LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

F1	Fórmula 1
GP	Grande Prêmio (de Fórmula 1)

LISTA DE SÍMBOLOS

Ω	Espaço amostral
ω	Evento pertencente ao espaço amostral Ω
$\mathbb{R}$	Conjunto dos números reais

SUMÁRIO

1 INTRODUÇÃO

1.1 A Fórmula 1

Fórmula 1 (F1, de forma abreviada) é a modalidade de automobilismo mais popular do mundo. O "Campeonato Mundial de Pilotos", torneio da modalidade que premia pilotos e equipes que conseguirem maior pontuação na temporada, existe desde 1950. Durante estas mais de sete décadas atraiu a atenção dos apaixonados por carros e o interesse das maiores montadores do planeta. A história da Fórmula 1 é inseparável da história de algumas gigantes do automobilismo como Ferrari, Williams, Renault e McLaren.

O alto valor de mercado das empresas que veem na modalidade uma forma de consolidar ainda mais suas marcas eleva o nível de competitividade da modalidade. Com isso, há investimentos milionários feitos pelas equipes todos os anos. Tais investimentos são baseados em pesquisas avançadas buscando o máximo desempenho nas pistas. Fatores como aerodinâmica, tração dos pneus, entre outros são estudados com afinco a fim de maximizar a performance.

[Tremayne 2006, 6] descreve o carro de Fórmula 1 como uma confecção das mais atualizadas técnicas de desenvolvimento de produto e pesquisa criada dentro de uma escala de tempo que é inevitavelmente muito curta, para ser enviada para uma competição acirrada da qual não há trégua durante 'a temporada'.

Não somente as marcas, mas também diversos pilotos ganharam notoriedade na história do esporte. Dentre eles o alemão Michael Schumacher, vencedor de sete temporadas, e o inglês Lewis Hamilton, também detentor de sete títulos e recordista de vitórias em corridas - com 103 conquistas. Há também ídolos brasileiros como Nelson Piquet - vencedor de três temporadas, Emerson Fittipaldi - duas vezes campeão, Rubens Barrichello - detentor do recorde de corridas disputadas - 322, e o maior deles, Ayrton Senna - também com três títulos.

Uma temporada de Fórmula 1 consiste em corridas - conhecidas com Grandes Prêmios (GPs). Os GPs são realizados em diversos locais do mundo, em pistas geralmente construídas especialmente para a prática da modalidade. Na temporada de 2022 ocorreram disputas nos continentes americano, asiático, europeu e oceânico. Em temporadas anteriores houve corridas no continente africano.

1.1.1 Campeonato Mundial de Fórmula 1 de 2022

Em 2022, o prêmio de melhores piloto e montadora foram dadas para o condutor e equipe com maior número de pontos ao longo de 22 Grandes Prêmios durante o ano. Disputaram o título 10 equipes. Cada equipe participa com 2 carros e, consequentemente, 2 pilotos.

Tabela 1 – Tabela de pontuação do Campeonato de Fórmula 1 de 2022

Corrida		Corrida Sprint	
Posição	Pontuação	Posição	Pontuação
1	25	1	8
2	18	2	7
3	15	3	6
4	12	4	5
5	10	5	4
6	8	6	3
7	6	7	2
8	4	8	1
9	2	Volta mais rápida da corrida	
10	1	1	1

Conforme dados da Tabela 1, por Grande Prêmio, de modo geral, é possível pontuar entre 0 e 26 pontos - 25 na corrida, e 1 de bonificação de volta mais rápida da corrida. No ano de 2022, os Grandes prêmios de Emília-Romanha, Áustria e São Paulo contaram também com a corrida sprint, possibilitando até mais 8 pontos no evento.

Como não é escopo do trabalho falar sobre cada piloto individualmente, e a fim de evitar referenciá-los de formas distintas ao longo deste projeto, serão utilizados os sobrenomes dos pilotos na forma como são apresentados na transmissão televisiva globalmente.

Os pilotos podem ser substituídos no decorrer do campeonato. Não é comum, e não é uma estratégia interessante pois cada piloto pontua individualmente e quando um piloto novo entra na disputa ele não herda os pontos do seu antecessor. O piloto Albon - da equipe Williams teve um problema de saúde e foi substituído por de Vries no Grande Prêmio da Itália. Ainda, Vettel - da equipe Aston Martin foi diagnosticado com COVID no começo da temporada e não disputou as duas primeiras provas do ano - Grande Prêmio do Bahrein e Grande Prêmio da Arábia Saudita, sendo substituído por Hulkenberg.

Tabela 2 – Tabela de Grandes Prêmios do Campeonato de Fórmula 1 de 2022

Nome do Evento	Data do Evento	Piloto Vencedor
Grande Prêmio do Bahrein	20/3/2022	Leclerc - Ferrari
Grande Prêmio da Arábia Saudita	27/3/2022	Verstappen - Red Bull Racing
Grande Prêmio da Austrália	10/4/2022	Leclerc - Ferrari
Grande Prêmio Emília-Romanha	24/4/2022	Verstappen - Red Bull Racing
Grande Prêmio de Miami	8/5/2022	Verstappen - Red Bull Racing
Grande Prêmio da Espanha	22/5/2022	Verstappen - Red Bull Racing
Grande Prêmio de Mônaco	29/5/2022	Perez - Red Bull Racing
Grande Prêmio do Azerbaijão	12/6/2022	Verstappen - Red Bull Racing
Grande Prêmio do Canadá	19/6/2022	Verstappen - Red Bull Racing
Grande Prêmio da Inglaterra	3/7/2022	Sainz - Ferrari
Grande Prêmio da Áustria	10/7/2022	Leclerc - Ferrari
Grande Prêmio da França	24/7/2022	Verstappen - Red Bull Racing
Grande Prêmio da Hungria	31/7/2022	Verstappen - Red Bull Racing
Grande Prêmio da Bélgica	28/8/2022	Verstappen - Red Bull Racing
Grande Prêmio da Holanda	4/9/2022	Verstappen - Red Bull Racing
Grande Prêmio da Itália	11/9/2022	Verstappen - Red Bull Racing
Grande Prêmio de Singapura	2/10/2022	Perez - Red Bull Racing
Grande Prêmio do Japão	9/10/2022	Verstappen - Red Bull Racing
Grande Prêmio dos EUA	23/10/2022	Verstappen - Red Bull Racing
Grande Prêmio da Cidade do México	30/10/2022	Verstappen - Red Bull Racing
Grande Prêmio de São Paulo	13/11/2022	Russel - Mercedes
Grande Prêmio de Abu Dhabi	20/11/2022	Verstappen - Red Bull Racing

A Tabela 2 mostra quais foram os eventos da temporada de 2022, com a data de realização e o piloto vencedor. Trata-se de um campeonato longo, com duração de 8 meses e corridas em circuitos espalhados pelo mundo. As equipes precisam levar isso em consideração ao fazer o planejamento, e adaptar-se conforme o andamento do campeonato.

Tabela 3 – Resultados por equipe do Campeonato de Fórmula 1 de 2022

Posição	Equipe	Vitórias	Pontos
1	Red Bull Racing	17	759
2	Ferrari	4	554
3	Mercedes	1	515
4	Alpine	0	173
5	McLaren	0	159
6	Alfa Romeo	0	55
7	Aston Martin	0	55
8	Haas F1 Team	0	37
9	AlphaTauri	0	35
10	Williams	0	8

Tabela 4 – Resultados por piloto do Campeonato de Fórmula 1 de 2022

Posição	Piloto	Equipe	Vitórias	Pontos
1	Verstappen	Red Bull Racing	15	454
2	Leclerc	Ferrari	3	308
3	Perez	Red Bull Racing	2	305
4	Russel	Mercedes	1	275
5	Sainz	Ferrari	1	246
6	Hamilton	Mercedes	0	240
7	Norris	McLaren	0	122
8	Ocon	Alpine	0	92
9	Alonso	Alpine	0	81
10	Bottas	Alfa Romeo	0	49
11	Vettel	Aston Martin	0	37
12	Ricciardo	McLaren	0	37
13	Magnussen	Haas F1 Team	0	25
14	Gasly	AlphaTauri	0	23
15	Stroll	Aston Martin	0	18
16	Tsunoda	AlphaTauri	0	12
17	Schumacher	Haas F1 Team	0	12
18	Zhou	Alfa Romeo	0	6
19	Albon	Williams	0	4
20	Latifi	Williams	0	2
21	de Vries	Williams	0	2
22	Hulkenberg	Aston Martin	0	0

As tabelas 3 e 4 mostram que a temporada de 2022 foi dominada pela equipe Red Bull Racing, vencedora de 17 dos 22 Grandes Prêmios. Ferrari e Mercedes foram as outras a conseguir vitórias em eventos - 4 e 1 -, respectivamente. Especificamente, o piloto Verstappen foi o responsável por 15 das 17 vitórias de sua equipe.

1.2 Objetivos

Este trabalho tem como objetivo estudar as variáveis que impactam o tempo necessário para completar uma volta em uma corrida de Fórmula 1. Para tal é feita uma análise exploratória dos dados e o tratamento dos mesmos a fim de aumentar a confiabilidade da análise.

Em seguida são listados os principais fatores que influenciam na variável de interesse - o tempo para completar uma volta. A estrutura dos dados também é levada em conta.Desta forma, são propostos modelos lineares mistos com efeitos aleatórios cruzados, que visam explicar o tempo da volta considerando a influência dos pilotos e do circuito da corrida.

Os modelos são submetidos a critérios de comparação adequados para este tipo de estudo, definindo-se um modelo mais apropriado. Em cima deste modelo observa-se a influência das variáveis explicativas escolhidas no tempo necessário para completar uma volta. Em seguida são elencados, com base no estudo, desafios e possíveis melhorias para trabalhos futuros.

1.3 Tratamento dos Dados

Para o presente trabalho foram utilizados dados da API para desenvolvedorr Ergast, de uso livre para fins não comerciais. Esta ferramente disponibiliza dados desde o primeiro Campeonato Mundial de Fórmula 1, em 1950 e está em constante evolução. Estes dados são referência na comunidade de desenvolvedores do ecossistema da Fórmula 1 - servindo de base para sites e aplicativos.

Serão considerados apenas 20 pilotos. Os pilotos de Vries e Hulkenberg terão seus dados agregados aos de seus companheiros de equipe Albon e Vettel, que disputaram 21 e 20 das 22 corridas, respectivamente. Para não gerar dúvidas, serão utilizados Albon* e Vettel* no restante do trabalho.

Ainda, foram desconsiderados da análise deste trabalho os Grandes Prêmios de: Emília-Romanha, Japão, Mônaco e Singapura. A razão para isso foi pelo fato de haver chuva durante as corridas. Embora tenhamos dados de presença ou ausência de chuva, estes são insuficientes para medir o impacto na corrida. Para fazê-lo de forma eficiente seria necessário dados pluviométrico durante cada volta além de informação sobre a capacidade de absorção do asfalto.

Desta forma, o objeto deste trabalho serão as demais 18 corridas realizadas no ano de 2022. Ainda, para garantir a confiabilidade das análises foram:

- retiradas as voltas nas quais o piloto entra ou sai do *pit stop*.
- retiradas as voltas com e imediatamente após *Safety Car/ Virtual Safety Car*.
- mantidas apenas voltas sob bandeiras verde e amarela

A saber, um piloto dirige-se ao *pit stop* conforme instrução da equipe de engenheiros/ mecânicos, geralmente para troca de pneus. Devido a alta velocidade, os pneus são muito exigidos, causando um rápido desgaste e perda de performance. Outro motivo para fazer a parada pode ser a necessidade de ajuste devido a algum acidente. Na temporada de 2022 não era permitido reabastecer, então o combustível necessário para completar a corrida já foi calculado pelas equipes previamente.

O *Safety Car* tem como objetivo garantir a segurança dos pilotos. Considera-se que há risco, por exemplo, quando ocorrem acidentes que acarretem obstrução da via e condições climáticas adversas. Nestas situações o carro de segurança entra na pista e os pilotos são submetidos há regras específicas, como não ultrapassagem e limite de velocidade.

Em alguns casos, entende-se que o carro de segurança não precisa estar fisicamente presente, bastanto que os pilotos saibam dos riscos e sigam as regras, fazendo-se uso

do *Virtual Safety Car*. A bandeira verde está presente quando não há perigo na pista. Já a bandeira amarela significa que há perigo, porém a corrida, a princípio, não será interrompida.

