From Zero to Hero

Stuart Coles

4 April, 2024

Sports Modelling

- One of the things that makes our job interesting is that although the objectives are pretty much the same for every sport, the details of the modelling depend very much on the nature and rules of each individual sport.

Sports Modelling

- One of the things that makes our job interesting is that although the objectives are pretty much the same for every sport, the details of the modelling depend very much on the nature and rules of each individual sport.
- Even for an individual sport, the methods may vary according to rule variations.

Sports Modelling

- One of the things that makes our job interesting is that although the objectives are pretty much the same for every sport, the details of the modelling depend very much on the nature and rules of each individual sport.
- Even for an individual sport, the methods may vary according to rule variations.
- Certain classes of ski racing have especially challenging rules.

The 'Two Manche' Issue

- There are 4 main categories of alpine ski racing: Slalom and Giant Slalom (technical disciplines); Super G and Downhill (speed disciplines).

The 'Two Manche' Issue

- There are 4 main categories of alpine ski racing: Slalom and Giant Slalom (technical disciplines); Super G and Downhill (speed disciplines).
- For Slalom and Giant Slalom each race consists of 2 runs (manche).

The 'Two Manche' Issue

- There are 4 main categories of alpine ski racing: Slalom and Giant Slalom (technical disciplines); Super G and Downhill (speed disciplines).
- For Slalom and Giant Slalom each race consists of 2 runs (manche).
- Only the fastest 30 athletes from the first manche qualify for the second manche.

The 'Two Manche' Issue

- There are 4 main categories of alpine ski racing: Slalom and Giant Slalom (technical disciplines); Super G and Downhill (speed disciplines).
- For Slalom and Giant Slalom each race consists of 2 runs (manche).
- Only the fastest 30 athletes from the first manche qualify for the second manche.
- The times from the 2 manche are added and the winner is the athlete with the smallest overall race time.

From Zero to Hero

In February, in the Slalom event at Chamonix, Swiss skier Daniel Yule became the first skier in history to win any World Cup race after finishing 30th in the first manche, 1.93 seconds behind the fastest racer Clement Noel.

From Zero to Hero

"Absolutely incredible, I got really lucky, staying 30th after the first run but then I managed to ski an amazing second run and... wow, it's just unbelievable. It was a long wait down here, but a nice one"

Lucky, but how Lucky?

Do you think the chances of Yule winning in Chamonix after finishing 30th in first manche, with a deficit of nearly 2 seconds, were roughly:

1. 1 in 10?
2. 1 in 100?
3. 1 in 1000?
4. 1 in 10000?

Lucky, but how Lucky?

Do you think the chances of Yule winning in Chamonix after finishing 30th in first manche, with a deficit of nearly 2 seconds, were roughly:

1. 1 in 10?
2. 1 in 100?
3. 1 in 1000?
4. 1 in 10000?

And how would you go about calculating the probability?

What's Unique about Ski races?

- In general there is a tendency for snow conditions to deteriorate as a manche progresses.
- To allow for this the following rules are adopted. . .

World Cup Rules

1. In the first manche the stronger a racer - according to current world rankings - the lower their start number (subject to a small amount of randomisation).

World Cup Rules

1. In the first manche the stronger a racer - according to current world rankings - the lower their start number (subject to a small amount of randomisation).
2. The fastest 30 racers from the first manche qualify for the second manche.

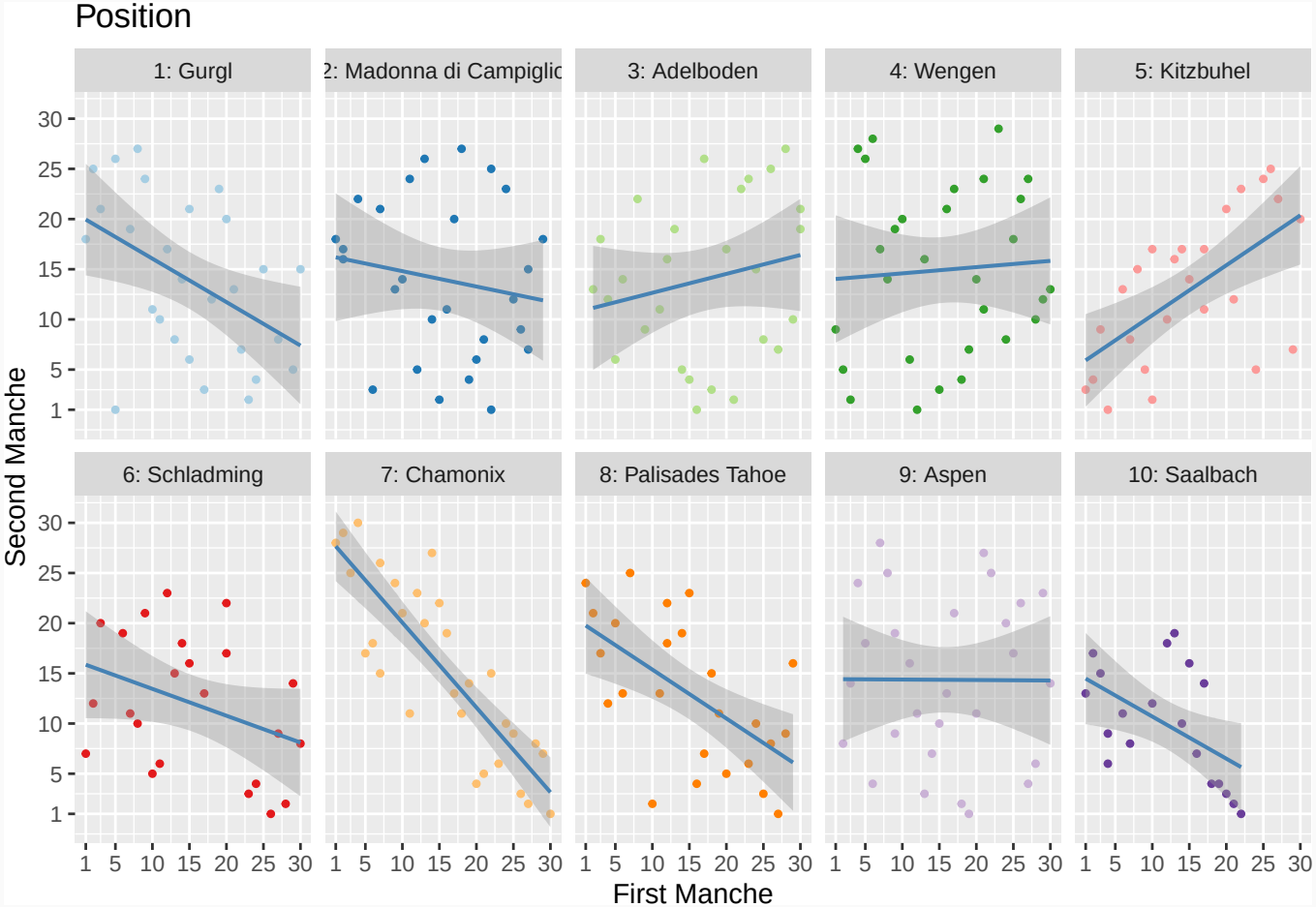
World Cup Rules

1. In the first manche the stronger a racer - according to current world rankings - the lower their start number (subject to a small amount of randomisation).
2. The fastest 30 racers from the first manche qualify for the second manche.
3. Starting positions in the second manche are determined by finishing positions in the first manche: the racer who finished 30th goes first, followed by the racer who finished 29th and so on, until the racer who finished first in the first manche goes last (30th) in the second manche.

