Faster identification of faster Formula 1 drivers via time-rank duality

John Fry*, Tom Brighton, Silvio Fanzon
Centre for Mathematical Sciences, School of Natural Sciences, University of Hull, Hull, HU6 7RX, UK

1. Introduction

Modelling Formula (1) races is an interesting econometric problem (Bell et al., 2016; van Kesteren and Bergkamp, 2023) of significant wider interest (Maurya, 2021). It is of interest to separate out driver-level and car-level effects. Previously, such an analysis has only been possible over longer time periods (Bell et al., 2016; Eichenberger and Stadelmann, 2009; van Kesteren and Bergkamp, 2023). Here, we present a solution that requires only one season of previous data.

Formula (1) races are most easily modelled assuming car finishing times are independently and exponentially distributed random variables. Under this assumption race-winning probabilities can be written down in closed form. This tractability also enables relatively easy model calibration via bookmakers’ odds. However, this approach is at odds with much of the publicly-available race data. Final race-finishing times are typically unavailable as lapped cars do not typically finish the full race distance. In contrast, the most convenient way of modelling publicly-available race data is regression modelling of the final finishing position (Eichenberger and Stadelmann, 2009). Thus, in this paper, we combine both modes of analysis — an approach we term time-rank duality.

The layout of this paper is as follows. Section 2 outlines a probabilistic approach to modelling race-finishing times and model calibration via bookmakers’ odds. Section 3 establishes theoretical duality between this probabilistic approach and regression modelling of the final rank. Firstly, we show that a regression model for ranks can be used to estimate race-winning probabilities and then to an equivalent exponential-distribution parameterisation using the method in Section 2. Secondly, we show that, under the simplifying assumption of homoscedasticity, regression parameters can be reverse-engineered from a set of race-winning probabilities e.g. those corresponding to a given exponential-distribution parameterisation or a set of bookmakers’ odds. Section 4 discusses empirical regression modelling of historical results. Section 5 combines both approaches to enable quicker identification of individual driver-level effects. Section 6 concludes. A mathematical appendix is contained at the end of the paper.

2. Probabilistic approach to modelling finishing times

Models based around the exponential distribution are amongst the most convenient ways to model Formula (1) races. This is due to its tractability alongside its usage in classical applied probability models. A related formulation based on the Weibull distribution is explored in Appendix. Whilst its non-constant hazard function may be more physically realistic, the Weibull distribution may be more cumbersome in applications due to its additional shape parameter.

Suppose, for the sake of simplicity, that a race consists of $n$ cars whose finishing times $T_1, T_2, ..., T_n$ are independent exponential distributions with parameters $\lambda_1, \lambda_2, ..., \lambda_n$. Independence is a common simplifying assumption in sports models (Scarf et al., 2019) but may be difficult to justify empirically. A standard result in probability theory (Grimmett and Stirzaker, 2020) gives:

Proposition 1.

i. If $T_1, T_2, ..., T_n$ are independent and exponentially distributed with parameters $\lambda_1, \lambda_2, ..., \lambda_n$, then
The minimisation in (1) can be done numerically. Results of the procedure applied to bookmakers’ data are shown in Table 1. The R code

\[
\min \{T_1, T_2, \ldots, T_n\} = \exp \left( \sum_{i=1}^{n} \lambda_i \right)
\]

\[
\text{if } X \text{ and } Y \text{ are independent exponential distributions with parameters } \lambda_X \text{ and } \lambda_Y \text{ then}
\]

\[
Pr(X \leq Y) = \frac{\lambda_X}{\lambda_X + \lambda_Y}
\]

\[
\text{Consider the Formula (1) race with independent and exponentially distributed finishing times as outlined above. Then}
\]

\[
Pr(\text{Car } i \text{ wins}) = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}
\]

Proposition 1 shows that given a sequence of win probabilities \(p_1, p_2, \ldots, p_n\), calculated e.g. from bookmakers’ odds, we can estimate the parameters \(\lambda_i\). This can be done by minimising the Residual Sum of Squares (RSS):

\[
\text{RSS} := \sum_{i=1}^{n} \left( \frac{\lambda_i}{\lambda_1 + \ldots + \lambda_n} - p_i \right)^2.
\]

The minimisation in (1) can be done numerically. Results of the procedure applied to bookmakers’ data are shown in Table 1. The R code and data to reproduce these results is openly available on GitHub. In Table 1 odds can be converted to probabilities as follows. The win probability corresponding to odds of 25/1 for Lewis Hamilton victory can be calculated via

\[
1 - \frac{1}{p} = 25; \quad p = \frac{1}{26}
\]

Win probabilities for the remaining drivers are calculated similarly, and then renormalised (Strumbelj, 2014) so that they sum to 1. These renormalised win probabilities are given in the fourth column of Table 1. Estimated \(\lambda\) values from the minimisation in (1) are in the fifth column.

