Generalized Linear Models: “Non-linear” Terms

Goals:

Learn how to construct and solve GLMs with non-linear terms
Illustrate the process of fitting a GLM model for a concrete applied problem
Illustrate possible tradeoffs between model performance and problem constraints

Assignments:

Use the data on France Soccer Championship 2007-2008.

a) Construct a GLM to predict the probability to win a game (response) using the current standing of both teams (explanatory). (Decide whether you need the quadratic terms in this problem; justify your choice).
b) Discuss the model results. What is the predicted dependence of the game result on the team standing? Does this supports/contradicts the possibility of bribing? Why or why not?
c) Find 95% CIs for all model parameters.
d) Find a 95% CI for the probability to win when the team score is 50 and the opponent’s score is 10.

Reports: Assignments require printed report, which will consist of R-results (do not print the entire session, only the necessary results!) and plots. Describe briefly the theoretical background for the methods you use, including necessary formulas, and make short statements about result interpretation. Consult instructor if you have any questions about the level of detail or formatting of your report.

Reports are due on December 3rd

Essential R commands:

Session management:

```
       help()            ls()              getwd()
       setwd()           library()         data()
       save()            load()            read.table()
       class()           names()           rm()
```

Vectors:

```
       c()               seq()             rep()
       factor()          cbind()           rbind()
```

Data summaries:

```
      mean()             sd()              median()
      quantile()         summary()
```

Graphs:

```
      par()              plot()            points()
      lines()            mosaicplot()      text()
```

GLMs:

```
      glm()              family()          summary.glm()
      predict.glm()
```

Bribery in two-party games

In this Lab, we will be working with data from an experiment focused at bribes in a two-party game. The bribing problem is very important in sports and contract assignments; it can be associated with bid rigging, *etc.* Accordingly, it is important to study the human behavior and decision making strategies related to bribing. The experimental data used here has been recently collected at the University of Pittsburgh by Prof. A. Matros.

Experiment description

In this experiment, the subjects are involved in a two-party game between “player 1” and “player 2”. Each game may result in one of three possible outcomes: player 1 wins, player 2 wins, or there is a tie. After each game, the winner receives 2 points, loser receives 0 points, the tie results in 1 point to each player.

The outcome of each game depends on negotiation between the two players: (a) The game can be fair, meaning that the outcome will be determined by tossing a “tree-sided coin”, with probability 1/3 for each possible outcome; (b) Player 1 (Player 2) can offer Player 2 (Player 1) a monetary prize (bribe) in order to agree that game will be won by Player 1 (Player 2). If the bribe is accepted the game *might* be won by the paying party -- a player can accept bribe and then decide to play a fair game. If the bribe is rejected – the game is fair as in (a).

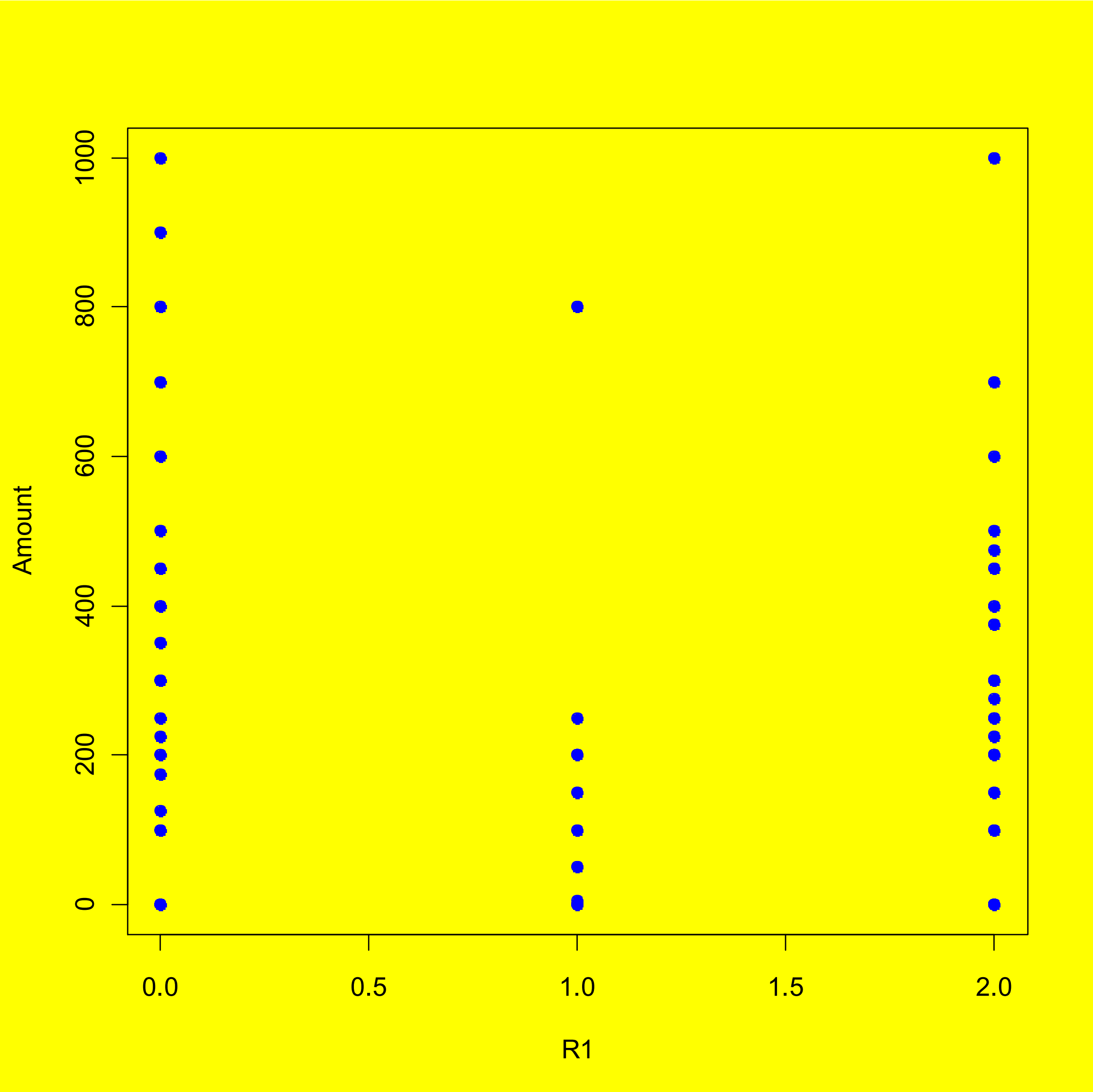
Each experiment participant plays two games with every other participant; the total number of earned points is recorded. At the end of experiment, the players receive monetary rewards, which depend on the total number of points accumulated (the more points, the larger the reward). At the beginning of experiment, each player is given some amount of money, which (s)he can keep or use for bribes. Thus, the total amount received after the experiment equals the sun of the game reward and the money left after bribing.

