Sapienza
Università di Roma

MSc Data Science

Bayesian Analysis of MotoGP race results to extrapolate rider's skill and constructor advantage

Statistical Methods for Data Science II

Professor:
Luca Tardella

Student:
Arturo Ghinassi

Academic Year 2021/2022

Contents

1 Introduction and Goals of the Analysis

The goal of this paper is to implement a Bayesian model that is able to quantify and discern, in a sports racing context, the skill of the rider from the advantage given by the constructor. To be able to do that we will use a multilevel Beta regression that models the individual race success as the proportion of outperformed competitors, as described in van Kesteren and Bergkamp, 2022. However, the purpose of the model will not be to predict the results of the race but we will focus on the a posteriori coefficients referred to the strength of a rider and of a bike.

We would like to demonstrate, by looking at the explained variance of the coefficients, how in the case of MotoGP the real difference is made by the rider rather than by the bike unlike F1 where the car has a bigger impact. Two different models will be used: a basic one and one that takes into account also the race weather.

The outcome of a race, or better the proportion of outperformed competitors (that we will refer as **POC**), is modeled as the combination of the riders ability and the bike strength. Of course in such extreme sports there are some intangible factors that cannot be estimated through a model, such as the rider's feeling with the bike or the feeling with the team, that have a huge impact on performance. For these reasons we can expect that the model may sometimes have imperfections in the posterior predictive part due to mathematically inexplicable factors.

The benchmark results obtained by Bell et al. in 2016 and van Kesteren and Bergkamp, 2022 state that the constructors effects accounts for around 86% of the variance in the model in the case of F1. What we expect instead from the model applied on our data is a drastic reduction of this percentage in favour of the riders ability.

2 Data preparation

We applied the model to the MotoGP 2016-2021 seasons data that were scraped from the official MotoGP web page and available at the following link as *csv* files.

The columns of interest are **year, race sequence, rider, constructor, position and weather**, plus we added the columns of the POC and the smoothed POC (the target variable of the model). We then removed the non-finishing races for each rider and ended up with a total of 1839 records.

In MotoGP often teams change name when they change sponsor even if the team is actually the same. Fpr this reason we renamed the Teams with a unique name in the case in which the difference was given only by a change in the sponsor of the team: for example we renamed *Ducati Team, Ducati Lenovo Team and Mission Winnow Ducati* all as *Ducati*.

This are the first five records of the dataset:

	year	sequence	rider	constructor	position	weather	POC	POC smooth
1	2016	1	Jorge Lorenzo	Yamaha Factory	1	Dry	1.00	0.97
2	2016	1	Andrea Dovizioso	Ducati	2	Dry	0.93	0.90
3	2016	1	Marc Marquez	Repsol Honda Team	3	Dry	0.86	0.83
4	2016	1	Valentino Rossi	Yamaha Factory	4	Dry	0.79	0.77
5	2016	1	Dani Pedrosa	Repsol Honda Team	5	Dry	0.71	0.70

3 EDA

First of all let's have some insights on the data. As we can see in **Fig. 1** there are a consistent number of races where not all the riders arrived at the end line (22 riders) in fact in MotoGP crashes are far more frequent than in F1.

Even if the x-axis goes up to 23 we still considered only 22 riders because in MotoGP often there are some riders called "Wildcards" that did just a few number of races that we have not considered in the model.

We also took a look at the distribution of the wet races and as we can see the number of wet races during the 5 years considered is around the 10%.

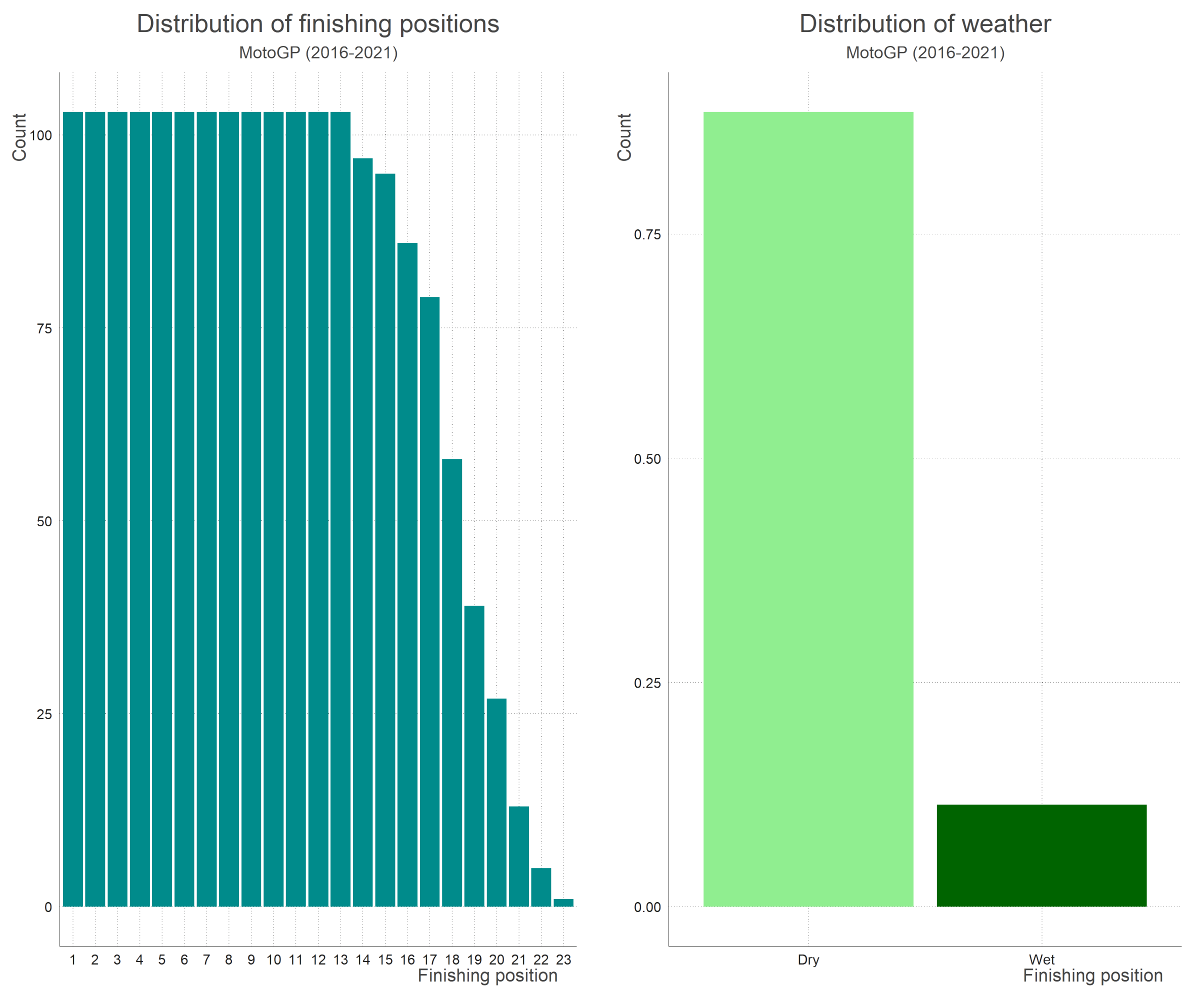

Figure 1: Distribution of the finishing positions from 2016 to 2021

We can also take a look at the distribution of the smoothed proportion of outperformed competitors and its evolution in time for three specific drivers: Andrea Dovizioso, Jack Miller and Marc Marquez.

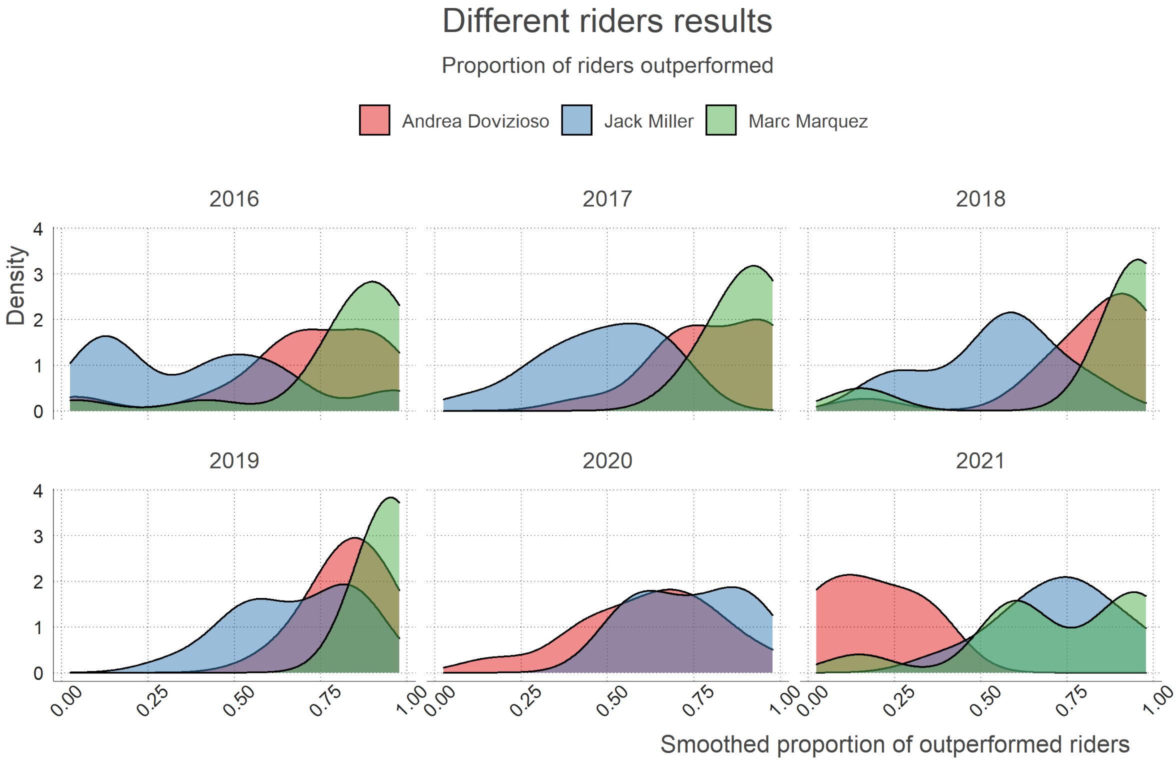

Figure 2: Distribution of the smoothed proportion of outperformed riders from 2016 to 2021

Is interesting how the distribution of Andrea Dovizioso and Jack Miller changed, in fact in 2020 Dovizioso left Ducati and his score dropped while the constant improve of Miller brought him to join Ducati in 2021 and we can see how his score increased, in fact as we will see Ducati is one of the best constructors of the last 6 years and this explains this change in distributions.

4 Proposed Model

The proposed model is a multilevel Beta regression to estimate the smoothed POC but, as said before, what we are more interested in is the mean of the Beta distribution that is obtained as a sum of the rider skill and constructor advantage.

