Who is the best Formula One driver?

Erik-Jan van Kesteren & Tom Bergkamp

Photo by Bill Stephan on Unsplash

“Why would you like that?”

“Such a stupid sport!”

“It’s all about the cars anyway!”

Is it?

Who is the best Formula One driver?

A Bayesian multilevel Beta regression model

Erik-Jan van Kesteren & Tom Bergkamp

Outline

1. Data collection & pre-processing
2. Measuring performance
3. Model development
4. Model estimation & checks
5. Inferences & conclusions

Photo by José Pablo Domínguez on Unsplash

The data

Formula One 101

The hybrid era of F1

- Rules & regulations change
- Most notably: engine regulations
- 2014-2021 is known as the “hybrid era”
- Every year/season has around 21 races

The constructors

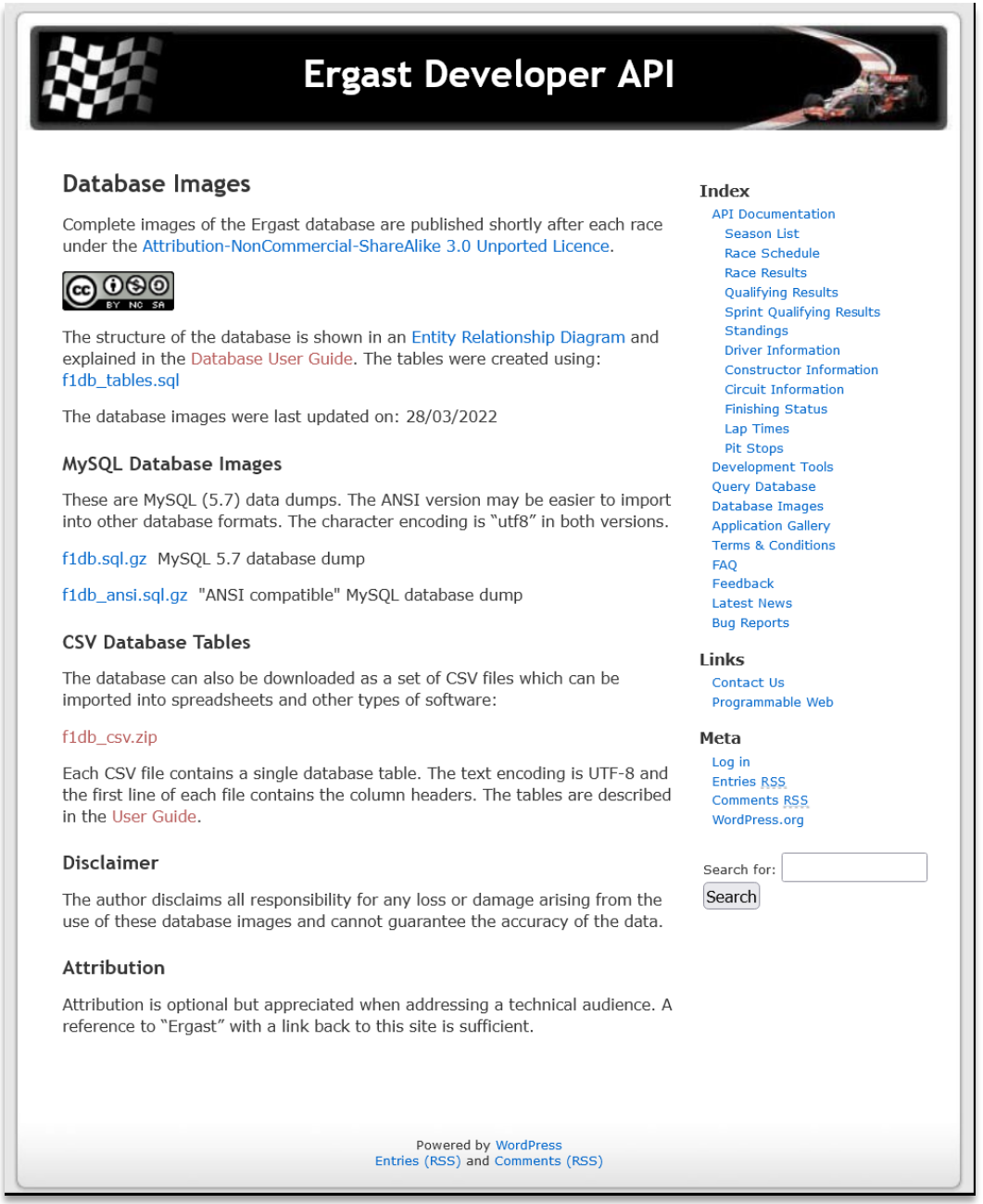
- There are around 10 teams
- Two cars each, new cars each year
- Teams change owners & name
- New teams enter sometimes
- Ferrari, Mercedes & Red Bull are rich

The drivers

- Drivers usually enter F1 with a “junior” team
- They change team: everyone wants to go to the best teams
- Drivers are compared to their teammates a lot

The data

- We want data about all race results of the hybrid era
- It exists! **Ergast**
  - An application programming interface (API)
  - A full data dump
  - Updated after every race
  - We got permission to use & reupload on our code repo!



The data: preprocessing

- Download Ergast data
- Join race results with driver properties & circuit properties
- Add information scraped from **Wikipedia**:
  - Weather (wet or dry race)
  - Circuit type (street circuit / permanent circuit)
- Some cleanup & variable type conversion

The data: tidy!

```
# A tibble: 3,267 x 9
   driver          constructor  year round circuit     position weather_type circuit_type status
   <fct>           <fct>       <int> <int> <chr>          <int> <fct>        <fct>        <fct>
 1 rosberg         mercedes     2014     1 albert_park        1 dry          street       Finished
 2 kevin_magnussen mclaren      2014     1 albert_park        2 dry          street       Finished
 3 button          mclaren      2014     1 albert_park        3 dry          street       Finished
 4 alonso          ferrari      2014     1 albert_park        4 dry          street       Finished
 5 bottas          williams     2014     1 albert_park        5 dry          street       Finished
 6 hulkenberg      force_india  2014     1 albert_park        6 dry          street       Finished
 7 raikkonen       ferrari      2014     1 albert_park        7 dry          street       Finished
 8 vergne          toro_rosso   2014     1 albert_park        8 dry          street       Finished
 9 kvyat           toro_rosso   2014     1 albert_park        9 dry          street       Finished
10 perez           force_india  2014     1 albert_park       10 dry          street       Finished
# ... with 3,257 more rows
```

Measuring “performance”

The best race result is finishing first

The worst race result is finishing last

Sometimes, a contestant does not finish at all!?

