Model Checking

Owen G. Ward

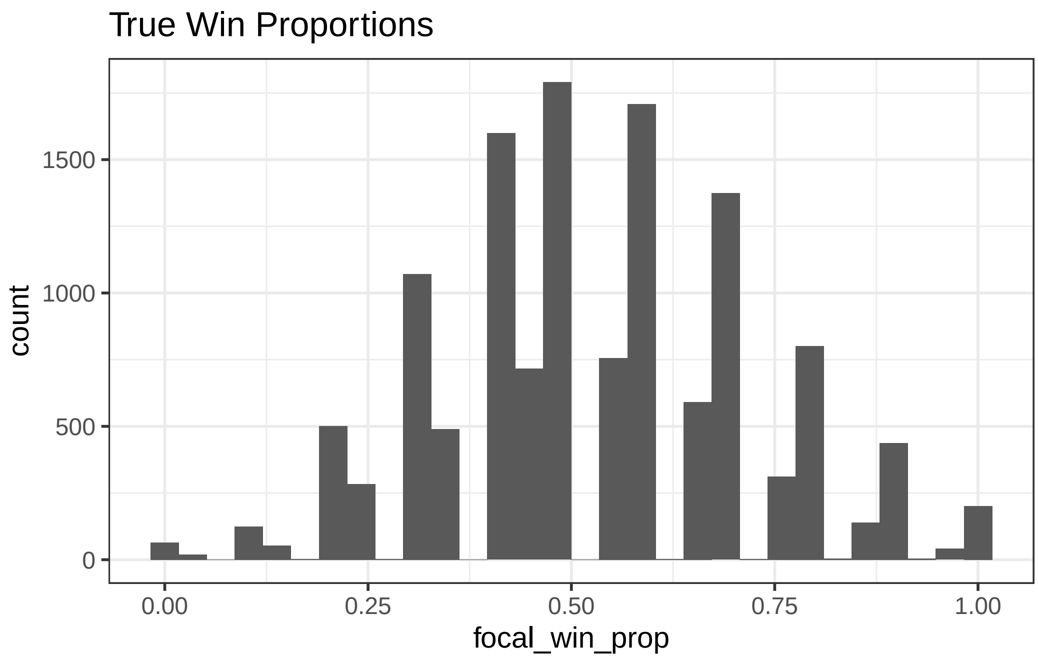
4/17/23

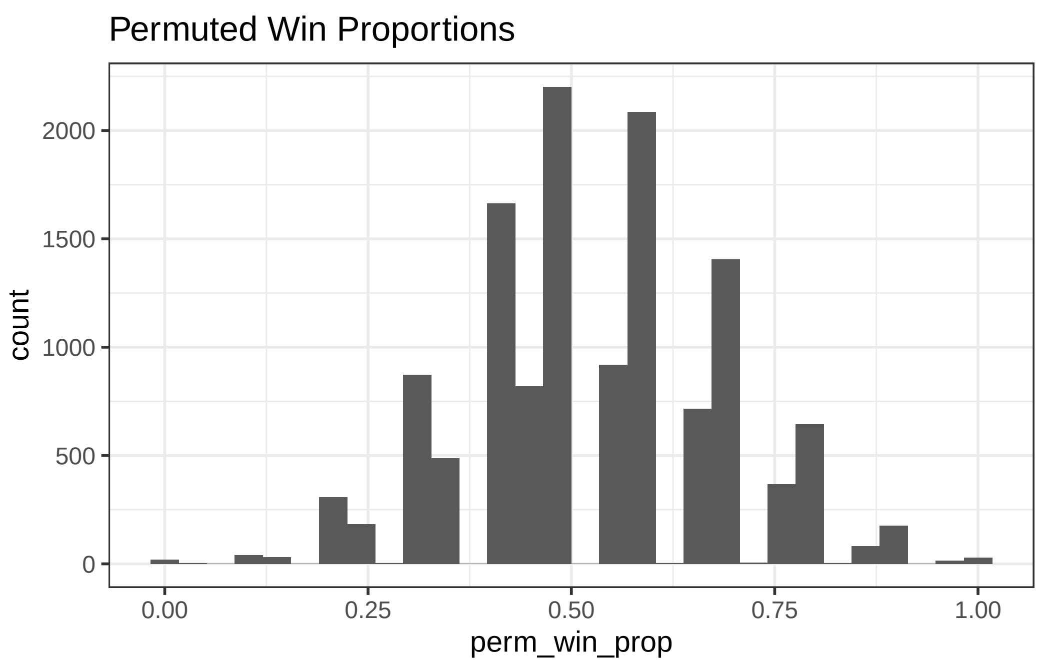
Identifying Problems

An issue was identified with the current implementation, where the `previous_win_history` used, which was supposed to be the proportion of the past n games won compared to the players overall win percentage, contained the result of the current game (the outcome). When this is corrected the strong equal positive winner effects are removed, and we now see a variable effect, with some players still displaying positive winner effects. However, several of the players no longer show such an effect.