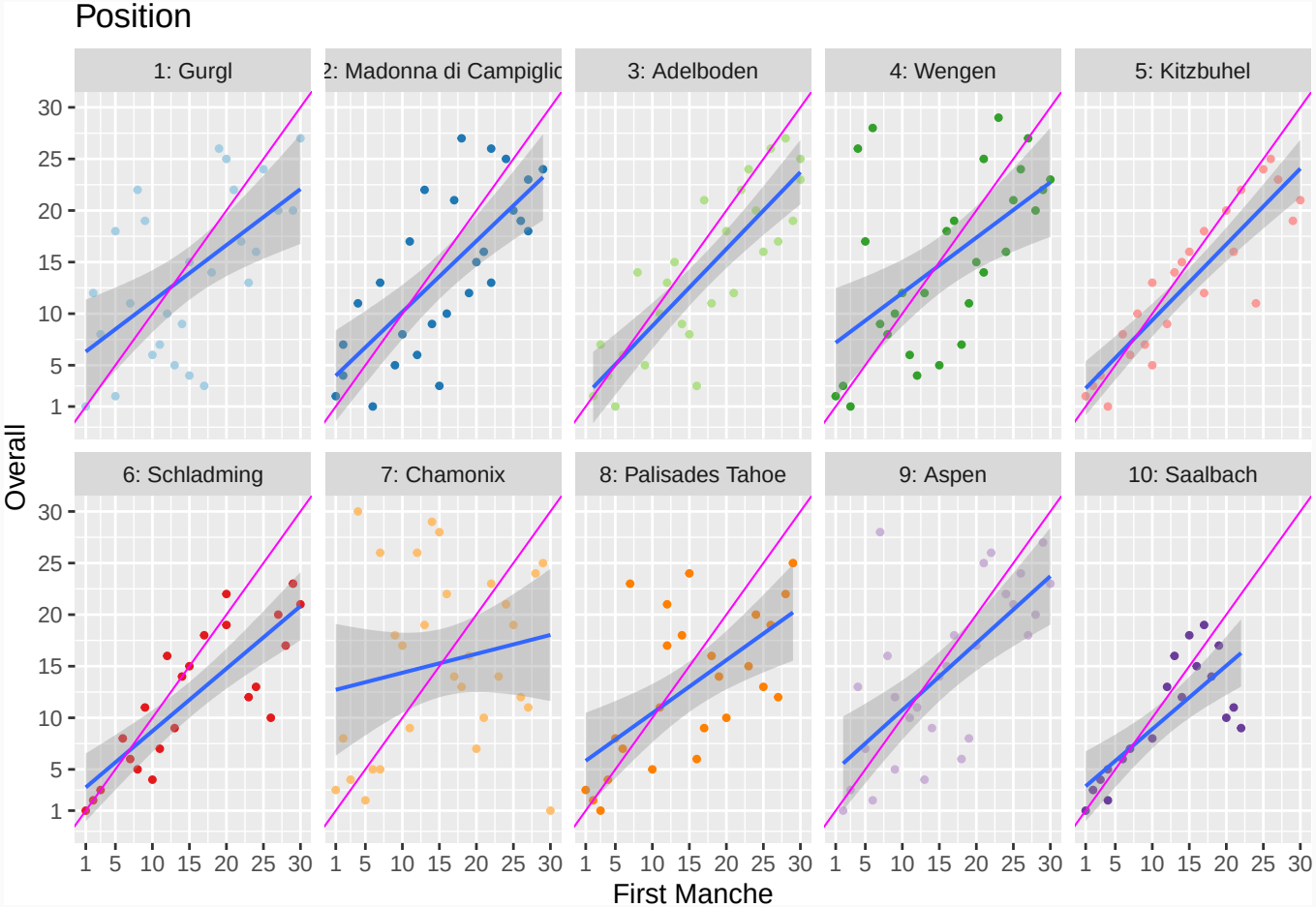
World Cup Rules

1. In the first manche the stronger a racer - according to current world rankings - the lower their start number (subject to a small amount of randomisation).
2. The fastest 30 racers from the first manche qualify for the second manche.
3. Starting positions in the second manche are determined by finishing positions in the first manche: the racer who finished 30th goes first, followed by the racer who finished 29th and so on, until the racer who finished first in the first manche goes last (30th) in the second manche.
4. In this way, the strongest skiers get the best conditions in the first manche, while the slowest skiers in the first manche get the best conditions in the second manche.