Dealing with non-finishes

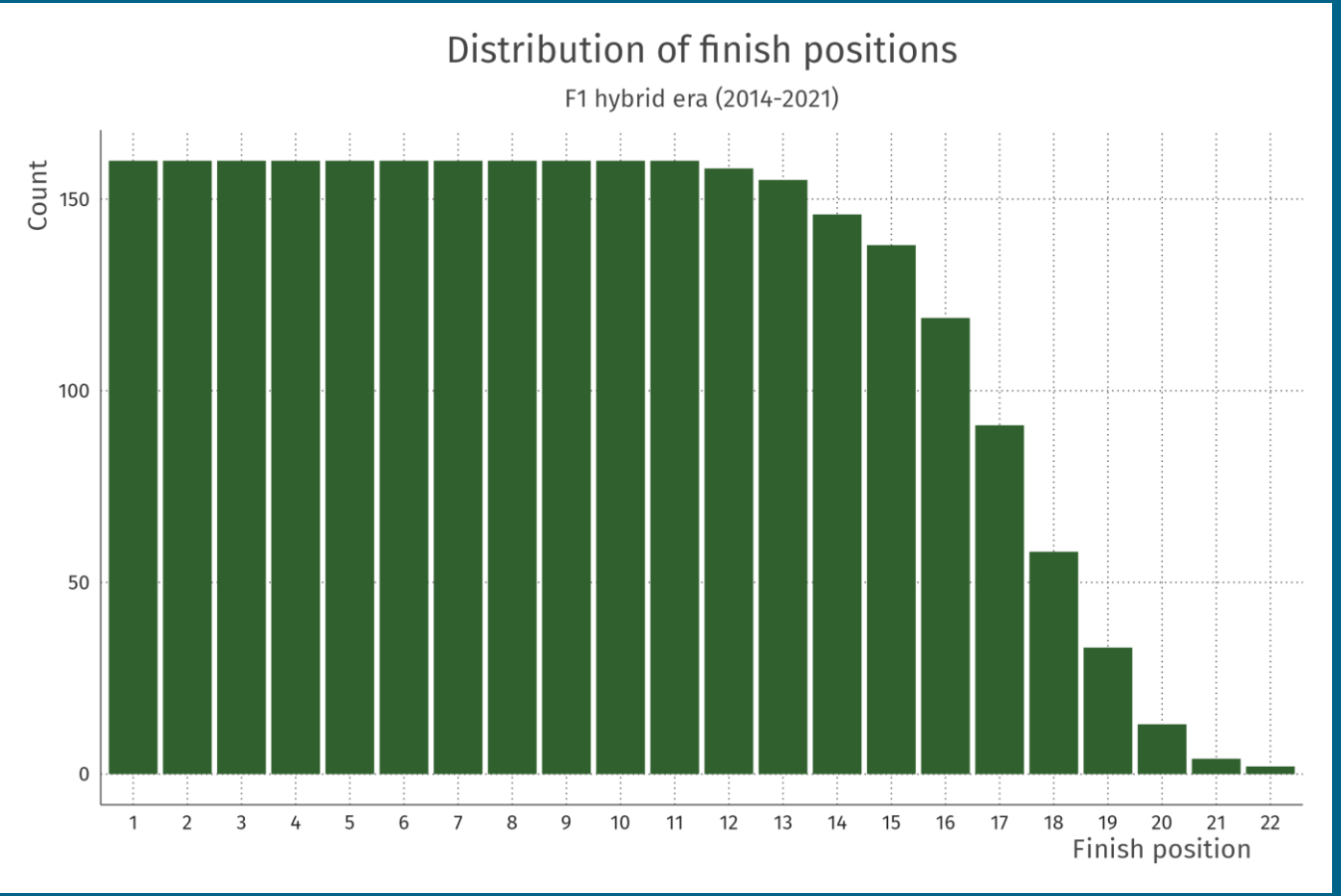
Many incidents in F1

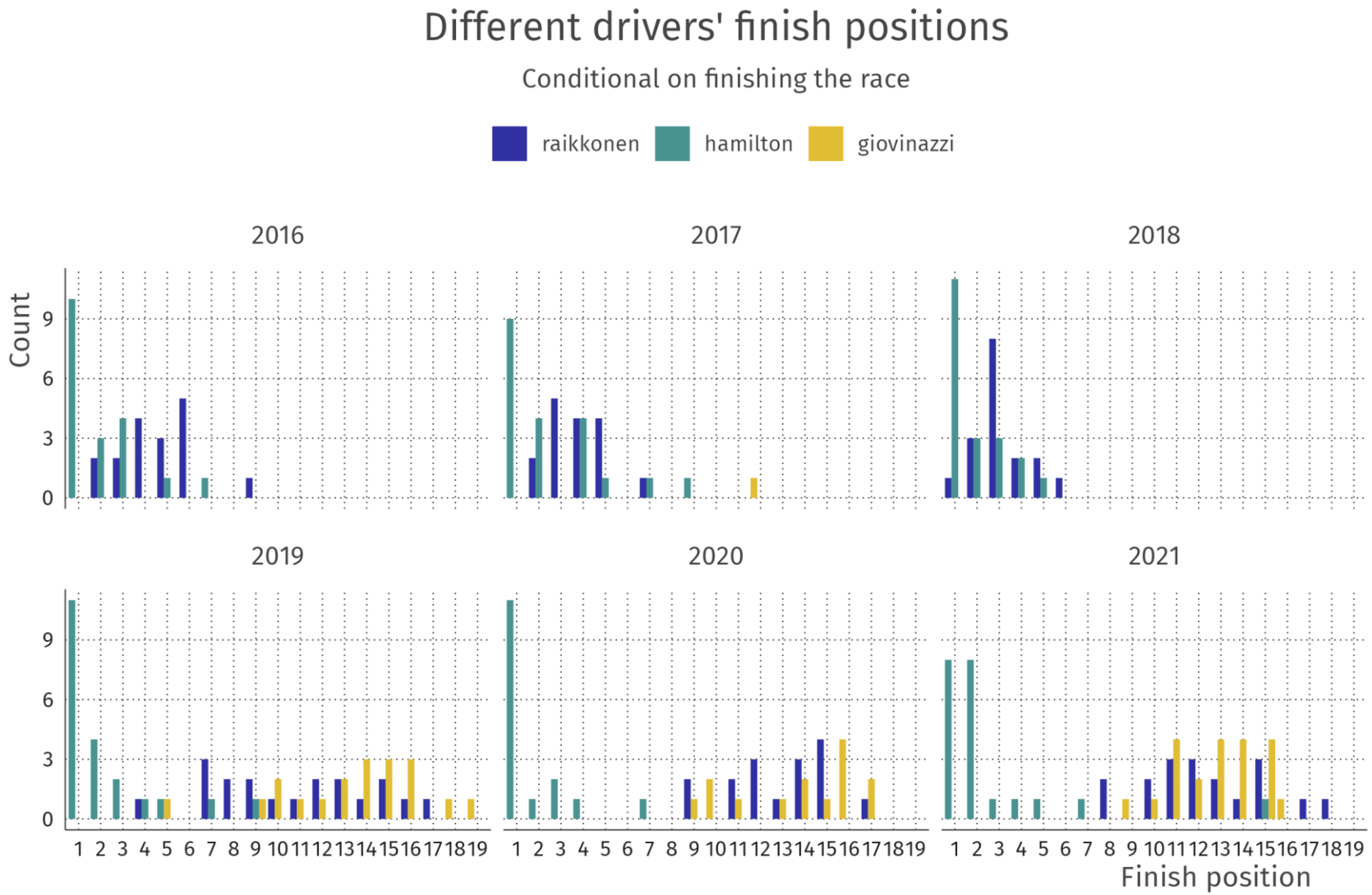
- Drivers make mistakes
- Constructors forget screws
- Random stuff happens
- The world is complicated

We don't have a way to attribute not-finishing to drivers or constructors or randomness!

