Count Data

A tale of Poisson and predicting football results

Hello

My name is Vernon Gayle.

I am Professor of Sociology and Social Statistics at the University of Edinburgh and I am part of the ESRC National Centre for Research Methods

At the current time, due to the restrictions placed on us by the COVID-19 pandemic the National Centre for Research Methods are unable to deliver any standard face-to-face research methods training courses.

I hope that sometime in the near future you will be able to join us in person at the University of Edinburgh.

The film that follows is about the analysis of count data.

Consider the following -

How many times did you go to the cinema last year?

How many people has your best friend slept with?

How many goals have your favourite football team scored this season?

The answer to any of these questions is likely to be a count, which means it is a positive whole number (i.e. an integer). The number must be positive because you can't have -2 visits to the cinema, and it must be a whole number because you can't make a ¾ (or .75)

visit to the cinema. Similarly, sexual partners and goals are also counted in positive whole numbers rather than fractions or decimals.

Examples of social science data that take the form of positive counts are legion. For example, how many burglaries take place in a neighbourhood, how many women under twenty gave birth last year, or how many cases of a disease were diagnosed?

Indeed, the question how many ‘any things’ will usually be answered with a count.

Given the prevalence of count data in the social sciences, for many years it has puzzled me why most social scientists know very little about analysing count data.

In reality social science data analysts tend to know more about analysing either binary (i.e. 0,1) outcomes or continuous (i.e. metric) measures.

The Poisson distribution is integral to analysing count data. The Poisson distribution is named after French mathematician Siméon Denis Poisson, who was a Fellow of the Royal Society of Edinburgh and his name is one of the 72 names inscribed on the Eiffel Tower.

Lord Tennyson laments that in spring a young man’s thoughts turn to love.

I am a middle-aged football fan, and by contrast, my thoughts often turn to the final game of the season.

My interests lie in Scottish League Two, the fourth tier of Scottish men’s professional football.

I am a Stirling Albion fan and here I am pictured with our mascot Bino the Bear.

I am going to use an example that I developed before the final day of the football season in Scottish League Two, back in 2018.

I am going to follow an analytical approach that was used to analyse data in the English Premier League by Professor Sir David Speigelhalter.

Here I am at the Royal Statistical Society, teaching a course, and Professor Spiegelhalter, who was the President of the society, came in to say hello.

In my view David is the greatest living British statistician.

You may well have seen him on TV during the current COVID-19 pandemic, and he is a recognisable voice on Radio 4.

David is a really nice fella and he is also great fun.

For example, he is the reigning (first and only), world champion in Loop. This is a version of pool invented by Alex Bellos and played on an elliptical table with a single pocket in the green baize.

League football matches either end in the home team winning (a home win), the away team winning (an away win) or a draw (when both teams have scored the same number of goals).

Many games end without either team scoring, and typically games end with each team scoring only a few goals.

There are occasional ‘shockeroonies’ for example when a team will suffer a six nil defeat. There are also occasional ‘goalfests’ where both teams ‘stick it in the onion bag’ half a dozen times.

In 1885 Arbroath thrashed Bon Accord 36 - 0, and in 1984 Stirling Albion beat Selkirk 20 - nil.

Routinely most games end with a modest number of goals despite the large number of opportunities to score.

Here is the example.

As Cher would say – Let me turn back time.

As the last day of the 2017/18 football season approached, the winner of Scottish League Two was still undecided.

My own wee football club (Stirling Albion) was due be battling for a place in the playoff competition.

When consuming a traditional half-time pie, I have often ruminated on the veracity of a statistical approach to predicting match outcomes and final scores.

Here is a list of the five games that made up the last day of the season.

In this example I am going to bring some statistical thinking to the prediction of the outcomes of these matches and to predicting the final scores.

To make things interesting I consulted a fellow fan, a guy that has followed the club since his teens.

Every fan thinks that they are an expert but I prefer to consider this as ‘pseudo-‘ expert knowledge.

Here are his predictions.

I also constructed a set of random predictions decided by a seven sided dice.

Let us now take a look at the data that have been generated by the games played in the league during season.

Montrose are at the top of the table with an impressive 76 points and Cowdenbeath are at the foot of the table with only 22 points.

The final game of the season for Stirling Albion will be played at home. As a fan I would like to think that our ground, Forthbank, is a modern day manifestation of a Roman Colosseum where football foes are routinely vanquished, and pies and Bovril refresh the senators and equites.

The reality is somewhat different. So far this season we have played 17 home games and won only 8.

Stirling Albion will be playing local rivals Stenhousemuir on the final day of the season.

Stirling have lost four of their last five games, compared with Stenhousemuir who have only lost once in their last five matches.

I would like to predict the likely outcome of this forthcoming match and the other games on the final day of the season.