Mesmo após os filtros iniciais, alguns outros problemas foram facilmente identificados. A volta de número 3 do Grande Prêmio da Inglaterra, foi interrompida logo na primeira volta, e apesar do filtro remover as voltas 1 e 2 a volta de número 3 foi impactada, uma vez que a corrida foi reiniciada nesta volta, com os pilotos saindo do repouso. No Grande Prêmio da Itália, um incidente com o carro do piloto Ricciardo afetou os tempos da 47ª volta.

No Grande Prêmio da Saudita, na 48ª volta, ocorreu um acidente entre os pilotos Albon* e Stroll. Apesar da colisão, o piloto Stroll conseguiu completar a volta, porém seu tempo foi comprometido. Algo semelhante ocorreu com o piloto Vettel* na 39ª volta do GP da Áustria, em um acidente provocado pelo piloto Gasly. Estas voltas também foram desconsideradas em nosso estudo.

A Fórmula 1 é uma competição que exige o máximo dos pilotos e dos carros, e incidentes são comuns. Outras voltas também poderiam ser consideradas *outliers* e removidas do presente estudo, porém optou-se por remover apenas os exemplos acima. Todas as voltas possuem circunstâncias únicas, por isso modelá-las é um desafio.

1.4 Análise Exploratória

Nesta seção serão explicadas as principais variáveis utilizadas neste trabalho a fim de explicar o tempo de uma volta em corrida. Estas variáveis sabidamente impactam na nossa varável de interesse, e são levadas em consideração nas análises das equipes de Fórmula 1.

1.4.1 *Grid Position*

A cada Grande Prêmio são realizadas 3 provas qualificatórias - Q1, Q2 e Q3, no mesmo circuito da corrida. Essas provas definem a posição inicial do piloto na corrida, usando como critério seu melhor tempo de volta. Isso dá aos pilotos que largam nas primeiras posições uma enorme vantagem, pois a ultrapassagem é uma das tarefas mais difícies na disputa automobilística.

O termo *grid position* é utilizado para definir esta colocação, e o primeiro colocado nas qualificatórias larga na *pole position*, isto é, primeira posição. O *grid position* é um ótimo indicativo da corrida pois é consequência de um teste do piloto e seu carro no local da corrida.

1.4.2 Pneus

Por corrida, as equipes podem usar 3 tipos de pneus anunciados pela organização do evento dentre 5 possíveis. Em caso de chuva, outros 2 são disponibilizados. Como os eventos com chuva foram desconsiderados desta análise, iremos nos restringir ao primeiro caso. As informações abaixo foram retiradas do site da Pirelli, empresa responsável pelo fornecimento de pneus do campeonato. [Pirelli]

- C1: o pneu mais duro, escolhido para circuitos que desgastam mais os pneus. Desenvolvido para atingir máxima resistência ao calor, é capaz de percorrer distâncias maiores com baixa diminuição de performance;
- C2: o segundo mais duro, porém um pouco mais versátil que o C1. Funciona para pistas mais quentes e rápidas como o C1;
- C3: está presente em todas as corridas. Pode ser nominado pela organização como pneu duro, pneu médio ou macio, e portanto é extremamente versátil;
- C4: desenvolido para performar bem em pistas apertadas e com muitas curvas. O desgaste deste pneu é alto e ele atinge o pico de performance rapidamente;
- C5: o pneu mais macio, ideal para pistas mais lentas que desgastam pouco o pneu e requer que o piloto manobre o carro constantemente nas curvas.

Em outras palavras, quanto mais macio for o pneu mais fácil será para o piloto realizar as curvas e manobrar o veículo. No entanto, o desgaste também será maior, e em poucas voltas a aderência do pneu ao asfalto do circuito pode comprometer-se. Analogamente, quanto mais duro for o pneu maior será sua durabilidade e resistência ao calor, sendo esta caractéristica ideal para pistas com retas longas e altas velocidades.

1.4.3 Circuitos

Cada circuito tem suas características únicas, demandando adaptabilidade dos pilotos e equipes ao longo do campeonato. Esta variabilidade acarreta em diferenças no tempo de volta em cada pista, tanto na medida de valor central - mediana, como na concentração dos tempos do evento, conforme mostrado na figura abaixo.

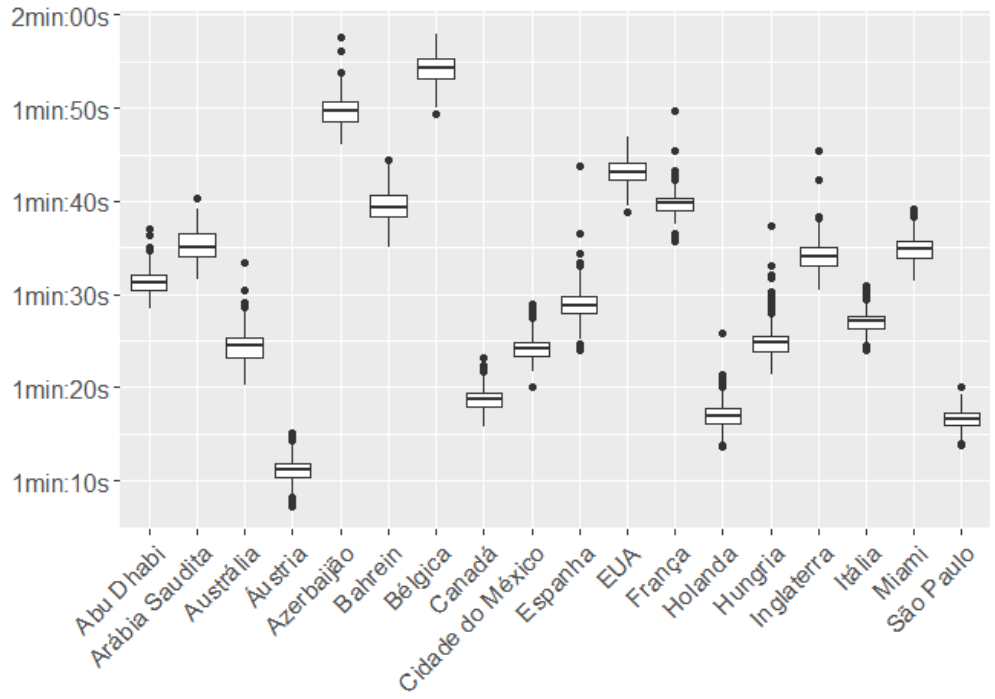

Figura 1 – Boxplot dos tempos de volta por Grande Prêmio

No GP da Áustria a maior parte das voltas foram completadas em pouco mais de 1 minuto e 10 segundos. Já no GP da Bélgica, para que o mesmo ocorresse, foi necessário cerca de 40 segundos a mais. Isto nos mostra que tratar as voltas de diferentes pistas como dados independentes não é uma boa estratégia para traçar um modelo que explique o tempo de volta. A seguir serão listados alguns dados que corroboram esta afirmação.

Tabela 5 – Características dos circuito

Local do Evento	Pneu mácio	Pneu médio	Pneu duro	Comprimento do circuito (km)	Voltas
Bahrein	C3	C2	C1	5,412	57
Arábia Saudita	C4	C3	C2	6,174	50
Austrália	C5	C3	C2	5,303	58
Miami	C4	C3	C2	5,412	57
Espanha	C3	C2	C1	4,675	66
Azerbaijão	C5	C4	C3	6,003	51
Canadá	C5	C4	C3	4,361	70
Inglaterra	C3	C2	C1	5,891	52
Áustria	C5	C4	C3	4,318	71
França	C4	C3	C2	5,842	53
Hungria	C4	C3	C2	4,381	70
Bélgica	C4	C3	C2	7,004	44
Holanda	C3	C2	C1	4,259	72
Itália	C4	C3	C2	5,793	53
EUA	C4	C3	C2	5,513	56
Cidade do México	C4	C3	C2	4,304	71
São Paulo	C4	C3	C2	4,309	71
Abu Dhabi	C5	C4	C3	5,281	58

Vale observar que o circuito do GP da Áustria possui 4,318km de comprimento, enquanto o circuito do GP da Bélgica possui 7,004km. O comportamento visto no boxplot ocorre, em grande parte, por este fator. Outro fator relevante é o conjunto de pneus adotados, definidos conforme características específicas da pista.

Outra informação importante sobre a pista é dada pela escolha do conjunto de pneus da pista. Para o GP do Bahrein, por exemplo, foi utilizado o conjunto de pneus mais duros: C3, C2 e C1, dando a entender que o desgaste nesta pista é acentuado. No GP do Canadá, por outro lado, foi feito o contrário, com a escolha de pneus mais macios. O mesmo pneu, C3, foi utilizado como pneu macio em Bahrein e como pneu duro no Canadá, tamanha a diferença entre as pistas.

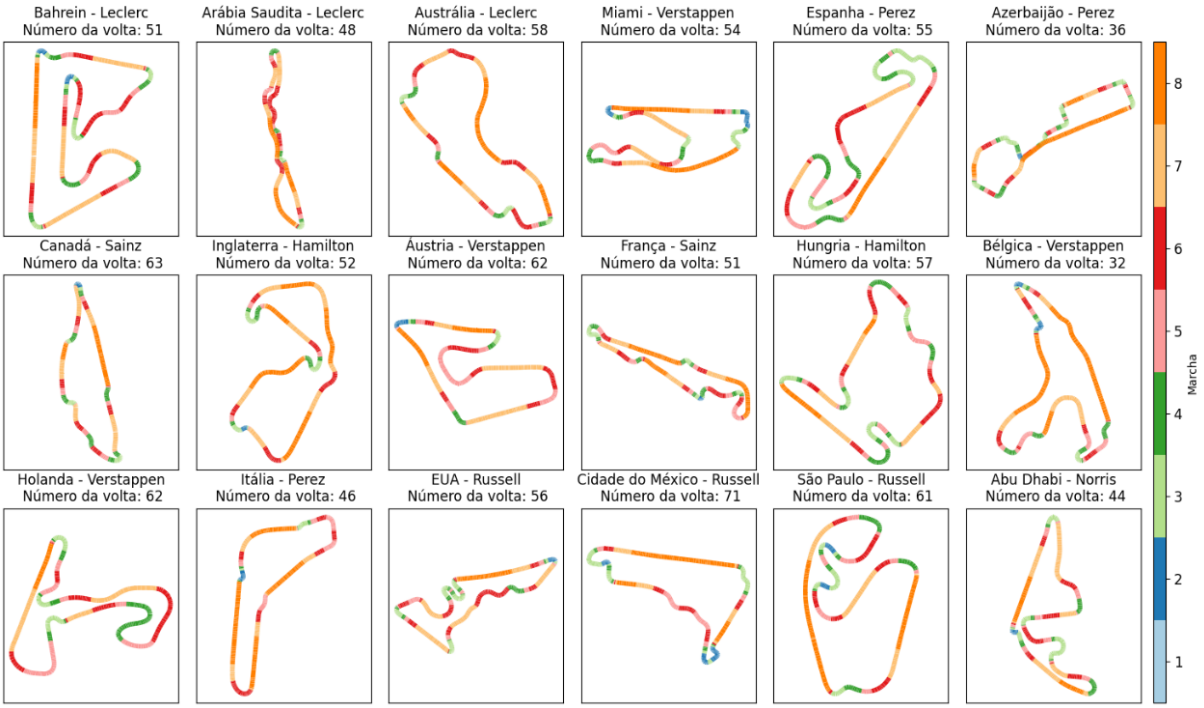

Figura 2 – Visualização das trocas de marcha nas voltas mais rápidas de cada Grande Prêmio

No entanto, o tamanho do circuito não é o único fator que pode impactar no tempo da volta. Cada pista possui características únicas, como explicitado na Figura 2, que mostra as voltas mais rápidas de cada um dos eventos realizados no ano. Nota-se também que as voltas mais rápidas tendem a ser realizadas no final das corridas, quando os carros já estão mais leves, pela quantidade reduzida de combustível.

Para passar pelas curvas da maneira mais eficiente, o piloto necessitar mudar de marcha constantemente, com algumas curvas exigindo até mesmo que ele troque para a primeira marcha, e consequentemente reduza consideravelmente a velocidade. Há ainda outros fatores que fazem com que a pista impacte no tempo da volta, como aderência e abrasividade do asfalto, sendo seus efeitos difícies de estimar.

1.4.4 Pilotos

Conforme já mostrado anteriormente houve um domínio, tanto por pilotos quanto por equipes, no campeonato em 2022. Esta vantagem em pontos é consequência dos resultados na pista, uma vez que o primeiro piloto a completar a corrida é também aquele que obtém menor média do tempo das voltas naquela corrida.