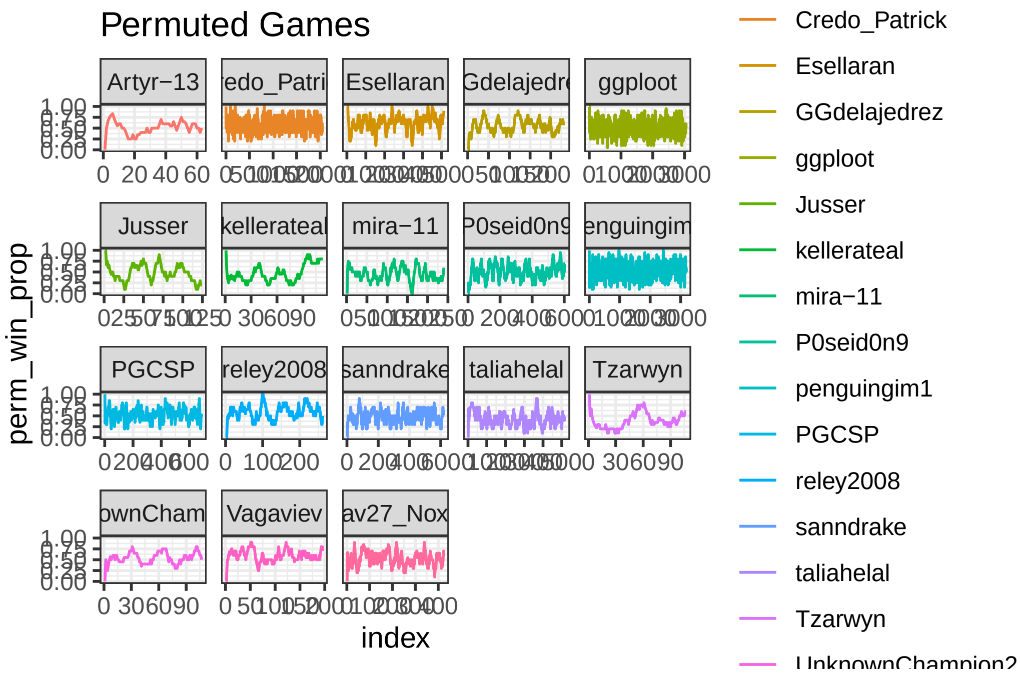
Permuting Data

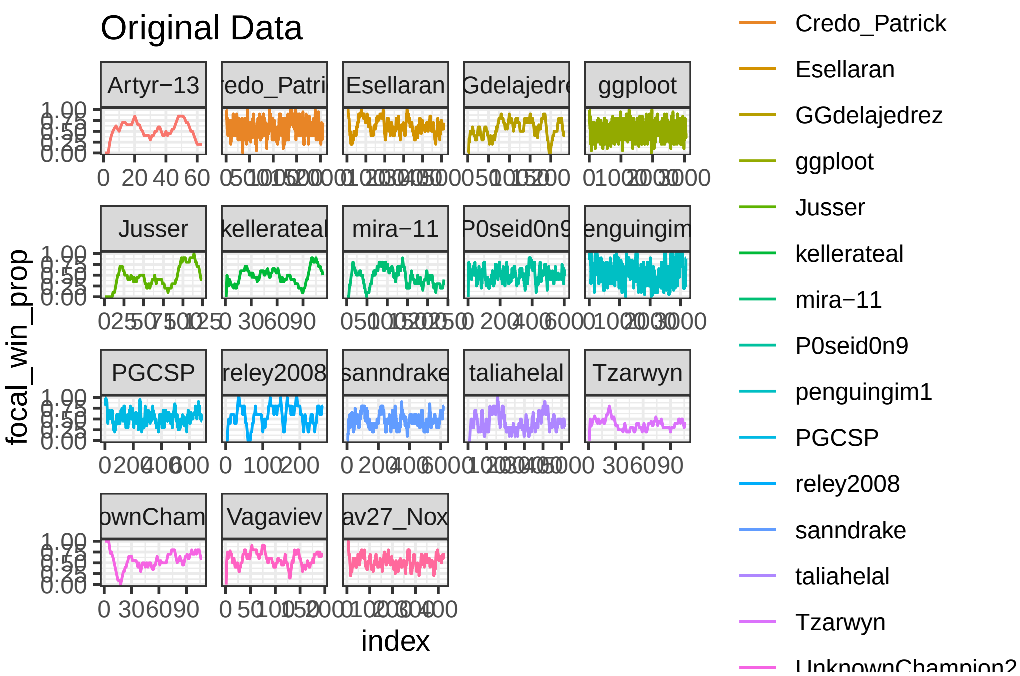
A quick sanity check on the existing model is to see if there are still strong winner and loser effects present when we permute the ordering of the events. This can be checked quite easily.





Can also plot the current previous n win proportion over time. Under the permuted model this should look more like random noise.

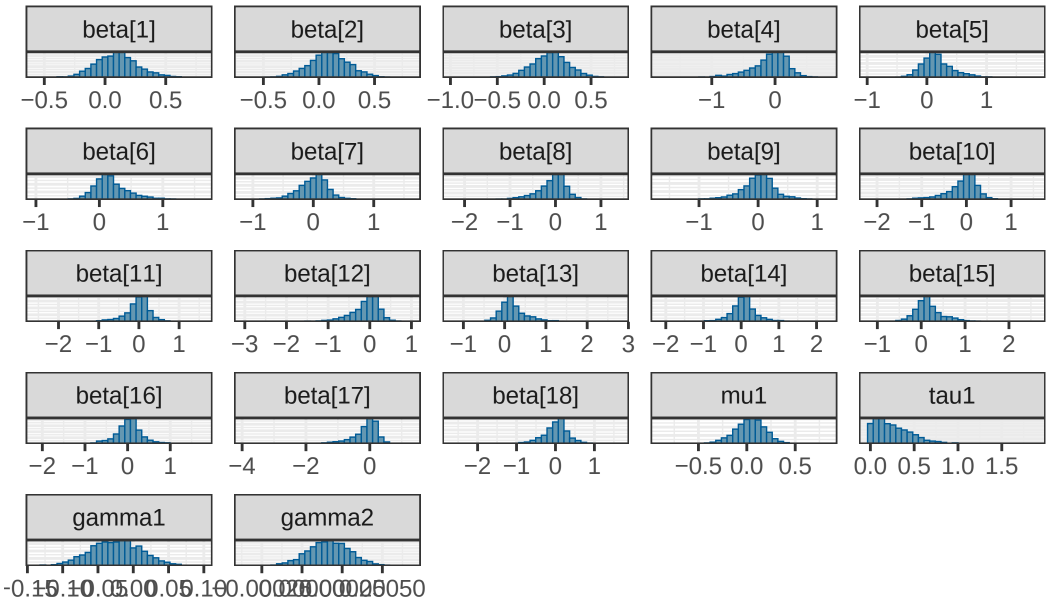




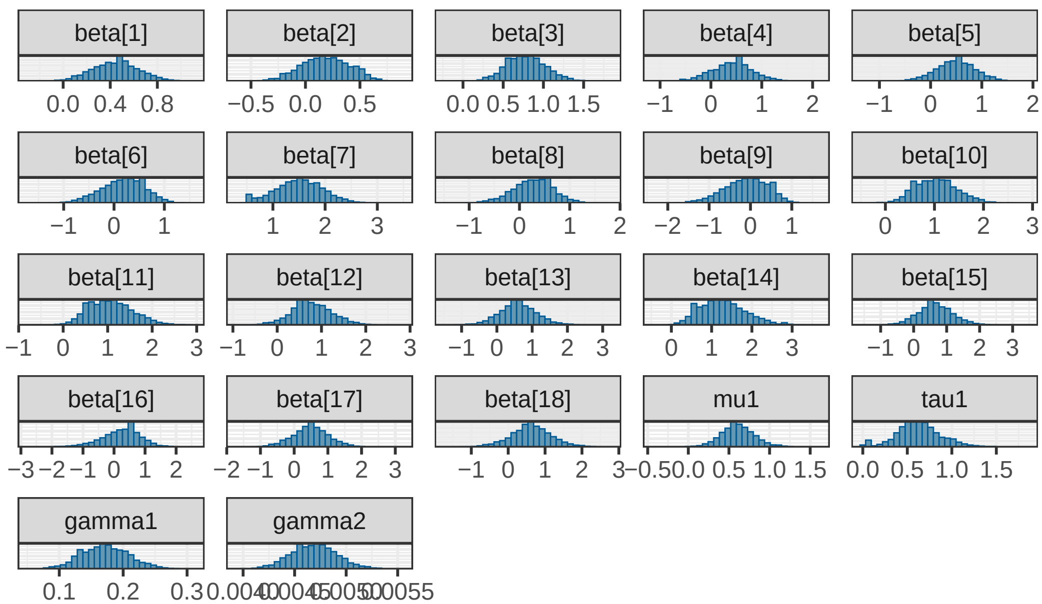
Now we fit the stan model to the permuted data and investigate what the effects are.

Then we want to compare the estimates from this permuted model to the true data.

