Am I Washed Up?

..the dilemma

Can we predict the peak time to cast an actor in order to maximize revenue?

Things I was thinking about

Does it benefit a director to choose his or her actors based on the length of time since their last film?

Are actors that appear in movies more frequently beneficial to the success of their movie, or should the director go out on a limb to give a has-been actor an opportunity for a comeback?

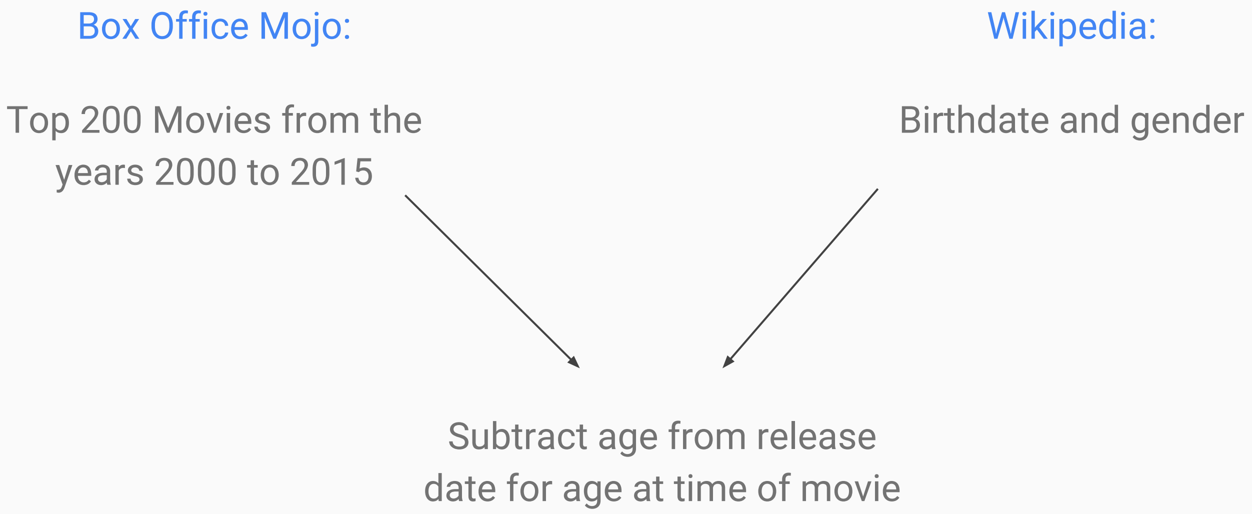
Does this kind of data align in any way?

Approach

I will be using linear regression to determine if there is a relationship between the time passed since an actor’s last movie and the overall success of the movie they have been casted in.

Using this model, can we predict the peak time between movies in which to cast an actor in order to gross the most revenue?

Data Sources:



Box Office Mojo:

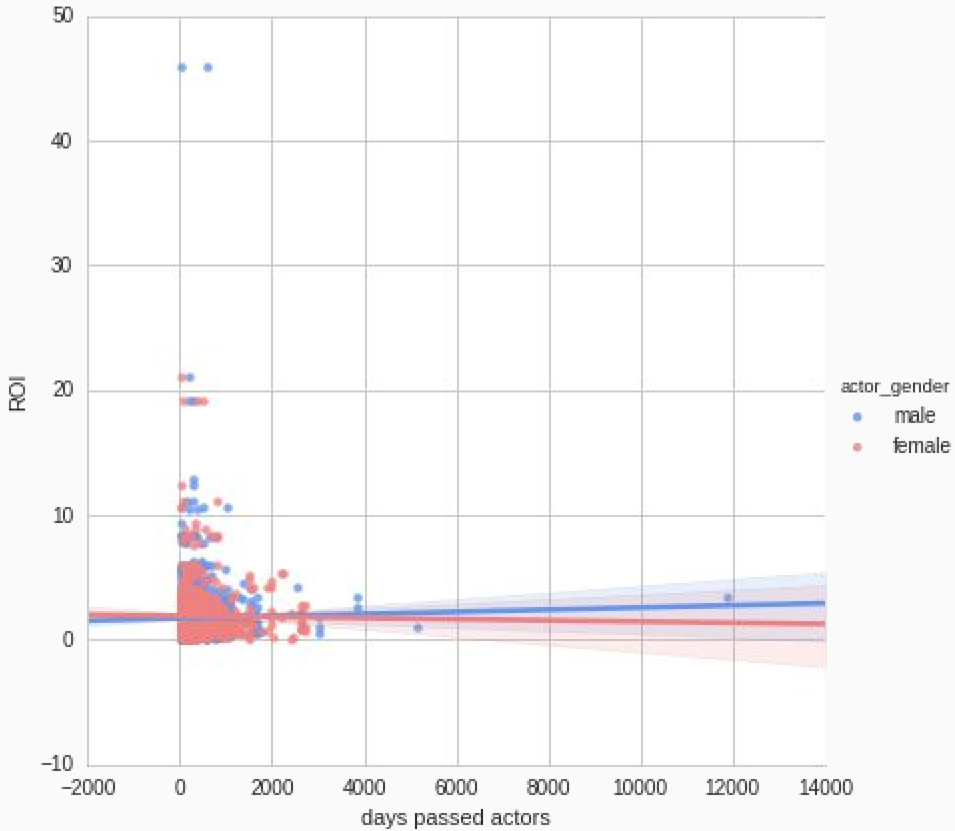
Top 200 Movies from the years 2000 to 2015

