Predicting Goals per Minute as a Function of Time

Andrew Kinsman (24/3/17)

(All code used can be found, with annotations, in the accompanying R file.)

Data Preparation

The data comprise two csv files, "minutes" (20785 goal timing observations) and "results" (7842 observations of individual games) which relate to football matches played across Europe between Aug 2012 and May 2015. There is no common identifier between the two sets of data, so first we should check the quality of the data to ensure that the two datasets relate to the exact same games and the data they contain are precisely consistent with one another. (A unique game identifier was added to each game in the results.csv to help spot any anomalies.)

In the process of linking the two datasets together, two data quality issues were discovered:

1. Game 4642 (Cagliari-AS Roma on 23/09/12) is missing from the minutes.csv file, even though three goals were scored. This game was therefore omitted in all subsequent analyses.
2. There are 20785 recorded goals in minutes.csv, but (after allowing for the omission of the Cagliari-AS Roma game) only 20781 goals in results.csv, once we add up home goals and away goals. Two most plausible explanations for this discrepancy might be:
    a. Incorrect recording of the home goals/away goals data in results.csv, leading to an under-accounting of the goals scored. (We shall proceed as if this were the case.)
    b. The minutes.csv has 33 exact duplicates. Naturally, some duplicates are to be expected since on occasion two goals will be recorded for the same minute (especially the 45th/90th minutes). However, it is possible that four of these duplications may simply be erroneous.

Additionally, the games that finished 0-0 only appear in the results data and not the minutes data, since there were obviously no goal times to be recorded. These 0-0 games need to be kept careful track of, since they will need to be added to the games in which goals did occur in order to calculate averages and probabilities.

Exploratory Data Analysis and Naïve Model

The "naïve" prediction of the number of goals per minute assumes that the rate of goals scored is constant over the course of the game. This naïve prediction model predicts 0.0295 goals per minute (note the very slight discrepancy with the number given on the supplied task description; which is down to the way that the data quality issues have been addressed).

Let's take a look at a histogram and see whether this assumption of a constant rate of goals over time is realistic (*fig.1*). Two major issues stand out:

1. There is a slight upward trend in goals as the game progresses. This could be caused by combinations of a number of reasons: players get tired, there might be an imbalance of players between the teams due to sendings off or injuries, one or both team(s) may be throwing caution to the wind trying to get an equaliser or winner etc.
2. There are two very large peaks (for minutes 45 and 90) and two very large troughs (for minutes 1 and 46). These outliers are a function of the way that football games unfold. Here the 45th minute is actually “the 45th minute plus any stoppage time” and the 90th minute is “the 90th minute plus any stoppage time”. Naturally, stoppage time could run to several minutes or more, thus greatly widening the window of opportunity for a “45th” or “90th” minute goal. Furthermore, goals in the 1st and 46th minute are relatively rare because each half begins with the ball on the centre spot with each team having 11 men “behind" the ball.

