How Pace Deficit and Starting Grid Relate to Finishing Position in Formula 1*

Suneel Chandra Vanamu

2025-10-23

This paper investigates how a driver's on-day pace—measured by the **best-lap deficit** (the difference between a driver's quickest lap and the race's overall fastest lap)—and **starting grid position** (qualifying rank on the starting grid) relate to **finishing position** in Formula 1 races. The goal is to quantify how much each of these straightforward metrics helps predict race outcomes, without resorting to complex multivariable models.

Using data from the **Formula 1 Database (F1DB)** covering the **2014–2024 hybrid-engine era**, when power-unit regulations stabilized and race formats were relatively consistent, I fit two simple linear regressions: best-lap pace deficit predicting finishing position, and starting-grid position predicting finishing position.

The hybrid-era restriction helps ensure comparability across seasons, though **track-specific differences** (such as overtaking difficulty or circuit length) remain unmodeled and may affect the results. I summarize each model's slope, 95% confidence interval, and diagnostics including $\mathbf{R^2}$, **adjusted** $\mathbf{R^2}$, and **residual standard error** (in finishing-position units).

1 Introduction

Results in Formula 1 can be influenced by many changing factors such as strategy, traffic, and weather, but two straightforward indicators are almost always center stage: a driver's raw pace and their starting grid position. This report focuses on a simple but practical question: How do a driver's deficit on their best lap and their starting grid position influence where they finish? In everyday terms, if a driver is slower on their best lap or starts farther back, do they also tend to finish farther back?

This question matters because teams invest significant effort in qualifying for clean air, and commentators often emphasize speed as the key difference-maker. Establishing a single, quantitative relationship between pace, grid position, and results allows analysts to discuss performance without relying on complex multivariable models. To make fair comparisons across different seasons and circuits, this analysis focuses on the **2014–present hybrid-engine era**, when Formula 1 introduced consistent power-unit regulations that made cars and race formats more comparable. However, track-specific differences—such as overtaking difficulty, circuit length, and layout—remain unmodeled and may affect the results.

*Project repository: https://github.com/Suneel1508/MATH261A-project1

Using the **F1DB** data set, I construct a driver-race table for this hybrid-era period and fit two **single-predictor linear regression models**: (1) best-lap deficit predicting finishing position, and (2) starting grid position predicting finishing position.

The **best-lap deficit** is defined as the time difference (in seconds) between a driver's fastest lap during the race and the overall race's quickest lap. For each model, I report how closely the predictor aligns with finishing order (R^2 and adjusted R^2), the model's typical miss in finishing positions (residual standard error), and the average change in finishing position for a one-unit change in the predictor (slope with a 95% confidence interval). I also include scatter plots with regression lines and basic residual checks to assess linearity and fit quality.

The rest of the paper is organized as follows: **Section 2** describes the dataset and key variables; **Section 3** summarizes exploratory analysis; **Section 4** builds and tests the models; **Section 5** presents results and interpretation; and **Section 6** concludes with key findings and practical implications.

2 Data and Variables

This analysis uses the public Formula 1 Database (F1DB contributors 2025) covering races from **2014 through 2024**, corresponding to the **hybrid-engine era**. All officially classified race results are included; every driver who completed a race or was given a finishing position by the FIA appears once per race. Data files were read and tidied in R (R Core Team 2024) using `readr`(Wickham, Hester, and François 2024), `janitor`(Firke 2023), and `tidyr`(Wickham and Girlich 2024), and graphics were created with `ggplot2`(Wickham 2016).

To ensure comparability across seasons and tracks, only championship races from this era are included. For each driver in each race, three core fields are used:

- **Finishing position (response):** Integer rank, where 1 = winner.
- **Best-lap pace deficit (seconds, predictor A):** The difference between a driver's fastest lap time and the race's overall fastest lap. Smaller values indicate faster pace.
- **Starting-grid position (predictor B):** The driver's qualifying rank at the race start (1 = pole position; higher numbers = further back).

This dataset covers hundreds of races and thousands of driver-race observations, making it suitable for exploring how simple single-predictor models perform across varied conditions.