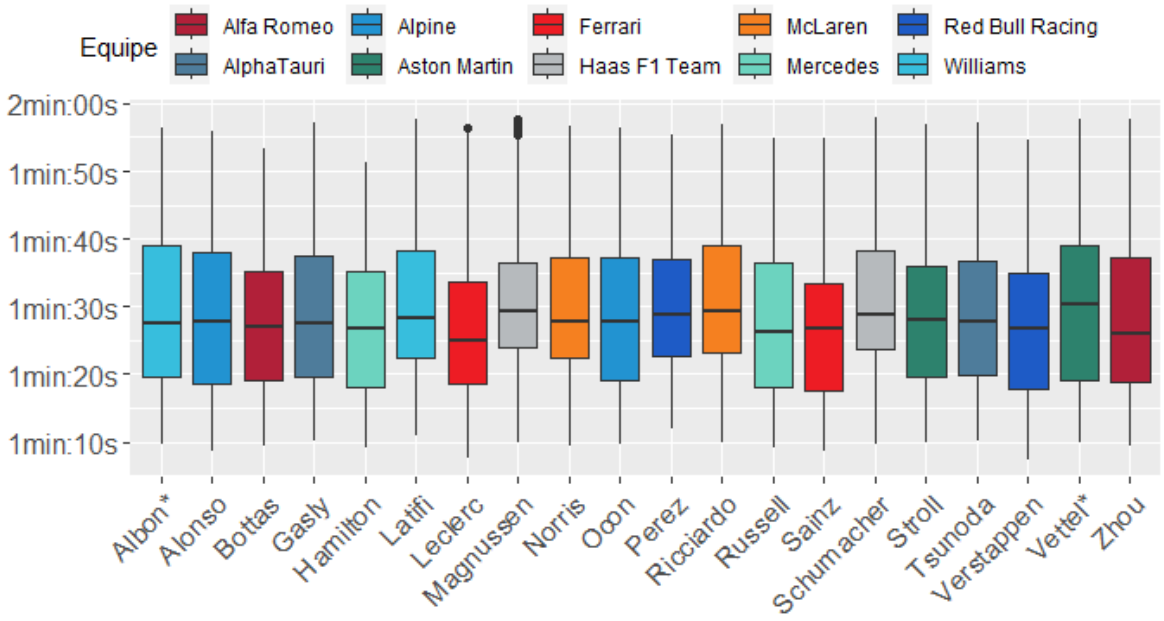

Figura 3 – Boxplot dos tempos de volta por piloto

A Figura 3 mostra o comportamento dos tempos de volta dos pilotos nas corridas ao longo do ano de 2022. Como não foi isolada a influência dos circuitos, é mais difícil de enxergar as diferenças, e, embora possa não parecer, elas são grandes, uma vez que as corridas são muitas vezes decididas por segundos, ou até mesmo milésimos.

Ainda, apesar do carro ser resultado da tecnologia da equipe, é possível notar que pilotos de mesma equipe performam de maneiras distintas. Um exemplo seriam os pilotos Verstappen e Perez, ambos da *Red Bull Racing*. Esta diferença fica mais evidente no primeiro quartil dos tempos, com quase 5 segundos separando-os.

1.4.5 Comportamento das voltas

A Figura 4, a seguir, mostra os tempos de volta dos pilotos no GP da Holanda com diferentes pneus. As linhas tracejadas em vermelho delimitam os *stints*. Toda vez que um piloto faz um *pit stop* - geralmente para troca de pneus, inicia-se um novo *stint*.

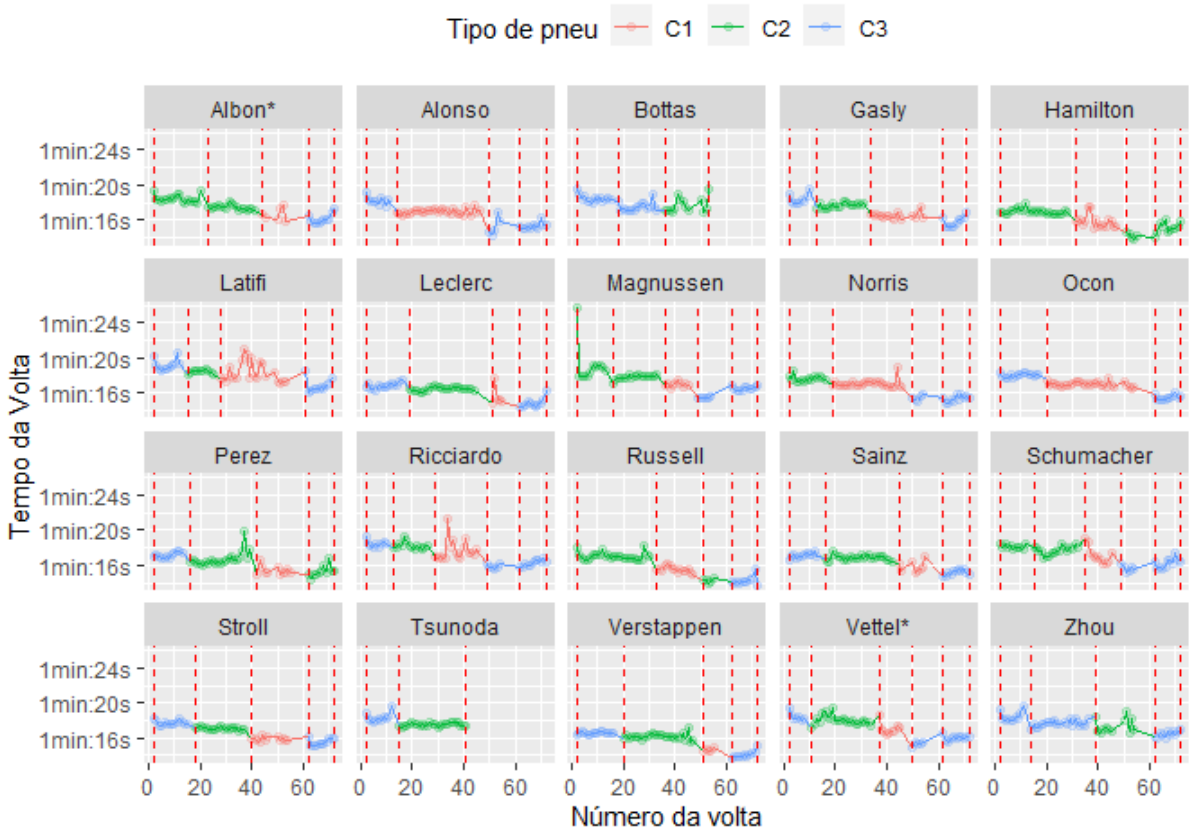

Figura 4 – Comportamento ao longo das voltas de cada piloto no Grande Prêmio da Holanda

Os tempos tendem a ser menores conforme o número de voltas aumenta, em acordo com o que foi visto sobre as voltas mais rápidas dos eventos. Isso se dá pois a quantidade de combustível cai conforme é consumido pelo veículo, fazendo com que ele fique mais leve. Ainda, a cada *stint*, costuma ocorrer uma redução imediata no tempo devido a substituição de um pneu desgastado, e conforme o novo pneu é exigido, o tempo volta a subir.

Espera-se ainda que o desgaste ocorra mais rapidamente nos pneus mais mácios, e mais lentamento nos pneus duros. No caso acima, o pneu mais duro é o C1, representado em vermelho, e o mais mácio é o C3 representado em azul, concordando com a Tabela 5 apresentada anteriormente.

Para que o comportamento da figura 4 ocorra é preciso, porém, especificar um evento, e individualizar os pilotos. A Figura 5 mostra como seria se desconsiderássemos estes fatores e explorasse diretamente a relação de números de voltas e tempo para completar a volta volta. Os dados ficam muito dispersos, e embora ainda seja possível notar uma diminuição conforme aumenta o número de voltas, especialmente nas últimas voltas, esta relação é muito fraca.

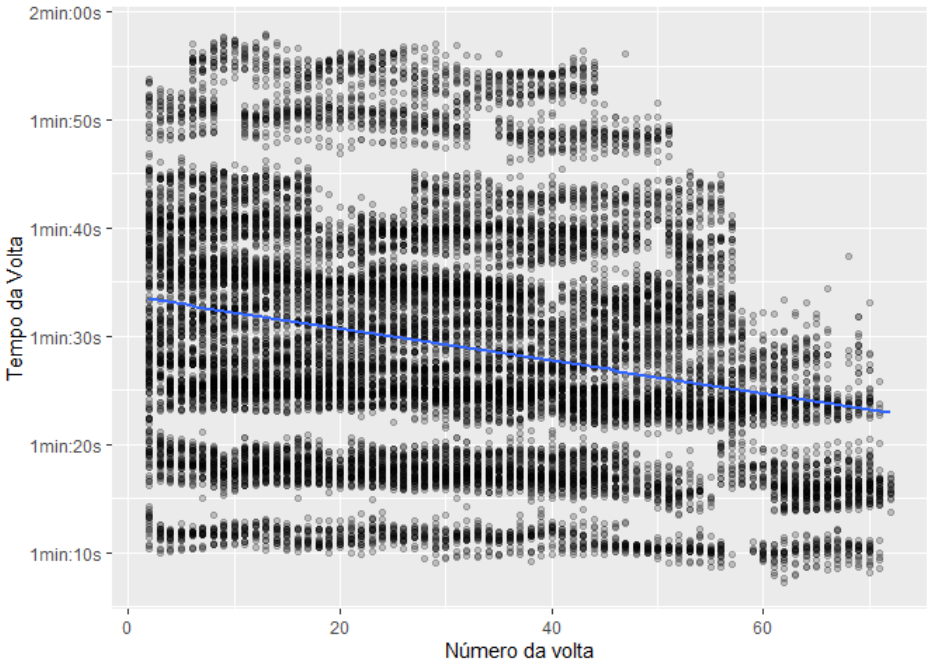

Figura 5 – Gráfico dos tempos de volta

A Figura 6 representa os mesmos dados da Figura 5, porém agrupados por circuitos, cada um representado por uma cor. Fica clara a formação de grupos, conforme já explicitado pela Figura 1. A Figura 7 mostra a dependência causada pelos circuitos ainda mais claramente representando-os separadamente, e desta forma tornando a dispersão consideravelmente mais concentrada.

