Faculty of Science, Technology and Medicine

Overtaking in Formula 1

Fiona Seny

Bachelor thesis submitted to obtain the degree of Bachelor in Mathematics

Supervised by Prof. Dr. Christophe LEY

Academic year 2024-2025

Contents

1 Introduction **1**
- 1.1 Regression . . . 2
- 1.2 Poisson regression . . . 10
- 1.3 Log likelihood . . . 11

2 Overtaking in Formula 1 **15**
- 2.1 Formula 1 . . . 15
  - 2.1.1 Introduction in F1 . . . 15
  - 2.1.2 Rules and Regulations . . . 17
- 2.2 Theoretic framework . . . 18
- 2.3 Factors influencing the potential of overtaking . . . 19
- 2.4 Fixed effects models . . . 20
  - 2.4.1 Track fixed effects . . . 25
  - 2.4.2 Season fixed effects . . . 27
  - 2.4.3 Race and driver race-specific fixed effects . . . 27
- 2.5 Model . . . 31
- 2.6 Baseline results from the article (de Groote, 2021) . . . 33
- 2.7 Conclusion . . . 37

3 Simulation of data **38**
- 3.1 Synthetic data . . . 38
- 3.2 Python Code: F1 2024 Season Simulation . . . 38
  - 3.2.1 Python code for the simulation of data . . . 39
  - 3.2.2 Analysis of the Python code . . . 43
- 3.3 Ranking of the drivers based on our simulation . . . 45
  - 3.3.1 Python code used for the ranking . . . 45
  - 3.3.2 Python code used for the points distribution . . . 46
  - 3.3.3 Results . . . 47
- 3.4 Estimation of the different parameters . . . 48
  - 3.4.1 Python code used for the estimation of the parameters . . . 48
  - 3.4.2 Representation of the estimated parameters . . . 49
  - 3.4.3 Explanation and conclusion about the simulation of data and estimation of the parameters . . . 50

Sources **52**

1 Introduction

This thesis is about overtaking in Formula 1, the world's most prestigious motor racing competition. We aim to analyze which different factors such as DRS, pit-stops, tires, rules,... influence the number of overtakes during a Formula 1 race.

Therefore, we will introduce in section 2 the sport Formula 1 and explain the different rules and regulations. Moreover, we will explain the different fixed effects used in our analysis about overtaking.

Since our analysis is based on the article "Overtaking in Formula 1 during the Pirelli era: A driver-level analysis" (de Groote, 2021), we will explain the model and the main baseline results of the article. We use the Poisson regression model to determine the expected number of places gained and lost by a driver on a specific track and during a specific season.

In section 3, we will generate synthetic data following the Poisson distribution. We simulate data of the 2024 Formula 1 season, consisting of 24 races. Moreover, using our simulated data we are able to rank the drivers depending on their average performance. In addition, we will estimate the different parameters of the regression model, we used to generate our synthetic data.

Before starting with the analysis of the data about overtaking in Formula 1, we will give an explanation of regression, especially Poisson regression. In addition, we will describe two methods to estimate the parameters used in the Poisson regression i.e. least squares estimation and maximum likelihood estimation. Moreover, we will explain the likelihood function.

1.1 Regression

Regression is a statistical method, which is used in different disciplines such as finance, testing automobiles, neuroscience... etc. This method helps us to find a relation between a dependent variable, which is the main factor we are trying to understand or predict, and one or more independent variables (explanatory variables), which are the factors that may have an impact on our dependent variable. To fit a linear model to observed data, we need to determine if there exists a relation between the variables of interest, which means that there exists some significant association between the two variables. In addition, we use the correlation coefficient, a value between -1 and 1, to determine the strength of the association between the variables. On the one hand, if we have a negative correlation value, we have a negative association between variables, which means that increasing values in one variable corresponds to decreasing values in the other variable. On the other hand, if we have a positive correlation value, we have a positive association, which means that increasing values in one variable corresponds to increasing values in the other variable. However, if the correlation value is close to 0, we have no association between the variables.
Let's consider a set of observations $(x_1, y_1), (x_2, y_2)...(x_n, y_n)$, we can compute the correlation coefficient using the following formula:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)}},$$

where S_x, S_y are the sample standard deviation, S_{xy} is the sample covariance and $\bar{x}, \bar{y}$ are the sample means.
Regression analysis is used to find trends in the data and provides us with an equation for a graph, used to make predictions about our data.

Moreover, we often use linear regression, which creates a linear relationship between two variables. The graph of linear regression is a straight line, where the slope defines how the change in one variable impacts the other one.
We have the following definition of Simple Linear Regression Model: We have the parameters β_0, the intercept of the true regression line (average value of Y when $x = 0$), β_1, the slope of the true regression line (expected average change in Y associated with a one unit increase in the value of x) and the variance σ^2. For any fixed value of the independent variable X, the dependent variable Y is a random variable following the model equation:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where the quantity ϵ is the 'error' (random deviation), a random variable assumed to be symmetrically distributed with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma_\epsilon^2 = \sigma^2$. The reasons for this error ϵ can for example be outliers and high leverage points, correlation of error terms, collinearity or an effect of deleted variables in the model. Let us recall the definition of an outlier and a high leverage point.

An outlier in a regression model is a point with an extreme y-value relative to the regression line. A high leverage point is a data point with an extreme x-value relative to the other data points. However, this point may still follow the general linear trend of the data (Figure 1).
Moreover, a high leverage point can also be considered as an outlier (Figure 1) if, in addition to the extreme x-value relative to the other data points, we also have an extreme y-value relative to the regression line. If a high leverage point still follows the general linear trend (extreme x-value still lies close to the regression line), it can not be an outlier.

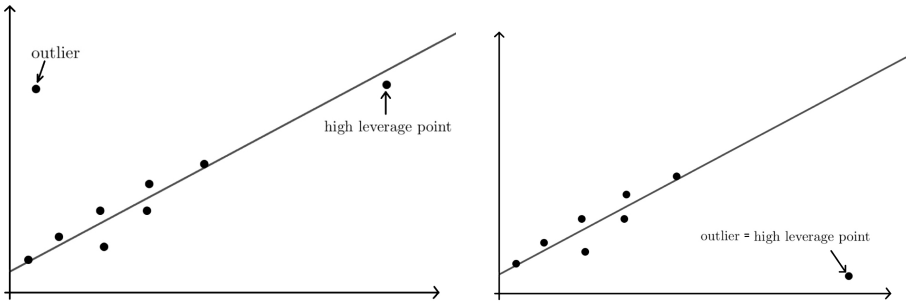

Figure 1: The first graph shows the situation where we have an outlier and a high leverage point. In the second graph, we have the situation where a high leverage point is an outlier.

Let us look at a graph of a true regression line (Figure 2), by considering n independent observations, where the points $(x_1, y_1), \ldots (x_n, y_n)$ are scattered around the line.

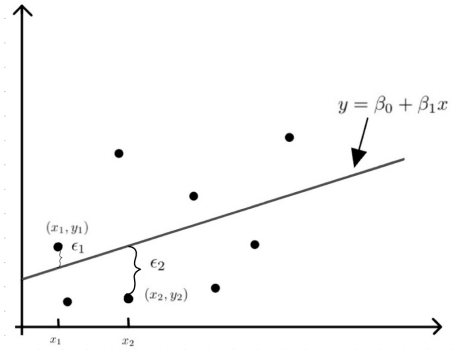

Figure 2: Graph of a true regression line

We can calculate the sample mean and variance of Y, when X is viewed as controlled by the experimentator or when X is considered to be a random variable. In the first case, we have the following equations: $E(Y) = \beta_0 + \beta_1 X$ and $Var(Y) = \sigma^2$. In the second case, we consider the conditional mean and conditional variance of Y given $X = x$ as follows: $E(Y|x) = \beta_0 + \beta_1 x$ and $Var(Y|x) = \sigma^2$.
Moreover, our model is completely described when we know the values of β_0, β_1 and σ^2. However, these parameters are generally unknown and ϵ is unobserved. The determination of the linear regression model depends on the estimation of the parameters. In addition, to determine the parameters, there are various methods, such as least squares estimation or maximum likelihood estimation.

Least squares estimation
For the least squares estimation we consider a sample of sets of n observation, $(x_1, y_1), \ldots, (x_n, y_n)$, which satisfy the simple linear regression model. Hence we can write:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{with } i \in \{1, \ldots, n\}$$

To estimate the parameters β_0 and β_1, we minimize the sum of squares of the difference between the observations and the line in the scatter diagram. Recall that a scatter diagram pairs numerical data and creates a relation between them. The diagram has one variable on each axis. For the least square estimation we have three types of differences depending on the perspective. Firstly, we have the vertical difference between the observations and the line of the scatter diagram (Figure 3), which is also called method of direct regression.

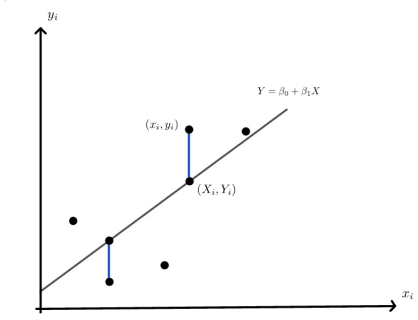

Figure 3: vertical difference between the observations and the line of the scatter diagram

This direct regression method minimizes the sum of squares, using the following formula:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2, \quad \text{with respect to } \beta_0 \text{ and } \beta_1.$$

Moreover, we can calculate the partial derivatives of $S(\beta_0, \beta_1)$ with respect to β_0 and β_1:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i$$

Hence, we obtain the solutions of β_0 and β_1 by setting

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0 \text{ and } \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

Let us solve both equations:

$$\begin{aligned}
& \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \\
\Leftrightarrow \quad & -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\
\Leftrightarrow \quad & n\bar{y} - n\beta_0 - n\beta_1 \bar{x} = 0 \\
\Leftrightarrow \quad & \beta_0 = \bar{y} - \beta_1 \bar{x}
\end{aligned}$$

and:

$$\begin{aligned}
& \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0 \\
\Leftrightarrow \quad & -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \beta_0 x_i - \sum_{i=1}^{n} \beta_1 x_i^2 = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} y_i x_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \bar{y} x_i + \beta_1 \sum_{i=1}^{n} \bar{x} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} x_i (y_i - \bar{y}) + \beta_1 \sum_{i=1}^{n} x_i (\bar{x} - x_i) = 0 \\
\Leftrightarrow \quad & \beta_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} x_i (y_i - \bar{y}) \\
\Leftrightarrow \quad & \beta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}
\end{aligned}$$

Hence, we have $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$.

Moreover, let us have a look at the coefficient R-squared, noted R^2. R^2 is defined as a number, indicating how well the independent variable in a statistical model explains the variation in the dependent variable. This number ranges from 0 to 1 (Figure 4).

If $R^2 = 1$, the model is a perfect fit to the data, i.e. all the variability in the dependent variable is explained by the independent variable.
If $R^2 = 0$, the independent variable does not explain the variability of the dependent variable.
If R^2 is a number between 0 and 1, we can interpret the R^2 coefficient as follows: "$R^2 \cdot 100\%$ of the variation in y is reduced by taking into account predictor x".

Note that if the model is overfitted, the R^2 value can be misleading. Therefore, we should use the R^2 value with other statistics and context.
Let us examine a figure showing a weak relation between y and x (Figure 5).
We have two lines, a horizontal line placed at the average response $\bar{y}$. The second line is the estimated regression line $\hat{y}$. Moreover, the slope of $\hat{y}$ is not very steep, hence as x increases we do not have much of a change in the average response y.

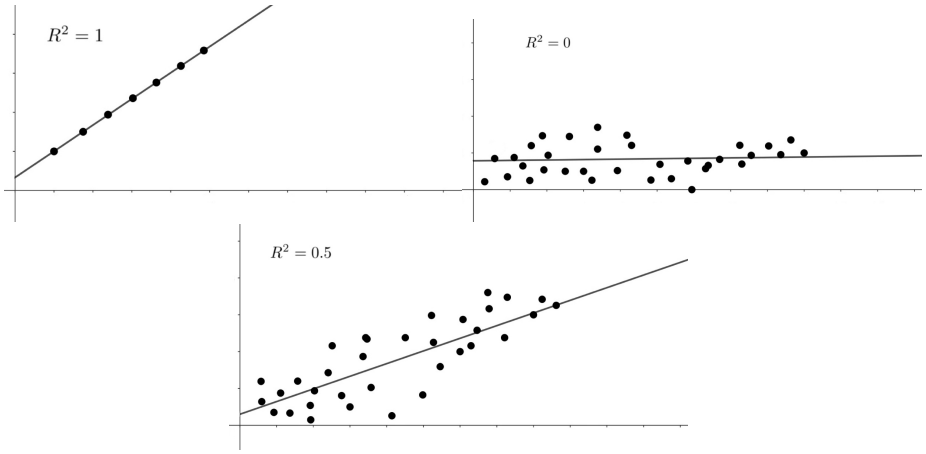

Figure 4: The first figure on the left shows the situation where the model is a perfect fit to the data, hence $R^2 = 1$. The picture on the right displays the situation where $R^2 = 0$, meaning that the independent variable does not influence the dependent variable. The last picture shows the situation where R^2 is a number between 0 and 1.

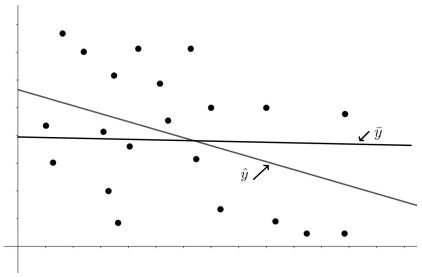

Figure 5: Regression plot

We have the following formula to calculate the coefficient R^2:

$$R^2 = 1 - \frac{\text{unexplained variation}}{\text{total variation}}.$$

In addition, we can also use the following notation for the formula of the coefficient R^2:

$$R^2 = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}},$$

SSR is the regression sum of squares, which quantifies how far the estimated sloped regression line $\hat{y}_i$ is from the horizontal line $\bar{y}$ (mean).

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SSE is the error sum of squares (unexplained variation), which quantifies how much data points y_i vary around the estimated regression line $\hat{y}_i$. We start by finding the regression line to visualize the relation between the dependent and independent variables. This allows us to calculate the predicted values, then subtract the actual values and to square the results. Finally, we sum all the squared errors and obtain the unexplained variation (SSE).

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SSTO is the total sum of squares (total variation), which quantifies how much the data points y_i vary around the mean $\bar{y}$. For SSTO, we subtract the average actual value from each of the actual values, square the results and sum them.

$$\text{SSTO} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Coming back to the different types of difference, we can consider the horizontal difference between the line of the scatter diagram and observations (Figure 6), then we obtain the reverse (or inverse) regression method.

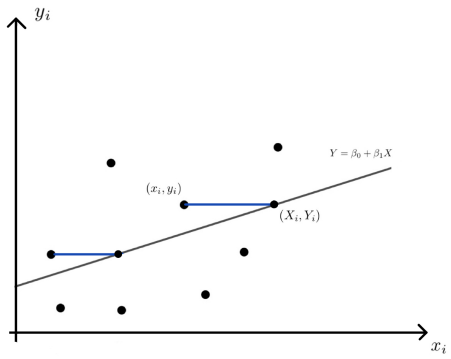

Figure 6: horizontal difference between the line in the scatter diagram and observations

Finally, we have the perpendicular distances between the observations and the line in the scatter diagram (Figure 7), which gives us the orthogonal regression or major axis regression method.

