Formula1 Models: How to improve the speed of the F1

Zhirong Lu

Table of Contents

[Abstract 3](#_Toc115122187)

[Formula1 Models: How to improve the speed of the F1 4](#_Toc115122188)

[F1 cars’ speed in the history of F1: technological and rules improvement 4](#_Toc115122189)

[Overall pattern. 4](#_Toc115122190)

[Rule Changes. 5](#_Toc115122191)

[Difference in drivers 7](#_Toc115122192)

[Which design philosophy is better 8](#_Toc115122193)

[Track temperature 10](#_Toc115122194)

[Pit stops: the change in pit stops and the effect of pit stop lengths 12](#_Toc115122195)

[Pit stops faster in more recent races. 12](#_Toc115122196)

[Use pit stop lengths to predict finishing position. 13](#_Toc115122197)

[Effects of leading the race 14](#_Toc115122198)

[References 16](#_Toc115122199)

[Appendix 17](#_Toc115122200)

Abstract

Formula 1, or F1 championship has long been one of the most popular and fastest motorsports in the world, with the combination of driving skills of the drivers, who are praised to be the twenty best drivers in the world, the racing strategy of the team, and the excellent manipulation of engineering in the design of the car. However, most of the data in F1 are kept secret in the hands of the teams since they do not want it to leak out to their competitors. Thus, fans have no access to these data. Luckily, since race results and some telemetry data are available to the public, there is an alternative: using statistical methods. As a result, factors that may affect the speed or the performance of an F1 car--technology improvement and rules, different driver, design philosophy, track temperature, length of pit stops, and whether the driver is leading the race or not—and how they affect the speed of F1 cars can be introspected by analyzing linear regression models and anova models built with data available to the public.

Keywords: linear regression, anova model, ancova model, Formula 1, motorsports

Formula1 Models: How to improve the speed of the F1

In the research, the data of race results and lap times of every lap in every race are from <https://ergast.com/mrd/db/>

The rules are from the official website of FIA, the organization that holds F1 each year: <https://www.fia.com/regulation/category/110>

The telemetry data, including live speed versus distance data and weather data are from F1’s official website.

## F1 cars’ speed in the history of F1: technological and rules improvement

## **Overall pattern.**

**Chart

Description automatically generated**

Using the dataset from F1 past races, I was able to track lap time in Q1 qualifying sessions in all races according to year from 1994 to 2022. Q1 is a session in qualifying, when all drivers use the maximum power engine mode to do the fastest lap possible in order to have a better grid start position at the start of the race. As a result, the qualifying lap times are the lap times that would most directly reflect the “raw” speed of the car, since in the race, the lap time might be affected by factors like tire compound or tire age.

However, while the 2022 data is recorded, since the F1 2022 season is not finished when this paper is drafted, the data of F1 2022 season will not be included in the research.

We can build an ancova model that uses year to explain the change in lap time, with qualitative terms in the model to eliminate the effects of different driver and the circuit. In the model, represents the change of F1 lap time according to year.

According to the model, we can run the regression below:

**Text

Description automatically generated**

According to the model, we can see that a linear relationship of speed change and year is significant.

As a result, , which means over the years, the lap time of F1 cars decrease, in other words, F1 cars are faster than before.

## **Rule Changes.**

However, there are clear patterns of lap time increasing and decreasing in certain small periods. The reason is that during the years 1994 to 2021, rules of F1 have changed a lot. Since every time there is a major rule change, the design of the car might vary greatly; the cars in 2021 are very different from cars in 1994. During the history of F1, there have been many rule changes, such as the different aerodynamics regulations, the introduction of turbo supercharging, and the change in tire suppliers. As the time of the data included in this paper, there has been three major rule changes:

In 1994, there is a ban of driver aids; in 2009, the cars were forced to have less aerodynamic devices (the prohibition of all aerodynamic devices other than front and rear wings) and slick tires are reintroduced; then, the most recent big rule change, is the new V6 turbo engines, started to be used in 2014.

