

A football tipping scheme

1. Introduction. This is a proposed project which uses Statistics, specifically, *multiple regression analysis*, to carry out a football tipping scheme.

In parts of Australia, especially where Australian Rules football dominates, the past-time of footy tipping is extremely widespread in the workplace. There are many models and methods for approaching the problem of how best to produce effective tips. All are essentially statistical in nature, being based on drawing inferences from observed data. The method outlined here is based upon a conceptually simple regression model. Many extensions and variations on it are possible.

The model envisages *ability scores* $\alpha_1, \dots, \alpha_{16}$ for the 16 AFL teams, as well as *home-ground advantage factors* $\beta_1, \dots, \beta_{16}$. Here, α_i is the 'ability' of the i th team, and β_i is a measure of the HGA, the advantage it gets when playing at home against another team. The model is

$$y_{ij} = \alpha_i - \alpha_j + \beta_i I_{ij} + \text{error},$$

where y_{ij} is a measure of the outcome when home-team i plays team j , the margin (positive or negative) by which i beats j , or a transformation of that margin; and I_{ij} is +1 if team i really does get a HGA when playing against team j . Thus, for example, this indicator factor I_{ij} is zero whenever Port plays Adelaide, or whenever two Melbourne clubs meet.

This is a *multiple linear regression model*. A *simple* linear regression model fits a straight line to data, by least squares, and estimates just two parameters, a slope and an intercept. The principle of fitting multiple regression is just the same, but with more parameters estimated. We propose to fit this model using the multiple linear regression operations in the statistical package Minitab.

There are several statistical issues to be considered, including how to deal with 'outliers', unusually large observations which have the capacity to unduly influence the results. These correspond to games which are won (and lost) by huge margins.

The following Sections discuss these points:

- Introduction to Minitab;
- simplifying the regression model;
- entering the data into Minitab;
- methods to modify the effects of outliers;
- how to estimate current and not past form; and
- using the results to make predictions about points margins and probabilities of winning and losing.

It's far from foolproof: randomness plays a big part, and upsets occur frequently. However, the record of tips produced by the system is quite good. Anecdotally, it

seems that winners of football tipping competitions achieve a success-rate between 65% and 70%. The success rate of the proposed scheme up to round 10 is 64.3%. There is a lot of scope for variation of the simple system introduced here.

2. Introduction to Minitab. Minitab is an excellent statistical package for the teaching of Statistics. Indeed, it is the most frequently used package world-wide, for this purpose. A 30-day free trial can be downloaded from the Minitab website www.minitab.com, or alternatively, purchase of the textbook used for the University of South Australia course *Statistical Analysis in Business* comes with a CD containing Minitab.

We have a powerpoint presentation *Footy Tipping.ppt* containing an Introduction to Minitab. This is available upon request by sending a message to bruce.brown@unisa.edu.au

3. Simplifying the regression model. A total of 32 parameters, ie 16 ability parameters and 16 HGAs, is too many. There are ways of reducing this rather large number.

First, note that the ability parameters $\{\alpha_i\}$ can only be determined for teams compared to each other, so it is necessary to set one value to a constant, and determine the other values relative to it. We will do this by setting $\alpha_1 = 0$ for the Adelaide Crows. Therefore all estimated ability scores will measure the ability of teams compared to the Crows. This leaves 15 'ability' parameters to be estimated.

Next, consider how home ground advantage, or HGA, can be described. It is generally accepted that the effect works in two directions, the home team gets a boost through the presence of its supporters, and the away team suffers through absence of its supporters, and perhaps having to travel interstate. To describe both effects would require too many parameters. A much simpler scheme envisages that, with the rationalisation of grounds in Melbourne, all Victorian teams are on an equal footing when playing in Melbourne. It appears that Geelong still retains a HGA when playing at Skilled Stadium. Therefore, HGAs can be modelled by β parameters only for the grounds of non-Melbourne clubs. This means that the ability parameters α are all relative to games played in Melbourne. Thus, non-Victorian clubs will record α values a little lower than their general ability warrants, but for their home games, the addition of a HGA β parameter will raise the predicted level of their performance.

4. Fitting the model using Minitab. The model is called a multiple linear regression model because the predicted response is a linear function of unknown parameters. The coefficients are known values, in our case always +1 or -1. Values of +1 indicate a home team, and possibly, a HGA factor. Values -1 indicate an away team.

The Minitab file, containing the record of past outcomes of games in order to fit this model, is called *AFL.mtp*, and is available upon request. In it, there is a column for each unknown α or β parameter, and the rows correspond to individual games. There is also a 'Margin' column, containing the points margin result of each game. Another column records the Round numbers, from early in the season through to recent rounds.

