STA 309

Homework 6

Problem 1 - Cheese

Question

Build a model for Q(sales volume) in terms of price (P), store level dummy variables, and a dummy variable for whether or not there was a display of cheese.

Answer the following questions:

- What is the price elasticity of demand for Borden Sliced Cheese in no display weeks? Interpret this number in a single sentence.
- Does price elasticity for borden cheese appear to be changed by the presence of in-store-display? (Remember the interaction terms). Can there be a possible economic explanation for the result?
- What price should Kroger's charge for cheese in no-display weeks to optimize gross profits?
- Adjusting for store-level differences and differences in price from week to week, how much higher or lower do sales seem to be in display weeks vs. non-display weeks, on average across all stores?

Approach

To solve this problem, I will fit a linear regression model, over a Monte Carlo simulation of 1,000 simulations, taking into account the interaction term of price and display. To determine the optimal price that should be used for the cheese, I used the optimize function in R, in addition to calculating the profit, over a Monte Carlo simulation of 1,000 simulations.

Results

Figure 1

Name	Lower	Upper	Estimate
Intercept	8.7952398	9.125855985	8.9919814
Log(Price)	-1.1806401	-0.876845135	-1.0663825
Display	0.3186949	0.723196234	0.5070631
Price:Display	-0.1281303	0.009072035	-0.0581441

The table above shows all of the predictive variables for the linear regression model, along with their corresponding lower and upper confidence intervals (at 95% confidence) with the respective estimates. With the dummy variable Display, which denotes whether or not the cheese was on display.

Figure 2

Name	Upper	Lower	Estimate
result	6.422082	9.999957	9.999943

The table above shows the result of what should be charged for non-display weeks after optimizing the price, with the respective confidence intervals.

Conclusion

- The price elasticity is as follows: when price of cheese goes up by 1%, the demanded volume of cheese goes down by between 1.18% and 0.88%, when not looking at the interaction. On days when there is a display, the volume of cheese demanded goes down by 1.05% and 0.87%, however, because the interaction term goes between 0, it is statistically insignificant. The range denoted (1.05% and 0.87%) does not take into account the display dummy variable, which will also affect the volume, on days when there is a display, by between 0.31% and 0.72%.

- There does seem to be change of the price elasticity for Borden in the presence of a display, it seems to go down by between .31% and .72%, without taking into account the interaction, and between .19% to .73% when taking into account the interaction between price and the dummy variable display. Possible economic reasons for this could be that, typically, when something is on a display they are running a promotion and could potentially give a higher opportunity cost for choosing a substitute good over Borden's milk, or maybe simply the display proved to be a good marketing tool and put the cheese in people's mind when shopping.

- The price that Kroger's should charge for cheese in no-display weeks should be between $6.42 and $9.99, estimated to be $9.99, as denoted in Figure 2 of the results.
- On display weeks, sales seem to be higher by between .19% to .73%, when taking into account the dummy variable and the interaction between price and the display dummy variable, estimated to be about 0.45%.

Problem 2

Question

Compare the out-of-sample performance (using RMSE) for four models, a small model, a big model, a huge model, and a feature-engineered model, used to predict whether or not a booking has children on it.

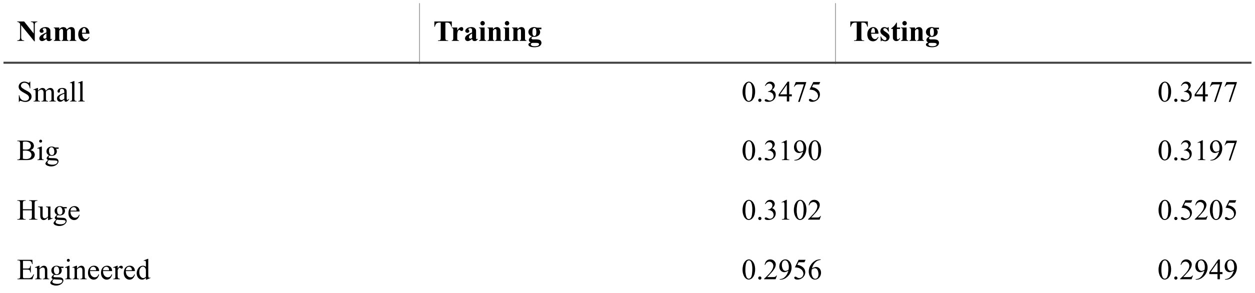
Approach

To compare all of these models, I created four linear models, with each of their respective variable inputs.

