Name: Alden Butzke

**BA 355: Business Analytics Case 1, Part 1**

Goal: To estimate the probability that the Denver Broncos beat the Seattle Seahawks on 9/12/22. More generally, estimate this probability for any game given the point spread.

The Broncos are currently a six point favorite (as of 8/30/22 at 8am) for the first game of the season, but the end result is by no means guaranteed. How would you estimate their chances, understanding that either team can win?

1. Combine that data for years 2013 – 2016 from the Excel file on the course webpage. You should have a total of 1068 rows. (A normal NFL season has 267 games including playoffs and the Super Bowl. 267\*4 = 1068)
2. Sort the data by point spread from largest (-26.5) to smallest (0).
3. Label column A as “Point Spread”, column B as “Favorite’s Score”, column B as “Underdog’s Score” and column D as “Over/Under” – we won’t need column D for now but might want it later.
4. Determine whether the favorite actually won the game (*not beat the spread*) for each point spread that occurred. In column E, subtract column C from column B. If this number is positive, it means the favored team won the game. If this number is negative, it means the underdog pulled off the upset. If this number is zero (and there are a couple of them), it means the game ended in a tie. Label this column as “Difference”.
5. In column, convert the result in column E to binary: a 1 if the favorite won the game and a 0 if the underdog won (or tied – for now we’re calling a tie a “win” for the underdog -- <https://en.wikipedia.org/wiki/1968_Yale_vs._Harvard_football_game>). Hint: =if. Label this column as “Win or Loss”.
6. Now, combine all the results for each possible point spread. For example, you should have 20 results from games with a 10-point spread; of these, the favorite won 16, but lost 4. What does this say about the *probability* that a 10-point favorite will win a game? Does 16/20 = 80% seem like a reasonable estimate? Combine the results so that for each possible point spread you know how many games there were (e.g. 20), how often the favorite won (e.g. 16) and the estimate for the probability (e.g. 80%). The best way to do this is by using the **PivotTable** function in MS Excel. If you don’t know how to do this, ask me or a classmate for help. If you do know how to do this, please share your expertise with your classmates.
7. In the end, you want a table of data with point spreads ranging from 0 to 26.5 in one column (but not every point is available, for example 26.5 is the only data point greater than 20) and the estimated probability of winning in the next column.

Now, use this data to make some estimates:

1. **i) What proportion of 3 point favorites actually won?**

55.73%

**ii) Is this a good estimate for the probability that a future 3 point favorite will win?**

There were 131 total games played in the 3 game category, this should be sufficiently many observations to conclude that 55.73% is a good estimate for winning potential.

iii) What is the 95% confidence interval for the true proportion? [Extra Credit]

The 95% CI for 3 point spread is 47.22% through 64.23%.

iv) Interpret this interval. [Extra Credit]

95% of the time, the true/population win percentage will be within these bounds.

1. **Repeat A. for 7 point favorites.**

72.22%, There were 54 games played in the 7 point spread category. This is significantly

less than the 3 point spread but should still be enough to appropriately represent the population.

The CI is 60.28% through 84.17% with the pop mean falling within these bounds 95% of the time.

1. **Repeat A. for 14 point favorites.**

The data say there is a 100% chance of a 14 point team winning. This should be an immediate red flag because 100% probabilities don’t exist, and accordingly, we can see that there were only 4 games played in this category. That is too low to conclude we can accurately predict the population proportion with this data alone. The CI is also meaningless, being 100%-100%. However, we can contextualize this trend and cautiously conclude that a 14 point team has a much better chance of winning than a 3 point and 7 point spread, however we don’t have enough data to accurately measure the wins without more context.

1. **Does the trend make sense? Do higher favorites have a better chance of winning?**

This trend does make sense because higher favorites will have much high odds of winning. However, the trend line is not currently fitted well because it predicts that once a team has a 13 point spread, it has the same odds of winning as a team with a 26.5 point spread. That’s an absurd result, especially considering that the point spreads don’t increase linearly.

1. **The Broncos are currently a six point favorite. According to your results, what percent of the time do teams favored by six win the game?**

64.29%

1. **Now, graph the point spreads (ranging from 0 to the max) versus the win probabilities and then fit a linear regression line to the data. Determine and interpret the equation of the line and the coefficient of determination (r2). Print out the graph with the line on it. How much does the probability of winning the game increase for every additional unit of point spread? At what point does the prediction from the line stop making sense?**

Y=0.026x+0.5402,

B0 = 0.5402 and B1 = 0.026

For each additional unit of point spread, a team’s probability of winning increases by 2.6%. This line doesn’t make any sense because we can observe non linear trends in the data, our intercept should be .5, and there should be an asymptote at 1. The R-squared value is 0.766, indicating that point spread comprises 76.6% of the variability in win proportion.

1. **Now fit a line from 0 to 12.5 points (up to where all probabilities become 100%) forcing the y-intercept to be 50%. (Why?). What is the equation of this line? Interpret both the y-intercept and the slope.**

Y=0.03x + 0.5

B1 = 0.03 and B0= 0.5 (because the known probability of a 0 spread game is .5 – though this is not the true probability, it is merely randomness because the information to determine the true probability is unknown to us at present. Ie when I see a pregnant woman, the probability of a boy or a girl to me is 0.5, to her and her doctor, it is likely not .5)

When point spread goes up by 1, win proportion is expected to go up by 3%.