(There are some other experimental conditions that will not affect our analysis and are not discussed here.)

Our goal is to understand, using the game outcomes, whether people involved in this experiment tend to give bribes and if so, in what situation a bribe is offered.

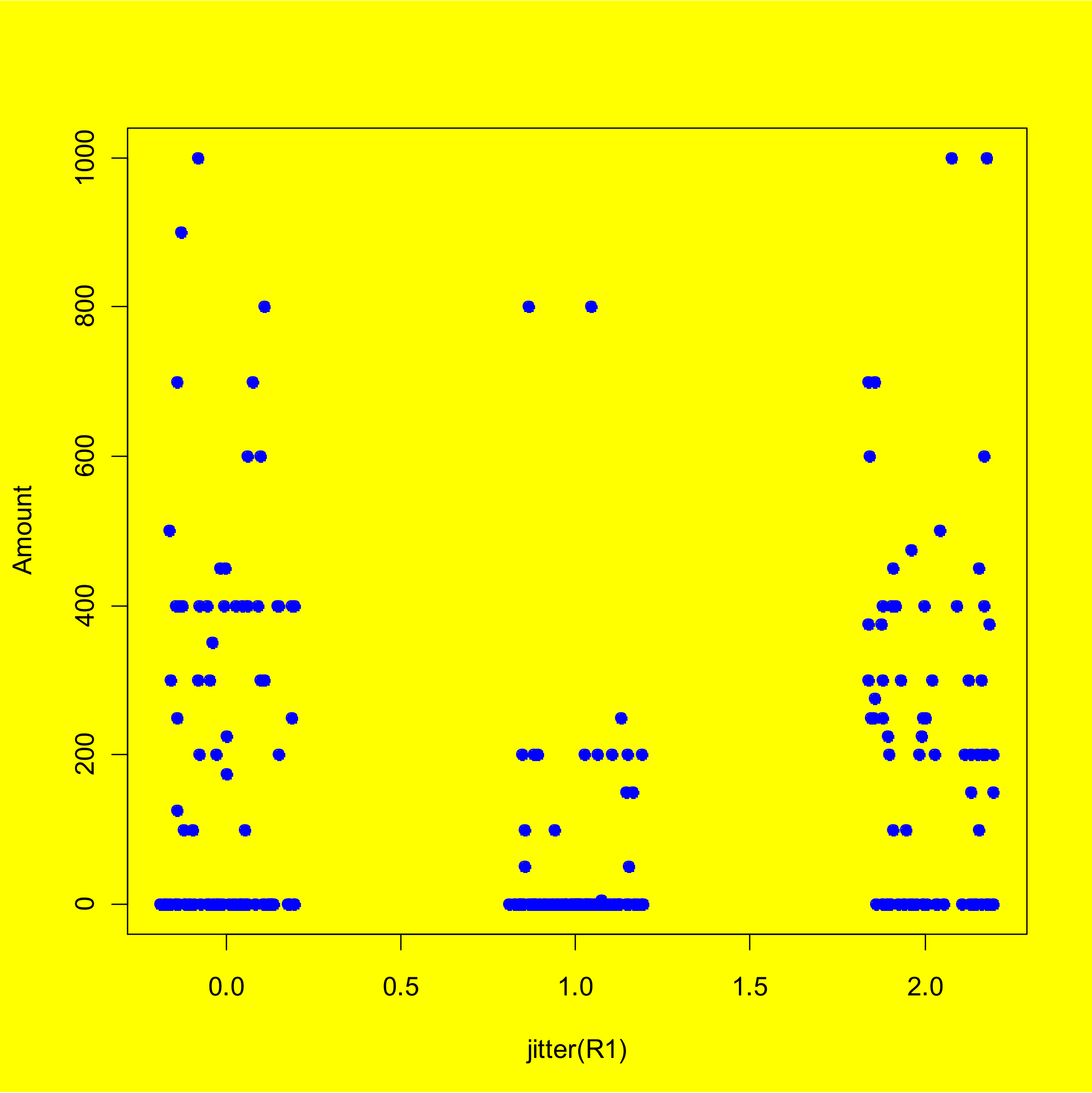
The experimental data are collected in the file ‘bribery.txt”. Each row corresponds to a single game; the file has five columns:

P1	is the total number of points for Player 1 before this game;
P2	is the total number of points for Player 2 before this game;
R1	is the number of points earned by Player 1 in this game;
R2	is the number of points earned by Player 2 in this game;
Amount	is the amount of bribe accepted by the winner of this game.