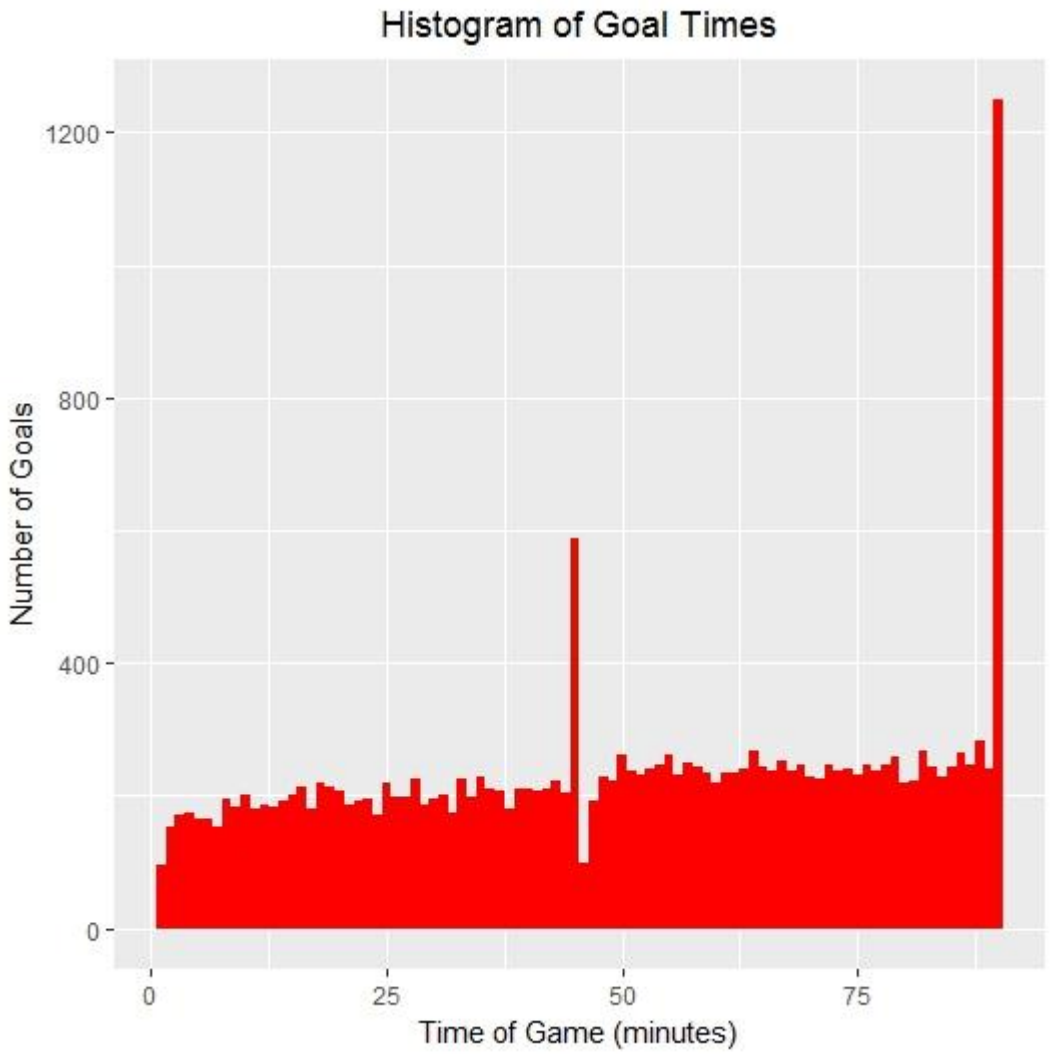

Figure 1

Clearly the naïve model fails to capture either the upward trend in goals over time or these four marked upward and downward spikes.

Let’s drill down into the data to quantify the extent of these spikes.

Top five most common minutes in which a goal is scored:

	Minute	Goals	Goal Probability
1	90	1249	0.15929091
2	45	588	0.07499043
3	88	282	0.03596480
4	64	268	0.03417931
5	82	268	0.03417931

More than twice as many goals are scored in the 90th minute as in the 45th minute, and more than twice as many goals are scored in the 45th minute as during any other single minute. In fact, the last minute of each half has a total 0.23 goal expectancy, which is significantly larger than the ten least common minutes combined.

Bottom ten least common minutes in which a goal is scored:

	Minute	Goals	Goal Probability
1	1	95	0.01211580
2	46	98	0.01249841
3	2	152	0.01938528
4	7	154	0.01964035
5	6	164	0.02091570
6	5	166	0.02117077
7	3	170	0.02168091
8	24	170	0.02168091
9	4	173	0.02206351
10	32	174	0.02219105

Notably, this second list contains the first seven minutes of the game (plus the 46th minute, of course).