4.1 Basic Model

For each rider r and for each constructor c we specify two parameters: the long term skill/advantage and the seasonal one.

$$y_{rcs} \sim Beta(\mu_{rcs},\ \phi),\ \phi = dispersion$$

$$\mu_{rcs} = \beta_r + \beta_{rs} + \beta_c + \beta_{cs}$$

$$\beta_r \sim N(0, \sigma_r^2)$$

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

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

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

The parametrization of the Beta uses a dispersion parameter ϕ and a mean parameter μ_{rcs} instead of the standard α and β parametrization. The reason is because with this parametrization the model allows us to easily interpret the mean. The logit link function ensures that the average driver at an average team with an average seasonal form will on average have $\mu_{dcs} = 0$, which translates into a probability of 0.5 of beating other drivers:

$$logit(\mu_{rcs}) = \frac{1}{1 + e^{-\mu_{rcs}}} = 0.5, \quad \mu_{rcs} = 0$$

.

We will use these parameters in order to rank the riders with respect to their ability and the constructor with respect to their advantage. The single parameters represent deviations from the average case, so negative values mean worse than average skill/advantage, and positive values mean better than average skill/advantage. The seasonal parameters that we also include instead represent yearly deviations from the long-term average skill/advantage. For all the previous explanations it is clear that we only partially care about the output variable because the real focus is on the skill and advantage parameters for each rider and each team. It is implicit in the specification of the model that a rider's skill is independent of the constructor advantage, in fact the rider skill does not change when the rider moves to a different constructor. The main problem with this approach is that we are not considering that in motorsport may happen that a rider's riding style is better suited to one bike than another so the assumption that is made is very strong and sometimes limiting.

Since we have no particular beliefs about what the real skill is but we rather want the model to tell us for each model we will use the default priors.

4.2 Enhanced model

In previous works it has been shown that several predictors may change the race results (Bell et al., 2016). For this reason we propose a variation of the original model that considers also the weather conditions of the race.
The reason behind this choice lies in the fact that racing in wet conditions requires greater skill on the part of the rider. In fact, it often happens that in wet race conditions the finish ranking reserve some surprises, this is due both to a greater probability of accidents but also above all to the individual skills of the riders in knowing how to adapt in uncommon conditions.
We therefore expect that by inserting the variable of the conditions in which the race took place (dry or wet) the model will be able to better discern the individual abilities of the rider given that in those conditions the result will rely less on the bike and more on the rider. Following the paper approach we represent this knowledge by splitting the rider average skill parameter into a random intercept parameter γ_{0r} and a random slope parameter γ_{1r} as follows:

$$\beta_r = \gamma_{0r} + \gamma_{1r} \cdot weather$$

Considering the weather as 0 for a dry race and 1 for a wet race we can resume the rider's ability like this:

$$\beta_r = \begin{cases} \gamma_{0r} & ;\ if\ dry \\ \gamma_{0r} + \gamma_{1r} & ;\ if\ wet \end{cases}$$

4.3 Model comparison

To compare the models we used efficient leave-one-out cross-validation (LOO) that computes the expected log posterior density (ELPD) for each model, which is an alternative to the standard information criteria in Bayesian model comparison and is suggested in the case of weak priors. In terms of ELPD the weather model seems to be, as we expected, much better than the basic. This result is completely different from the one of the reference paper where the basic model resulted as the best one. This may be due to the fact that the model that manages to better capture a dominant factor such as the skill of the rider gets better results than the basic one. However, it must be taken into account that, since ours is not a predictive model, this comparison only gives us a general idea. To really understand which model is the best we will use the results obtained.

	$ELPD_\Delta$	se_Δ	ELPD	se_{ELPD}
Weather	0.00	0.00	626.57	32.56
Basic	-33.02	10.49	593.55	43.05

5 Basic Model

The first model implemented is the basic one in which we don't take into account the weather conditions as feature to infer the rider's skill.

The model was estimated using the software package *brms* with the default priors for all parameter types. We used 4 Monte Carlo Markov Chains with 10000 iterations and a fixed burn-in of 1000 observations.

As said before we used a Beta family regression with zero intercept, the parameters we have to estimate are the standard deviations for the riders skill and the constructor advantage and the family dispersion parameters.

The model will output not only the values of skill and advantage for each rider, season and constructor but also the standard deviations of the distributions of the parameters that are the ones we are more interested in in order to evaluate the impact.

```r
fit_basic <- brm(
  formula = prop_trans ~ 0 + (1 | rider) + (1 | rider:year) +
                             (1 | constructor) + (1 | constructor:year),
  family = Beta(),
  data   = mgp,
  backend = 'cmdstanr',
  control = list(adapt_delta = 0.9),
  chains = 4,
  cores  = 4,
  warmup = 1000,
  iter   = 10000,
  save_pars = save_pars(all = TRUE))
```

5.1 Convergence check

First of all we have to check the convergence of the chains, to do so we can take a look at the trace plots and see if they arrive at a steady state. The summary of the model suggests also to look at the Rhat value for each estimated parameter, a value smaller than 1.05 ensures convergence.

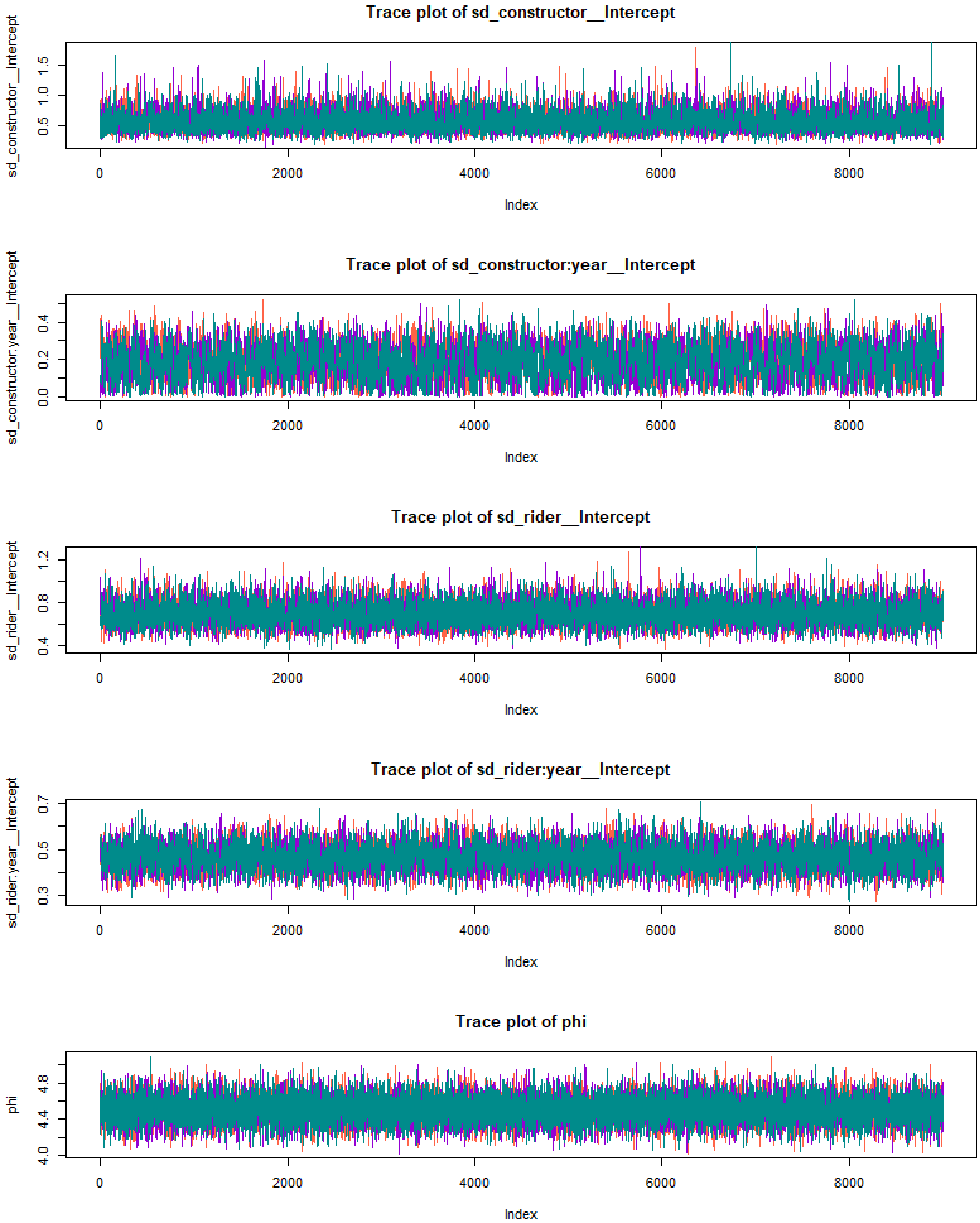

Figure 3: Chains trace plots for each parameter to estimate

As its clear from the plots all the chains are converging and overlapping as expected from an appropriately converged model. Thus can be seen also better by smoothing the traces through the running means.