SIMPLIFYING ASSUMPTION

- Look only at *finished* races




Different drivers' finish positions
Conditional on finishing the race
raikkonen
hamilton
giovinazzi
2016
2017
2018
2019
2020
2021
Count
0
3
6
9
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Finish position

Which is a better performance?

Which is a better performance?

A

Finishing 14th in a race where 20 people finish

Which is a better performance?

A

Finishing 14th in a race where 20 people finish

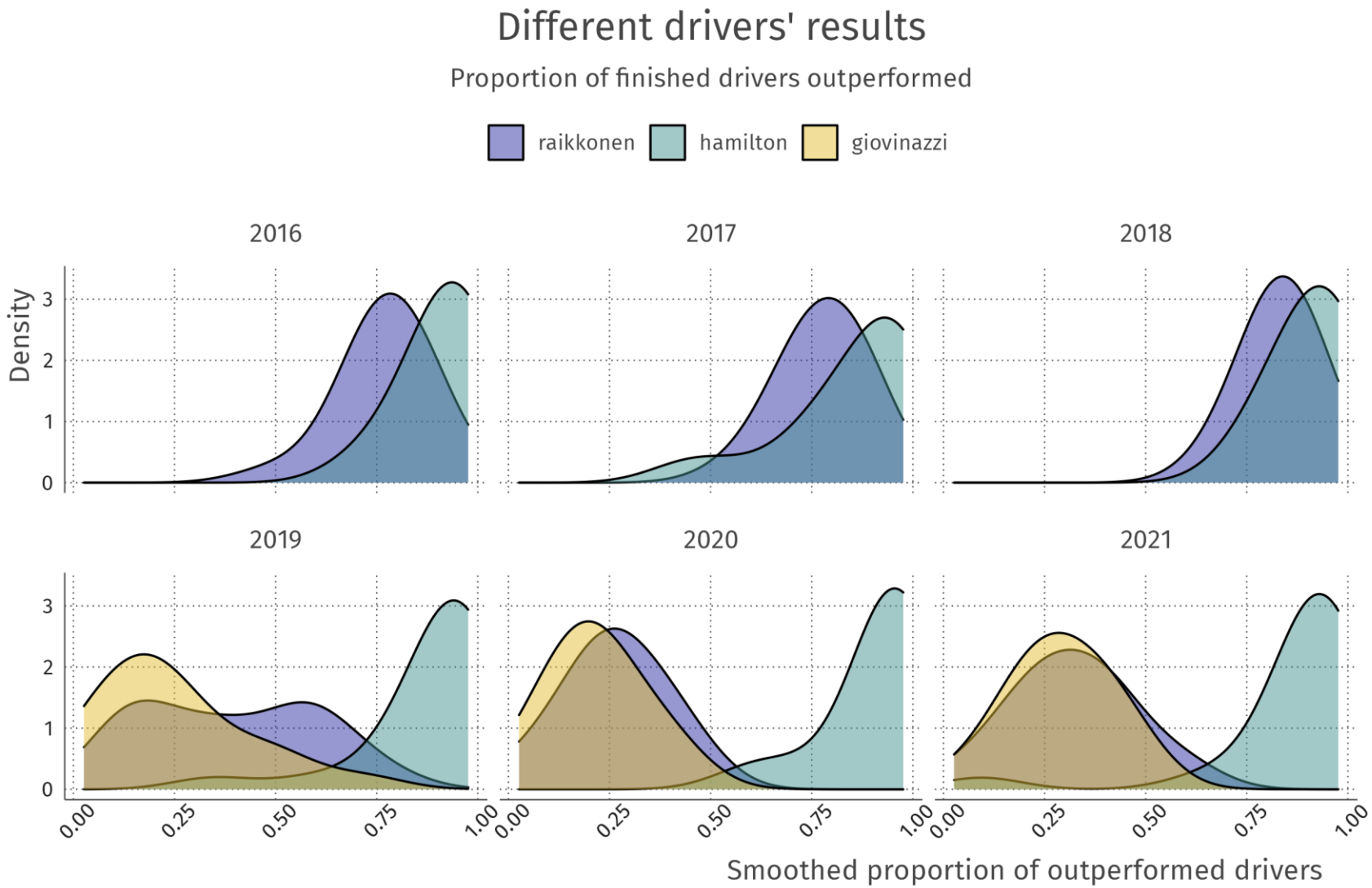
B

Finishing 14th in a race where 14 people finish

Solution: the proportion

We measure performance as the *proportion of drivers beaten*

- For each driver for each race, the proportion of drivers who finished behind the driver's position
- Smooth this estimate using recommended procedure from the R package betareg


Different drivers' results
Proportion of finished drivers outperformed
raikkonen
hamilton
giovinazzi
2016
2017
2018
2019
2020
2021
Density
0
1
2
3
0.00
0.25
0.50
0.75
1.00
Smoothed proportion of outperformed drivers

Modeling performance

Model for a proportion

- We use Beta regression (Ferrari & Cribari-Neto, 2004)
- Similar to exponential-family generalized linear model
- Outcome distribution: Beta with non-standard parameterization (mean and precision parameters)

$$p(y) = Beta(\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\big((1-\mu)\phi\big)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$$

- We make a linear model for $\text{logit}(\mu)$

Model for a proportion

Mean performance is the result of four components

1. Average driver skill: β_d
2. Driver form in a season: β_{ds}
3. Average constructor (car) advantage: β_c
4. Constructor form in a season: β_{cs}

So for driver d, constructor c in season s:

$$logit(\mu_{dcs}) = \beta_d + \beta_{ds} + \beta_c + \beta_{cs}$$

Multilevel model

49 drivers, 19 constructors, 8 seasons, 19-22 races

$$y_{dcsr} \sim Beta(\mu_{dcs}, \phi)$$

i.i.d. races within season, driver, constructor

$$logit(\mu_{dcs}) = \beta_d + \beta_{ds} + \beta_c + \beta_{cs}$$

Mean as function of 4 components:

$$\beta_d \sim \mathcal{N}(0, \sigma_d^2)$$

Driver average skill random intercept

$$\beta_{ds} \sim \mathcal{N}(0, \sigma_{ds}^2)$$

Driver seasonal form random intercept

$$\beta_c \sim \mathcal{N}(0, \sigma_c^2)$$

Constructor average advantage random intercept

$$\beta_{cs} \sim \mathcal{N}(0, \sigma_{cs}^2)$$

Constructor seasonal form random intercept

Parameter interpretation

$$logit(\mu_{dcs}) = \beta_d + \beta_{ds} + \beta_c + \beta_{cs}$$

- Proportion outcome per race so no intercept needed
- Parameters are on the *log odds-ratio scale*
- If all parameters are 0, then we have:

$$E[y_{dcsr}] = \mu_{dcs} = logit^{-1}(0 + 0 + 0 + 0) = 0.5$$

- This is an average driver with an average car in an average season ➡ 0.5 as result makes sense!