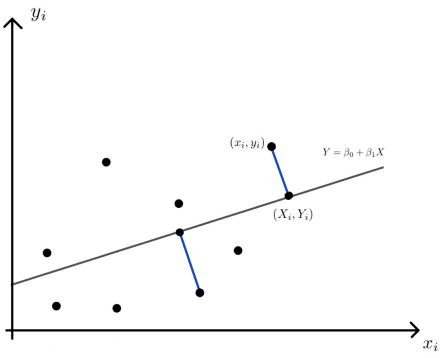

Figure 7: perpendicular distance between observations and the line in the scatter diagram

Furthermore, instead of minimizing the distance between the observations and the line of the scatter diagram, we can use the reduced major axis regression method (Figure 8). By definition, this method minimizes the sum of the areas of rectangles defined between the observed data points and the nearest point on the line in the scatter diagram, to estimate the regression coefficients.

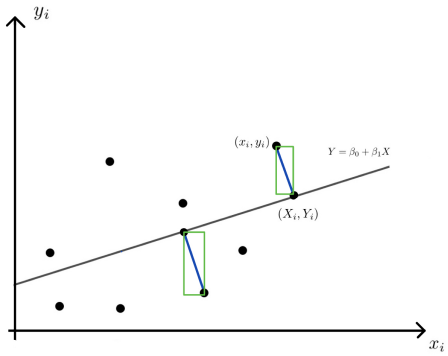

Figure 8: reduced major axis regression method

Maximum likelihood estimation
We will have a closer look at the maximum likelihood estimation in the section 1.3 Log likelihood.

1.2 Poisson regression

Poisson regression assumes a Poisson distribution, which is a discrete probability distribution with parameter λ. This parameter is the mean number of events. For the Poisson distribution the variance is equal to the mean. Moreover, the Poisson distribution is used to predict the number of events occurring within a given interval of time or space. The probability that k events occur in the same interval is: $P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$.
Poisson regression is usually used to model count data, which are discrete data with only nonnegative values, because the coefficients are exponential. As we are using the Poisson distribution, we usually collect data that happen within a certain interval of time or space. Moreover, using the formula for the linear regression, seen above, we obtain the following expression:

$$E(Y_i) = \beta_0 + \beta_1 X_i,$$

which gives us the idea to apply to count data. However, if we model the mean of the distribution as a linear function in X, then we risk to get negative counts and the variability will not be a function in X. Hence, to avoid the normal regression errors, we often use a ln or log transformation because the Poisson regression contains no error term as the mean equals the variance for the Poisson distribution. The ln transformation describes the relationship between the dependent and independent variable. We have the following formula:

$$\ln(E(Y_i)) = \ln(\lambda_i) = \beta_0 + \beta_1 X_i,$$

where the observed counts Y_i follow the Poisson distribution with parameter λ_i for $i \in \{1, \ldots, n\}$. The Poisson parameter is given as a function of the independent variables.
Let us have a look at the technical conditions of the Poisson regression, which allow us to conclude that the Poisson regression is an effective model to use. Firstly, we see that the graph is a line because the mean is a linear function of X: $ln(\lambda_i) = \beta_0 + \beta_1 X_i$. Moreover, the observations are often described by a simple random sample, hence the observations are independent. In addition, the response variable is a count variable and we have no error term.

Poisson regression can be considered as a generalized linear model (GLM), which generalizes linear regression by relating the linear model to the response variable via a link function. In our case, the link function for the Poisson regression is the log function. Consider Y to be the response variable following the Poisson distribution. Let $x \in \mathbb{R}^n$ be a vector of independent variables, then the regression model is of the following form:

$$\log(E(Y|x)) = \alpha + \beta x,$$

where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^n$.

Finally, we can compare the Poisson regression to the linear regression. Let us write down the expression for the Poisson regression : $\ln(E(Y_i)) = \beta_0 + \beta_1 X_i$, where $Y_i \sim P(\exp^{\beta_0+\beta_1 X_i})$ and the normal regression with log transformation: $E(\ln(Y_i)) = \beta_0 + \beta_1 X_i$, where $\ln(Y_i) \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$. One difference between the Poisson and normal regression is that the expected value or average of the logs (normal regression) does not equal the log of the averages (Poisson regression). Another difference is that the variability of the two models is calculated differently, because the likelihood functions are different and hence we obtain different estimates for the parameters.

1.3 Log likelihood

To understand the Log likelihood, we first need to explain the term likelihood. Likelihood is used to describe the process of determining the best data distribution given a specific situation in the data. Therefore, we often use the likelihood ratio test, which compares two models and the improvement with respect to likelihood value. If we add more variables to a linear model, then the likelihood value will improve. We have the following definition for likelihood: Given a random sample $Y_1, \dots, Y_n$ from a discrete distribution D with an unknown parameter θ, we define the likelihood function by

$$L(\theta; y_1, \dots, y_n) = p^{\theta}_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = p^{\theta}_{(y_1)} p^{\theta}_{(y_2)} \cdots p^{\theta}_{(y_n)},$$

where p^{θ} is the Probability Mass Function (PMF) of each Y_i. If $Y_1, \dots, Y_n$ come from a continuous distribution, we set

$$L(\theta; y_1, \dots, y_n) = f^{\theta}_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = f^{\theta}_{(y_1)} f^{\theta}_{(y_2)} \cdots f^{\theta}_{(y_n)},$$

where f^{θ} is the Probability Density Function (PDF) of each Y_i.
Recall the definition of the Probability Mass Function: Let X be a discrete random variable with range $R_X = \{x_1, x_2, \dots\}$ finite or countably infinite. The function $P_X(x_k) = P(X = x_k)$, for $k \in \{1, 2, 3, \dots\}$ is called the probability mass function of X.
Moreover, we have the following definition for the Probability Density Function: Consider a continuous random variable X with an absolutely continuous cumulative distribution function (CDF) $F_X(x) = P(X \leq x), \forall x \in \mathbb{R}$. The function $f_X(x)$ defined by $f_X(x) = \frac{dF_X(x)}{dx} = F_X^{'}(x)$, if $F_X(x)$ is differentiable at x, is called the Probability Density Function of X.
Moreover, we have the Poisson Log-Linear model, also called Poisson regression model. By definition, it is a model for n responses $Y_1, ..., Y_n$ that take integer count values. Each Y_i is an independent Poisson random variable with parameter λ_i and $\log \lambda_i$ is a linear combination of the covariates corresponding to the i^{th} observation. We treat the covariates as fixed constants and parameters are the regression coefficients β. This model is often used in neuroscience to detect spikes. Hence, we consider n time windows of length Δ, Y_i are the number of spikes in the i^{th} time window.

In addition, $Y_1 \ldots, Y_n$ are independent random variables with $Y_i \sim P(\lambda_i \Delta)$. The parameter λ_i controls the spiking rate in the i^{th} time window and may be influenced by a stimuli present in this i^{th} window of time. However, to encode this stimuli, applied in the i^{th} time window by a set of p covariates $x_{i1}, \ldots, x_{ip}$, a model for the Poisson rate, with parameter λ is given by:

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Suppose a data set with two variables X and Y with the corresponding regression equation between the two. The likelihood tells us how likely it is that we will get a similar dataset. In addition, we can measure the likelihood by taking the log of the likelihood value and we obtain the Log likelihood. The higher the value of the Log likelihood, the better is the model.
Let Y follow the Poisson distribution with parameter λ, which is determined by a set of p predictors. Hence, we get the following expression for λ:

$$\lambda = \exp\{\boldsymbol{\beta}^T \boldsymbol{X_i}\}$$

We consider $\boldsymbol{X_i}$ and $\boldsymbol{\beta}$ to be two vectors: $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ and $\boldsymbol{X_i} = \begin{pmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix}$.
For the i^{th} observation, the Poisson regression model is given by:

$$P(Y_i = y_i | \boldsymbol{X_i}, \boldsymbol{\beta}) = \frac{e^{-\exp\{\boldsymbol{\beta}^T \boldsymbol{X_i}\}} \exp\{\boldsymbol{\beta}^T \boldsymbol{X_i}\}^{y_i}}{y_i!}$$

Moreover, for a sample of size n, we get the likelihood for a Poisson regression:

$$L(\beta; y, (X_1, ..., X_n)) = \prod_{i=1}^{n} \frac{e^{-\exp\{\boldsymbol{\beta}^T \boldsymbol{X_i}\}} \exp\{\boldsymbol{\beta}^T \boldsymbol{X_i}\}^{y_i}}{y_i!},$$

and we get the Log likelihood:

$$l(\beta) = \sum_{i=1}^{n} y_i \boldsymbol{\beta}^T \boldsymbol{X_i} - \sum_{i=1}^{n} \exp\{\boldsymbol{\beta}^T \boldsymbol{X_i}\} - \sum_{i=1}^{n} \log(y_i!)$$

Let us consider the situation where $\boldsymbol{\beta}$ and X_i are of the following form:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \text{ and } X_i = \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

We obtain the formula of the linear regression line: $\boldsymbol{\beta}^T X_i = \beta_0 + \beta_1 x_i$.
To determine the parameters β_0, β_1 and σ^2, we can now use the method of maximum likelihood estimation.

For this estimation, we suppose that ϵ_i for $i \in \{1, \ldots, n\}$ are independent and identically distributed and they follow the normal distribution $N(0, \sigma^2)$. We have again the equation of the linear regression model:

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i, \text{with } i \in \{1, \ldots, n\}$$

and the observations y_i for $i \in \{1, \ldots, n\}$ follow $N(\beta_0 + \beta_1 x_i, \sigma^2)$.
To estimate the parameters β_0, β_1 and σ^2 we can maximize the likelihood function $L(x_i, y_i; \beta_0, \beta_1, \sigma^2)$ or the log likelihood function $\ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)$:

$$L(x_i, y_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

and

$$\ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2) = -\left(\frac{n}{2}\right) \ln 2\pi - \left(\frac{n}{2}\right) \ln \sigma^2 - \left(\frac{1}{2\sigma^2}\right) \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Let us solve the equation to find the parameters β_0, β_1 and σ^2:

$$\begin{aligned}
& \frac{\ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = 0 \\
\Leftrightarrow \quad & \left(\frac{1}{\sigma^2}\right) \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\
\Leftrightarrow \quad & n\bar{y} - n\beta_0 - n\beta_1 \bar{x} = 0 \\
\Leftrightarrow \quad & \beta_0 = \bar{y} - \beta_1 \bar{x}
\end{aligned}$$

and:

$$\begin{aligned}
& \frac{\ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = 0 \\
\Leftrightarrow \quad & \left(\frac{1}{\sigma^2}\right) \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \beta_0 x_i - \sum_{i=1}^{n} \beta_1 x_i^2 = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} y_i x_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \bar{y} x_i + \beta_1 \sum_{i=1}^{n} \bar{x} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} x_i (y_i - \bar{y}) + \beta_1 \sum_{i=1}^{n} x_i (\bar{x} - x_i) = 0 \\
\Leftrightarrow \quad & \beta_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} x_i (y_i - \bar{y}) \\
\Leftrightarrow \quad & \beta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}
\end{aligned}$$

and:

$$\begin{aligned}
& \frac{\ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = 0 \\
\Leftrightarrow \quad & -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0 \\
\Leftrightarrow \quad & \frac{n}{2\sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
\Leftrightarrow \quad & n\sigma^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
\Leftrightarrow \quad & \sigma^2 = \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{n}
\end{aligned}$$

Hence, we have the following results: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n}$

2 Overtaking in Formula 1

2.1 Formula 1

2.1.1 Introduction in F1

Nowadays Formula 1, the world's most prestigious motor racing competition, gets more and more popular, capturing the attention of millions of spectators around the world. The reason why Formula 1 is called Formula 1 is two-fold. We use the term "Formula" to describe a set of rules such as the car design, engine size, component usage and so on. All competitors have to follow those rules. The number 1 denotes that it is the premier formula.
Formula 1 exists since 1950, celebrating its 75 years in 2025 with a unique opening show in London at the beginning of the season. To get a seat in Formula 1, most drivers first participate in carting competitions, Formula 3 and Formula 2, where the cars are smaller and slower than Formula 1 cars. Moreover, to make motorsport more diverse, inclusive and accessible, F1 Academy Racing was created in 2023. F1 Academy allows female drivers to race against each other and explore their own motorsport journeys.

Formula 1 is considered to be the highest class of international racing for single-seater racing cars. Every driver wants to be the best, pushing themselves and their racing car to the very limit. In order to do that, they need to be in top shape, battling extreme g-forces, making daring decisions in the blink of an eye while driving at a speed up to 370km/h. In addition, to the individual performance of the driver, Formula 1 is also a team sport. There are different engineers designing the car and also the pit-crew being able to change all 4 tires in 2-3 seconds. As pit-stop we define the situation where a car stops in the pit stalls during a race for usually a change of tires or even a quick maintenance or mechanical repairs.
Drivers compete to win the F1 Driver's Championship and they fight together with their team for the F1 Constructor's Championship.

In total, Formula 1 consists, as of 2025, of 10 different teams with each team 2 drivers. Each driver has an assigned number, which was introduced in 1973, and the number 1 is assigned to the reigning world champion.
Some teams such as Ferrari (Figure 9) and McLaren competed since the beginning of Formula 1, other teams such as Haas entered the competition for the 2016 season.

The F1 calender of 2025 consists of 24 race weekends (distance of every race is almost always equal to the fewest number of laps that exceed 305 km in total), including six F1 Sprint races (distance is equal to the fewest number of laps needed to exceed a distance of 260km) (Figure 10). Each race is called Grand Prix and they are held in different locations around the world. Some locations were always part of the F1 calender such as Monaco or Silverstone, while others are new additions such as Shanghai or Miami.

Figure 9: Evolution of the F1 car over the years

Both race weekend and sprint weekend start on Thursday with a mediaday, where the drivers give interviews answering questions about the past and the upcoming Grand Prix. During the race weekend, we have two free practices (each one lasting for an hour) on Friday, a third free practice and qualifying on Saturday and the race is held on Sunday. For sprint weekend, we have only one free practice on Friday together with the sprint qualifying. On Saturday, we have the sprint race and qualifying and on Sunday the race.

Place	Points
1st	25
2nd	18
3rd	15
4th	12
5th	10
6th	8
7th	6
8th	4
9th	2
10th	1

Place	Points
1st	8
2nd	7
3rd	6
4th	5
5th	4
6th	3
7th	2
8th	1

Figure 10: The first picture shows the points distributed during a race and the second picture shows the points distributed during a sprint race.

2.1.2 Rules and Regulations

The rules and regulations in Formula 1 are set by the FIA ("Fédération internationale de l'Automobile"). Changes in rules and regulations have as primary reasons to do with the safety of the drivers, such as better facilities and equipment available in case of an accident. A lot of regulations are changed in order to slow down the cars to a level where a Grand Prix can be driven relatively safely. For example if cornering speed is too high, an accident in a corner could result in the death of a driver.

