The Classical Model for Race Rankings

Stuart Coles

7 April, 2022

Markets for Ski Racing

- Last year I gave a talk on modelling ski races;
- I focused then on the margin of the winning race time;

Market Prices: Winning Time Margin

Vantaggio del vincitore		
0.01 - 0.10 sec. **3.50**	0.11 - 0.20 sec. **3.75**	0.21 - 0.30 sec. **4.50**
0.31 - 0.50 sec. **5.00**	0.51 - 0.75 sec. **8.00**	0.76 - 1.00 sec. **13.00**
1.01 - 1.50 sec. **34.00**	1.51 sec. o più **101.00**	ex aequo **34.00**

Markets for Ski Racing

- Last year I gave a talk on modelling ski races;
- I focused then on the margin of the winning race time;
- The other available markets related to race rankings;

Market Prices: Winner

Vincente		
B. Feuz **3.50**	M. Mayer **4.50**	D. Paris **5.00**
V. Kriechmayr **11.00**	M. Franz **11.00**	C. Innerhofer **11.00**
J. Clarey **17.00**	All other skiers not listed **26.00**	A. Sander **34.00**
K. Jansrud **34.00**	R. Baumann **34.00**	C. Janka **41.00**
M. Odermatt **41.00**	J. Ferstl **51.00**	T. Ganong **67.00**
M. Bailet **81.00**	O. Striedinger **81.00**	N. Hintermann **101.00**
N. Allegre **101.00**	M. Muzaton **151.00**	J. Goldberg **151.00**
D. Schwaiger **151.00**		

Market Prices: Podium Place

Beat Feuz (SUI) arriverà fra i primi 3? (l'atleta deve gareggiare)	
Sì **1.60**	No **2.20**

Dominik Paris (ITA) arriverà fra i primi 3? (l'atleta deve gareggiare)	
Sì **1.90**	No **1.80**

Vincent Kriechmayr (AUT) arriverà fra i primi 3? (l'atleta deve gareggiare)	
Sì **4.25**	No **1.18**

Market Prices: Head-to-Head

Testa a testa B. Feuz/M. Mayer	
B. Feuz **1.60**	M. Mayer **2.20**

Testa a testa C. Innerhofer/M. Franz	
C. Innerhofer **1.75**	M. Franz **1.95**

Head-to-Head A. Sander/M. Odermatt	
A. Sander **1.67**	M. Odermatt **2.07**

Markets for Ski Racing

- Last year I gave a talk on modelling ski races;
- I focused then on the margin of the winning race time;
- The other available markets related to race rankings;
- This is equally relevant to any race market - horses, F1, cycling etc.

Statistical Setup

Objective: given a set of competitors $C_1, C_2, \ldots, C_N$, calculate the probabilities that:

1. C_k is the winner;
2. C_k finishes in the first 3;
3. C_{k_1} beats C_{k_2}.

Additionally, the probability that:

4. The first three places go to $C_{k_1}, C_{k_2}, C_{k_3}$ (either in that order or an arbitrary order).

Available information

1. Rankings from previous races;
2. Possibly, but not necessarily, race times from previous races;
3. Covariate information about competitors;
4. Covariate information about race course and conditions.

Race Times

- The race times themselves are of no interest, other than in determining the rankings;

Race Times

- The race times themselves are of no interest, other than in determining the rankings;
- They may or may not be available from previous events;

Race Times

- The race times themselves are of no interest, other than in determining the rankings;
- They may or may not be available from previous events;
- If available, we could model the race times of the individual competitors, but this might be complicated by:

1. dependence on track, weather conditions etc etc.
2. dependence between the competitors.

Race Times

- The race times themselves are of no interest, other than in determining the rankings;
- They may or may not be available from previous events;
- If available, we could model the race times of the individual competitors, but this might be complicated by:

1. dependence on track, weather conditions etc etc.
2. dependence between the competitors.

- Therefore ignore race times and model just the ranks of past and future races.