Before we try and develop a more sophisticated model, there are few other interesting details. First, the average number of goals per game (i.e. the sum of all the goals per minute probabilities) is 2.651, with 56.33% of goals coming in the second half. Finally, the probability of a 0-0 draw is 0.0802.

Linear Modelling

Here we shall focus on a linear modelling approach, first randomly splitting the data into a training set and a validation set, with 70% of games being allocated to the former. We already know that there are some key outlier elements to the data (the downward and upward spikes at the start and end of each half — see the scatterplot in fig.2), which are going to make trying to model these data problematic. (In the diagram the regression line is systematically underestimating early goals and systematically overestimating late goals, failing to cope well with the presence of outliers at the start and end of each half.)

For the purposes of illustration let's see what happens if we ignore the influential aspects of these spikes and just try to model the number of goals per minute directly as a function of time (*model 1*).

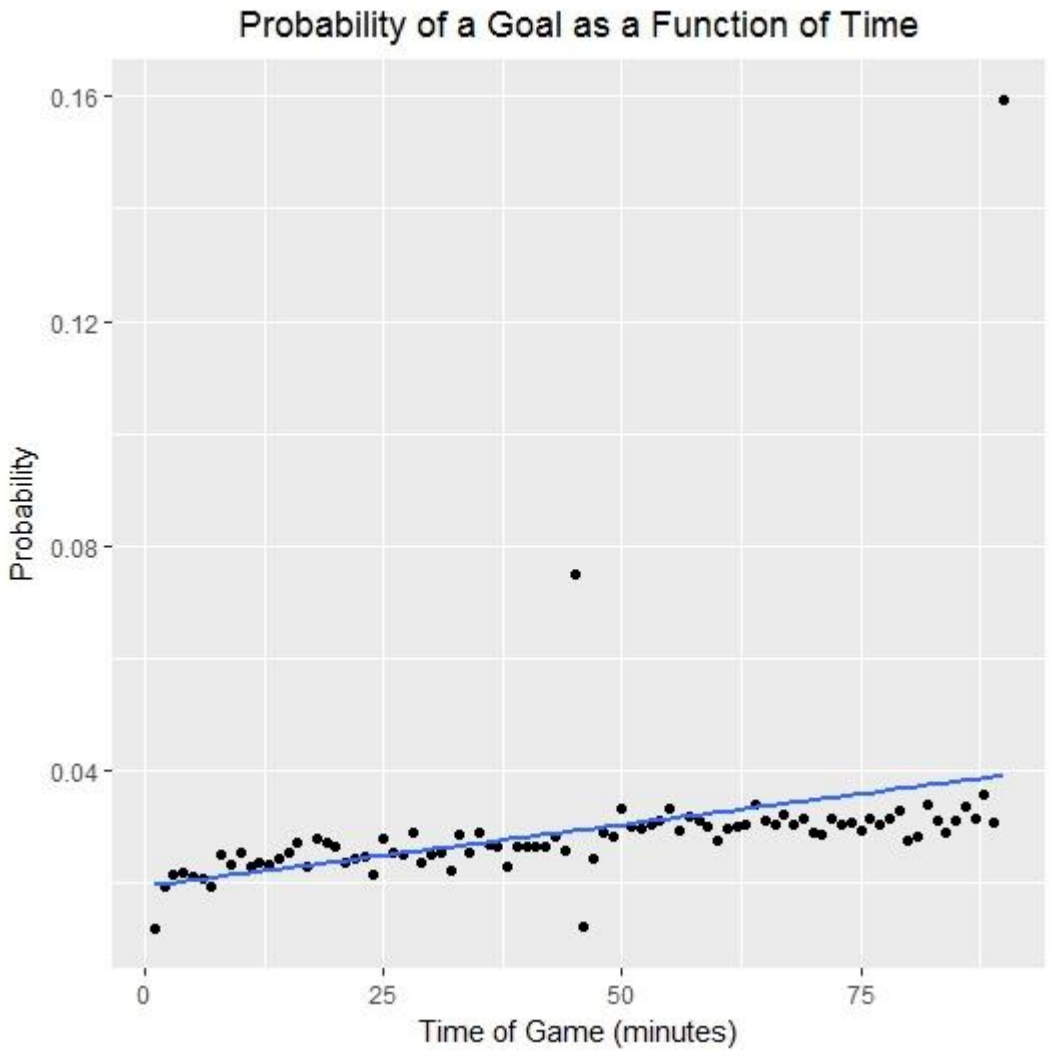

Figure 2

Model 1

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.915e-02  3.063e-03   6.254 1.41e-08 ***
minute      2.228e-04  5.846e-05   3.811 0.000256 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01441 on 88 degrees of freedom
Multiple R-squared:  0.1416,  Adjusted R-squared:  0.1319
F-statistic: 14.52 on 1 and 88 DF,  p-value: 0.0002563
```

Clearly this model is badly flawed, as demonstrated by an R-squared of only 0.14. Only 14% of the variance in probabilities is being explained by the model.

Let’s see what happens if we first remove the four outliers and then try to fit a linear model *(model 2*).

Model 2

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.205e-02  5.575e-04   39.55   <2e-16 ***
minute      1.230e-04  1.067e-05   11.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.002542 on 84 degrees of freedom
Multiple R-squared:  0.6129,  Adjusted R-squared:  0.6083
F-statistic:   133 on 1 and 84 DF,  p-value: < 2.2e-16
```

This time we see a respectable R-squared of over 0.61. How should we interpret the coefficients to identify the goal probability is for any particular minute? Here are two worked examples:

Goals per minute probability in the 2nd minute is 0.02205 + (2*0.0001230) = 0.0223

Goals per minute probability in the 89th minute is 0.02205 + (89*0.0001230) = 0.0330

Let’s now see what happens when we plot these data (*fig. 3*).