Let us have a closer look at the regulation changes over the years:

In the first ten years of Formula 1, safety was almost a non-issue. Most tracks didn't have safety features, death was considered as an acceptable risk for winning races.
Moreover, the technological progress was extremely slow compared to nowadays. There were no weight limits for the engine. A progress happened when alcohol-based racing fuels were banned and replaced with commercial petrol.
The 1960s were the beginning of drivers lobbying for safer racing. Therefore the FIA executed safety inspections of the track, which were previously done by local authorities. Moreover, they introduced protective helmets, overalls and a flag signalling code was established.
From 1966 to 1969, the cockpit was redesigned to allow a quick evacuation and FIA introduced some recommendations such as seat harness, fire-resistant clothing and shatterproof visors. Moreover, straw bales were banned from being used as safety barriers.
In the 1970s, Formula 1 cars became faster and the safety measures were not adapted, hence causing a lot of deaths.
However, in 1978, because of the commitment of a lot of drivers, circuits were improved with safety features, and new circuits with better safety features were constructed. Those safety features are for example double guard rails in place, three meter grass verges, sand traps, barriers between pit lane and track. The track width and surface also play an important role for the safety. Moreover, spectators need to keep a distance of minimum three meters behind the guard railings.
In addition, new roles were introduced such as race supervisors, marshals and signalers. Drivers need to be able to evacuate from their cockpit in less than five seconds. Two mirrors were introduced to the cars and a better fire extinguisher on board was required.
In the early 1980s, ground effects which created a huge amount of downforce causing a lot of accidents, were banned. The aluminium material for chassis construction was replaced by the much stronger carbon fibre, reducing the number of fatalities.
Unfortunately, in 1994 FIA banned virtually all the performance enhancing electronic technology, such as ABS or traction control, which was needed by the teams and caused the cars to become nervy and edgy to drive. Due to more horsepower and less car stability, a lot of fatal accidents happened, such as the 3 accidents during the San Marino Grand Prix in 1994, causing a severely injured

driver and two deaths including the death of Ayrton Senna. From there on, safety became Formula 1's number one concern.
Since 2010, refuelling the car during the race is no longer allowed. The introduction of HANS i.e. head and neck support in 2003 and halo (driver crash-protection system consisting of a curved bar placed above the driver's head to protect it) in 2018 were the last 2 major safety improvements.
Besides the safety issues, FIA faced another problem in 2008, namely dramatic cost savings due to the world economic crisis. A lot of teams could no longer afford to invest a huge amount of money and some teams such as Honda, Toyota or BMW withdrew from the sport. FIA was able to control the cost escalations thanks to implemented budget caps. The budget cap also called cost cap is a financial regulation to limit the amount of money teams can spend annually on designing and operating their cars.

In 2011, another big change for the aerodynamic of the cars was the introduction of the driver adjustable rear wing. This rear wing also known as DRS (drag reduction system) helps the drivers to overtake, by gaining speed. It can only be used in the activation zone and the driver must be within 1 second of the next car in the detection zone.
For 2026, there are new regulations planned to make the sport more agile, competitive, sustainable and safer. These regulations change the aerodynamic of the car, increase for example the battery power and also the use of 100% sustainable fuels.

2.2 Theoretic framework

In the following analysis, we want to look into the detail of overtaking in Formula 1, to find out which variables are able to increase the number of overtakes. Note that overtaking varies a lot on a race-to-race basis. Therefore, we need overtaking data.

For our analysis we use the article "Overtaking in Formula 1 during the Pirelli era: A driver-level analysis" (de Groote, 2021) as a reference.

According to the article we define the notion of an overtake: an overtake is a change of position on track. However, an on track position change such as position change on the first lap, position change due to drivers yielding (i.e. surrendering the place without fighting) or position change because of car problems are not considered to be an overtake. The number of overtakes is derived from lap charts in combination with actual race footage. Actual race footage is needed because passes may be obscured in lap charts by pit-stops or retirements. Note that for our analysis, we don't consider wet races i.e. a race where at least one driver uses wet-weather tires at some point in the race. We only analyze dry races to avoid the influence of bad weather.
Moreover we also exclude drivers who retire from a race before the end of the first lap.

Based on the article, the data set contains information about the number of times any driver passed or was passed by another driver. We need the timing information such as race order at the end of the first lap and fastest lap-time. In addition, we also take into account race strategies for example the number of pit-stops and timing of the first pit-stop.

2.3 Factors influencing the potential of overtaking

We have two factors which we can not observe directly: the overtake friendliness of a track and the amount of wake turbulence produced by the cars.
Wake turbulence is a chaotic airflow, generated by the F1 car. This airflow has a significant impact on the performance of a following car, i.e. it experiences a loss of downforce or grip and an increase of drag. Moreover, wake tubulence also influences the cooling system of the enginge and brakes, leading to an overheating of the car behind. Therefore, the driver has more problems to maintain control of his car and needs to keep more distance to the car in front.
Our analysis focuses mainly on observable factors, that may influence overtaking on the driver-level, to indirectly estimate those unobservable factors. Let us have a look at the different categories of observable factors:

1) Effectiveness of DRS

2) Reliability of the car and field size:
Note that larger field size and better reliability of the car (i.e. lower breakdown rates) increase the number of cars on track, and hence increase the likelihood of overtaking.

3) Pit-stops:
In relation with pit-stops, let us first define three important notions to describe the different positions of a driver on track. On the one hand, we define as backmarker a car falling behind in a race, often they are lapped by the faster cars. On the other hand, front-runners are the race leaders, fighting to win the race. Moreover, we define the midfield consisting of the teams which are fighting for lower points.

Pit-stops mix up the order of the cars and have a strong effect on overtaking. We have the pit-lane overtaking and change in track position because of different speed of cars, which creates new on-track battles and increases overtaking. Moreover, pit-stops allow the teams to do an undercut or overcut to overtake other cars. On the one hand, we define the undercut as a strategy where a driver pits earlier than his rival who is in front, to gain an advantage by using fresher tires. On the other hand, the overcut is a strategy to overtake where a driver stays out on track longer than his rival, who pits earlier. This allows the driver to be faster because of the cleaner air and lighter fuel load, and he is able to gain a track position.

Another important factor is the timing of the first pit-stop. If the first round of pit-stops starts early (especially if front-runners pit earlier), drivers lose more positions because the field is still close and it results in more overtaking. Unscheduled pit-stops, for example to repair accident damage, induce overtaking especially if fast cars are involved.

4) Cars being out of position (i.e. mixing) after start and pit-stop:
If a driver starts in a worse position than what he would be based on his expected race pace, i.e. stronger race pace compared to qualifying pace, the driver is more likely to overtake another car and less likely to be overtaken. In addition, we observe the opposite if a driver qualifies in a better place than expected, his qualifying pace is stronger than race-pace. Moreover, we also have the possibility to be out of position at the beginning of the race because of a bad start or grid penalty.

5) Safety car:
On the one hand, the safety car is able to eradicate the gaps between the cars, which increases the amount of on-track battles and overtakes at the restart. On the other hand, the virtual safety car preserves the gaps between the cars and hence has less impact on racing. Therefore, we ignore virtual safety cars in our analysis. In addition, we don't consider safety cars after the start of the race or at the very end of the race.
We use a dummy variable to determine whether or not a driver is likely to stop again and a safety-car dummy, which is not directly observed. Note that the safety-car dummy is a variable which gives us the information if there was a safety-car period (no virtual-safety-car period) in between two green-flag periods, hence it does not count the number of safety-car periods. Moreover, we consider a cut-off value of 75% of the race distance, which will be the amount where we can expect the last pit-stops of the race.

2.4 Fixed effects models

Before looking into the detail of the formulas used in the article (de Groote, 2021), we start with an explanation of the different categories of fixed effects used to analyze overtaking in Formula 1.
First let us define the term fixed effects in general before explaining the different fixed effects used in our analysis in F1. We need to distinguish between the fixed effects, random effects and mixed effects.

The fixed effects model is a method to control variables. We have two different types of fixed effects: entity fixed effects, where the variables do not change over time but across entities, and we have time fixed effects, where the variables change over time but not across entities. Entities can for example be individuals, families, companies or countries.

In the following, we mainly use entity fixed effects for our analysis.
Therefore, using fixed effects we are able to observe the individual characteristics of entities and to control their impact on the desired outcome.
Moreover, fixed effects estimation allows us to consider data on individuals with multiple observations and to estimate effects only for those variables changing during the observation period.
To analyze data with fixed effects, we use a regression model with dependent and independent variables. The independent variables are binary variables also called time-invariant dummy variables, i.e. a variable that takes a binary value (0 or 1) to indicate the absence or presence of an effect (Equation 1).

Let us compare linear regression and fixed effects model. On the one hand, we use linear regression to analyze the relation between dependent and independent variables. Moreover, we usually use cross-sectional data in linear regression, where we observe different groups such as individuals or countries at a specific period of time i.e. one point in time.
On the other hand, fixed effects are often considered as a specific type of linear regression, used to control variables. In addition, we mainly use panel data in the fixed effects model, which means that we consider data from different groups over many points in time.

The fixed effects regression model (i.e. entity fixed effects regression model) is:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}, \tag{1}$$

where β_0, β_1 and β_2 are parameters, i is the notation for entity and t the notation for time. We denote the error term with u_{it} and it has conditional mean zero, i.e. $E(u_{it}|X_{i1}, X_{i2}, ..., X_{iT}) = 0$.
Y_{it} is the outcome variable, X_{it} is the variable we are interested in measuring the causal effect on Y. Moreover, Z_i are the unobserved variables, which only depend on the entities and are constant over time.

We can rewrite this formula in a different way. Let us combine β_0 and $\beta_2 Z_i$ as α_i, we can consider α_i as the fixed effect for the entity i and we obtain the following formula:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}. \tag{2}$$

To estimate the parameter β_1 we use equation (2). We compute the average of this equation and obtain:

$$\frac{1}{T}\sum_{t=1}^{T} Y_{it} = \beta_1 \frac{1}{T}\sum_{t=1}^{T} X_{it} + \alpha_i + \frac{1}{T}\sum_{t=1}^{T} u_{it}.$$

Note that, since α_i does not depend on t, we have $\frac{1}{T}\sum_{t=1}^{T} \alpha_i = \alpha_i$. We can rewrite the average equation as follows:

$$\bar{Y}_{it} = \beta_1 \bar{X}_{it} + \alpha_i + \bar{u}_{it}. \tag{3}$$

Now we compute (2)-(3):

$$\begin{aligned} Y_{it} - \bar{Y}_{it} &= \beta_1 X_{it} - \beta_1 \bar{X}_{it} + \alpha_i - \alpha_i + u_{it} - \bar{u}_{it} \\ \Rightarrow Y_{it} - \bar{Y}_{it} &= \beta_1 (X_{it} - \bar{X}_{it}) + (u_{it} - \bar{u}_{it}) \end{aligned}$$

Let $\tilde{Y}_{it} = Y_{it} - \bar{Y}_{it}$, $\tilde{X}_{it} = X_{it} - \bar{X}_{it}$ and $\tilde{u}_{it} = u_{it} - \bar{u}_{it}$. We use the least square estimation to find the expression for β_1. We need to minimize the sum of squared residuals and to compute the derivative with respect to β_1 and set it to 0:

$$\frac{\partial}{\partial \beta_1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{Y}_{it} - \beta_1 \tilde{X}_{it})^2 \right) = 0$$

$$\begin{aligned} \Leftrightarrow \quad & \frac{\partial}{\partial \beta_1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{Y}_{it}^2 - 2\tilde{Y}_{it}\beta_1 \tilde{X}_{it} + \beta_1^2 \tilde{X}_{it}^2) \right) = 0 \\ \Leftrightarrow \quad & -2 \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{Y}_{it} \tilde{X}_{it} - \beta_1 \tilde{X}_{it}^2 \right) = 0 \\ \Leftrightarrow \quad & \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{Y}_{it} \tilde{X}_{it} = \sum_{i=1}^{N} \sum_{t=1}^{T} \beta_1 \tilde{X}_{it}^2 \\ \Leftrightarrow \quad & \beta_1 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{Y}_{it} \tilde{X}_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{X}_{it}^2} \end{aligned}$$

Hence, we obtain that $\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{Y}_{it} \tilde{X}_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{X}_{it}^2}$.
Moreover, we can estimate β_0 which is the average of individual effects by using $\hat{\beta}_0 = \bar{Y}_{it} - \hat{\beta}_1 \bar{X}_{it}$.

We can also use the entity and time fixed effects regression model. We have the following formula:

$$Y_{it} = \alpha_i + \beta_1 X_{it} + \delta_t + u_{it}$$

where i is the notation for entity and t the notation for time, β_1 is a parameter, δ_t is the coefficient for the time regressors and α_i is the fixed effect for the entity i. We denote the error term with u_{it}. Y_{it} is the outcome variable, X_{it} is the variable we are interested in measuring the causal effect on Y.

The random effects model is used to observe variability and differences between different entities in a larger group. Note that the term "random" does not imply randomness, but rather that we consider that each level is drawn from a random variable which originates from an underlying distribution.
For the random effects model the factors are not fixed but we group them as random variables and obtain different levels of factors. The estimation of random effects allows us to get information about specific levels, but also population, absent or unobserved levels. The levels of random effects are considered as representative samples from a larger collection of levels.
We have the following formula:

$$Y_{in} = \mu + \alpha_i + \epsilon_{in},$$

where n is the number of observations and i stands for the i^{th} group. We have that Y_{in} is the response variable and μ the overall mean. Moreover, $\alpha_i \sim N(0, \sigma_\alpha^2)$ are the random effects for groups and $\epsilon_{in} \sim N(0, \sigma_\epsilon^2)$ is the residual error.
Let us have a look at the advantages of fixed and random effects. On the one hand, fixed effects are ideal to control time-invariant factors. They are also used to reduce the confusion of the relation between independent and dependent variables. By using fixed effects in our analysis, we remove the influence of unobserved factors and therefore we get a more accurate estimation of the desired effects compared to the random effects model.
Moreover, fixed effects models allow us to analyze entity-specific effects and the unique impact of different variables on the desired outcome.
However, fixed effects models mainly focus on the within-entity variations and therefore it is difficult to generalize the estimations to populations outside the observed entities.
In addition, fixed effects regression models reduce possible bias (no accurate representation of the population) compared to the random effects model because of unmeasured, unchanging variables that may be correlated with the variables of interest.

On the other hand, we have some advantages of random effects. For example random effects allow us to capture unobserved heterogeneity and to explain variations in the outcome, which can not be explained by observable variables. Moreover, using random effects, our estimations of parameters get more precise compared to the estimations with fixed effects. This precision is due to the fact that random effects are correlated with independent variables.
Another advantage of random effects models is that these models are able to capture a larger range of variation in the population because the effects vary across different groups.

However, one main disadvantage of random effects is that we are not able to estimate and interpret entity-specific effects (i.e. variations and nuances within individual characteristics) because random effects mainly capture unobserved heterogeneity.

The mixed effects models (MEMs) are defined as a class of models which are built on linear or generalized linear models. Moreover, MEMs rely on different factors such as dependence, heterogeneity and non-linearity of variables.

Let us first look at the linear mixed effects models, which are extensions of linear regression models, where the data is collected and summarized in groups. Moreover, MEMs mix fixed effects and random effects. On the one hand, the terms of fixed effects are usually the linear regression part and on the other hand, the random effects correspond to the individual level randomly drawn from a population. Mixed effects models describe a relation between a response variable and independent variables (Equation 4).
In addition, the coefficients vary with respect to one or more groups.
Using the mixed effects models, we obtain a covariance structure related to the fact that we associate the random effects to observations with the same level of a grouping variable.
We have the following model for linear mixed effects model:

$$Y = X\beta + Zb + \epsilon, \tag{4}$$

where Y is the n-by-1 response vector and n the number of observations. Moreover, β is a p-by-1 fixed effect vector and X an n-by-p fixed effect matrix. We have that b is a q-by-1 random effects vector and Z a n-by-q random effects matrix. The error term ϵ is the n-by-1 observation error vector.
Note that the random effects vector b and the error term ϵ, which are independent, have the following distributions with σ^2 the error variance:

- $b \sim N(0, \sigma^2 D(\theta))$, where D is a symmetric and positive semidefinite matrix and parameterized by a variance component vector θ.
- $\epsilon \sim N(0, \sigma^2 I)$, where I is an n-by-n identity matrix

We can consider different mixed effects models. Depending on the context these models are called multilevel models (parameters vary at more than one level) or hierarchical models (relation between lower level variables and higher, more general level of factors). Mixed effects models include factors which can be multilevel, hierarchical or crossed (i.e. every level of one factor occurs in every level of the other factor).