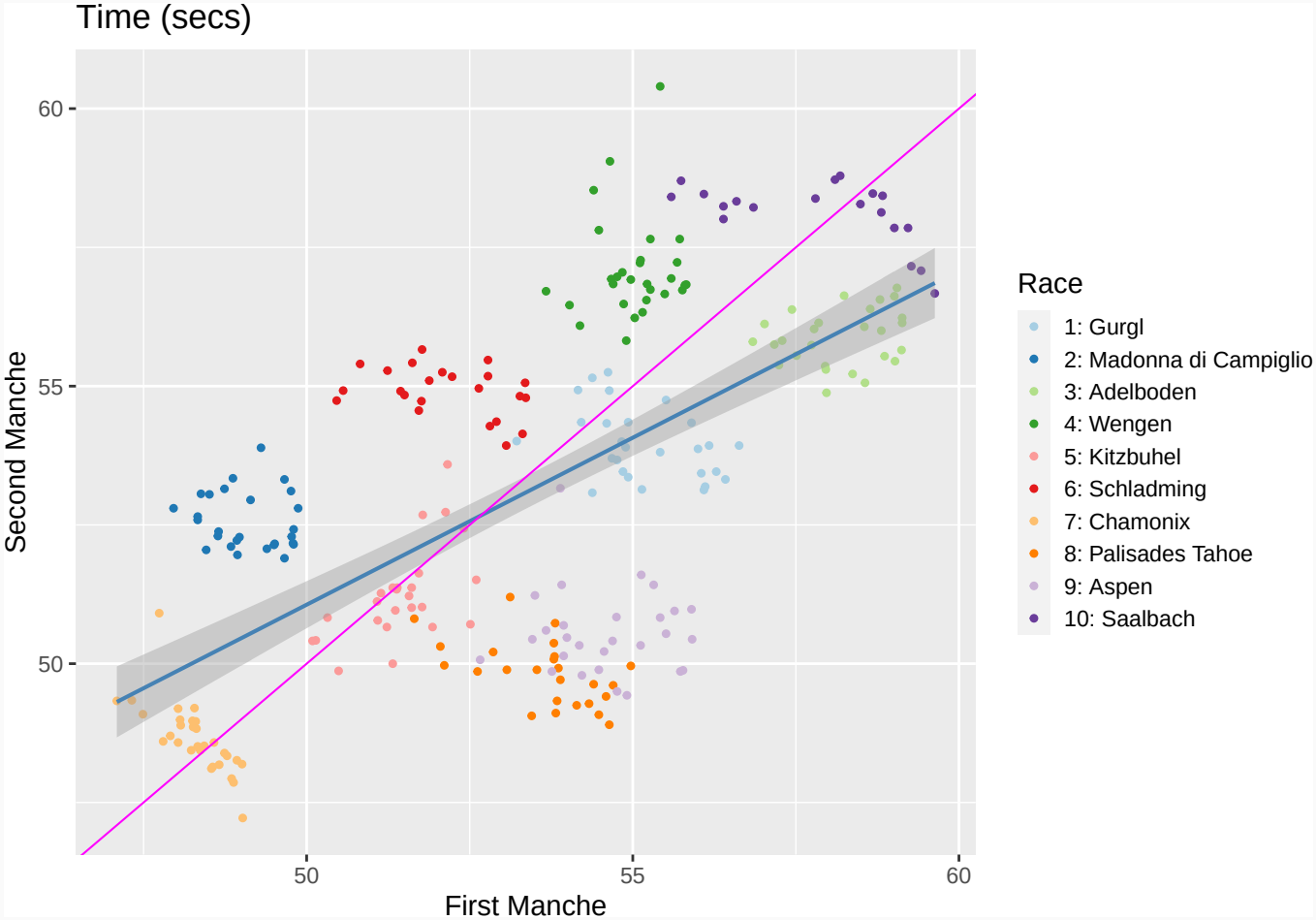
Graphical analysis 1



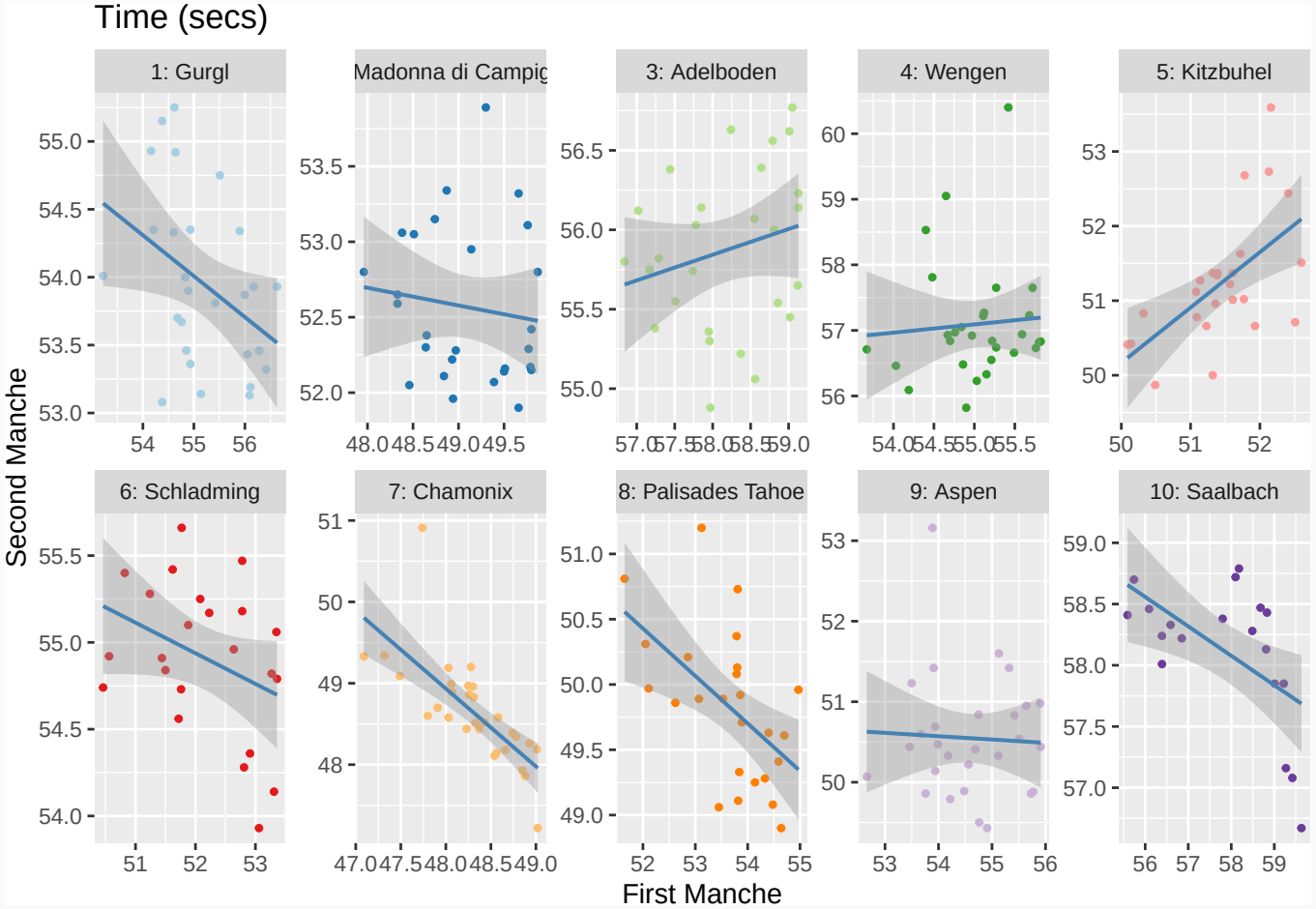
Graphical analysis 2



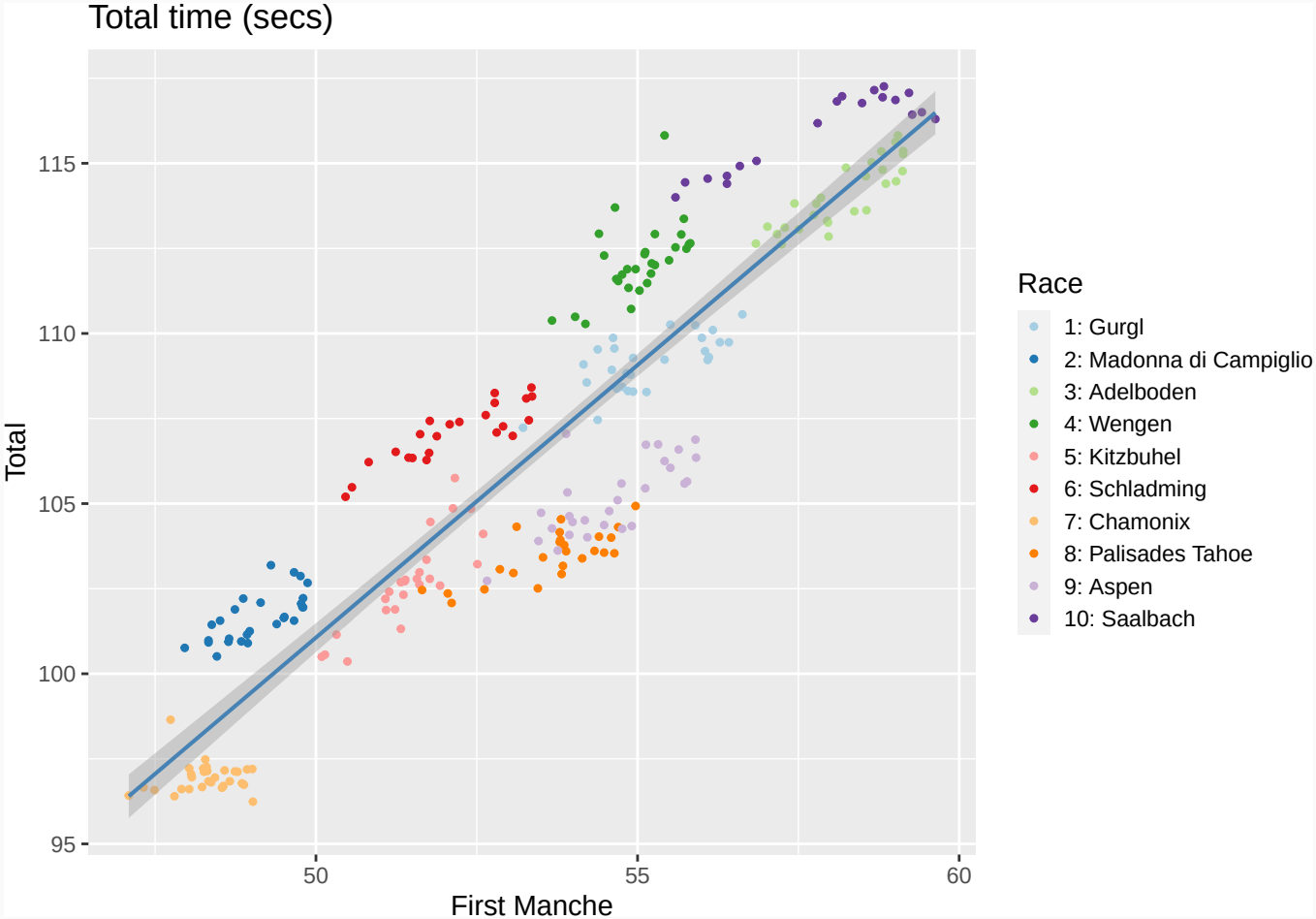
Graphical analysis 3



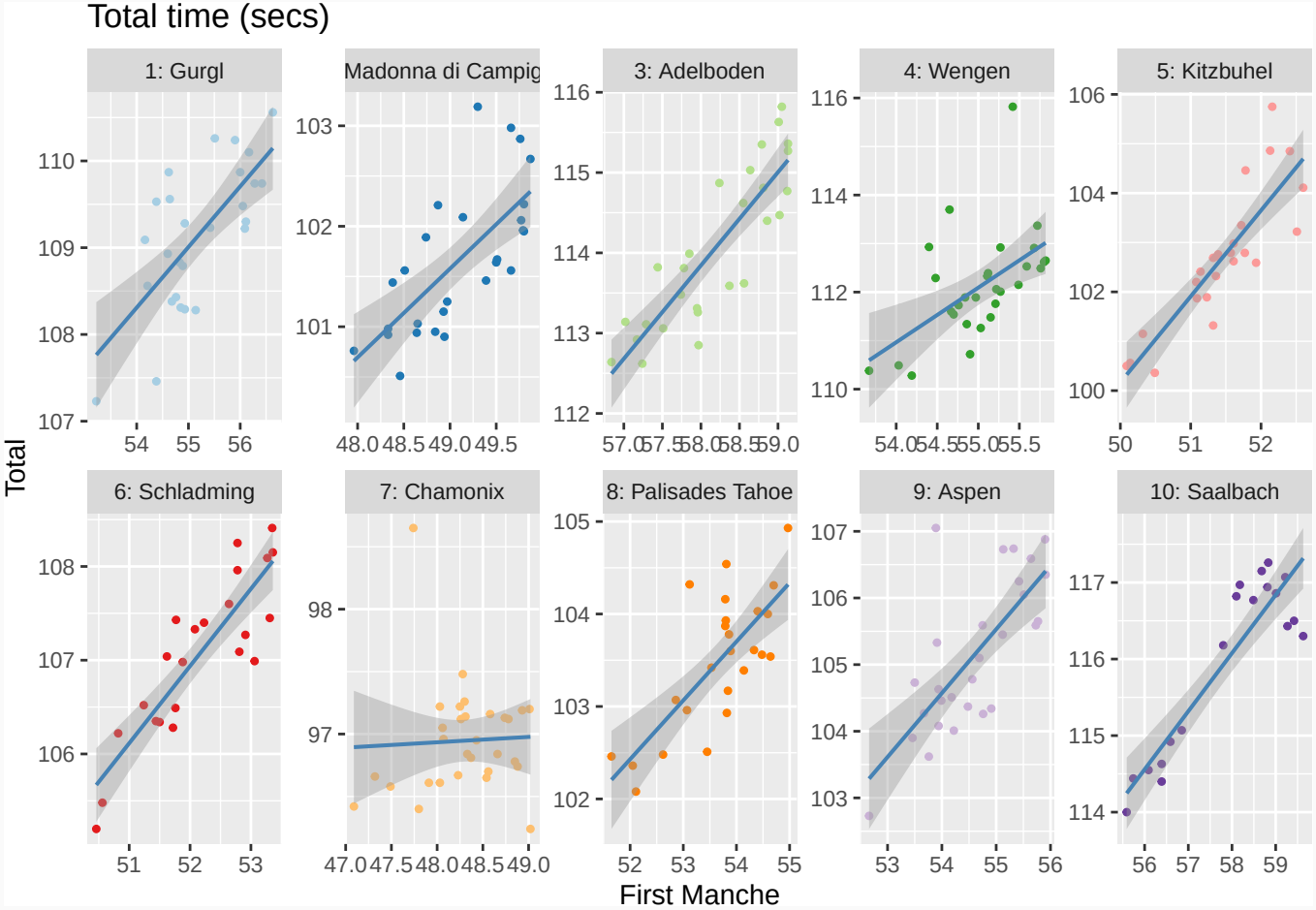
Graphical analysis 4



Graphical analysis 5



Graphical analysis 6



Apparent Features

1. Race times in second manche are not easily predictable from race times in first manche, mostly due to course changes from manche to manche.

Apparent Features

1. Race times in second manche are not easily predictable from race times in first manche, mostly due to course changes from manche to manche.
2. Times in second run are often - though not always - negatively correlated with times in first run. (Three effects: strong racers are strong for both races; regression to mean; deterioration of snow.)