By looking into the lap times in these three different eras, it is not hard to find that ever since there is a major rule change, the lap time of the cars might become longer, which means the cars may get slower than before. The reason of that is that the teams are not used to the new rules or do not have a very deep understanding of them. Then, over the years of development, the cars get faster and faster as they got more used to designing cars in that rule. The pattern of F1’s change in speed might be more clearly seen if we can build separate models for these three eras separated by major rule changes.

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

1994-2008 2009-2013 2014-2021

## Difference in drivers

There have been many heated debates in F1 when a driver wins a race. Whether the race win is caused by the driver’s excellent driving skills or just because the car is better designed than other cars on the grid. This problem can be solved by building ancova models to expose the real effect of drivers and the effect of cars by eliminating the effect of other terms (year, circuit). We can build the model below and run the regression:

Note: here, car effect is called “constructorId”, since in F1, every team is a constructor, and it constructs two cars with the same design for the two drivers to drive in the race at the same time.

Table

Description automatically generated with medium confidence

We can see that by eliminating the effects of other terms, both driver and car make a significant difference on the speed of the car. This conclusion, both drivers and cars are significant in affecting lap time, align with the consensus of F1 fans. If a good driver has a bad car, he or she might have great performance, but might not have great lap time, while if a bad driver has a good car, he or she cannot have great lap time either. Only when good drivers have good cars can they win championships.

## Which design philosophy is better

There are two main design philosophies in F1—more downforce, so the car can have more grid that would enable cars to go through the corners at a higher speed, but also more drag, so the cars might be slower in the straight; and less downforce, but less drag, so the cars can be faster in the straight, but slower in the corners.

We can test which design philosophy is better by analyzing the telemetry live data of speed versus distance in qualifying Q3, which could best reflect the fastest possible lap time of the car of the team Racing Point in six of the same race tracks that had races in both 2019 and 2020 season. The team changed its design philosophy from more downforce to less drag, and the driver did not change.

Chart, histogram

Description automatically generated Chart

Description automatically generated Chart

Description automatically generated

Chart

Description automatically generated Chart, histogram

Description automatically generated Chart

Description automatically generated with low confidence

By using the model below, we can run the regression.

Since in the original dataset, the value distance variable in the telemetry data is not the same for all laps even in the same circuit, we cannot compare the speed of the car in two years at the same position on the track. We can conduct a linear interpolation to make the interval in distance, of each speed collected, the same in all laps, so as to compare the performance of the car in 2020 and 2019 by comparing their instantaneous speed at the same position on the track in the two years.

A picture containing text

Description automatically generated

Since the p-value of design philosophy is smaller than 0.05, we can conclude that there is a significant effect for design philosophy. Then, a TuckeyHSD plot could be drawn.

Chart, box and whisker chart

Description automatically generated

There is statistical significance that the car in 2020 is faster than the car in 2019, since the 95% confidence interval of the speed of the 2020 car minus the speed of the 2019 car does not include 0 and is all positive, so the 2020 car is estimated to have a higher instantaneous speed on the track, which means that the low drag but low downforce is design philosophy made the car faster than the more drag but more down force design philosophy.

=0.3216. The plots of the model are shown in figure below.

Chart

Description automatically generated Chart, line chart

Description automatically generated Chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

However, the R square is not high. As a result, we can assume that is so low because a lot of variation is not caused by the design philosophy, it is caused by the feature of the circuit--corners and straights are not the same in every racetrack, and this is not an abnormal thing in this model.

Furthermore, since the car is more developed over the one year, the difference in design philosophy might be masked by the advancement in the car.

## Track temperature

In this model, I included the telemetry data of the three different tracks, Monza, Silverstone, and Spa, in 2018, 2019, and 2020. Then, we can use the backward elimination method to yield the final model. First, the model includes all the weather data and effects of year and circuit, and their interaction terms with each other. The is 0.9899. Then, the terms with p-value less than 0.05 are eliminated. Finally, the model got an of 0.9902.