A Simple Model for Race Times

Assume (for the moment) that the race times for competitors $C_1, \ldots, C_N$ are independent and exponentially distributed:

$$X_k \sim \mathrm{Exp}(\lambda_k), \quad k = 1, \ldots, N$$

An Aside: Min-Stability of the Exponential Distribution

If $X \sim \mathrm{Exp}(\lambda_1)$, $Y \sim \mathrm{Exp}(\lambda_2)$ and X and Y are independent, then

$$Z = \min(X, Y) \sim \mathrm{Exp}(\lambda_1 + \lambda_2)$$

An Aside: Min-Stability of the Exponential Distribution

If $X \sim \mathrm{Exp}(\lambda_1)$, $Y \sim \mathrm{Exp}(\lambda_2)$ and X and Y are independent, then

$$Z = \min(X, Y) \sim \mathrm{Exp}(\lambda_1 + \lambda_2)$$

Proof:

$$\begin{aligned} P(\min(X, Y) > z) = P(X > z, Y > z) &= \exp(-\lambda_1 z) \exp(-\lambda_2 z) \\ &= \exp(-(\lambda_1 + \lambda_2) z) \end{aligned}$$

Min-Stability of the Exponential Distribution

This result generalizes in the obvious way, so that if $X_j \sim \mathrm{Exp}(\lambda_j)$ for $j = 1, \ldots, N$, with $X_1, \ldots, X_N$ independent,

$$\min(X_1, \ldots, X_N) \sim \mathrm{Exp}\left(\sum_{j=1}^{N} \lambda_j\right)$$

Race Winner Probabilities

Fundamental result (race winner model):

$$P(R_1 = C_k) = \frac{\lambda_k}{\sum_{j=1}^{N} \lambda_j}$$

Proof of Fundamental Result

$$P(R_1 = C_k) = \int_{x=0}^{\infty} P(X_k = x) P(\min_{j \neq k}\{X_j\} > x) dx$$

Proof of Fundamental Result

$$P(R_1 = C_k) = \int_{x=0}^{\infty} P(X_k = x) P(\min_{j \neq k}\{X_j\} > x) dx$$

But, by min-stability:

$$\min_{j \neq k}\{X_j\} \sim \text{Exp}\left(\sum_{j \neq k} \lambda_j\right)$$

Proof of Fundamental Result

$$
\begin{aligned}
P(R_1 = C_k) &= \int_{x=0}^{\infty} \left\{ \lambda_k \exp(-\lambda_k x) \exp\left(-\sum_{j \neq k} \lambda_j x \right) \right\} dx \\
&= \int_{x=0}^{\infty} \left\{ \lambda_k \exp\left(-\sum_{j=1}^{N} \lambda_j x \right) \right\} dx \\
&= \left[-\frac{\lambda_k}{\sum_{j=1}^{N} \lambda_j} \exp\left(-\sum_{j=1}^{N} \lambda_j x \right) \right]_{x=0}^{\infty} \\
&= \frac{\lambda_k}{\sum_{j=1}^{N} \lambda_j}
\end{aligned}
$$

Invariance Property

- This is a nice result, linking the exponential model for race times to a multinomial model for rankings.

Invariance Property

- This is a nice result, linking the exponential model for race times to a multinomial model for rankings.
- Though seemingly limited by the exponential assumption, it actually has much wider applicability...

Invariance Property

- Suppose that h is a monotonic increasing function.

Invariance Property

- Suppose that h is a monotonic increasing function.
- Then, $h(X) < h(Y)$ if and only if $X < Y$.

Invariance Property

- Suppose that h is a monotonic increasing function.
- Then, $h(X) < h(Y)$ if and only if $X < Y$.
- It follows that the fundamental result

$$P(R_1 = C_k) = \frac{\lambda_k}{\sum_{j=1}^{N} \lambda_j},$$

  and any other result that concerns just the ranks, also holds true if race times are a monotonic transformation of exponentially distributed variables.

Invariance Property

- For example, it holds true if race times are log-exponential...

Invariance Property

- For example, it holds true if race times are log-exponential...
- If $X \sim \mathrm{Exp}(\lambda)$ then $Z = \log X$ has an extreme value (Gumbel) distribution (for minima).