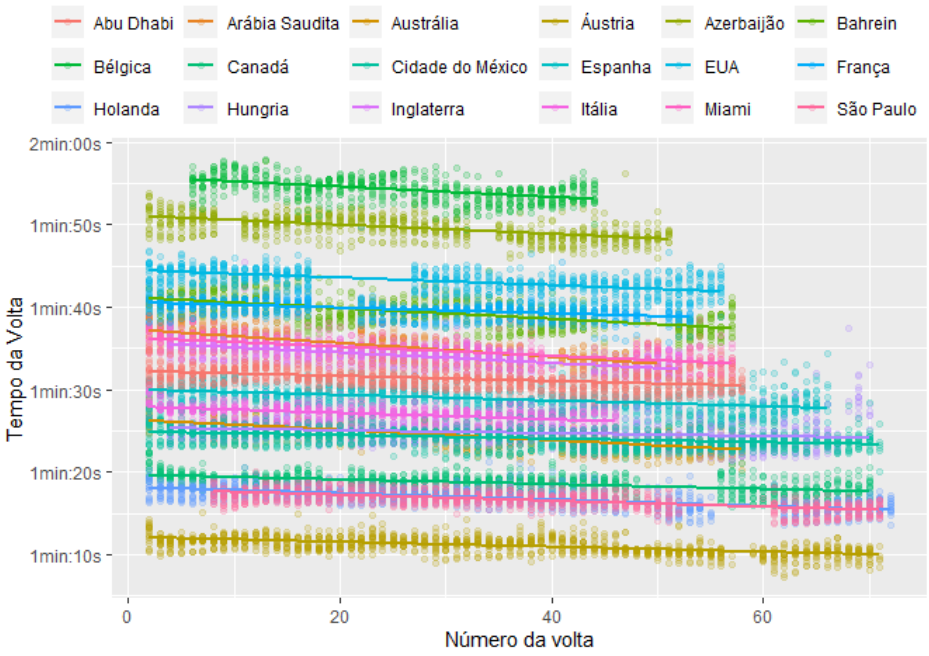

Figura 6 – Gráfico dos tempos de volta com cores específicas para cada circuito

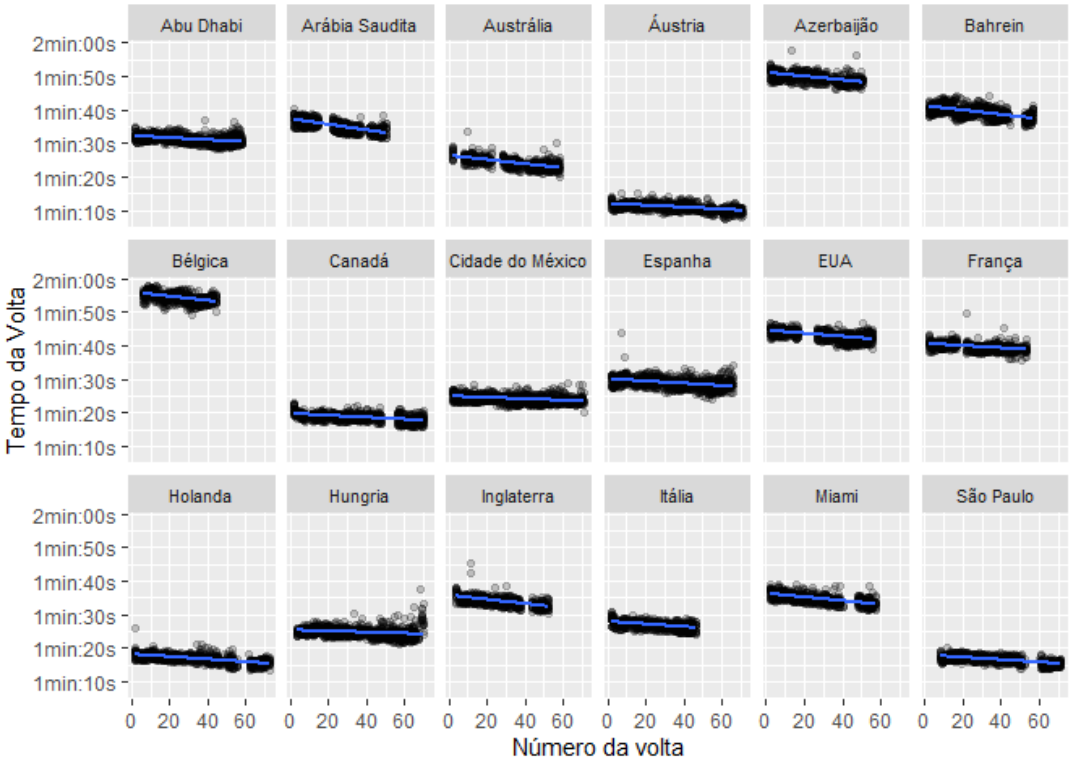

Figura 7 – Gráfico dos tempos de volta por Grande Prêmio

1.4.6 Outros fatores

Outras covariáveis sabidamente impactam na nossa variável de interesse, e são levadas em consideração pelas equipes durante as corridas. No entanto, é difícil estimar seus impactos, e incluir muitas variáveis em nosso modelo acarretaria em *overfitting*, pois, como exemplo, ocorrências de chuvas se deram em situações específicas.

Para evitar isso, poderia pensar-se em agregar dados de outros anos, no entanto o esporte é muito dinâmico, com mudanças de regras constantes. Os tempos de voltas de 2021 e 2022, por exemplo, não são imediamente comparáveis, pois as regras para as configurações do veículo mudaram consideravelmente, afetando a variável de interesse.

Fatores climáticos, como temperatura - do ar e da pista, umidade, pressão, velocidade e direção do vento e até mesmo a estratégia da equipe, como optar por fazer menos paradas e portanto forçar o piloto a permanecer por mais tempo com o mesmo pneu, confiando que mesmo em desvantagem ele será capaz de se defender e não ser ultrapassado, são algumas destas covariáveis. Devido a natureza dos modelos lineares mistos, conforme explicados no seguinte capítulo, embora não sejam explicitadas no modelo, o impacto delas afetará os efeitos aleatórios.

2 REVISÃO DE LITERATURA

Neste capítulo serão discutidos conceitos empregados neste trabalho. Dentre os tópicos abordados teremos: Análise de Regressão, Modelos Lineares, Modelos Lineares Mistos - explicando conceitos de dados aninhados e cruzados, e efeitos fixos e aleatórios - e Estatística Bayesiana.

2.1 Análise de Regressão

A Análise de Regressão é uma ferramenta estatística utilizada quando é razoável pensar que certas quantidades se relacionam, de maneira que seja possível prever um valor de interesse através das quantidades explicativas.

Exemplo 1 Supondo que um carro esteja na posição inicial S_0, e desloca-se com aceleração constante por um perído de tempo t, medido em segundos. Sejam S_f, v_0 e a: a posição após t segundos, a velocidade inicial no instante $t = 0$ e a aceleração, respectivamente.

Sabe-se que $S_f = S_0 + v_0\mathrm{t} + \frac{at^2}{2}$. Podemos escrever $S_f = f(S_0, v_0, t, a)$, ou seja, tratamos a variável S_f como variável dependente, e explicitamos uma relação com as variáveis independentes S_0, v_0, t e a. O exemplo acima é de uma relação determinística. A análise de regressão é utilizada em eventos estocásticos, quando há incerteza e não é possível precisar o valor que assumirá nossa quantidade de interesse, sendo essa sempre representada numericamente.

Exemplo 2 Consideremos a variável de interesse seja o tempo que um carro leva para, saindo do repouso, atingir a velocidade de 100 km/h.

Faz sentido pensar que o custo do motor do veículo influencie neste tempo. Outra variável que pode explicar este tempo é o peso do veículo. No entanto, não conseguimos uma fórmula direta para esta relação, como no exemplo anterior. Existem diversos outros fatores que influenciam na nossa variável de interesse de maneira que não conseguimos precisar, como a capacidade real do motor, uma vez que o preço é apenas um indicativo, o destreza do piloto, e a aderência dos pneus à pista são alguns desses.

Trata-se de um evento aleatório, por possuir influência indeterminadas de variáveis conhecidas e variáveis desconhecidas. Ainda, nossa variável de interesse, o tempo para atingir a velocidade de $100km/h$ é númérica, e especificamente, contínua. Este é o tipo de caso tratado na análise de regressão. Por meio de métodos estatísticos, conforme exemplificado na seção a seguir, busca-se estimar os efeitos de covariáveis usando dados

amostrados de eventos observados.

2.2 Modelos Lineares

Conforme explicado anteriormente, a análise de regressão tem como objetivo explicar uma variável aleatória através de uma relação com outras covariáveis. O tipo de relação e de modelo escolhido depende dos dados, podendo ser polinomial, exponencial, logarítmica, ou até mesmo combinações de diferentes famílias de funções.

Os modelos lineares se destacam, sendo muito usados para diversos tipos de problemas, são simples, rápidos para treinar e de fácil interpretação. Ainda, costumam ter variância relativamente baixa, possuindo menor tendência a *overfitting*. Generalizações destes modelos foram estudadas e validadas tornando-os capazes de se adaptar a diferentes problemas.

Modelos com apenas 1 variável independente são chamados de modelos lineares simples, e o ajuste é feito através de uma reta, sendo a distância de cada ponto para a reta, fixado o valor da variável dependente, o erro do modelo. De maneira geral, incluindo modelos lineares com mais covariáveis, o ajuste é feito por hiperplanos em um espaço de dimensão p + 1, com p representa o número de variáveis independentes.

Sob abordagem frequentista, uma forma de determinar a reta ou hiperplano que ajustará os dados é oo método de mínimos quadrados. Esta técnica busca encontrar os coeficientes que minimiza a soma dos quadrados das diferenças entre os valores observados e os valores previstos pelo modelo, e pode ser estudada com mais detalhes em [Charnet et al. 1999].

Sejam $\mathrm{y_i}$ e $\hat{\mathrm{y}}_\mathrm{i}$ os valores observados e os valores previstos pelo modelo, respectivamente. Algebricamente, o método de mínimos quadrados escolhe o hiperplano de maneira a minimizar $\sum_{i=1}^{n}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2$, com n representando o total de observações. O modelo linear simples assume a forma $\hat{\mathrm{y}}_\mathrm{i} = \beta_0 + \beta_1 x + \epsilon_i$, onde β_0 representa o intercepto e β_1 representa o coeficiente angular da variável independente x, e ϵ_i representa o erro do modelo.

2.3 Modelos Lineares Mistos

Modelos Lineares Mistos são extenões dos modelos lineares mais básicos, utilizada quando a estrutura dos dados analisados possui aninhamentos. Também são conhecidos como modelos lineares hierárquicos ou modelos lineares multiníveis, e são abordados de forma detalhada em [Gelman e Hill 2006].

2.3.1 Dados aninhados e cruzados

Estrutura de dados aninhadas são encontradas comumente em diversos estudos. Trata-se de dados nos quais cada observação pertence a um grupo, e acredita-se que este grupo possua um efeito importante na variável resposta. Trata-se de uma estrutura hierárquica.

Dados cruzados são referentes a estruturas nas quais dois ou mais grupos são agrupados de forma a conter todas as possíveis combinações. Dados cruzados não são uma estrutura hierárquica, em num modelo com cruzamento em um nível hierárquico ocorre o agrupamento desses fatores, sendo a combinação deles um nível hierárquico. Este cruzamento é utilizado quando assume-se que não há independência nas unidades estudadas.

2.3.2 Efeitos mistos

Os modelos lineares mistos possuem este nome por possuírem parâmetros de efeitos fixos e parâmetros de efeitos aleatórios, sendo portanto caracterizados como modelos lineares de efeitos mistos.

Efeitos fixos e aleatórios dizem respeito à natureza das variáveis que estão sendo modeladas. O primeiro assume que o efeito da variável é fixo para todas as unidades estudadas, e são utilizados para inferir sobre uma variável de interesse.

Por outro lado, os efeitos aleatórios são característicos de unidades que permite-se variar aleatoriamente. Isso ocorre para explicar variações desconhecidas e não explicadas pelo modelo, quando acredita-se que os dados não sejam independentes.

Exemplo 3 Deseja-se estimar quantas voltas um piloto de Fórmula 1 irá completar antes de parar para troca de pneus. Podemos considerar um modelo com um efeito fixo representado pela variável que representa o tempo de vida do pneu em uso. No entanto, sabe-se que cada pista possui uma interação do asfalto com o pneu diferente, exigindo o pneu de formas diferentes e difícies de prever.

Um efeito aleatório no intercepto para cada circuito pode ser interpretado como diferentes médias do tempo de duração para cada pista. Pode-se ainda assumir um efeito aleatório no coeficiente angular do tempo de vida do pneu em uso no modelo. Este efeito dá a entender que o tempo de vida do pneu impacta marginalmente de forma diferente na nossa variável de interesse.

A linha preta tracejada nas Figura 8 e 9 mostra como seria a reta de ajuste utilizando o método de mínimos quadrados para todos os dados, sem diferenciação por circuito. Quando adicionamos um efeito aleatório no intercepto (Figura 8), o ajuste aos dados é evidentemente melhor, ilustrando dependência das médias da quantidade de voltas

que o piloto irá completar antes da parada por circuito.