### 3. Econometric modelling of the final race ranking

Empirical Formula (1) data are most commonly listed in terms of the rank rather than the strict finishing times. The analysis of historical race data is therefore most easily accomplished by regression modelling of the final rank obtained (Eichenberger and Stadelmann, 2009). This implicitly assumes a Gaussian model for sporting outcomes (Scarf et al., 2019).

Consider two related problems. Firstly, suppose that there are \(n\) cars in the race and the final ranking \(r_i\) of car \(i\) can be approximated by a normal distribution: \(r_i \sim N(\mu_i, \sigma_i^2)\). The approximate probability that car \(i\) wins the race is given by

\[
p_i = Pr(r_i \leq 1.5) = \Phi \left( \frac{1.5 - \mu_i}{\sigma_i} \right),
\]

where \(\Phi(\cdot)\) denotes the standard normal CDF. Secondly, suppose we are given a sequence of win probabilities \(p_1, p_2, \ldots, p_n\) for Cars 1, 2, \ldots, \(n\). Under the simplifying assumption of \(\sigma_i^2 = \sigma^2\), equivalent to the classical normal linear regression model (Fry and Burke, 2022), from Eq. (2) set

\[
\Phi \left( \frac{1.5 - \mu_i}{\sigma} \right) = p_i; \quad \mu_i = 1.5 - \sigma \Phi^{-1}(p_i).
\]

Since the sum of the ranks is equal to \(\frac{n(n+1)}{2}\) summing Eq. (3) over \(i\) gives

\[
\frac{n(n+1)}{2} = 1.5n - \sigma \sum_{i=1}^{n} \Phi^{-1}(p_i); \quad \sigma = \frac{n - \frac{n}{2}}{\sum_{i=1}^{n} \Phi^{-1}(p_i)}.
\]

Combining Eqs. (3)–(4) therefore gives the estimated \(\mu_i\) values corresponding to the given win probabilities \(p_i\). Table 2 applies this approach to estimate a set of \(\mu_i\) and \(\hat{\sigma}^2\) regression parameters for the bookmakers’ data in Table 1.

### 4. Regression modelling of historical results

In this section we calibrate the model to historical results (observed race rankings) from the 2022 season which was the last fully-completed season at the time of writing. This follows a similar approach to modelling historical results in Fry et al. (2021). Following Eichenberger and Stadelmann (2009) we regress the finishing position against the dummy variables corresponding to each of the constructors. We then use stepwise regression (Fry and Burke, 2022) to automatically choose the best model. We constrain all models fitted to include a dummy variable indicating the teams’ second (less-favoured) driver. Forwards and stepwise regression choose the same model indicated below in

\[\text{R codes and data files are made available on the third authors’ GitHub page.}\]
Table 3. In contrast, backward selection suggests a more complex
model. However, an F-test, not reported, is non-significant suggest-
ing the simpler model in Table 3 should suffice. Negative and signifi-
cant parameters in Table 3 indicate more successful constructors with lower
expected final finishing positions.

5. Distilling driver-level effects

From the regression output in Table 3 a 95% confidence interval for the
second driver term is

Second driver confidence interval = (−0.581, 1.013).

The upper value of 1.013 produced in (5) is physically meaningful.
Suppose race orderings are completely determined by the level of the
car. In this case positions 1–2 would be occupied by the best car,
positions 3–4 by drivers of the second best car, positions 5–6 by drivers
of the third best car etc. A difference in the average ranking greater
than 1 indicates that the quality of the leading driver is sufficient to be
able to outperform the next best car on the grid.

Thus combining Eq. (5) with implied regression parameters in Ta-
ble 2 a difference between two drivers of the same team bigger than
1.013 implies an extraordinary level of performance beyond the qual-
ity of the car. Comparing drivers in this way suggests two drivers
Max Verstappen (Red Bull) and Fernando Alonso (Aston Martin) out-
perform their respective cars. Past academic research has previously
highlighted Verstappen’s level of performance as historically significant
(van Kesteren and Bergkamp, 2023).