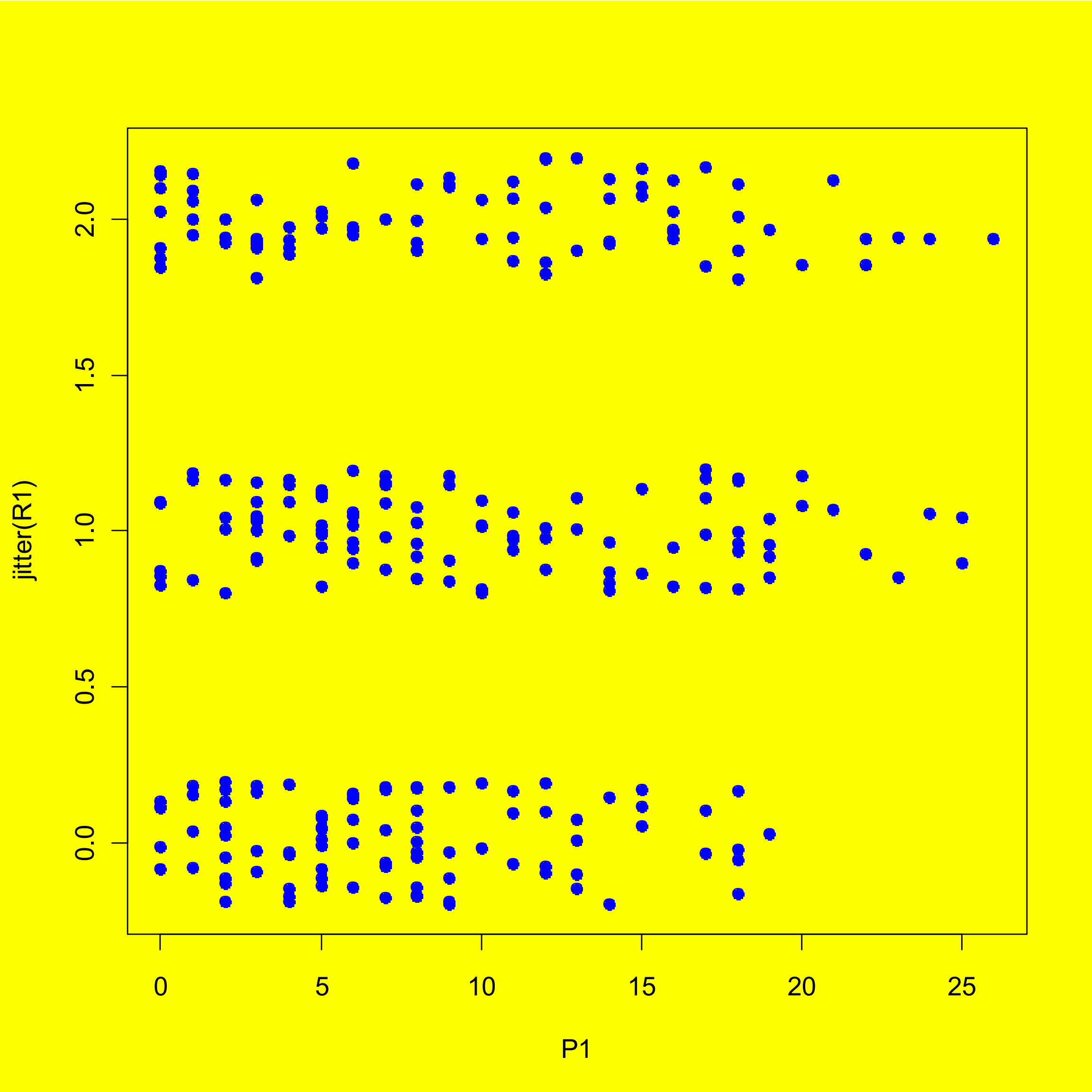
Here, the bribe amount is shown as a function of the game result. It is clear that significant bribes are associated with games with no tie (one player is the winner). Does this mean that bribes *cause* the game not to result in a tie?

```
par(bg='yellow')
plot(R1,Amount,pch=19,col='blue')
```



This is another version of the plot at page 4; it uses jitter for x-axis (game result) and makes the plot more informative, since it separates some overlapping points.