Permuted Data



True Data




```
## for the permuted data
prob_positive(fit3_ave)
```

```
beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]
0.71300  0.70475  0.63675  0.43525  0.75200  0.76200  0.53200  0.45525
```

```
 beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] beta[15] beta[16]
 0.52675  0.46900  0.51400  0.44725  0.76300  0.59750  0.71775  0.56225
beta[17] beta[18]
 0.43175  0.56350
```

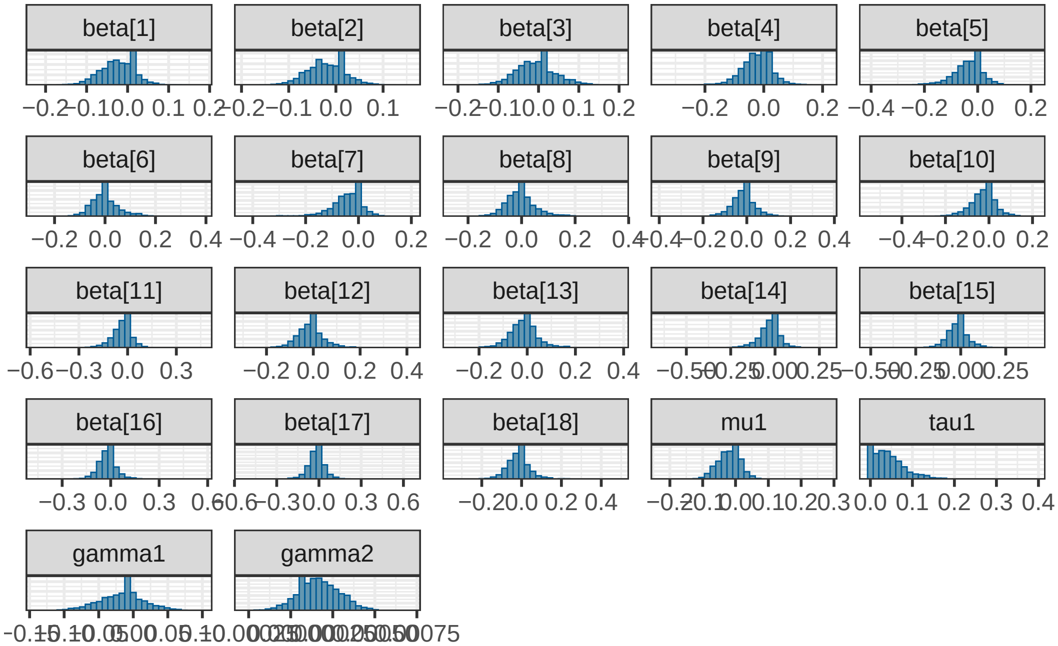
```r
## for the real data
prob_positive(fit3_orig)
```

```
 beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]
 0.98825  0.78700  0.99925  0.84750  0.89525  0.67200  1.00000  0.75475
 beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] beta[15] beta[16]
 0.44675  0.99300  0.99400  0.95675  0.87475  0.99700  0.89225  0.67125
beta[17] beta[18]
 0.82375  0.87375
```

Permuted data gives effects which are all zero on average, which is reasonable. This would be expected as permuting the ordering of the games should remove any possible winner effects. Similarly, the probability of positive effects is closer to 0.5, which would be expected under a model with no effects.

Another way to confirm that having no history should lead to no winner effects is shown below. We now replace the historic win proportion with uniform draws on the interval $[-1, 1]$. The estimated winner effects under such a history are shown below.

We also see that the probability of these effects being positive is close to 0.5, as would be expected under this model.




```
 beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]
 0.33900  0.35125  0.46900  0.36750  0.30950  0.50850  0.27575  0.48425
 beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] beta[15] beta[16]
 0.41875  0.36550  0.32800  0.44900  0.42400  0.34150  0.42275  0.39000
beta[17] beta[18]
 0.41950  0.42600
```

Initial Model

For completeness, we include the initial model output here again, to show the results.

$$P(y_{ij} = 1) = \frac{1}{1 + \exp(-(\alpha_j + \beta_j x_{ij} + \gamma z_{ij}))}$$

where:

- α_j is a player level random effect
- β_j is a player level random effect, accounting for the win ratio x_{ij}
- γ is a fixed effect of game level coefficients