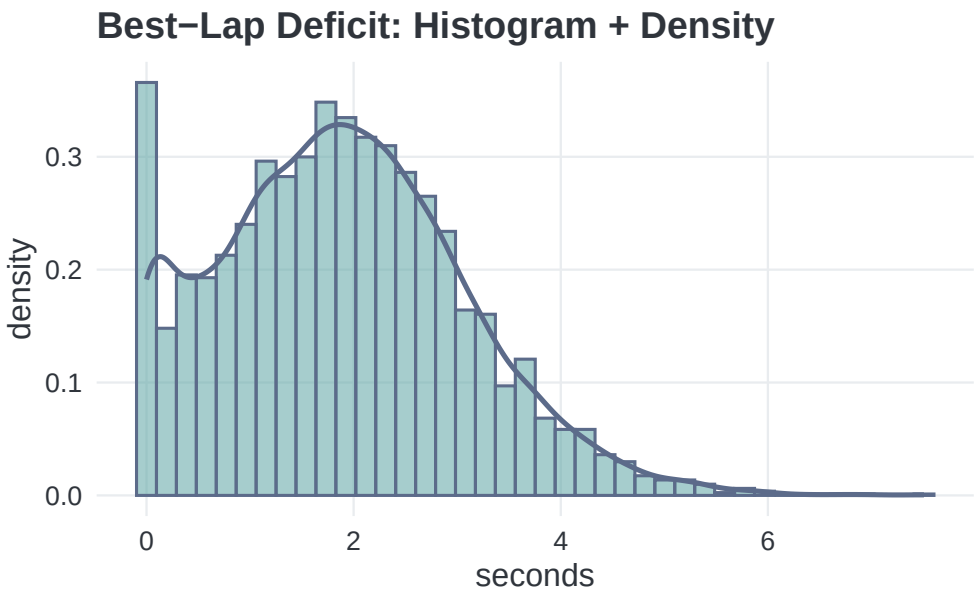

Figure 1: Distribution of best-lap deficits (2014–2024). Most drivers record lap times within 1–2 seconds of the fastest lap; a small fraction are several seconds slower due to incidents or slower pace.

The histogram above shows the distribution of **best-lap deficits** across all hybrid-era races. The overlaid smooth **density line** represents a *kernel density estimate*—a continuous approximation of how the values are distributed, computed by averaging nearby observations to reveal the overall shape. Most drivers are within a few seconds of the race's fastest lap, with a long right-hand tail reflecting incidents, tire wear, or traffic.

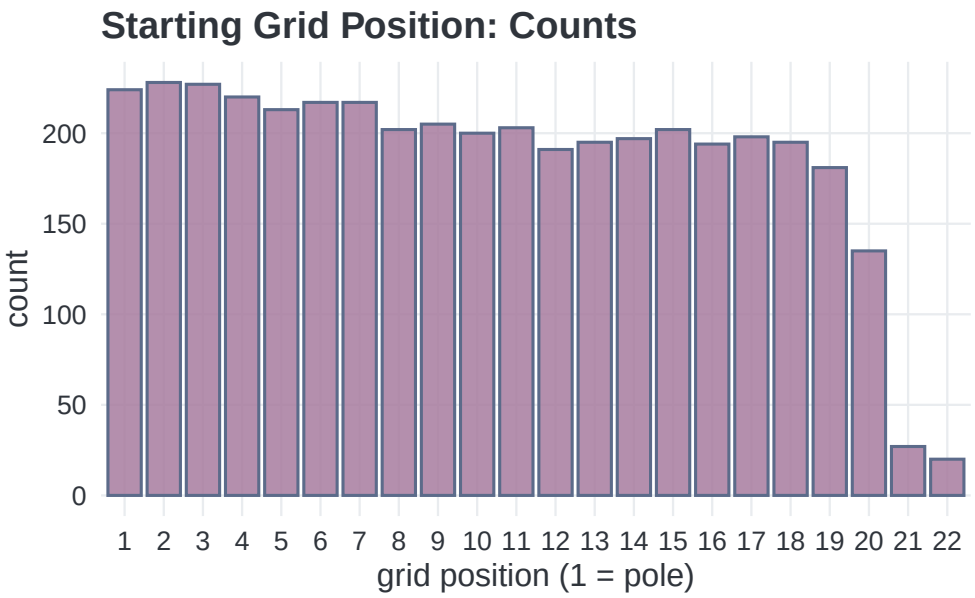

Figure 2: Starting grid positions used (2014–present). Discrete ranks produce vertical bands in later scatterplots.

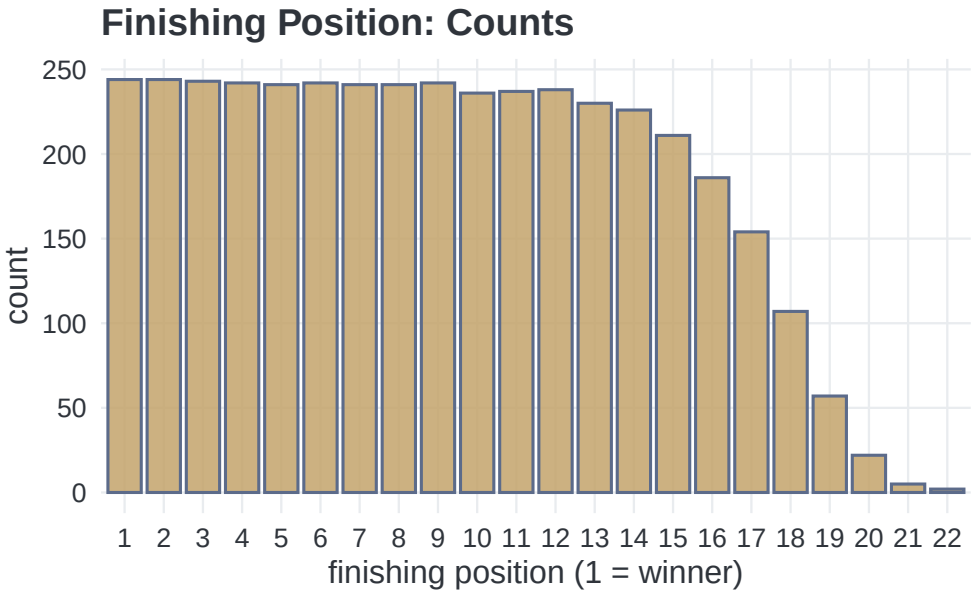

Figure 3: Finishing positions (1 = winner). Winners create a spike at 1; mass spreads gradually toward the midfield and backmarkers.