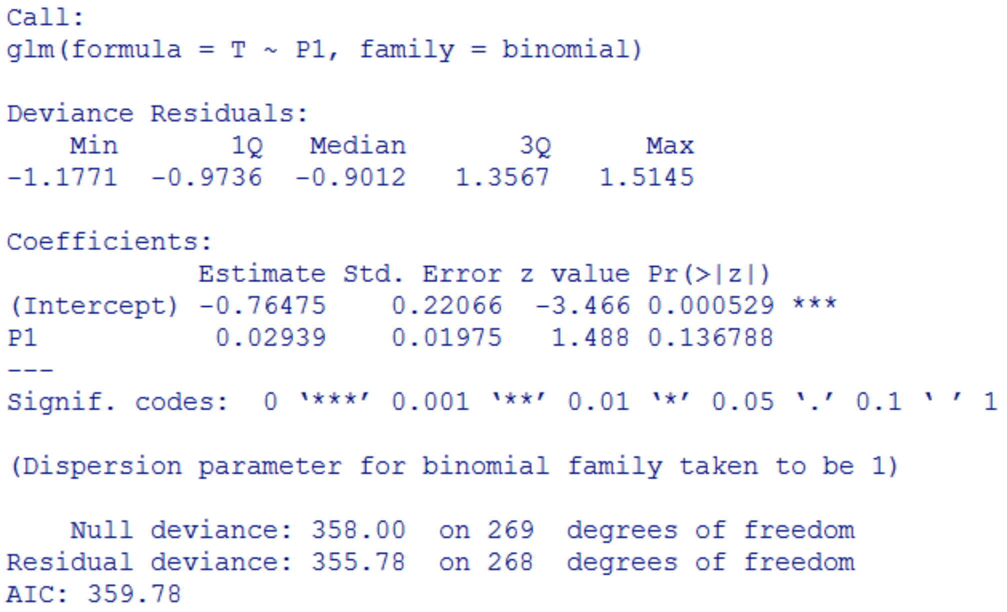
```
par(bg='yellow')
plot(jitter(R1),Amount,pch=19,col='blue')
```



Game result as a function of the current standing (the total amount of points) for the Player 1. This graph suggests that there is no dependence between the current standing of Player 1 and the game result.

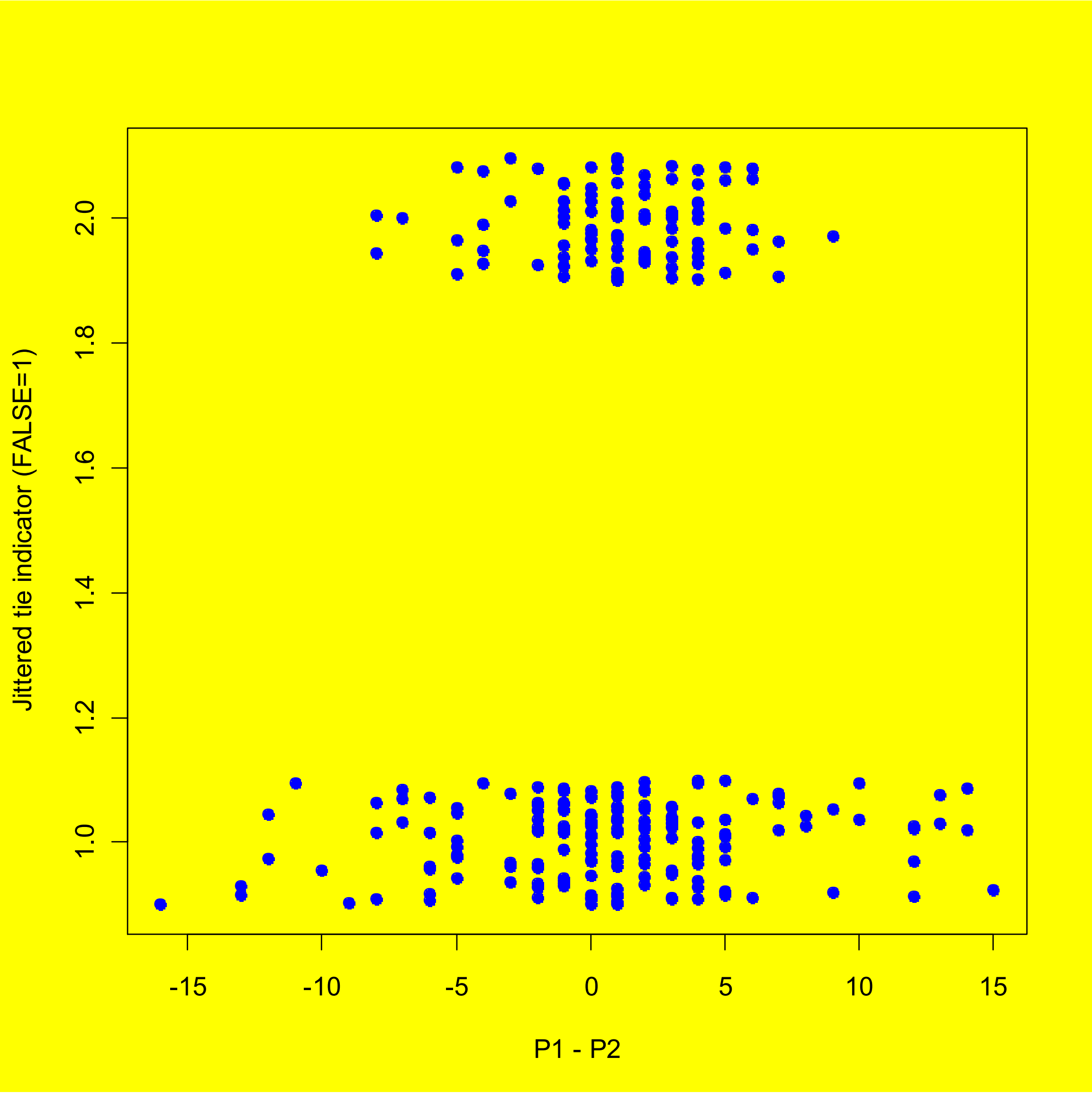
```
par(bg='yellow')
plot(P1,jitter(R1),pch=19,col='blue')
```

```
T<-factor(R1==1) # Tie indicator
g<-glm(T~P1,family=binomial)
```