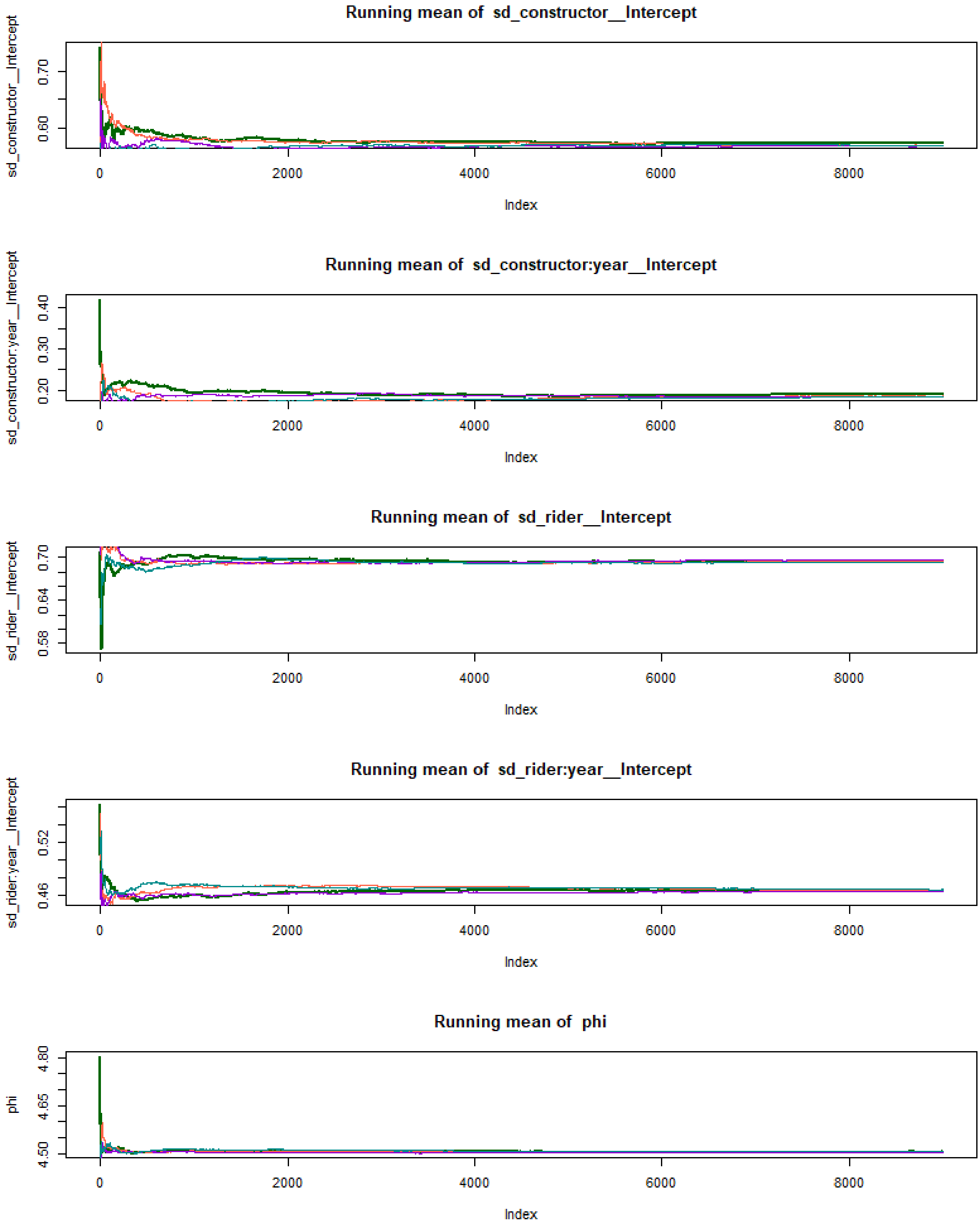

Figure 4: Running means

We also check the Rhat hist of the parameters, if the distribution is converging the value is smaller than 1.05 and this condition is respected by the model as we can see from the histogram.

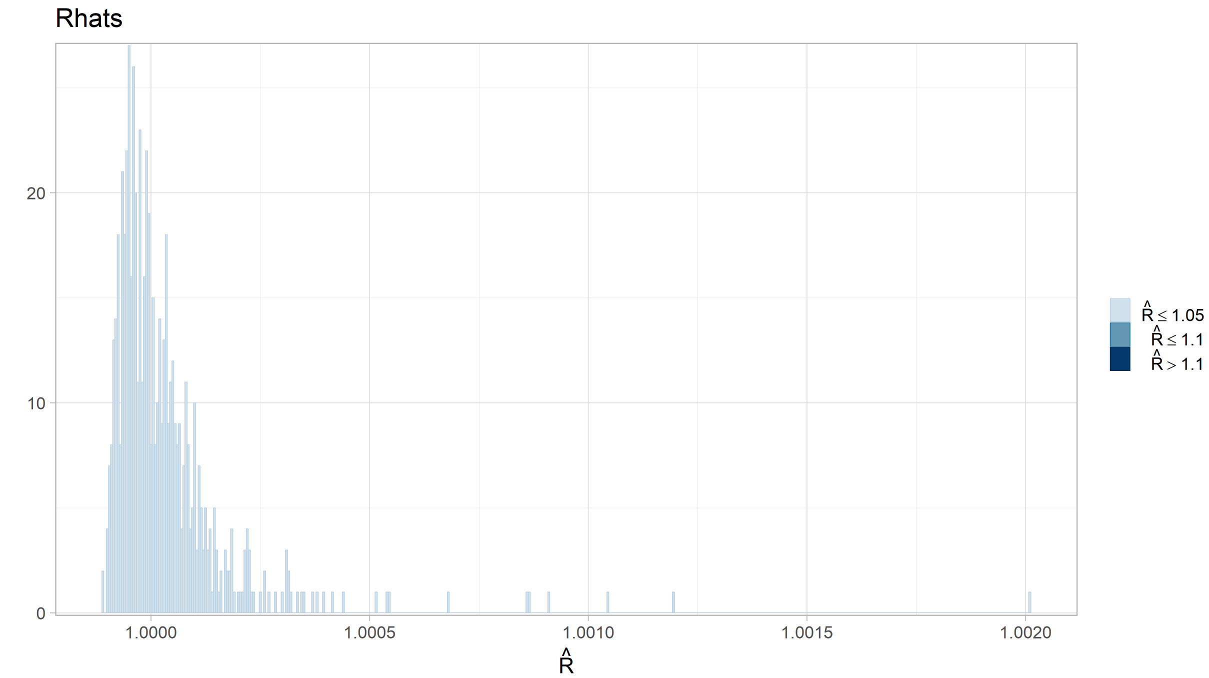

Figure 5: Diagnostics

Let's also take a look at the effective sample size for the most relevant parameters, the model suggests at least a value of 100 both for ESS Tail and Bulk for each chain:

	Rhat	Bulk ESS	Tail ESS
constructor	1.00	15023.84	21437.70
constructor form	1.00	3998.58	8788.75
rider	1.00	12990.39	21354.89
rider form	1.00	8529.72	18344.03

5.2 Posterior Check

Next we want to check the posterior predictive distributions both on the POC and the rank scale, we've only considered riders that completed more than one fourth of the races (5 races). In order to simulate data on the rank scale, the proportion of drivers beaten needs to be back-transformed into a rank with the following procedure:

1. For each driver-team-year under consideration, obtain a posterior sample of the "proportion of drivers beaten"

2. Consider these proportions as expectations for the relative finishing positions: transform these proportions into ranks by sampling each rank with a given probability for each driver accounting for the probability distribution of the finishing positions.

Starting from the POC scale we can see how for the majority of the riders the posterior distributions are overlapping quite well but for some riders such as Zarco, P.Esaprgaro, Morbidelli and Mir are not overlapping really well, how can we explain this?

Simply by going to better "study" what happened that season. If we look at Zarco's results in the previous seasons we can see that riding a Yamaha (a strong bike) he always obtained podiums and scored good points while in 2019 he had no podiums and moved from Yamaha to KTM (weaker bike), so the model considers Zarco a skilled rider (long right tail) but the results of the season were poor and the same reasoning can also be made for P. Espargaro.

More interesting are the exploits of Morbidelli and Yamaha Petronas that had a surprising season in 2019, and also the case of Mir that at his first season in MotoGP was able to score many points and obtain high placements.

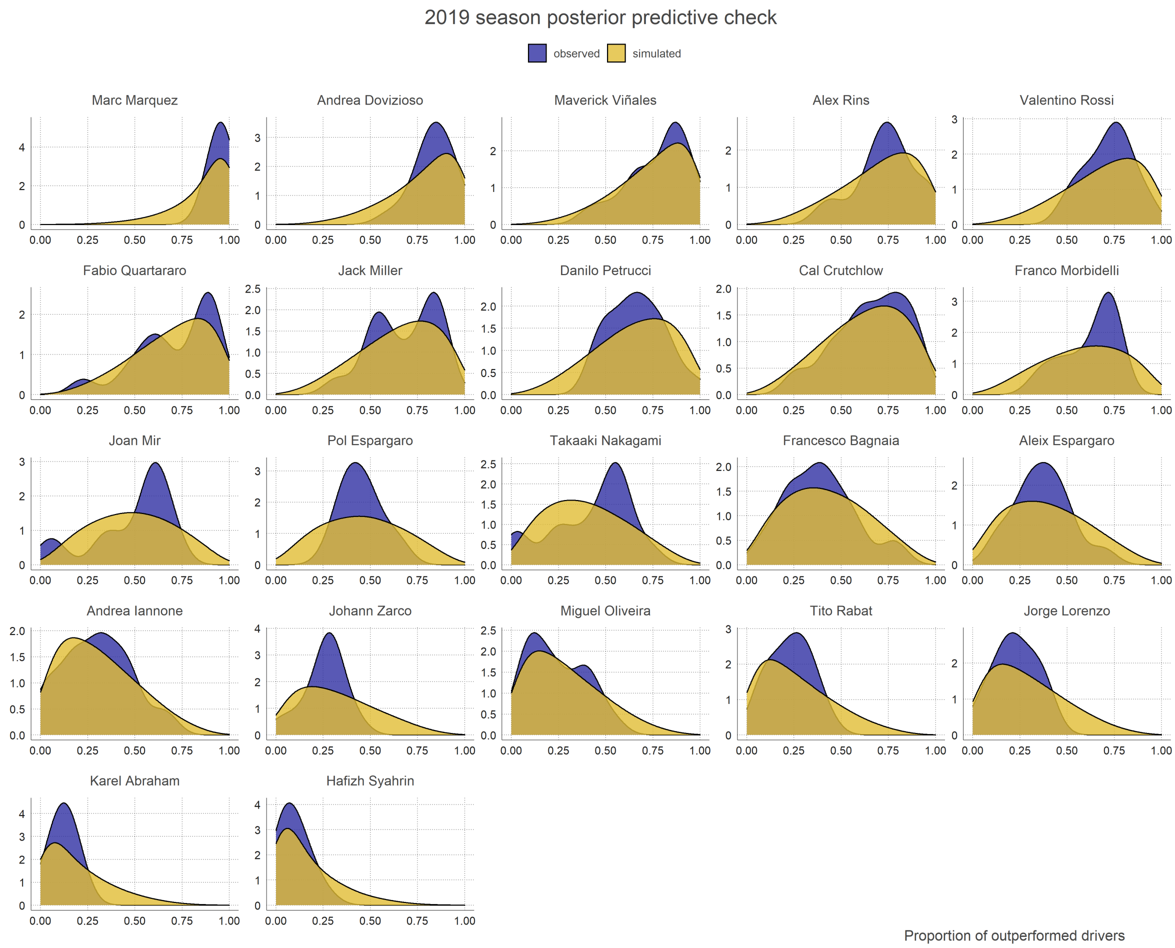

Figure 6: Posterior predictive check on the POC scale

Moving now to the rank scale the model performs quite well for all riders except for Jorge Lorenzo. Also in this case we can explain the "error" deepening its season.

In 2019 Lorenzo moved from Ducati to Honda after obtaining the previous season 4 podiums and 3 wins but moving to Honda he never arrived in Top 10 because he never found the right feeling with the bike and suffered this heavily during the season.

In poor words this means the model had higher expectations given the rider skill but for reasons that the model can't catch this expectations were not met.