Wikipedia:

Birthdate and gender

Subtract age from release date for age at time of movie

Linear Regression of Days Passed Since Last Movie and ROI



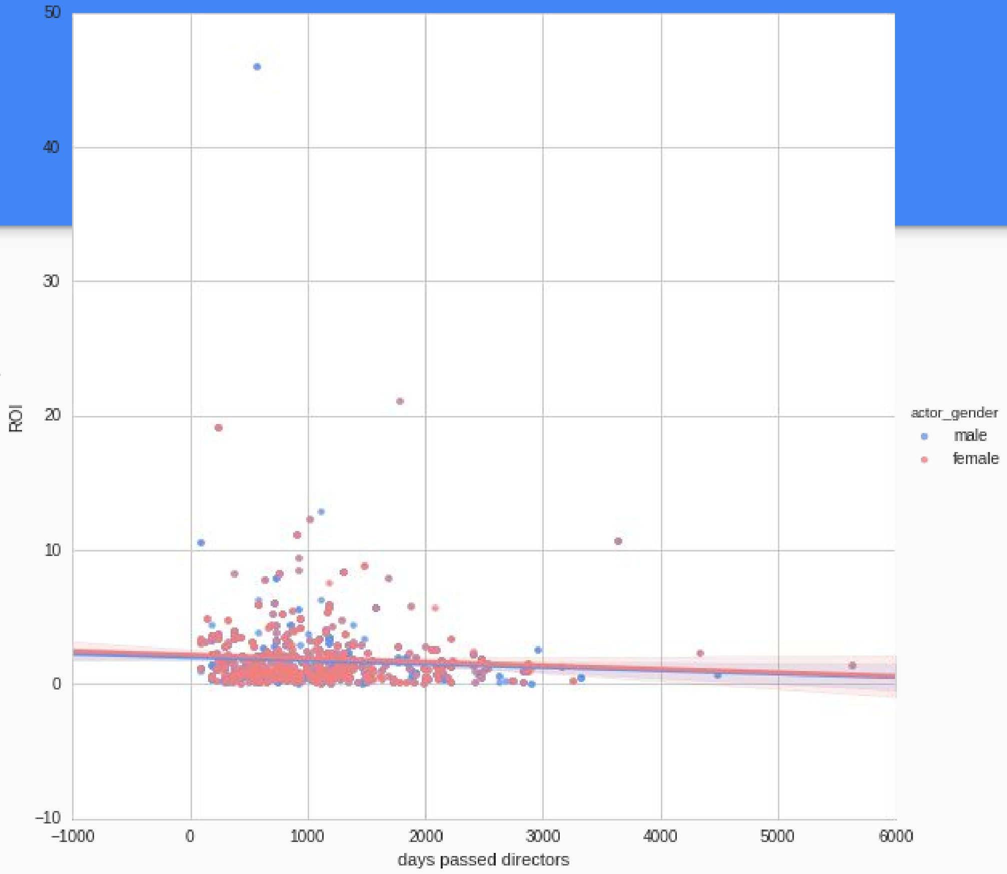
At first glance, there does not appear to be any correlation. ..there isn’t.