Here is my final model:

Table

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

In the regression, the residual vs fitted plot shows a clear pattern, which indicates that there is no concern of collinearity at .

This model can be extremely helpful for F1 teams to create strategies in the race since the tires of F1 have a working range of about 90-110 Celsius, and the tires need an amount of time to heat up in order to reach this temperature, which may be affected by the track temperature. However, the weather broadcast may only contain things like temperature and rains, but not the direct track temperature. As a result, the teams will have to find out what the track temperature might be in the next few minutes to find out which type of tires to give the driver in the next pit stop.

## Pit stops: the change in pit stops and the effect of pit stop lengths

Pit stops faster in more recent races. The pit stops in the 1950s might have taken 1 minute, while ones in 2010 might only take 2 or 3 seconds. As the technology of pit stops improves, we can also make the hypothesis that pit stop lengths are shorter in more recent races.

pit stop lengths are shorter in more recent races.

We use the data from 2011 to 2021 when F1 banned the time of refueling during pit stops, without outliers out of Q1-1.5IQR, and Q3+1.5IQR, probably caused by damage repairing, so the time of pit stops only means the time to change four tires, as the boxpolt below shows.

Chart, box and whisker chart

Description automatically generated

And we can run the regression.

Text, letter

Description automatically generated

Then, the result of TuckeyHSD is shown below:

Though the variation seems very large, there is in fact not a very large fluctuation since the unit is milliseconds and the pit stop length is counted as the time interval from when the driver enter the pit lane to the time the drive exits the pit lane. Racetracks have pit lanes of various lengths, and there might be different locations of races in different seasons. So that difference is acceptable.

But the result got a coefficient of year as 105 milliseconds with an of 0.59. That means pit stops are estimated to be 105 milliseconds slower each year. This counter-intuitive result is because of the fluctuation of the length of pit lanes each year, and that the tire changing speed is reaching the limit of human body. Humans could change a tire in 2 seconds minimum, but it could be extremely hard to go faster than that since the process requires getting a tire down, moving it away, moving a new tire, and getting the new tire on the car.

Use pit stop lengths to predict finishing position. It can be reasonable to predict finishing position using length of pit stops, since the team that devoted a lot of resources into pit stops in order to shorten the pit stop lengths might be very rich so that they could also afford devoting a lot of money to develop a faster car.

There is no relationship between the finishing position and the length of pit stops

Then, we can run the regression.

**Text

Description automatically generated**

We can see that there is a significant effect of pitstop length and the interaction term of which time of pit stop it is (the first or second stop a driver pit in a race) to the finishing position.

## Effects of leading the race

In Formula 1, it is widely believed that when leading the race, the driver can have faster lap times, since the air in front is “clean”, not affected by aerodynamic devices in front. Moreover, the back car driver might be affected by the front car driver’s driving style, including breaking points and so on. We can test this by the model below.

There is no relationship between lap and whether the driver is leading the race.

Text

Description automatically generated

Since the residuals are not normal, we can do bootstrapping.

Chart, line chart

Description automatically generated Chart, histogram

Description automatically generated

Text, letter

Description automatically generated

The resulting is 0.427, which is acceptable, and the p-value of “leading a race” is smaller than 0.05. Thus, we can conclude that there is a difference in lap time when the driver is leading the race.

References

Peter Wright; Tony Matthews, "Formula 1 Technology," in Formula 1 Technology , SAE, 2001, pp.i-xviii.