The other two variables, **starting grid position** and **finishing position**, are discrete rankings rather

than continuous values, so bar charts are used instead of boxplots.

The **starting-grid position** distribution is roughly uniform across the grid, as expected after many seasons of racing. Each position from pole (1) to the back (around 22) appears frequently, producing vertical bands in later scatter plots.

The **finishing-position** distribution naturally tapers toward the rear of the field, since every race has one winner and many midfield or back-marker finishers. The gradual decline in counts illustrates how retirements and lap-downs reduce the number of classified finishers. These counts also highlight that the difference between nearby ranks (e.g., finishing 4th vs. 6th) is not equivalent in frequency to differences farther down the order (e.g., 18th vs. 20th).

To prepare the dataset, I append lap-time results to compute each driver's best-lap deficit per race, then keep one row per driver-race combination containing finishing position, grid position, and lap deficit. The analysis includes all classified drivers from **2014–2024**, filtered to exclude non-classified results such as **DNFs, DSQs, and DNSs**.

These plots provide two useful perspectives. First, they show **scale**—finishing and grid positions are integer ranks, while pace deficit is measured in seconds. Second, they show **shape**—positions are count-based and skewed toward lower ranks (winners), while pace deficits form a smooth right-tailed distribution. Together, these visuals establish context for the regression models in the next section.

3 Quick Bivariate Analysis

Before fitting any models, I first visualize how each predictor relates to finishing position and check for possible relationships between the predictors themselves. This step does not attempt to prove causation—it simply explores the direction and strength of the relationships, overall spread, and whether any curvature or outliers are visible. I use scatterplots created with *ggplot2*(Wickham 2016) and dplyr(Wickham et al. 2023) to examine (a) **best-lap pace deficit versus finishing position**, (b) **starting-grid position versus finishing position**, and (c) **starting-grid position versus best-lap pace deficit.**

3.1 Best-lap deficit → finishing position

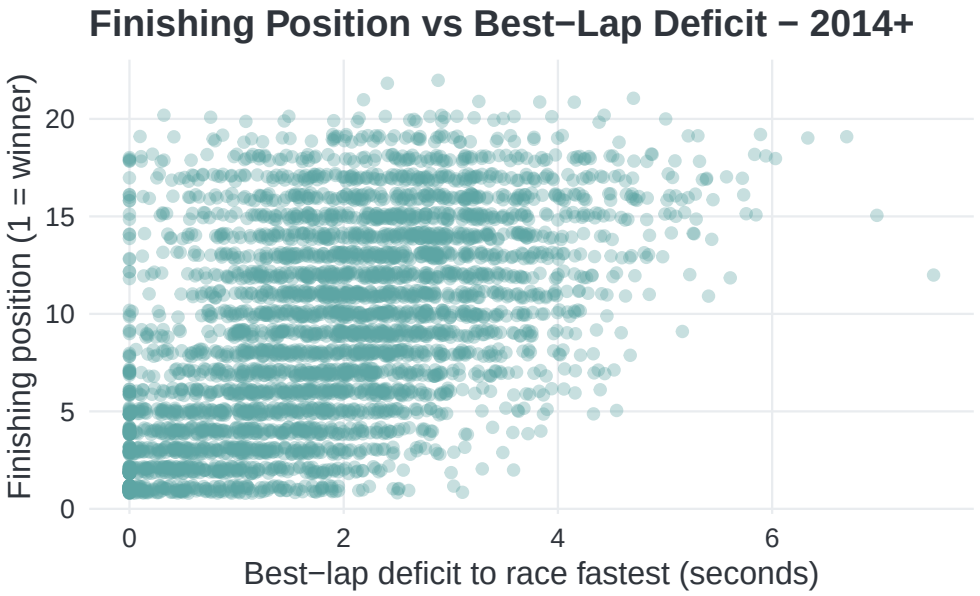

Figure 4: Finishing position versus best-lap pace deficit (2014–2024). Each point represents a driver–race result. Larger deficits are generally associated with lower finishing positions, though variation remains high.

The scatterplot shows finishing position versus best-lap pace deficit. A clear upward trend appears: drivers with larger best-lap deficits generally finish farther back in the field. The relationship is **meaningful but noisy**—many factors such as pit strategy, tire wear, or safety cars can blur the pattern. Near zero deficit, most finishes occur near the front, but not all; once the deficit exceeds about three to four seconds, front-running finishes are rare. Overall, the relationship between lap-time deficit and finishing rank is **moderately positive**, suggesting that slower pace usually, but not always, corresponds to worse finishing results.

