Analyzing Driver Fatalities in Great Britain 1969-84: A Resampling-Based Analysis

Vishal Nair - 1740105

02/09/2025

Abstract

Road safety is a crucial public health concern, with vehicle fatalities being a significant source of mortality worldwide. This study analyzes the "Seatbelts" dataset, which records monthly traffic accident statistics in Great Britain from January 1969 to December 1984, to assess the factors influencing driver fatalities. The dataset includes key variables such as kilometers driven, petrol prices, and the introduction of seatbelt laws in 1983.

Using resampling techniques—specifically permutation tests, non parametric-bootstrap we assess the impact of seasonality, seatbelt legislation, and economic factors on road fatalities. Our findings indicate that seasonal patterns affect accident rates, the seatbelt law significantly reduced driver fatalities, and petrol prices influence road safety trends. By employing rigorous non-parametric techniques, this study provides robust statistical validation of key road safety policies and economic influences on traffic fatalities.

0.1 Introduction and Dataset Description

0.1.1 Background and Motivation

Road traffic accidents remain a leading cause of mortality globally, prompting governments to implement various safety regulations to mitigate risks. The introduction of mandatory seatbelt laws in Great Britain in 1983 marked a significant policy intervention aimed at reducing driver fatalities. However, the effectiveness of seatbelt laws, the influence of fuel prices, and seasonal variations in accident rates require empirical validation.

To investigate these issues, we analyze the Seatbelts dataset, which spans 16 years (1969–1984) and contains monthly data on traffic fatalities and economic factors. The dataset, originally compiled by the Department of Transport in Great Britain, enables an extensive statistical analysis of driver fatalities and the external influences shaping road safety.

0.1.2 Dataset Overview

The dataset consists of 192 monthly observations, covering the period from January 1969 to December 1984. Key variables include:

- **DriversKilled**: The number of drivers killed in road accidents each month.
- **kms**: The total distance driven in kilometers (thousands).
- **PetrolPrice**: The price of petrol per unit.
- **law**: A binary variable (0 before 1983, 1 after 1983) indicating whether the seatbelt law was in effect.
- **front and rear**: Number of front-seat and rear-seat passenger fatalities.
- **VanKilled**: The number of van occupant fatalities.

In the data pre-processing step, we enhance the dataset by adding Year and Month columns to facilitate time-based analysis. This dataset allows us to examine critical road safety trends, policy effectiveness, and economic influences on driver fatalities over time.

0.2 Problem Statement, Objectives, and Hypotheses

Road safety remains a critical public health concern, with driver fatalities serving as a key metric for evaluating the effectiveness of traffic regulations. The introduction of seatbelt laws aimed to reduce fatalities, yet their true impact requires rigorous statistical validation. Additionally, external factors such as seasonal variations, driving distances, and fuel prices may influence accident rates. This study leverages statistical resampling techniques to assess these relationships using the *Seatbelts* dataset.

0.2.1 Objectives

This analysis focuses on four key objectives:

- **Assess the impact of seatbelt laws on driver fatalities**: Determine whether the introduction of mandatory seatbelt laws in 1983 significantly reduced fatalities.
- **Examine seasonal patterns in driver fatalities**: Identify whether fatalities exhibit systematic monthly variations.
- **Compare front and rear-seat passenger fatalities**: Investigate whether fatalities differ significantly between front-seat and rear-seat passengers.
- **Evaluate the relationship between driver fatalities and external factors**: Analyze the effect of kilometers driven and petrol prices on road fatalities.

By applying permutation tests and bootstrap confidence intervals, we aim to provide robust statistical evidence supporting or refuting these claims.