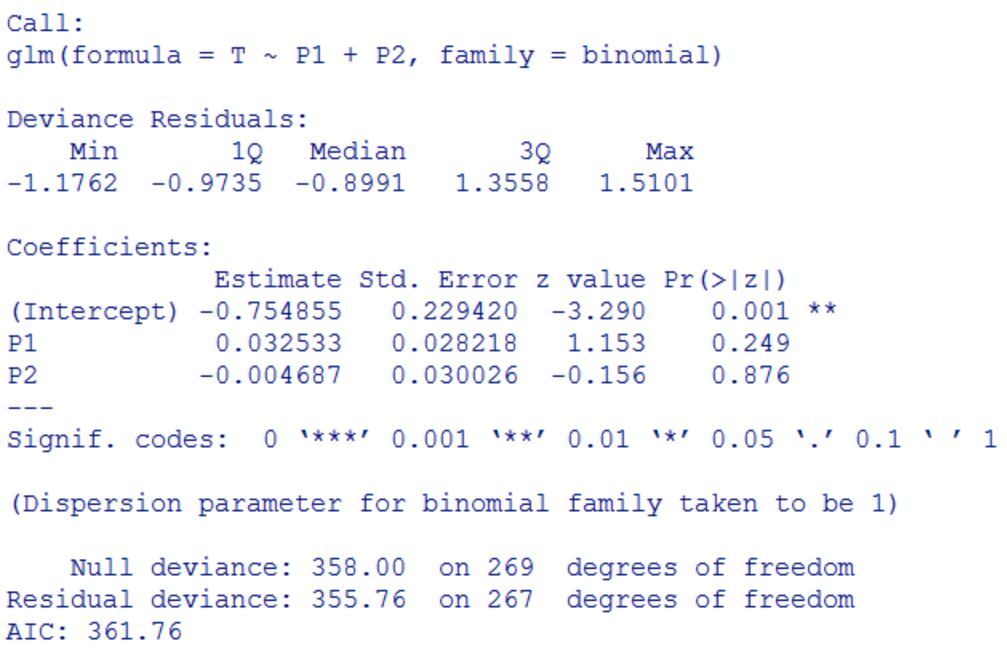
```
Call:
glm(formula = T ~ P1, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-1.1771  -0.9736  -0.9012   1.3567   1.5145  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.76475    0.22066  -3.466 0.000529 ***
P1           0.02939    0.01975   1.488 0.136788    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 358.00  on 269  degrees of freedom
Residual deviance: 355.78  on 268  degrees of freedom
AIC: 359.78
```

The observation from the graph on page 6 is supported by formal analysis. The effect of parameter P1 (points earned by player 1) on T (tie indicator) is not significant.



Tie indicator (FALSE = 1) as a function of the point difference for two players. The graph suggests that tie only happens when both players have comparable number of points earned (within 7 from one another). When one is doing significantly better than the other (more than 7 points difference), the game tends not to result in a tie. This cannot happen if all the games are fair, since in this case the game outcome is independent of player standing in the tournament. Thus, we have a possible indication of bribery.

In the next page we try to test this observation by formal analysis via GLM.

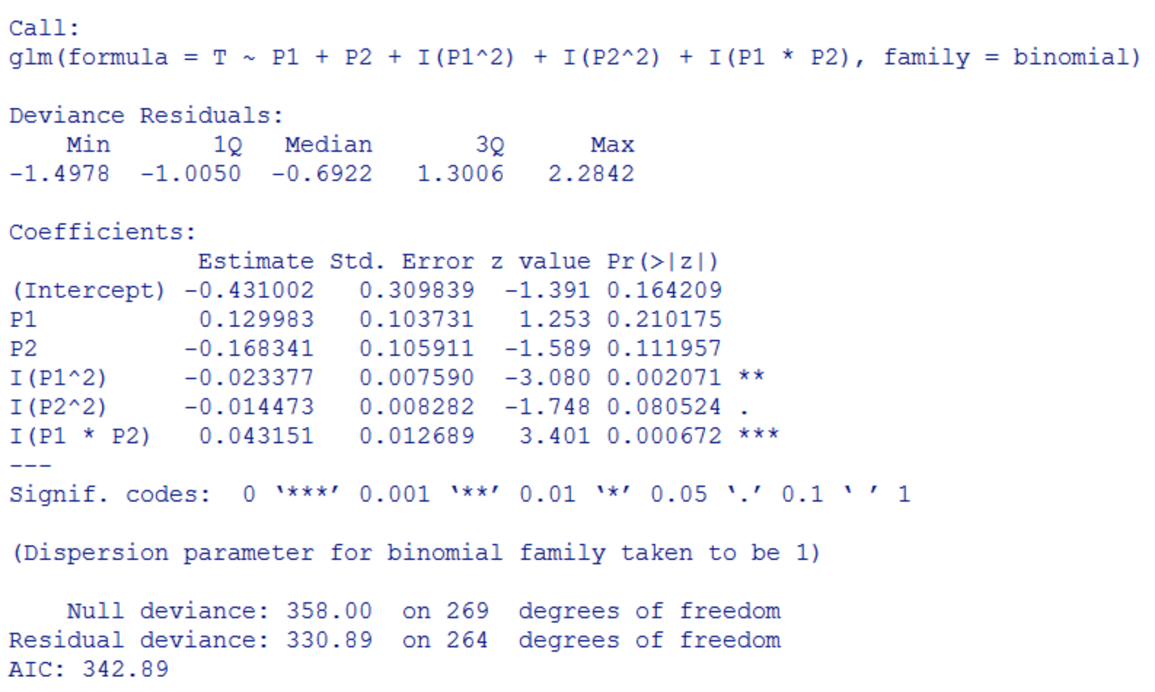


```
Call:
glm(formula = T ~ P1 + P2, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1762  -0.9735  -0.8991   1.3558   1.5101

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.754855   0.229420  -3.290    0.001 **
P1           0.032533   0.028218   1.153    0.249
P2          -0.004687   0.030026  -0.156    0.876
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 358.00  on 269  degrees of freedom
Residual deviance: 355.76  on 267  degrees of freedom
AIC: 361.76
```