Estimating this model

The data again

```
# A tibble: 2,677 x 5
   driver          constructor  year round prop_trans
   <fct>           <fct>       <int> <int>      <dbl>
 1 rosberg         mercedes     2014     1      0.964
 2 kevin_magnussen mclaren      2014     1      0.893
 3 button          mclaren      2014     1      0.821
 4 alonso          ferrari      2014     1      0.75
 5 bottas          williams     2014     1      0.679
 6 hulkenberg      force_india  2014     1      0.607
 7 raikkonen       ferrari      2014     1      0.536
 8 vergne          toro_rosso   2014     1      0.464
 9 kvyat           toro_rosso   2014     1      0.393
10 perez           force_india  2014     1      0.321
# ... with 2,667 more rows
```

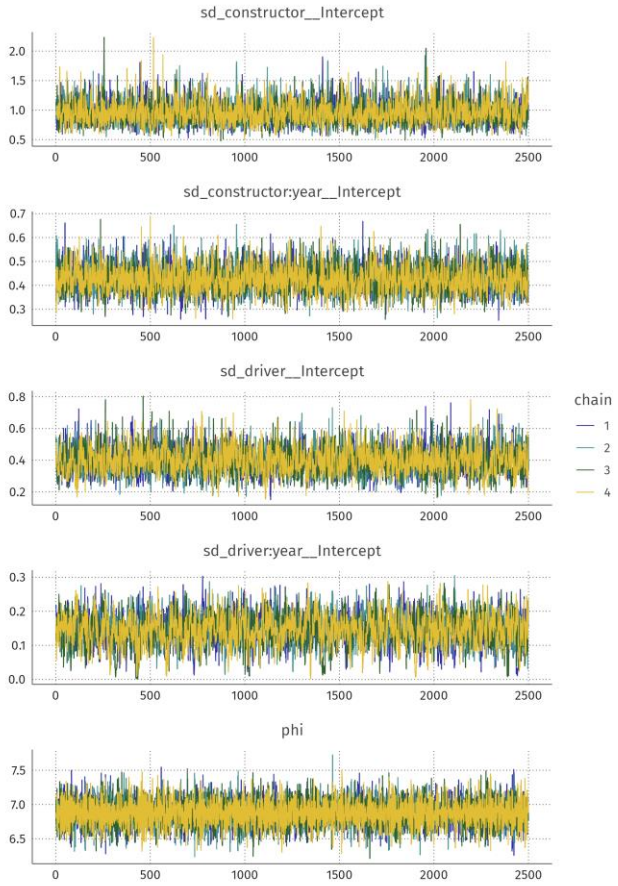
Model estimation in R

```
library(brms)

frm <-
  prop_trans ~ 0 + (1 | driver) + (1 | driver:year) +
                   (1 | constructor) + (1 | constructor:year)

fit <- brm(
  formula = frm,
  family  = Beta(),
  data    = f1_dat
)
```

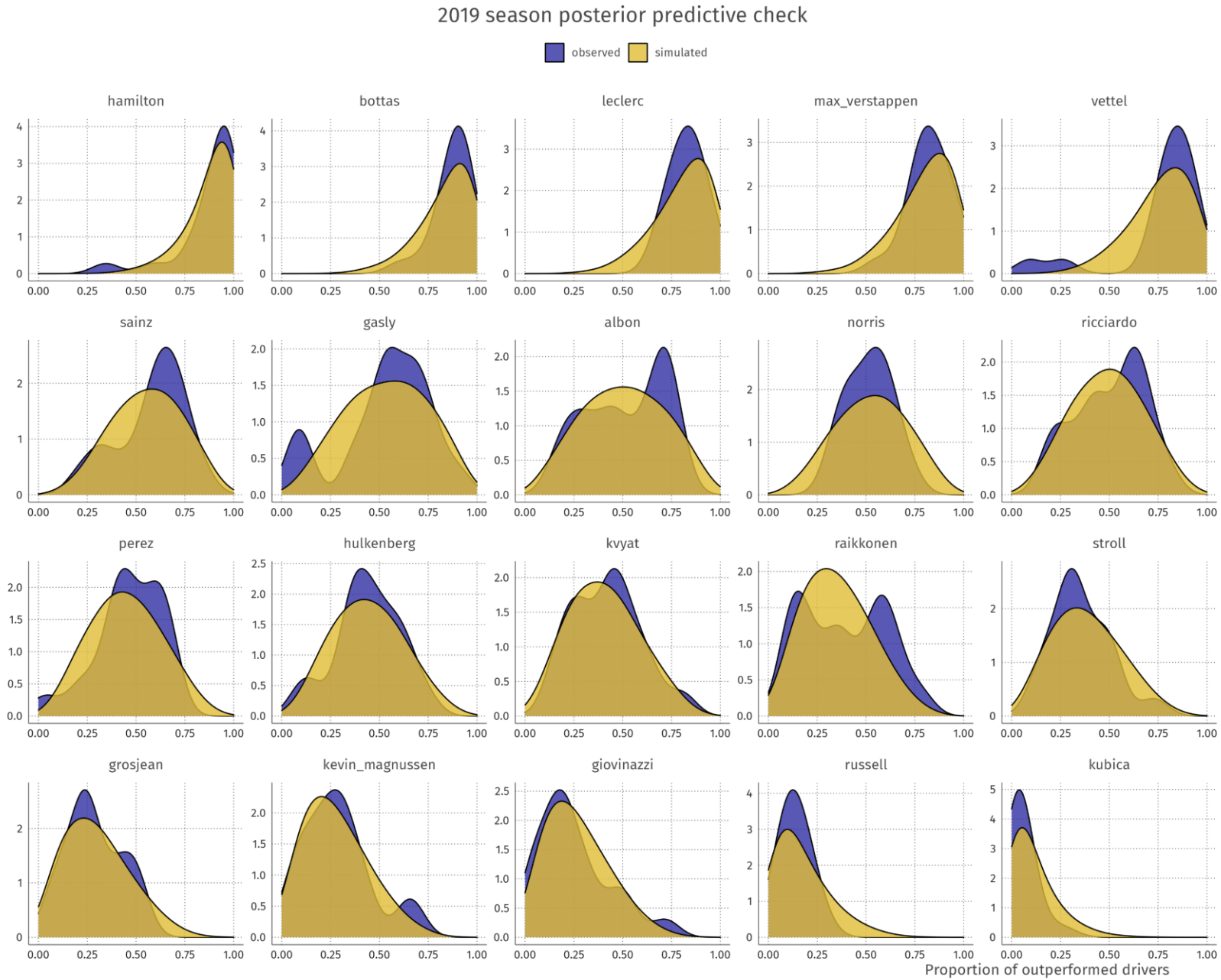
Obligatory convergence check



- 10k posterior samples across 4 chains
- 1000 burn-in iterations
- Everything looks good!
- Rhats all <1.01
- Etc.