Apparent Features

1. Race times in second manche are not easily predictable from race times in first manche, mostly due to course changes from manche to manche.
2. Times in second run are often - though not always - negatively correlated with times in first run. (Three effects: strong racers are strong for both races; regression to mean; deterioration of snow.)
3. Chamonix was unusual in that overall race time was virtually uncorrelated with first manche time.

Apparent Features

1. Race times in second manche are not easily predictable from race times in first manche, mostly due to course changes from manche to manche.
2. Times in second run are often - though not always - negatively correlated with times in first run. (Three effects: strong racers are strong for both races; regression to mean; deterioration of snow.)
3. Chamonix was unusual in that overall race time was virtually uncorrelated with first manche time.
4. Daniel Yule's performance at Chamonix does stand out as exceptional, though the overall pattern of results in that particular race suggest such an achievement was plausible.

How did Daniel Yule win?

- One reason Daniel Yule achieved his win at Chamonix was the rapidly deteriorating snow, so that racing first in the second manche was a huge advantage.

How did Daniel Yule win?

- One reason Daniel Yule achieved his win at Chamonix was the rapidly deteriorating snow, so that racing first in the second manche was a huge advantage.
- The second reason is that he is currently one of the strongest slalom ski racers, who just happened to have a poor time in the first manche due to a mistake.

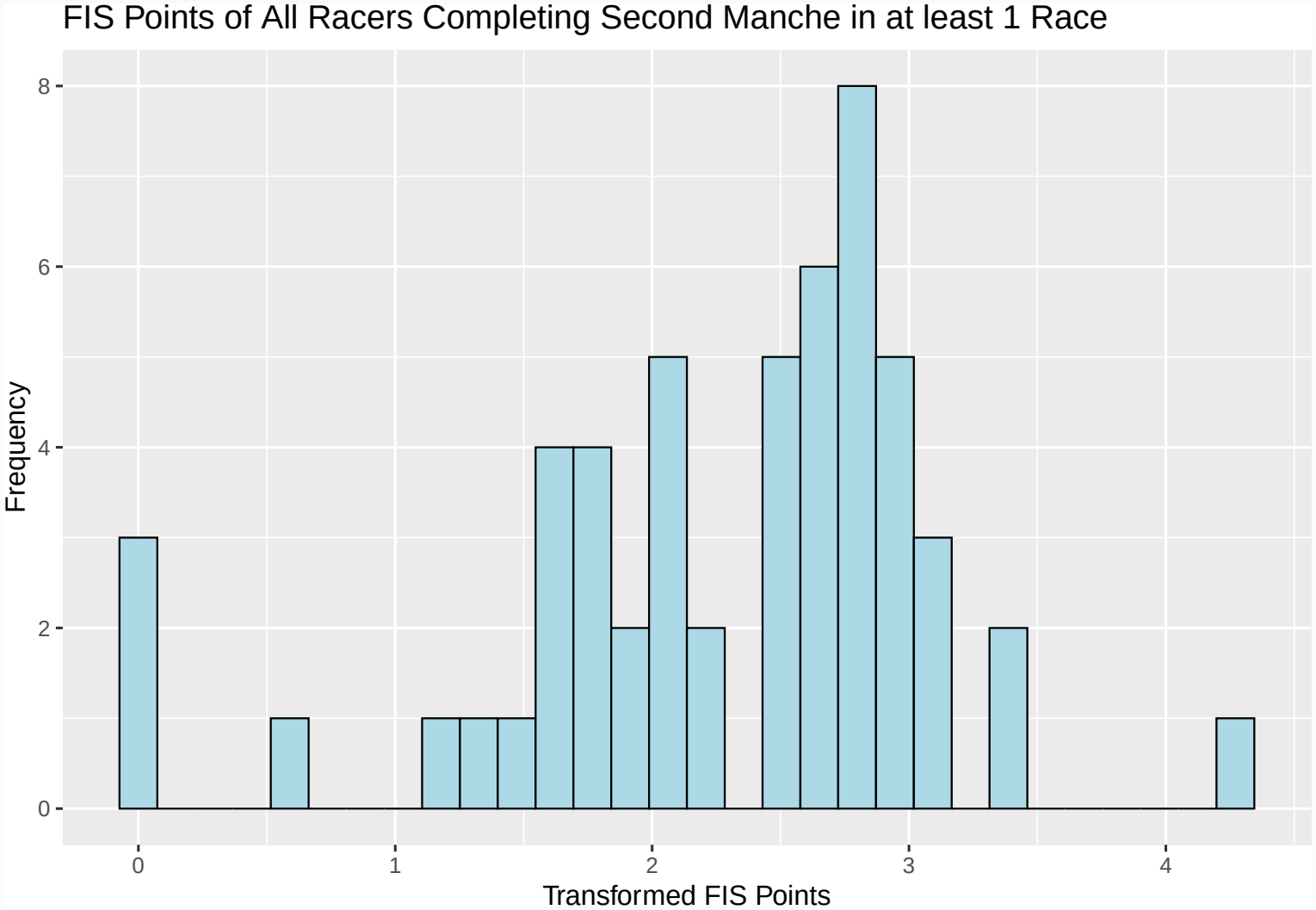
How did Daniel Yule win?

- One reason Daniel Yule achieved his win at Chamonix was the rapidly deteriorating snow, so that racing first in the second manche was a huge advantage.
- The second reason is that he is currently one of the strongest slalom ski racers, who just happened to have a poor time in the first manche due to a mistake.
- A final factor is that the time difference between the 1st and 30th racer in the first manche was relatively small.