1. **Use the equation of the line from G above to estimate the probability that 3 point, 7 point and 14 point favorites will win a game. Compare these to the empirical estimates from the data you found in parts A, B and C.**

The regression from G predicts, 61.82%, 72.22%, and 90.42%. The data itself shows 55.73%, 72.22%, and 100%. The regression has artificially high predictions for 3 and 7, but actually provides a more reasonable estimate of 14 than the data does. This is likely because the higher number of observations in the lower point spreads have forced some variability into the prediction for 14, where the data itself does not have enough variability in the 14 point category. This is the same reason 3 and 7 are estimated more poorly, the lack of variability in the higher point spreads has falsely cancelled the variability in the lower ones.

The new equation (Y=0.03x + 0.5) predicts 59%, 71%, and 92% respectively.

1. **What is the probability for the favored Broncos? How does this compare to part E?**

69.62% from the linear and 68.7% logistic regression compared to 64.29% from the data. The regression is a worse prediction because of the artificially cancelled variability leading to a higher prediction even despite the weighted predictions.

1. **The graph is clearly non-linear. Use Excel to try to fit a better curve to the data. Soon we will develop our own function that should fit very well in class.**

Fitting a polynomial works in the short term, but that is a quirk of the data because it quickly stops fitting after looking 30 periods forward.

1. **Repeat what you did in part F above, but now there are a lot more data points. For example, the point spread of 10 isn’t just one point with probability 80% it’s actually n = 20 points at point spread 10 with probability 80%. Use all 1068 data points. Calculate the coefficients of the new line to three decimal places – it will only be slightly different from what you got in part F.**

Chart, scatter chart

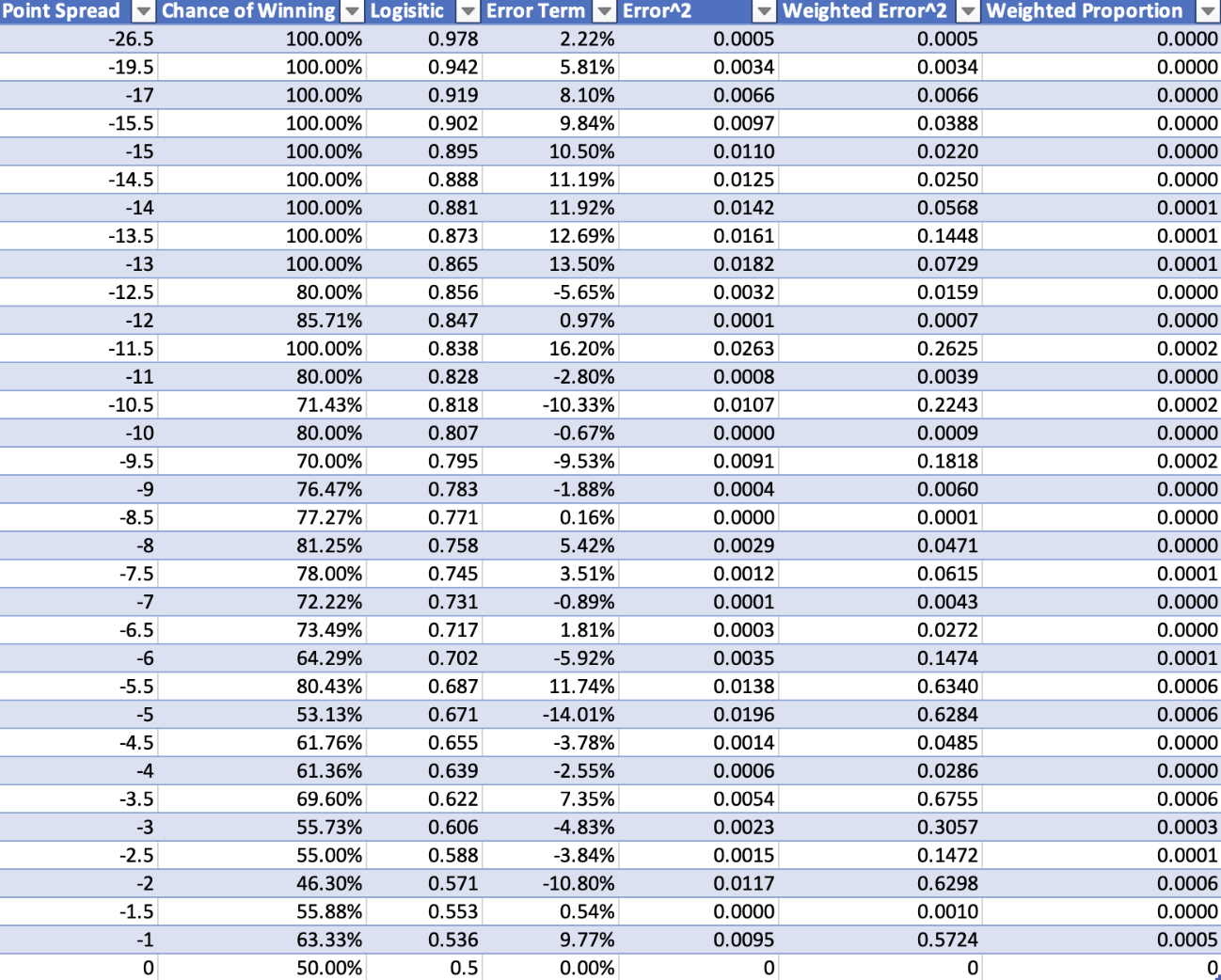
Description automatically generated

A:0

B. 0.143

10 pt = 79.5%

**Deliverable:** Type up all of your results into a clear document that addresses each part above. The due date is TBD and will depend on the class’s progress and overall level of effort.



[Eventual Extra Credit: This data comes from the degenerate gambling website [www.goldsheet.com](http://www.goldsheet.com). This site stops recording data after the 2017 season. Find another reliable (for the internet) source that has data for the 2018, 2019 and 2020 seasons.]