The Human Limit

Modeling World Record Probability in Super Mario Bros. Speed running

Duc Huy Nguyen (Thomas)
Theo Schouten

STAT 333 Final Project
Deliverable 2: Statistical Modeling Analysis

November 26, 2025

1. Introduction and Historical Context

Super Mario Bros. speed running has evolved dramatically since the first recorded world record (WR) of 325 seconds set in June 2002. The current WR, set by Niftski on October 22, 2025, stands at 294.448 seconds, only 0.416 seconds away from the Tool-Assisted Speed Run, a speed run that slows the game down to play perfectly, of 294.032 seconds. This represents approximately 11 frames at 60 FPS that separate human performance from theoretical perfection. As shown in Figure 1, the rate of WR improvement has slowed considerably as players approach this asymptotic limit; this is the Human Limit.

The reason this is possible is because of frame rules. Super Mario Bros. operates on a system where the game only checks for level completion every 21 frames (0.35 seconds). This means improvements must come in discrete 21-frame increments in most levels, making fractional-second improvements extraordinarily difficult. However, it also means that the world record only needs to match the frame rule of the Tool Assisted Speed run and not the frame. Still, the Any% (completion speed run) category requires flawless execution across eight major checkpoints: World 1-1, 1-2, 4-1, 4-2, 8-1, 8-2, 8-3, and 8-4. Each subsequent stage has differing success rates due to difficulty, mounting pressure, and other factors. Due to the conditional nature of progression (one cannot reach 8-4 without completing all prior stages), this makes the probability of a World Record astronomically small.

2. Data and Modeling Approach

Our dataset consists of 1,353 attempts by current WR holder Niftski, with binary outcomes (1 = success, 0 = failure) recorded for each of the eight major checkpoints. The empirical success rates decline sharply across stages: 61.3% for 1-1, 44.0% for 1-2, 35.6% for 4-1, 4.7% for 4-2, 2.2% for 8-1, 0.8% for 8-2, 0.7% for 8-3, and only 0.07% for 8-4 (1 success in 1,353 attempts).

We model the probability of achieving a world record (completing W8_4 successfully) as a function of performance across all prior checkpoints. Our primary predictor is `success_rate`, defined as:

$$\text{success\_rate} = \frac{W1\_1 + W1\_2 + W4\_1 + W4\_2 + W8\_1 + W8\_2 + W8\_3}{7}$$

This composite measure captures overall consistency and skill level leading up to the final stage. We explore two modeling approaches to handle the extreme rarity of the outcome event (only 1 success in 1,353 trials): a zero-inflated binomial model (ZIP) and Firth's penalized logistic regression. Our third option for a model is a Bayesian model, which hasn't been fully built yet.

3. Model 1: Zero-Inflated Binomial (Unsuccessful)

Given the extreme prevalence of zeros in W8_4, we initially attempted a zero-inflated binomial (ZIB) model using `glmmTMB`. The model specification was:

$$\begin{aligned}
\text{Binomial component:} \quad & \text{logit}(p_i) = \beta_0 + \beta_1(\text{success_rate}_i) \\
\text{Zero-inflation component:} \quad & \text{logit}(\psi_i) = \gamma_0 + \gamma_1(\text{success_rate}_i)
\end{aligned}$$

where p_i represents the probability of success conditional on not being a structural zero, and ψ_i represents the probability of being a structural zero.

Results: The fitted model produced `NaN` (Not a Number) values for all parameter estimates and standard errors. This failure occurs because with only a single success event, the maximum likelihood estimation cannot converge. The zero-inflation framework attempts to distinguish between "structural" zeros (impossible successes) and "sampling" zeros (possible but unlucky attempts), but our data provides insufficient information to estimate these components separately. The model output showed:

```
Conditional model:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)          NaN        NaN     NaN      NaN
success_rate         NaN        NaN     NaN      NaN

Zero-inflation model:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)          NaN        NaN     NaN      NaN
success_rate         NaN        NaN     NaN      NaN
```

This demonstrates that zero-inflated models, while theoretically appropriate for excess-zero data, are impractical when the outcome is so rare that maximum likelihood estimation breaks down.

4. Model 2: Firth's Penalized Logistic Regression (Successful)

To address the separation problem inherent in our rare-event data, we employed Firth's penalized-likelihood logistic regression using the `logistf` package. This approach adds a penalty term, which prevents infinite estimates and provides finite, stable parameter estimates even with perfect or near-perfect separation.

The model takes the form:

$$\text{logit}(P(W8\_4 = 1)) = \beta_0 + \beta_1(\text{success\_rate})$$

with penalized log-likelihood:

$$\ell_{\text{pen}}(\beta) = \ell(\beta) + \frac{1}{2}\log|\mathcal{I}(\beta)|$$

where $\mathcal{I}(\beta)$ is the Fisher information matrix.

Results: The Firth model successfully converged and yielded interpretable results:

- **Intercept** (β_0): -11.67 (p = 0.0013), indicating extremely low baseline probability
- **success_rate** (β_1): 9.82 (p = 0.0436), showing strong positive association

The significant positive coefficient for `success_rate` confirms that a World Record can be predicted based on position in the run (earlier world successes). For Niftski's single successful run (where all prior stages were completed, giving success_rate = 1.0), the model predicts:

$$P(\text{W8\_4} = 1 \mid \text{success\_rate} = 1.0) = \frac{1}{1 + e^{-(-11.67 + 9.82 \times 1.0)}} \approx 0.135$$

This suggests that the probability of a World Record is 13.5% when Niftski reaches 8-4 with perfect prior-stage completion. A Hosmer-Lemeshow goodness-of-fit test yielded p = 0.998, which indicates excellent model fit to the observed data.

The use of the composite `success_rate` predictor rather than individual stage indicators serves multiple purposes: it reduces model complexity and avoids multicollinearity, as later stages are conditional on earlier ones; it provides a single, interpretable measure of "momentum" or skill level; and it stabilizes estimation by aggregating information across stages rather than attempting to estimate seven separate coefficients with minimal variation in the outcome.

5. Future Work: Bayesian Hierarchical Model

While the Firth model provides satisfactory point estimates, a fully Bayesian approach would better quantify uncertainty in this low-data regime. We plan to implement a hierarchical Beta-Binomial model that treats each stage's success probability as a random effect drawn from a common distribution, allowing us to model the conditional progression through stages explicitly. This approach will provide posterior distributions for the probability of achieving a WR in the next N attempts via the formula:

$$P(\text{WR in } N \text{ attempts}) = 1 - (1 - \hat{p})^N$$

where $\hat{p}$ is the posterior mean probability per attempt. This Bayesian framework is currently under development and will be featured in our final deliverable.

6. Conclusion

We have demonstrated that Firth's penalized logistic regression successfully models the probability of Nifski achieving a Super Mario Bros. Any% WR. The model confirms that previous attempts across all prior stages are a strong predictor of final-stage success (WR), with approximately 13.5% success probability conditional on perfect prior execution. As the speed running community continues to approach the theoretical human limit, such models become increasingly valuable for estimating timelines for future milestones and, more importantly, adding an interactive aspect for the community to get excited about in the monotonous grind of a World Record.