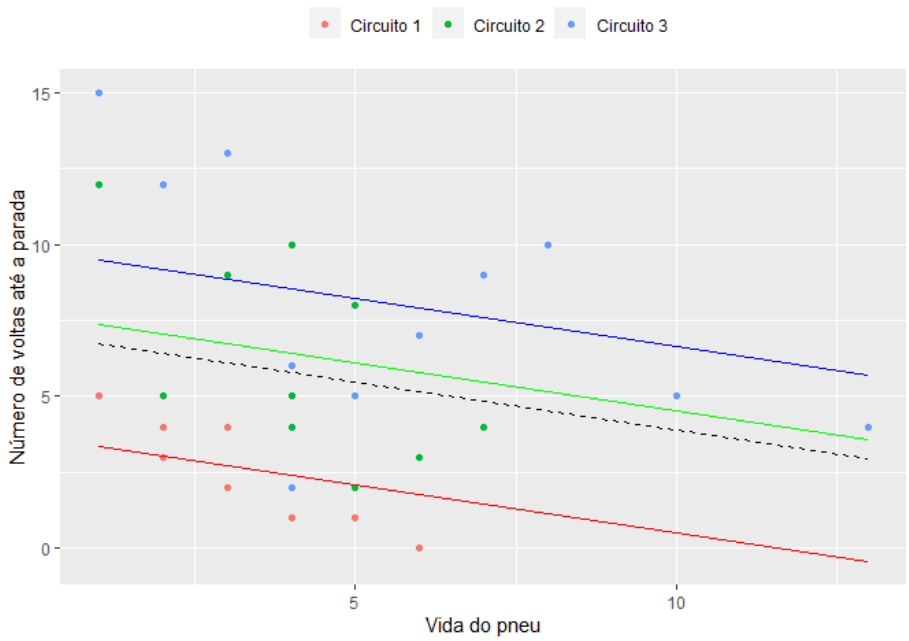

Figura 8 – Modelo com efeito aleatório no intercepto

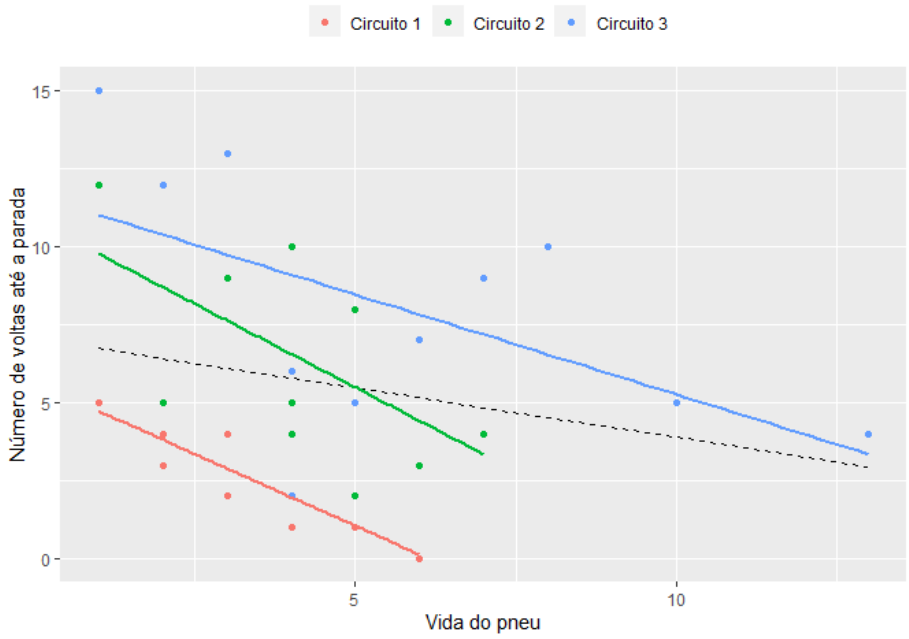

Figura 9 – Modelo com efeito aleatório no intercepto e no coeficiente angular da variável explicativa

A Figura 9, acima, representa um modelo com efeitos aleatórios no intercepto e no coeficiente angular. Pode-se destacar o circuito 2, no qual o número de voltas até a parada decai mais rapidamente que nos demais circuitos. Embora os coeficientes angulares das retas que ajustam os dados dos circuitos 1 e 3 possam parecer próximos, ainda são bem diferentes do modelo inicial, sem efeitos aleatórios.

2.4 Estatística Bayesiana

A metologia bayesiana difere-se da frequentista uma vez que a primeira entende probabilidade com uma medida de credibilidade de um dado evento, enquanto a segunda interpreta como uma frequência ao repetir um experimento diversas vezes. Estatística bayesiana considera conhecimentos iniciais sobre a probabilidade de eventos, que são incorporados na distribuição a priori. Estes conhecimentos podem vir de estudos anteriores ou opiniões de especialistas. Sua origem advém do Teorema de Bayes, que diz que: $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^{\complement})P(A^{\complement})}$.

Exemplo 4: Suponha que um piloto vença, em média, 30% de suas corridas. Sabe-se ainda que quando ele vence, 10% das vezes chove, e quando ele perde 25% das vezes chove. A probabilidade de chuva em uma corrida é de 15%. Qual a probabilidade de vitória dado que choveu? Usando o teorema de Bayes, representa-se o problema como:

$$\begin{aligned} P(vitória|chuva) &= \frac{P(chuva|vitória)P(vitória)}{P(chuva|vitória)P(vitória) + P(chuva|derrota)P(derrota)} \\ &= \frac{0,1 \times 0,3}{0,1 \times 0,3 + 0,25 \times 0,7} = \frac{0,03}{0,03 + 0,175} \approx 0,1463 = 14,63 \end{aligned}$$

Utilizando o teorema de Bayes e a informação a priori de chuva, obtêm-se uma probabilidade de vitória de aproximadamente 14,63%, menos da metade da probabilidade esperada quando esta informação é desconsiderada. Na abordagem bayesiana para estimação de parâmetros busca-se obter a distibuição a posteriori, que pode ser entendida uma medida de credibilidade para uma distribuição de parâmetros $\mathbf{\Theta}$ dadas observações $\mathbf{x}$.

$$p(\mathbf{\Theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{\Theta})p(\mathbf{\Theta})}{\int_{\Theta} p(\mathbf{x}|\theta)p(\theta)d\theta} \overset{\theta}{\propto} p(\mathbf{x}|\mathbf{\Theta})p(\mathbf{\Theta})$$

Desta forma, a distribuição a posteriori é proporcional ao produto da verossimilhança e da distribuição a priori, sendo uma combinação dos dados com o conhecimento inicial. A escolha pelo modelo bayesiano não se dá apenas quando se tem informação relevante a priori. Em alguns casos usa-se prioris chamadas de não informativas, permitindo uma estimação objetiva, uma vez que a informação relevante é obtida através dos dados.

Modelos frequentistas usam de estimativas pontuais e intervalos de confiança. Opta-se por utilizar o método bayesiano mesmo sem informação relevante a priori pois assim é possível obter a distribuição a posteriori que descreve detalhadamente o modelo de parâmetros, incluindo média, variância, e momentos da distribuição. Ainda, nem sempre é possível determinar de forma analítica a distribuição a posteriori, e é comum utilizar de métodos de Monte Carlo Markov Chain (MCMC) para gerar amostras da distribuição. Mais detalhes sobre a abordagem bayesiana podem ser consultados em [Bolstad e Curran 2016].

3 MODELOS PROPOSTOS

Neste capítulo será discutida a metodologia para modelagem dos tempos de volta no Campeonato Mundial de Fórmula 1 de 2022. Serão discutidas a estrutura dos dados, diferentes propostas de modelos, o procedimento de inferência realizado e critérios de comparação para os modelos.

3.1 Estrutura dos dados

As voltas de uma corrida de Fórmula 1 podem ser agrupadas por piloto, assim como as voltas de cada piloto podem ser agrupadas por circuito (Grande Prêmio). De maneira análoga, as voltas de um conjunto de corridas pode ser agrupada por circuito, e as voltas em cada circuitos podem ser agrupados por piloto.

Acontece que apesar de existirem 2 grupos não há 3 níveis de hierarquia. Isso se daria apenas se os pilotos pudessem ser aninhados por circuitos ou vice-versa, o que não é o caso neste estudo. Ao considerarmos um modelo linear misto levando em conta as influências dos pilotos e dos circuitos, temos um modelo com efeitos aleatórios cruzados, conforme mostrado abaixo. Vale ressaltar que esta estrutura é inerante aos dados, e independe do modelo adotado.

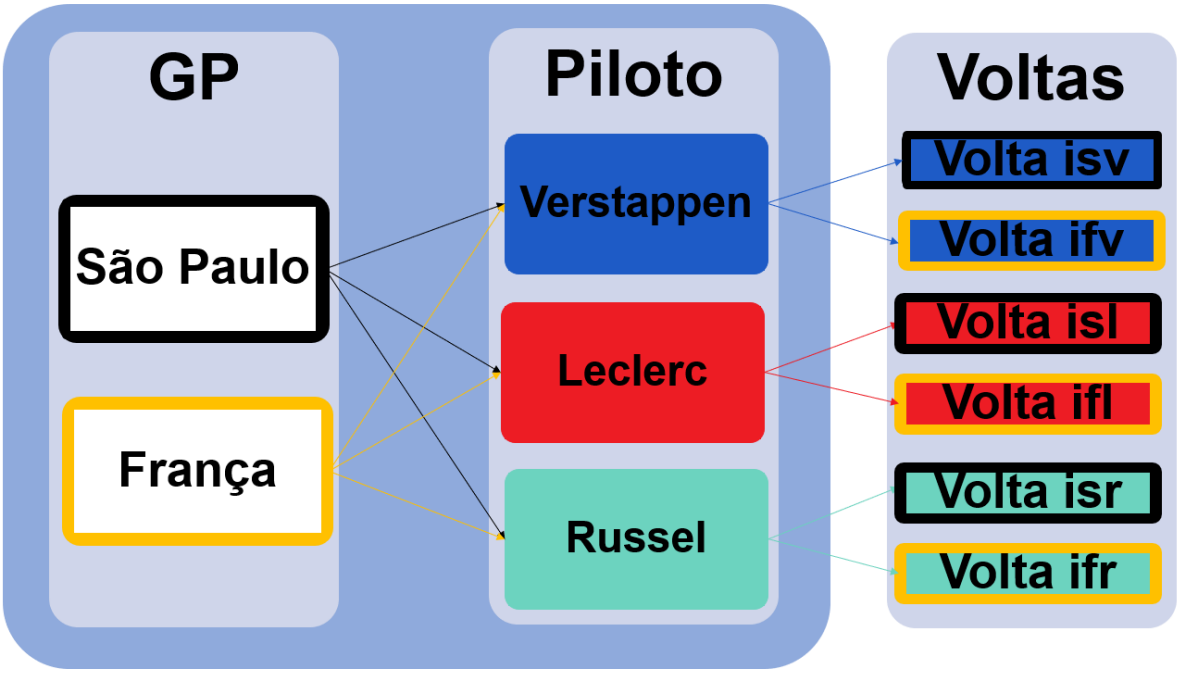

Figura 10 – Exemplo de voltas aninhadas por circuito e piloto de maneira cruzada

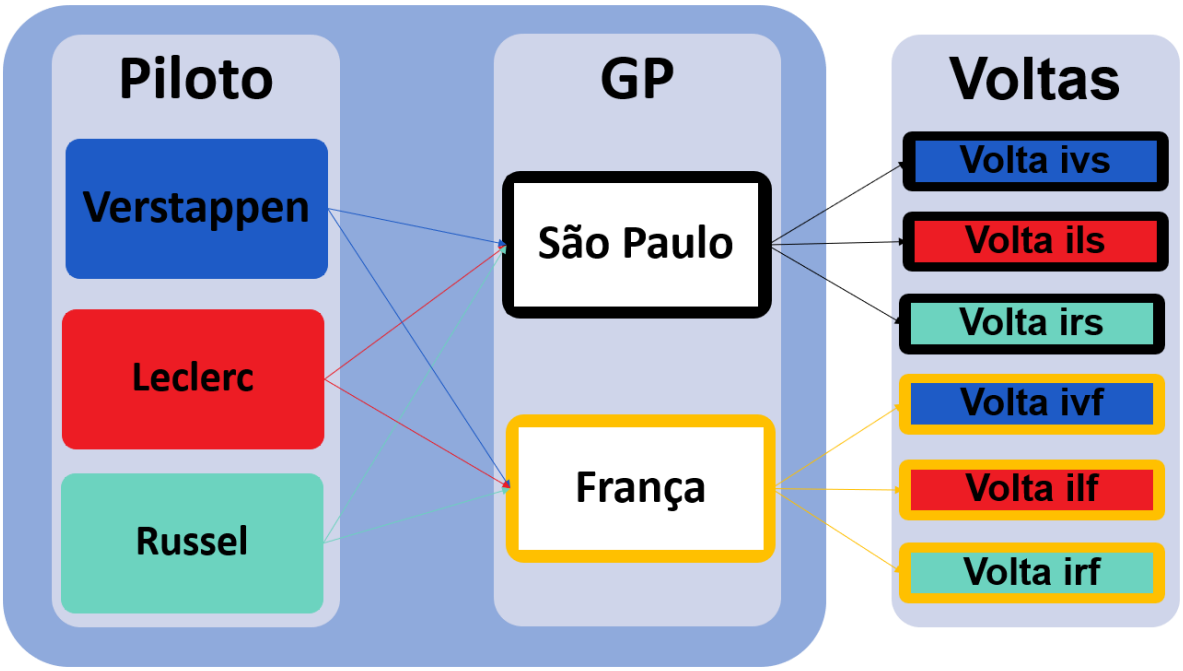

Figura 11 – Exemplo de voltas aninhadas por piloto e circuito de maneira cruzada

As Figuras 10 e 11 mostram que independentemente do agrupamento considerar piloto em cada evento ou evento em cada piloto, o objeto de estudo final não muda, são as mesmas voltas. Caso houvesse um aninhamento entre pistas e pilotos, teríamos pilotos diferentes disputando em cada um dos circuitos.