We then partially pool the α and β coefficients, so

$$\alpha_j \sim \mathcal{N}(\mu_2, \tau_2)$$

$$\mu_2 \sim \mathcal{N}(0, 5)$$

$$\tau_2 \sim \mathcal{C}auchy(0, 5)$$

and similarly

$$\beta_j \sim \mathcal{N}(\mu_1, \tau_1)$$

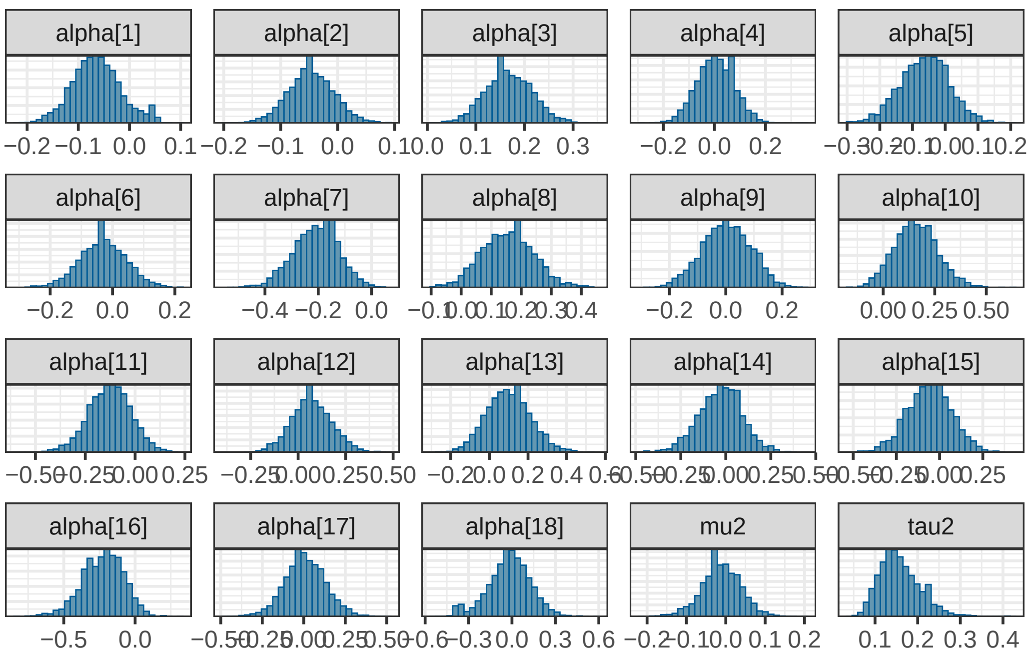
$$\mu_1 \sim \mathcal{N}(0, 5)$$

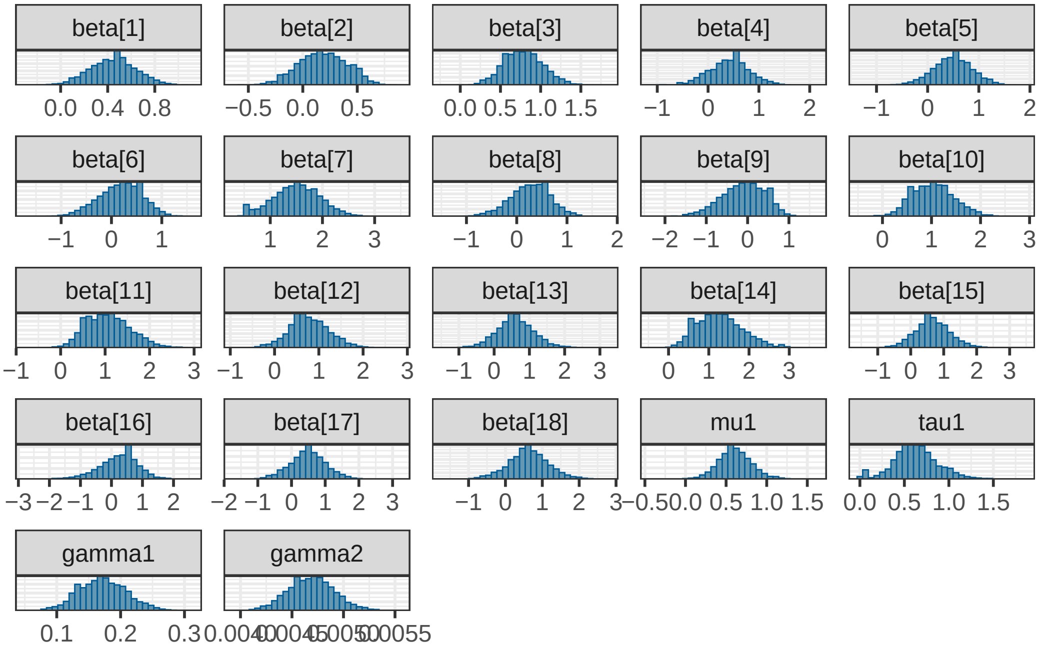
$$\tau_1 \sim \mathcal{C}auchy(0, 5)$$

We rescale the win ratio x_{ij}, by the average win ration of player j. This means β_j captures how a players win probability changes as their historic performance deviates from the overall average win ratio.

For z_{ij} we incorporate 2 covariates, namely the colour of the focal player and the ELO difference between the focal player and their opponent. So that means we have γ_1, which tells us how the win probability changes in going from black to white, and γ_2 gives how the win probability changes as the focal player has a better ELO than their opponent.

We show the posterior estimates for these parameters below.






```
prob_positive(fit3_ave$draws())
```

```
 beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8] 
 0.98825  0.78700  0.99925  0.84750  0.89525  0.67200  1.00000  0.75475 
 beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] beta[15] beta[16] 
 0.44675  0.99300  0.99400  0.95675  0.87475  0.99700  0.89225  0.67125 
beta[17] beta[18] 
 0.82375  0.87375 
```

Interpreting this output

Given this output, we can get the probability a player who is black will win against an equally ranked player, when their current win ratio is exactly their average will be $\frac{1}{1+\exp(-\alpha_j)}$. So a positive α_j indicates a player favoured to win even when playing black. The overall μ_2 tells us about the global probability of winning as black against an equally ranked opponent, which is just below 0, indicating a slightly less than 50% chance of winning. If a player is playing as white their win probability will be $\frac{1}{1+\exp(-(\alpha_j+\gamma_1))}$, so γ_1 being positive indicates the win probability increases when playing white, which makes sense. Playing white makes a player who has 50% chance of winning as black have approximately a 54% chance of winning with white. Wikipedia says the increase in win probability from playing white has been estimated to be 4-6% so quite similar.

Similarly, γ_2 tells us about how the ELO of the focal player being larger than their opponent influences their win probability. The estimate here indicates that a large ELO difference makes a player very likely to win, as would be expected. For example, a player who has a 50% chance of winning against an opponent with the same ELO will have about a 61% chance of winning against an opponent with an ELO score 100 points lower.

How big are these effects

We then want to quantify how much these estimated effects change the expected win probability. We have seen the size of the effects above for the non history parameters. The largest posterior mean for a $\beta \approx 1.56$. Since the normalized win proportion is between -0.6 and 0.6, that means that for a game where this player has a 50% chance of winning with a normalized average win proportion of 0, if their normalized win proportion is 0.5, the win probability will increase by $\approx 19\%$. This would occur, for example, if they win 50% of their total games but have won all 10 of their last 10. This is a similar size effect to the increase in win probability which would come from playing an opponent with an ELO ranking that is 170 points lower.

Similarly, a normalized win history 0.25 above average corresponds to a difference of approximately 80 in ELO scores.

The underlying parameter μ_1 has posterior mean of 0.606. A normalised win history of 0.25 with a winner effect of this size would increase the win probability by about 4%, a similar effect as playing as white instead of black, all other covariates being equal.

Separating out history

We want to next look at the history as both the current win streak, up to but not including the previous game, along with the result of the previous game, to see if we can identify how potential winner effects can be decomposed between these two components.

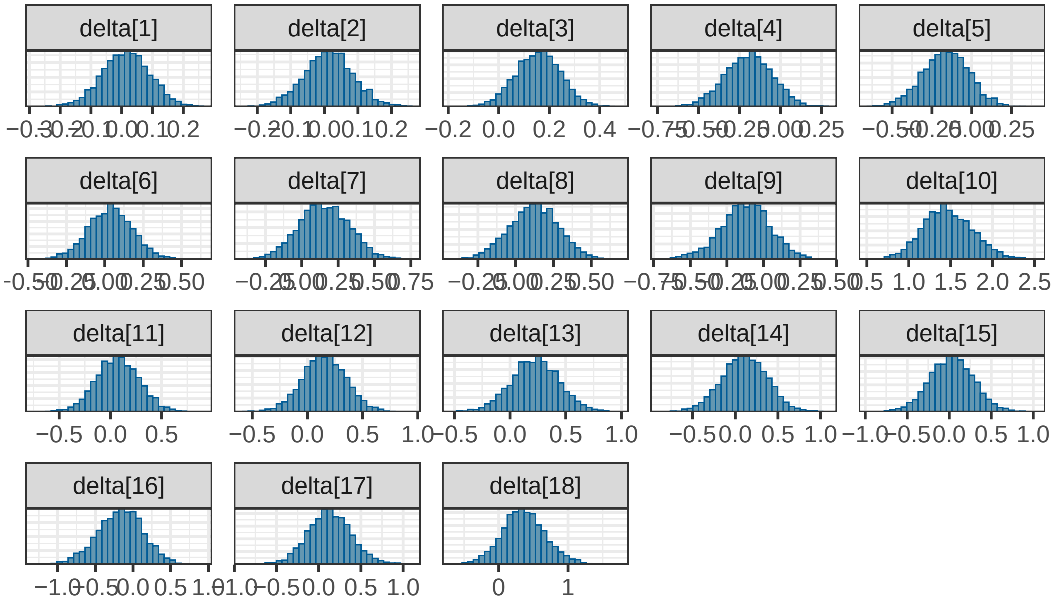
We do this with a model of the form

$$P(y_{ij} = 1) = \frac{1}{1 + \exp(-(\alpha_j + \beta_j x_{ij} + \delta_j x^*_{ij} + \gamma z_{ij}))},$$