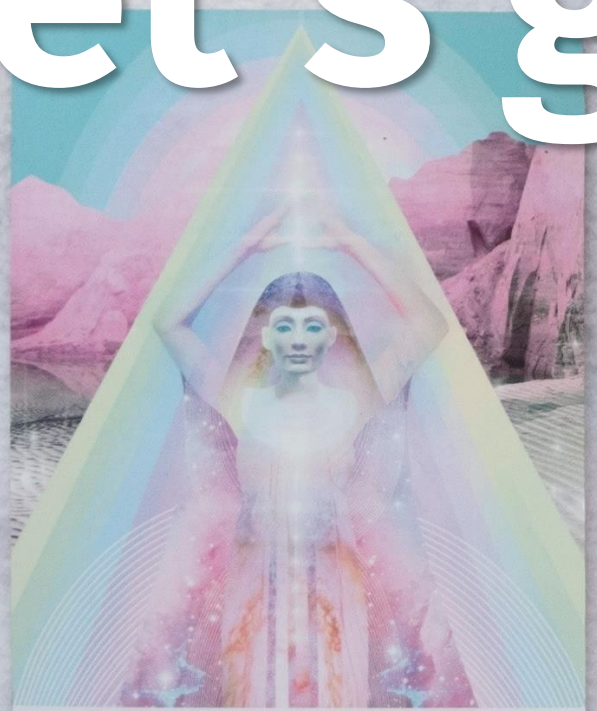
Posterior predictive check

- Does the model make sense?
- Compare observed proportion to model-implied proportion


2019 season posterior predictive check
observed
simulated
hamilton
bottas
leclerc
max_verstappen
vettel
sainz
gasly
albon
norris
ricciardo
perez
hulkenberg
kvyat
raikkonen
stroll
grosjean
kevin_magnussen
giovinazzi
russell
kubica
Proportion of outperformed drivers

Inference & conclusions

Let’s get mystical



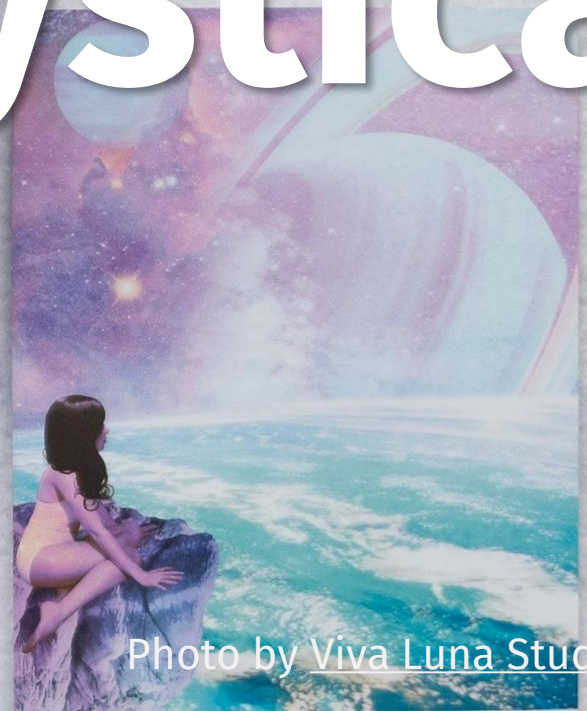
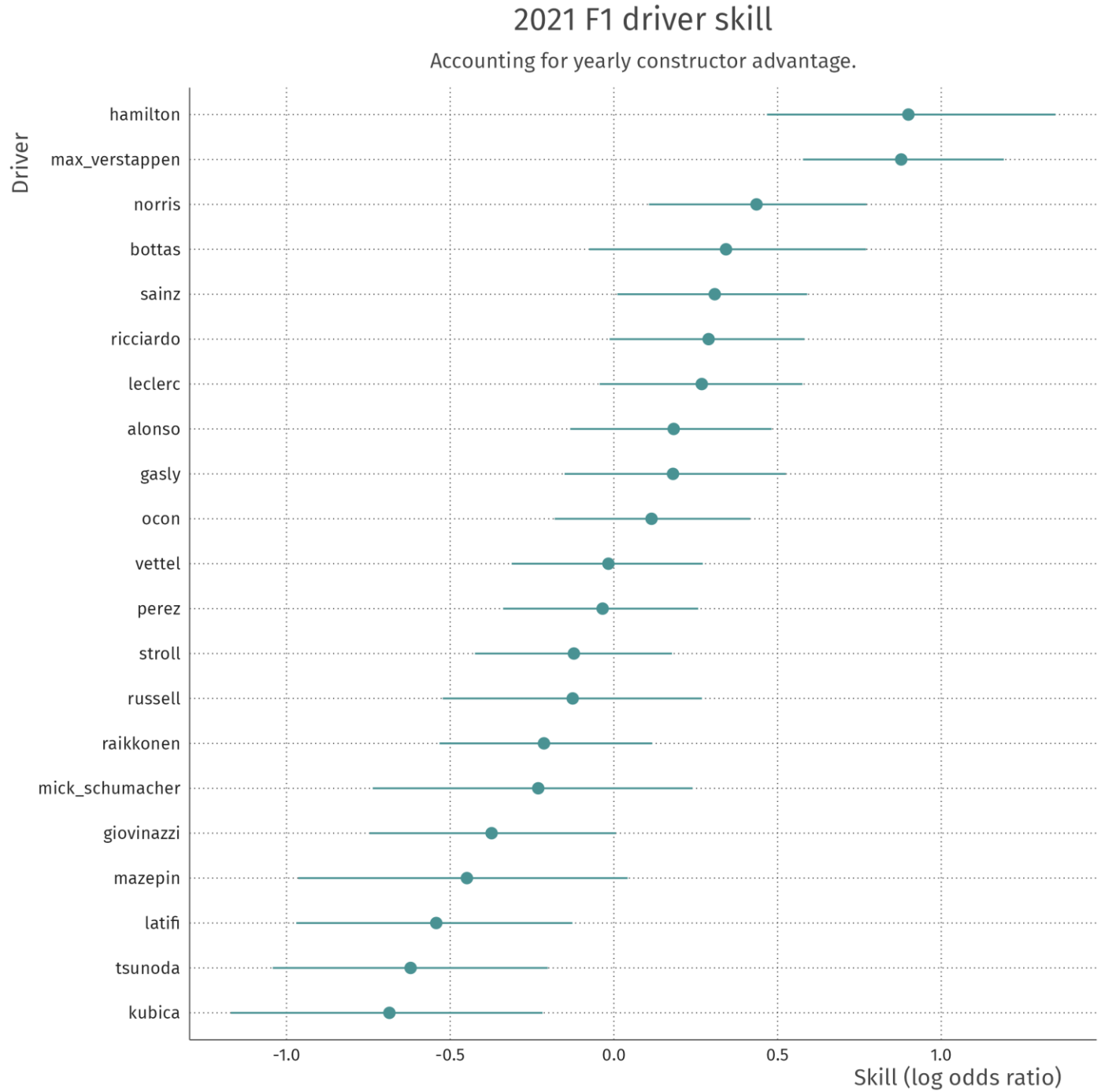


Photo by Viva Luna Studios on Unsplash

Dear oracle model,

How skilled was each driver in 2021?