3.2 Modelos

Modelo 1 Assume-se que o tempo para completar a volta é dado em função de variáveis dependentes sem considerar efeitos específicos do circuito e do piloto. Trata-se, portanto, de um modelo linear múltiplo com variáveis preditivas apenas em um nível. A variável y_{ijc} representa a o tempo necessário para completar a i-ésima volta do piloto j no circuito c. As variáveis independentes consideradas foram (x_1) número da volta; (x_2) quantidade de voltas dadas pelo pneu em uso; e $(\boldsymbol{x_3})$ tipo de pneu em uso. Desta forma, temos que $x_{1ijc} = i$, e, portanto:

$$\mathrm{y}_{ijc} = \boldsymbol{\beta_0 x_{3}}_{ijc} + \beta_1 i + \beta_2 x_{2ijc} + \epsilon_{ijc} \quad \epsilon_{ijc} \sim N(0, \sigma^2) \tag{3.1.1}$$

Tem-se que $\boldsymbol{\beta_0}$ representa os diferentes interceptos do modelo e também o coeficiente angular da variável (x_3), β_1 e β_2 são os coeficientes angulares para as variáveis (x_1) e (x_2), respectivamente. Já ϵ_{ijc} representa o erro do modelo. $\boldsymbol{\beta_0}$ e $\boldsymbol{x_3}$ são vetores, uma vez que há 5 possibilidades de pneu.

Modelo 2 Assume-se um modelo linear misto permitindo que o intercepto varie por circuito em função de suas características específicas conhecidas. A saber, considerou-se a variável (x_5) tamanho do circuito, medido em km. Ainda, adiciona-se a variável *grid position*, que é única para cada combinação de circuito e piloto.

$$\begin{aligned} \mathrm{y}_{ijc} &= \boldsymbol{\beta_0}_{jc} + \beta_1 i + \beta_2 x_{2ijc} + \beta_4 x_{4ijc} + \beta_5 x_{5ijc} + \epsilon_{ijc} \quad \epsilon_{ijc} \sim N(0, \sigma^2) \\ \boldsymbol{\beta_0}_{jc} &= \boldsymbol{\beta_0 x_3}_{ijc} + \beta_4 x_{4jc} + \beta_5 x_{5c} \end{aligned} \tag{3.2.1}$$

$$\mathrm{y}_{ijc} = (\boldsymbol{\beta_0 x_3}_{ijc} + \beta_4 x_{4jc} + \beta_5 x_{5c}) + \beta_1 i + \beta_2 x_{2ijc} + \epsilon_{ijc} \tag{3.2.2}$$

Modelo 3 Assume-se um modelo linear misto assim como no **Modelo 2**, porém permite-se um efeito aleatório no nível do circuito. Este efeito explica as influências dadas por características do circuito não observadas. O modelo é dado por:

$$\begin{aligned} \mathrm{y}_{ijc} &= \boldsymbol{\beta_0}_{jc} + \beta_1 i + \beta_2 x_{2ijc} + \epsilon_{ijc} \quad \epsilon_{ijc} \sim N(0, \sigma^2) \\ \boldsymbol{\beta_0}_{jc} &= \boldsymbol{\beta_0 x_3}_{ijc} + \beta_4 x_{4jc} + \beta_5 x_{5c} + u_c \quad u_c \sim N(0, \sigma_u^2) \end{aligned} \tag{3.3.1}$$

$$\mathrm{y}_{ijc} = (\boldsymbol{\beta_0 x_3}_{ijc} + \beta_4 x_{4jc} + \beta_5 x_{5c} + u_c) + \beta_1 i + \beta_2 x_{2ijc} + \epsilon_{ijc} \tag{3.3.2}$$

Modelo 4 Adiciona-se ao **Modelo 3** outro efeito aleatório, desta vez no nível do piloto. Este efeito explica as influências dadas por características não observadas do piloto.

$$\begin{aligned} \mathrm{y}_{ijc} &= \boldsymbol{\beta_0}_{jc} + \beta_1 i + \beta_2 x_{2ijc} + \epsilon_{ijc} \quad \epsilon_{ijc} \sim N(0, \sigma^2) \\ \boldsymbol{\beta_0}_{jc} &= \boldsymbol{\beta_0 x_3}_{ijc} + \beta_4 x_{jc} + \beta_5 x_{5c} + u_c + v_j \quad u_c \sim N(0, \sigma_u^2) \quad v_j \sim N(0, \sigma_v^2) \end{aligned} \tag{3.4.1}$$

$$\mathrm{y}_{ijc} = (\boldsymbol{\beta_0 x_3}_{ijc} + \beta_4 x_{4jc} + \beta_5 x_{5c} + u_c + v_j) + \beta_1 i + \beta_2 x_{2ijc} + \epsilon_{ijc} \tag{3.4.2}$$

Modelo 5 Adiciona-se ao **Modelo 4** um efeito aleatório no coeficiente angular da variável ($x1$) que representa o número de voltas.

Tal efeito foi escolhido pelo fato de cada circuito ter um número específico de voltas. Com isso, a preparação para as corridas varia, principalmente na quantidade de combustível em cada veículo. Por exemplo, estar na 44ª volta no Grande Prêmio da Bélgica significa estar terminando a corrida, e o carro encontra-se quase sem gasolina e muito mais leve. No GP do Brasil, com 71 voltas para completar a corrida, esta relação é diferente.

$$\begin{aligned} \mathrm{y}_{ijc} &= \boldsymbol{\beta_0}_{jc} + \beta_{1c} i + \beta_2 x_{2ijc} + \epsilon_{ijc} \quad \epsilon_{ijc} \sim N(0, \sigma^2) \\ \boldsymbol{\beta_0}_{jc} &= \boldsymbol{\beta_0 x_3}_{ijc} + \beta_4 x_{jc} + \beta_5 x_{5c} + u_c + v_j \quad u_c \sim N(0, \sigma_u^2) \quad v_j \sim N(0, \sigma_v^2) \\ \beta_{1c} &= \beta_1 + w_c \quad w_c \sim N(0, \sigma_w^2) \end{aligned} \tag{3.5.1}$$

$$\mathrm{y}_{ijc} = (\boldsymbol{\beta_0 x_{3}}_{ijc} + \beta_4 x_{4jc} + \beta_5 x_{5c} + u_c + v_j) + (\beta_1 + w_c)i + \beta_2 x_{2ijc} + \epsilon_{ijc} \tag{3.5.2}$$

Observação: Não foi testado modelo com covariáveis apenas no nível do piloto uma vez que a competição de Fórmula 1 é muito dinâmica e decidida muitas vezes por detalhes tecnológicos. Ex-campeões como Alonso e Vettel, com muitos anos de experiência, no ano de 2022 não brigavam pelo título muito por conta do carro de suas equipes. Equipes dominantes também mudam constantemente ao longo das temporadas.

3.3 Procedimento de Inferência

Inferência realizada sobre a metologia bayesiana. Com base no teorema de Bayes obteve-se a distribuição a posteriori dos parâmetros do modelo, proporcional ao produto da função de verossimilhança e a distribuição a posteriori. Foram escolhidas prioris não informativas. Vale ressaltar que $\mathrm{y}_i \sim N(\mu, \sigma^2)$. No modelo 5, por exemplo, $\mu = (\beta_0 + \beta_4 x_{jc} + \beta_5 x_{5c} + u_c + v_j) + (\beta_1 + w_c)i + \beta_2 x_{2ijc} + \boldsymbol{\beta_3 x_{3}}_{ijc}$ e $\sigma^2 = Var(\epsilon_{ijc})$ satisfazem esta especificação.

Dessa forma, foram escolhidas prioris da distribuição Normal, pois esta combinação possui características interessantes que facilitam a análise do modelo. Sejam N_c o número de circuitos, N_j o número de pilotos, e $N_{i[jc]}$ o número de voltas dadas pelo piloto j no circuito c. A função de verossimilhança $l(\boldsymbol{\Theta}; \boldsymbol{y})$ e a distribuição a priori podem ser descritas como:

$$l(\boldsymbol{\Theta}; \mathbf{y}_{ijc}) = \prod_{c=1}^{N_c} \prod_{j=1}^{N_j} \prod_{i=1}^{N_i} p(\mathbf{y}_{ijc}|\boldsymbol{\Theta})$$

$$p(\boldsymbol{\Theta}) = \prod_{c=1}^{N_c} \prod_{j=1}^{N_j} \prod_{i=1}^{N_i} p(\mu, \sigma^2|\beta_{0jc}, \beta_{1c}, \beta_2, \boldsymbol{\beta_3}) p(\beta_{0jc}) p(\beta_{1c}) p(\beta_2) p(\boldsymbol{\beta_3}) p(\sigma^2)$$

Obtém-se a distribuição a posteriori $p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{x}) \overset{\theta}{\propto} l(\boldsymbol{\Theta}; \mathbf{y}_{ijc}) \times p(\boldsymbol{\Theta})$. Utilizou-se métodos de Monte Carlo via Cadeia de Markov para amostrar da distribuição a posteriori através do software JAGS [Plummer 2003] por meio de um pacote da linguagem de programação R [R Core Team 2022]. Para gerar as amostras das distribuições a posterioir utilizou-se 2 cadeias e 15000 iterações, sendo as primeiras 5000 descartadas. Ainda, a cada 10 iterações foi armazenada 1 delas, sendo o total de iterações salvas para cada observação dado por $2 \times \frac{15000-5000}{10} = 2000$.

3.4 Critérios de Comparação

3.4.1 DIC

O *Deviance Information Criterion* (DIC) é uma medida de seleção de modelo que é amplamente utilizada em modelos Bayesianos. Ele foi proposto por Spiegelhalter et al [Spiegelhalter et al. 2002] e é definido como a diferença entre o deviance médio da amostra observada e o deviance médio esperado pelo modelo.

O deviance é uma medida para verificar a qualidade do ajuste do modelo com base nos dados observados. Ele é dado por $D = -2 \log L$, com L representando a verossimilhança do modelo, quanto maior o valor de L (e consequentemente menor o valor de D), melhor o ajuste aos dados observados. O DIC é dado por $DIC = D_{médio} + p_D$, com $D_{médio}$ representando o deviance médio da amostra observada e p_D é uma medida para o número de parâmetros do modelo, de forma a penalizar modelos mais complexos. Quanto menor o valor do DIC, melhor o ajuste aos dados observados.

O DIC também fornece um intervalo de confiança para o deviance, permitindo melhor controle da incerteza. Ele é utilizado sob o paradigma bayesiano pois permite avaliar o modelo sem depender de um conjunto específico de parâmetros, uma vez que é calculado a partir da distribuição a posteriori dos mesmos.

3.4.2 Erro Quadrático Médio

O erro quadrátivo médio (EQM) é uma medida de erro definada como a média dos quadrados da diferença entre os valores previstos pelo modelo e os valores observados. Ou seja, é o mesmo utilizado no método dos mínimos quadrados visto anteriormente, representado algebricamente por $\frac{1}{n}\sum_{i=1}^{n}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2$.

Uma de suas características é penalizar mais fortemente os erros maiores, uma vez que todas as diferenças são elevadas ao quadrado e depois somadas. Desta forma, uma diferença de 100 adicionaria 10000 ao EQM, enquanto 10 diferenças de 10 adicionariam 1000. Sua interpretação é bem direta, de forma geral, quanto menor o EQM melhor o ajuste ao modelo.

Essa relação, no entanto, assim como em todos os critérios de comparação, deve considerar se não está ocorrendo *overfitting*.Um EQM muito próximo a 0 pode indicar que o modelo está especifico aos dados utilizados e não seria possível generalizá-lo. Vale ressaltar que $\frac{1}{n}\sum_{i=1}^{n}(|\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}|)$ também seria uma alternativa, e é conhecido por erro absoluto médio (EAM). Além de penalizar mais erros maiores, o EQM é diferenciável, fazendo com que sua escolha seja preferível em muitos casos.