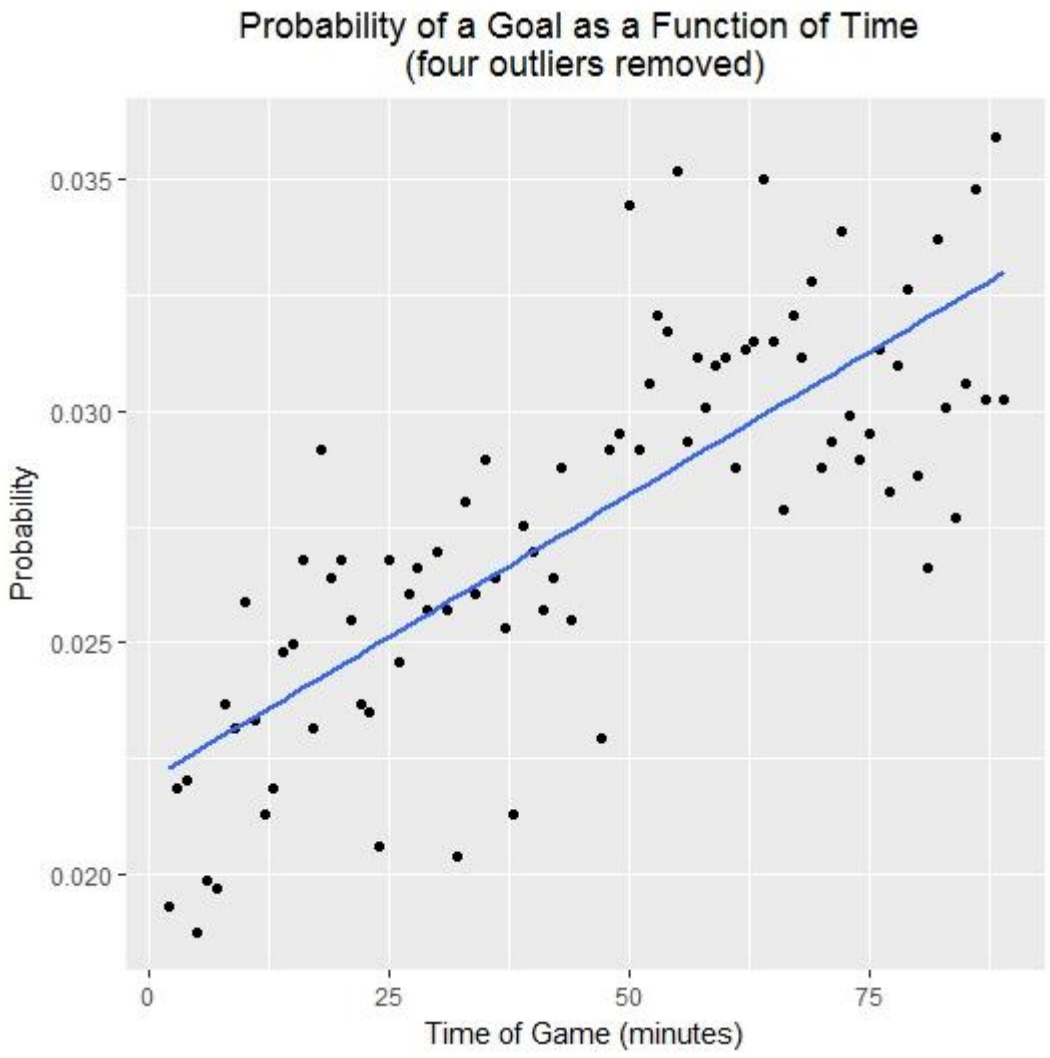

Figure 3

While there are still some issues (the model is clearly overestimating the probability of a goal between the 2nd and 7th minutes), overall this linear model does a reasonable job of fitting the data.

However, we still need to incorporate the start and end of each half into the model. This can be achieved using indicator variables. We create one indicator variable for the 1st minute, another for the 45th, another for the 46th and finally one for the 90th. Each of these indicators is set to 1 (i.e. TRUE) if we are looking at that particular minute and zero otherwise. Thus, in the 2nd minute all the indicator variables are set to zero and have no effect on the model, but for the 45th minute the indicator variable for the 45th minute is set to 1 and the other three indicator variables are set to zero. In this way, we can capture the unique effects of each of the four outlier variables (*model 3*).