- A small model that only uses the market_segment, adults, customer_type, and is_repeated_guest variables
- A big model that uses all of the possible predictors except arrival_date, main effects only
- A huge model that uses all of the predictors, except arrival_date, and all possible pairwise interactions
- A model, the same as the big, with a feature engineered month of the year predictor.

Results

Figure 1



Name	Training	Testing
Small	0.3475	0.3477
Big	0.3190	0.3197
Huge	0.3102	0.5205
Engineered	0.2956	0.2949

This table shows the in-sample RMSE and the out-of-sample RMSE for all of the models detailed in the approach section

Conclusion

The model to use to predict children in bookings should be the one that is feature engineered with the month as it had the lowest in and out of sample RMSE.

Problem 3

Question

Replicate Ng's approach using the 2018-19 data to answer the following questions:

1. What is your estimated probability distribution of win/lose/draw for a match between Liverpool v. Tottenham?

2. What about Manchester v. Arsenal?

Approach

To answer this question I created a function in R that would use the Random Poisson distribution function that ran over a Monte Carlo simulation of 100,000 simulations, reporting the amount of ties and wins for each team, respectively, while taking into account the attack strength, the weakness strength, and calculating the average goal-per-game rate for each respective team using these measurements. Mathematically, I used the same approach as David and Yin in that I calculated the average number of goals scored, then calculated the average goals per fame for away games and for home games. Then I used the Poisson distribution (as rpois) and calculated the percentage of ties, home wins, and away wins, respectively. I calculated the probability of a 2-1 victor for the home team in each of the two games using the following formula:

$$prob = sum(t1 > t2)/100{,}000$$

Where the probability is the amount of times that the Poisson distribution for the home team was higher than the Poisson distribution for the away team, then divided it by the amount of Monte-Carlo simulations that were ran for the model.

Results

Figure 1

Liverpool	Tied	Tottenham
0.85947	0.06313	0.0774

The table above shows the probability distributions for Liverpool v. Tottenham, respectively, and the probability of a tie, in decimals.

Figure 2

Manchester City	Tie	Arsenal
0.94684	0.02446	0.0287

The table above shows the probability distributions for Manchester City v. Arsenal, respectively, and the probability of a tie, in decimals.