For the analysis in F1, the article (de Groote, 2021) uses four different categories of fixed effects: track fixed effects, season fixed effects, race fixed effects and driver race-specific fixed effects.
We use fixed effects in our analysis because this method allows us to observe the individual characteristics of the track, season, race and driver and to observe their specific impact on overtaking.
Let us have a closer look at the different fixed effects used:

2.4.1 Track fixed effects

Track fixed effects describe the unique characteristics of each circuit that can influence the outcome of a race. We consider for example the layout of a track, the elevation, the surface type or the length.
The layout of a track plays an important role in the analysis of overtaking such as overtaking-friendly tracks, which consist of long straights and long braking zones (Figure 11).

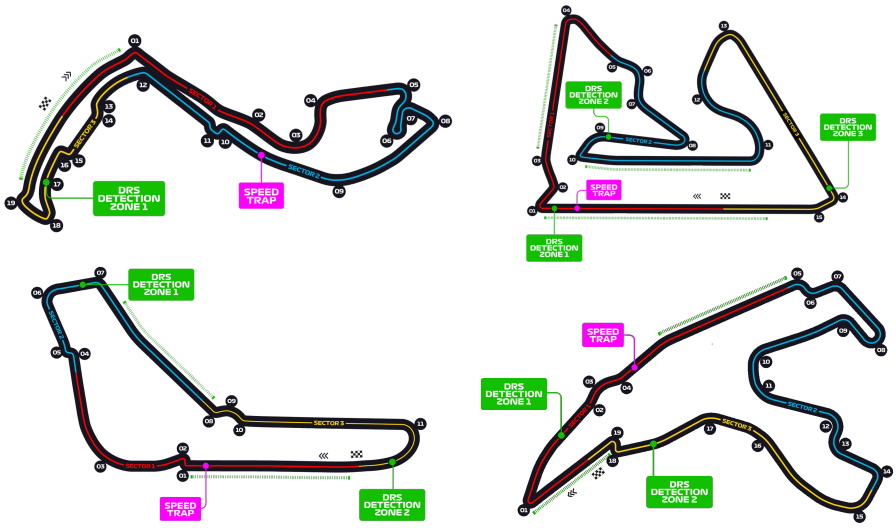

Figure 11: The first layout shows the circuit of Monaco (reference track), which is a twisty street circuit and the worst track for overtaking because there is not enough space to overtake and it is a slowest circuit. The second layout is the track in Bahrain, which consists of long straights and long breaking zones to promote overtaking. The two last circuits are Monza and Spa, which are ideal to overtake because there is enough space on track, drivers have DRS coming out of a corner and long breaking zones.
Note that each track is divided into three sectors (marked in the pictures with different colors) with approximately the same length. This division is useful for the teams to easier compare the performance of the drivers.

Before going into the detail of how the surface of the track influences the tires, let us first have a look at the different tires used in Formula 1. Since 2011, Pirelli exclusively supplies F1 teams with tires after some previous involvements in the 1950s, '80s and '90s.
The 18-inch tires (marked on the outside in different colors) range six slick compounds (used in dry weather conditions) from hardest to softest, i.e. C1, C2, C3, C4, C5, C6.

Soft tires, marked in red, are the fastest tires but are likely to wear out quickly. Medium tires, marked in yellow, are a compromise between the soft and hard compound. They last longer than the softs but are slower. Finally, the hard tires, marked in white, last the longest but provide the least grip.
Moreover, Pirelli also provides intermediates (marked in green) and full wets (rainy weather, marked in blue) (Figure 12).
At every race weekend, Pirelli picks three compounds to be used, taking into account track characteristics and weather conditions. Moreover, each team has 13 sets of dry weather tires available for a race weekend and during a race every driver has to use at least two different slick compounds.

Figure 12: Different tires marked in their respective color

Note that the roughness, composition and even color of the track have an influence on the grip levels of the cars. Most surfaces of racetracks are made of asphalt, applied on top of layers of crushed rock and stone. The surface has a significant impact on durability and tire grip, especially resurfacing a track reduces the grip.
Another important aspect of track fixed effects is the track evolution with each lap, which refers to the gradual increase in grip. This is due to the fact that the tires deposit rubber on the track, which finally improves lap times.
Track temperature also influences the tire grip.
On the one hand, higher track temperature often brings tires into their working range which increases the grip. However, extremely high track temperatures cause overheating of tires. This overheating reduces the grip and hence induces graining and blistering.
On the other hand, low track temperatures make it difficult for the rubber to deform and therefore generate friction.

To conclude, we can say that we consider track fixed effects to give specific information about the characteristics of each track. The most important information is the overtaking-friendlyness of a track i.e. the layout of a track.

2.4.2 Season fixed effects

We consider season fixed effects because they describe different factors that are constant during a season but may change across seasons. These changes are for example different regulations, other tire compounds, engine performance or aerodynamic rules.

2.4.3 Race and driver race-specific fixed effects

In the analysis of overtaking from the article (de Groote, 2021), race fixed effects and driver race-specific fixed effects are composed of different variables. These variables are usually derived from observed variables and used as auxiliary variables. Note that we use three indices to describe the variables: i determines the driver, j the track and t the season. Moreover, the race fixed effects are uppercase letters and the driver race-specific effects are lowercase letters.
Let us have a look at the auxiliary variables:

- Race distance in number of laps; N_{jt}
- Number of cars running at the end of the first lap; C_{jt}
- Unscheduled stop in first 5 laps; S^a_{ijt}
- Unscheduled stop after lap 5; S^z_{ijt}

Note that the letters a and z for the variables S^a_{ijt} and S^z_{ijt} are chosen randomly.

- Position at the end of the first lap; p_{ijt}
- Ranking based on fastest lap of the weekend; r_{ijt}
- Number of laps completed; n_{ijt}
- Lap at which first pit-stop is made; l_{ijt}
- Percentage of race distance at which the first pit-stop is made; $k_{ijt} = \frac{l_{ijt}}{N_{jt}}$
- Unknown strategy (u stands for unknown); $q^u_{ijt} = \begin{cases} 1 \text{ if } \frac{n_{ijt}}{N_{jt}} < \frac{3}{4} \\ 0 \text{ else} \end{cases}$

Note that standardization (namely z-score standardization) used for the variables is a method of transforming variables so that they have a mean of zero and standard deviation of one. We adjust the data values which allows us to compare different variables.
For the z-score standardization we have the following formula:

$$z = \frac{(x - \mu)}{\sigma},$$

where x is the original data point, μ is the mean of the data and σ is the standard deviation of the data.

To do the standardization we subtract the mean of each variable from each data point and divide it by the standard deviation.
To ensure standardization and consistency in data representation, we use equal interval in our measurements. Therefore, we can also compare standardization across different measurement units.
Since standardization eliminates variations caused by differences in measurements, it allows us to compare, analyze and interpret numerical data by transforming the data to a common scale or format. Moreover, using standardized data enables us to identify outliers and detect trends. Hence it is commonly used in regression analysis and hypothesis testing.
Another method used for standardization is the $\min - \max$ scaling, used to scale the data to range between 0 and 1. To obtain this rescaling, we subtract the minimum value from data points and divide it by the range $X_{max} - X_{min}$. Hence, we have the following formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}},$$

where X' is the standardized value and X is the original value. Moreover, X_{min} is the minimal value in the dataset and X_{max} is the maximum value in the dataset .

We have three more auxiliary variables:

- Standardized rank [0,1]; $\rho_{ijt} = 1 - \frac{r_{ijt}-1}{C_{jt}-1}$
  The fastest driver of the weekend has a standardized rank ρ_{ijt} of 1 and the slowest driver has a standardized rank of 0.
- Standardized position [-1,1]; $\pi_{ijt} = \frac{2(p_{ijt}-1)}{C_{jt}-1} - 1$
  The standardized position π_{ijt} ranges from -1 to 1 in an ascending order where the leader has the lowest number.
- Standardized position [0,1]; $\psi_{ijt} = 1 - \frac{p_{ijt}-1}{C_{jt}-1}$
  The leader at the end of the first lap has a standardized position ψ_{ijt} of 1 and the driver running last at the end of the first lap has a standardized position of 0.

The standardized rank and position are normalized to a $[0, 1]$ or a $[-1, 1]$ interval by dividing the ranks with the number of cars running at the end of the first lap.
It is helpful to have the variables with a mean of zero because ψ_{ijt} is directly compared to ρ_{ijt} and π_{ijt} interacts with the pit-stop timing.

Race fixed effects

We use race fixed effects to analyze if a car is qualifying pace biased or race pace biased.
On the one hand, a car is qualifying pace biased if it qualifies in a good position

but loses a lot of places during the race because of a slow race pace. A qualifying pace biased car performs well over one single fast lap.
On the other hand, a race paced biased car usually qualifies in a lower position but gains a lot of positions during the race. Hence, this car is more consistent and efficient during the race.
This difference of pace can be due to the different tire degradation of the cars or to a good balance of the car with high fuel.
Let us consider the situation where a driver with a strong qualifying pace car obtains a grid penalty, he is able to overtake a lot of cars during the race if he has a strong race pace car. Some reasons to obtain a grid penalty are for example impeding another driver during qualifying (3-place grid penalty) or multiple power unit changes (10-place grid penalty).

Moreover, another factor of race fixed effects are weather conditions such as the wind. If the wind is blowing from the front, the maximum speed of the car will decrease because the car needs to overcome a greater momentum of air in the opposite direction. In addition, the car consumes more fuel and the engine needs to work harder against the wind resistance.
However, this situation also has advantages: if the car is on the straight line, the air brakes the car and allows the driver to brake later before the corner and with less force which reduces the brake wear. Moreover, DRS is more effective and allows the driver to overtake more easily.
If the wind is blowing from behind, we observe an increase in the maximum speed of the car in less time, hence the car uses less fuel. However, on the straight line, the driver needs to break earlier which increases brake wear. In addition, DRS is less effective because there is not much difference in speed between the car using DRS to overtake and the car right in front.
If the wind is blowing from the side, the car becomes more unstable especially in corners. In addition, side wind can be dangerous if there are sudden wind gusts, because the driver suddenly loses stability and needs to react quickly.

For example the circuit Silverstone in Great Britain, which is located in a place with few natural obstacles, is known for its sudden and strong wind gusts causing the drivers to spin, drive through the gravel or even end up in the wall.

For our model the race fixed effects are composed of 6 different variables. Note that these race fixed effects are usually the sum of auxiliary variables.
1) The number of cars running at the end of the first lap; C_{jt}.
2) Race-average proportion of laps completed; $\mu_{jt} = \frac{\sum_i n_{ijt}}{N_{jt}}$.
3) Race-average distance at which the first pit-stop is made; $\lambda_{jt} = \sum_i \left(\frac{l_{ijt}}{N_{jt}}\right) |q^u_{ijt} = 0$.
4) One or more safety car periods (dummy variable); S_{jt}.
5) Number of unscheduled stops in the first 5 laps; $U^a_{jt} = \sum_i S^a_{ijt}$.
6) Number of unscheduled stops after lap 5; $U^z_{jt} = \sum_i S^z_{ijt}$.

Driver race-specific fixed effects

In our analysis, we also consider driver race fixed effects to take the personality of a driver and his driving style into consideration. For example, Max Verstappen is known for his aggressive driving style or Nico Rosberg usually driving more carefully and therefore he was often overtaken.

Most driver race-specific fixed effects are directly observed, except for the pit-stop and position variables. For our model the driver race-specific fixed effects are composed of 11 different variables:
1) Proportion of laps completed; $v_{ijt} = \frac{n_{ijt}}{N_{jt}}$.
2) Number of scheduled pit-stops; q_{ijt}.
3) Percentage of drivers pitting more often (m stands for more); q^m_{ijt}.
4) Percentage of drivers pitting less often (l stands for less); q^l_{ijt}.
5) Unknown strategy (noted with u); $q^u_{ijt} = \begin{cases} 1 \text{ if } \frac{n_{ijt}}{N_{jt}} < \frac{3}{4} \\ 0 \text{ else} \end{cases}$

6) Pit-stop mixing; $\chi_{ijt} = \pi_{ijt}(k_{ijt} - \lambda_{jt})$.
We obtain positive values if either a back-marker pits late or a front-runner pits early. Both situations cause mixing of the field and hence it induces overtaking. However, if χ_{ijt} is negative (i.e. one of the terms of the product is negative), then we have less mixing and hence less overtaking is expected. In this case the variable χ_{ijt} is split into a positive part χ^h_{ijt} and negative part χ^l_{ijt}.
7) Low pit-stop mixing (noted with l); $\chi^l_{ijt} = \begin{cases} -\chi_{ijt} \text{ if } \chi_{ijt} < 0 \\ 0 \text{ if } \chi_{ijt} \geq 0 \end{cases}$, where χ^l_{ijt} takes the negative values of the pit-stop mixing χ_{ijt} and positive numbers are treated as zero.
8) High pit-stop mixing (noted with h) ; $\chi^h_{ijt} = \begin{cases} 0 \text{ if } \chi_{ijt} < 0 \\ \chi_{ijt} \text{ if } \chi_{ijt} \geq 0 \end{cases}$, where χ^h_{ijt} takes the positive values of the pit-stop mixing χ_{ijt} and negative numbers are treated as zero.

9) Position compared to expected position; $\delta_{ijt} = \psi_{ijt} - \rho_{ijt}$.
Moreover, if δ_{ijt} is positive, it means that a driver is running in a worse position than where he is expected and if δ_{ijt} is negative, the driver is running a better position than where he is expected. In this case, δ_{ijt} is separated into a positive part Δ^w_{ijt} and negative part Δ^b_{ijt}.

10) Better (noted with b) than expected position at the end of the first lap; $\Delta^b_{ijt} = \begin{cases} -\delta_{ijt} \text{ if } \delta_{ijt} < 0 \\ 0 \text{ if } \delta_{ijt} \geq 0 \end{cases}$, where δ_{ijt} is the position compared to the expected position.
11) Worse (noted with w) than expected position at the end of the first lap; $\Delta^w_{ijt} = \begin{cases} 0 \text{ if } \delta_{ijt} < 0 \\ \delta_{ijt} \text{ if } \delta_{ijt} \geq 0 \end{cases}$, where δ_{ijt} is the position compared to the expected position.