3.2 Starting grid position → finishing position

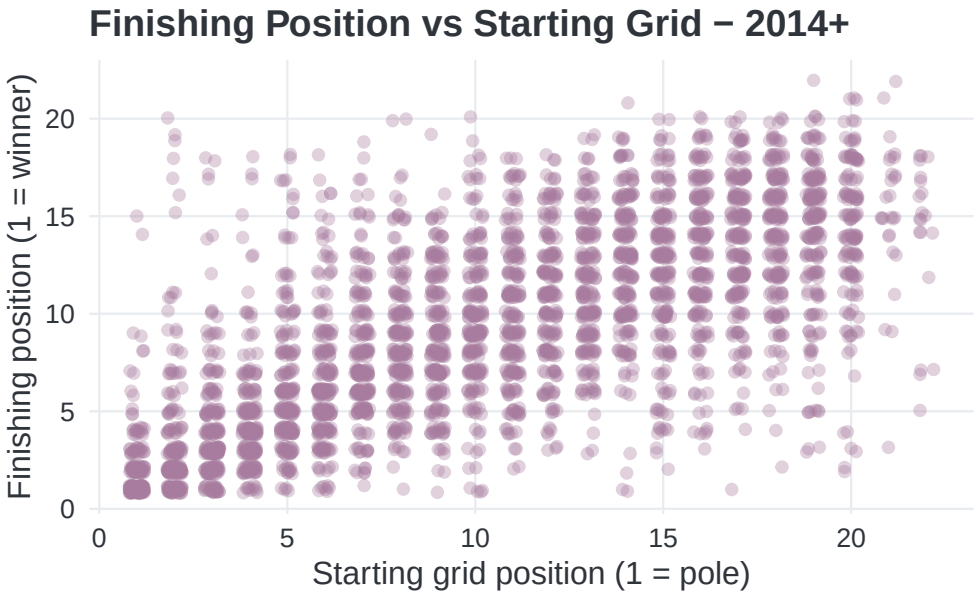

Figure 5: Finishing position versus starting-grid position (2014–2024). Finishing ranks rise steadily with deeper grid positions, showing a clear monotonic trend

This plot displays finishing position versus starting-grid position. A staircase pattern emerges: as starting-grid position increases (further back on the grid), finishing position also tends to worsen. The vertical stripes occur because grid ranks are discrete integers, but the overall upward slope shows a strong monotonic relationship. Compared with the pace-deficit plot, the spread here is narrower—especially among front rows—indicating that grid position explains finishing results more cleanly than lap-time deficit.

3.3 Are the predictors related?

Table 1: Spearman Rank correlation between starting grid and best-lap deficit (2014–2024).

Metric	Estimate
Spearman	0.51

Finally, I examine whether the two predictors—starting-grid position and best-lap pace deficit—are correlated. The scatterplot below shows a **moderate positive relationship**: drivers starting farther back tend to have slightly larger lap-time deficits, but the link is not one-to-one.

To quantify this association, I compute the **Spearman rank correlation coefficient ( )**, which measures the strength of a *monotonic* relationship between two ranked variables without assuming

linearity. A value of **= 0.51** indicates a moderate positive association: in general, drivers who start deeper on the grid also tend to record larger best-lap deficits, although there is substantial variation.

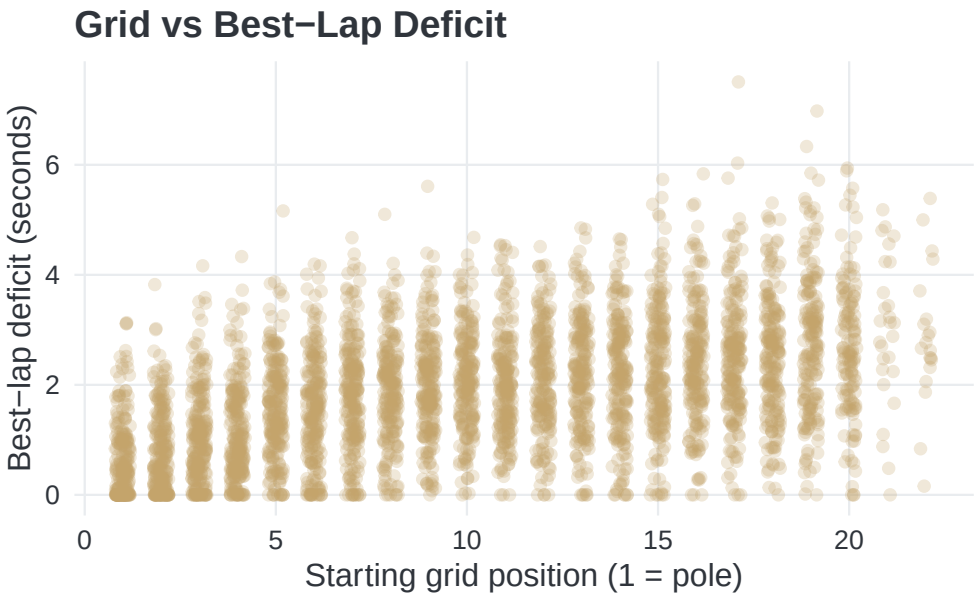

Figure 6: Starting-grid position versus best-lap pace deficit (2014–2024). Moderate positive association: cars starting farther back often have larger pace deficits, but with wide variability.

This exploratory check confirms that the predictors are related but not redundant—each contributes distinct information about finishing position. In later sections, I quantify their predictive strength using linear regression models and report goodness-of-fit measures including the coefficient of determination (R^2), adjusted R^2, and residual standard error.

4 Methods

To answer the question "How do a driver's best-lap pace deficit and starting-grid position each affect **finishing position**?", I fit **two separate one-predictor linear regressions** using data from the 2014-2024 engine era. The models are intentionally simple so that effects can be interpreted in plain units of positions and seconds.