Invariance Property

- For example, it holds true if race times are log-exponential...
- If $X \sim \mathrm{Exp}(\lambda)$ then $Z = \log X$ has an extreme value (Gumbel) distribution (for minima).
- This might have some justification as a model for the fastest race time, but is no easier to justify than exponential for individual race times.

Invariance Property

- For example, it holds true if race times are log-exponential...
- If $X \sim \text{Exp}(\lambda)$ then $Z = \log X$ has an extreme value (Gumbel) distribution (for minima).
- This might have some justification as a model for the fastest race time, but is no easier to justify than exponential for individual race times.
- Nonetheless, in the context of horse racing, Henery (1984) writes

  the exponential is intrinsically suspect as a model for race times, but the logarithm of an exponential turns out to be a very fair representation

Other Probabilities

Other probabilities can also be calculated based on the (possibly transformed) exponential setup . . .

Head-to-Head

$$P(C_{k_1} \text{ beats } C_{k_2}) = \frac{\lambda_{k_1}}{\lambda_{k_1} + \lambda_{k_2}}$$

Podium Places

$$P(R_1 = C_{k_1}, R_2 = C_{k_2}, R_3 = C_{k_3}) = \frac{\lambda_{k_1}}{\sum_{S_1} \lambda_k} \times \frac{\lambda_{k_2}}{\sum_{S_2} \lambda_k} \times \frac{\lambda_{k_3}}{\sum_{S_3} \lambda_k}$$

where

$S_1 = \{1, \dots, N\}$,

$S_2 = S_1 \setminus k_1$,

$S_3 = S_1 \setminus \{k_1, k_2\}$

Log-Likelihood Based on Ranks

- This final probability also gives us a means to estimate the λ_k using the log-likelihood for a history of observed ranks:

$$\ell = \sum_{\text{Races}} \left\{ \sum_{j=1}^{3} \log \lambda_{k_j} - \log(\sum_{S_1} \lambda_k) - \log(\sum_{S_2} \lambda_k) - \log(\sum_{S_3} \lambda_k) \right\}$$

Log-Likelihood Based on Ranks

- This final probability also gives us a means to estimate the λ_k using the log-likelihood for a history of observed ranks:

$$\ell = \sum_{\text{Races}} \left\{ \sum_{j=1}^{3} \log \lambda_{k_j} - \log(\sum_{S_1} \lambda_k) - \log(\sum_{S_2} \lambda_k) - \log(\sum_{S_3} \lambda_k) \right\}$$

- More generally, based on the top M race positions

$$\ell = \sum_{\text{Races}} \left\{ \sum_{j=1}^{M} \log \lambda_{k_j} - \sum_{m=1}^{M} \left[\log(\sum_{S_m} \lambda_k) \right] \right\}$$

Strategy

- Use this rank model, treating λ_k as a measure of the 'strength' of competitor k.
- The bigger the value of λ_k, the stronger competitor C_k.
- Model λ_k as a function of covariates relating to competitor and race.
- Could presumably use Machine Learning techniques with this framework to optimize variable selection.

Connections

Related to

- Bradley-Terry Model
- Plackett-Luce Model
- Discrete Choice Models (McFadden) in socio-economics.

Connections

- Other models can also be constructed this way using distributions other than exponential - e.g. Normal.
- But, closed form expressions are generally unavailable and numerical methods or approximations are required.
- The case with just 2 competitors can easily be studied analytically though. . .

Gaussian Model - 2 Competitors

- Suppose race times for C_1 and C_2 are $N(\mu_1, \mu_1^2)$ and $N(\mu_2, \mu_2^2)$.