2.5 Model

Our analysis on overtaking relies on count data. Therefore, we use the Poisson regression to model how often each driver passes or gets passed in each race. Since the total number of places gained and lost matter for the analysis, they are estimated simultaneously. Moreover, these estimations share some race and driver race-specific fixed effects (see 2.4 Fixed effects models).
We have the following two formulas to estimate the number of places gained and lost by a driver:
1) To compute the expected number of places gained (G_{ijt}) by driver i on track j in season t we use:

$$\log(E(G_{ijt}|x_{ijt})) = R\alpha + R_g\gamma + D\beta + D_g\eta + \tau_j + \phi_t$$

2) To compute the expected number of places lost (L_{ijt}) by driver i on track j in season t we use:

$$\log(E(L_{ijt}|x_{ijt})) = R\alpha + R_l\delta + D\beta + D_l\theta + \tau_j + \phi_t$$

For both formulas, τ_j is the vector for track fixed effects and ϕ_t the vector for season fixed effects. In addition, we consider R to be a vector of race fixed effects and D to be a vector of driver race-specific fixed effects. Moreover, we add R_g as vector of race fixed effects and D_g as vector of driver race-specific fixed effects where g stands for the number of places gained by a driver i in track j in season t. Similarly, R_l is a vector of race fixed effects and D_l is a vector of driver race-specific fixed effects where l stands for the number of places lost by a driver i in track j in season t.
Moreover, α and β are estimated vectors, which are the same in both equations. Note that the vectors γ and η are only estimated for positions gained and the vectors δ and θ are only estimated for positions lost.

To interpret the results obtained during the calculations of the number of places gained and lost by every driver we use the incidence-rate ratio of the baseline model.
Let us have a look at the term of incidence rate. The incidence rate is not only used in Formula 1 but also in other domains. It refers to the rate at which a new event occurs over a specified period of time. Hence it refers to the frequency at which new events occur. This method is often used to describe the frequency of a disease or accidents but also financial phenomena.
Moreover, the incidence rate allows experts to predict future incidents and prepare appropriate plans because the incidence rate only uses new cases instead of previously detected ones. In order to determine the incidence rate of a particular event, experts consider the number of new cases as a proportion of the population at risk.
If two rates of an incidence are computed during the same time period we can determine the incidence-rate ratio, which refers to the ratio of two different rates.

If we obtain an incidence-rate ratio of 1, we can conclude that there is no difference between the exposed and unexposed group. On the one hand, an incidence-rate ratio greater than 1 indicates a positive association, i.e. in the case of a disease the incident rate in the exposed group is higher than in the unexposed group. On the other hand, an incidence-rate ratio less than 1 indicates a negative association, i.e. in the case of a disease the incident rate in the exposed group is lower than in the unexposed group.
In Formula 1, we use the incidence-rate ratio to compare the rate of certain events such as overtaking between two groups. For example one group using soft tires and the other hard tires, or one group doing more pit-stops than the other group.

For the analysis in Formula 1, we can use the incidence-rate ratio as follows:
The incidence-rate ratios are the exponents of the regression coefficients and they indicate the ratio at which the number of overtakes is affected by a unit increase of a control variable.
If the incidence-rate ratio is equal to 1, we don't observe an increase or decrease of the number of overtakes. Moreover, if the incidence-rate ratio is greater than 1, we observe an increase of the number of overtakes and the incidence-rate ratio less than 1 indicates a decrease in overtakes.
For dummy variables the incidence-rate ratio indicates the rate at which overtaking increases if the dummy variable goes up from 0 to 1.
Moreover, some continuous variables are log-transformed, hence the estimated coefficients can be interpreted as elasticities.
We use elasticities to measure the responsiveness of the dependent variable to changes in the independent variables. While using a linear regression model, we obtain coefficients which indicate the expected change in the dependent variable for a one-unit change in the independent variable.
Hence a positive coefficient means that as the independent variable increases, so does the dependent variable. However, if the coefficient is negative, we have that the independent variable increases and the dependent variable decreases.

Let us consider the formula of the linear regression model: $Y = \beta_0 + \beta_1 X + \epsilon$, where Y is the dependent variable and X the independent variable. Moreover, β_0 is the intercept of the true regression line, β_1 is the slope of the true regression line and ϵ is the error term.
Hence, we can derive the elasticity from a regression analysis and we obtain the following formula:

$$E = \frac{\frac{dY}{dX} \cdot X}{Y},$$

where Y is the dependent variable, X the independent variable and $\frac{dY}{dX} = \beta_1$ is the estimated coefficient from the regression.

2.6 Baseline results from the article (de Groote, 2021)

Let us look at the incidence-rate ratios of figure 13:
For the computation in the first column (1) we use track and season fixed effects. Moreover, we also consider the number of cars, reliability, number of pit-stops made by a driver, the strategic variation and the amount of mixing at the end of the first lap.
In the second column (2) we add the information about the timing of the first pit-stop and unscheduled stops. Moreover, in the third column (3) we also take the mixing due to pit-stops into account. We observe in the three columns a significant reduction in overtaking in 2014, 2015, 2017 and 2018 compared to the reference year 2011.
Moreover, we can observe that for the three columns we have a slight decline in overtaking from 2011 to 2014, but a significant drop in 2015 and from there on less overtaking.
These reductions in overtaking could be explained by some season rule changes such as the introduction of turbo engines, fuel and fuel-flow limit and the reduction of the height of the nose of the car in 2014. Moreover, we also reduced the height of the nose even more in 2015. In 2017, the Formula 1 cars became wider and the fuel limit increased to 105kg. In addition, in 2018 Formula 1 introduced triple DRS zones and the HALO system.

	(1)		(2)		(3)	
Season						
2011	Reference		Reference		Reference	
2012	0.966	(0.053)	0.965	(0.054)	0.971	(0.055)
2013	1.030	(0.059)	0.942	(0.059)	0.964	(0.061)
2014	0.869**	(0.059)	0.857**	(0.061)	0.868**	(0.061)
2015	0.693***	(0.060)	0.666***	(0.063)	0.675***	(0.063)
2016	0.947	(0.061)	0.897	(0.065)	0.904	(0.066)
2017	0.643***	(0.059)	0.606***	(0.057)	0.620***	(0.058)
2018	0.751**	(0.070)	0.762***	(0.072)	0.747***	(0.071)

Figure 13: (figure from the article (de Groote, 2021)) Number of overtakes per car (incidence-rate ratios). Note that the Asterisks depict significance levels: *10% significance level, **5% significance level, ***1% significance level

Note that the values in brackets in figure 13 refer to the standard error. Standard error (SE) is usually used to refer to the standard error of the mean, which estimates how much the sample mean deviates from the actual mean of a population. We use the following formula to compute the standard error: $SE = \frac{\sigma}{\sqrt{n}}$, where σ is the empirical standard deviation and n the sample size.
Standard deviation is used to determine the accuracy of a sample, because a large value of SE indicates more uncertainty and a relatively small SE value compared to the estimates such as $0,053$ indicates a more precise estimate.
Since all the values of the standard error in figure 13 are relatively small compared to the estimates, we can conclude that there is a decreasing of the number of overtakes from 2011 to 2018.

Let us have a closer look at the influence on overtaking of different factors in figure 14.
In the first column (1), we can clearly see that overtaking increases with the percentage of laps completed. This increase in overtaking can be explained by the fact that the more laps are completed, the more pit-stops are happening and mixing the order of the cars. Moreover, at the end of the race the tires are degrading (i.e. losing grip), which allows drivers on fresher tires to overtake.

Note that we can also observe this clear increase in overtaking in the column (2) and (3). Computing the percentage of laps completed for column (2), we obtain: $\log(2,462) \approx 0.89$. This means that if we consider a 10% increase in completed laps, we obtain approximately 9% more overtakes. Hence, we see that if a driver completes more laps, the number of overtakes increases.

Now consider column (2) where we introduced the timing of the pit-stops. We observe that earlier pit-stops increase the number of overtakes (see 2.3, pit-stop explanation). This trend is also observed in column (3). Moreover, we observe that the number of unscheduled pit-stops increases the probability of being overtaken. This increase can be explained by looking at the reasons of an unscheduled pit-stop. An unscheduled pit-stop often happens when a car has mechanical problems such as changing damaged front wing. This causes a longer pit-stop and therefore the other drivers on track are gaining time on the driver in the pits.

Note that in the analysis on the probability of overtaking the unscheduled pit-stop dummy interacts with the driver's rank based on his expected race pace. Hence if a driver with a high race pace pits late, he is more likely to overtake other drivers because of the advantage of fresher and faster tires.
Not only the timing of the pit-stops plays an important role in the analysis but also the number of pit-stops.

Let us have a look at column (3) where we include mixing due to pit-stops. Looking at the timing of stops, we see that the incidence-rate ratios for overtaking and being overtaken in case of high mixing is higher than the incidence-rate ratios for overtaking and being overtaken in case of low mixing. This observation can be explained due to the fact that the timing of a pit-stop influences mixing. Especially, if front-runners pit earlier than backmarkers and the pit-stops are early in the race. At that moment the field is still close and this results in more mixing and hence more overtaking.

Let us have a look at the probability of being overtaken and overtaking. Since the incidence-rate ratio of the percentage of pitting more often is greater than 1 for the probability of overtaking and smaller than 1 for the probability of being overtaken, we can conclude that the more a driver is pitting, the more he will overtake others because of the advantage of new, fresh tires and he will be less often overtaken.

Hence a driver who pits less often, will overtake less cars but will be overtaken by the others.

	(1)		(2)		(3)	
Season						
2011	Reference		Reference		Reference	
2012	0.966	(0.053)	0.965	(0.054)	0.971	(0.055)
2013	1.030	(0.059)	0.942	(0.059)	0.964	(0.061)
2014	0.869**	(0.059)	0.857**	(0.061)	0.868**	(0.061)
2015	0.693***	(0.060)	0.666***	(0.063)	0.675***	(0.063)
2016	0.947	(0.061)	0.897	(0.065)	0.904	(0.066)
2017	0.643***	(0.059)	0.606***	(0.057)	0.620***	(0.058)
2018	0.751**	(0.070)	0.762***	(0.072)	0.747***	(0.071)
Race fixed effects						
Number of cars (log)	1.072	(0.309)	0.913	(0.279)	0.947	(0.285)
Overall laps completed						
<87%	Reference		Reference		Reference	
87–90%	1.287***	(0.078)	1.296***	(0.079)	1.279***	(0.077)
90–93%	1.288***	(0.075)	1.319***	(0.079)	1.325***	(0.078)
93–96%	1.267***	(0.074)	1.366***	(0.084)	1.355***	(0.082)
>96%	1.094	(0.071)	1.150**	(0.078)	1.140**	(0.076)
Timing of the first stop						
First 20% of the race			Reference		Reference	
20–25%			0.862***	(0.043)	0.872***	(0.044)
25–30%			0.814***	(0.048)	0.818***	(0.049)
30–40%			0.792***	(0.065)	0.823**	(0.068)
>40%			0.764***	(0.075)	0.791**	(0.078)
Safety car	1.408***	(0.054)	1.402***	(0.058)	1.396***	(0.057)
Driver race-specific effects						
Percentage of laps completed (log)	1.852***	(0.098)	2.462***	(0.207)	2.428***	(0.212)
Number of pit-stops	1.132***	(0.045)	1.058	(0.057)	1.061	(0.058)
On probability of being overtaken						
Percentage pitting more often	2.158***	(0.195)	2.206***	(0.215)	1.882***	(0.192)
Percentage pitting less often	0.845*	(0.088)	0.996	(0.110)	0.963	(0.107)
Unknown strategy			1.785***	(0.208)	1.724***	(0.205)
Number of unscheduled early stops			1.101***	(0.034)	1.104***	(0.033)
Number of unscheduled late stops			1.012	(0.014)	1.011	(0.014)
Timing of stops: high mixing					16.504***	(7.056)
Timing of stops: low mixing					0.602	(0.453)
Position after first lap: better	7.953***	(1.466)	7.325***	(1.357)	6.770***	(1.258)
Position after first lap: worse	0.451***	(0.074)	0.409***	(0.068)	0.315***	(0.056)
On probability of overtaking						
Percentage pitting more often	0.809**	(0.076)	0.711***	(0.073)	0.739***	(0.061)
Percentage pitting less often	1.562***	(0.139)	1.730***	(0.170)	1.874***	(0.189)
Unknown strategy			1.007	(0.121)	1.101	(0.133)
Unscheduled early stop × rank			3.037***	(0.482)	3.201***	(0.562)
Unscheduled late stop × rank			1.774***	(0.206)	1.793***	(0.207)
Timing of stops: high mixing					0.998	(0.480)
Timing of stops: low mixing					0.093***	(0.057)
Position after first lap: better	0.445***	(0.098)	0.524***	(0.114)	0.526***	(0.114)
Position after first lap: worse	8.192***	(0.797)	8.483***	(0.868)	9.119***	(1.036)
Track fixed effects	yes		yes		yes	
Number of observations	5864		5864		5864	
Log likelihood	−11398		−11279		−11187	

Note: Displayed are the incidence-rate ratios. Asterisks depict significance levels. *10% significance level, **5% significance level, ***1% significance level.

Figure 14: (figure from the article (de Groote, 2021)) Number of overtakes per car (incidence-rate ratios). Note that we need to swap the values of the percentage pitting more often respectively less often on the **probability of being overtaken** with the values of the percentage pitting more often respectively less often on the **probability of overtaking**.

Let us analyze in a more detailed way the incidence-rate ratio in relation with the track fixed effects. As already mentioned, we consider Monaco as reference track because it is the worst track for overtaking.
Considering figure 15, the table starts with the circuits where overtaking is difficult such as Singapore and Albert Park. These two are street circuits, hence twisty tracks and not enough space to overtake. At the lower part of the table, we find the circuits with the highest overtaking rate such as Shanghai or Sakhir. These tracks are the most overtake-friendly because they consist of long straights and long braking zones.
However, we have some circuits (Yeongam, Sochi and Buddh), which also consist of long straights and long braking zones, hence they should be overtake-friendly, but they are at the beginning of the table and overtaking is relatively difficult. Let us have a closer look at the circuit of Buddh which was designed to be the most challenging for drivers. It consists of 16 turns and has some noticeable changes in elevation. It rises fourteen meters within the first three corners. Moreover, it has a long pitlane (more than 600 meters) which increases the duration of a pit-stop. Hence the driver looses a lot of time while pitting. All these factors increase the difficulty for the drivers to overtake. This shows that the layout of a track influences the number of overtakes.

Circuit	(1)		(2)		(3)	
Hungaroring	1.736	(0.267)	1.622	(0.251)	1.637	(0.253)
Singapore	2.491	(0.357)	2.333	(0.336)	2.359	(0.339)
Albert Park	2.791	(0.409)	2.744	(0.405)	2.755	(0.406)
Yeongam	2.838	(0.407)	2.676	(0.393)	2.746	(0.402)
Nürburgring	2.942	(0.472)	3.008	(0.506)	3.084	(0.514)
Autódromo Hermanos Rodriguez	3.091	(0.468)	2.789	(0.425)	2.848	(0.430)
Sochi	3.157	(0.489)	3.092	(0.477)	3.220	(0.499)
Buddh	3.303	(0.532)	2.972	(0.489)	2.918	(0.478)
Circuit Gilles Villeneuve	3.544	(0.500)	3.429	(0.483)	3.378	(0.478)
Suzuka	3.599	(0.510)	3.480	(0.492)	3.475	(0.491)
Valencia	3.706	(0.576)	3.646	(0.571)	3.725	(0.586)
Red Bull Ring	3.745	(0.544)	3.443	(0.512)	3.358	(0.494)
Catalunya	4.003	(0.581)	3.630	(0.518)	3.647	(0.520)
Silverstone	4.007	(0.562)	3.937	(0.559)	3.915	(0.558)
Spa-Francorchamps	4.041	(0.545)	3.835	(0.518)	3.877	(0.523)
Yas Marina	4.055	(0.546)	3.789	(0.511)	3.730	(0.505)
Interlagos	4.282	(0.616)	4.134	(0.593)	4.116	(0.589)
Monza	4.402	(0.587)	4.342	(0.596)	4.243	(0.580)
Sepang	4.419	(0.648)	3.725	(0.565)	3.771	(0.562)
Circuit of the Americas	4.533	(0.646)	4.382	(0.621)	4.337	(0.615)
Hockenheimring	4.821	(0.776)	4.294	(0.684)	4.330	(0.688)
Baku	4.945	(0.755)	4.611	(0.721)	4.640	(0.730)
Sakhir	5.461	(0.767)	4.705	(0.669)	4.780	(0.673)
Paul Ricard	5.859	(1.065)	5.514	(1.049)	5.789	(1.120)
Shanghai	6.024	(0.821)	5.647	(0.762)	5.664	(0.764)
Istanbul Park	8.412	(1.473)	7.621	(1.336)	7.677	(1.349)

Figure 15: (figure from the article (de Groote, 2021)) incidence-rate ratio of overtaking considering different circuits.