More Investigation

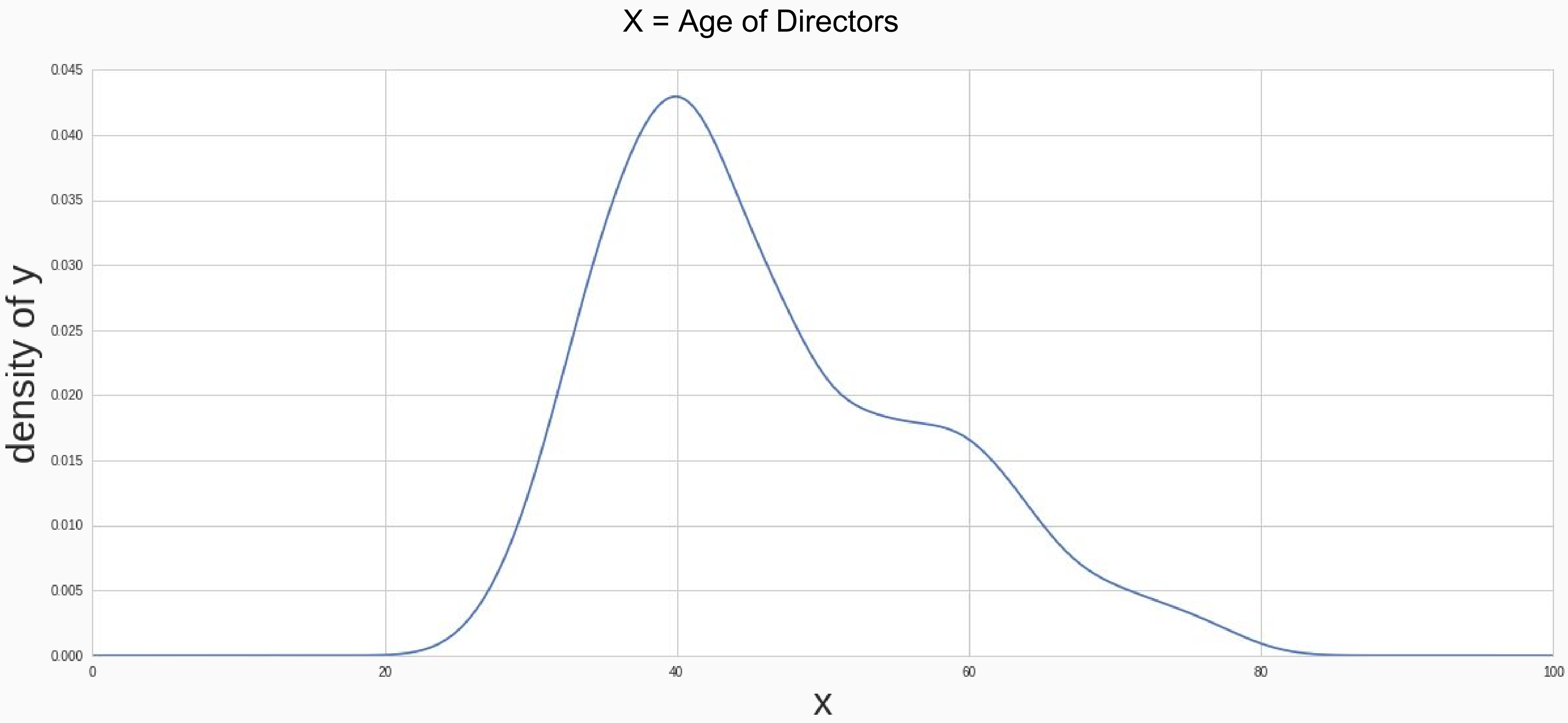
As it turns out, actors' time since their last movie had no effect on how well the movie was going to perform.

Looked into director attributes and other actor characteristics.

Director features seemed to have better predictive qualities than actors.