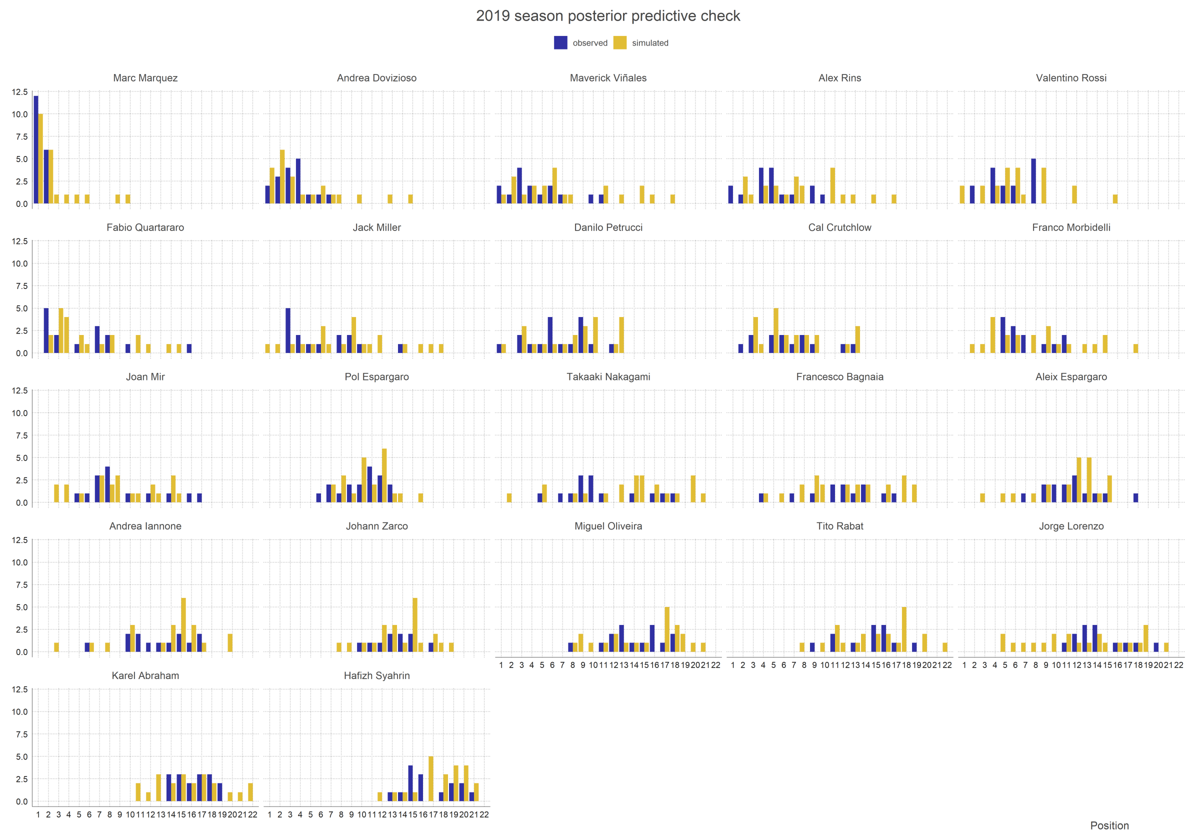

Figure 7: Posterior predictive check on the rank scale

5.3 Model inference

To make inference we focused on the 2021 season and tried to infer which rider was the most skilled during the season accounting also for the constructor advantage.

5.3.1 Inference on rider skill

In order to produce a riders ranking, we obtained the posterior means and 89% credible intervals of $\beta_r + \beta_{rs}$ for the season 2021.

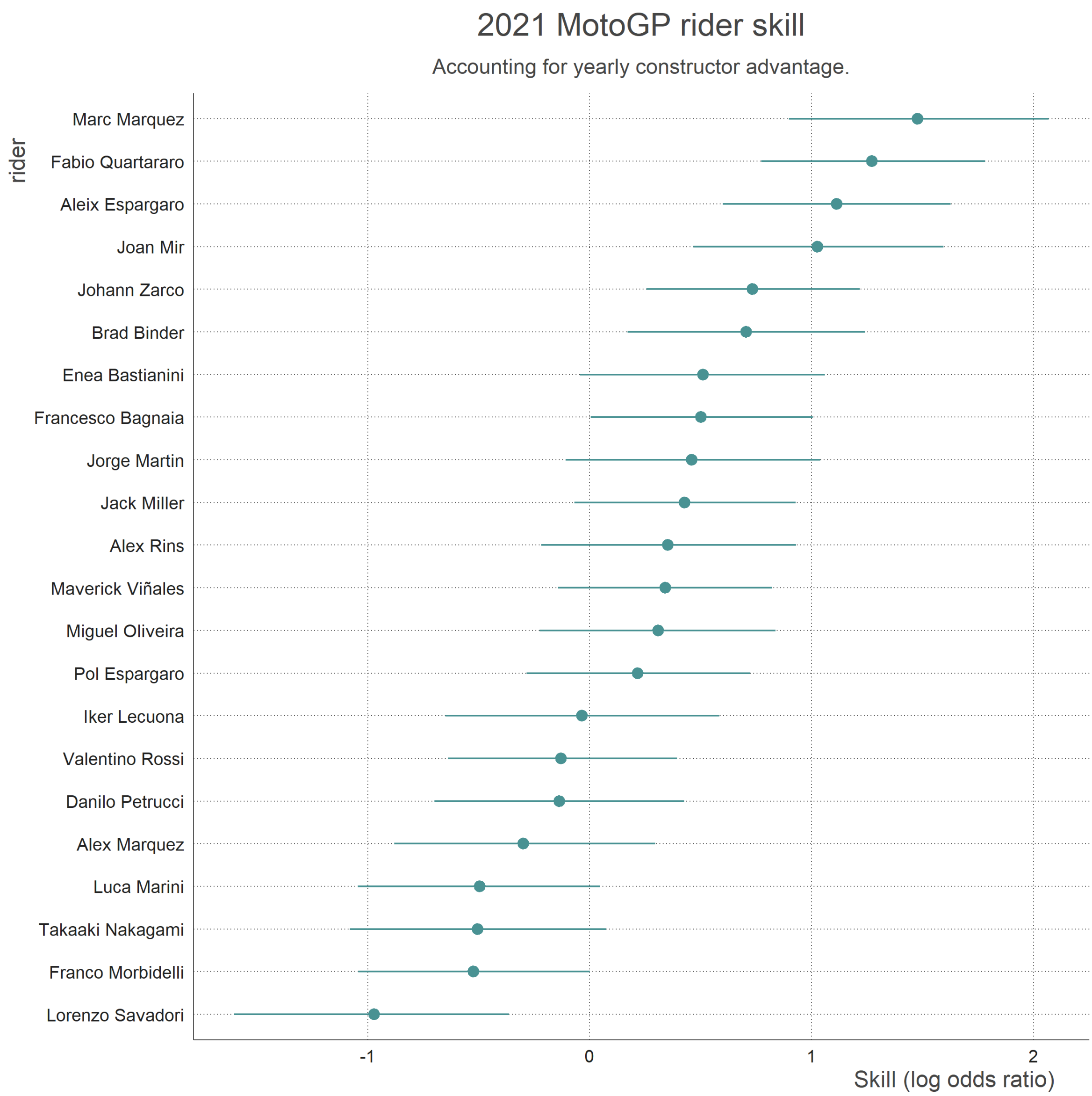

Figure 8: Riders ranking for 2021 season

Could be surprising that Marc Marquez (7th position in the final standings) is ranked above Fabio Quartararo (2021 World Champion) but in this case the model is not wrong, on the contrary it would seem to have fully grasped the skill of the riders. In fact before suffering a terrible injury in 2020 that forced him to end the season, Marc Marquez was without doubt the best rider on the grid competing in all races for the podium (in 2019 he registered 12 wins out of 19 races) and moreover doing this on a Honda that we will see later is not one of the best bikes.

Coming back in 2021 after a year of stop he definitely won the comparison with his teammate Pol Espargaro that has got only one podium in 17 races while Marc got 4 with 3 wins in 14 races still suffering for the repercussions of the injury.

Another interesting plot is about the riders skill trajectory in the different seasons where we can clearly see what we have said before about Marquez strength and Lorenzo's expectations for 2019. We focused in particular on Jorge Lorenzo, Andrea Dovizioso, Valentino Rossi, Fabio Quartararo, Marc Marquez and Maverick Viñales:

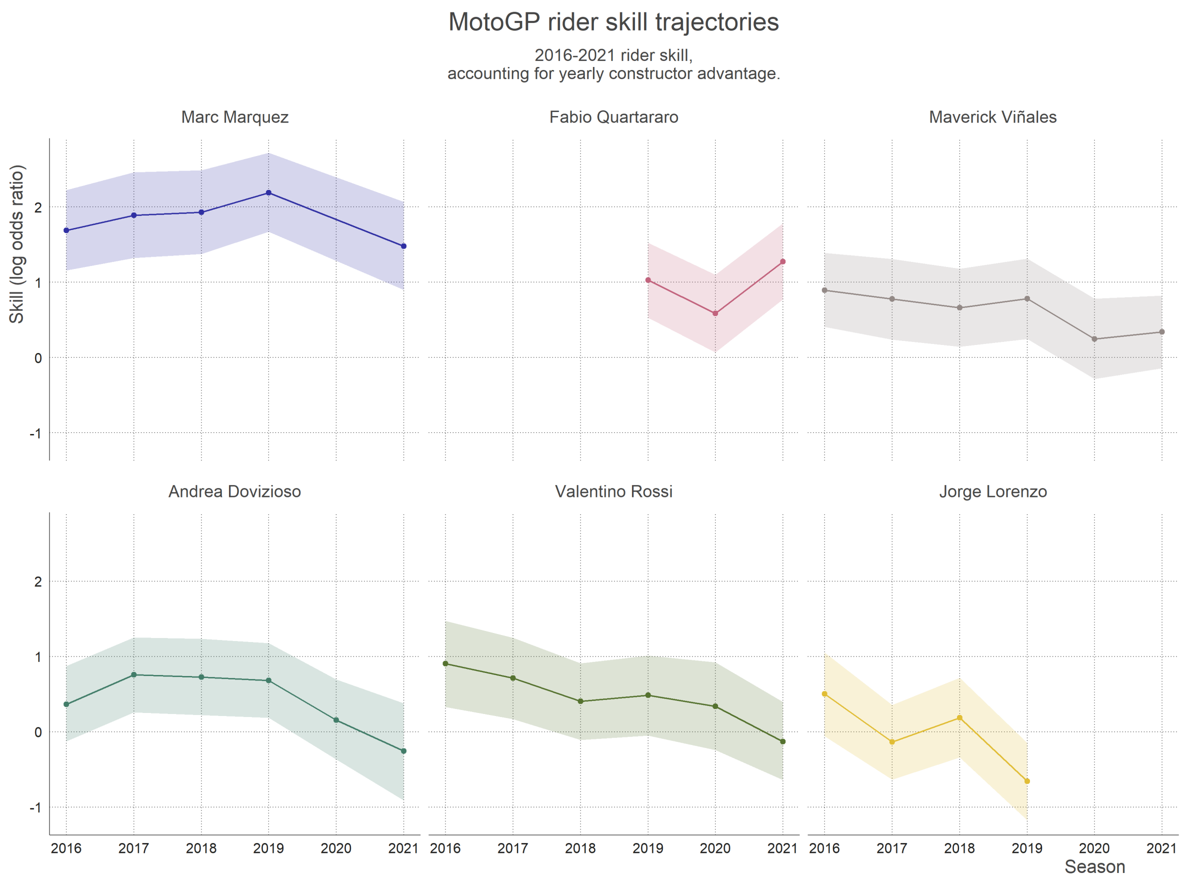

Figure 9: Riders skill trajectories

It is clear how riders at the end of their career have a downward trajectory, while drivers such as Quartararo that is expected to be the next "big thing" in motorsport have an upward trajectory.