0.3 Hypotheses

To rigorously analyze the impact of seatbelt laws and other factors on driver fatalities, we define the following hypotheses:

- **H1: Impact of Seatbelt Laws on Driver Fatalities**
  - **Null Hypothesis (H_0)**: The introduction of seatbelt laws had no significant effect on the number of driver fatalities.
  - **Alternative Hypothesis (H_A)**: The introduction of seatbelt laws significantly reduced the number of driver fatalities.
- **H2: Seasonal Patterns in Driver Fatalities**
  - **Null Hypothesis (H_0)**: Driver fatalities are uniformly distributed across months, showing no seasonal variation.
  - **Alternative Hypothesis (H_A)**: Driver fatalities exhibit significant seasonal variation.
- **H3: Comparison of Front and Rear Passenger Fatalities**
  - **Null Hypothesis (H_0)**: There is no significant difference between front-seat and rear-seat passenger fatalities.
  - **Alternative Hypothesis (H_A)**: There is a significant difference between front-seat and rear-seat passenger fatalities.
- **H4: Relationship Between Driver Fatalities and External Factors**
  - **Null Hypothesis (H_0)**: The number of kilometers driven and petrol prices have no significant effect on driver fatalities.
  - **Alternative Hypothesis (H_A)**: The number of kilometers driven and petrol prices significantly affect driver fatalities.

These hypotheses will be tested using permutation tests, two-way ANOVA, and bootstrap confidence intervals to ensure statistical robustness.

0.3.1 Prior Information Available

Several prior studies suggest that mandatory seatbelt laws lead to a reduction in road fatalities. For example:

- Research in Great Britain demonstrates that seatbelt mandates have significantly reduced fatalities. After the seatbelt law was introduced in 1983, there was a 29% reduction in fatal injuries and a 30% reduction in serious injuries among front-seat passengers.
- Seasonal trends in Great Britain reveal an increase in road accidents during autumn and winter due to challenging driving conditions.
- Economic studies suggest that higher petrol prices lead to reduced travel, potentially decreasing accident rates.
- Studies show rear-seat passengers have a 25–75% lower risk of fatality compared to front-seat occupants in most crash scenarios.

0.3.2 Exploratory Data Analysis

To gain preliminary insights into the dataset, we conducted exploratory data analysis (EDA). This step helps to understand the distribution of key variables, identify patterns, and assess relationships between different factors.

Effect of Seatbelt Law on Driver Fatalities

To assess the impact of the seatbelt law, we compared driver fatalities before and after its introduction. Figure 1 illustrates a noticeable reduction in fatalities post-law implementation.

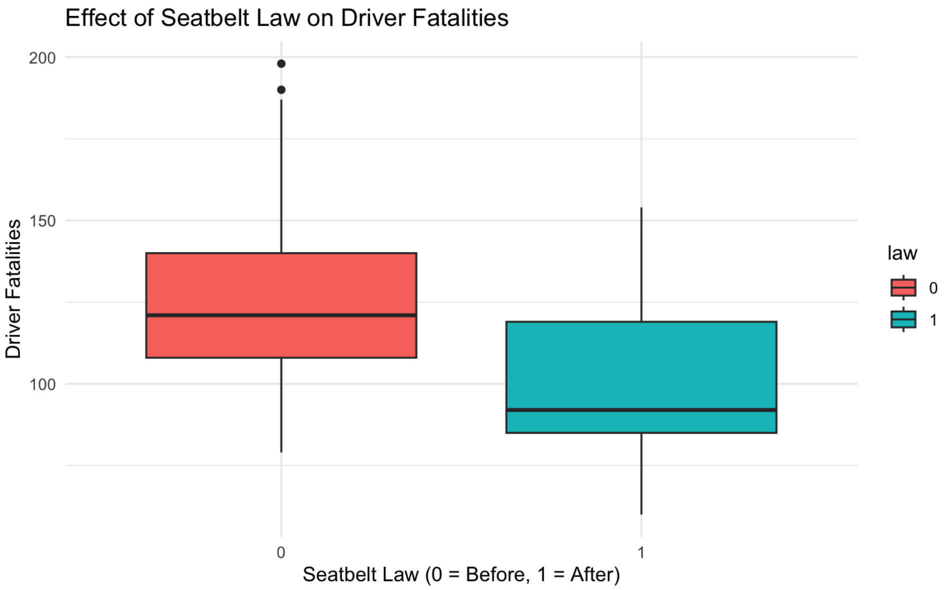

Figure 1: Comparison of driver fatalities before and after the seatbelt law.

Seasonality in Driver Fatalities

Figure 2 presents monthly trends in driver fatalities. A clear seasonal pattern is observed, with fatalities peaking in winter months. This aligns with prior studies suggesting increased accident risks due to adverse weather conditions.