4 RESULTADOS

4.1 Comparação de Modelos

Tabela 6 – Critérios de comparação para os modelos ajustados

Modelo	$D_{médio}$	p_D	DIC	EQM (milisegundos^2)
Modelo 1	396621,5	1157,0	397778,5	130593,3
Modelo 2	387336,2	7045,9	394382,1	73919,0
Modelo 3	269660,7	41,4	269702,1	54,4
Modelo 4	266256,3	58,3	266314,5	44,1
Modelo 5	264150,5	75,2	264225,7	38,8

O Modelo 5 apresenta os menores valores de DIC e EQM, sendo por estes critérios o que melhor ajusta os dados ao modelo. Isso não é muito informativo, uma vez que os modelos foram agregando variáveis e efeitos em ordem crescente, sendo o Modelo 1 o mais básico e o Modelo 5 o mais complexo, portanto esse comportamento já era esperado.

Caso fosse adicionada alguma informação não relevante ao modelo, o ajuste seria de forma a considerá-la não significativa. Vale destacar, porém, a inclusão do intercepto do circuito, feita inicialmente no Modelo 3, que foi capaz de melhorar consideravelmente o ajuste. Ainda, embora os valores de $D_{médio}$ tenham diminuído conforme foram incluídas variáveis nos modelos, o mesmo não ocorre para p_D, que cresce nos modelos 2, 4 e 5, penalizando a complexidade dos mesmos. No entanto, o DIC permaneceu decrescente, sendo um indicativo de que a penalização foi compensada pela melhoria no ajuste.

4.2 Análises dos resultados

A seguir serão analisados os resultados com base no **Modelo 5**. A Figura 12 mostra 2 efeitos esperados: o número da volta impacta inversamente o tempo da volta, ou seja, quanto maior o número da volta, menor o tempo para completá-la, e quanto maior a vida do pneu, maior o desgaste e menor a aderência, acarretando maiores tempos, numa relação direta.

Os pneus mais duros também acarretam em maiores tempos no modelo, ainda que esta relação seja mais complexa e dependente de outros fatores como circuito e estratégia do piloto. Embora os intervalos de credibilidade dos pneus C3 e C5 incluam o 0, sendo portanto não significativos, a variável tipo de pneu inclui os 5 tipos e faz sentido considerá-los.

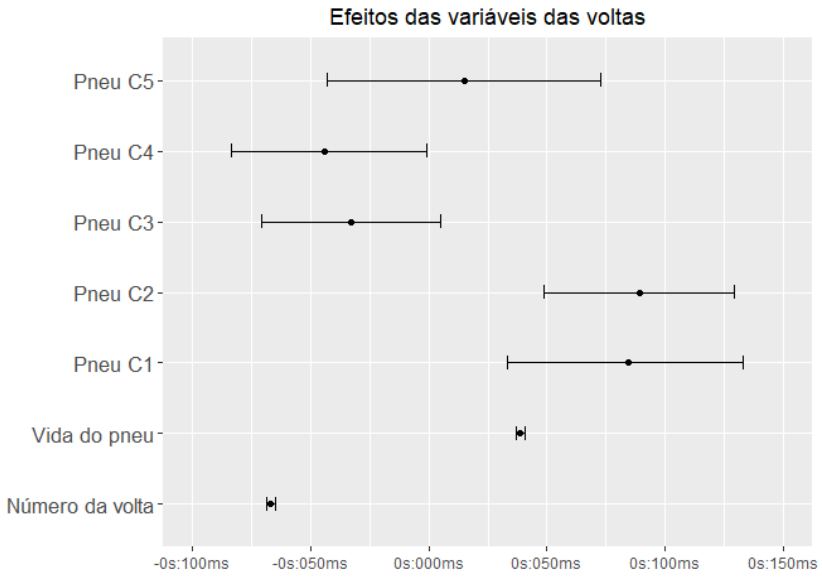

Figura 12 – Distribuições a posteriori dos coeficientes de regressão das variáveis número da volta, vida do pneu e tipo de pneu, médias e intervalos de 95% de credibilidade

A Figura 13 mostra também dois comportamentos esperados. Tanto o tamanho da curva quanto a posição no grid possuem uma relação direta com o tempo de completar a volta. O *grid position* é definido de acordo com os tempos no próprio circuito da corrida dos pilotos, ficando os melhores colocados nas primeiras posições, então é esperado que os últimos tenham um tempo mais alto. Vale ressaltar que o intervalo de 95% de credibilidade do *grid position* é muito menor que o do tamanho do circuito, estando seu efeito esperado mais concentrado.

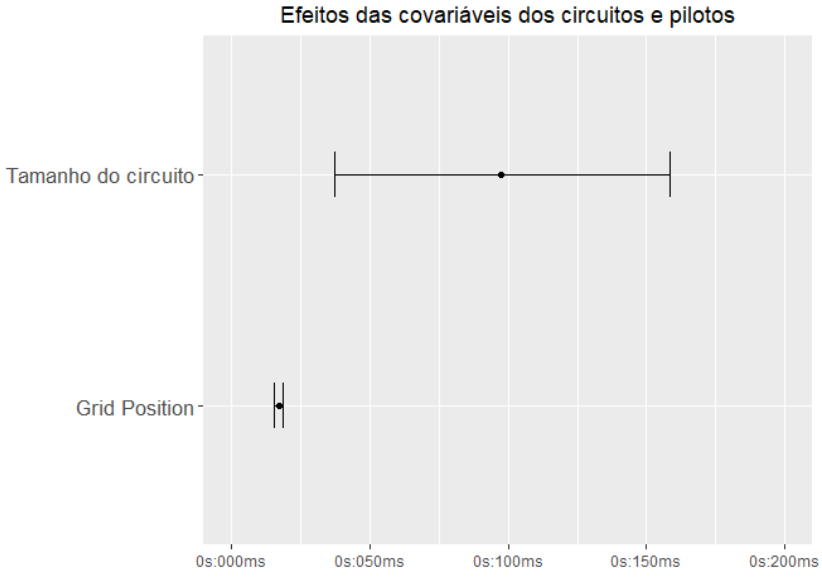

Figura 13 – Distribuições a posteriori dos coeficientes de regressão das variáveis *grid position* e tamanho do circuito, médias e intervalos de 95% de credibilidade

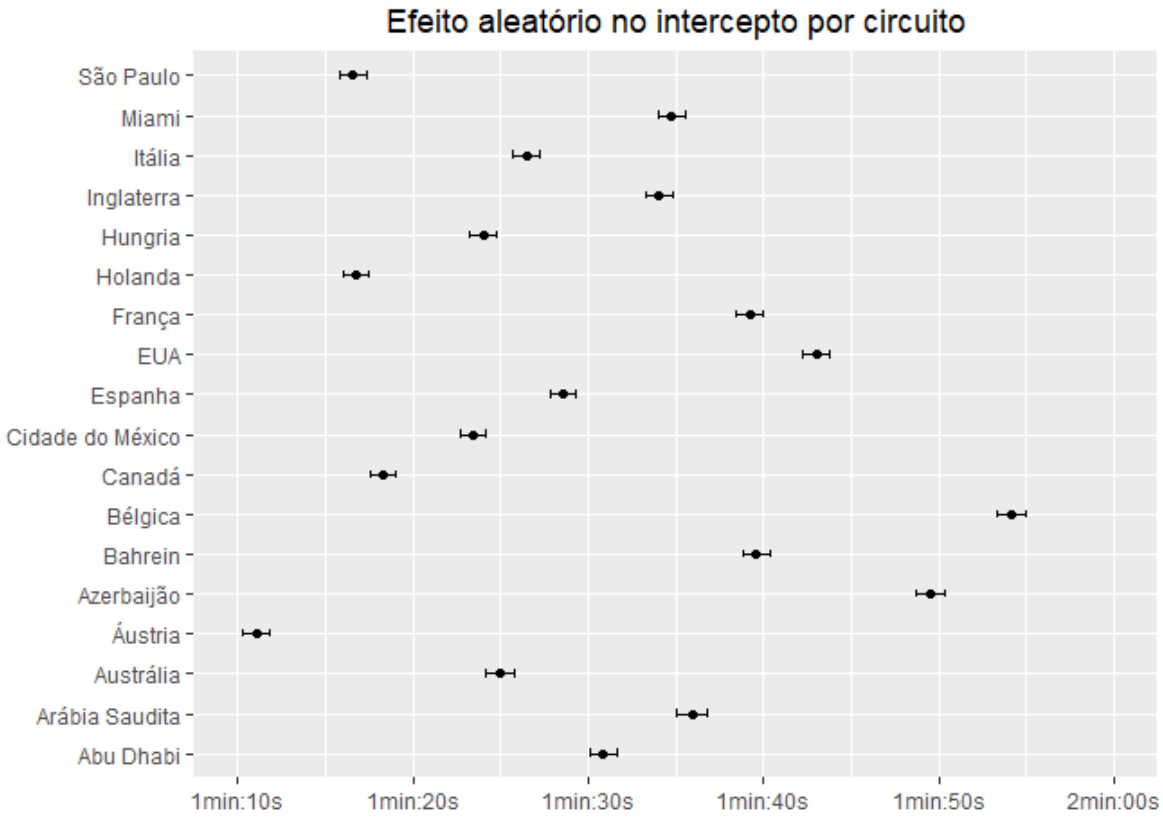

Figura 14 – Distribuições a posteriori dos efeitos aleatórios no intercepto do circuito, médias e intervalos de 95% de credibilidade

O efeito aleatório nos intercepto por circuito refletem a Figura 1, onde foi verificado um comportamento bastante heterogênio dos tempos de voltas em diferentes circuitos. Este efeito aleatório é interpretado como uma média para os diferentes circuitos. Foi visto na definição do modelo que a variável tipo de pneu compõe o intercepto.

No entanto, como visto nas figuras 11 e 12, os intervalos de 95% de credibilidade para os coeficientes dos diferentes pneus estão -0,2 e 0,2 segundos e os intervalos de credibilidade para os coeficientes das variáveis *grid position* e tamanho do circuito estão entre 0 e 0,2 segundos. O coeficiente dos efeitos aleatórios no intercepto por circuito, porém, estão entre 1 minuto e 10 segundos e 2 minutos, com pelo menos 95% de credibilidade, sendo responsável pela maior parte das médias da voltas.

Já o efeito aleatório nos interceptos por piloto na Figura 15 refletem a Figura 3, que indicava heterogeneidade nos tempos de volta dos diferentes pilotos. Este efeito aleatório somado aos efeitos nos coeficientes explicados anteriormente são interpretados como as médias dos diferentes pilotos. Os intervalos de credibilidade de 95% ficam entre -2 e 2 segundos e é fácil notar diversos intervalos cujas interseções são vazias, corroborando a importância da inclusão deste efeito no modelo.