6. Conclusions

It is interesting to separate out driver-level and car-level effects in
Formula (1) racing. Previously, such an analysis has only been possible
over longer time periods (Bell et al., 2016; Eichenberger and Stadel-
mann, 2009; van Kesteren and Bergkamp, 2023). Here, we present a
solution that requires only one season’s worth of previous data. We
combine a probabilistic approach based on the exponential distribution
with econometric modelling of the ranks (Eichenberger and Stadel-
mann, 2009). Both approaches enable the race-winning probabilities
to be exactly solved analytically. Equating race-winning probabilities
means that both approaches can be seen as equivalent to each other.
Results suggest that of the current crop of drivers Max Verstappen
and Fernando Alonso out-perform the level of the car that they drive.
Results match previous suggestions that Verstappen’s performance level
is historically significant (van Kesteren and Bergkamp, 2023). Future
work will adjust the above models to account for cars that fail to finish
races. Substantial interest in the analytical modelling of sports remains
(Singh et al., 2023).

Data availability

Data will be made available on request.

Acknowledgement

The authors would like to acknowledge helpful and supportive
comments from an anonymous reviewer. The usual disclaimer applies.

Appendix. Mathematical proofs

In Proposition 2 we consider race times to be independent Weibull
distributions with common shape parameter $k$. This is a small technical
extension of Proposition 1, where finishing times are exponential. We
present the proof for Proposition 2 below, noting that Proposition 1 is
the special case of $k = 1$ in Proposition 2.

Proposition 2.

i. If $T_1, \ldots, T_n$ are independent and Weibull distributed with parameters
$(\lambda_1, k), \ldots, (\lambda_n, k)$ then

$$\min\{T_1, T_2, \ldots, T_n\} \sim \text{Weibull}\left(\sum_{i=1}^{n} \lambda_i, k\right).$$

ii. If $X \sim \text{Weibull}(\lambda_X, k)$ and $Y \sim \text{Weibull}(\lambda_Y, k)$ and $X$ and $Y$
are independent then

$$P(X \leq Y) = \frac{\lambda_X}{\lambda_X + \lambda_Y}.$$

iii. Consider the Formula (1) race with independent and Weibull dis-
tributed finishing times as outlined above. Then

$$P(\text{Car } j \text{ wins}) = \frac{\lambda_j}{\sum_{i=1}^{n} \lambda_i}.$$

Proof of Proposition 2.

i. $P(T_i \geq x) = e^{-\lambda_i x^k}$. Since all the $T_i$ are independent
$P(T_1 \geq x, \ldots, T_n \geq x) = e^{-(\lambda_1 + \cdots + \lambda_n)x^k}$.

This gives

$$P(\text{min}\{T_1, \ldots, T_n\} \leq x) = 1 - e^{-\sum_{i=1}^{n} \lambda_i x^k}.$$

ii. Since $f_x(x) = k\lambda X^{k-1}e^{-\lambda x^k}$ and $f_y(y) = k\lambda Y^{k-1}e^{-\lambda y^k}$
$P(X \leq Y) = \int_{0}^{\infty} \int_{0}^{\lambda Y^k} k^2 \lambda X^{k-1}y^{k-1}e^{-\lambda x^k}e^{-\lambda y^k} \, dx \, dy$

$$= \int_{0}^{\infty} \int_{0}^{\lambda Y^k} k^2 \lambda X^{k-1}y^{k-1}e^{-\lambda x^k} \, dy \, dx$$

$$= \int_{0}^{\infty} \lambda Y^{k-1} e^{-\lambda x^k} \, dY \, dx$$

$$= \int_{0}^{\infty} \lambda Y^{k-1} e^{-\lambda x^k} \, dY \, dx - \int_{0}^{\infty} \lambda Y^{k-1} e^{-\lambda x^k} \, dx \, dY$$

$$= 1 - \frac{\lambda Y}{\lambda X + \lambda Y}.$$

iii. For the sake of argument suppose $j = 1$. Then

$$P(\text{Car } 1 \text{ wins}) = P(T_1 \leq \min\{T_2, T_3, \ldots, T_n\}).$$

Now $T_1$ and $\min\{T_2, T_3, \ldots, T_n\}$ are independent with

$T_1 \sim \text{Weibull}(\lambda_1, k)$; $\min\{T_2, T_3, \ldots, T_n\} \sim \text{Weibull}\left(\sum_{i=2}^{n} \lambda_i, k\right)$.

Hence the result follows from part ii. \qed

References


Eichenberger, R., Stadelmann, D., 2009. Who is the best Formula 1 driver? An economic

Press, Maidenhead.