Gaussian Model - 2 Competitors

- Suppose race times for C_1 and C_2 are $N(\mu_1, \mu_1^2)$ and $N(\mu_2, \mu_2^2)$.
- It's easy to show that

$$P(R_1 = C_1) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\mu_1^2 + \mu_2^2}}\right)$$

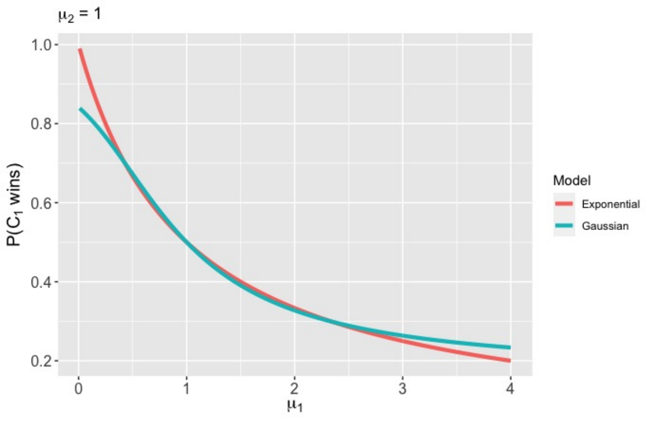
Gaussian Model - 2 Competitors

- Suppose race times for C_1 and C_2 are $N(\mu_1, \mu_1^2)$ and $N(\mu_2, \mu_2^2)$.
- It's easy to show that

$$P(R_1 = C_1) = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\mu_1^2 + \mu_2^2}}\right)$$

- With, for example, $\mu_2 = 1$, a comparison of win probabilities with the Exponential model is as follows...

Gaussian model - 2 Competitors



Parametric Models for λ

Basic:

$$\lambda = \exp(\beta' x)$$

where x is a vector of competitor-specific covariates (e.g. current world ranking; home country indicator; age; ...)

Parametric Models for λ

Enhanced:

$$\lambda = \exp(r\beta' x)$$

where $r = \exp(\alpha' y)$ and y is a vector of race-specific covariates.

$r = 1$: ‘standard’

$r = 0$: ‘all competitors equivalent’.

$0 < r < 1$: ‘reduced impact of competitor-based covariates.’

$r > 1$: ‘increased impact of competitor-based covariates.’

Illustrative Model Fit

Crude example of fit to Men's Downhill, using $M = 5$ positions and the basic model.

Covariates (all on log scale) are:

1. Downhill world ranking (DHR).
2. Super Giant Slalom world ranking (SGR).
3. Giant Slalom world ranking (GSR).
4. Slalom world ranking (SR)
5. Rolling average of time difference to winner (TDW).
6. Rolling average of DNF's. (DNF).

Illustrative Model Fit

Par	Est	SE	z
DHR	-1.002	0.068	-14.73
SGR	-0.135	0.067	-2.01
GSR	-0.068	0.044	-1.55
SR	-0.011	0.056	-0.20
TDW	-0.673	0.140	-4.81
DNF	-0.026	0.052	-0.50

Illustrative Model Fit

Predictions for Garmisch Men's Downhill:

Name	Surname	Price	Win Prob %	ROI	Place
Beat	Feuz	3.5	15.5	-0.46	2
Matthias	Mayer	4.5	22.1	-0.01	3
Dominik	Paris	5.0	17.8	-0.11	1
Vincent	Kriechmayr	11.0	10.2	0.12	11
Max	Franz	11.0	1.4	-0.84	4
Christof	Innerhofer	11.0	1.5	-0.38	5

Illustrative Model Fit

- Model is doing something right in that the highest 3 win probabilities went to the first 3 finishers.

Illustrative Model Fit

- Model is doing something right in that the highest 3 win probabilities went to the first 3 finishers.
- However, bets on each of these skiers had negative value.

Illustrative Model Fit

- Model is doing something right in that the highest 3 win probabilities went to the first 3 finishers.
- However, bets on each of these skiers had negative value.
- Most value was on Kriechmayr. However, he finished 11th, so maybe model is lacking some crucial information that was available to marlket.

Summary

- The basic rank model here is intuitive, simple and links to several other models.
- Though it can be derived from an exponential model for race times, it has wider applicability and is perhaps robust to this specification.
- In applications, the choice of covariates is, as always, critical.
- Even with the correct covariates, the model suitability requires assessment.