**Predictive Fuel Consumption Regression**

**Methods and Data Used**

Based on the 39 observations, we developed a multivariate regression model with fuel consumption per 10 km as the dependent variable. We assume that variables that we classify as mainly associated with driving behaviour, such as offences made (speeding, braking, cornering, acceleration) and running idle are influenced more by drivers’ choices.

The multiple regression is useful to assess the relative influence of individual variables when the others are constant to find out which has the most impact. However, in real-life situations, other variables do have some influence on each other one way or another, for example a change in speed affects total duration of the trips and the fuel consumption. Since variables change simultaneously, multivariate regression is appropriate since it assesses the independent effect of one variable controlled for other variables.

1. **Dependent Variable:** Total fuel consumption per month
2. **Independent Variables:**
3. Total distance covered per month
4. Total duration when the ignition was On that was not idling (speed >0 recorded during each trip in Detailed trip report)
5. Idling time (speed = 0 from ignition On until Off)
6. Speeding
7. Cornering
8. Acceleration
9. Braking
10. Total number of offences *= speeding + cornering + acceleration + braking*

**3. Methods and Concepts Used:**

1. **Best Subset Selection Method:** We first performed Best Subset Selection to identify promising predictor variables. However, this approach generates a list of models for different variable combinations, requiring further analysis to choose the most suitable model.
2. **Validation Set Approach and 5-fold Cross Validation:** To assess prediction errors, we estimated Mean Square Error (MSE) using both indirect and direct methods. Indirect estimation involved calculating RSS, adjusted R-squared, Cp, and BIC to select the best model. For direct estimation, we used Validation Set Approach and 5-fold Cross Validation to train and test models on different data subsets, ultimately selecting the model with the lowest average MSE.
3. **Lasso and Ridge Regression:** We applied ridge and lasso regression to regularize coefficient estimates towards zero, aiming to prevent overfitting. Model performance was compared using adjusted R-squared to determine the optimal approach.

**Regression Result Analysis and Recommendations**

The derived multilinear equation is as follows:

**Total Fuel Consumed** = 0.07373039 \* total distance + 0.10376002 \* total duration - 1.92080252 \* speeding - 6.70395044 \* acceleration - 0.53071289 \* cornering - 0.10097106 \* total idling

The coefficients in the equation offer insights into the relationships between independent variables and fuel consumption. A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship. For example, a longer duration on the road (total duration) corresponds to higher fuel consumption. Additionally, driving behaviour variables such as speeding, acceleration, braking, and cornering significantly impact fuel consumption.

To provide user interaction with the model, we developed a Shiny application using the R Shiny library.

**Conclusion and Future Work**

In conclusion, our multivariate regression model offers valuable insights into fuel consumption patterns driven by driving behaviour. While our analysis provides statistically significant results, it's important to note that the coefficients themselves don't quantify the magnitude of the relationships. Real-world complexities involve interactions between variables that may influence the model's outcomes.

For future work, we recommend exploring alternative modelling methods to further refine predictions and recommendations. While our current model is informative, other techniques can offer complementary insights. Some possibilities include:

1. **Machine Learning Algorithms:** Investigate machine learning algorithms such as Random Forest, Gradient Boosting, or Neural Networks. These methods could capture non-linear relationships and interactions that might be missed by traditional regression models.
2. **Time Series Analysis:** Consider incorporating time series analysis techniques if the dataset expands to include temporal information. Time-based patterns and trends could impact fuel consumption, making this analysis valuable for long-term prediction.
3. **Clustering and Segmentation:** Group similar vehicles or driving behaviours into clusters. This could lead to more targeted strategies for optimizing fuel efficiency based on distinct driving profiles.
4. **Feature Engineering:** Explore creating new features based on domain knowledge. For instance, deriving a "smoothness" metric from acceleration and braking patterns might better capture the impact on fuel consumption.
5. **External Data:** Incorporate external data sources such as weather conditions, road infrastructure, or traffic data. These factors could enhance the model's accuracy by accounting for additional influences on fuel consumption.

By considering these alternative modelling methods, we can gain a comprehensive understanding of the fuel consumption dynamics and potentially uncover insights that align with specific scenarios.

For future endeavours, addressing limitations such as data quality and external validity would be pivotal. Gathering more data over time could significantly enhance the accuracy and robustness of the model, enabling more confident predictions and actionable recommendations.