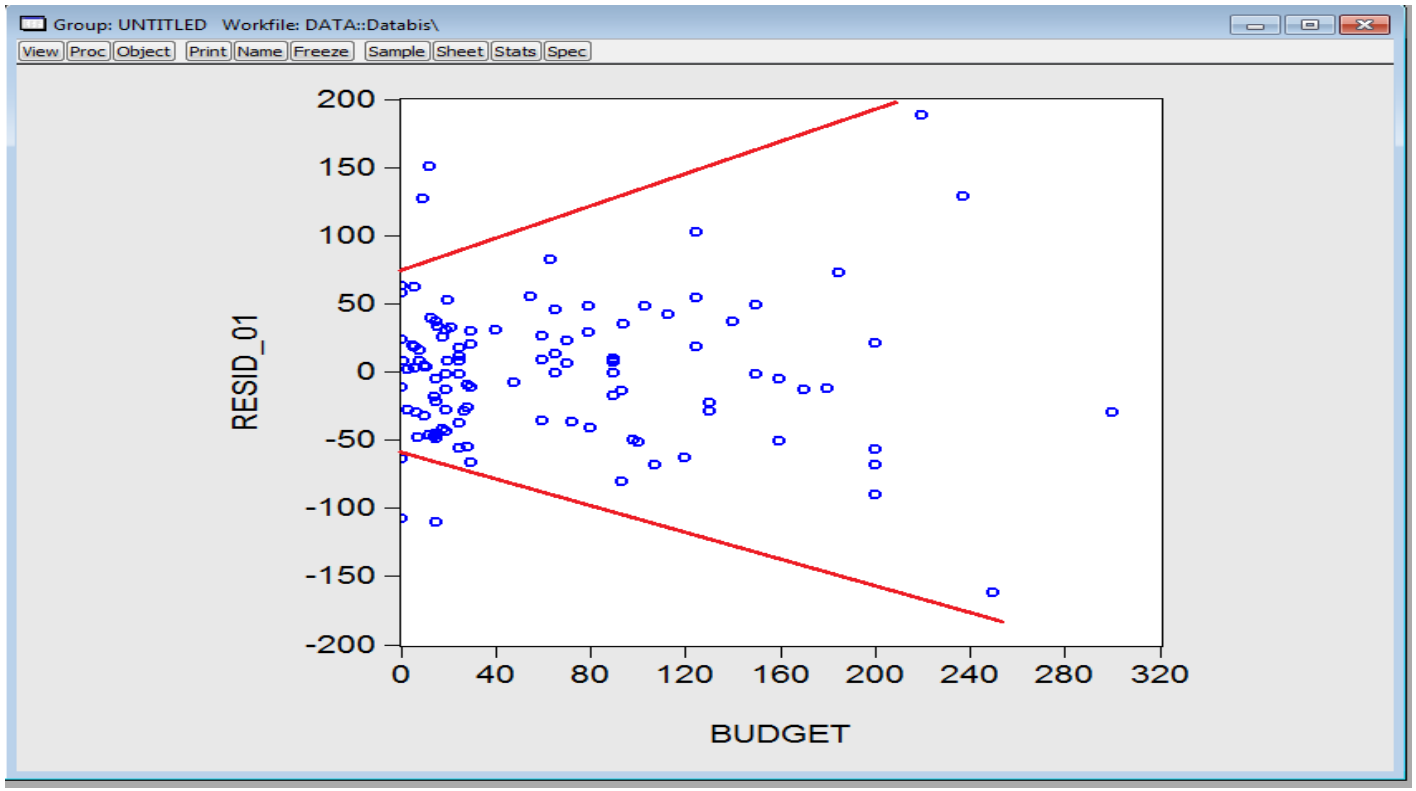

Figure: It seems that the sample contains heteroscedasticity

Heteroskedasticity

Heteroskedasticity Test: Breusch-Pagan-Godfrey
Null hypothesis: Homoskedasticity

F-statistic	1.375470	Prob. F(24,83)	0.1457
Obs*R-squared	30.73165	Prob. Chi-Square(24)	0.1617
Scaled explained SS	35.25142	Prob. Chi-Square(24)	0.0648