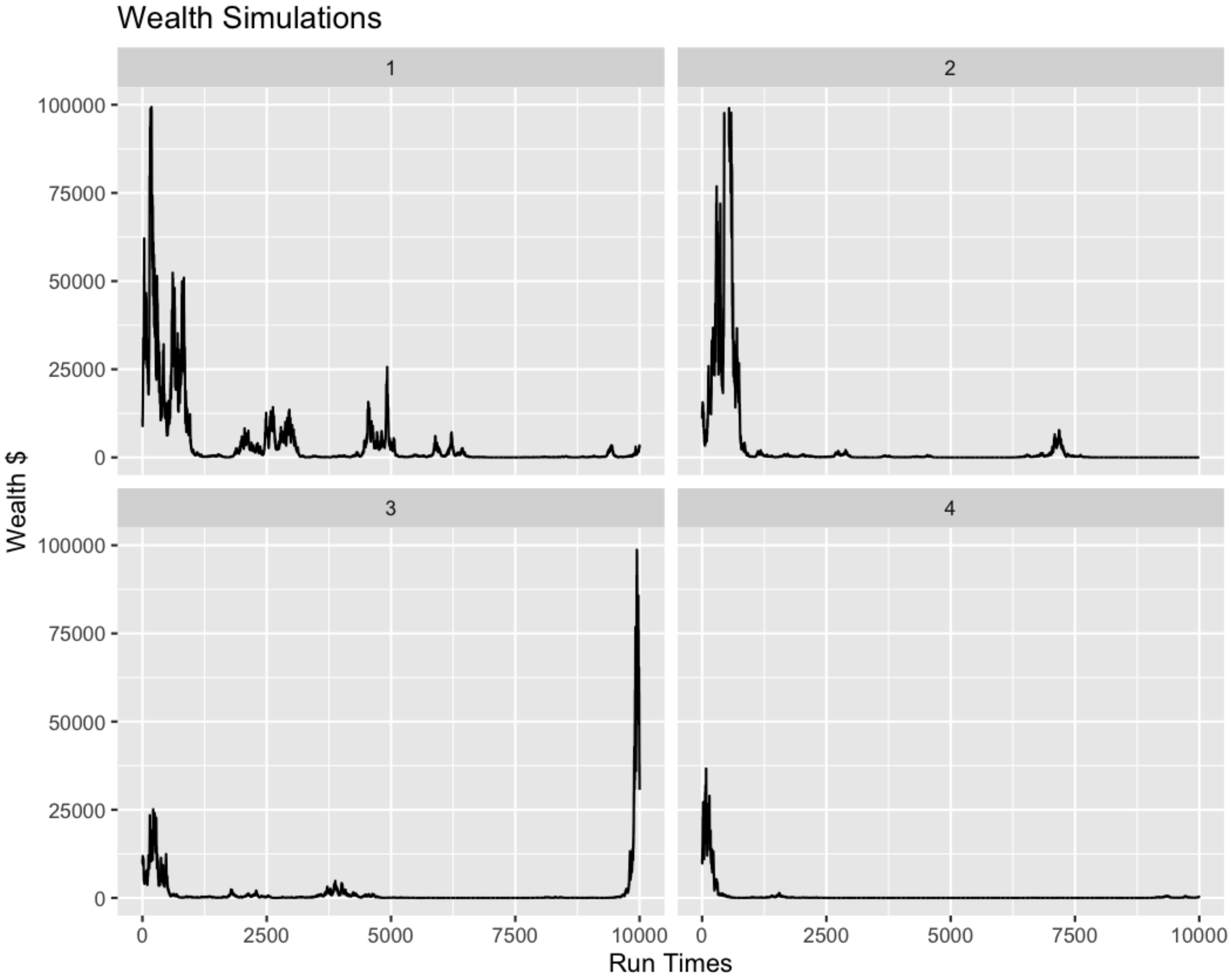
Conclusion

The probability distribution for Liverpool v. Tottenham is 86% Liverpool, 7% Tottenham, and a 6% probability for a tie, making it so that Liverpool is the likely winner for the match.

The probability distribution for Manchester City v. Arsenal is 95% Manchester City, 3% Arsenal, and 2% a tie, making it so that Manchester City is the likely winner for the match.

