Predicting the Banker's Offer in Deal or No Deal

Jake Watts

I. Introduction

Deal or No Deal was a television game show hosted by Howie Mandel that was on air from 2005 to 2019. In Deal or No Deal contestants have the opportunity to play for a prize that ranges anywhere from one cent to one million dollars. The game involves 26 cases, each representing a different amount of money. The only players are the contestant and the banker. The contestant's objective is to maximize winnings while the banker's objective is to leave with the least amount of money possible. At the start of the game the player selects one of the 26 briefcases containing an unknown amount of money. Then the player selects cases to be removed from play during each round. At specified intervals the banker offers the player a specific amount of money to stop playing the game. The player can choose to accept the offer of keep playing in hopes of earning more. If the player does not accept any offer from the banker, the game is over when there is only one case remaining. The player can then take the money in either the case they selected or the one remaining case.

In this paper, I investigate the functional relationship of the banker's offer and attempt to predict the offer that the banker will give at specific points in the game. The banker's offer amount is an important decision because if the offer is too high the player will accept and possibly make more from the banker than they would from playing. While if the offers are too low the player will never accept and could win more than the banker wants by finishing the game. Therefore, the bank offer is ideally an equilibrium point to minimize the risk of the television show paying out an extraordinary amount of money to the contestants.
Estimating the banker's offer will give insight into both the bankers and players perception of risk and expected earnings or losses.

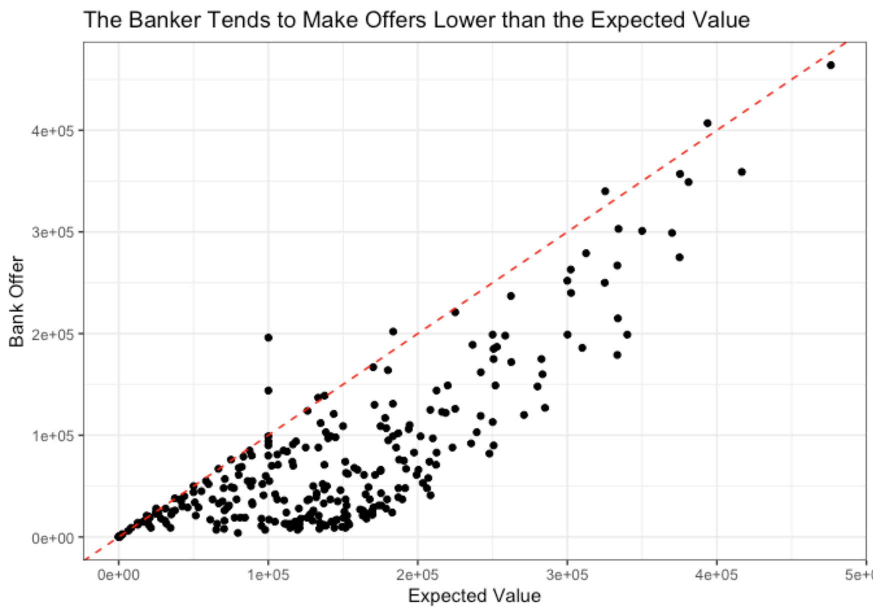
The banker's formula or decision-making process is not publicly known. It is also not certain whether the banker's offer follows a strict formula to calculate his offer or whether the banker uses his discretion to decide the offer. There are not any published papers that estimate the functional form of the banker's offer however there are several bloggers online that have attempted to find the formula for the bankers offers. To determine the validity of the proposed formulas, I divide the data into a training and testing set. Then I model the data and compare my model with another model proposed online to see which formula performs the best on the testing set. I use data from the first season of the show. The data is provided by the American Economic Association and was used in a paper entitled "Deal or No Deal? Decision Making under Risk in a Large-Payoff Game Show" that was published by the journal. I find that the banker's offer is best modeled by a non-parametric function of the expected value of the cases and the round of the offer.