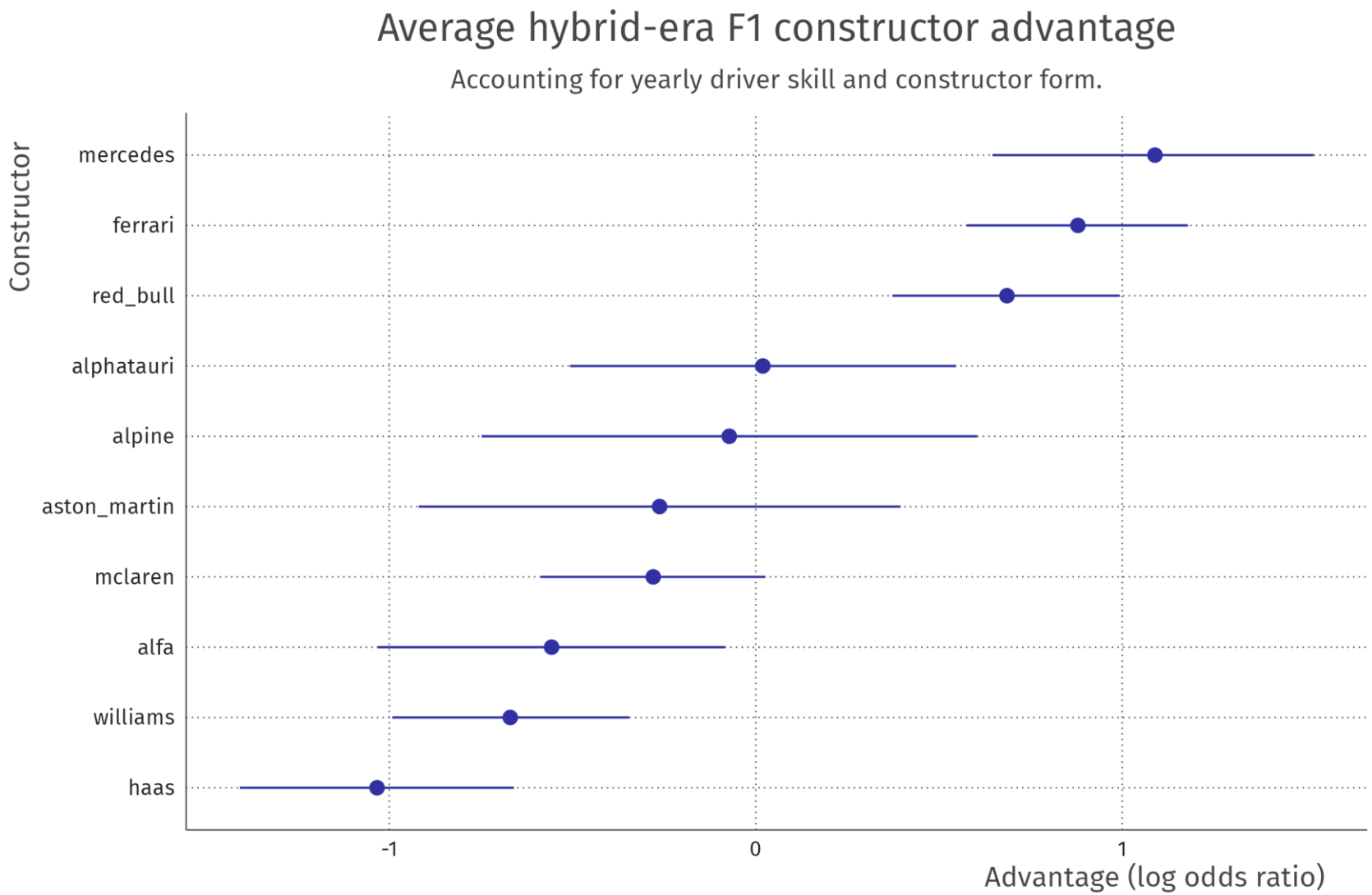

2021 F1 driver skill
Accounting for yearly constructor advantage.
Driver
hamilton
max_verstappen
norris
bottas
sainz
ricciardo
leclerc
alonso
gasly
ocon
vettel
perez
stroll
russell
raikkonen
mick_schumacher
giovinazzi
mazepin
latifi
tsunoda
kubica
-1.0
-0.5
0.0
0.5
1.0
Skill (log odds ratio)

According to the model:

- Hiring Hamilton as opposed to an average driver yields an odds-ratio of defeating other drivers of 2.46 [1.60 – 3.86]
- If your team with an average driver beats 50% of the competitors on average, then with Hamilton you will beat 71%
- 2.46 / (1 + 2.46) = 0.71 [0.61 – 0.79]

Dear oracle model,

How did the constructors compare in the hybrid era?


Average hybrid-era F1 constructor advantage
Accounting for yearly driver skill and constructor form.
Constructor
mercedes
ferrari
red_bull
alphatauri
alpine
aston_martin
mclaren
alfa
williams
haas
-1
0
1
Advantage (log odds ratio)

According to the model:

- If you are an average driver at an average team, moving to hybrid-era Mercedes yields an average odds-ratio of 2.97 [1.91 – 4.58]
- If you placed middle-of-the road before, you will now on average beat three quarters of the field!