However, formal analysis does not seem to support the observation of page 8. The coefficients P1 (player 1 points) and P2 (player 2 points) are not significant. The residual deviance is about the same as the null deviance.

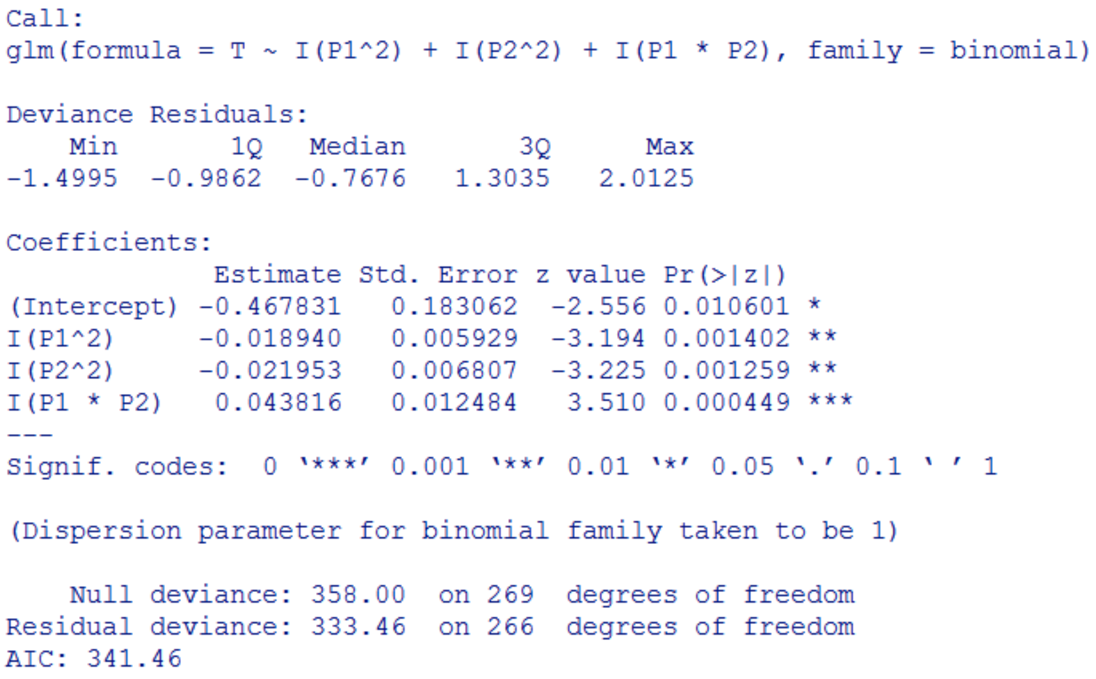
This is not surprising though, since no *linear* combination of P1 and P2 can describe the pattern of page 8: the probability of having a tie is low when P1-P2 is low or high, and high, when P1-P2 is about 0. This is a non-linear effect!

Let's try to catch this effect by using a non-linear term is out GLM.



```
Call:
glm(formula = T ~ P1 + P2 + I(P1^2) + I(P2^2) + I(P1 * P2), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4978  -1.0050  -0.6922   1.3006   2.2842

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.431002   0.309839  -1.391 0.164209
P1           0.129983   0.103731   1.253 0.210175
P2          -0.168341   0.105911  -1.589 0.111957
I(P1^2)     -0.023377   0.007590  -3.080 0.002071 **
I(P2^2)     -0.014473   0.008282  -1.748 0.080524 .
I(P1 * P2)   0.043151   0.012689   3.401 0.000672 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 358.00  on 269  degrees of freedom
Residual deviance: 330.89  on 264  degrees of freedom
AIC: 342.89
```

Here, we allowed for quadratic terms in our GLM. The result suggests that the linear terms P1 and P2 are not significant, while all quadratic terms are significant. Let us re-estimate the model only using quadratic terms.