Model 3

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.205e-02  5.575e-04  39.548  < 2e-16 ***
minute       1.230e-04  1.067e-05  11.533  < 2e-16 ***
min_1       -1.014e-02  2.601e-03  -3.901 0.000193 ***
min_45       4.676e-02  2.557e-03  18.287  < 2e-16 ***
min_46      -1.532e-02  2.557e-03  -5.990 5.03e-08 ***
min_90       1.267e-01  2.601e-03  48.711  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002542 on 84 degrees of freedom
Multiple R-squared:  0.9745,  Adjusted R-squared:  0.973
F-statistic: 641.8 on 5 and 84 DF,  p-value: < 2.2e-16
```

Here we see that all four of the indicator variables are statistically significant, with the model producing a very high R-squared of 0.9745. Let’s see how we interpret the coefficients when each of the indicator variables come into effect in turn:

Goals per minute probability in the 1st minute is 0.02205 + (1*0.0001230) - 0.01014 = 0.0120

Goals per minute probability in the 45th minute is 0.02205 + (45*0.0001230) + 0.04676 = 0.0743

Goals per minute probability in the 46th minute is 0.02205 + (46*0.0001230) - 0.01532 = 0.0124

Goals per minute probability in the 90th minute is 0.02205 + (90*0.0001230) + 0.1267 = 0.1598

If we compare this model to the naïve model, then for the training data the final model has an RMSE of 0.002 compared to 0.015 for the naïve model, a huge improvement.

However, the real test of any model is how well it can make predictions. In order to assess whether the model can make good predictions on “unseen data”, we now turn to the validation set that was held out for this purpose. We can compare the predictions made by the model to the actual probabilities that appear in the validation set and then calculate a “pseudo r-squared” to try and

assess how much our model may be overfitting the data. Whereas the R-squared for the original model was 0.9745, the pseudo r-squared when the model is applied to the validation set is 0.9599, a small decline in performance but certainly not enough to suggest that the model might be badly overfitting. The linear model with indicator variables (model 3) is doing a good predictive job.

Earlier we saw in *fig.3* that there was a systematic overestimation of goal probability for the 2nd-7th minutes. One last model was attempted which incorporated six further indicator variables, one for each of those minutes. However, none of those indicator variables were found to be statistically significant at the 95% confidence level, so the more parsimonious model 3 was preferred.

Conclusions

Clearly the naïve model that predicts the number of goals per minute as a constant function of time is too simplistic, failing to capture either the upward trend in goals as the game progresses or the outlier issues related to the 1st, 45th, 46th and 90th minutes. Instead we chose to tackle the problem by means of a simple linear model that contained indicator variables that attempt to capture the vagaries of the start and end of each half.

It would certainly be possible to extend this model in various ways, depending on what additional data are available, for example:

1. This model was built on data from several European leagues. It is possible that some leagues are generally more goal-friendly than others, in which case either a model for each league could be created or the league could be added to the existing model as a categorical variable.
2. Likewise, some teams (or even some referees) may be prone to higher-scoring games on average than others, which could be dealt with in the same way.
3. It is also worth considering whether the number of goals in a game varies if there is a mismatch in the skill levels of the two teams, or if (say) both teams are far more proficient as an attacking force than they are in defence.
4. Finally, perhaps more goals are scored in matches played in good weather (or fewer goals scored in matches played in atrocious weather).

In addition, there may be events that happen “in running” that affect the probability of more goals in a game. Are goals more likely when there is currently a one-goal difference (forcing the team that is behind to press forward), or when there is an imbalance in the number of players after a sending-off? And is the number of future goals in a match affected by the number of total (or home, or away) goals that have already been scored in that game (there is a long-standing school of thought that goals become more likely “once the deadlock has been broken”)? Again, all these factors could potentially be incorporated into the model if the necessary “in-running” data are available.

Appendix

Here are the model’s rounded predictions for the number of goals scored in each individual minute of a game:

Minute	Goals scored
1	0.0120
2	0.0223
3	0.0224
4	0.0225
5	0.0227
6	0.0228
7	0.0229
8	0.0230
9	0.0232
10	0.0233
11	0.0234
12	0.0235
13	0.0236
14	0.0238
15	0.0239
16	0.0240
17	0.0241
18	0.0243
19	0.0244
20	0.0245
21	0.0246
22	0.0248
23	0.0249
24	0.0250
25	0.0251
26	0.0252
27	0.0254
28	0.0255
29	0.0256
30	0.0257
31	0.0259
32	0.0260
33	0.0261
34	0.0262
35	0.0264
36	0.0265
37	0.0266
38	0.0267
39	0.0268
40	0.0270
41	0.0271
42	0.0272
43	0.0273
44	0.0275
45	0.0743
46	0.0124
47	0.0278
48	0.0280
49	0.0281
50	0.0282
51	0.0283
52	0.0284

53	0.0286
54	0.0287
55	0.0288
56	0.0289
57	0.0291
58	0.0292
59	0.0293
60	0.0294
61	0.0296
62	0.0297
63	0.0298
64	0.0299
65	0.0300
66	0.0302
67	0.0303
68	0.0304
69	0.0305
70	0.0307
71	0.0308
72	0.0309
73	0.0310
74	0.0312
75	0.0313
76	0.0314
77	0.0315
78	0.0316
79	0.0318
80	0.0319
81	0.0320
82	0.0321
83	0.0323
84	0.0324
85	0.0325
86	0.0326
87	0.0328
88	0.0329
89	0.0330
90	0.1598