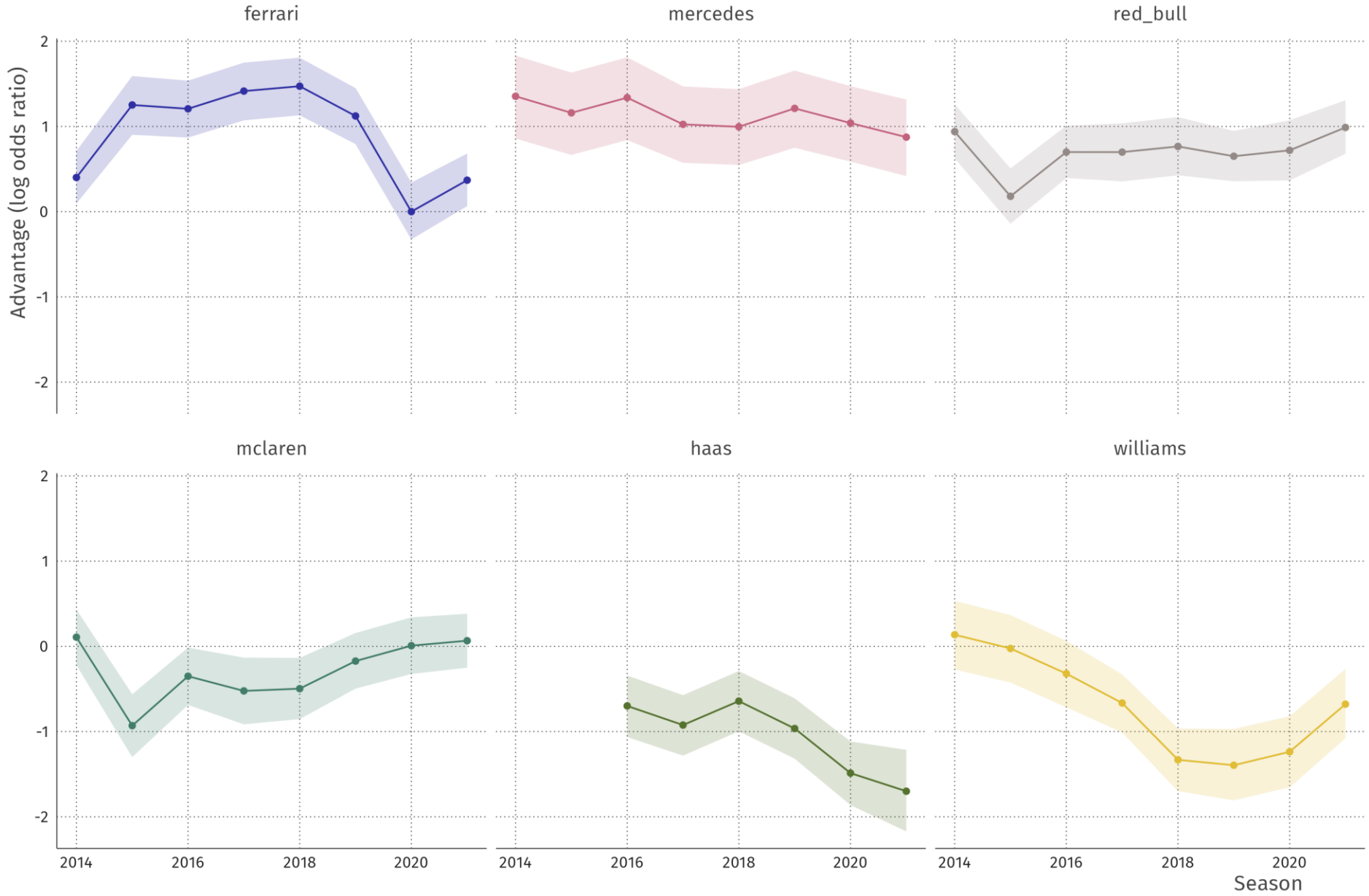
- 2.97 / (1 + 2.97) = 0.75 [0.66 – 0.82]

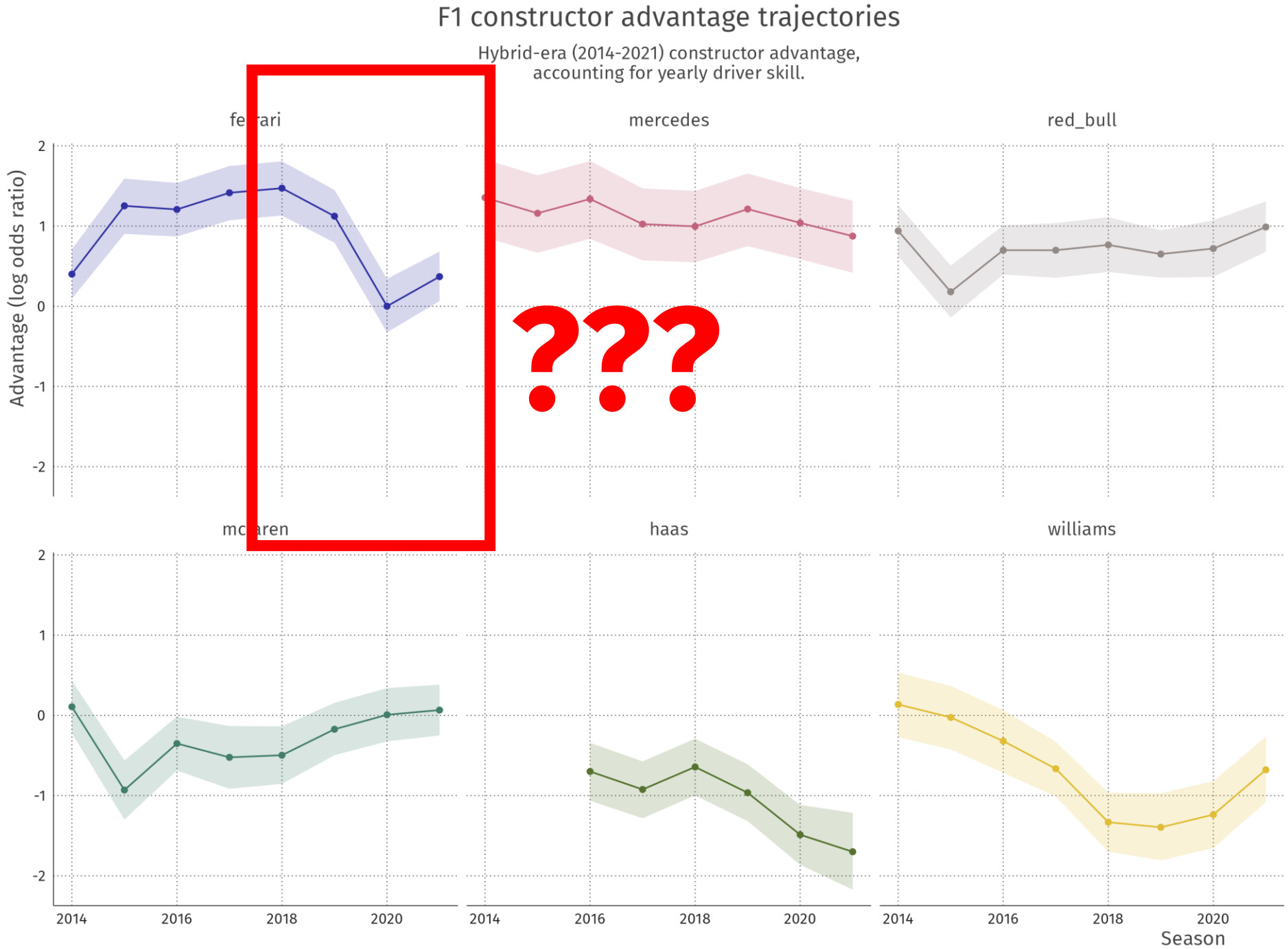
Dear oracle model,

How did this change over time?