4.1 Model A — Best-lap pace deficit predicting finishing position

Let FinishPos_i be driver i's finishing position (1 = winner) and Deficit_i the driver's best-lap pace deficit in seconds (their best race lap minus the race's fastest lap). I fit a simple linear model:

$$\text{FinishPos}_i = \beta_0 + \beta_1 \, \text{Deficit}_i + \varepsilon_i$$

Here, β_0 is the intercept—the expected finishing position when the deficit is zero (that is, when a driver's best lap matches the race's fastest lap).
β_1 is the **average change in finishing position per additional second of deficit** (for example, $\beta_1 = 1.4$ means that each extra second of pace deficit corresponds on average to finishing about 1.4 places farther back).
The error term ε_i captures race-specific noise not explained by the model, such as pit strategy, weather, or incidents.

I report the **coefficient of determination (R^2)**, the **adjusted R^2**, the **residual standard error (RSE)**, and the **slope estimate with a 95% confidence interval (CI)**.
- R^2 measures the proportion of variation in finishing position explained by the model.
- Adjusted R^2 penalizes unnecessary predictors; for single-predictor models it is almost identical to R^2.
- The **Residual SE** gives the model's typical miss, in the same units as the response (finishing positions).
- The **95% CI** around the slope represents the range of values consistent with the data, assuming approximately normal or large-sample t-based errors.

4.2 Model B — Starting grid position → Finishing position

Let FinishPos_i be the finishing position and Grid_i the starting-grid slot (1 = pole; higher numbers = farther back). The model is

$$\text{FinishPos}_i = \gamma_0 + \gamma_1 \, \text{Grid}_i + \eta_i$$

Here, γ_1 is the **average change in finishing position per one grid place farther back** (we expect a positive value: deeper grid positions usually lead to worse finishes).
γ_0 is the intercept, mainly anchoring the line near the observed range (e.g., predicted finish at grid 1–2).
As with Model A, I summarize R^2, adjusted R^2, RSE (in finishing-position units), and the slope with a 95% CI to describe effect size and uncertainty.

4.2.1 Model Assumptions

Both models are estimated with ordinary least squares (OLS) using `lm()` in R (R Core Team 2024) and reported with **knitr** (Xie 2024).
The analysis relies on the standard OLS assumptions:

1. **Linearity:** The average finishing position changes linearly with each predictor (pace deficit or grid slot).
2. **Homoscedasticity:** The spread of residuals is roughly constant across predictor values.
3. **Normality of residuals:** Residuals are approximately normally distributed, allowing valid confidence intervals and t-tests under large-sample conditions.
4. **Independence:** Each observation is treated as independent. In reality, finishing positions within a race are **not fully independent** (if one driver wins, no one else can). This dependence can slightly understate uncertainty, so confidence intervals should be viewed as approximate.

These assumptions are evaluated qualitatively using residual-diagnostic plots in the Results section. Because each model uses only one predictor, the focus is clarity: to quantify how much **one straightforward metric**—pace deficit or grid position—shifts finishing results, expressed as positions per second or positions per grid place, while acknowledging unmodeled race-level factors.

5 Results

We report results for two simple regressions on the 2014–present hybrid era:

- **Best-lap pace deficit predicting finishing position**
- **Starting-grid position predicting finishing position**

For each, I summarize the slope (effect size), how much variance is explained (R^2 and adjusted R^2), and the model's typical miss in positions (Residual Standard Error, RSE).
The **red line** in each scatter plot is the fitted regression line showing the average trend, while the **grey shaded band** is its **95% confidence interval (CI)** — the range of plausible mean predictions given sampling uncertainty.
A 95% CI means that, under the model's assumptions, if we refit this model many times on new samples, about 95% of the intervals would capture the true average slope.

5.1 Best-lap pace deficit predicting finishing position

According to the model, the average change in finishing place per +1 second of pace deficit is **2.349** positions (95% CI [**2.238, 2.461**]).
The fit explains **0.29** of the variation in finishing order (adjusted R^2 = **0.29**), with a Residual SE of **4.33** positions (variance **18.71**).

In plain terms, if the line predicts P9, actual results often fall about **±4.3** positions around that.
This matches the earlier exploratory plots: pace deficit is a meaningful but noisy predictor — it captures general trends but not race-to-race volatility.

5.2 Starting grid position → finishing position

Here the slope is **0.655** positions per grid place farther back (95% CI [**0.637, 0.673**]). This model explains **0.554** of the variation in finishing position (adjusted R^2 = **0.554**), with a Residual SE of **3.41** positions (variance **11.66**).

The tighter error and higher R^2 reflect that **grid position** is a stronger, steadier predictor of finishing order than pace deficit.