2.7 Conclusion

In the first part of this thesis, we started explaining the statistical method of regression, used to analyze the relation between independent and dependent variables. We mainly used the Simple Linear Regression Model in our analysis, by considering the following formula: $Y = \beta_0 + \beta_1 X + \epsilon$, where Y is the dependent variable and X the independent variable. In addition, ϵ is the error term and β_0 and β_1 are parameters. We estimated these two parameters using two different methods: least square estimation and maximum likelihood estimation and obtained the same result.
Moreover, we studied in more detail Poisson regression assuming that the dependent variable follows a Poisson distribution. Poisson regression, which uses count data, predicts the number of events occurring within a given interval of time or space. Using linear regression and Poisson regression, we explained the Log Likelihood, which is used to describe the process of determining the specific data distribution given a certain data set.

In the second part of this thesis, we introduced Formula 1, explaining the different regulations and rules before starting the theoretical framework of the analysis.
Our main goal is to understand which factors are able to impact the number of overtakes in Formula 1. Therefore, we mainly considered the effectiveness of DRS, reliability of the car, field size, pit-stops, cars out of position and safety car as main factors in our analysis. Moreover, we introduced the fixed effects model to better control our variables and we considered four categories of fixed effects namely track-, season-, race- and driver specific fixed effects.
After our study of the baseline results from the article (de Groote, 2021), we can conclude that we observe a clear decrease in the number of overtakes from 2011 to 2018. However, it is difficult to pinpoint the specific reason causing this decrease because the number of overtakes is influenced by many different factors. These factors cannot always be controlled, for example the timing of the first pit-stop. If a driver has front wing damage, he is forced to do an unscheduled pit-stop, which in revenge causes more mixing during the race and can increase the number of overtakes.
Moreover, we discovered that the layout of a track does not guarantee more overtaking, even if the track has long straights and long braking zones.

Finally, we can conclude that overtaking is one of the main reasons making Formula 1 exciting. Even if it is difficult to influence all factors to increase the number of overtakes, it is clear that the number of overtakes is influenced by the number of pit-stops. One solution to increase the number of overtakes is to increase the number of mandatory pit-stops, which will be the case at the Monaco (worst track to overtake) Grand Prix 2025. FIA confirmed that a two-stop strategy with two different tire compounds is mandatory for the Monaco race.

3 Simulation of data

In this section, we will generate data of a Formula 1 season using synthetic data and try to estimate the parameters. Let us first explain synthetic data.

3.1 Synthetic data

Synthetic data are data that are artificially simulated and not generated by real world events. The process to generate real world data is difficult, expensive and time-consuming. Therefore, we use the technology of synthetic data, which allows us to easily, quickly and digitally generate data in the desired amount. We also generate synthetic data to meet certain conditions that may not be found in the real data.

To simulate the data we use different algorithms and systems which depend on data to function. Moreover, synthetic data are often used in test data sets to validate mathematical models and train machine learning. Since synthetic data mimics real data sets, it allows companies to generate a lot of training data without spending a lot of time and money.

Depending on the tools and algorithms we use, we can distinguish three different techniques to create synthetic data.
First, we can simulate data using a specific distribution, in our case it will be the Poisson distribution to simulate the data of a Formula 1 season. This method allows us to produce a data distribution that resembles real world data.
Another technique to simulate data is agent-based modeling, which examines how different agents such as people, mobile phones or computer programs communicate or interact with one another.
A third method is generative models i.e. algorithms which generate synthetic data that replicates statistical properties of the real world. Using the training data, these models learn the statistical patterns and use this knowledge to generate new data similar to the original one.

3.2 Python Code: F1 2024 Season Simulation

In the following we will generate synthetic data following the Poisson distribution. We generate data of the 2024 Formula 1 season. Therefore, we start by defining the 20 drivers, the 10 different teams with their corresponding car performance and the race calendar. For the car performance, we use the ranking from 1 to 3 i.e. least reliable car to best reliable car. Moreover, we also rank the drivers depending on their driving skills and character. We do 100 simulations of the season to get more realistic results.

3.2.1 Python code for the simulation of data

```python
import pandas as pd
import numpy as np

# Set driver/team info
drivers_teams = {
    'Max Verstappen': 'Red Bull Racing',
    'Sergio Perez': 'Red Bull Racing',
    'Lewis Hamilton': 'Mercedes',
    'George Russell': 'Mercedes',
    'Charles Leclerc': 'Ferrari',
    'Carlos Sainz': 'Ferrari',
    'Lando Norris': 'McLaren',
    'Oscar Piastri': 'McLaren',
    'Fernando Alonso': 'Aston Martin',
    'Lance Stroll': 'Aston Martin',
    'Pierre Gasly': 'Alpine',
    'Esteban Ocon': 'Alpine',
    'Yuki Tsunoda': 'RB',
    'Daniel Ricciardo': 'RB',
    'Kevin Magnussen': 'Haas',
    'Nico Hulkenberg': 'Haas',
    'Zhou Guanyu': 'Sauber',
    'Valtteri Bottas': 'Sauber',
    'Alex Albon': 'Williams',
    'Logan Sargeant': 'Williams'
}

team_performance = {
    'Red Bull Racing': 3,
    'Mercedes': 3,
    'Ferrari': 3,
    'McLaren': 2,
    'Aston Martin': 2,
    'Alpine': 2,
    'RB': 1,
    'Haas': 1,
    'Sauber': 1,
    'Williams': 1
}

# 2024 race calendar
race_calendar = [
    ('Bahrain', 57, 'circuit'),
    ('Saudi Arabia', 50, 'street'),
    ('Australia', 58, 'street'),
    ('Japan', 53, 'circuit'),
    ('China', 56, 'circuit'),
    ('Miami', 57, 'street'),
    ('Emilia-Romagna', 63, 'circuit'),
    ('Monaco', 78, 'street'),
    ('Canada', 70, 'street'),
    ('Spain', 66, 'circuit'),
    ('Austria', 71, 'circuit'),
    ('Great Britain', 52, 'circuit'),
    ('Hungary', 70, 'circuit'),
```

```
    ('Belgium', 44, 'circuit'),
    ('Netherlands', 72, 'circuit'),
    ('Italy', 53, 'circuit'),
    ('Azerbaijan', 51, 'street'),
    ('Singapore', 62, 'street'),
    ('USA', 56, 'circuit'),
    ('Mexico', 71, 'circuit'),
    ('Brazil', 71, 'circuit'),
    ('Las Vegas', 50, 'street'),
    ('Qatar', 57, 'circuit'),
    ('Abu Dhabi', 58, 'street')
]

# Driver talent scores (scale: 1 = low, 3 = elite)
driver_talent = {
    'Max Verstappen': 3,
    'Sergio Perez': 2,
    'Lewis Hamilton': 3,
    'George Russell': 2,
    'Charles Leclerc': 3,
    'Carlos Sainz': 2,
    'Lando Norris': 3,
    'Oscar Piastri': 2,
    'Fernando Alonso': 3,
    'Lance Stroll': 1,
    'Pierre Gasly': 2,
    'Esteban Ocon': 2,
    'Yuki Tsunoda': 2,
    'Daniel Ricciardo': 2,
    'Kevin Magnussen': 1,
    'Nico Hulkenberg': 2,
    'Zhou Guanyu': 1,
    'Valtteri Bottas': 2,
    'Alex Albon': 2,
    'Logan Sargeant': 1
}


def simulate_season(sim_id):
    all_races_data = []
    drivers = list(drivers_teams.keys())
    teams = [drivers_teams[d] for d in drivers]
    car_perf = [team_performance[t] for t in teams]
    n = len(drivers)

    for race_name, laps, layout in race_calendar:
        # Layout-based probabilities
        safety_car_prob = 0.5 if layout == 'street' else 0.3
        dnf_prob = 0.15 if layout == 'street' else 0.08  # Higher
    DNF probability on street circuits

        # Race-level randomness
        safety_car = np.random.choice([0, 1], p=[1 -
    safety_car_prob, safety_car_prob])
        penalised = np.random.binomial(1, 0.1, n)
        penalty_drop = penalised * np.random.choice([1,2,3,5],size=
    n,p=[0.4,0.3,0.2,0.1])
```

```
        # Compute qualifying performance score
        talents = [driver_talent[d] for d in drivers]

        qualifying_score = (np.array(talents)+ np.array(car_perf)+
    np.random.normal(0, 0.15, n))  # small noise for variability

        # Weighted components for realistic qualifying variation
        talents = np.array([driver_talent[d] for d in drivers])
        car_perf_array = np.array(car_perf)

        # Lower score = better position, so we rank in reverse
        base_start = pd.Series(-qualifying_score).rank(method='
    first').astype(int).values

        # Adjust with penalty
        adjusted_raw = base_start + penalty_drop
        # Assign unique adjusted starting positions (lower
    adjusted_raw = better position)
        adjusted_start = pd.Series(adjusted_raw).rank(method='first
    ').astype(int).values

        # Pit stops
        unsched_before_5 = np.random.poisson(0.1, n)
        unsched_after_5 = np.random.poisson(0.3, n)
        total_pitstops = np.clip(unsched_before_5 + unsched_after_5
     + np.random.poisson(1, n), 1, 5)

        # Determine DNFs
        # Base DNF probability + higher for lower performance cars
    + randomness
        dnf_probs = np.clip(
            dnf_prob
            + (3 - np.array(car_perf)) * 0.02  # Lower performance
    cars more likely to DNF
            ,0, 0.25  # Keep probabilities between 0% and 25%
        )
        dnf = np.random.binomial(1, dnf_probs)

        # Overtaking & being overtaken (only for drivers who finish
    )
        aggression = np.random.normal(0, 0.005, n)
        aggression = aggression-np.mean(aggression)

        # Encode track layout effect consistently
        is_street = 1 if layout == 'street' else 0
        layout_effect = -0.2 * is_street


        lin_pred = (
        0.5 + 0.02 * (laps-60) #laps vary widely from 44 to 78 ->
    rough average
             + 0.25 * (np.array(car_perf) - 2)
        # + 0.1 * (np.array(car_perf) == 3)  # Small boost for top-
    tier cars
        # + 0.07 * (np.array(car_perf) == 2)  # small boost for mid
    -tier
        + layout_effect
```

```
        - 0.15 * unsched_before_5
        + 0.1 * unsched_after_5
        + 0.1 * np.log1p(total_pitstops)
        + 0.5 * safety_car
        + aggression
        )

        rate=np.exp(lin_pred)
        overtakes = np.random.poisson(rate) * (1 - dnf)  # DNF drivers can't overtake

        lam_overtaken = (
            2.5 - 0.3 * np.sqrt(overtakes)
            + (0.3 if layout == 'street' else -0.2)
            #+ np.random.normal(0, 0.5, n)
        )
        lam_overtaken = np.clip(lam_overtaken, 0.1, None)
        times_overtaken = np.random.poisson(lam_overtaken) * (1 - dnf)  # DNF drivers can't be overtaken

        # Result
        position_change = -overtakes + times_overtaken
        adjusted_finish = np.clip(adjusted_start + position_change.round().astype(int), 1, 20)

        # Set DNF drivers to position 100 (or any high number to indicate they didn't finish)
        adjusted_finish = np.where(dnf == 1, 100, adjusted_finish)

        # Rank the finishing positions (DNFs will be at the end)
        adjusted_finish = pd.Series(adjusted_finish).rank(method='first').astype(int).values

        race_df = pd.DataFrame({
            'simulation_id': sim_id,
            'race': f'2024-{race_name}',
            'driver': drivers,
            'team': teams,
            'car_performance': car_perf,
            'penalised': penalised,
            'penalty_grid_drop': penalty_drop,
            'starting_position': base_start,
            'adjusted_starting_position': adjusted_start,
            'finishing_position': adjusted_finish,
            'dnf': dnf,
            'overtakes': overtakes,
            'times_overtaken': times_overtaken,
            'unsched_pitstops_before_5': unsched_before_5,
            'unsched_pitstops_after_5': unsched_after_5,
            'total_pitstops': total_pitstops,
            'safety_car': safety_car,
            'race_distance': laps,
            'track_layout': layout_effect,
            'street_circuit':is_street,
            'aggression': aggression
        })

```

```
208         all_races_data.append(race_df)
209
210     return pd.concat(all_races_data, ignore_index=True)
211
212 #       Run multiple simulations
213 simulations = 100  # Change this to 1000+ for more realistic
    results
214 all_simulations = pd.concat([simulate_season(sim_id) for sim_id in
    range(1, simulations + 1)], ignore_index=True)
215
216 #       Save to CSV
217 all_simulations.to_csv("f1_2024_full_season_sim10_with_dnf.csv",
    index=False)
218 print("    Monte Carlo season saved as '
    f1_2024_full_season_sim10_with_dnf.csv'")
```

3.2.2 Analysis of the Python code

Let us have a look at our "lin pred" (i.e. linear combination of the predictors) expression in the Python code. This expression defines the expected number of overtakes for each driver in a race, modeled as a Poisson rate with parameter λ influenced by different factors:

1. race distance in number of laps
2. car performance (reliability of the car)
3. track layout
4. unscheduled pit-stop before lap 5
5. unscheduled pit-stop after 5 laps
6. total number of pit-stops
7. safety car dummy
8. aggression of the driver i.e. driver specific fixed effects

Each one of these parameters influences the predicted number of overtakes for each driver. Note that the track layout has an important influence not only on the overtake chances but also on the pit-stop strategy and the probability of having a safety car.
If we consider a street circuit then the overtaking chances are lower than for a not street circuit. Moreover, pit-stop strategies are often riskier for street circuits because of tight pit entries and exits. Therefore, we have less pit-stops for street circuits than for not street circuits.
In addition, the probability of a safety car is $\approx 50\%$ higher for a street circuit than for a "normal" circuit.

