In January of 2024, Saudi Arabia hosted the Dakar Rally, an epic off-road endurance event spanning thousands of kilometers across the rugged terrain. With drivers navigating deserts, mountains, and dunes, this race poses formidable challenges for all sorts of vehicles, such as bikes, cars, and trucks. Our focus will be on exploring the biker’s ranking based off the racers time statistics and driver information.

1. Given the data, fit a linear model that uses Hours, Variation Hours, Penalty Hours, Penalty Minutes to predict driver’s ranking after the sixth stage has been completed in the rally.

A screenshot of a computer

Description automatically generatedA table with numbers and symbols

Description automatically generated

*\*Model has a F-statistic: 101.9 on 4 and 117 DF\**

1. Record the R-squared of the model. Show and explain work.

# R2 = 1 - SSE / SST

SSE = 33737

SST = (35395+50998+22590+8590+33737)

(R2 = 1 - (SSE / SST)) =  **0.7770339**

b. Interpret the Coefficient for Hours and Penalty Hours in the context of the problem.

For hours, every hour that is added to drivers overall time will increase overall ranking by 0.25667, given all other variables remain constant. For penalty, for every hour added to a drivers penalty time, will increase their overall ranking by 0.69841.

2. Now let make a model that predicts the probability of a driver ranking in the top five. We will be using Hours, Variation Hours, Penalty Hours, Penalty Minutes, along with Stages.

1. Fit the model, and record R-code below.

dakarRally\_bikes\_data %>%

mutate(top5 = ifelse(Rank <= 5, 1, 0)) %>%

select(top5, Hours, Variation\_Hours, Penalty\_Hours, Stage) %>%

glm(top5 ~ Hours + Variation\_Hours + Penalty\_Hours + Stage,

data = ., family = binomial)

1. How does a model where we only use Hours, Variation Hours, and Penalty Hours, along with Stages to predict top5 compared to the model in part a. Show work and Rcode.

modRed <- glm(topFive ~ Hours + Variation\_Hours + Penalty\_Hours + Stage,

data = dakarRally\_bikes\_data, family = binomial)

modFull <- glm(topFive ~ Hours + Variation\_Hours + Penalty\_Hours + Penalty\_Minutes + Stage,

data = dakarRally\_bikes\_data, family = binomial)

anova(modRed, modFull, test = "Chisq")

The Full model proves to be more significant with a **p = 0.000152**

3. Since this race spans 13 stages, winning a single stage doesn't guarantee overall victory. Provide an R summary output listing the top 5 drivers by their average ranking, including their name, country, average ranking, average hours, variation hours, penalty hours, and the number of stages raced, ensuring completion of all stages. Also, record the top driver and their average ranking, along with the corresponding R-code. \*Remember to also include the minutes and seconds variables as well for a more accurate outcome\*

dakarRally\_bikes\_data %>%

group\_by(Country, Driver) %>%

summarise(count = n(),

average\_Rank = mean(Rank),

Mean\_Hours\_Driven = ((mean(Hours)\*60 + mean(Minutes) + (mean(Seconds)/60)) / 60),

Mean\_Hours\_Penalty = (mean(Penalty\_Hours)\*60 + mean(Penalty\_Minutes) + (mean(Penalty\_Seconds)/60)) / 60,

Mean\_Hours\_Variation = (mean(Variation\_Hours)\*60 + mean(Variation\_Minutes) + (mean(Variation\_Seconds)/60)) / 60) %>%

filter((count > 1)) %>%

arrange(average\_Rank) %>%

head()