5.2.1 Model summaries

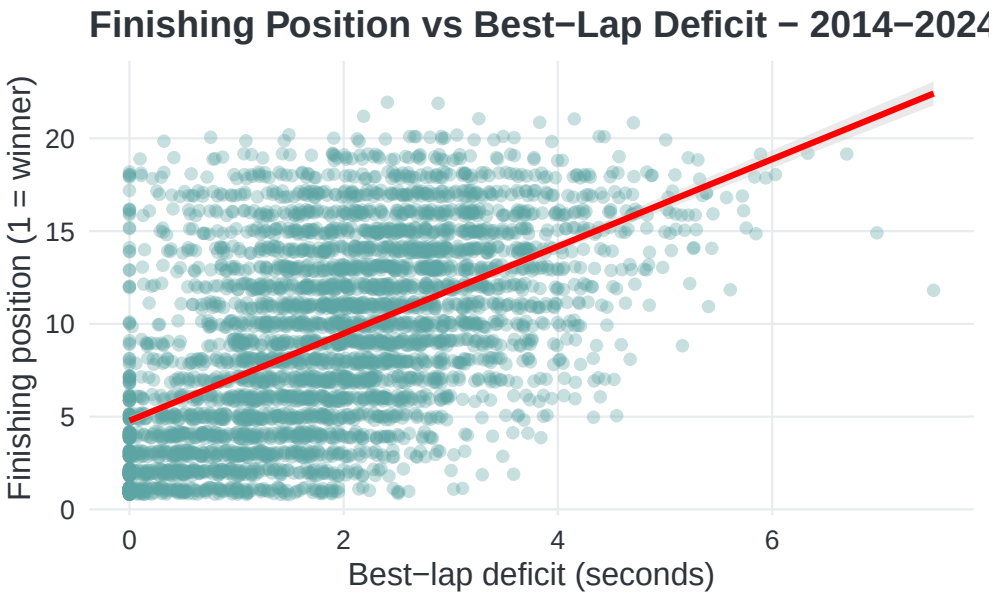

Figure 7: Finishing position vs best-lap deficit with linear fit. The red line shows the fitted average relationship; the grey band shows its 95% confidence interval. Upward slope quantifies positions lost per second of deficit; variability reflects race dynamics.

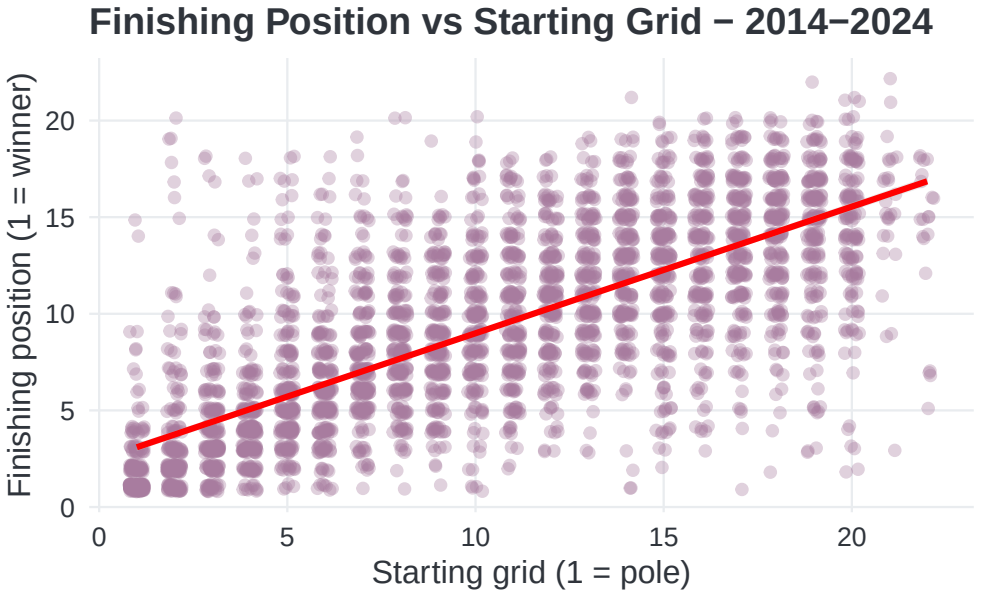

Figure 8: Finishing position vs starting grid with linear fit. The red line shows the fitted mean; the grey band shows the 95% CI for that mean. Grid position has a clear, steady upward trend, indicating that starting farther back usually means finishing farther back.

Table 2: Model summaries: number of observations (n), R^2, adjusted R^2, residual standard error (in positions), and residual variance.

Model	n	r2	adj_r2	rse_pos	rse_sq
A: Best-lap deficit → finish	4175	0.290	0.290	4.325	18.708
B: Grid → finish	4091	0.554	0.554	3.415	11.662

Table 3: Slope estimates (effect on finishing position) with 95% confidence intervals.

Model	Term	Estimate	Std. Error	CI low	CI high
A: Best-lap deficit predicting finish	Best-lap deficit (seconds)	2.349	0.057	2.238	2.461
B: Starting grid predicting finish	Starting-grid position	0.655	0.009	0.637	0.673