Appendix

Code used in solving the models included in the paper:

%%Track Temperature

TT1\_1=read.csv("/Users/guaarmstrong/Downloads/example3.csv")

TT1\_1$year<-2020

TT1\_1$circuitId<-1

TT1\_2=read.csv("/Users/guaarmstrong/Downloads/TT1-2.csv")

TT1\_2$year<-2019

TT1\_2$circuitId<-1

TT1\_3=read.csv("/Users/guaarmstrong/Downloads/TT1-3.csv")

TT1\_3$year<-2018

TT1\_3$circuitId<-1

TT2\_1=read.csv(\*\*\*)

TT2\_1$year<-2020

TT2\_1$circuitId<-2

TT2\_2=read.csv(\*\*\*)

TT2\_2$year<-2019

TT2\_2$circuitId<-2

TT2\_3=read.csv(\*\*\*)

TT2\_3$year<-2018

TT2\_3$circuitId<-2

TT3\_1=read.csv(\*\*\*)

TT3\_1$year<-2020

TT3\_1$circuitId<-3

TT3\_2=read.csv(\*\*\*)

TT3\_2$year<-2019

TT3\_2$circuitId<-3

TT3\_3=read.csv(\*\*\*)

TT3\_3$year<-2018

TT3\_3$circuitId<-3

example3$Time<-NULL

example3$Rainfall<-NULL

regmodel=lm(TrackTemp~AirTemp+Humidity+Pressure+WindSpeed+AirTemp:as.factor(circuitId)+Humidity:as.factor(circuitId)+Pressure:as.factor(circuitId)+AirTemp:Humidity+AirTemp:Pressure+AirTemp:WindSpeed+Humidity:Pressure+Humidity:WindSpeed+Pressure:WindSpeed+as.factor(circuitId)+as.factor(year),data=WeatherTotal)

summary(regmodel)##model1

regmodel=aov(sqrt(TrackTemp)~AirTemp+Humidity+Pressure+AirTemp:as.factor(year)+Humidity:as.factor(year)+Pressure:as.factor(year)+WindSpeed:as.factor(year)+AirTemp:as.factor(circuitId)+Humidity:as.factor(circuitId)+Pressure:as.factor(circuitId)+WindSpeed:as.factor(circuitId)+AirTemp:Humidity+AirTemp:Pressure+Humidity:Pressure+Pressure:WindSpeed+as.factor(circuitId)+as.factor(year),data=WeatherTotal)

summary(regmodel)

regmodel=lm(sqrt(TrackTemp)~AirTemp+Humidity+Pressure+AirTemp:as.factor(year)+Humidity:as.factor(year)+Pressure:as.factor(year)+WindSpeed:as.factor(year)+AirTemp:as.factor(circuitId)+Humidity:as.factor(circuitId)+WindSpeed:as.factor(circuitId)+AirTemp:Humidity+AirTemp:Pressure+Humidity:Pressure+Pressure:WindSpeed+as.factor(circuitId)+as.factor(year),data=WeatherTotal)

summary(regmodel)

plot(regmodel)

all=regsubsets(TrackTemp~.,nbest=5,data=example3)

ShowSubsets=function(regout){

boxplot(example3$TrackTemp)

WeatherTotal<-rbind(TT1\_1,TT1\_2,TT1\_3,TT2\_1,TT2\_2,TT2\_3,TT3\_1,TT3\_2,TT3\_3)

Monze<-rbind(TT1\_1,TT1\_2,TT1\_3)

Silverstone<-rbind(TT2\_1,TT2\_2,TT2\_3)

Spa<-rbind(TT3\_1,TT3\_2,TT3\_3)

#design phylosophy

var<-approx(Firstlap19$Distance,Firstlap19$Speed, xout=seq(1,7050,5))$y # warning

fir19<-data.frame(var,seq(1,7050,5))

var<-approx(Firstlap20$Distance,Firstlap20$Speed, xout=seq(1,7050,5))$y # warning

fir20<-data.frame(var,seq(1,7050,5))

fir19$designPh<-2019

fir19$circuitId<-1

fir20$designPh<-2020

fir20$circuitId<-1

library(mosaic)