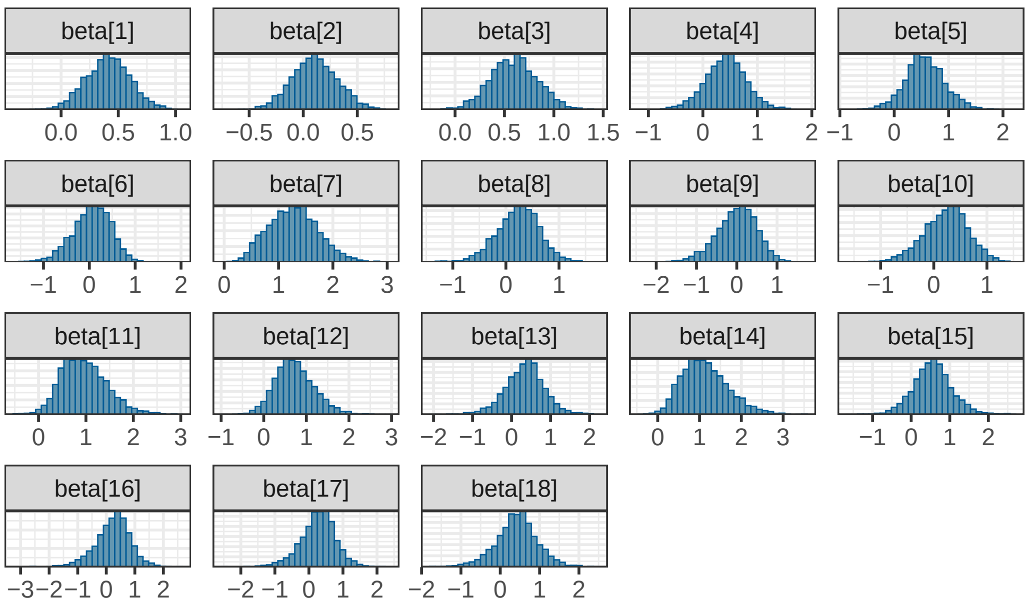
where now x^*_{ij} is the result of the **previous** game played by player j, and x_{ij} is their normalised win ratio over the previous n games, not including the single previous game. Note that x^*_{ij} is not normalised, so can take on values $0, 0.5, 1$.

When we fit this model we can see the posterior distributions of these δ parameters, which correspond to the role the **single previous** game has in the win probability for the next game.

Effect of Previous Game



Effect of Previous n games, excluding most recent game



We can then again compare these estimated effects. Most of these δ have posterior means between $[-0.2, 0.2]$. $\delta = 0.2$ corresponds to winning the previous game leading to an increase in the win probability of 5%, while holding all other variables fixed.

It is maybe not immediate how to compare the effect sizes for β and δ here.

```r
## the beta parameters (previous n, not last game)
prob_positive(fit4)
```

```
 beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8] 
 0.98400  0.68625  0.99325  0.86875  0.94450  0.57075  0.99975  0.73100 
 beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] beta[15] beta[16] 
 0.52975  0.71400  0.98850  0.95750  0.76525  0.99500  0.87350  0.68200 
beta[17] beta[18] 
 0.72775  0.80575 
```

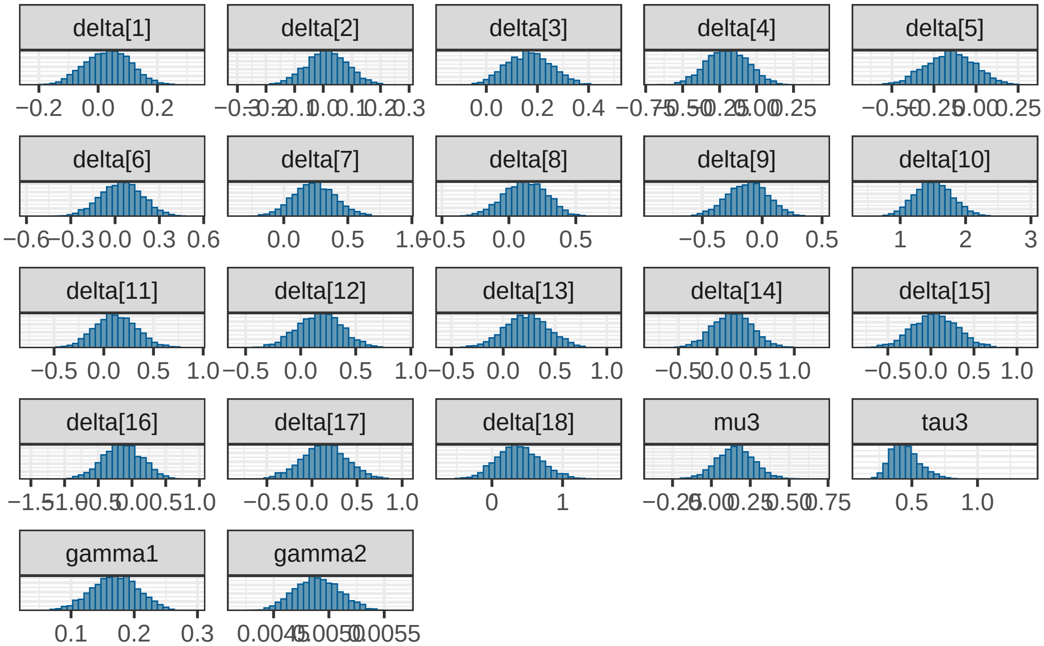
```r
## the delta parameters (single game)
prob_positive(fit4, param = "delta")
```

```
 delta[1]  delta[2]  delta[3]  delta[4]  delta[5]  delta[6]  delta[7]  delta[8] 
  0.57700   0.57375   0.95725   0.09000   0.12550   0.60475   0.81250   0.76925 
 delta[9] delta[10] delta[11] delta[12] delta[13] delta[14] delta[15] delta[16] 
  0.22650   1.00000   0.63225   0.78125   0.85200   0.65825   0.54800   0.29475 
delta[17] delta[18] 
  0.69875   0.88475 
```

This seems to indicate that the previous game on its own does not have as clear an effect, but that the streak of previous games shows stronger evidence for a winner effect.

We will also fit another model, using only the previous game as history, to this same dataset.