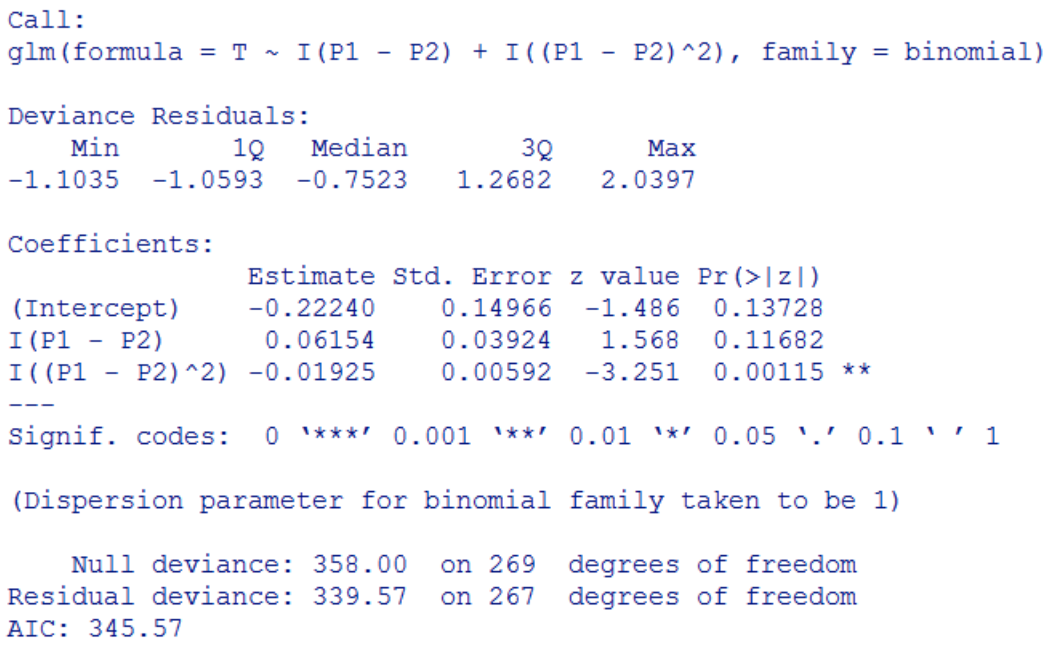
```
Call:
glm(formula = T ~ I(P1^2) + I(P2^2) + I(P1 * P2), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-1.4995  -0.9862  -0.7676   1.3035   2.0125  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.467831   0.183062  -2.556 0.010601 *  
I(P1^2)     -0.018940   0.005929  -3.194 0.001402 ** 
I(P2^2)     -0.021953   0.006807  -3.225 0.001259 ** 
I(P1 * P2)   0.043816   0.012484   3.510 0.000449 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 358.00  on 269  degrees of freedom
Residual deviance: 333.46  on 266  degrees of freedom
AIC: 341.46
```

In this model, all the terms are significant. The suggested equation for the probability *p* of having a tie is

$$\begin{aligned} \text{logit}(p) &= -\ 0.47 - 0.02(P1)^2 - .02(P2)^2 + 0.04(P1*P2) \\ &= -\ 0.47 - 0.02(P1-P2)^2 \end{aligned}$$

Notice that this equation confirms our intuitive feeling from page 8 that the probability of a tie depends on (P1−P2), which is the difference in players' standing in the tournament. Here, this parameter has naturally resulted from a model, and has not been forced by us. This is very good news, since this parameter *makes sense* for our problem.

Finally, let us run a model using explicitly this parameter.



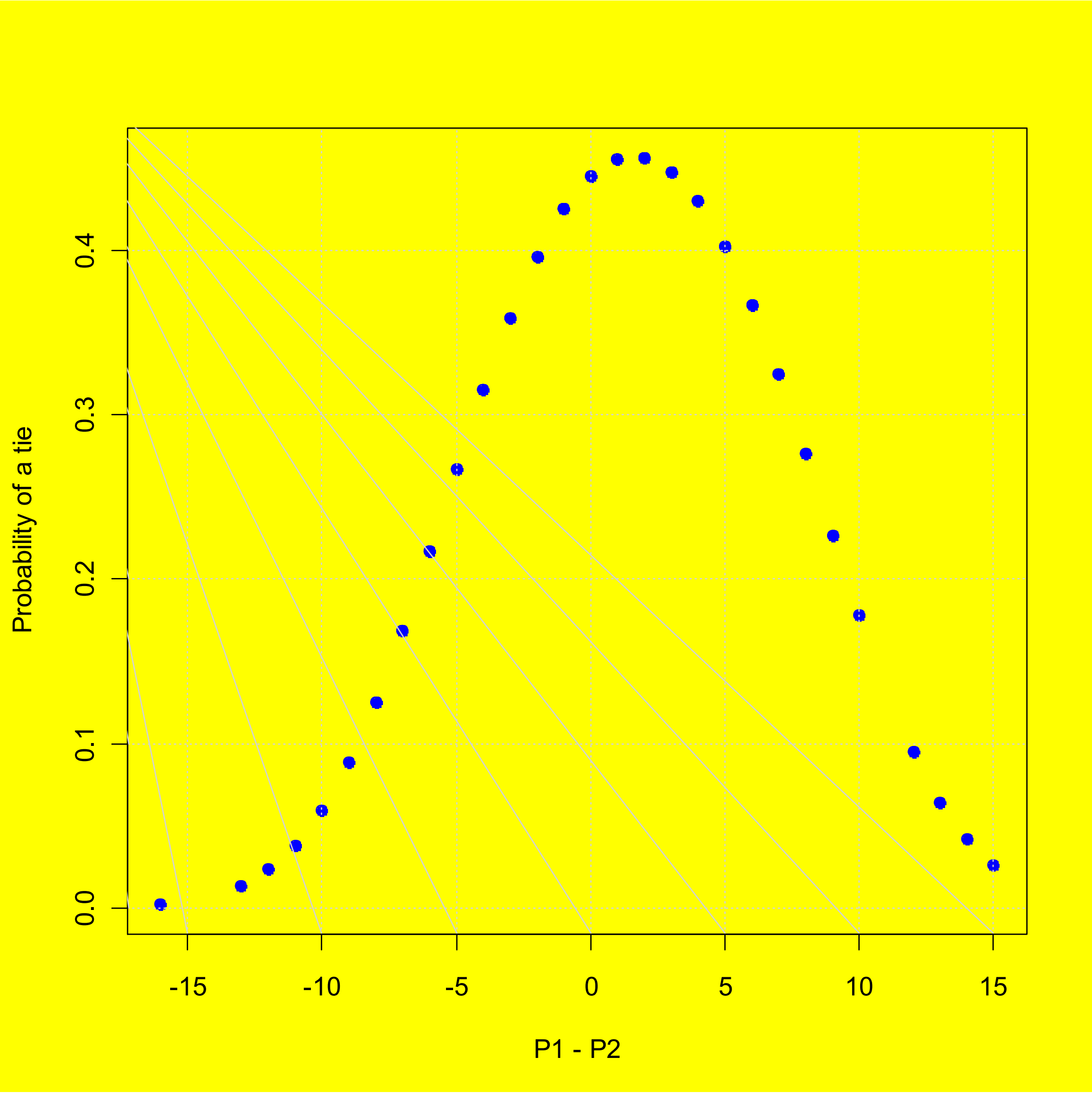
```
Call:
glm(formula = T ~ I(P1 - P2) + I((P1 - P2)^2), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1035  -1.0593  -0.7523   1.2682   2.0397

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.22240    0.14966  -1.486  0.13728
I(P1 - P2)      0.06154    0.03924   1.568  0.11682
I((P1 - P2)^2) -0.01925    0.00592  -3.251  0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 358.00  on 269  degrees of freedom
Residual deviance: 339.57  on 267  degrees of freedom
AIC: 345.57
```