We use the following regression expression for the "lin pred" (Figure 16):

$$\begin{aligned}\text{lin pred} = {} & 0.5 + 0.02 \cdot (\text{laps} - 60) + 0.25 \cdot (\text{car performance} - 2) + \text{layout effect} \\ & - 0.15 \cdot (\text{unscheduled pit-stop before lap } 5) + 0.1 \cdot (\text{unscheduled pit-stop after } 5 \text{ laps}) \\ & + 0.1 \cdot \log(\text{total number of pit-stops}) + 0.5 \cdot \text{safety car} + \text{aggression}\end{aligned}$$

Component	**Interpretation/ Justification**
0.5	Baseline overtaking rate i.e. assumed base for a midfield driver.
$0.02 \cdot$ (**laps** $- 60$)	Each extra lap adds ≈ 0.02 overtakes i.e. if the number of completed laps increases so does the number of overtakes. Since the number of laps of the different races vary widely from 44 to 78, we use 60 as the rough average.
$0.25 \cdot$ (**car performance** $- 2$)	Car performance effect for example a high-performance car such as Red Bull gains $\approx 0.4 - 0.8$ overtakes.
layout_effect **i.e. is_street =** 1 **if layout == `street` else** 0 **layout_effect =** $-0.2 \cdot$ **is_street**	Track layout describes if it is a street circuit or not or if it is a overtake friendly circuit with long straights and long braking zones. We use **is_street =** 1 **if layout == `street` else** 0 to convert the layout string (`street` = 1 or `circuit` = 0) into a binary variable. In addition, **layout_effect =** $-0.2 \cdot$ **is_street** applies a penalty of -0.2 for street circuits in our model.
$-0.15 \cdot$ (**unscheduled pit-stop before lap 5**)	Early unscheduled pit-stop usually because of a mechanical problem, often causes the car to be less competitive and to loose a lot of places.
$0.1 \cdot$ (**unscheduled pit-stop after 5 laps**)	Later unscheduled pit-stop allows a possible recovery drive i.e. even if the driver looses some places during the pit-stop, he is able to overtake more cars because of fresher tires.
$0.1 \cdot \log($**total number of pit-stops**$)$	The more a driver is pitting, the more he will overtake others because of the advantage of new, fresh tires. We use $\log$(total number of pit-stops) to obtain more realism in our simulation.
$0.5 \cdot$ **safety car**	Safety car eradicates the gaps between the drivers and increases the number of overtakes.
aggression	Driver specific fixed effects i.e. personality of a driver and driving style.

Figure 16: Interpretation of components used in "lin pred" expression

Note that we define the rate as exp(lin pred) to ensure that the predicted mean is positive and matches the Poisson logic.

Finally, we can generate the number of overtakes using the following expression: overtakes $=$ np.random.poisson(rate) $\cdot (1 - \text{dnf})$, where we consider that dnf (i.e did not finish) drivers cannot overtake anymore.
This expression simulates actual overtakes from the expected count.

3.3 Ranking of the drivers based on our simulation

Before doing the estimation of the parameters used in our simulation of synthetic data, we want to rank the twenty drivers across a Formula 1 season with 100 simulations.

We start by computing the average position of a driver during a specific race for 100 repetitions. Hence, we obtain the driver's average rank of a race. In addition, we rank the drivers from 1 to 20 i.e. best average position to worst average position.

Moreover, since we know how many points are distributed to each driver according to his position, we can compute the average amount of points each driver received for each race.

3.3.1 Python code used for the ranking

```python
import pandas as pd

# Load the simulation results
all_simulations = pd.read_csv("f1_2024_full_season_sim10_with_dnf.csv")

# Calculate mean finishing position per driver per race
mean_positions = (
    all_simulations
    .groupby(['race', 'driver'])['finishing_position']
    .mean()
    .reset_index()
    .rename(columns={'finishing_position': 'mean_finishing_position'})
)

# Rank drivers per race
mean_positions['race_rank'] = (
    mean_positions
    .groupby('race')['mean_finishing_position']
    .rank(method='min')
)

mean_positions.sort_values(by=['race', 'race_rank'], inplace=True)

# Save to CSV
mean_positions.to_csv("f1_2024_mean_positions_and_rankings.csv", index=False)
print("     Mean positions and rankings saved as 'f1_2024_mean_positions_and_rankings.csv'")
```

3.3.2 Python code used for the points distribution

```python
import pandas as pd

# Load full simulation results (one row per driver per race per simulation)
all_simulations = pd.read_csv("f1_2024_full_season_sim10_with_dnf.csv")

# F1 points system for top 10
points_table = {
    1: 25,
    2: 18,
    3: 15,
    4: 12,
    5: 10,
    6: 8,
    7: 6,
    8: 4,
    9: 2,
    10: 1
}

# Assign race points per simulation (based on actual finishing position)
all_simulations['points'] = all_simulations['finishing_position'].map(points_table).fillna(0).astype(int)

# Average points per driver per race (across simulations)
avg_points_per_race = (
    all_simulations
    .groupby(['race', 'driver'])['points']
    .mean()
    .reset_index()
    .rename(columns={'points': 'average_points'})
)

# Round for readability
avg_points_per_race['average_points'] = avg_points_per_race['average_points'].round(2)

# Sort for clarity
avg_points_per_race = avg_points_per_race.sort_values(by=['race', 'average_points'], ascending=[True, False])

# Save the result
avg_points_per_race.to_csv("f1_2024_average_points_per_driver_per_race.csv", index=False)
print("     Average points per race per driver saved as 'f1_2024_average_points_per_driver_per_race.csv'")
```

3.3.3 Results

In the following, let us have a look at the ranking and point distribution of the Abu Dhabi race. Using our two Python codes for the ranking and the points distribution, we obtain similar tables for each race of the 2024 Formula 1 season. Looking at figure 17, we can conclude that the lower the mean finishing position of a driver across all simulations is the better is his rank. For example, on the one hand Max Verstappen has a mean finishing position of 4.75, therefore he is rank 1 for the Abu Dhabi race and he scores an average of 16.14 points. On the other hand, Zhou Guanyu, who is rank 19 in the Abu Dhabi race, has an average finishing position of 16.58 and scores 0 points on average.

We observe that the Formula 1 point system only rewards the top 10 drivers i.e. the drivers who are consistently in the top 10. The average amount of points drops drastically for the drivers outside the top 10.

race	driver	mean_finishing_position	race_rank
2024-Abu Dhabi	Max Verstappen	4.75	1.0
2024-Abu Dhabi	Charles Leclerc	5.21	2.0
2024-Abu Dhabi	Lewis Hamilton	5.21	2.0
2024-Abu Dhabi	Sergio Perez	5.54	4.0
2024-Abu Dhabi	Carlos Sainz	6.17	5.0
2024-Abu Dhabi	George Russell	6.28	6.0
2024-Abu Dhabi	Lando Norris	7.09	7.0
2024-Abu Dhabi	Fernando Alonso	8.12	8.0
2024-Abu Dhabi	Esteban Ocon	10.18	9.0
2024-Abu Dhabi	Pierre Gasly	10.3	10.0
2024-Abu Dhabi	Oscar Piastri	10.74	11.0
2024-Abu Dhabi	Lance Stroll	12.49	12.0
2024-Abu Dhabi	Valtteri Bottas	13.17	13.0
2024-Abu Dhabi	Nico Hulkenberg	13.33	14.0
2024-Abu Dhabi	Daniel Ricciardo	13.7	15.0
2024-Abu Dhabi	Yuki Tsunoda	13.73	16.0
2024-Abu Dhabi	Alex Albon	14.91	17.0
2024-Abu Dhabi	Kevin Magnussen	15.82	18.0
2024-Abu Dhabi	Zhou Guanyu	16.58	19.0
2024-Abu Dhabi	Logan Sargeant	16.68	20.0

race	driver	average_points
2024-Abu Dhabi	Max Verstappen	16.14
2024-Abu Dhabi	Lewis Hamilton	14.38
2024-Abu Dhabi	Charles Leclerc	14.13
2024-Abu Dhabi	Sergio Perez	11.09
2024-Abu Dhabi	George Russell	9.86
2024-Abu Dhabi	Carlos Sainz	9.8
2024-Abu Dhabi	Lando Norris	8.02
2024-Abu Dhabi	Fernando Alonso	6.78
2024-Abu Dhabi	Esteban Ocon	2.9
2024-Abu Dhabi	Oscar Piastri	2.77
2024-Abu Dhabi	Pierre Gasly	2.75
2024-Abu Dhabi	Lance Stroll	0.72
2024-Abu Dhabi	Nico Hulkenberg	0.42
2024-Abu Dhabi	Valtteri Bottas	0.42
2024-Abu Dhabi	Yuki Tsunoda	0.37
2024-Abu Dhabi	Daniel Ricciardo	0.25
2024-Abu Dhabi	Alex Albon	0.19
2024-Abu Dhabi	Kevin Magnussen	0.01
2024-Abu Dhabi	Logan Sargeant	0.0
2024-Abu Dhabi	Zhou Guanyu	0.0

Figure 17: The first table shows the average finishing position i.e. rank of a driver for the Abu Dhabi race. The second table shows the average amount of points each driver obtained for the Abu Dhabi race.

3.4 Estimation of the different parameters

3.4.1 Python code used for the estimation of the parameters

In the following, we use the generated data in section 3.2 to estimate the parameters from the regression expression.

```python
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import numpy as np

# Load simulated data
df = pd.read_csv("f1_2024_full_season_sim10_with_dnf.csv")
df["log_total_pitstops"]=np.log1p(df["total_pitstops"])
# Define the regression formula (treat track_layout as categorical
    without dummies)
df["race_distance"]=df["race_distance"]-60
df["car_performance"]=df["car_performance"]-2

formula = (
    'overtakes ~ car_performance + race_distance + '
    ' + aggression '
   '+unsched_pitstops_before_5+unsched_pitstops_after_5 +
    log_total_pitstops + safety_car + '
    'street_circuit'
)

# Store results across simulations
coefficients = []
standard_errors = []

for sim_id, season in df.groupby("simulation_id"):
    # Fit a Poisson GLM (for count data: overtakes)
    model = smf.glm(
        formula=formula,
        data=season,
        family=sm.families.Poisson()  # Overtakes are count data
    ).fit()

    coefficients.append(model.params)
    standard_errors.append(model.bse)

# Aggregate results across simulations
coef_df = pd.DataFrame(coefficients)
stderr_df = pd.DataFrame(standard_errors)

summary_df = pd.DataFrame({
    "variable": coef_df.columns,
    "avg_coef": coef_df.mean(),
    "avg_std_err": stderr_df.mean()})

# Save to CSV
summary_df.to_csv("poisson_avg_parameters10.csv", index=False)
print("    Regression results saved to 'poisson_avg_parameters10.
    csv'")
```

Let us consider the following log-linear model expression, i.e. the log of the expected number of overtakes per driver i per race:

$$\begin{aligned}\log(\lambda_i) = \beta_0 &+ \beta_1 \cdot (\text{laps}_i - 60) \\ &+ \beta_2 \cdot (\text{car\_performance}_i - 2) \\ &+ \beta_3 \cdot \text{unsched\_pitstops\_before\_lap\_5}_i \\ &+ \beta_4 \cdot \text{unsched\_pitstops\_after\_5\_laps}_i \\ &+ \beta_5 \cdot \text{total\_pitstops}_i \\ &+ \beta_6 \cdot \text{safety\_car}_i \\ &+ \beta_7 \cdot \text{track\_layout\_street}_i \\ &+ \beta_8 \cdot \text{aggression}_i\end{aligned}$$

where β_0 is the intercept, which is the baseline log-rate and β_j for $j \in \{1, ..., 8\}$ is the estimated coefficient for the corresponding value of the predictors.

3.4.2 Representation of the estimated parameters

We use the Python code to estimate the parameters following the Poisson distribution and we compute the average of the parameters. Hence we obtain the following table (Figure 18) with average of the coefficients and average standard error:

variable	avg_coef	avg_std_err
Intercept	0.396248933434811	0.1192655853734926
car_performance	0.27266912209063543	0.04014374439705097
race_distance	0.019941587133015644	0.003902343481830647
aggression	0.5803936738430949	6.8595700892232445
unsched_pitstops_before_5	-0.1480917913162359	0.11784208537713686
unsched_pitstops_after_5	0.09993262079308135	0.0657472518104715
log_total_pitstops	0.08536246309286016	0.12647816571058718
safety_car	0.5029629066975272	0.07411728484240113
street_circuit	-0.2774995746604652	0.07502540785899757

Figure 18: Average of estimated parameters (i.e. coefficients) and average standard error

To better compare the true parameters obtained during the simulations and the estimated parameters, we represent them in a histogram (Figure 19).

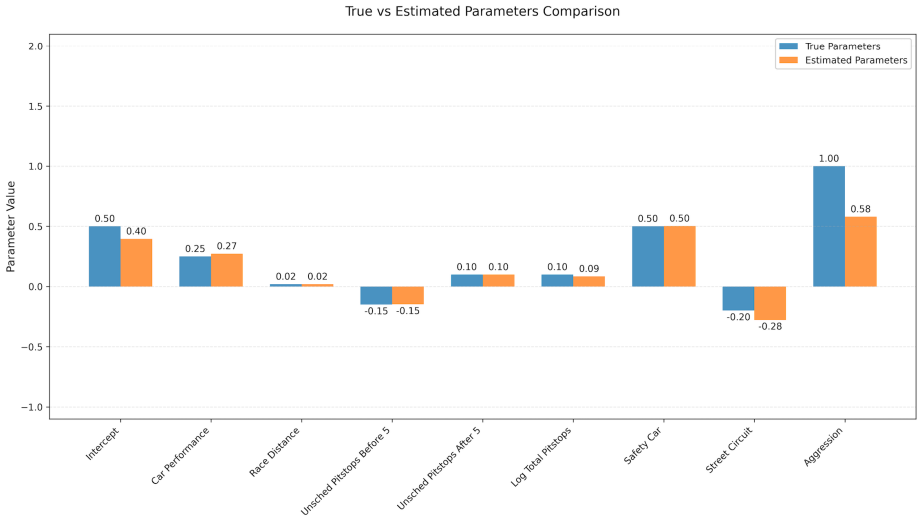

Figure 19: Comparison between true and estimated parameters

3.4.3 Explanation and conclusion about the simulation of data and estimation of the parameters

In figure 19 the blue bars represent the true parameters we used in our simulation and the orange bars represent the estimated parameters from Poisson regression on the synthetic data.
The closer the orange bars i.e. estimated parameters are to the blue bars i.e. simulated parameters, the better is our regression model.
Looking at the true parameters and the estimations we can conclude the following:

- Intercept: We observe a slight underestimation which can be explained due to the fact that the intercept depends on how the other variables are centered.
- Car performance: We observe that both values are close hence our model detects correctly that the more a car is reliable, the more it can overtake.
- Race distance (i.e. laps): We observe that both values are identical, which indicates that our model detects that the more laps are completed the more we observe overtaking.

- Unscheduled pit-stop before lap 5: We observe the same value for the true and estimated parameter, therefore we can deduce that our regression model recovers that early unscheduled pit-stops often due to mechanical issue reduce the number of overtaking.
- Unscheduled pit-stop after 5 laps: We observe the same value for the simulated and estimated parameter, hence our model detects that more late pit-stops increase the number of overtakes.
- Log of the total number of pit-stops: We have a small underestimation, but our model still detects that more pit-stops increase the number of overtaking.
- Safety car: We observe that both values of the parameters are identical, hence we can conclude that safety car causes the field to get closer together and therefore provokes more overtaking.
- Street circuit: We observe an overestimation of this parameter i.e. street circuits don't reduce the number of overtaking as much as our model claims.
- Aggression: We observe a large underestimation, it is difficult for our regression model to detect the effect of aggression.

To conclude, we can summarize that in the first part of section 3, we generated Formula 1 data based on the car performance, race distance, pit-stops (unscheduled and total number of pit-stops), safety car, track layout, and aggression. Moreover, we used randomized inputs such as penalties and dnf (i.e. did not finish) which allows the model to be more realistic. Using our regression model, we can control the effect of each variable to simulate the total number of overtakes.

