Econometrics Session 4 - R Exercise

Prof. Jan Sebastian Nimczik

You are a data analyst working in the movie industry in Hollywood. You have at your disposal a dataset, *hollywood.csv*, on 74 movies which opened in the US. It contains containing data on movie characteristics and performance in the box office. These are the variables in the data set:

Variable name	Variable Description
movie	name of movie
budget	production budget ($)
openinggross	gross revenues from movie's box office opening ($)
openingtheaters	number of theaters in which the movie opened
knownstory	=1 if movie is based on a known story (e.g. a book), = 0 otherwise
sequel	yes if movie is a sequel, no otherwise
opensummer	=1 if movie opened during the summer, =0 otherwise
openholiday	=1 if movie opened on a public holiday, =0 otherwise
openchristmas	=1 if movie opened during Christmas holidays, =0 otherwise
criticsopinion	index of critics' ratings of the movies ranging from 0 (horrible) to 100 (brilliant)
comedy	=1 if the movie is a comedy, =0 otherwise

You want to understand which variables are associated with high gross revenues (openinggross).

1. Write down a regression model, which includes the dependent variable on the left-hand side of the equation and the available dependent variables on the right-hand side of the equation. Provide some common-sense (economic) intuition for the inclusion of each of these variables in your regression model, explaining the direction of relationship you conceptually anticipate between each explanatory variable and your dependent variable.
2. Estimate this model; show your regression output. In your view, does this model do a "good job" of explaining the data? Justify your answer.
3. Include an interaction term between opensummer and openholiday. How do you interpret the coefficient on the interaction? Include an interaction term between the criticsopinion and sequel. How do you interpret the coefficient on the interaction? Include a polynomial of degree two in openingtheaters in your model (i.e. opiningtheaters and the squared number of theaters). How do you interpret the coefficients? Do these additional terms increase the adjusted R^2?

Now, you are hired by a Hollywood producer who is considering investing in a new movie production. The proposed film project will require a production of $40 million; this budget is

non-negotiable. The movie is based on a book and is a fresh project (i.e., it is not a sequel). Since it is a comedy, the producer intends to open the film during the summer in 2,800 theaters across the U.S.; he does not plan to open on a public holiday. The producer has informed you that he needs to make at least $10 million in gross revenues from the movie's box office opening to make this investment worthwhile.

Your task is to help him decide whether or not--or under what circumstances--he should undertake this movie project using regression analysis.

4. Predict opening box office revenues for the proposed film project, along with the 95% *confidence* interval. (Hint: you are not provided information on the critics' opinion you need for this prediction, so consider different scenarios.)

5. According to your analysis, should the producer invest in this movie project? If so, under what circumstances?