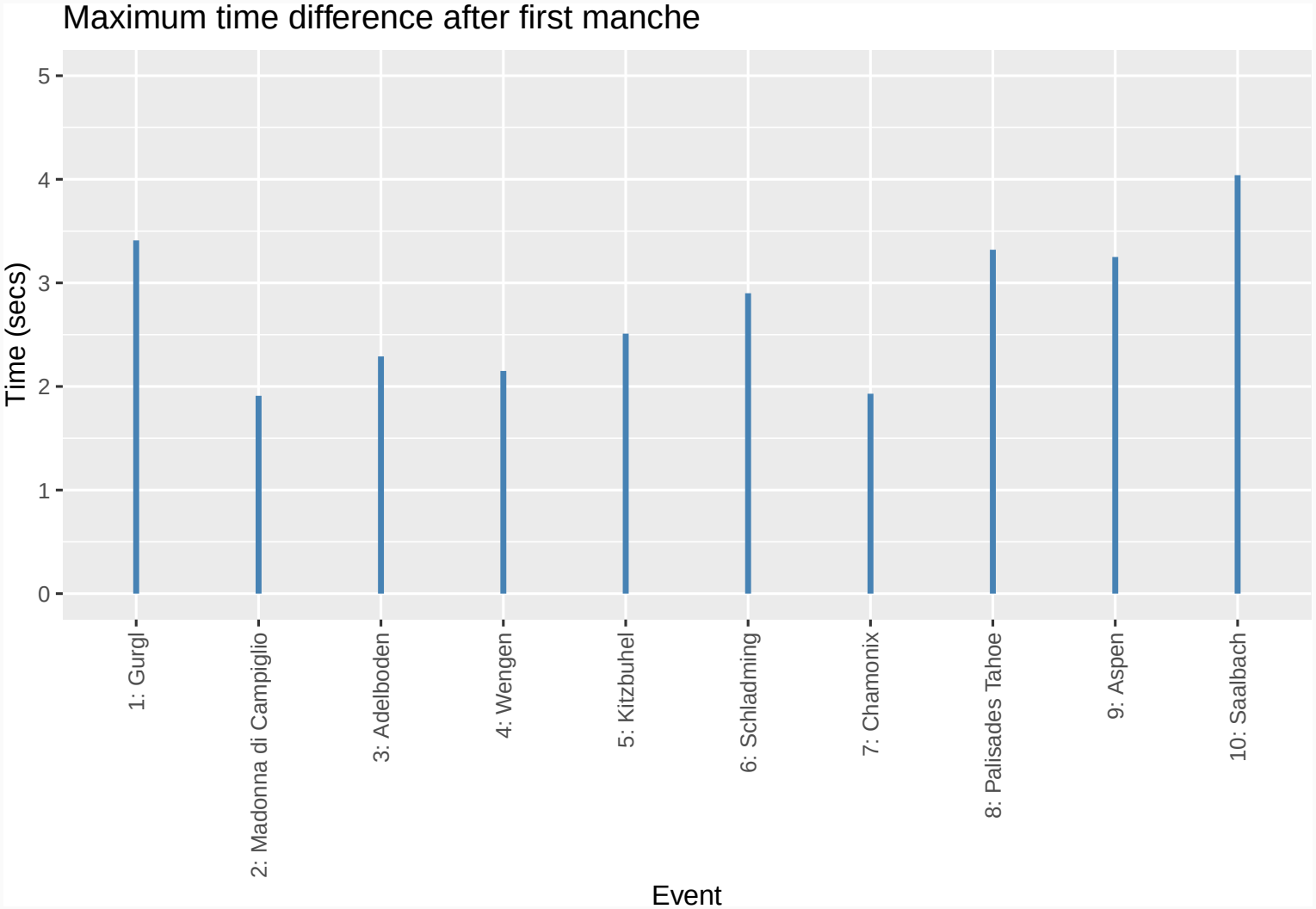
How did Daniel Yule win?

- One reason Daniel Yule achieved his win at Chamonix was the rapidly deteriorating snow, so that racing first in the second manche was a huge advantage.
- The second reason is that he is currently one of the strongest slalom ski racers, who just happened to have a poor time in the first manche due to a mistake.
- A final factor is that the time difference between the 1st and 30th racer in the first manche was relatively small.

Racer Effects



Time Difference Effect



Modelling Strategy

1. Model race results (ranks), not times, since there is no basis for predicting times.
2. Adjust standard models for ranks to allow for observed times in first manche.
3. Optionally include racer strength and starting position as covariates.
4. Optionally include results (ranks) from first manche.

A Standard Model for Race Times

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

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

Race Winner Probabilities: Standard Model

Denoting by R_1 the identity of the race winner:

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

Proof of Standard 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 Standard 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 properties of Exponential distribution:

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

Proof of Standard Result

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

Important features

1. Result remains true if each X_i is a monotonic transform of an exponential variable.

Important features

1. Result remains true if each X_i is a monotonic transform of an exponential variable.
2. Result extends (by memoryless property of the exponential distribution) to provide the joint probability of the complete rankings.

Important features

1. Result remains true if each X_i is a monotonic transform of an exponential variable.
2. Result extends (by memoryless property of the exponential distribution) to provide the joint probability of the complete rankings.

3. This latter aspect leads to a log-likelihood based on complete race result:

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

where S_m is set of racers outside of the top $m-1$ positions.

A Rank-Offset Model

To enable the inclusion of times from the first manche, assume that total race times are as follows:

$$Y_k = X_k + a_k, \quad k = 1, \ldots, N$$

where the X_k are independent and exponentially distributed,

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

and the a_k are known constants.

A Rank-Offset Model

To enable the inclusion of times from the first manche, assume that total race times are as follows:

$$Y_k = X_k + a_k, \quad k = 1, \ldots, N$$

where the X_k are independent and exponentially distributed,

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

and the a_k are known constants.

Under these conditions, what is $P(R_1 = C_k)$?

A Rank-Offset Model

Like before:

$$P(R_1 = C_k) = \int_{y=a_k}^{\infty} P(Y_k = y) P(\min_{j \neq k}\{Y_j\} > y) dy$$

Preliminary Result

Let

$$Z = \min(Y_1, \ldots, Y_m)$$

$$\begin{aligned} P(Z > z) &= P(Y_1 > z, \ldots, Y_m > z) \\ &= \prod_{i=1}^{m} P(X_i > z - a_i) \\ &= \prod_{i=1}^{m} \exp(-\lambda_i (z - a_i)_+) \end{aligned}$$

where $x_+ = \min(x, 0)$.

Main Result

Without loss of generality, assume that

$$a_1 \leq a_2 \leq \ldots \leq a_N$$

and let

$$d_j = a_j - a_k$$

for $j = 1, \ldots, N$. Also set $d_{N+1} = \infty$.

Main Result

Then

$$P(R_1 = C_K) = \sum_{j=0}^{m-k} I_j$$

where

$$I_j = \frac{\lambda_k}{\sum_{i=1}^{k+j} \lambda_i} \exp(\sum_{i=1}^{k+j} \lambda_i d_i - \lambda_k d_k) \left(\exp(-\sum_{i=1}^{k+j} \lambda_i d_{k+j}) - \exp(-\sum_{i=1}^{k+j} \lambda_i d_{k+j+1})) \right)$$

However!

- It's difficult to generalise the model to obtain the probability for the complete set of rankings.

However!

- It's difficult to generalise the model to obtain the probability for the complete set of rankings.
- It pains me to say it, but my girlfriend found a recursive formula to calculate this probability.

However!

- It's difficult to generalise the model to obtain the probability for the complete set of rankings.
- It pains me to say it, but my girlfriend found a recursive formula to calculate this probability.
- Brilliantly, she was able to exploit the memoryless property of the exponential distribution, thereby avoiding complicated integrals.

However!

- It's difficult to generalise the model to obtain the probability for the complete set of rankings.
- It pains me to say it, but my girlfriend found a recursive formula to calculate this probability.
- Brilliantly, she was able to exploit the memoryless property of the exponential distribution, thereby avoiding complicated integrals.
- Unfortunately, each recursion has conditional branches with multiple recursive function calls. A race with 20 racers is just about manageable; with 30 racers each likelihood calculation is impossibly slow.

Additionally...

The model is no longer robust to the assumption of an exponential distribution.

Gumbel Alternative to Classic Model

$$Y_k = \alpha \log E_k + \beta_k + a_k, \quad k = 1, \ldots, N$$

where the E_k are unit exponential, the β_k are race/racer specific effects and the a_k are the first manche times.