5.3.2 Inference on constructors advantage

In the case of the constructors we focused on: Repsol Honda Team, Ducati, Team SUZUKI ECSTAR, Yamaha Factory, Aprilia Racing Team, LCR Honda.

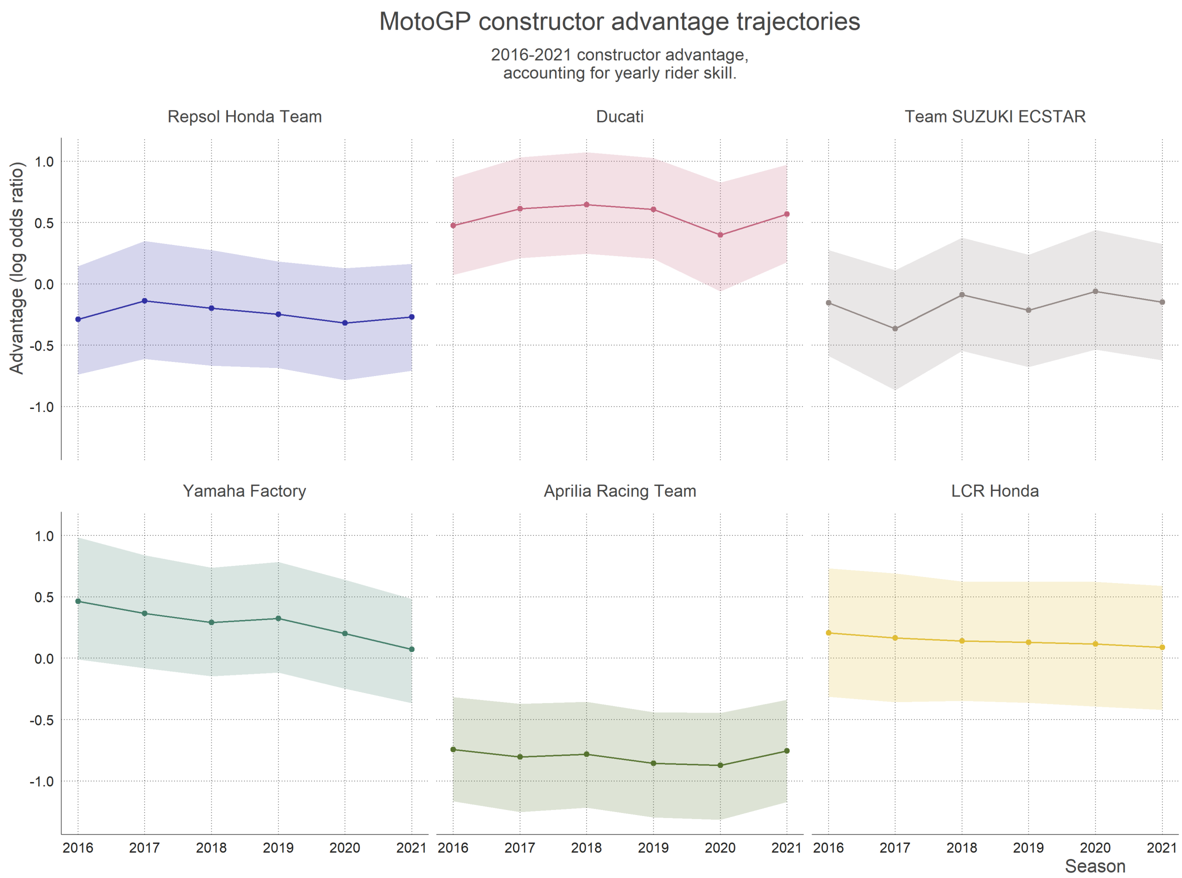

Figure 10: Riders ranking for 2021 season

The model faithfully reproduces reality, in fact Ducati and Yamaha have always been among the strongest constructors in recent years. For Honda, on the other hand, we can be satisfied with the result as in this case the bike turned out to be strong only when it was driven by Marquez and terrible with other riders.

The most interesting result is the one concerning the LCR Honda team which in fact ranked third in the 2021 ranking below:

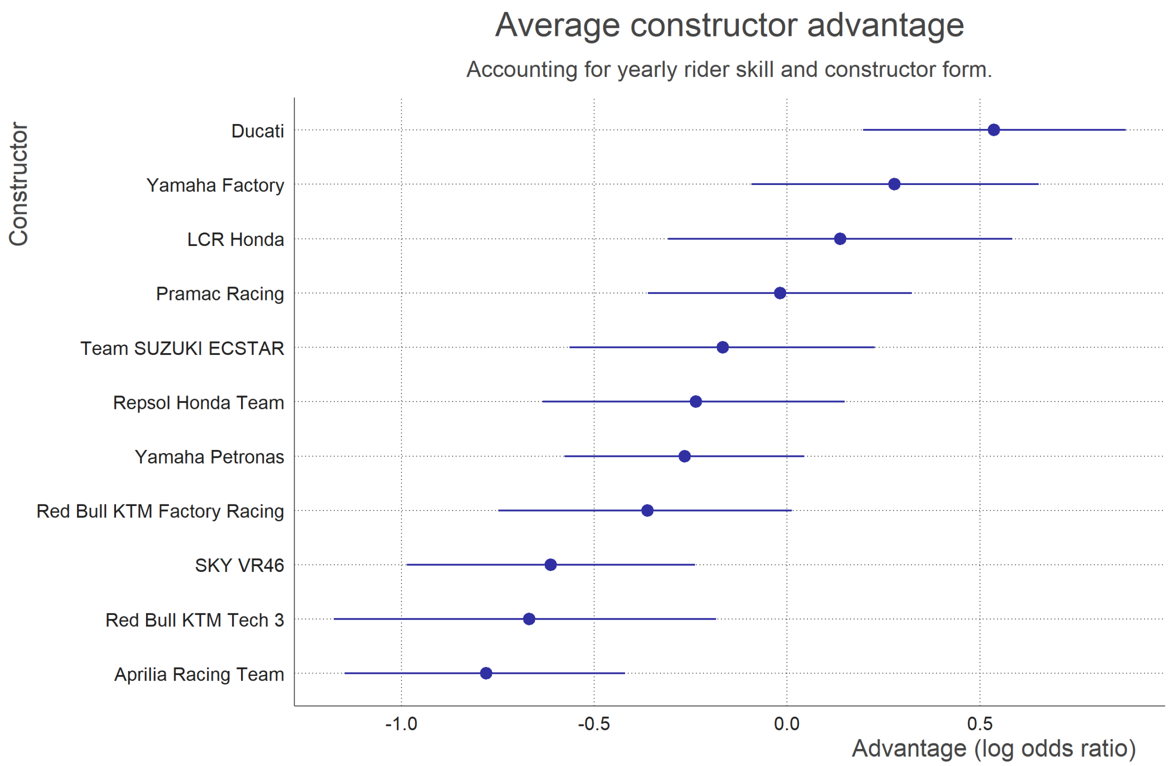

Figure 11: Riders ranking for 2021 season

The reason of such a high rank could be found in the fact that below average riders (Nakagami and A. Marquez) were able to reach often the Top 10 or placements close to the podium.

5.3.3 Relative contributions of riders and constructors

In order to investigate the contributions of drivers and constructors to the race results, we look at the summary of the model that outputs the estimates for the standard deviations and the MC errors ($\hat{\sigma}_{posterior}/\sqrt{N_{eff}}$) for all chains merged together and we focus on the riders and constructors parameters. The obtained results in terms of estimation error are in line with the ones obtained in the paper, we can see this as a proof that the model work correctly.

	Estimate	Est.Error	LB	UB
constructor	0.57	0.16	0.36	0.85
constructor form	0.19	0.09	0.03	0.33
rider	0.70	0.11	0.54	0.88
rider form	0.46	0.06	0.38	0.56

The standard deviation of the rider is larger than that for the constructor. This means that on average, the rider has a larger impact on race results than the bike.

To quantify the relative importance of long-term constructor advantage compared to rider skill we get it directly from the numerical summaries. The posterior estimates for the variances are shown below:

	Estimate	LB	UB
constructor variance	0.34	0.23	0.44
rider variance	0.66	0.56	0.77

Following the methodology of Bell et al. that has been applied in the paper, this means that rider effects account for around 66% of the variance in the model, which is quite the opposite of the results obtained by van Kesteren and Bergkamp where the contribution of the car was 86%. This is exactly the result we wanted to obtain as it confirms our initial hypothesis.

An interesting fact is that the two parameters with the highest correlation are the form parameters with a negative correlation equal to 0.4.

	rider	constr.	rider form	constr. form
rider	1.00	-0.04	-0.02	-0.12
constr.	-0.04	1.00	-0.03	-0.04
rider form	-0.02	-0.03	1.00	-0.40
constr. form	-0.12	-0.04	-0.40	1.00

This comes directly from the definition of the model. The latter in fact considers the seasonal outcome as a sum of complementary contributions, so if the rider's contribution increases, a reduction in the merit of the bike follows. This is not correct generally where instead one might think that the two coefficients are positively rather than negatively correlated but makes perfectly sense for the model.

The overall performance of the 2021 season combining riders ability and constructors advantage is shown in the following plot and except for few riders (with not great surprise, Marquez) it follows the actual ranking at the end of the season:

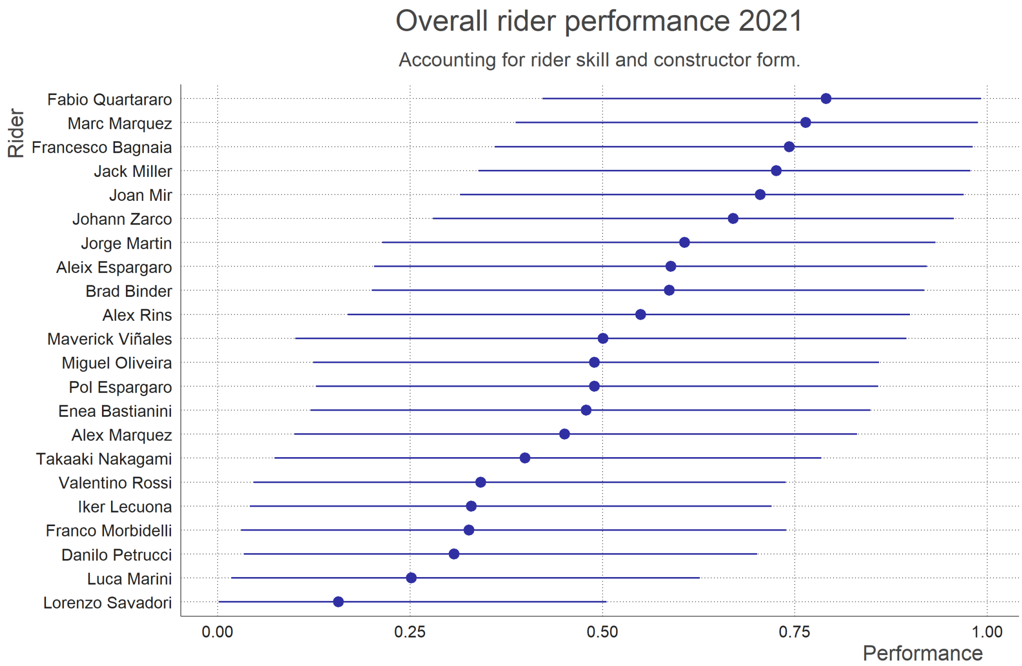

Figure 12: Overall performance for 2021 season

	rider	constructor	position
1	Fabio Quartararo	Yamaha	1
2	Francesco Bagnaia	Ducati	2
3	Joan Mir	Suzuki	3
4	Johann Zarco	Ducati	4
5	Jack Miller	Ducati	5
6	Marc Marquez	Honda	6
7	Brad Binder	KTM	7
8	Aleix Espargaro	Aprilia	8
9	Maverick Viñales	Aprilia	9
10	Miguel Oliveira	KTM	10
11	Alex Rins	Suzuki	11
12	Pol Espargaro	Honda	12
13	Enea Bastianini	Ducati	13
14	Jorge Martin	Ducati	14
15	Takaaki Nakagami	Honda	15
16	Alex Marquez	Honda	16
17	Franco Morbidelli	Yamaha	17
18	Iker Lecuona	KTM	18
19	Danilo Petrucci	KTM	19
20	Luca Marini	Ducati	20
21	Valentino Rossi	Yamaha	21
22	Lorenzo Savadori	Aprilia	22

5.3.4 Counterfactual inference

Using samples from the posterior distributions of the parameters, we can answer some counterfactual questions about the riders in the model. The main approach for this is by comparing the POC for different configurations of the predictors, we considered a race in 2018 where Dovizioso is competing for Yamaha and Marquez for EG VDV and computed the posterior distribution of the difference δ of the predicted POC for Dovizioso in a Yamaha and Marquez in a EG Marc VDS:

```
dovizioso_yamaha <- posterior_predict(fit_basic, tibble(
  year = 2018, constructor = "Yamaha Factory", rider = "Andrea Dovizioso"))

marquez_EG <- posterior_predict(fit_basic, tibble(
  year = 2018, constructor = "EG 0,0 Marc VDS", rider = "Marc Marquez"))

delta <- marquez_EG - dovizioso_yamaha
```

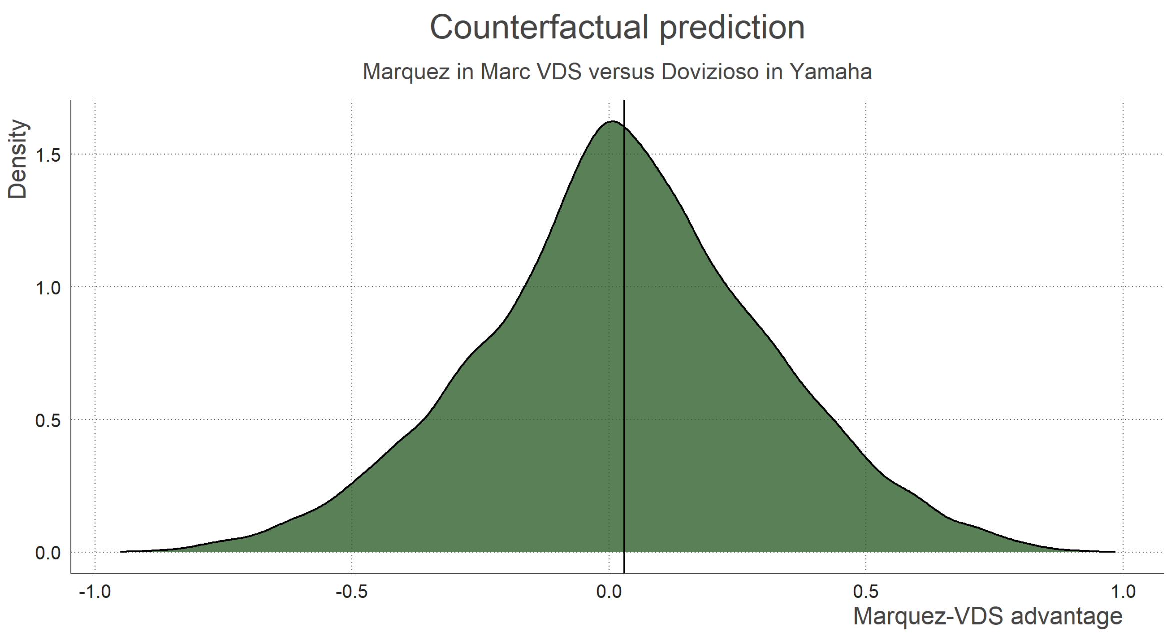

Figure 13: Counterfactual prediction, Dovizioso in Yamaha vs Marquez in VDV

It is not surprising, given the impact of the rider skill on the final race outcome, that even on a high performing bike Dovizioso in still not able to beat Marc Marquez, that in average has a slightly positive advantage, on a much poorer bike and this is due to the high skill that the model has attributed to Marquez. Below the code:

6 Weather model

In this section we implement the enhanced model that has been previously presented by adding the influence of the weather conditions.

6.1 Convergence check

Looking at the Rhats we can confidently state that the distributions are convergent and the model works well.

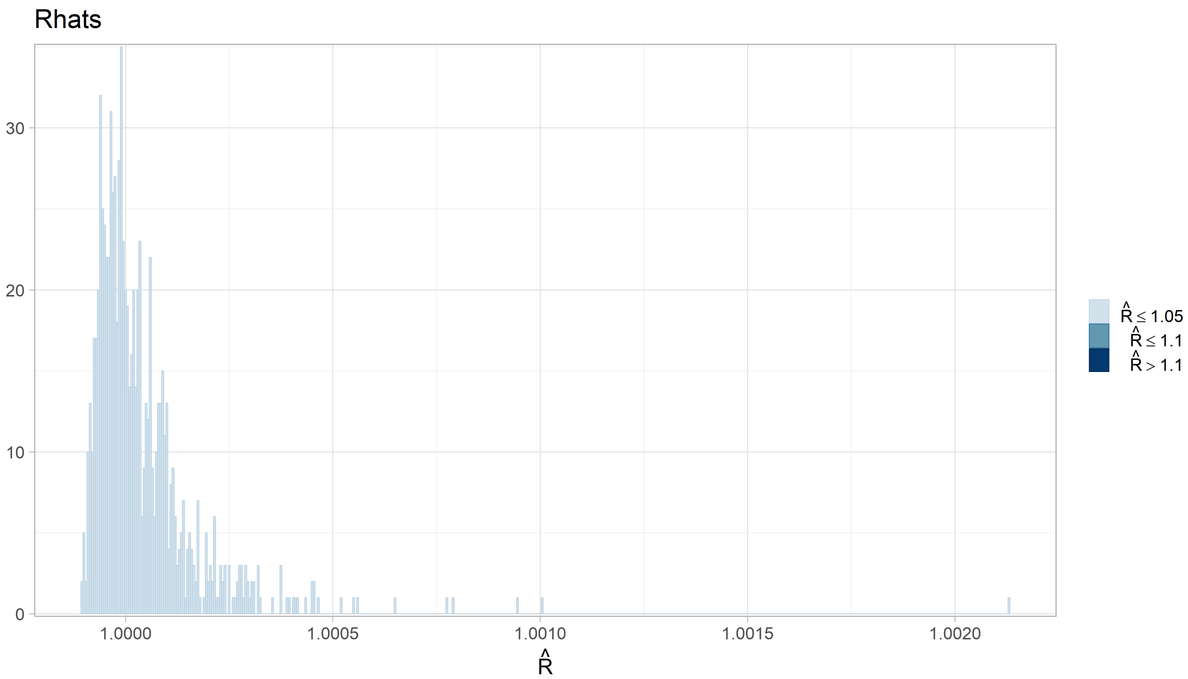

Figure 14: Diagnostics

As before let's take a look also at the ESS:

	Rhat	Bulk ESS	Tail ESS
constructor	1.00	10601.70	18511.01
constructor form	1.00	2943.63	7473.26
rider intercept	1.00	10756.49	17646.15
rider slope	1.00	15087.65	23497.51
rider form	1.00	7092.82	14993.16

Below we can also see the trace plots and the running means that confirm what said before.