II. Methods

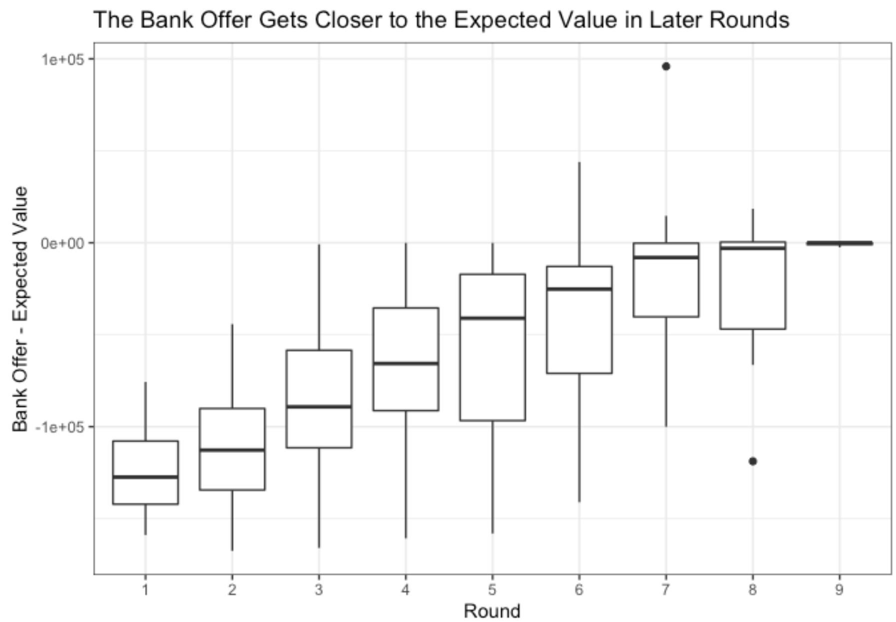
The dataset used for this analysis contained data from the first season of Deal or No Deal. The data contains 402 rows each representing a round of Deal or No Deal from 51 different contestants. The dataset contains 37 row including demographic information, such as name, age, gender and education level, as well as information about the game such as round number, bank offer, amount won and a binary column for each of the 26 cases to indicate whether that case was eliminated or not in the current round. The data was also initially split into two different sheets as there were a few contestants in one sheet that had their highest case value 1.5 million rather than one million. To prepare the dataset for analysis, I combined the two datasets together and calculated the expected value of the cases remaining in game play and randomly chose 70% of the players to go into the training data and the remaining in testing. The expected value is given by the formula below where n is the number of cases remaining and p is the probability that a remaining case was selected by the contestant.

$$E[RemainingCases] = \sum_{i=1}^{n} p * CaseValue_i$$

While exploring the training dataset, I found two variables that seemed to have a strong relationship with the banker's offer. The first being the expected value of the offer. The graph below shows the relationship between the expected value and the banker's offer. From the graph below we can see that there is a positive relationship between the expected value and the bankers offer. The dotted red line represents the point at which the expected value is equal to the banker's offer. In general, the banker makes offers lower than the expected value, although there are some exceptions shown in the graph that are above the expected value. The relationship also appears to be nonlinear as the banker's offer tends to be closer to the expected value for points at the high or low end of the graph but further below expected value on the rest of the graph. Due to the non-linear relationship between expected value and bank offer, I chose to use a non-parametric model to estimate the relationship between these variables.



The other variable that appeared to have a strong relationship with the banker's offer was the round of the game in which the offer was being made. This relationship also seemed to explain some of the variation in difference between the bank offer and expected value which is shown in the graph below. As the round increases the difference between the expected value and the banker's offer appears to shrink. This is likely because the number of cases remaining in each round decreases the banker and the contestant will both be able to more accurately guess which cases the contestant has chosen. It can also be seen from the graph that the variance appears to be the highest in the middle rounds while the variance in the first and last few rounds is comparatively small. Based on my exploratory data analysis, I decided that the most appropriate model would be a generalized additive model so that I could include both a predictor that has a non-linear relationship and predictor that is stored as a factor with nine levels.



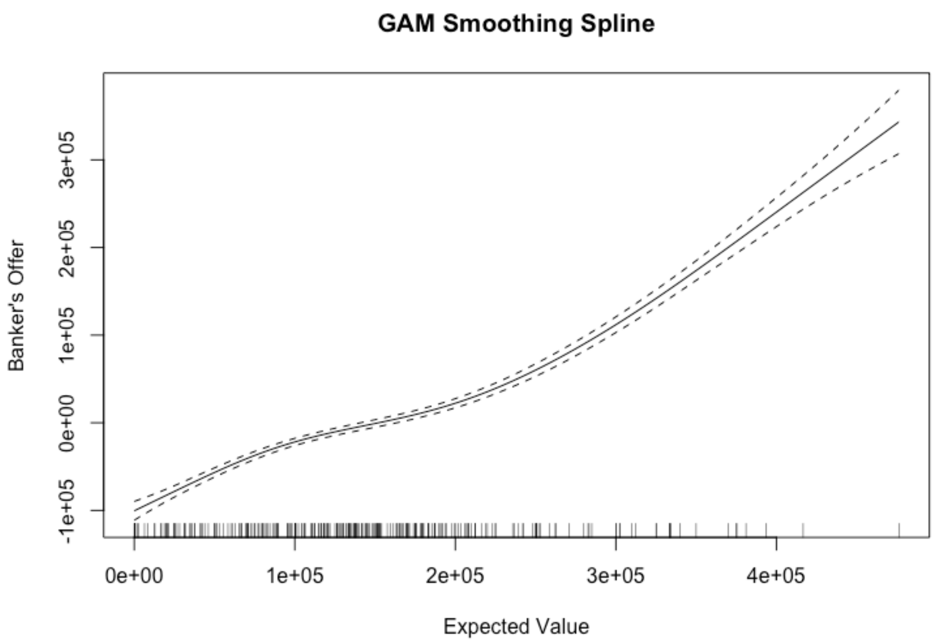
To access the performance of the model after training my generalized additive model, I test the predictions of the model and calculate the mean squared prediction error. I then compare my model with the most prominent proposed formulas for the banker's offer that I could find. Since the models are simply being used as a benchmark to gauge my model's performance I will not go into as great of detail for this model as my own, however I think it's valuable to consider how others have attempted to calculate the banker's offer. The most detailed analysis comes from a blog post from Texas Tech professor Samuel Bradley entitled "Deal or No Deal Banker's Formula". The formula he gives is a linear regression of the following form:

$$\text{Banker's Offer} = \beta_0 + \beta_1 ExpectedValue + \beta_2 (ExpectedValue)^2 + \beta_3 MaxCaseValue + \beta_4 NumberOfCasesRemaining + \beta_5 (NumberOfCasesRemaining)^2$$

My goal in training and testing the model was to create one that performs on par with this model as the author claimed this model accurately explained the majority of the variation in the banker's offer. I also hypothesized that a simpler functional form was possible because I found it unlikely that the studio used such a complicated functional form to make offers. My hypothesis was that the banker saw the expected value of the cases and deviated from this value based on heuristics rather than an exact formula.

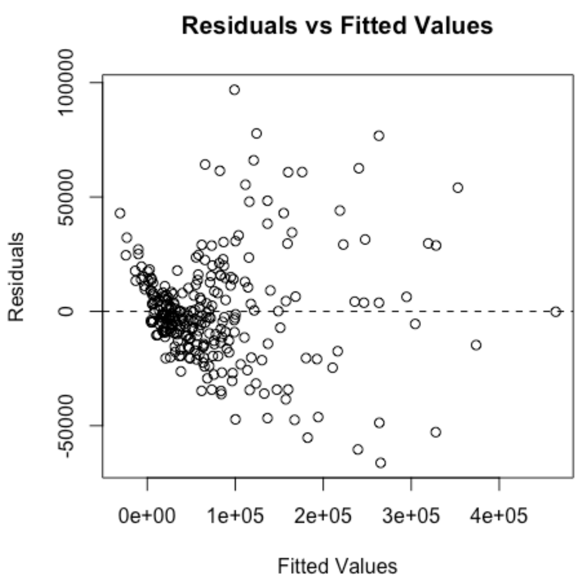
III. Results

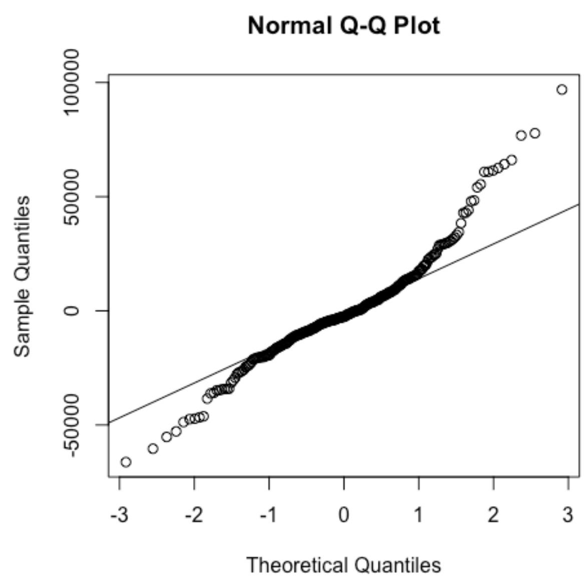
The final generalized additive model includes the expected value as a non-parametric predictor and the rounds treated as a factor variable. The graph below shows the smoothing spline created using the training data. The smoothing spline appears to capture the relationship between the expected value and the banker's offer relatively accurately. The expected value is close to the banker's offer for the high and low values and lower than the banker's offer in the middle for middle values. In this sense, the pattern of the spline appears similar to the true relationship. It also appears that the confidence interval, represented by the dashed lines, becomes wider for high expected values, particular when the expected value is greater than four hundred thousand.