#design phylosophy222

x <- Secondlap19$Distance

y <- Secondlap19$Speed

var<-approx(x,y, xout=seq(1,7050,5))$y # warning

sec19<-data.frame(var,seq(1,7050,5))

var<-approx(Secondlap20$Distance,Secondlap20$Speed, xout=seq(1,7050,5))$y # warning

sec20<-data.frame(var,seq(1,7050,5))

sec19$designPh<-2019

sec19$circuitId<-2

sec20$designPh<-2020

sec20$circuitId<-2

plot(var~seq(1,7050,5))

var<-approx(Thirdlap19$Distance,Thirdlap19$Speed, xout=seq(1,7050,5))$y # warning

thi19<-data.frame(var,seq(1,7050,5))

var<-approx(Thirdlap20$Distance,Thirdlap20$Speed, xout=seq(1,7050,5))$y # warning

thi20<-data.frame(var,seq(1,7050,5))

thi19$circuitId<-3

thi19$designPh<-2019

thi20$designPh<-2020

thi20$circuitId<-3

var<-approx(Forthlap19$Distance,Forthlap19$Speed, xout=seq(1,7050,5))$y # warning

for19<-data.frame(var,seq(1,7050,5))

var<-approx(Forthlap20$Distance,Forthlap20$Speed, xout=seq(1,7050,5))$y # warning

for20<-data.frame(var,seq(1,7050,5))

for19$circuitId<-4

for19$designPh<-2019

for20$designPh<-2020

for20$circuitId<-4

var<-approx(Fifthlap19$Distance,Fifthlap19$Speed, xout=seq(1,7050,5))$y # warning

fif19<-data.frame(var,seq(1,7050,5))

var<-approx(Fifthlap20$Distance,Fifthlap20$Speed, xout=seq(1,7050,5))$y # warning

fif20<-data.frame(var,seq(1,7050,5))

fif19$circuitId<-5

fif19$designPh<-2019

fif20$designPh<-2020

fif20$circuitId<-5

var<-approx(Sixthlap19$Distance,Sixthlap19$Speed, xout=seq(1,7050,5))$y # warning

six19<-data.frame(var,seq(1,7050,5))

var<-approx(Sixthlap20$Distance,Sixthlap20$Speed, xout=seq(1,7050,5))$y # warning

six20<-data.frame(var,seq(1,7050,5))

six19$circuitId<-6

six19$designPh<-2019

six20$designPh<-2020

six20$circuitId<-6

designPh\_total<-rbind(fir19,fir20,sec19,sec20,thi19,thi20,for19,for20,fif19,fif20,six19,six20)

plot(Speed~Distance,type="l",col=designPh,data=designPh\_total)

regmodelDesignPh=aov(var~as.factor(designPh)+as.factor(circuitId)+as.factor(seq.1..7050..5.),use="complete.obs",data=designPh\_total)

regmodelDesignPh=lm(sqrt(var)~as.factor(designPh)+as.factor(circuitId)+as.factor(seq.1..7050..5.),use="complete.obs",data=designPh\_total)

pitStops=read.csv("/Users/guaarmstrong/Downloads/f1db\_csv/pit\_stops.csv")

races=read.csv("/Users/guaarmstrong/Downloads/f1db\_csv/races.csv")

pitStops\_total<-merge(pitStops,races,by="raceId",all.x=T)

#find Q1, Q3, and interquartile range for values in column A

Q1 <- quantile(pitStops\_total$milliseconds, .25)

Q3 <- quantile(pitStops\_total$milliseconds, .75)

IQR <- IQR(pitStops\_total$milliseconds)

#only keep rows in dataframe that have values within 1.5\*IQR of Q1 and Q3

no\_outliers <- subset(pitStops\_total, pitStops\_total$milliseconds> (Q1 - 1.5\*IQR) & pitStops\_total$milliseconds< (Q3 + 1.5\*IQR))

#view row and column count of new data frame

dim(no\_outliers)

regmodel\_pit=lm(milliseconds~year+as.factor(circuitId)+as.factor(driverId),data=no\_outliers,use="complete.obs")

summary(regmodel\_pit)