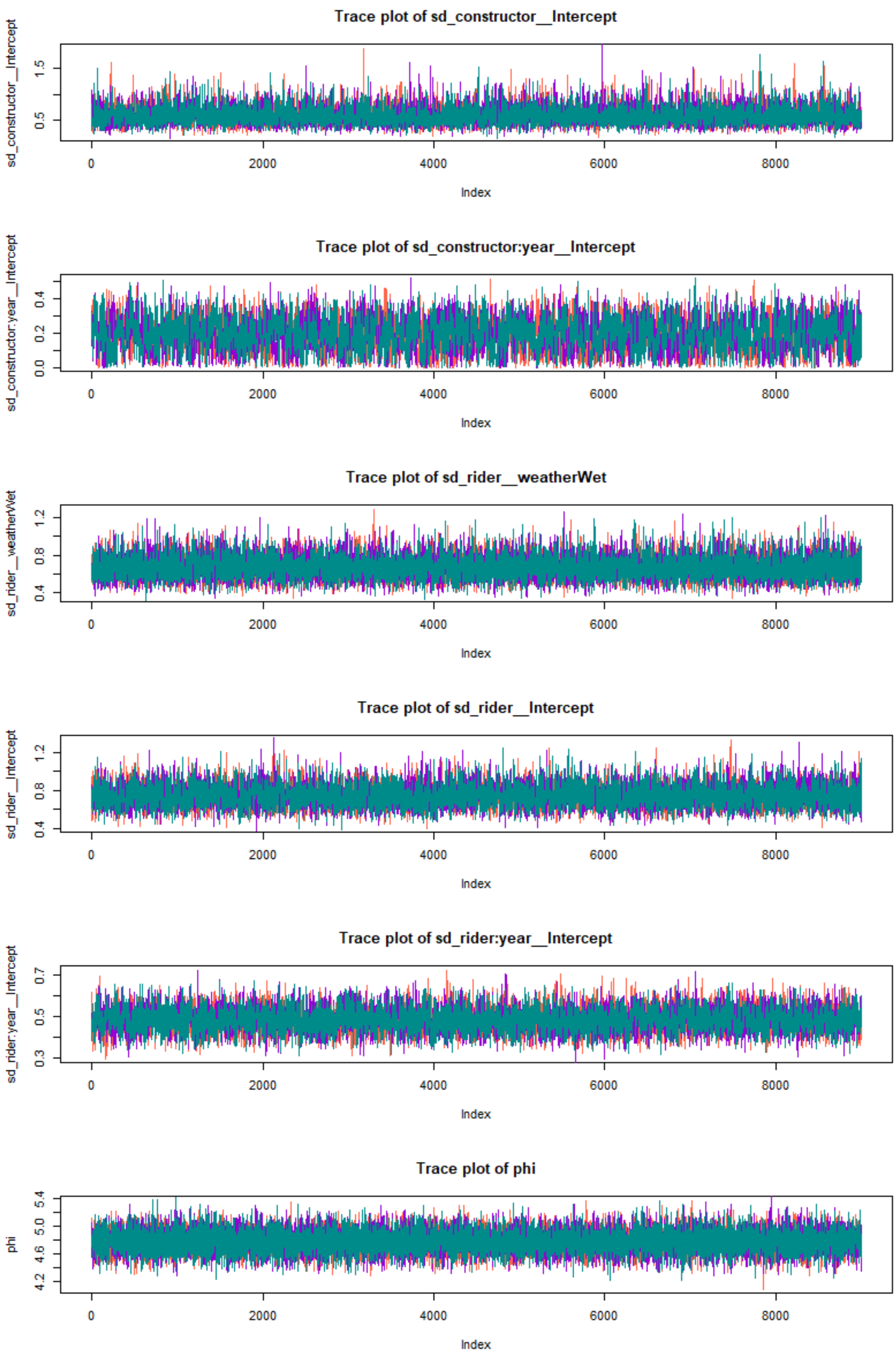

Figure 15: Chains trace plots for each parameter to estimate

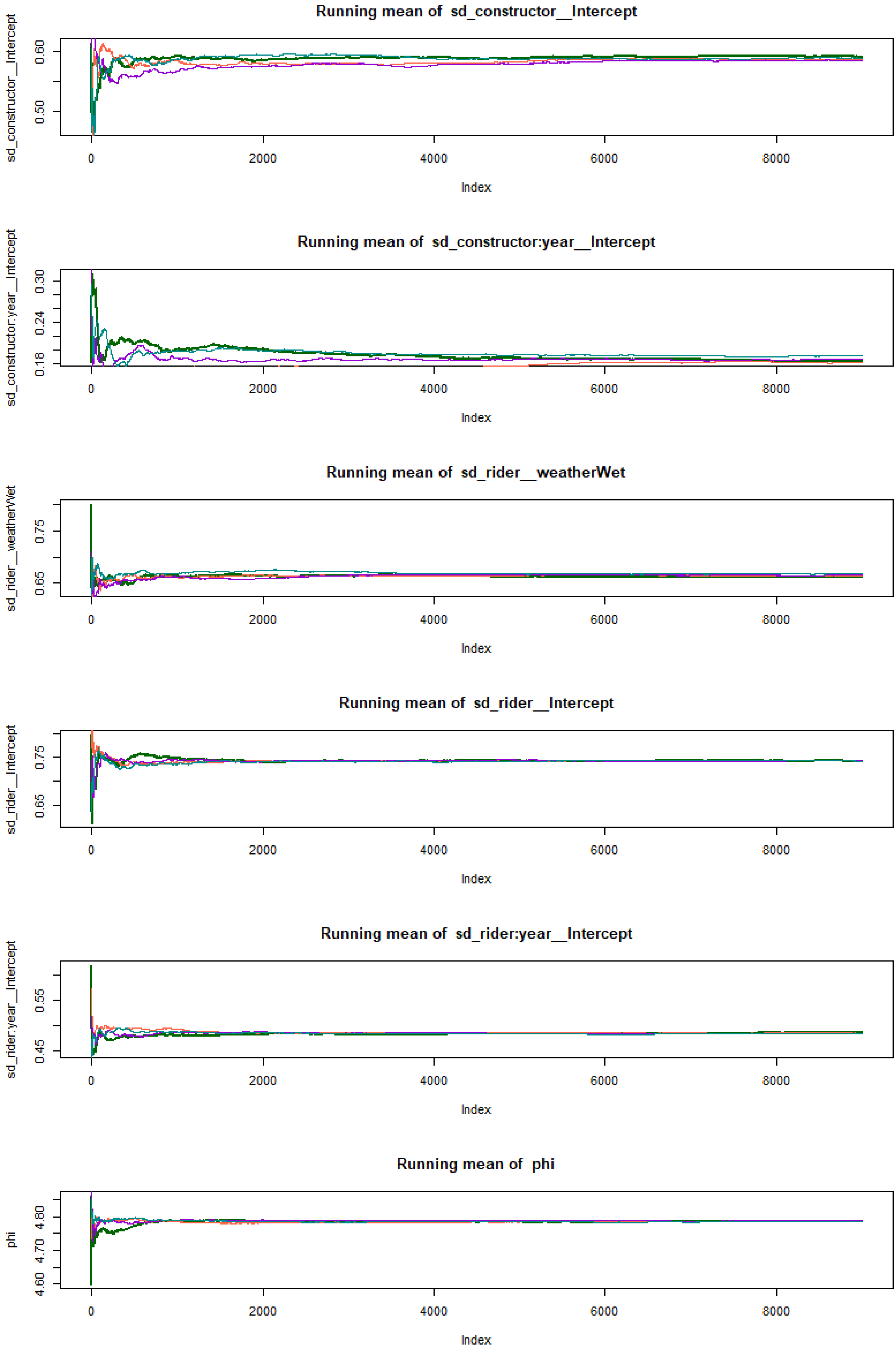

Figure 16: Running means

6.2 Posterior check

For what concerns the posterior check it is interesting to see how the distributions are varying depending on the weather conditions. In the case of wet conditions we are able to find who are the most skilled riders when it comes to wet races and of course given that in a season there are not only wet races the distributions will not be well overlapped.

Figure 17: Posterior predictive check on the POC scale for Dry races

We can observe some "boosted" densities due to the new predictor. In fact riders as Petrucci or Miller are known to be very good riders on the wet while it seems like is not the same for Lorenzo.

Figure 18: Posterior predictive check on the POC scale for Wet races

It is even more clear in this second plot where Petrucci and Miller are predicted as two of the best riders in absolute in wet conditions, while Jorge Lorenzo is projected as one of the worst and this is historically true. The following table gives us an insight on those riders scores during wet races compared to dry ones:

	rider	Avg. wet placement	Avg. wet POC	Avg. dry placement	Avg. dry POC
1	Danilo Petrucci	5.56	0.75	9.65	0.50
2	Jack Miller	6.12	0.71	8.17	0.59
3	Jorge Lorenzo	12.00	0.32	7.78	0.61

As before we also check the posterior distribution on the rank scale and this time it seems the model is able to better overlap with the observed results, in fact there are no specific cases were the model is completely off target.

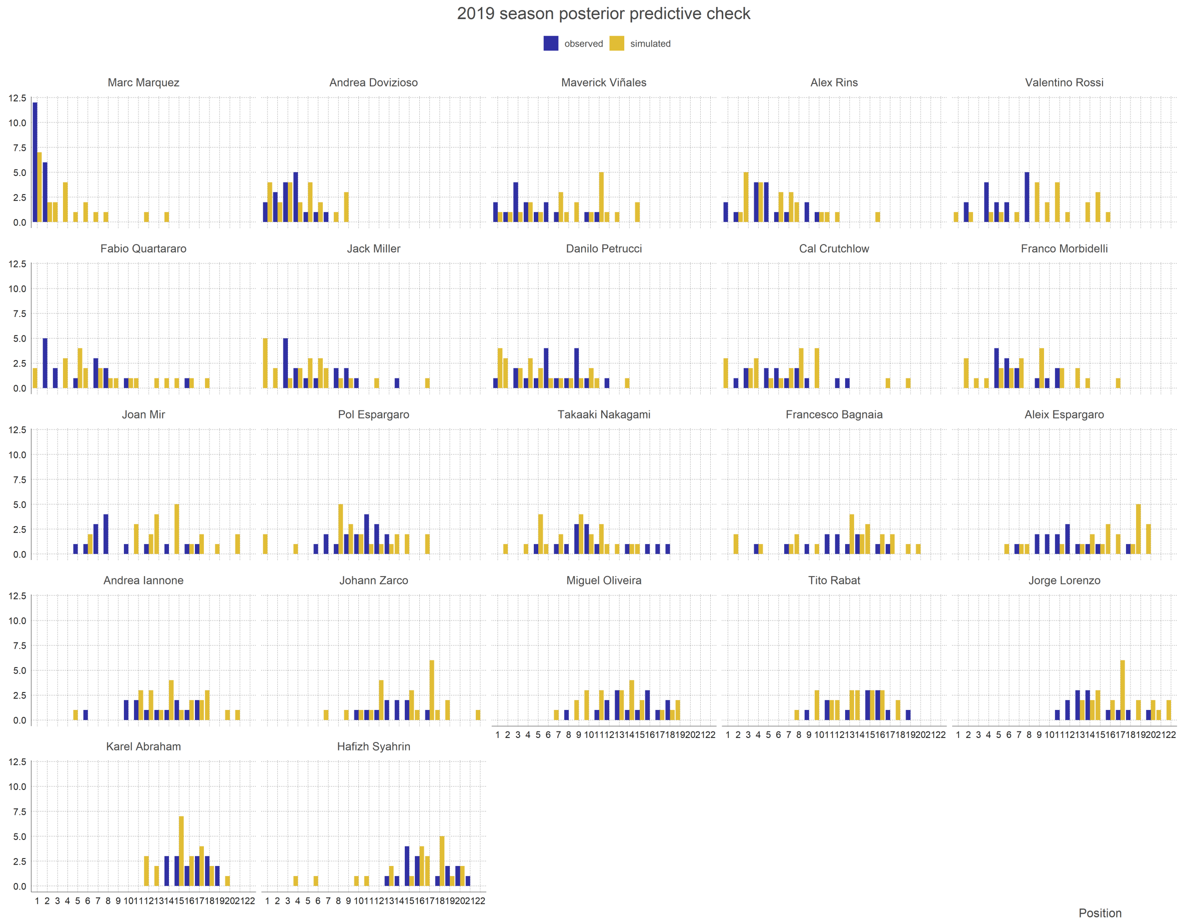

Figure 19: Posterior predictive check on the rank scale

6.3 Model inference

To compare the two models outcome we will reproduce the same inferential steps on the new model but we will focus on the most relevant differences, for this reason we have not included the inference on constructors advantage because it is independent from the weather conditions.