Gumbel Alternative to Classic Model

With this set-up:

Gumbel Alternative to Classic Model

With this set-up:

1. $$P(R_1 = C_k) = \frac{\exp\{-(\beta_k + a_k)/\alpha\}}{\sum_{j=1}^{N} \exp\{-(\beta_j + a_j)/\alpha\}}$$

Gumbel Alternative to Classic Model

With this set-up:

1. $$P(R_1 = C_k) = \frac{\exp\{-(\beta_k + a_k)/\alpha\}}{\sum_{j=1}^{N} \exp\{-(\beta_j + a_j)/\alpha\}}$$

2. The result does now easily extend to the probability of joint ranks.

Gumbel Alternative to Classic Model

With this set-up:

1. $$P(R_1 = C_k) = \frac{\exp\{-(\beta_k + a_k)/\alpha\}}{\sum_{j=1}^{N} \exp\{-(\beta_j + a_j)/\alpha\}}$$

2. The result does now easily extend to the probability of joint ranks.
3. Results derive directly from the memoryless property of the Exponential distribution.

Gumbel Alternative to Classic Model

With this set-up:

1. $$P(R_1 = C_k) = \frac{\exp\{-(\beta_k + a_k)/\alpha\}}{\sum_{j=1}^{N} \exp\{-(\beta_j + a_j)/\alpha\}}$$

2. The result does now easily extend to the probability of joint ranks.
3. Results derive directly from the memoryless property of the Exponential distribution.
4. When the $a_k = 0$, this model is a transform of the standard exponential model, but includes additionally the scale parameter α.

Models for β_k

Optional covariates for β_k included in a linear predictor:

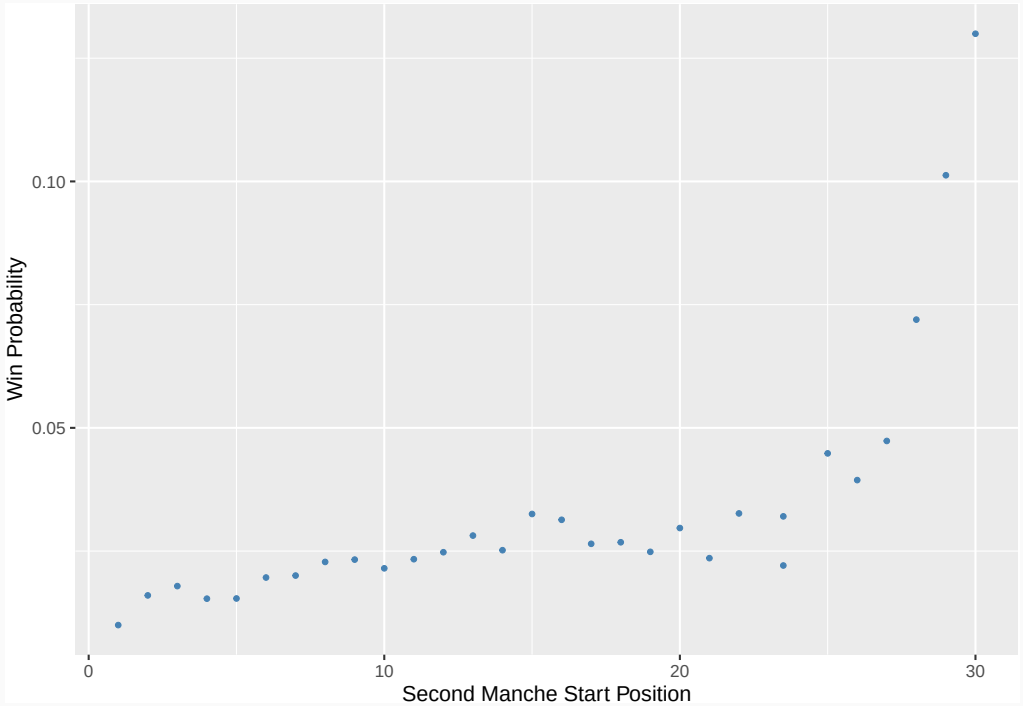
- Racer points at start of season (proxy for individual racer effect) - included on a log scale.
- Starting number in manche.

Potential Problems with Confounding

- In the first manche, start position is strongly confounded with FIS points.
- In the second manche, there will also be confounding, but in opposite direction.
- My motivation for including results from first manche was to balance out these confounding effects.
- From a simulation study, the model proves to be identifiable - albeit with lower precision - in the presence of these confounding effects (even just using second manche data.)

Results

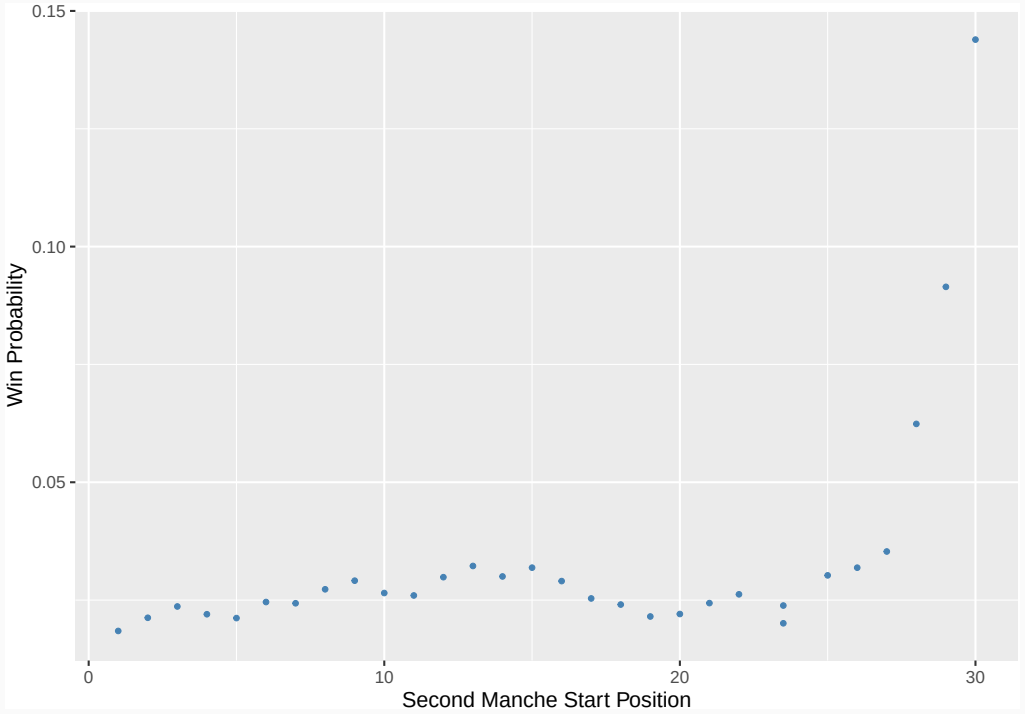
Model 1: Universal position effect; include first manche.



$P(\text{Yule win}) = 0.010$.