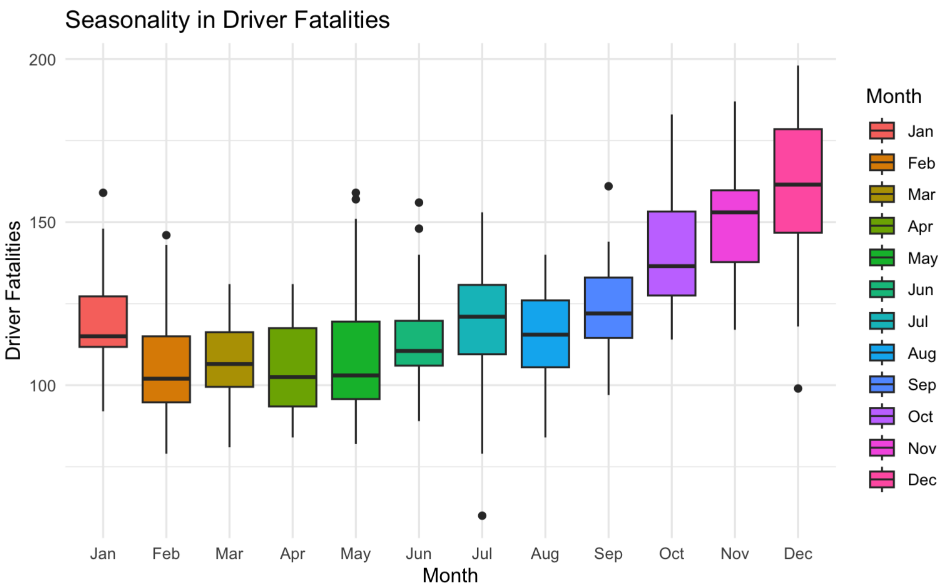

Figure 2: Monthly trends in driver fatalities, indicating seasonality.

Front vs. Rear Passenger Fatalities

A significant difference was observed between front and rear passenger fatalities (Figure 3).

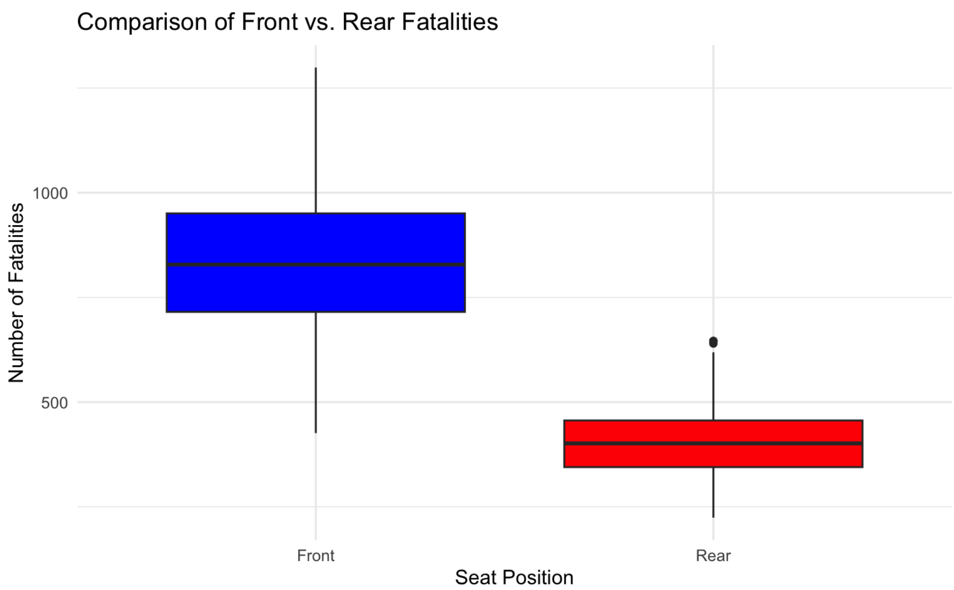

Figure 3: Comparison of fatalities for front and rear-seat passengers.

Correlation Analysis

A correlation heatmap (Figure 4) was generated to examine relationships between key variables. