Inferential Statistical Analysis for Generating a Prediction Model

To develop the best model for predicting Metacritic review scores on video games, I need to verify which variables have a statistically significant relationship with Critic Scores variable. A chi-squared goodness of fit test allows us to test the statistical significance of a relationship between categorical variables. I will run the test on several variables within the data frame to see which variables should be included when creating the final model. To be clear, for each variable that is tested our null and alternate hypothesis is as follows:

H0: There is no relationship between the variable and Game Quality (Categorized Critic Scores).

Ha: There is a relationship between the variable and Game Quality (Categorized Critic Scores).

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Chi Statistic | P-Value | Expected Value |
| Genre | 61.9599558931 | 0.00166336164951 | 33 |
| Platform | 26.4503350982 | 0.493729754709 | 27 |
| Rating | 33.5884026466 | 0.00010543964857 | 9 |
| Developer size | 11.1890509966 | 0.0107463201464 | 3 |
| Critic Count | 436.95798475 | 3.28801958791e-08 | 288 |
| Publisher size | 14.2250362929 | 0.00261430484993 | 3 |

After running a couple of chi-squared goodness of fit tests, I found that the variables 'Genre', 'Rating', and 'Critic Count' have a statistically significant relationship with 'Game Quality'(Categorized Critic Scores). The 'Platform' variable doesn't have a statistically significant relationship with 'Game Quality' based on the p-value of 'Platform' surpassing .05 alpha threshold. The test of 'Developer size' (Whether a game was made by a major developer or not) and 'Publisher size'(Whether a game is published by a major publisher or not) relationship with 'Game Quality' isn't valid since the expected value for both variables is 3 and need to be greater than 5 to be considered valid.

Next, I wanted to explore the relationship between critic score and YSR (years since platform release). I predict that there should be a positive correlation between YSR (the difference in platform release date and game release date). As console get older, scores should increase as developers become more familiar with each platforms capability. Surprisingly, I found almost no correlation between Critic Score and YSR. The calculated correlation from the data set is -0.04164. This leads me to believe that is should consider leaving YSR out of the variables used to construct my predictive model.