5.2.2 Residual diagnostics

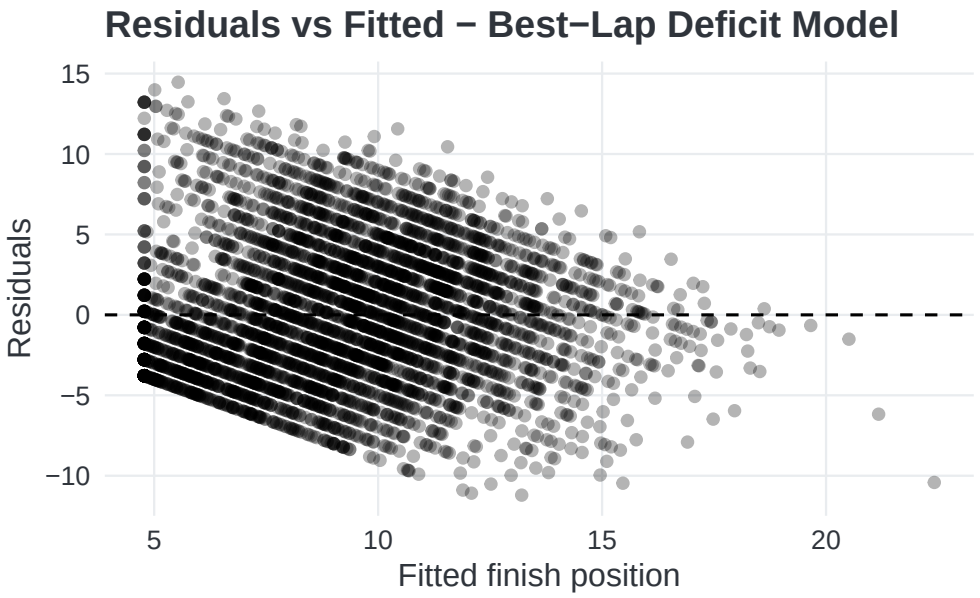

Figure 9: Residuals vs fitted values for the best-lap deficit model. The band is centered near zero, with mild structure at the extremes reflecting the bounded nature of finishing ranks (1 to 22).

Both residual plots show the expected **striping** pattern because finishing positions are **integer values**, not continuous ones. This discrete structure naturally creates horizontal bands but does not indicate a model problem. The points remain centered around zero without a widening funnel, so a linear mean with roughly constant variance appears reasonable. The pace model shows slightly more spread and mild curvature, while the grid model's residuals are more uniform — consistent with grid position being a cleaner predictor.

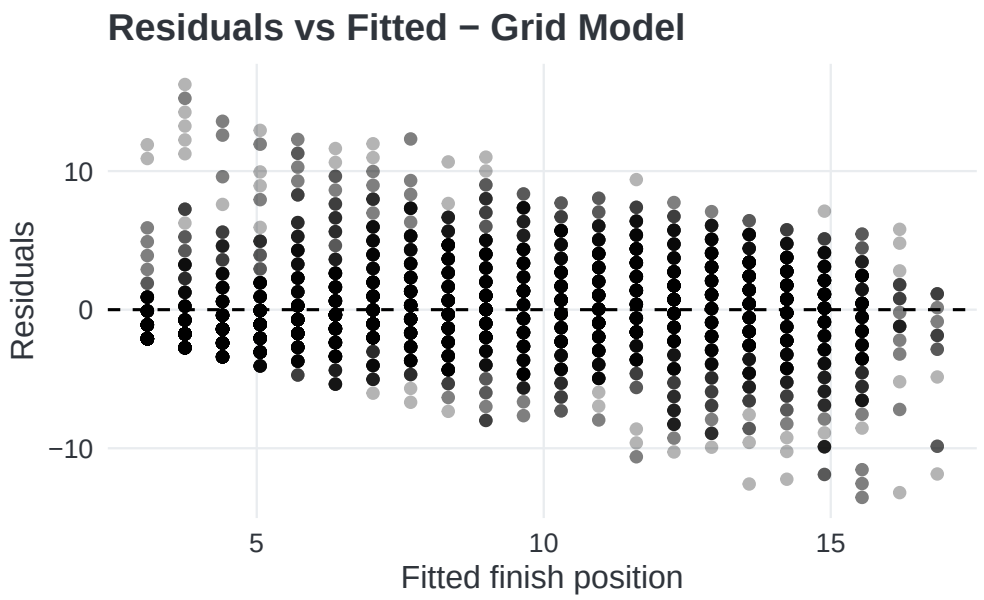

Figure 10: Residuals vs fitted values for the grid model. The band is tighter and more horizontal, showing roughly constant variance and no strong funneling.