The multiple regression model is fitted in Minitab by selecting *Stat > Regression > Regression*. A dialogue box then appears. The column 'Margin' is entered under Response, and typing in C2-C21 into the Predictors box enters the columns for all ability and HGA parameters. Under Options, turn off the Fit Intercept box, and under Results, make sure that at least the second radio button 'Regression equation ...' is chosen. Then hit OK.

This is the kind of output which is produced. (This was after just six rounds had been played).

Regression Analysis

The regression equation is

Margin = 45.6 Bris - 2.1 Carl - 8.8 Coll + 15.6 Ess - 0.6 Freo - 12.9 Geel
 + 6.4 Haw + 37.5 Melb + 15.8 North + 17.5 Port - 9.6 Rich + 53.2 StK +
 39.8 Syd - 5.5 WCE + 6.4 WB + 43.6 AAMI + 3.5 Gabba + 28.6 Subiaco +
 54.1 Skill - 23.5 SCG

Predictor	Coef	StDev	T	P
Noconstant				
Bris	45.63	29.94	1.52	0.139
Carl	-2.09	26.44	-0.08	0.938
Coll	-8.83	28.63	-0.31	0.760
Ess	15.62	28.49	0.55	0.588
Freo	-0.61	32.74	-0.02	0.985
Geel	-12.90	28.98	-0.45	0.660
Haw	6.38	27.61	0.23	0.819
Melb	37.48	26.35	1.42	0.166
North	15.79	24.65	0.64	0.527
Port	17.47	27.11	0.64	0.524
Rich	-9.60	23.17	-0.41	0.682
StK	53.18	25.28	2.10	0.044
Syd	39.84	33.09	1.20	0.239
WCE	-5.45	30.65	-0.18	0.860
WB	6.40	28.33	0.23	0.823
AAMI	43.58	27.17	1.60	0.120
Gabba	3.50	36.81	0.10	0.925
Subiaco	28.57	28.36	1.01	0.322
Skill	54.11	37.53	1.44	0.160
SCG	-23.52	36.05	-0.65	0.519

S = 38.29

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	20	49554	2478	1.69	0.099
Residual Error	28	41041	1466		
Total	48	90595			

Ignore the Analysis of Variance Table – this is too advanced. Notice the main features of the fitted model:

- The most able team, relative to Melbourne games, was St Kilda, with $\alpha = 53$. At that stage of the season, also Brisbane and Melbourne were performing well in Melbourne.
- There were large HGAs for both Skilled Stadium ($\beta = 54$) and AAMI stadium ($\beta = 44$). Sydney appeared to perform better away from home than at the SCG (or Telstra Stadium)!

- The item $S = 38$ indicates that the estimated standard deviation of the model's error term is 38. This confirms that there is a lot of random fluctuation in football results, and that tipping accurately is indeed difficult.
- Despite the Crows' apparently poor record at that stage of the season, they had played and lost to five very good teams. Their 'ability' in Melbourne was better than 6 other teams with negative α values (recalling that $\alpha = 0$ for the Crows). For games in Adelaide, when the HGA of AAMI stadium is added, Adelaide would be predicted to lose to only Brisbane and St Kilda, and Port would be predicted to beat all other teams.

5. Reducing the effects of outliers. Some games are won (and lost) by huge margins: St Kilda d Carlton by 108, Melbourne d Carlton by 105, St Kilda d West Coast by 101, Port d Essendon by 96, Kangaroos d Port by 92. These results have the capacity to unduly influence estimated parameters. Indeed, the issue of *robustness against the effects of outliers* is important in the practice of Statistics. A device to counteract the effect of outliers is to apply a transformation of the actual winning margins, which effectively downweights large margins. Any symmetric, monotone, S-shaped transformation will be suitable. In effect such a transformation will say that a winning margin of 40 points is less than twice as good as a margin of 20 points, and so on. The transformation we have used is given by

$$m = \frac{2x}{3 + \sqrt{9 + 12|x|}}, \quad \text{or} \quad x = 3(m + m^2),$$

where x is the actual points margin, and m is the transformed margin. In the Minitab file *AFL.mtp*, the values of m are recorded in the column named *scale*.

Using these transformed values, the model is fitted by selecting *Stat > Regression > Regression*, and entering *scale* as the Response variable.

All the fitted parameters will be now on an ' m -scale', and predicted margins of future games can be re-converted to a predicted points margin by applying the reverse transformation $x = 3(m + m^2)$.