Next we estimated the parameters of our regression model and we observed that most values of our parameters are close to the simulated value.
Hence, we can conclude that we were able to simulate and estimate the number of overtakes of a Formula 1 season, where the number of overtakes follows a Poisson distribution.

Sources

Regression

[1] Statistics How To Regression. *Regression Analysis*. https://www.statisticshowto.com/probability-and-statistics/regression-analysis/

[2] Investopedia. *Regression Definition*. https://www.investopedia.com/terms/r/regression.asp

[3] Harvard Business Review. *A Refresher on Regression Analysis*. https://hbr.org/2015/11/a-refresher-on-regression-analysis

[4] Save My Exams. *Outliers, High Leverage and Influential Points*. https://www.savemyexams.com/ap/statistics/college-board/20/revision-notes/exploring-two-variable-data/scatterplots-and-regression/outliers-high-leverage-and-influential-points/

[5] Yale University. *Linear Regression*. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[6] Analytics Vidhya. *Everything About Linear Regression*. https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/

[7] University of Colorado. *Regression Notes*. https://www.colorado.edu/amath/sites/default/files/attached-files/ch12_0.pdf

[8] IIT Kanpur. *Simple Linear Regression*. https://home.iitk.ac.in/~shalab/regression/Chapter2-Regression-SimpleLinearRegressionAnalysis.pdf

[9] ASQ. *Scatter Diagram*. https://asq.org/quality-resources/scatter-diagram

[10] Towards Data Science. *Obstacles in Linear Regression*. https://towardsdatascience.com/five-obstacles-faced-in-linear-regression-80fb5c599fbc/

[11] Probability. *Probability course*. https://www.probabilitycourse.com/chapter3/3_2_1_cdf.php

[12] Seltman, H. *Experimental Design and Analysis, Chapter 9*. Department of Statistics, Carnegie Mellon University. 2018. https://www.stat.cmu.edu/~hseltman/309/Book/chapter9.pdf

[13] Wikipedia. *Generalized linear model*. https://en.wikipedia.org/wiki/Generalized_linear_model

Poisson Regression

[14] ScienceDirect. *Poisson Regression.* https://www.sciencedirect.com/topics/psychology/poisson-regression

[15] Scribbr. *Poisson Distribution.* https://www.scribbr.com/statistics/poisson-distribution/

[16] Wikipedia. *Poisson Distribution.* https://en.wikipedia.org/wiki/Poisson_distribution

[17] YouTube. *Poisson Regression Video.* https://youtu.be/Obpz_Uvo2rQ?si=nUCJO7buZO6eUtIl

[18] Wikipedia. *Poisson Regression.* https://en.wikipedia.org/wiki/Poisson_regression

[19] Penn State University. *Poisson Regression Notes.* https://online.stat.psu.edu/stat501/lesson/t/t.3/t.3.1-poisson-regression#:~:text=For%20a%20sample%20of%20size,i%20%3D%201%20n%20log%20%E2%81%A1

[20] ST47S. *Poisson Regression.* http://st47s.com/Math150/Notes/poisson-regression.html

[21] Probability course. *Probability mass function.* https://www.probabilitycourse.com/chapter3/3_1_3_pmf.php

[22] Probability course. *Probability density function.* https://www.probabilitycourse.com/chapter4/4_1_1_pdf.php

[23] Stanford University. *Statistics 200: Lecture 27 Notes.* Department of Statistics, Stanford University. 2017. https://web.stanford.edu/class/archive/stats/stats200/stats200.1172/Lecture27.pdf

Likelihood Methods

[24] StatLect. *Log-Likelihood Function.* https://www.statlect.com/glossary/log-likelihood#:~:text=The%20log%2Dlikelihood%20function%20is%20typically%20used%20to%20derive%20the,as%20maximizing%20the%20likelihood%20function

[25] Analyttica. *Log-Likelihood Function Series.* https://medium.com/@analyttica/log-likelihood-analyttica-function-series-cb059e0d379#:~:text=Log%20Likelihood%20value%20is%20a,Likelihood%20values%20between%20multiple%20models.

[26] Simplilearn. *Difference Between Probability and Likelihood.* https://www.simplilearn.com/tutorials/statistics-tutorial/difference-between-probability-and-likelihood

[27] Aptech. *Beginner's Guide to Maximum Likelihood Estimation in Gauss.* https://www.aptech.com/blog/beginners-guide-to-maximum-likelihood-estimation-in-gauss/

[28] Gordan Žitković. *Likelihood Function.* Department of Mathematics, University of Texas at Austin. https://web.ma.utexas.edu/users/gordanz/notes/likelihood_color.pdf

R-squared

[29] Penn State University. *R-squared Explanation.* https://online.stat.psu.edu/stat462/node/95/

[30] Investopedia. *R-squared Definition.* https://www.investopedia.com/terms/r/r-squared.asp

[31] Graphpad. *R-squared .* https://www.graphpad.com/guides/prism/latest/curve-fitting/r2_ameasureofgoodness_of_fitoflinearregression.htm

Correlation

[32] Yale University. *Correlation Analysis.* http://www.stat.yale.edu/Courses/1997-98/101/correl.htm

[33] Cuemath. *Correlation Coefficient Formula.* https://www.cuemath.com/correlation-coefficient-formula/

Standardization

[34] Medium. *Feature Engineering 101.* https://medium.com/@brijesh_soni/feature-engineering-101-7cb68d293551#:~:text=Normalization%20can%20be%20done%20by,range%20between%200%20and%201.

[35] Alooba. *Standardization in Statistics.* https://www.alooba.com/skills/concepts/statistics/standardization/#:~:text=Standardization%20is%20a%20fundamental%20concept,analysis%20and%20effective%20decision%2Dmaking.

[36] Secoda. *How to Standardize Data.* https://www.secoda.co/learn/how-to-standardize-data#:~:text=Formula%3A%20The%20formula%20for%20z,standard%20deviation%20of%20the%20data

[37] EDUCBA. *Normalization Formula.* https://www.educba.com/normalization-formula/

Fixed effects

[38] Wikipedia. *Dummy Variables in Statistics*. https://en.wikipedia.org/wiki/Dummy_variable_(statistics)#:~:text=In%20regression%20analysis%2C%20a%20dummy,expected%20to%20shift%20the%20outcome.

[39] Wikipedia. *Cross-Sectional Data*. https://en.wikipedia.org/wiki/Cross-sectional_data

[40] StatHelp. *Cross-Sectional vs Panel Data*. https://www.stathelp.se/en/xtset_en.html#:~:text=Cross%20sectional%20data%20means%20that,over%20many%20points%20in%20time.

[41] ScienceDirect. *Fixed Effects Model*. https://www.sciencedirect.com/topics/social-sciences/fixed-effects-model#:~:text=Fixed%2Deffects%20estimation%20uses%20only,invariant%20individual%2Dspecific%20dummy%20variables.

[42] The Effect. *Fixed Effects Chapter*. https://theeffectbook.net/ch-FixedEffects.html

[43] Econometrics with R. *Fixed Effects Regression*. https://www.econometrics-with-r.org/10.3-fixed-effects-regression.html

[44] IJEBH. *Fixed vs Random Effects Meta-Analysis*. https://journals.lww.com/ijebh/fulltext/2015/09000/fixed_or_random_effects_meta_analysis__common.12.aspx#:~:text=In%20fixed%2Deffects%20models%2C%20we,(usually%20a%20normal%20distribution).

[45] Akif IIPS. *Understanding Random and Fixed Effects*. https://medium.com/@akif.iips/understanding-random-effect-and-fixed-effect-in-statistical-analysis-db4983cdf8b1

[46] Statistics by Jim. *Heterogeneity Basics*. https://statisticsbyjim.com/basics/heterogeneity/

[47] Data Science Medium. *Fixed Effect Regression Simply Explained*. https://medium.com/data-science/fixed-effect-regression-simply-explained-ab690bd885cf

[48] Investopedia. *Multicollinearity Definition*. https://www.investopedia.com/terms/m/multicollinearity.asp#:~:text=Multicollinearity%20exists%20whenever%20an%20independent,the%20statistical%20inferences%20less%20reliable.

[49] Princeton University. *Panel Data Analysis 101*. https://www.princeton.edu/~otorres/Panel101.pdf

[50] Shoemaker, K. *Mixed Effects Models II.* https://kevintshoemaker.github.io/NRES-746/MEM2.html

[51] Carnegie Mellon University. *Mixed Models Chapter.* https://www.stat.cmu.edu/~hseltman/309/Book/chapter15.pdf

[52] MathWorks. *Linear Mixed-Effects Models.* https://www.mathworks.com/help/stats/linear-mixed-effects-models.html

[53] Wikipedia. *Multilevel Model.* https://en.wikipedia.org/wiki/Multilevel_model

[54] APA Dictionary. *Hierarchical Model.* https://dictionary.apa.org/hierarchical-model

[55] Environmental Computing. *Mixed Models 2.* https://environmentalcomputing.net/statistics/mixed-models/mixed-model-2/#:~:text=Two%20factors%20are%20crossed%20when,categories%20for%20the%20two%20factors.

[56] Number Analytics. *Understanding Random Effects Models.* https://www.numberanalytics.com/blog/understanding-random-effects-model-data-analysis

[57] Wikipedia. *Fixed Effects Model.* https://en.wikipedia.org/wiki/Fixed_effects_model

[58] Princeton Library. *Fixed Effects in Stata.* https://libguides.princeton.edu/stata-panel-fe-re#:~:text=Entity%20fixed%20effects%20account%20for,model%20in%20panel%20data%20analysis.

[59] Number Analytics. *Fixed Effects Model Techniques.* https://www.numberanalytics.com/blog/fixed-effects-model-techniques

[60] Appinio. *Interval Scale in Market Research.* https://www.appinio.com/en/blog/market-research/interval-scale

[61] Econometrics. *Fixed effects regression.* https://www.econometrics-with-r.org/10.3-fixed-effects-regression.html

Incidence-rate ratio

[62] Investopedia. *Incidence Rate Definition.* https://www.investopedia.com/terms/i/incidence-rate.asp

[63] UNC Gillings School of Public Health. *Epidemiologic Research Methods.* https://sph.unc.edu/wp-content/uploads/sites/112/2015/07/nciph_ERIC3.pdf

[64] Statology. *Incidence Rate Ratio Explanation*. https://www.statology.org/incidence-rate-ratio/

[65] DataZip. *Introduction to Elasticity in Statistics*. https://datazip.io/blog/quantifying-the-flexibility-of-data-an-introduction-to-elasticity-in-statistics

[66] Jang, D. *Elasticity in Machine Learning*. https://medium.com/@jangdaehan1/elasticity-and-regression-analysis-in-machine-learning-a-comprehensive-expl

Standard error, Standard deviation

[67] University of Southampton Library. *Variance, Standard Deviation and Standard Error*. https://library.soton.ac.uk/variance-standard-deviation-and-standard-error#Main%20heading%203

[68] Investopedia. *Standard Error Definition*. https://www.investopedia.com/terms/s/standard-error.asp

Formula 1

[69] Catapult. *Race Strategy: F1 Track Surface*. https://www.catapult.com/blog/race-strategy-f1-track-surface

[70] Formula 1. *Essential F1 Guide: Drivers, Teams, Circuits*. https://www.formula1.com/en/latest/article/drivers-teams-cars-circuits-and-more-everything-you-need-to-know-about.7iQfL3Rivf1comzdqV5jwc

[71] Wikipedia. *List of Formula One Circuits*. https://en.wikipedia.org/wiki/List_of_Formula_One_circuits

[72] Wikipedia. *History of Formula One Regulations*. https://en.wikipedia.org/wiki/History_of_Formula_One_regulations

[73] Global Sports Advocates. *Understanding the F1 Cost Cap*. https://www.globalsportsadvocates.com/blog/understanding-the-f1-cost-cap.cfm#:~:text=The%20originally%20planned%20%24175%20million,for%20inflation%20in%20future%20years.

[74] World of Speed. *What Are Backmarkers in F1?*. https://worldofspeed.org/blog/what-are-backmarkers-in-f1/

[75] Reddit. *Frontrunners vs Midfield Discussion*. https://www.reddit.com/r/formula1/comments/2ewvqm/eli5_whats_the_major_difference_between/?rdt=43965#:~:text=Race%20leaders%2FFrontrunners%20are%20the,extent%20Williams%20and%20Ferrari%2FAlonso.

[76] Statathlon. *Pit Stop Strategy Analysis.* https://statathlon.com/analysis-of-the-pit-stop-strategy-in-f1/#:~:text=In%20F1%2C%20a%20pit%20stop,connected%20to%20the%20main%20track.

[77] Catapult. *Undercut vs Overcut Strategies.* https://www.catapult.com/blog/motorsports-race-strategy-undercut-overcut#:~:text=An%20undercut%20requires%20a%20clear,air%20before%20making%20their%20stop.

[78] LAS Motorsport. *Dirty Air in Formula 1.* https://las-motorsport.com/f1/blog/dirty-air-in-formula-1-understanding-the-turbulent-wake-effect/5063/#:~:text=Turbulent%20wake%20is%20the%20chaotic,performance%20of%20a%20following%20car.

[79] Autosport. *F1 Tyre Compounds Explained.* https://www.autosport.com/f1/news/f1-tyres-what-are-the-compounds-and-what-do-they-mean/10344284/

[80] Formula 1. *Beginner's Guide to F1 Tyres.* https://www.formula1.com/en/latest/article/the-beginners-guide-to-formula-1-tyres.61SvF0Kfg29UR2SPhakDqd

[81] Formula 1. *2026 Technical Regulations.* https://www.formula1.com/en/latest/article/fia-unveils-formula-1-regulations-for-2026-and-beyond-featuring-more-agile.75qJiYOHXgeJqsVQtDr2UB

[82] Wikipedia. *Halo Safety Device.* https://en.wikipedia.org/wiki/Halo_(safety_device)

[83] Everything F1. *2025 Pirelli Tyre Development.* https://www.everythingf1.com/pirelli-working-on-softer-c6-tires-for-2025-f1-season/

[84] F1 Experiences. *F1 Glossary.* https://f1experiences.com/de/blog/f1-glossary-a-z-most-commonly-used-terminology

[85] Statathlon. *Pit Stop Strategy Analysis.* https://statathlon.com/analysis-of-the-pit-stop-strategy-in-f1/

[86] Reddit. *F1 Technical Discussion.* https://www.reddit.com/r/F1Technical/comments/1753y1u/can_someone_please_explain_the_difference_between/?rdt=40421

[87] Red Bull Racing. *F1 Aerodynamics.* https://www.redbull.com/au-en/f1-technique-wind-on-car-aerodynamics

[88] RacingCircuits.info. *Buddh International Circuit.* https://www.racingcircuits.info/asia/india/buddh-international-circuit.html

[89] Wikipedia. *Buddh International Circuit*. https://en.wikipedia.org/wiki/Buddh_International_Circuit

[90] Motorsport.com. *Monaco GP Pit Strategy*. https://www.motorsport.com/f1/news/fia-two-pitstop-strategy-monaco-gp/10698674/

Synthetic data

[91] Techtarget *Synthetic data*. https://www.techtarget.com/searchcio/definition/synthetic-data#:~:text=Synthetic%20data%20is%20information%20that's,ML)%20and%20deep%20learning%20models.

[92] Wikipedia. *Synthetic date*. https://en.wikipedia.org/wiki/Synthetic_data