Density curve of the age of directors in my sample data set



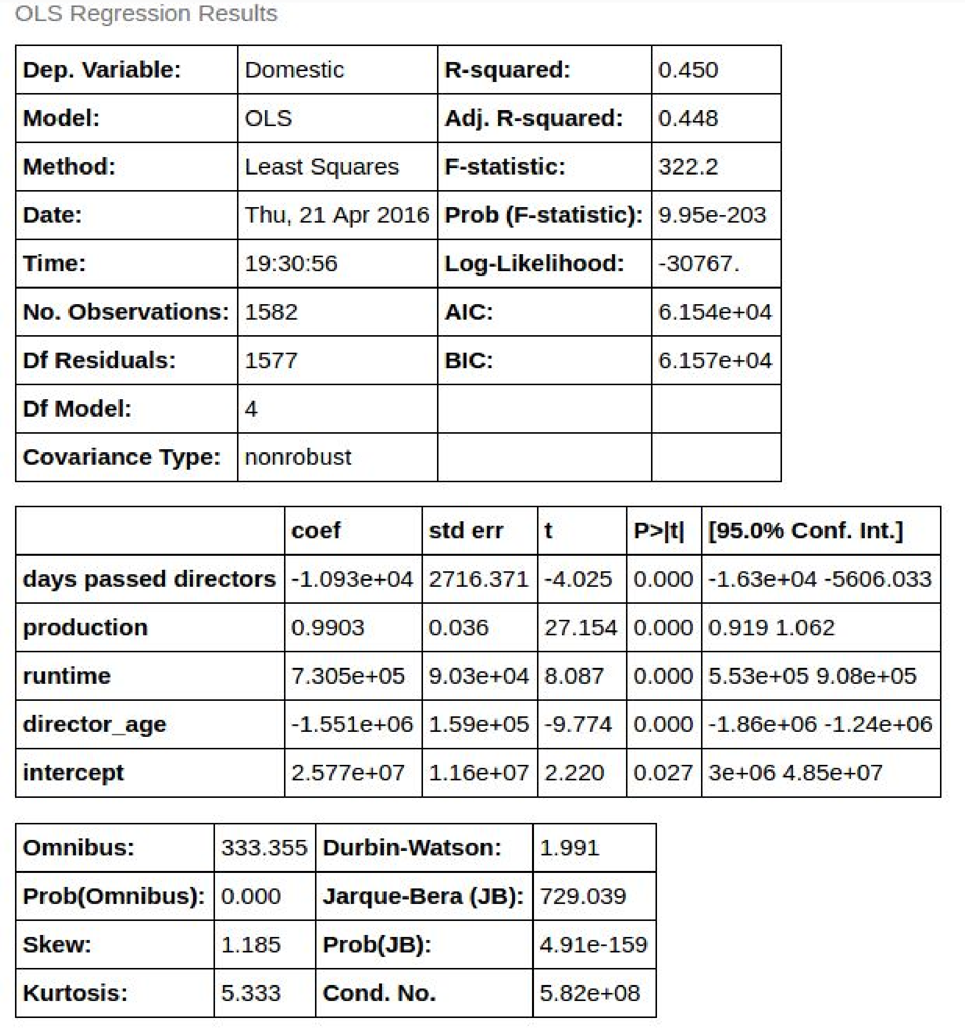
My Model

Dependant Variable:

Y = Total Domestic Gross

Features I decided to include:

1. Days passed since the director’s last movie
2. Age of director at time of movie
3. Production Budget
4. Movie runtime



OLS Regression Results

Dep. Variable:	Domestic	**R-squared:**	0.450
Model:	OLS	**Adj. R-squared:**	0.448
Method:	Least Squares	**F-statistic:**	322.2
Date:	Thu, 21 Apr 2016	**Prob (F-statistic):**	9.95e-203
Time:	19:30:56	**Log-Likelihood:**	-30767.
No. Observations:	1582	**AIC:**	6.154e+04
Df Residuals:	1577	**BIC:**	6.157e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
days passed directors	-1.093e+04	2716.371	-4.025	0.000	-1.63e+04 -5606.033
production	0.9903	0.036	27.154	0.000	0.919 1.062
runtime	7.305e+05	9.03e+04	8.087	0.000	5.53e+05 9.08e+05
director_age	-1.551e+06	1.59e+05	-9.774	0.000	-1.86e+06 -1.24e+06
intercept	2.577e+07	1.16e+07	2.220	0.027	3e+06 4.85e+07

Omnibus:	333.355	**Durbin-Watson:**	1.991
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	729.039
Skew:	1.185	**Prob(JB):**	4.91e-159
Kurtosis:	5.333	**Cond. No.**	5.82e+08

Other Interesting Stuff..

r2_score(y_test, y_predictTest) = 0.442935

ROI as a function of actor age

Director StripPlot

Other Considerations:

Normalizing the feature set in order to interpret them on a common scale and to align their distributions.

Inflation adjustments

Incorporate variable selection techniques to increase prediction accuracy