TukeyHSD(regmodel\_pit)

plot(TukeyHSD(regmodel\_pit))

plot(regmodel\_pit)

qqnorm(regmodel\_pit$residuals)

plot(no\_outliers$milliseconds~no\_outliers$year)

boxplot(milliseconds~year,data=no\_outliers)

plot(no\_outliers,aes(x=year,y=milliseconds,group=circuitId,color=circuitId)) + geom\_point(size = 0.1, shape = 8)

%lead

dim(no\_outliers)

regmodelLead=aov(milliseconds~as.factor(driverId)+as.factor(raceId)+lead,data=no\_outliers)

#regmodelLead=aov(milliseconds~as.factor(driverId)+as.factor(raceId)+lead+as.factor(driverId)\*lead,data=no\_outliers)

summary(regmodelLead)

plot(regmodelLead)

races=read.csv("/Users/guaarmstrong/Downloads/f1db\_csv/races.csv")

lead\_total<-merge(lapTimes,races,by="raceId",all.x=T)

regmodelLead=aov(milliseconds~as.factor(driverId)+as.factor(raceId)+as.factor(circuitId)+lead,data=lead\_total)

#regmodelLead=aov(milliseconds~as.factor(driverId)+as.factor(raceId)+lead+as.factor(driverId)\*lead,data=no\_outliers)

summary(regmodelLead)

plot(regmodelLead)

%speed trend

lapTimes=read.csv("/Users/guaarmstrong/Downloads/f1db\_csv\_2/qualifying.csv")

races=read.csv("/Users/guaarmstrong/Downloads/f1db\_csv/races.csv")

rawSpeed\_total<-merge(lapTimes,races,by="raceId",all.x=T)

rawSpeed\_total$Q1\_seconds<-as.numeric(rawSpeed\_total$q3Seconds)

rawSpeed\_total10<-subset(rawSpeed\_total,rawSpeed\_total$Q1\_seconds!="NA")

rawSpeed\_total10<-subset(rawSpeed\_total10,rawSpeed\_total$year!=2022)

#get rid of outliers

qplot(q3Seconds~year,color=rule,data=no\_outliers)

boxplot(Q1\_seconds~year, data=no\_outliers)

regmodelSpeedChange=lm(Q1\_seconds~as.factor(circuitId)+as.factor(driverId)+year, data=no\_outliers)

summary(regmodelSpeedChange)

era1<-subset(no\_outliers,year<2009)

era2<-subset(no\_outliers,year>=2009)

era2<-subset(era2,year<2014)

era3<-subset(no\_outliers,year>=2014)

era3<-subset(era3,year<2022)

plot(Q1\_seconds~year,data=no\_outliers)

regmodelera1=lm(Q1\_seconds~as.factor(circuitId)+as.factor(driverId)+year, data=era1)

summary(regmodelera1)

boxplot(Q1\_seconds~year,data=era1)

abline(regmodelera1)

regmodelera2=lm(Q1\_seconds~as.factor(circuitId)+as.factor(driverId)+year, data=era2)

summary(regmodelera2)

plot(regmodelera2)

boxplot(Q1\_seconds~year,data=era2)

abline(regmodelera2)

regmodelera3=lm(Q1\_seconds~as.factor(circuitId)+as.factor(driverId)+year, data=era3)

summary(regmodelera3)

plot(regmodelera3)

boxplot(Q1\_seconds~year,data=era3)

abline(regmodelera3)

%pit stops and finishing position

pitStopsandPionts\_total<-merge(pitStops,results,by="mergeId",all.x=T)

pitStopsandPionts\_total<-subset(pitStopsandPionts\_total,position!="\n"&position!="R")

pitStopsandPionts\_total$position<-as.numeric(pitStopsandPionts\_total$position)

regmodel\_pitStopsandPionts=lm(position~as.factor(circuitId.x)+as.factor(raceId.x)+as.factor(driverId.x)+as.factor(stop):milliseconds+milliseconds,data=pitStopsandPionts\_total,use="complete.obs")

summary(regmodel\_pitStopsandPionts)

plot(regmodel\_pitStopsandPionts)