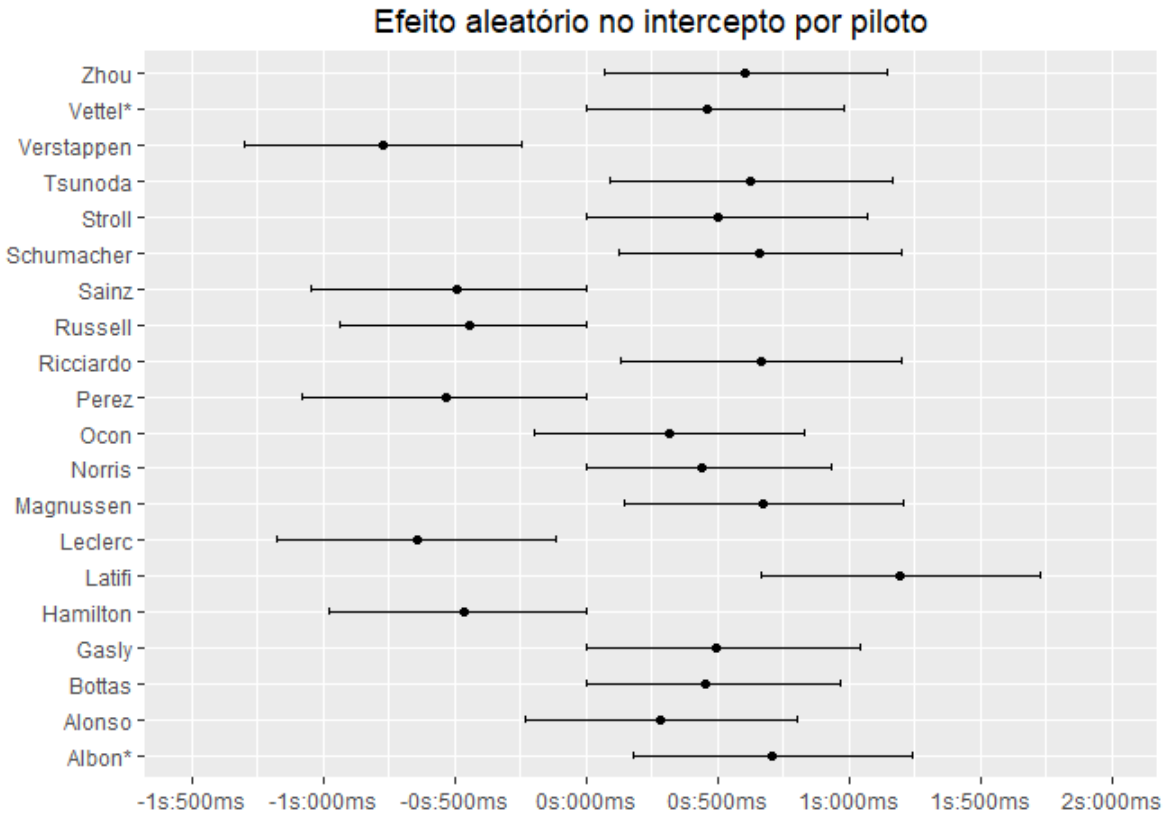

Figura 15 – Distribuições a posteriori dos efeitos aleatórios no intercepto do piloto, médias e intervalos de 95% de credibilidade

O efeito aleatório no coeficiente angular da variável número da volta é de interpretação mais delicada. Primeiramente, conforme visto na Figura 12, o intervalo de 95% de credibilidade para o número da volta está contido no intervalo de -0,05 e -0,1 segundos. Desta forma, como na Figura 16 verifica-se que os efeitos aleatórios nos coeficientes por circuito encontram-se entre -0,05 e 0,05 segundos, sem incluir 0,05 segundos, a soma dos efeitos é sempre negativa, e o comportamento esperado do tempo da volta diminuir ao longo da corrida é mantido.

Ainda, quanto menor o efeito na Figura 16, maior será o impacto do número da volta no tempo para completá-la, ou seja, no GP dos EUA espera-se que os tempos entre uma volta e outra caiam mais consideravelmente que nos demais. Os intervalos para alguns circuitos inclui o 0, mas, novamente, isso não é um problema pois é preciso analisar o efeito do fator circuito como um todo.

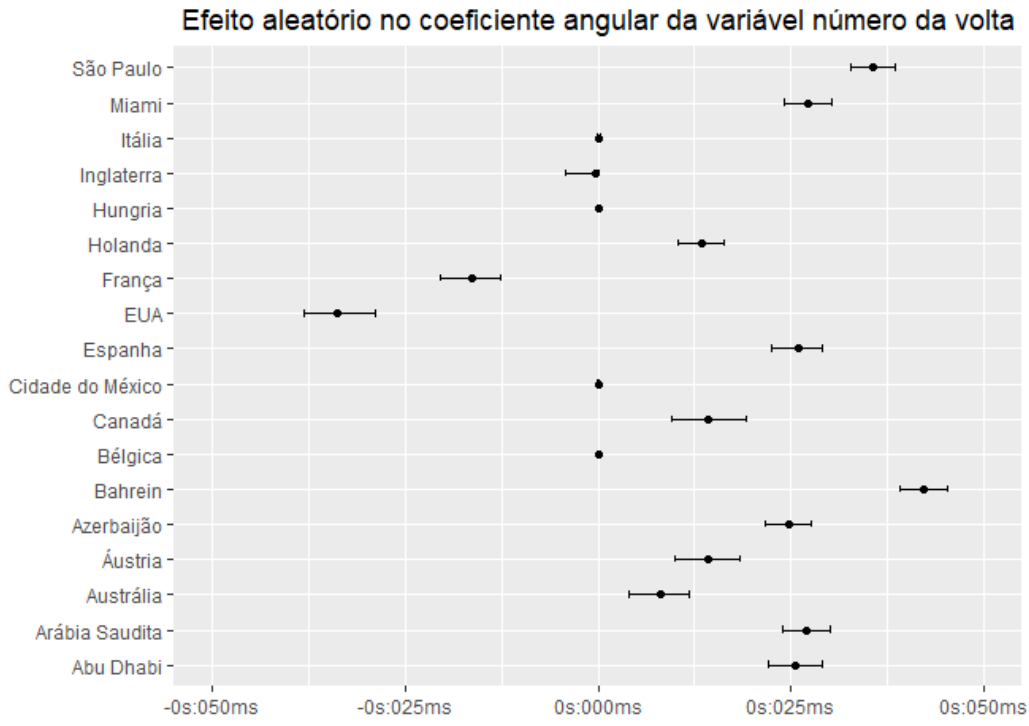

Figura 16 – Distribuições a posteriori dos efeitos aleatórios no coeficiente angular da variável número da volta, médias e intervalos de 95% de credibilidade

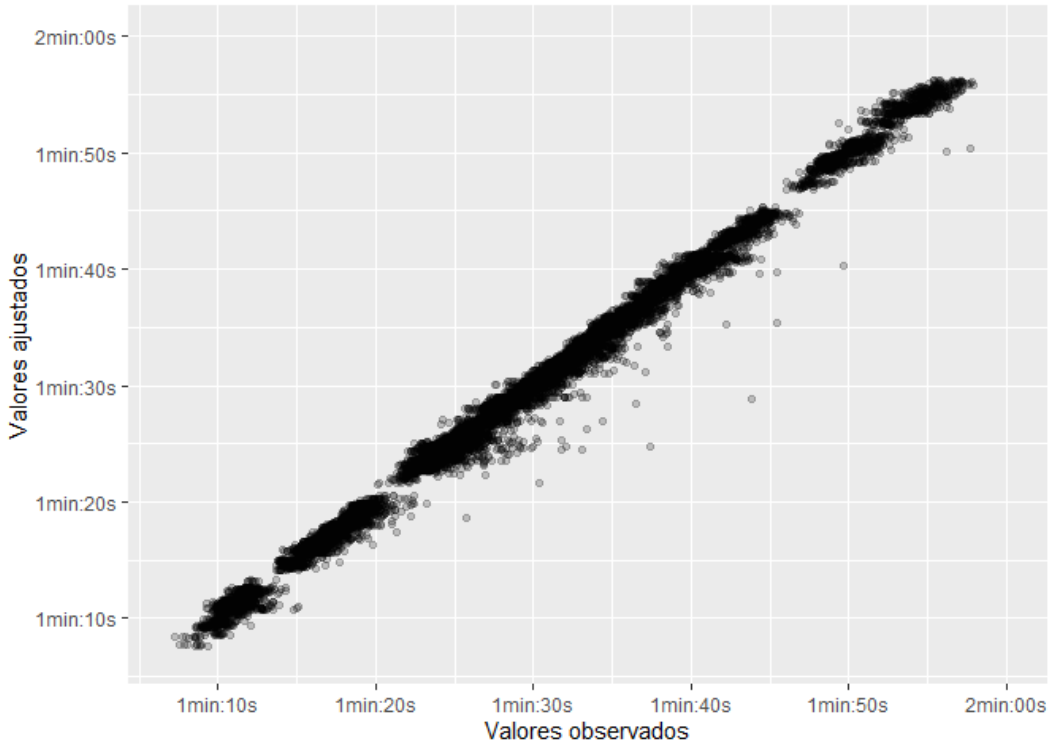

Figura 17 – Gráfico da média a posteriori representando os valores ajustados versus valores observados

A Figura 17 mostra resumidamente a capacidade do modelo de reproduzir os dados. É possível verificar que o gráfico se comporta, no geral, como uma reta, indicando um bom ajuste. Uma observação a ser feita é que os pontos que fogem deste comportamento são predominantemente casos nos quais os valores observados são maiores que os ajustados. Isso era esperado, pois são comuns na Formula 1 incidentes que acarretam em maiores tempos de voltas.

Conforme mencionado na etapa de tratamento dos dados, e também facilmente identificáveis na Figura 7, foram removidas algumas voltas com bases nos critérios já mencionados, porém outras que poderiam ser classificadas como outliers numa análise mais criteriosa foram mantidas neste estudo. Por outro lado, as voltas mais rápidas pouco se destacam, pois devido a alta competitividade a diferença para uma volta mediana no mesmo circuito costuma ser uma fração de segundo.

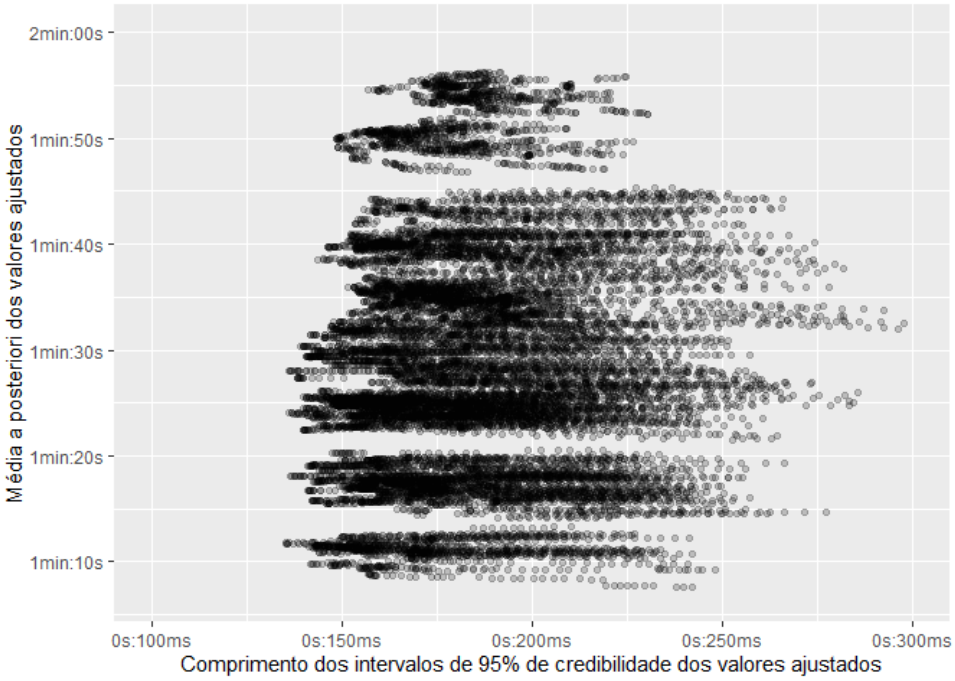

Figura 18 – Gráfico da média a posteriori representando os valores ajustados versus valores observados

Na Figura 17 foi utilizada apenas a média a posteriori dos valores ajustados uma vez que os comprimentos dos intervalos de credibilidade de 95% para os valores ajustados, dados pelas diferenças entre os quantis 97,5% e o 2,5% são muito curtos, não sendo possível visualizar adequadamente os intervalos de credibilidade no gráfico de valores ajustados versus observados.

Isso pode ser verificado na Figura 18, uma vez que os valores ajustados variam entre 1 e 2 minutos enquanto os comprimentos dos intervalos de credibilidade variam entre 0,1 e 0,3 segundos. Este comportamento condiz com as corridas uma vez que frações de segundos são determinantes em cada volta.

5 CONSIDERAÇÕES FINAIS

Falar sobre anos diferentes Falar sobre teste e treino

REFERÊNCIAS

BOLSTAD, W. M.; CURRAN, J. M. *Introduction to Bayesian statistics*. [S.l.]: John Wiley & Sons, 2016. Citado na página 19.

CHARNET, R. et al. Análise de modelos de regressão linear com aplicações. *Campinas: Unicamp*, 1999. Citado na página 16.

GELMAN, A.; HILL, J. *Data analysis using regression and multilevel/hierarchical models*. [S.l.]: Cambridge university press, 2006. Citado na página 16.

PIRELLI. *F1 tires: details and technical data*. <http://https://web.archive.org/web/20221020080408/https://www.pirelli.com/tires/en-us/motorsport/f1/tires>. Accessed: 2022-12-18. Citado na página 8.

PLUMMER, M. *JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling*. 2003. Citado na página 23.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>. Citado na página 23.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. Citado na página 24.

TREMAYNE, D. *The science of formula 1 design: expert analysis of the anatomy of the modern grand prix car*. [S.l.]: Haynes North America, 2006. Citado na página 1.