For the parametric coefficients, the first round was left out of the regression to avoid multicollinearity. The coefficients for the intercept and the remaining rounds in the regression were all significant at a reasonable significance level, with all p-values being lower than 0.001. Based on the coefficients for rounds it appears that, holding the expected value constant, the highest offer comes in round seven, since the model estimates an average offer that is 100,568 dollars higher than the offer in the first round which is estimated to be 20,376 dollars. The adjusted r-squared for the model is 0.912 with the deviance explained being 91.6 percent. The code provided in the appendix gives the full information for the table.

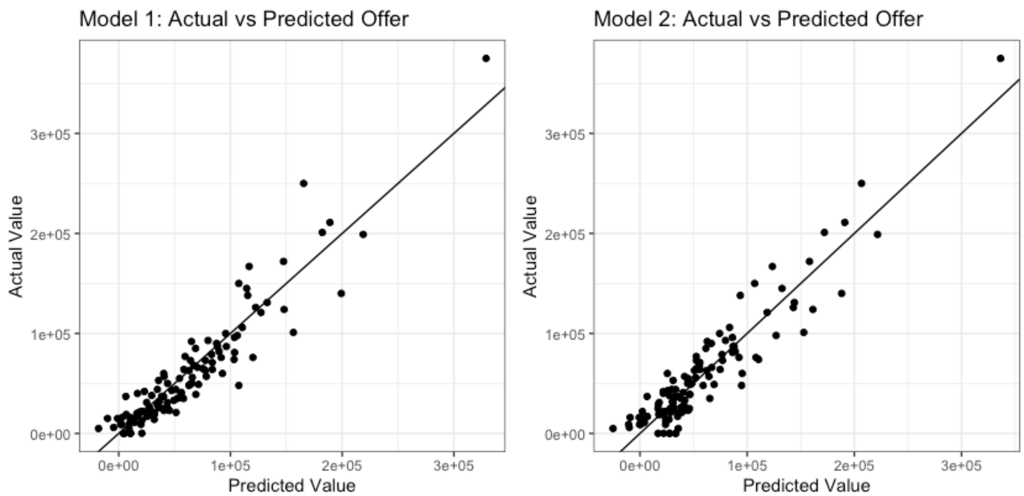
To test assumptions of the model I first plotted the fitted values against the residual values, which can be seen in the plot below on the left. The plot shows that the variance of the residuals is not independently distributed across the fitted values. The variance appears to increase as the fitted values increase. This was also evident in the increasing size of confidence interval in the smoothing spline graph. In addition, residuals on the far left of the graph all appear to be positive and decreasing. This is likely evidence that the smoothing spline for low expected values tends to under-predict. It is also worth noting that the majority of the points in the data are bunched up in the middle and it is the extreme points that tend to violate regression assumption. Also, this is shown in the normal Q-Q plot shown below on the right in which the points in the middle follow the Q-Q line closely while points on the left and right ends deviate.





After training my model, I predicted the bankers offer on the testing data using both my model and Samuel Bradley's model. The plots of the predicted offer versus the actual offer for both my model (Model 1) and Samuel Bradley's model (Model 2) are shown below. The mean squared predicted error Model 1 is 404,489,740 and for Model 2 is 403,511,794. Therefore

Model 2 performs slightly better in terms of the mean squared prediction error, although they are very close. Looking at the plots of actual versus predicted offer the two models also appear fairly similar. Model 1 appears to perform better for lower bank offers and worse on higher bank offers compared to Model 2. Model 1's poor performance is for higher values is likely due to the increasing variance shown in the residual plot. Therefore, depending on the size of the predicted value it may be preferable to use Model 1 predictions in the lower range and Model 2 for predictions in the upper range.



IV. Conclusions

Based on the results from testing my model I believe that my model does a sufficient job at predicting with a comparable performance to Samuel Bradley's model using a simpler functional form. However, there is some room for improvement as the model still does not perform well at extreme values, particularly for large offers. Perhaps an alternate smoothing method could improve the fit of the model. It would also be valuable to train and test the models using data from more than one season and perhaps to include additional variables to test whether or not they are related to the banker's offer. Increasing the scope of the analysis could yield more precise estimations, especially for evaluating predictions and adding more relevant variables may increase the explanatory power of models.

Despite some shortcomings of this analysis, using only the expected value, I was able to predict the banker's offer with a reasonable level of accuracy. It is also unlikely that the offer will ever be able to be predicted with 100 percent accuracy as I believe the banker does not use an exact formula to give his offer but rather takes the expected value and other relevant gameplay factors into account to decide what the offer will be. This would also prevent contestants from being able to cheat by discovering an exact formula before going on the show. From this analysis, I learned that the banker tends to make offers close to the expected value for small and large expected value but undervalues medium expected values. In addition, the later the round the more likely the banker is to give a fair offer. Further research into this topic could benefit from testing specific hypotheses from behavioral economics to see if the banker's offer can be explained by any behavioral or economic theory.