NFL Score Prediction Using Previous Game Results

**Michael Kerins – Project Report – 18/06/2021**

This report summarises my contribution to the project (describe via objectives below), undertaken with two of my peers: Aaron Tracey and Rhys Willis.

**Project Objectives**

* To accurately predict score difference for an NFL game, given the score difference of all previous games that same season.
* To use sound methodology to define, develop, diagnose, and evaluate a set of regression models to achieve objective 1.

**Motivation**

My primary motivation in undertaking this project was to apply statistical methodology to a real-world regression problem. Having just finished my second year as a mathematics undergrad (majoring in statistics), I felt that I had a solid foundation in statistical methods and was ready to take on my first major project. In hindsight I may have overestimated the strength of that foundation and had not anticipated many of the challenges the project threw up. A secondary motivator to do the project was to gain experience working as part of a team. That team consisted of two peers, Aaron Tracey (who had the original idea for the subject matter) and Rhys Willis.

**Summary of Method**

* December 2021: We began the project over the Christmas holidays by brainstorming precise objectives and identifying potential pitfalls. A rough timeline for the project was constructed: we aimed to complete the project by the end of June 2021. Finally, we researched how other NFL models worked to give us an idea of where to begin working. We decided to code the project through Python, using the Jupyter Notebook files. This allowed clear, accessible, and modulated code blocks.
* January - March 2021: The first model was coded collaboratively, in the evenings and over the weekends in the first half of the Spring trimester (Filename: ‘Model Iteration 1’). Rough performance analysis was undertaken. The methodology here was poor, but it brought to light the knowledge gaps needed to be filled before a second iteration could be coded. This model also made use of team statistics such as ‘Sacks’ and ‘First Downs’ in addition to previous score that season. We found that these statistics added little to model performance and decided to focus on previous scores. The model was object oriented which was useful for reuse purposes but was somewhat cumbersome. Statistics extracted from the model included success rate (correctly predicted the game winner), and the spread of residuals around the actual score. The former statistic was removed from the second model iteration as it was decided that the spread of residuals around the actual game scores would be used to measure the success of the model.
* March 2021: A report of the first model iteration was written. It was identified that we knew little about coding regression models in python. We decided to take a machine learning 'skills track' on Code Academy. I completed this course over the Spring trimester midterm.
* May 2021: The second major knowledge gap identified was that of regularized multiple linear regression. To give me a robust understanding of this topic I read 'Linear Regression Analysis: Theory and Computing' by Xin Yan and Xiaogang Su, after finishing my exams in May. This textbook took approximately three weeks to read and gave me a chance to the revise the linear algebra involved in multiple linear regression.
* May - June 2021: The second iteration of the model was coded. It was decided that we would each focus on different areas and produce separate models. This enabled us to avoid conflicts in model design and allowed each of us to get holistic experience coding a model. I decided to opt for a selection of simple multiple linear models with the aim of sound statistical implementation. On this occasion proper definition, diagnosis and analysis was undertaken (Filename: 'Prediction of NFL Games Using Only Game Results'). This file contains a full comparison of 7 model variations. All of these were restricted to the same predictor variables, and only multiple linear regression models were used. The file contains full descriptions of each model. Throughout this period, I continued to study statistical literature to overcome the challenges faced.
* June 2021: A more involved historical analysis of Models 1b and 4 performance was conducted (Filename: ‘Historical Analysis of Models 4 and 1b Performance’). Year by year and game by game analysis of the models was conducted.

**Difficulties Encountered**

1) The first difficulty that presented itself in this project was accessing reliable data. I had no experience taking data from the web and converting it into csv files which could be imported to python. As well as this, it was harder than expected to find comprehensive and uncorrupted historical data

2) Using Python to create regression models. This was a major knowledge gap. I had some experience creating regression models on R, but it was decided that this was not a flexible enough platform for the desired analysis. Therefore, we needed to familiarize ourselves with the available Python packages that would help.

3) High Dimensionality of the model, coupled with a low number of independent observations made accurate modelling challenging. This difficulty was expected and was in fact one of the reasons NFL was chosen to challenge ourselves. It was an interesting and insightful challenge to tackle and forced us to familiarize ourselves with Ridge Regression as a means of preventing over-fitting

4) Crossing the bridge from theory to practice presented difficulties throughout the project. The unforgiving nature of computer code plainly exposed inadequacies in our theory. A mistake was made (and subsequently fixed) while standardizing the residuals. The bias of ridge regression was overlooked, and we forgot to centre the residuals, as they would be if Ordinary Least Squares was used instead (This is an assumption of the linear model)

5) Examining seven models meant long and monotonous coding, as each piece of analysis was altered and repeated for each model.

6) Historical analysis of the models presented a computational challenge. Efficiency of the code had to be improved to allow the algorithm to run in a reasonable time (6.5 hours). As well as this, the results from this analysis had to be saved to avoid rerunning each time the Notebook was opened. This required learning to use the ‘json’ module.

**Evaluation & Takeaways**

I believe that the project has been successful in achieving the objectives outlined above. The final model choice does a good job predicting score difference in games, given the low number of training data points. In addition, the model is stable with a low number of outliers. The models were clearly defined, diagnosed, and evaluated using well known statistical methods. There were certainly areas where deeper investigation into model properties could have been conducted, however it was not the aim of this model to conduct an exhaustive evaluation. Instead, uncomplicated methods were employed to compare several models which provided much insight into the mechanics of linear regression. The statistical methods I have learned in my first two years studying maths have now been firmly grounded in context.

In addition to the listed objectives this project has expedited my theoretical and practical development as a statistician. The range of unforeseen difficulties compelled me to fill knowledge gaps and problem solve in unexpected ways. Specifically, the project has hugely improved my:

* Python programming skills
* Understanding of ridge regression, and its practical deployment
* Communication skills

Aside from these easily quantifiable takeaways, the project has given me a firm appreciation for the work of statisticians and has been a tremendously enjoyable process.

**Areas of Further Study & Project Extensions**

Although cliché, it is certainly true that for every question this project has answered for me, it has thrown up two more. There are many areas that I would like to (and intend to) study further including:

* Including higher order terms in the model. Interaction and quadratic terms may provide further insight into the relationship between variables.
* Dimensionality. Central to this regression problem was the problem of dimensionality. It is an intriguing issue which I would like to investigate further. Specifically, the trade-off between the added information of extra terms in the model and the increased dimensions is a potentially very rewarding issue to solve.
* Bayesian interpretation of them model. This seems to be a much more flexible framework. Investigating priors for the model could also prove very fruitful as it would allow for the inclusion of results from previous years.