We can see the estimated effects from this model below.




```r
prob_positive(fit5$draws(), param = "delta")
```

```
 delta[1]  delta[2]  delta[3]  delta[4]  delta[5]  delta[6]  delta[7]  delta[8]
  0.66475   0.56650   0.97350   0.08625   0.15300   0.61875   0.93025   0.76625
 delta[9] delta[10] delta[11] delta[12] delta[13] delta[14] delta[15] delta[16]
  0.24250   1.00000   0.71225   0.79750   0.85650   0.77300   0.57100   0.30975
delta[17] delta[18]
  0.70850   0.88925
```

The estimated effect sizes for δ are very similar to those estimated under the previous model, indicating that including the longer history provides an improvement over just the previous game (particularly given the model comparison below).

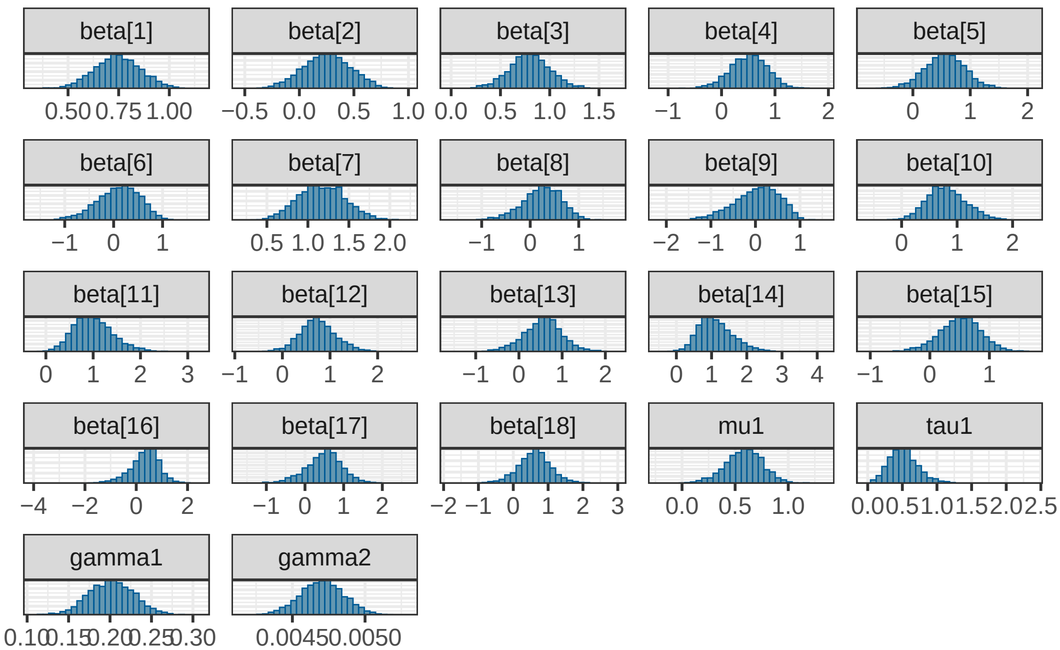
Consider all bullet games

Previously we considered only 60 second bullet games. We wish to consider if there is evidence for this effect across all bullet games.

```
# A tibble: 27,922 x 10
   White    Black Result UTCDate UTCTime White~1 Black~2 focal~3 focal~4 focal~5
   <chr>    <chr> <chr>  <chr>   <chr>   <chr>   <chr>     <dbl>   <dbl>   <dbl>
```

```
 1 valance  peng~ 0-1    2015.0~ 21:44:~ 1744    1500          0        1      1
 2 Georgian peng~ 0-1    2015.0~ 21:45:~ 1879    1838          0        1      1
 3 penguin~ Geor~ 1-0    2015.0~ 21:46:~ 1988    1870          1        1      1
 4 penguin~ kane~ 1-0    2015.0~ 21:53:~ 2056    1928          1        1      1
 5 kanepe   peng~ 0-1    2015.0~ 21:54:~ 1921    2108          0        1      1
 6 levtraru peng~ 0-1    2015.0~ 23:29:~ 2068    2144          0        1      1
 7 penguin~ levt~ 1-0    2015.0~ 23:31:~ 2190    2060          1        1      1
 8 penguin~ levt~ 1-0    2015.0~ 23:32:~ 2223    2053          1        1      1
 9 Kingscr~ peng~ 0-1    2015.0~ 23:34:~ 2347    2248          0        1      1
10 penguin~ King~ 0-1    2015.0~ 23:35:~ 2301    2334          1        0      0.9
# ... with 27,912 more rows, and abbreviated variable names 1: WhiteElo,
#   2: BlackElo, 3: focal_white, 4: focal_result, 5: focal_win_prop
```

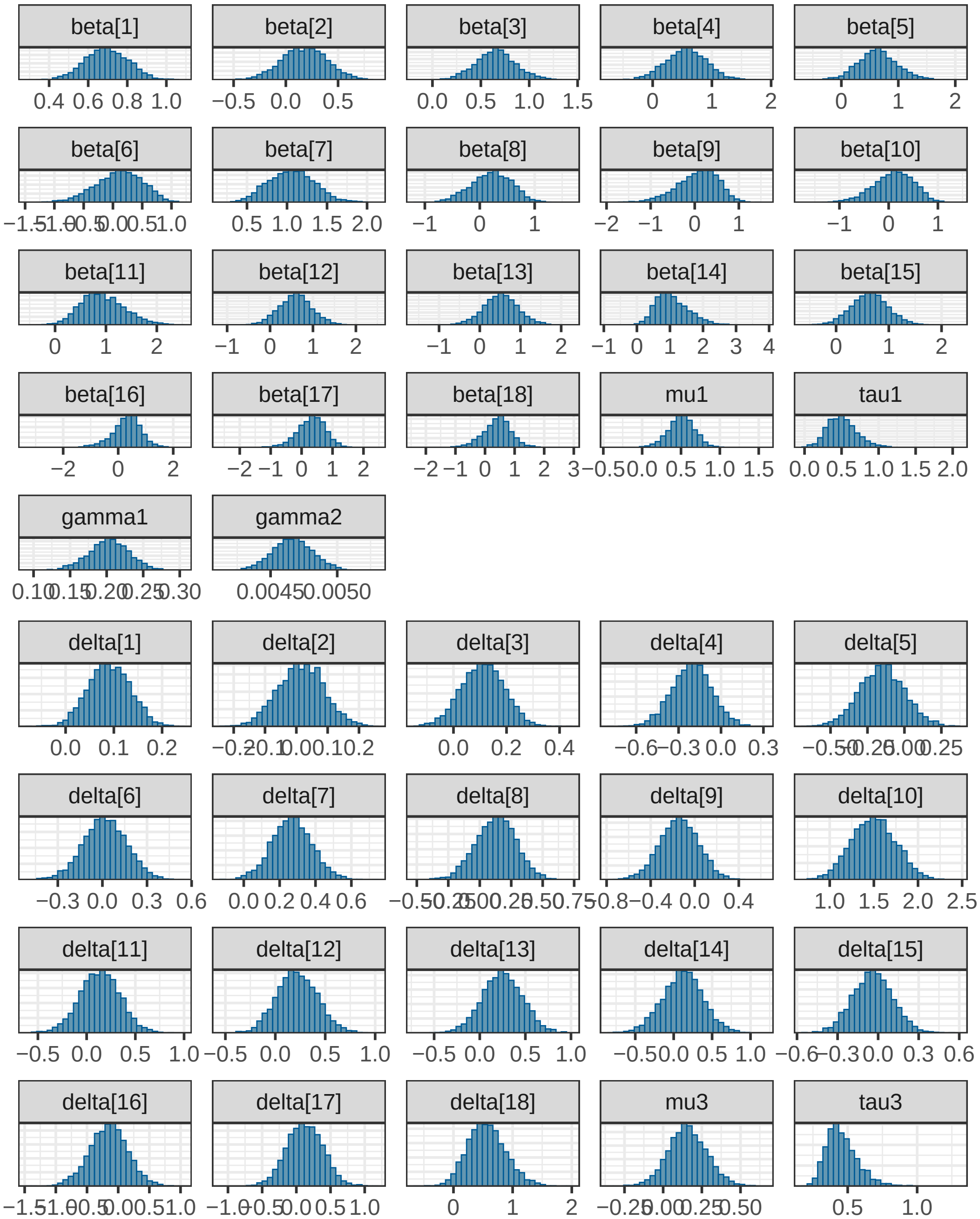
First Model all Bullet




```
 beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8] 
 1.00000  0.87025  1.00000  0.91925  0.93800  0.60675  1.00000  0.74300 
 beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] beta[15] beta[16] 
 0.55675  0.99075  0.99675  0.97450  0.88775  0.99575  0.91875  0.74875 
beta[17] beta[18] 
 0.84700  0.89775 
```

