Predicting the Ratings of Movie Remakes

Andrea Everett
10/4/16
Project Luther MVP

Motivation:

It has been said that in Hollywood, imitation is the most profitable form of flattery. While I may not be able to calculate profit, I aim to investigate what we can learn about viewer ratings of movie remakes based on a combination of information about the original film and various characteristics of the remake itself. To this end, I scraped the Internet for data on about 530 pairs of movie originals and remakes. I have a few specific hypotheses that I want to test, of which I will mention two here. First, I think that holding other characteristics of a film constant, viewer ratings should reflect the quality of the story. Therefore, if I can control for a number of important characteristics of the remake, viewer ratings for the original should reflect the quality of the shared story underlying both movies and should positively predict viewer ratings for the remake. Second, it is plausible that stories will not resonate as well on average in a different language or with a different culture as in the original language. If this is the case, then remakes for which the language is the same as the original should receive higher ratings than those for which it is different.

Description

Below I present a regression table for a simple linear regression that begins to test these hypotheses and also includes a couple of other independent variables. In this table:

y = imdbRating of the remake
x1 = imdbRating of the original
x2 = Award Nominations + Awards Received of the original movie
x3 = Runtime of the remake in minutes (length of film may affect user ratings)
x4 = An indicator variable for whether the original and remake are in the same language

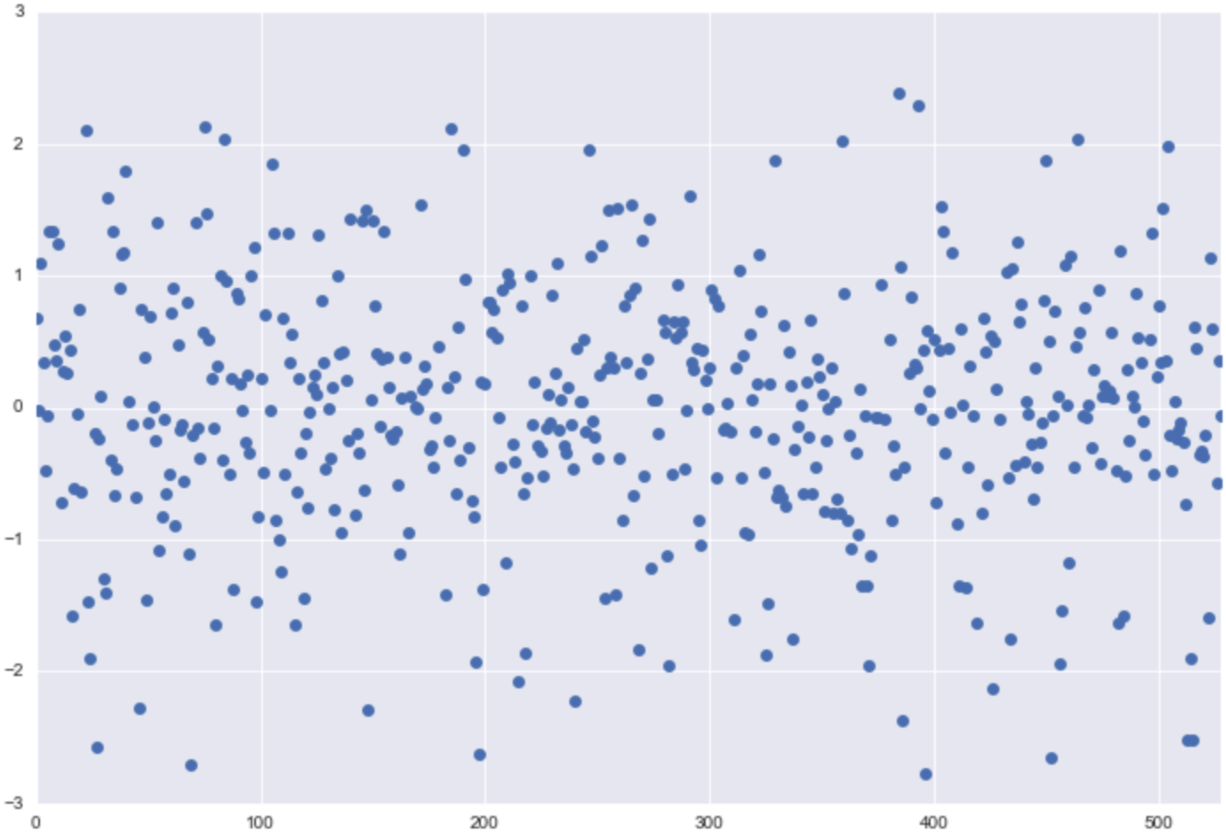
I haven't had time to do anything fancier than this yet. But we can see that these four variables (without yet considering potential non-linearity of the relationships) explain about 15% of the variation in the imdbRating of the remakes. I expect to be able to do better with some additional variables that I am working on. The residual plot below suggests to me that the residuals aren't random: their absolute values are generally concentrated more around the center of the plot than at the outer edges.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.145
Model:	OLS	**Adj. R-squared:**	0.138
Method:	Least Squares	**F-statistic:**	20.87
Date:	Tue, 04 Oct 2016	**Prob (F-statistic):**	6.57e-16
Time:	18:40:17	**Log-Likelihood:**	-664.53
No. Observations:	499	**AIC:**	1339.
Df Residuals:	494	**BIC:**	1360.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	2.5307	0.419	6.044	0.000	1.708 3.353
x1	0.4119	0.056	7.400	0.000	0.303 0.521
x2	-0.0044	0.003	-1.359	0.175	-0.011 0.002
x3	0.0065	0.002	3.876	0.000	0.003 0.010
x4	-0.0755	0.091	-0.826	0.409	-0.255 0.104

Plot of residuals from the above regression


3
2
1
0
−1
−2
−3
0
100
200
300
400
500