5.2.3 Q–Q plot check for normality

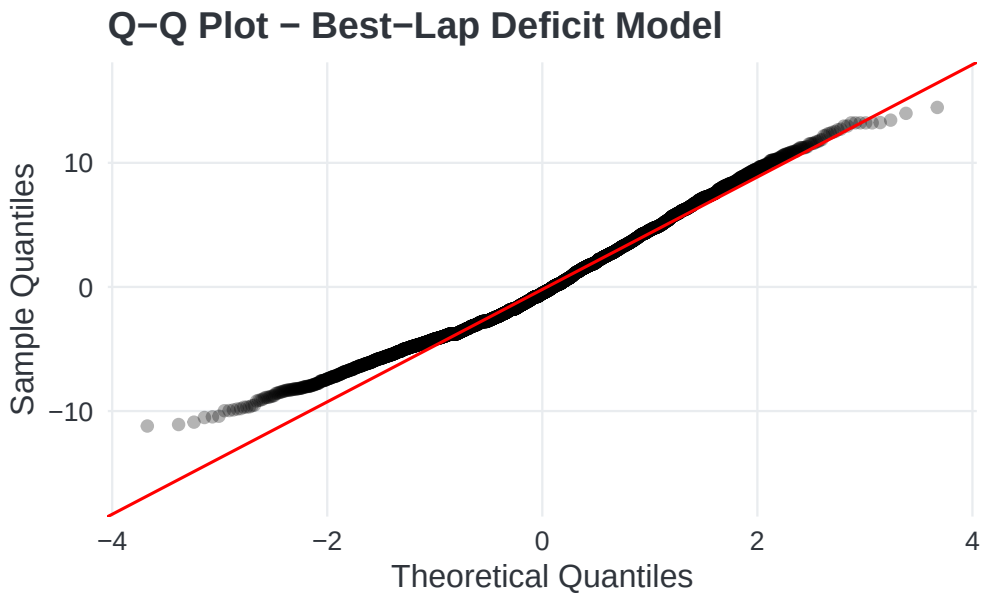

Figure 11: Q–Q plot of residuals for the best-lap deficit model. Points mostly follow the diagonal, suggesting residuals are approximately normal with mild tails.

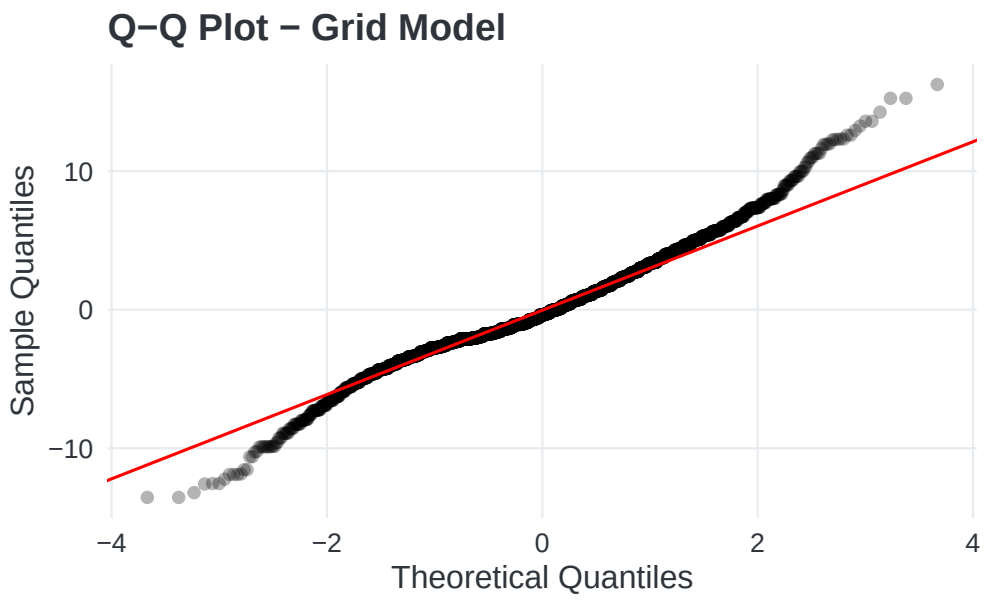

Figure 12: Q–Q plot of residuals for the grid model. Points follow the reference line closely, consistent with approximate normal residuals.

Both Q–Q plots show residuals that follow the reference line closely through the central range, indicating that the normality assumption for the error terms is broadly reasonable. Minor deviations appear at the tails—especially in the best-lap-deficit model—where a few races produced unusually large residuals. These slight tail departures are expected in competitive sports data and are unlikely to affect inference given the large sample size. Overall, the residuals display approximate normality, supporting the use of t-based confidence intervals in both models.

6 Discussion

This study used two simple linear regressions to examine how a driver's on-day pace (measured by best-lap deficit) and starting-grid position relate to finishing position in modern Formula 1. The results are clear: grid position is the stronger and steadier predictor, while best-lap deficit still carries meaningful information but with greater variability—consistent with the earlier bivariate and residual checks.

These models show that both pace and track position influence finishing outcomes, but in different ways. Grid position provides a structural advantage at the start, while the best-lap deficit—based on a driver's single fastest lap—offers only a snapshot of potential race pace rather than sustained performance. This helps explain why the pace–finish relationship is noisier. Quantitatively, the grid-position model explains about 55% of the variation in finishing order, while the pace-deficit model explains about 29%, showing how much stronger starting position is as a predictor of race results.

There are limitations. The analysis considers only one predictor at a time and does not adjust for track layout, safety cars, tire or driver strength, or weather. Finishing position is also an integer and

bounded variable, so small nonlinearities are expected. Because best-lap deficit represents only one lap's performance, it may not fully reflect race-long speed consistency.

Future work could extend the models to include track or season effects, team or driver indicators, or per-lap average pace metrics. Still, the practical message is straightforward: starting closer to the front reliably improves finishing position, and even being one second slower on the day's fastest lap tends to correspond to finishing several places farther back—quantified here in plain units of positions per grid place and positions per second.

References

F1DB contributors. 2025. "F1DB: Formula 1 Database (CSV Release and Documentation)." GitHub repository. https://github.com/f1db/f1db.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://doi.org/10.1007/978-3-319-24277-4.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Jim Hester, and Romain François. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.