Standard errors

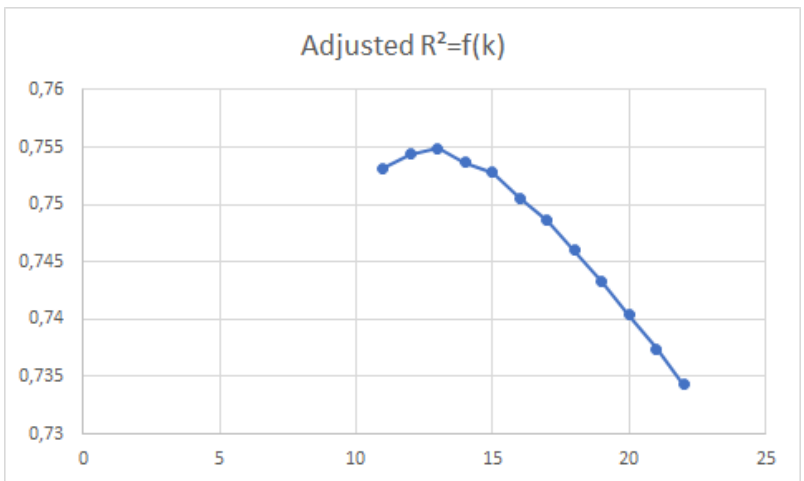

Table: Adjusted R squared

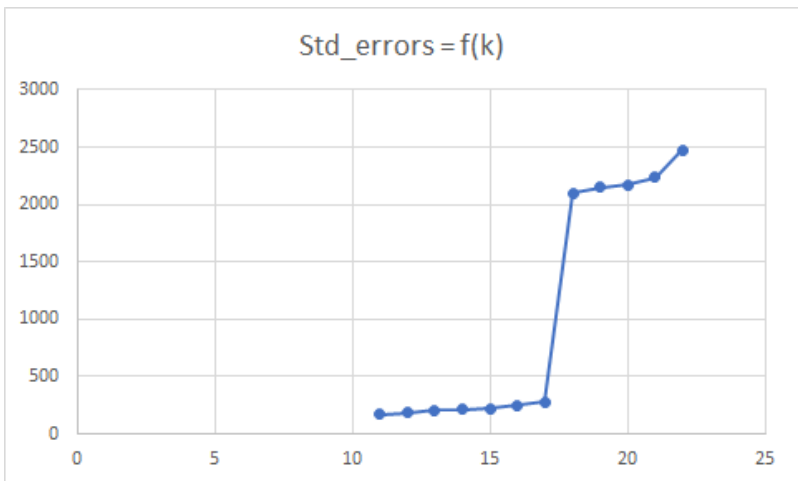

Table: Standard errors

Improvement of first model

We recomputed the box-office, but this time only against the variables that were significant i.e. budget, critics_tomato, europe, length, r, saga, supporting and trailers.

Computed equation

Dependent Variable: BOX
Method: Least Squares
Date: 04/10/18 Time: 19:01
Sample: 1 108
Included observations: 108

	Coefficient	Std. Error	t-Statistic	Prob.
C	-139.8412	31.67371	-4.415058	0.0000
BUDGET	0.741258	0.134018	5.531052	0.0000
CRITICS_TOMATO	1.016257	0.271338	3.745359	0.0003
EUROPE	-44.63463	21.34427	-2.091177	0.0391
LENGTH	0.416991	0.241335	1.727850	0.0871
R	-32.32459	13.88932	-2.327298	0.0220
SAGA	73.04915	17.26590	4.230833	0.0001
SUPORTING	-7.387048	3.769306	-1.959790	0.0528
TRAILERS	60.05990	13.09265	4.587298	0.0000

R-squared	0.754700	Mean dependent var	96.76655
Adjusted R-squared	0.734878	S.D. dependent var	114.2705
S.E. of regression	58.83794	Akaike info criterion	11.06711
Sum squared resid	342728.4	Schwarz criterion	11.29062
Log likelihood	-588.6237	Hannan-Quinn criter.	11.15773
F-statistic	38.07341	Durbin-Watson stat	1.882045
Prob(F-statistic)	0.000000		

Rentability

We created a second model, this time in order to study the rentability, which is the ratio between the box-office and the budget.

EViews modelization

Dependent Variable: BOX/BUDGET
Method: Least Squares
Date: 04/10/18 Time: 18:42
Sample: 1 108
Included observations: 108

	Coefficient	Std. Error	t-Statistic	Prob.
C	1154.717	12204.42	0.094615	0.9248
ACTION	187.6873	335.3609	0.559658	0.5772
ADVENTURE	67.70578	338.8216	0.199827	0.8421
COMEDY	119.6696	341.0622	0.350873	0.7265
CRITICS_TOMATO	1.721387	2.113304	0.814548	0.4176
DIRECTOR	1.939533	27.43265	0.070702	0.9438
DOCUMENTARY	-56.69779	406.6884	-0.139413	0.8894
DRAMA	174.1303	347.5416	0.501034	0.6176
FANTASTIC	75.89039	352.2889	0.215421	0.8299
HIGHER_BUDGET	44.77728	11.06993	4.044947	0.0001
HORROR	748.4930	368.2105	2.032786	0.0452
LENGTH	2.399396	2.026647	1.183924	0.2397
NEW_STORY	114.0454	101.8538	1.119697	0.2660
PG_13	-190.2847	148.2985	-1.283119	0.2029
R	-95.02911	152.9030	-0.621499	0.5359
SAGA	222.9842	121.2578	1.838926	0.0694
STAR	-25.31536	21.08197	-1.200806	0.2331
SUPORTING	-12.96895	28.22687	-0.459454	0.6471
THRILLER	298.3045	362.7012	0.822453	0.4131
TRAILERS	109.4800	104.4004	1.048655	0.2973
US	284.2005	150.2332	1.891730	0.0619
YEAR	-1.107652	6.100227	-0.181576	0.8563