Stirling Albion and playing Stenhousemuir in their final game.

[click]

Stirling have played 35 games.

[click]

Stirling have won 16, drawn 6 and lost 13 games.

Our opponents

[click]

Stenhousemuir, have also played 35 games.

[click]

Stenhousemuir have won 15, drawn 8 and lost 12 matches.

The first measure that we are going to construct is called

Attack Strength.

It is a measure of how good the team is at scoring goals.

At the current time Montrose who are top of the league have scored 59 goals and Cowdenbeath who are bottom of the league have only scored 23.

The average number of goals scored by each team in the league is 49 (i.e. the 10 teams have scored 490 goals in total).

Stirling Albion have scored 60 goals and Stenhousemuir have scored 55.

If we take a ratio of the team's goals scored over the league average then we have a measure of their *attack strength*, or the quality of their attack,

Stirling Albion 60/49 = 1.22

Stenhousemuir 55/49 = 1.12 .

We can infer that Stirling Albion score about 22% more goals than the league average, and Stenhousemuir score about 12% more than the league average.

The second measure that we are going to construct is called

Defensive Weakness –
how bad is the team is at defending (measured by conceding goals)

The average number of goals conceded by each team in the league is 49 (i.e. the 10 teams have scored 490 goals in total).

There is a beautiful symmetry here, simply because when one team score a goal the other team concede a goal.

If we take a ratio of the number of goals that the team conceded (goals against) over the league average for goals conceded then we have a measure of their *defensive weakness*, a measure of the quality of their defence,

Stirling Albion 51/49 = 1.04

Stenhousemuir 46/49 = 0.94 .

We can infer that Stirling Albion let in about 4% more goals than the league average, and Stenhousemuir let in about 6% fewer goals than the league average.

There are two further measures that are required.

Home Average *- The average number of goals home teams score.*

Away Average *- The average number of goals away teams score.*

The teams scored 255 goals in 175 games.

Therefore, the average number of goals that home teams score in the league is 1.46 (255/175).

Teams playing away from home scored 235 goals in 175 games.

The average number of goals that away teams scored in the league is 1.34 (235/175).

How many goals can we reasonably expect when Stirling Albion play Stenhousemuir?

Putting the information together we can now work out the expected goals for each team.

This is the information required for calculating the number of expected goals.

Stirling are playing at home.

The average number of goals scored by a home team is 1.46.

But Stirling are not an average team, they usually score about .22 more (their attack strength is 1.22).

They are also playing Stenhousemuir who have an effective defence and have only conceded 46 goals when the league average is 49.

Stenhousemuir have a defensive weakness of .94. Therefore, given Stirling's better than average scoring ability and Stenhousemuir's slightly better than average defence, I estimate that Stirling can expect to score 1.67 goals (1.46 x 1.22 x 0.94).

Stenhousemuir are playing away.

The average number of goals scored by an away team is only 1.34.

But Stenhousemuir are not an average team, they usually score about .12 more (their attack strength is 1.12).

They are also playing Stirling who have a slightly suspect defence and have conceded 51 goals when the league average is only 49. Stirling have a defensive weakness of 1.04.

Therefore, given Stenhousemuir's better than average scoring ability and Stirling's slightly weaker than average defence I estimate that Stenhousemuir can expect to score 1.56 goals (1.34 x 1.12 x 1.04).

Now that we have an expected number of goals for the two teams it is possible to plug this information into the Poisson formula.

In football once the referee blows the whistle and play commences, in the 90 minutes that follow, there are many chances to score a goal but few of these chances end in a success. In statistical terms we might consider this as a large number of trials, each with a low chance of success.

The poisson distribution expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and are independent of the time since the last event.

Here is the formula

λ lambda is the expected number of goals

e = 2.71828 (this Euler's number which is a mathematical constant)

k is the number of events (in this example 0 through to 6 goals)

k! is k factorial

when k = 6 it is 6 x 5 x 4 x 3 x 2 x1

[Speak it out]

Probability equals = $e^{-\lambda}$ (λ^k / k!)

Plugging the information for Stirling Albion into this formula…

For one goal the probability is 0.31

For two goals the probability is 0.26

The predicted probability of Stirling scoring zero goals is 0.19 or 19%;

The predicted probability of Stirling scoring one goal is 0.31 or 31% .

The predicted probability of Stirling scoring two goals is 26%.

And so on.

Here are the predicted probabilities for Stenhousemuir.

There is a 31% chance of Stirling Albion scoring 1 goal.

Stenhousemiur are also most likely to score only one goal (a 33% chance).

If we multiple 0.31 and 0.33 we can estimate the overall probability that

Stirling v Stenhousemuir will end 1 -1

0.31 x 0.33 = 0.103 .

This suggests that there is a 10% chance of a 1 - 1 result (i.e. a draw).