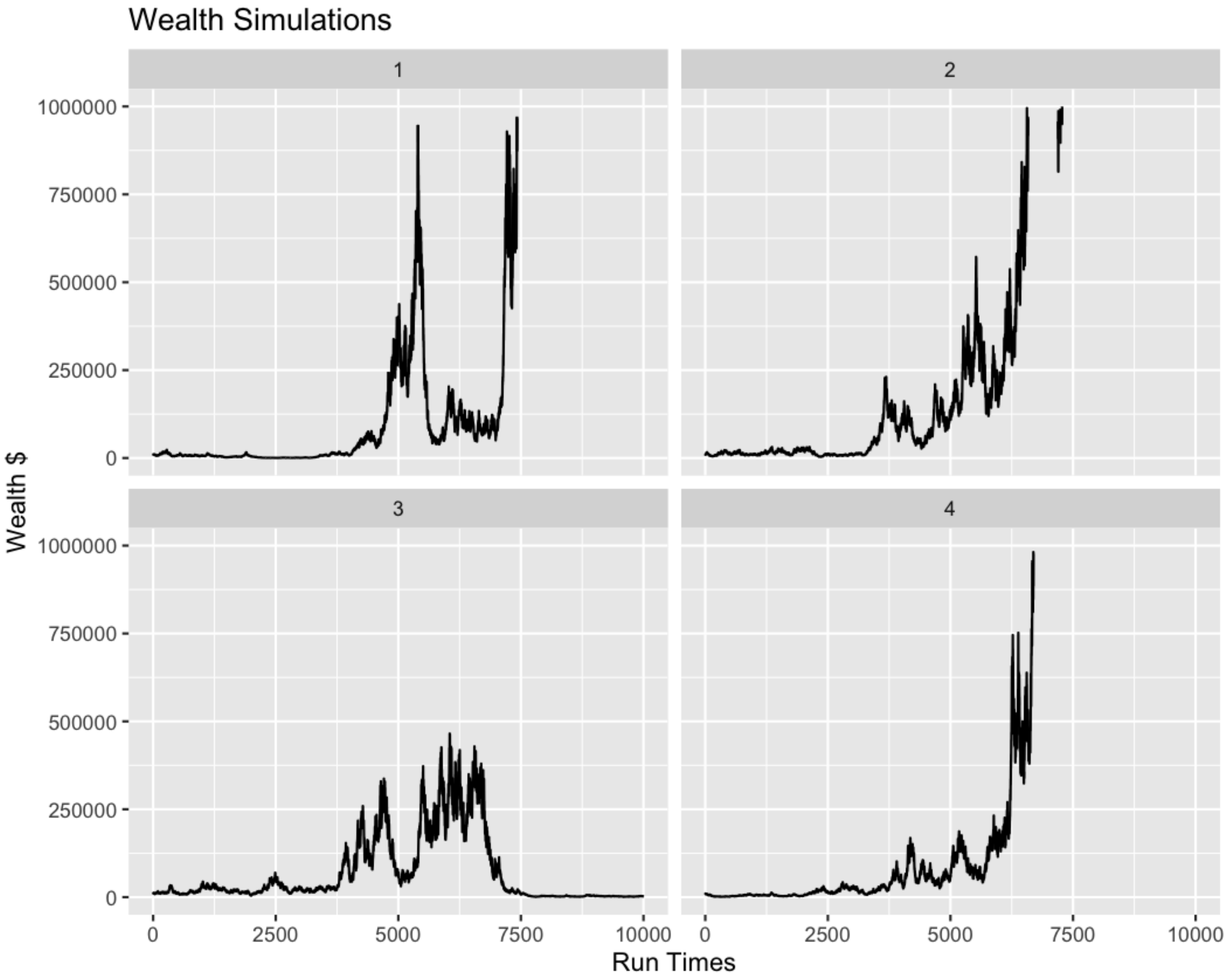
Problem 4

Part A



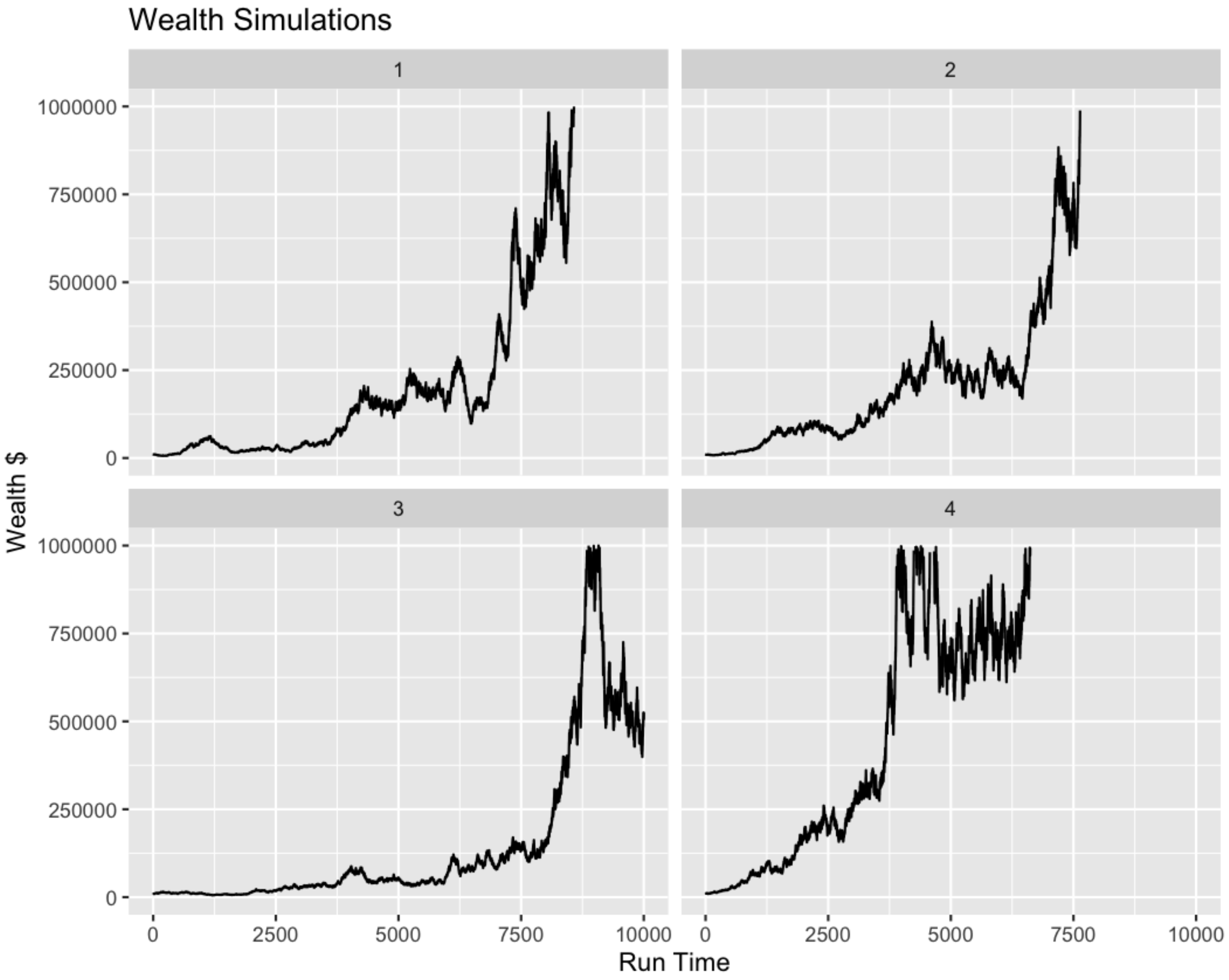
The graph above shows four plots, all of which took the wager the the same probabilities 10,000 times. The y-axis shows wealth in dollars (capped at $100,000) and the x-axis shows the run-times of the bet. It appears that all of the bets seem to go broke around the midway point to 2,500 simulations, except for plot 3. The third plot uses its very low wealth (some cents) and has a very lucky streak, which makes it attain some income back. The other three graphs all seem to go broke with no lucky streak to bring them back.

Part B



The chart above shows the wager taking place over 4 times, all of them going over 10,000 simulations with the same probability, however, they only leverage 5% of their overall wealth rather than the previous 10%. The x is the run time, the y is the wealth, and it is faceted by the time it was ran. The y axis is capped at $1 Million. It is clear that three of these wagers, at leveraging 5% of wealth per bet, came with a very lucky ending In that they were above $1 Million. Plot 3 did go broke but in terms of percentage from before, this model performed much better than the previous. There do seem to be trends of unlucky and lucky streaks, but it seems to even itself out to profit.

Part C



This graph shows my estimated c-value (wealth to be leveraged), which was set at 2% of the wealth, per bet, for 10,000 simulations. The x axis is the amount of simulations and the y axis is the wealth in dollars (capped at $1 Million). We see that all of these do not necessarily go broke, rather 3 out of the four graphs end with values that are above $1 Million and the one below $1 Million is still not broke. I would say that by using less leverage, the investments into the wager were safer, thus providing safer increases to the wealth of the person. I judged that this was the best c-value as I saw the increase in overall wealth from part A to part B, and went with 2% rather than 1% because, in my simulations, 1% did not provide the same metrics.