The final model is

$$\mathrm{logit}(p) = -0.22 + 0.06(P1-P2) - 0.02(P1-P2)^2$$

(You can check that the models without the intercept and (P1-P2) term, as well as models with a cubic term, have worse performance, according to both AIC and deviance.)

The prediction is shown in the next page.

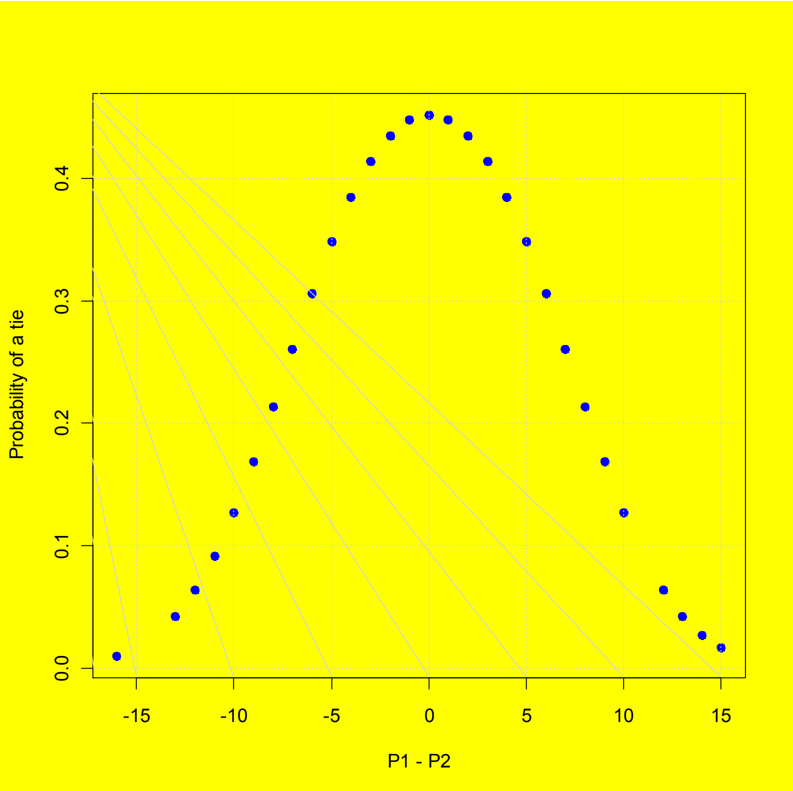


The predicted probability of having a tie as a function of the earned point difference between the two players.

The only problem with this model is that it gives prediction, which is asymmetric with respect to the players. To avoid this, we can force the model to be symmetric by leaving only (P1-P2)^2 term.

```
Call:
glm(formula = T ~ I((P1 - P2)^2) - 1, family = binomial)

Deviance Residuals:
   Min       1Q   Median       3Q      Max  
-1.177   -1.142   -0.826    1.186    1.942  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
I((P1 - P2)^2) -0.021247   0.005255  -4.043 5.28e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 374.30  on 270  degrees of freedom
Residual deviance: 343.91  on 269  degrees of freedom
AIC: 345.91
```

$$\text{logit}(p) = -\ 0.02(P1-P2)^2$$


Probability of a tie
P1 - P2