R-squared	0.405401	Mean dependent var	64.01138
Adjusted R-squared	0.260208	S.D. dependent var	468.8986
S.E. of regression	403.3051	Akaike info criterion	15.01689
Sum squared resid	13988328	Schwarz criterion	15.56325
Log likelihood	-788.9119	Hannan-Quinn criter.	15.23842
F-statistic	2.792158	Durbin-Watson stat	2.510650
Prob(F-statistic)	0.000451		

1/2

Variables	Predicted effect	Actual effect	Significant
action	/	+	no
adventure	/	+	no
comedy	/	+	no
critics_tomato	+	+	no
director	+	+	no
documentary	/	-	no
drama	/	+	no
fantastic	/	+	no
...			

2/2

Variables	Predicted effect	Actual effect	Significant
higher_budgets	-	+	yes
horror	/	+	yes
length	/	+	yes
new_story	+	+	yes
pg_13	/	-	yes
r	-	-	no
saga	+	+	yes
star	+	-	yes
supporting	+	-	no
thriller	/	+	no
trailers	+	+	yes
us	/	+	yes
year	/	-	no

Heteroscedasticity test

Heteroskedasticity Test: Breusch-Pagan-Godfrey
Null hypothesis: Homoskedasticity

F-statistic	3.303926	Prob. F(21,86)	0.0000
Obs*R-squared	48.22491	Prob. Chi-Square(21)	0.0006
Scaled explained SS	539.7074	Prob. Chi-Square(21)	0.0000

Improvement of second model

We recomputed the rentability, but this time only against the variables that were siginificant i.e. higher_budgets, horror, length, new_story, r, saga, star, trailers and us.

EViews modelization

Dependent Variable: BOX/BUDGET
Method: Least Squares
Date: 04/10/18 Time: 18:46
Sample: 1 108
Included observations: 108

	Coefficient	Std. Error	t-Statistic	Prob.
C	-748.4017	273.9199	-2.732191	0.0075
HIGHER_BUDGET	40.20555	9.092674	4.421752	0.0000
HORROR	598.9641	145.8630	4.106346	0.0001
LENGTH	2.217617	1.664107	1.332617	0.1857
NEW_STORY	101.5244	91.85325	1.105289	0.2717
PG_13	-135.9770	86.73875	-1.567661	0.1202
SAGA	186.7799	103.7846	1.799688	0.0750
STAR	-26.60021	18.45892	-1.441049	0.1528
TRAILERS	110.2166	86.21121	1.278448	0.2041
US	223.7647	130.9750	1.708453	0.0907

R-squared	0.382173	Mean dependent var	64.01138
Adjusted R-squared	0.325433	S.D. dependent var	468.8986
S.E. of regression	385.1158	Akaike info criterion	14.83299
Sum squared resid	14534791	Schwarz criterion	15.08133
Log likelihood	-790.9813	Hannan-Quinn criter.	14.93368
F-statistic	6.735599	Durbin-Watson stat	2.479994
Prob(F-statistic)	0.000000		

Comparison of the models

Forecasting

For "Batman V Superman"
Our prediction : 1 192 M$
2 weeks after the release of the movie: 710 M$

22/23

Table: iENAC14 model : Prediction/reality = 1.36

Reality

Box-office after 1 month: $321,322,593 , rentability : 1.29

Comparison of the models 1 : Batman vs Superman

First models

Box-office prediction = \$263,781,100

Ratio : 0.82 (adjusted R^2 : 0.734)

Box-office prediction = \$296,722,240

Ratio : 0.92 (adjusted R^2 : 0.735)

Second models

Rentability prediction = 217.43, ratio : 168.55

Rentability prediction = 40.68, ratio :31.5

Pretty bad modelization of rentability.

Comparison of the models 2 : Blair Witch Project

Real box office

$113,380,000

First models

1- Box-office prediction = $20,272,040
Ratio : 0.18 (adjusted R^2 : 0.734)
2- Box-office prediction = $173,379,800
Ratio : 1.53 (adjusted R^2 : 0.742)

Second models

1- Rentability prediction = 1109, ratio : 0.48
2- Rentability prediction = 1080, ratio : 0.46
Better modelization than the first movie.

Limits

Limits of our models