Results

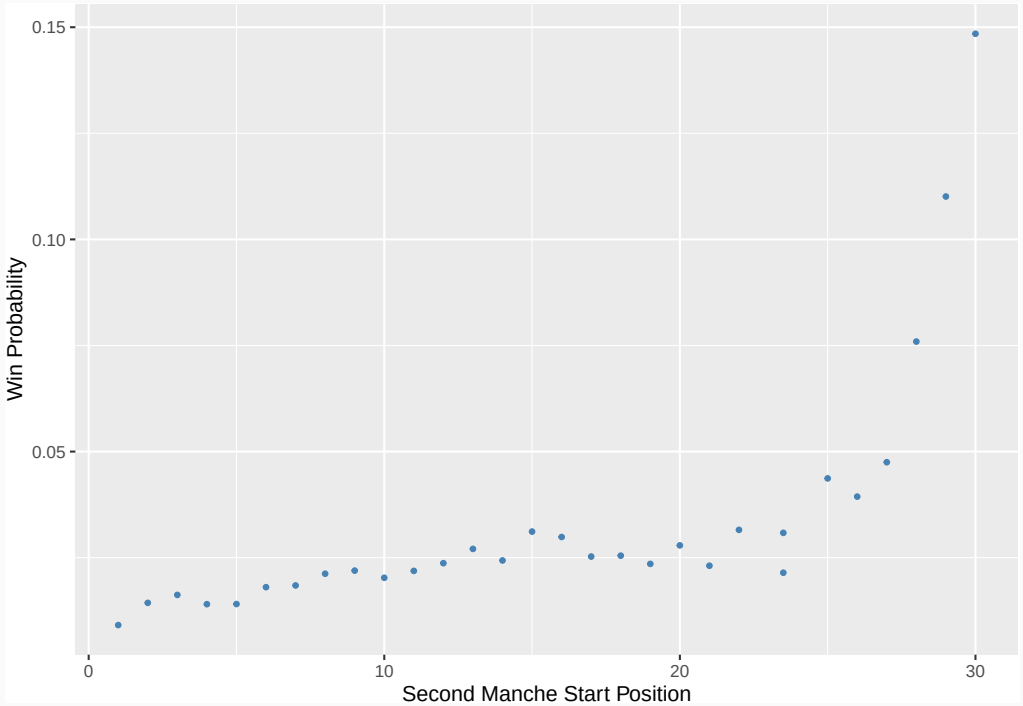
Model 2: Universal position effect; exclude first manche.



$P(\text{Yule win}) = 0.018.$

Results

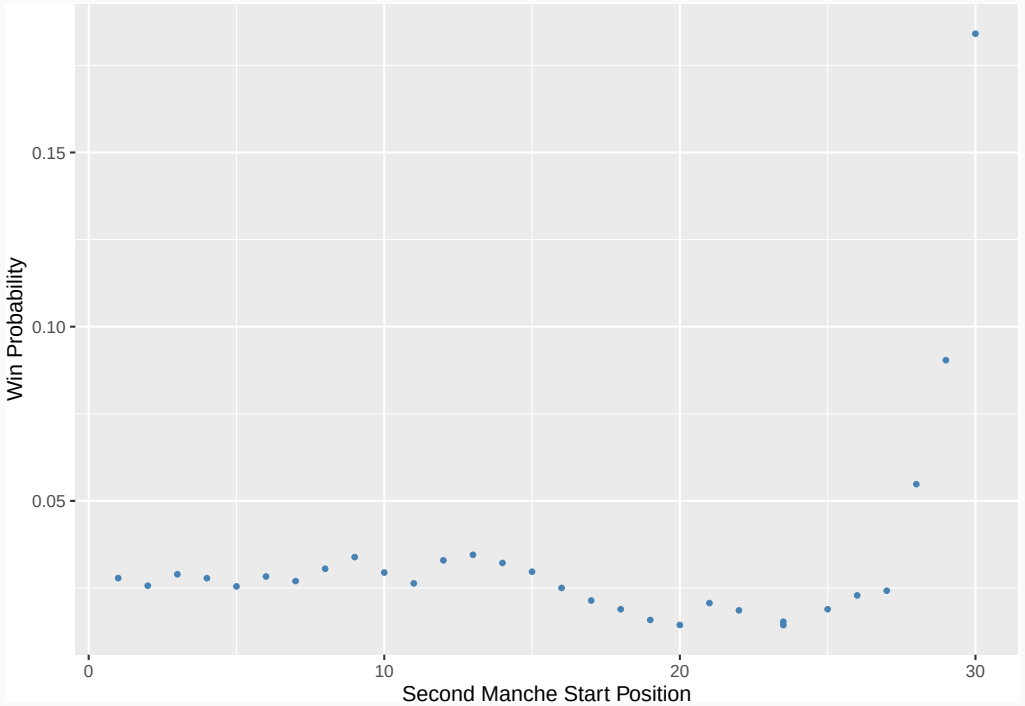
Model 3: Event-specific position effect; include first manche.



$P(\text{Yule win}) = 0.009.$

Results

Model 4: Event-specific position effect; exclude first manche.



$P(\text{Yule win}) = 0.028.$

Conclusions

1. There's variation between models, but Yule's win probability in Chamonix is generally in the range 1% - 3%.
2. Points effect not generally significant. Possibly due to confounding with start position, though simulations suggests inference is reliable anyway.
3. I'd expected the starting position regression parameter for Chamonix to be quite different from the other locations, but this didn't seem to be the case.

Summary

1. A rank offset model based on exponential race times turns out to be mathematically challenging, but feasible.
2. Computations, however, are prohibitively slow for competitions of 30 racers.
3. The Gumbel model is, in any case, a more natural framework for this type of development. (Natural, doesn't mean correct or accurate, though).
4. Models fitted to the ski data don't lead to entirely convincing parameter estimates, though with so few data it's very difficult to determine the cause for this effect.
5. It seems reasonable to conclude that Daniel Yule's win probability in Chamonix was of the order of 1 in 100.

But...

- The mathematics are elegant and fun, but statistically is it worth it?

But...

- The mathematics are elegant and fun, but statistically is it worth it?
- If the model is not robust to choice of distribution for race times, and if times from the first manche are required, why not just choose a distribution for the race times and model those directly?

But...

- The mathematics are elegant and fun, but statistically is it worth it?
- If the model is not robust to choice of distribution for race times, and if times from the first manche are required, why not just choose a distribution for the race times and model those directly?
- Admittedly, the results here then enable calculation of win probabilities etc. based on either the Exponential or Gumbel models.

But...

- The mathematics are elegant and fun, but statistically is it worth it?
- If the model is not robust to choice of distribution for race times, and if times from the first manche are required, why not just choose a distribution for the race times and model those directly?
- Admittedly, the results here then enable calculation of win probabilities etc. based on either the Exponential or Gumbel models.
- But plausibly there would be additional precision by basing inference directly on Exponential/Gumbel models for race times, rather than a likelihood that uses only the ranks.