In statistical terminology I have assumed that each event (i.e. goal) is independent.

As a fan I would be delighted for the match to end with Stirling winning 6 - 0, but I estimate that there is only a 0.001 chance of this result (0.01 x 0.21).

Here are the predictions for the result of the other four matches that will be played on the final day

[click through slides] Reading scores.

Over the years I've noticed some striking empirical regularities; birds fly, fish swim and colleagues sometime accuse me of using an outdated software package or programming language.

In this example I have used Python simply as a defence against the familiar accusation.

Here is the Python code running in a Jupyter notebook.

Other software and statistical languages are available.

The models outlined above, are simple Poisson models that are the product of only a few terms (e.g. home advantage x attack strength x defensive weakness) but we could extend these models?

Let us pause for a few moments so that you can consider this.

These models use data for the whole season, but we could put more emphasis on recent results.

We might also consider that some teams have a better (or worse) home advantage than the league average.

Also, there is no information on the composition of individual teams, for example new players may have joined during the season or some influential players may be injured.

It might even be advantageous to include other information, for example on the weather.

Or even the state of the pitch? Stenhousemuir, and a small number of other teams play all of their home games on synthetic pitches.

The sports betting companies use much more complex models that incorporate more information, and they also have football experts advising them.

[Blow a whistle]

[Click through slides]

The statistical method only predicted one correct score.

It did however predict the correct result for three of the five matches.

The statistical method beat the fan who only predicted two correct results.

The dice only managed one correct result.

Neither the fan (a pseudo-expert) or the dice predicted any correct scores.

We do not advocate using the methods above for gambling.

I stress, we do not advocate using the methods above for gambling.

There was a popular saying when I was a boy that it was not by chance that at my local bookmakers there were five windows for placing bets and only one window for collecting winnings.

The models outlined above are simple Poisson models that are the product of only a few terms (e.g. home advantage x attack strength x defensive weakness) but we could extend these models as I have noted above.

You might also have thought of some addition information that could be included in the analyses.

On further reflection an underlying problem is that any score combination (0 – 0 to six all) is one of 49 cells on a 7 by 7 grid.

Each specific score has a very low probability.

One technical extension might be to develop a set of confidence intervals to test the coverage of predictions.

It might also be prudent to check if the Poisson distribution is the most appropriate distribution to use when modelling the scores in lower division football matches.

Who knows I might event get around to doing some more work one of these seasons.

We have been discussing count data and thinking about how to use it an analysis.

I said at the start of this film that examples of social science data that take the form of positive counts are legion.

And I used the examples, of how many burglaries take place in a neighbourhood, how many women under twenty gave birth last year, or how many cases of a disease were diagnosed?

Indeed, the prosaic question how many ‘any things’ will usually be answered with a count.

I also said that given the prevalence of count data in the social sciences, for many years it has puzzled me why most social scientists know very little about analysing count data.

The technique known as Poisson regression estimates models of the number of occurrences (i.e. counts) of an event.

The Poisson distribution has been applied to diverse events.

Ladislaus Bortkiewicz analysed the number of soldiers kicked to death by horses in the Prussian army. This was probably the first use of this approach. Over the years I have read various, slightly pedantic, discussions as to whether or not the data were for officers only, or if it included mules and horses, but I must confess I don’t really care.

Clarke analysed patterns of hits by buzz bombs launched against London during World War II.

And, Thorndike analysed telephone connections to a wrong number.

If you are familiar with regression models or the generalised linear modelling framework.

Then you will have seen equations like this before.

In essence there is a left hand side (i.e. the outcome variable);

And a right hand side, with a set of explanatory variables. Here they are written as X1 through to Xk .

And, then finally there is an error term.

It is easy to make the conceptual leap to having a count variable as the outcome. and the regression model using information from the Poisson distribution.

There are several different models that are suitable for modelling count data.

The Institute for Digital Research and Education, at UCLA provide this excellent page with examples using Stata, SAS, SPSS, R and Mplus.

Here is an example of an empirical paper that has just been published that employs models for count data.

A stellar early career researcher, Dr Sarah Stopforth, and her colleagues model the number of school GCSE gained at grades A*–C.

They undertake a sensitivity analysis comparing alternative statistical models suitable for count data.

They use a negative binomial regression model rather than a Poisson model because there is evidence of over-dispersion.

Negative binomial regression can be used for over-dispersed count data, that is when the conditional variance exceeds the conditional mean.

In their dataset, there were high proportions of young people with zero counts, so a zero-inflated model was used.

In conclusion, we have been discussing count data and thinking about how to use it an analysis.

The prosaic question - how many ‘any things’ will usually be answered with a count.

Given its prevalence in social science research, it is worth learning about how to analyses count data.

I hope that watching this video and using the accompanying materials will help you to better understand count data and how it can be analysed.