How do these estimates compare to those with only the 60 second games? The model with all bullet games has all posterior means greater than 0. There is not much of an overall change between the estimated mean effect sizes between the two datasets. Using all bullet games also leads to smaller posterior standard deviations for the winner effect size for all players.

Separating out History

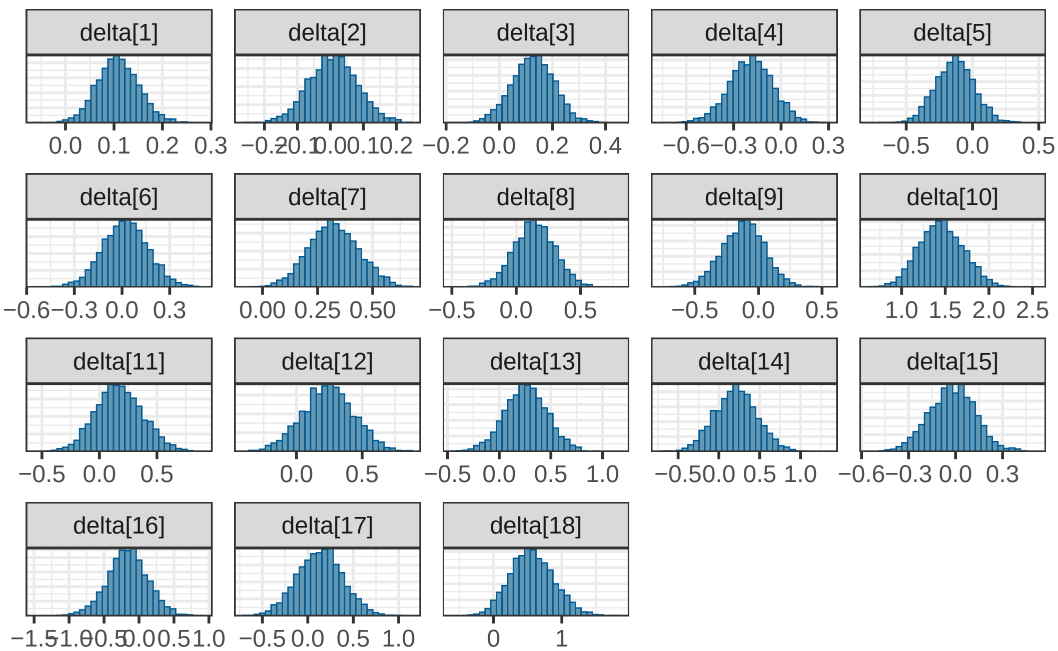


```
 beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]
 1.00000  0.78575  0.99900  0.94400  0.96675  0.62150  1.00000  0.71375
 beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] beta[15] beta[16]
 0.58925  0.61100  0.99075  0.93400  0.88825  0.99450  0.96425  0.75250
beta[17] beta[18]
 0.76500  0.80175
```

```
 delta[1]  delta[2]  delta[3]  delta[4]  delta[5]  delta[6]  delta[7]  delta[8]
  0.98350   0.57750   0.91200   0.06850   0.16300   0.54650   0.99175   0.78775
 delta[9] delta[10] delta[11] delta[12] delta[13] delta[14] delta[15] delta[16]
  0.21325   1.00000   0.74575   0.86225   0.87550   0.67800   0.39200   0.30800
delta[17] delta[18]
  0.67475   0.96950
```

Just previous game

We could also only use the previous game as a predictor, without using the rest of the history.




```
 delta[1]  delta[2]  delta[3]  delta[4]  delta[5]  delta[6]  delta[7]  delta[8]
  0.99475   0.54625   0.94200   0.08450   0.17725   0.56350   0.99850   0.81275
 delta[9] delta[10] delta[11] delta[12] delta[13] delta[14] delta[15] delta[16]
  0.23875   1.00000   0.79850   0.88275   0.89600   0.78175   0.44825   0.28575
delta[17] delta[18]
  0.70200   0.96550
```

Fitting using just the previous game as a predictor seems to give a reasonable model, with a similar interpretation as when we include the history also.

For this model, a posterior mean value for delta of 0.25 corresponds to 7% increase in the win probability having won the previous game compared to having lost it.

Not clear if this difference is due to excluding the win history or due to using a larger/different dataset.