6.3.1 Inference on rider skill

For what concernes the inference on rider's skill given the definition of the model we have focused in this case on the combined effect of the intercept and the slope as regards the ability of the riders to take into account the weather effect and make the most of the capabilities of the model.

	rider	weather est.	weather rank	basic est	basic rank
1	Marc Marquez	1.23	1	1.48	1
2	Jack Miller	1.20	2	0.43	10
3	Fabio Quartararo	0.93	3	1.27	2
4	Aleix Espargaro	0.71	4	1.11	3
5	Danilo Petrucci	0.67	5	-0.14	17
6	Pol Espargaro	0.50	6	0.22	14
7	Alex Rins	0.42	7	0.35	11
8	Alex Marquez	0.41	8	-0.30	18
9	Johann Zarco	0.40	9	0.73	5
10	Enea Bastianini	0.38	10	0.51	8
11	Miguel Oliveira	0.36	11	0.31	13
12	Jorge Martin	0.34	12	0.46	9
13	Joan Mir	0.33	13	1.03	4
14	Brad Binder	0.19	14	0.70	6
15	Francesco Bagnaia	0.08	15	0.50	8
16	Takaaki Nakagami	-0.10	16	-0.50	20
17	Maverick Viñales	-0.22	17	0.34	12
18	Luca Marini	-0.38	18	-0.50	19
19	Valentino Rossi	-0.52	19	-0.13	16
20	Franco Morbidelli	-0.58	20	-0.52	21
21	Iker Lecuona	-0.61	21	-0.04	15
22	Lorenzo Savadori	-0.68	22	-0.97	22

By introducing the weather factor, the riders ranking is completely upset. By flattening the gaps between the different bikes, the balance between the riders also changes, bringing out those who are more able to run in the wet, this does not necessarily mean that the latter are more skilled in general because it must be taken into account that wet races are only the 10% of the total. However, we can say that the model evaluates the skills of the riders well under these conditions and gives us a new perspective to evaluate the riders.

Another interesting fact concerns the trajectory of the riders' skill over the years. Turning our attention to the same riders as before, we observe how in the new model the trajectories are dampened both in the case of growth and decrease while maintaining the balance unchanged, we can interpret this by seeing the ability to drive in the wet as one of those skills that do not expire over time as the wet tends to eliminate the other factors and this is reflected in the damped course of the trajectories.

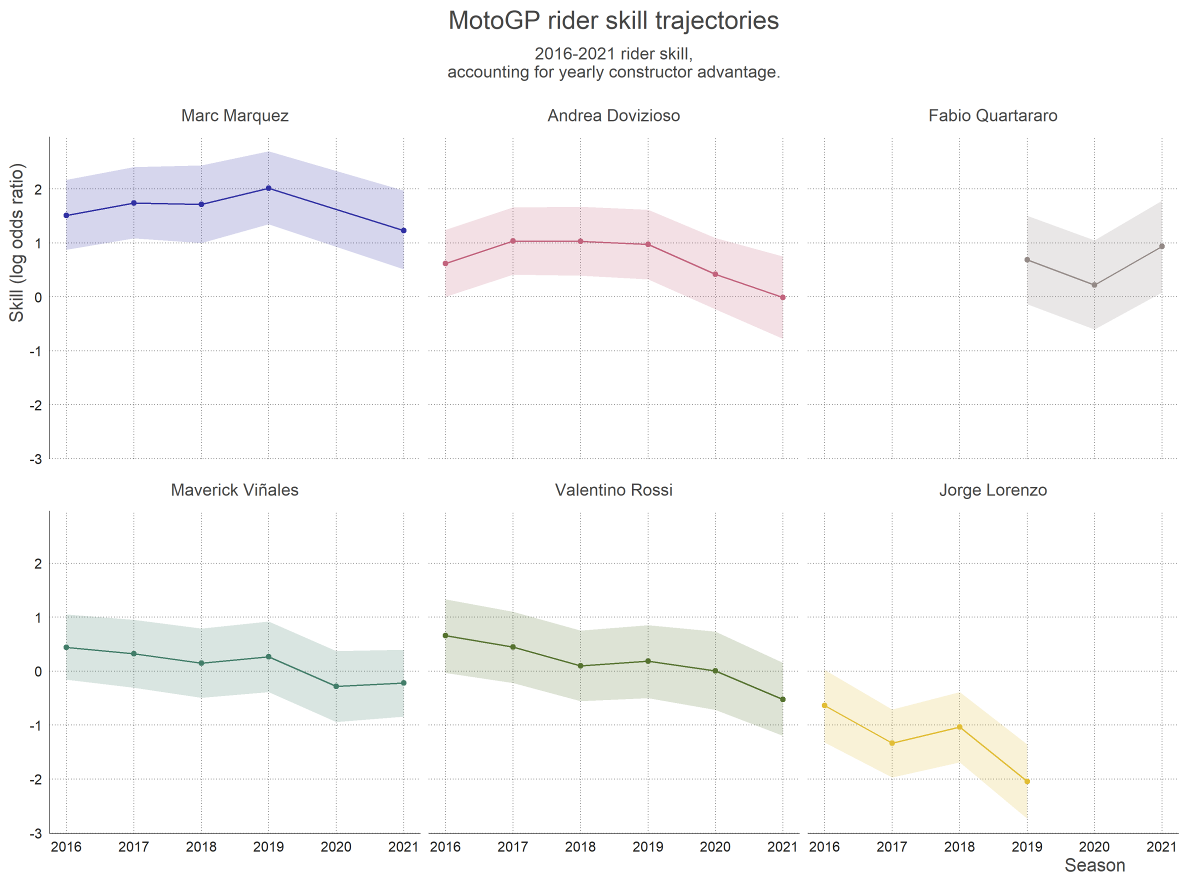

Figure 20: Riders skill trajectories

6.3.2 Relative contributions of riders and constructors

The most substantial differences can be found in the contributions of riders and constructors. It would seem that the model tends to increase the coefficients of almost all the riders thus leading to the importance of the rider far greater than that of the manufacturer. The main problem with this model is that since the advantage of the constructor is independent of the weather conditions, the only ones to benefit/lose from the addition of the predictor are the riders.

	Estimate	Est.Error	LB	UB
constructor	0.59	0.17	0.37	0.88
constructor form	0.19	0.09	0.03	0.33
rider intercept	0.74	0.11	0.58	0.94
rider slope	0.67	0.12	0.49	0.86
rider form	0.48	0.06	0.40	0.57

As before the estimate errors are inline with the original model and our basic one, the biggest change is in the contribution of the riders with respect of the constructor advantage. With respect to the basic model we have an increment of 1% in the contribution of the rider when considering dry races and this is perfectly in line with what we expected because when we are not considering the weather then the two models are basically the same.

In the case of wet races the contribution of the rider jumps to 76%, we can therefore say that the model has achieved the results we hoped for. In fact, introducing this variant, it was said that the gaps between different bikes are attenuated, leading to greater emergence of the rider's ability and this is exactly what happens in this case.

	Estimate	LB	UB
bike dry var	0.33	0.22	0.42
rider dry var	0.67	0.58	0.78
bike wet var	0.24	0.16	0.31
rider wet var	0.76	0.69	0.84

With regard to the correlation between the different parameters we have no surprises in fact the seasonal parameters are still the ones with the highest correlation as explained before.

	rider	constr.	rider form	constr. form	wet slope
rider	1.00	-0.03	-0.02	-0.13	0.12
constr.	-0.03	1.00	-0.04	-0.04	0.00
rider form	-0.02	-0.04	1.00	-0.40	0.01
constr. form	-0.13	-0.04	-0.40	1.00	0.02
wet slope	0.12	0.00	0.01	0.02	1.00

6.3.3 Counterfactual inference

We will reproduce now the same counterfactual example as before introducing the weather variable. This time we will answer the question whether if Dovizioso in a Yamaha is able to beat Marquez in a EG VDV in 2018 in the case of a **wet** race.

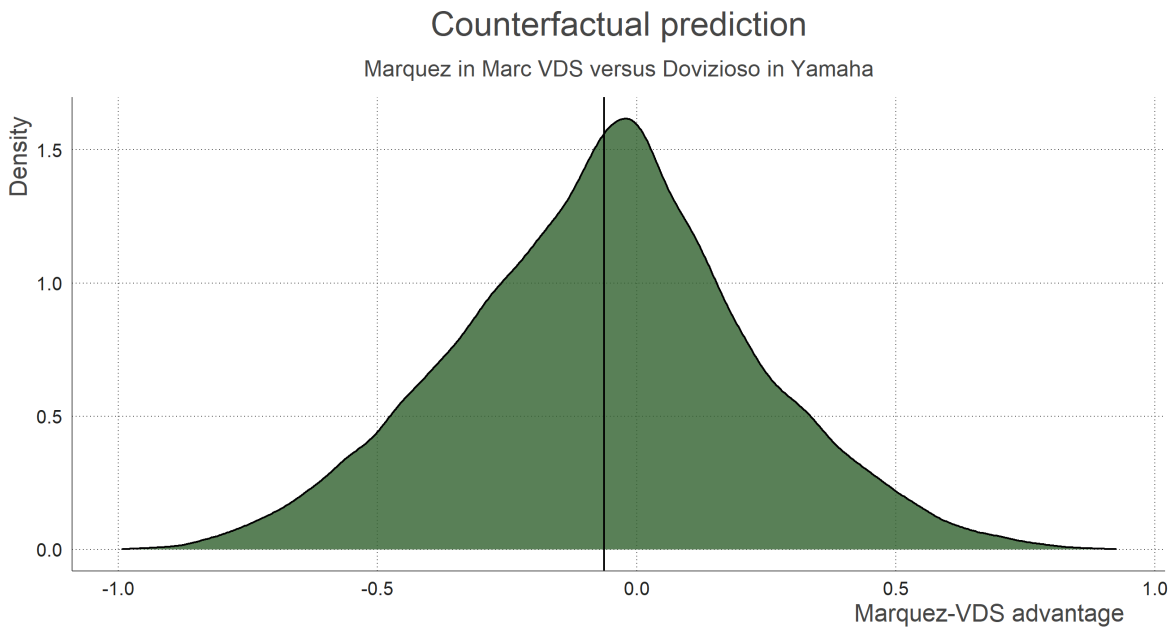

Figure 21: Counterfactual prediction, Dovizioso in Yamaha vs Marquez in VDV

This time we have a slightly advantage for Dovizioso and we are happy with this result. In fact, Dovizioso is an exceptional rider in the wet (as confirmed also by his direct rival in this statement) and in the last few years in which he fought for the title against Marquez he has achieved prestigious victories against the latter in the wet as well as at Motegi in 2017. Compared to the example with the basic model, there is a change in the result as in the wet the difference in skill between Marquez and Dovizioso is reduced despite the former being very strong in wet conditions.

7 Conclusions

Finally, we can be satisfied with the results obtained. In fact, both models achieved the results we expected, that is to demonstrate that in MotoGP the rider's ability is much more influential than the strength of the bike (for the base model) and that in the case of wet races this gap becomes even wider. Therefore, we believe that the second model is the most complete and suitable for estimating the contributions of the bike and rider in the outcome of a race also under different weather conditions.

The real strength of the model is in fact its bivalence for different weather conditions, so it also incorporates the basic model by extending and improving it to make it able to analyze in more detail.

References

(1) Erik-Jan van Kesteren and Tom Bergkamp, *Bayesian Analysis of Formula One Race Results: Disentangling Driver Skill and Constructor Advantage*, 2022.

(2) A. Bell, J. Smith, C. E. Sabel and K. Jones, *Formula for success: multilevel modelling of formula one driver and constructor performance, 1950–2014*, 2016.