6. Estimating current and not past form. To make sure that current form is being estimated, rather than form from the early rounds, the most recent results can be given a higher weighting. The proposed weights should be stored in a separate column in Minitab, and will be a function of the Round number. Rather than invent a new weighting scale, we have chosen simply to use the Round numbers themselves as weights. Thus the most recent rounds have the highest weights. The model is fitted using a *weighted* least-squares criterion.

In Minitab, after selecting *Stat > Regression > Regression*, under Options enter the column 'Round' under Weights.

Here is the output after 10 rounds of games, using the m transformation, and weighted regression.

Round 10 Regression Analysis

The regression equation is

scale = 1.70 Bris - 1.30 Carl - 0.54 Coll + 1.50 Ess - 1.30 Freo + 1.09 Geel - 2.02 Haw + 1.21 Melb - 0.72 North - 1.22 Port - 1.14 Rich + 3.60 StK + 0.63 Syd - 2.38 WCE - 1.44 WB + 2.41 AAMI + 1.16 Gabba + 4.57 Subiaco + 1.86 Skill - 1.11 SCG

Predictor	Coef	SE Coef	T	P
Noconstant				
Bris	1.704	1.344	1.27	0.210
Carl	-1.298	1.128	-1.15	0.254
Coll	-0.539	1.030	-0.52	0.603
Ess	1.504	1.137	1.32	0.191
Freo	-1.304	1.249	-1.04	0.301
Geel	1.088	1.263	0.86	0.392
Haw	-2.021	1.029	-1.96	0.054
Melb	1.213	1.096	1.11	0.273
North	-0.719	1.058	-0.68	0.499
Port	-1.224	1.039	-1.18	0.244
Rich	-1.135	1.072	-1.06	0.294
StK	3.600	1.109	3.25	0.002
Syd	0.630	1.523	0.41	0.680
WCE	-2.380	1.246	-1.91	0.061
WB	-1.444	1.127	-1.28	0.205
AAMI	2.406	1.220	1.97	0.053
Gabba	1.162	1.537	0.76	0.453
Subiaco	4.567	1.159	3.94	0.000
Skill	1.862	1.617	1.15	0.254
SCG	-1.114	1.592	-0.70	0.487

S = 5.185

The main features now are:

- The most able team, relative to Melbourne games, is still St Kilda, with $\alpha = 3.6$. It is far ahead of other teams, the next being Brisbane and (surprisingly) Essendon.
- There were large HGAs for both Subiaco ($\beta = 4.57$) and AAMI stadium ($\beta = 2.41$). The WA teams perform poorly in Melbourne but with the addition of HGAs, are formidable at Subiaco. Sydney still appears to perform better away from home than at the SCG ($\beta = -1.11$) or Telstra Stadium.
- The item $S = 5.2$ indicates that the estimated standard deviation of the model's error term is 5.2, on the transformed m scale. This is a large figure, and indicates that even when a predicted margin favours one team winning clearly, there is a chance that the other team will win; see the Prediction section below.
- The Crows are looking better: their 'ability' in Melbourne was better than 10 other teams with negative α values (recalling again that $\alpha = 0$ for the Crows). For games in Adelaide, when the HGA of AAMI stadium is added, Adelaide would be predicted to lose only to St Kilda, and Port would be predicted to lose to Brisbane, Essendon, Melbourne and St Kilda.
- Hawthorn are clearly the worst team on current form.

7. Carrying out predictions using the fitted model.

Details of prediction are included in the powerpoint presentation *Footy Tipping.ppt*, available upon request. Here is the calculation for two Round 11 games.

(i) Adelaide v Carlton. The predicted margin at AAMI stadium is $\hat{m} = 0 - (-1.298) + 2.406 = 3.704$, translating to a points margin of $\hat{x} = 3(\hat{m} + \hat{m}^2) = 52.27$, a comfortable win for the Crows. But, with an error standard deviation of 5.185, the probability that Carlton wins, using normal distribution calculations, is $\Pr\{5.185Z \geq 3.704\} = \Pr\{Z \geq 3.704/5.185 = 0.714\} = 0.237$, and Carlton still have a 24% chance of winning.

(ii) Brisbane v Port. The game is at the Gabba, and the predicted margin is $\hat{m} = 1.704 - (-1.224) + 1.162 = 4.090$, translating to a predicted points margin for Brisbane of $\hat{x} = 3(\hat{m} + \hat{m}^2) = 62.45$. The probability of Port winning is estimated to be $\Pr\{5.185Z \geq 4.09\} = \Pr\{Z \geq 4.09/5.185 = 0.789\} = 0.215$.

8. The record. Here is the season record of the transformed, weighted football tipping regression method.

Round	1	2	3	4	5	6	7	8	9	10	11
# winners	ND	NED	NED	3	6	5	6	5	5	6	

ND: no data, NED: not enough data