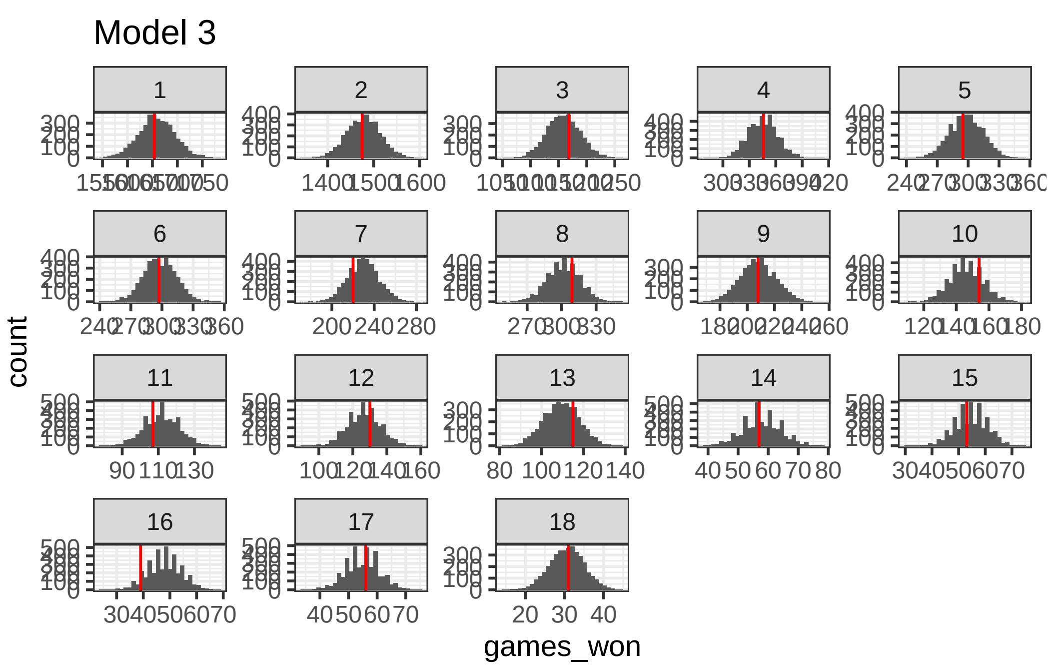
Model Checking

We now have a bunch of variations on the original model. We have fit these on both only 60 second games and all bullet games. We do some initial model checking, using only the smaller data initially (to speed up the computation).

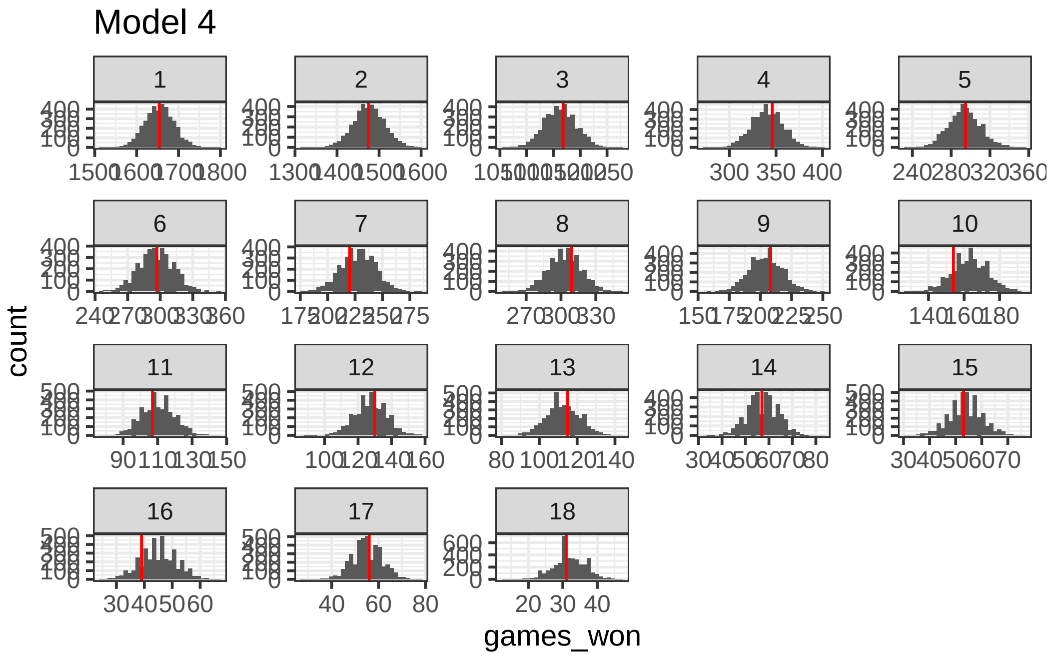
As a model check, we use each of the posterior draws to generate corresponding simulated outcomes, for each of the games, given the covariates. We can then use this to get a posterior predictive distribution for the number of games won by each of the players, which we show in the histograms below. Each red vertical line corresponds to the true number of games won by a player.

We see that the true values are very plausible under each of these models and provide one reasonable sanity check.

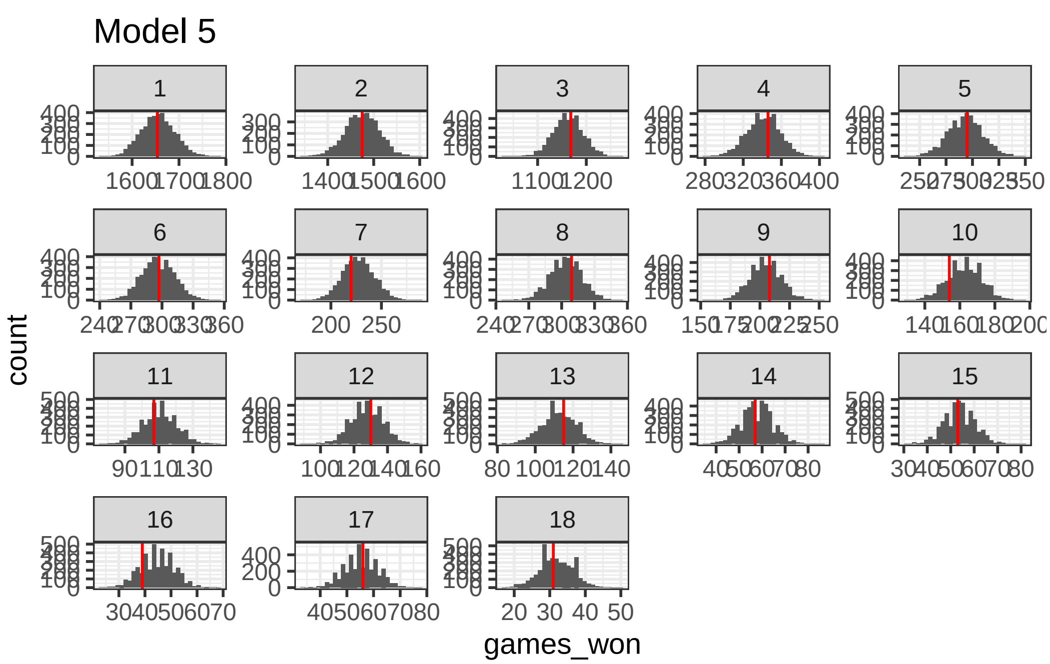
```
Warning: Dropping 'draws_df' class as required metadata was removed.
```



```
Warning: Dropping 'draws_df' class as required metadata was removed.
```




```
Warning: Dropping 'draws_df' class as required metadata was removed.
```



Model 5 appears to look as good as model 4, in terms of posterior predictive distribution for the output.

Model comparison

Finally, we also consider some model comparison between these 3 proposed models. Here we compute the leave one out cross validation. This allows us to compare across the models, with the general rule of thumb being a difference of more than 2 standard errors indicates a better model.

```
       elpd_diff se_diff
model2   0.0       0.0
model3 -10.0       5.1
model1 -19.9       6.0
```

Here we see that the second model, including both the previous game and the n games before that, is an improvement over the other two models. The model which considers only the single previous game is closer.

Further Ideas

- What about some sort of mixture of the previous game and the others before that?
- Some sort of decaying effect?
- Is there an interaction between the winner effect and the ability of the player, in terms of their ELO score?