F1 constructor advantage trajectories

Hybrid-era (2014-2021) constructor advantage,
accounting for yearly driver skill.




F1 constructor advantage trajectories
Hybrid-era (2014-2021) constructor advantage,
accounting for yearly driver skill.
ferrari
mercedes
red_bull
mclaren
haas
williams
???
Advantage (log odds ratio)
2
1
0
-1
-2
2014
2016
2018
2020
Season

FIA reaches 'settlement' with Ferrari following 2019 engine investigation

28 February 2020

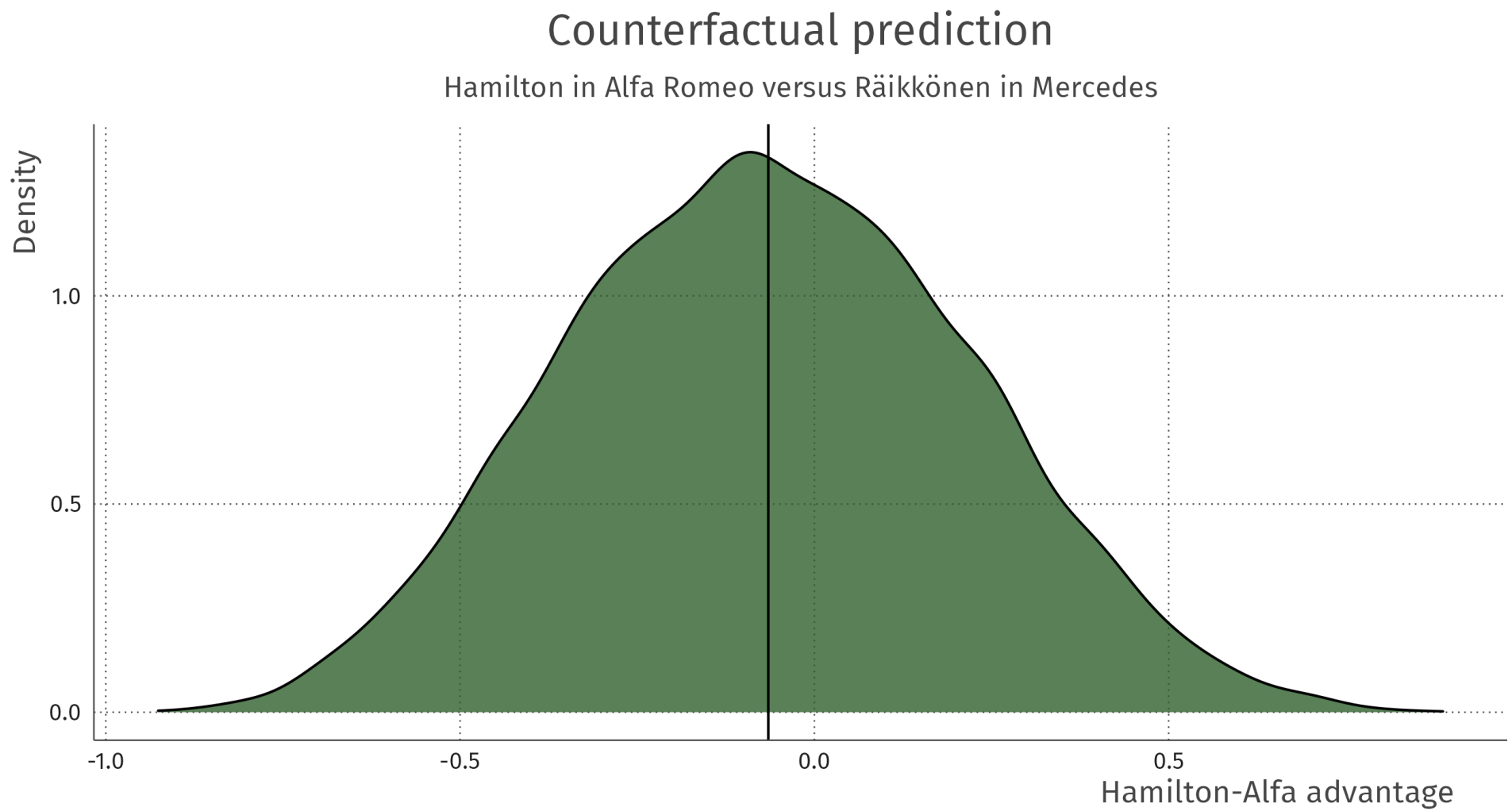
FIA FERRARI

Formula 1's governing body the FIA say they have reached a "settlement" with Ferrari following a long-running analysis into the Italian team's 2019-spec power unit.

Ferrari's power unit emerged as the class of the field in 2019, the red cars of Sebastian Vettel and Charles Leclerc displaying a power advantage over their rivals and enjoying impressive straight line speed.

Dear oracle model,

Would Hamilton in an Alfa Romeo car beat Räikkönen in a Mercedes in 2021?


Counterfactual prediction
Hamilton in Alfa Romeo versus Räikkönen in Mercedes
Density
1.0
0.5
0.0
-1.0
-0.5
0.0
0.5
Hamilton-Alfa advantage

Dear oracle model,

Is it weird to like Formula One?
Is Formula One a stupid sport?

Dear oracle model,

Is Formula One all about the car?

Car & driver contributions

Std. Dev.	Estimate	l-89% CI	u-89% CI
σ_c	0.95	0.69	1.29
σ_{cs}	0.43	0.34	0.52
σ_d	0.40	0.27	0.54
σ_{ds}	0.15	0.07	0.22

- Because components are all independent, we can do this:
  - $\sigma_c^2 + \sigma_{cs}^2 = 1.087$ (variance of car components)
  - $\sigma_d^2 + \sigma_{ds}^2 = 0.178$ (variance of driver components)
- Proportion of variance in performance due to the car:

$$\frac{1.087}{1.087+0.178} = \mathbf{86\%}$$

It’s not all about the car

Almost though... ☺

In conclusion...

- F1 performance model with open data
- Complex multilevel structure: races within seasons, drivers and constructors
- Measuring performance as proportion
- Bayesian Beta regression model with logit link:
  - Posterior predictive check
  - Estimates with uncertainty for many different quantities
- It’s not all about the car!

More info

All code + data:

`github.com/vankesteren/f1model`

Preprint:

`arxiv.org/abs/2203.08489`

Contact

e.vankesteren1@uu.nl
@vankesteren
t.l.g.bergkamp@rug.nl
@MisterBergkamp

Photo by Gustavo Campos on Unsplash

Questions?

Thoughts & issues

- Everything is relative & on logit scale just as with Elo score in chess
- Are parameters identified?
  - If driver stays with same team & teammate the whole time, we cannot disentangle $\beta_d, \beta_c, \beta_{cy}$!
  - How do the priors affect these?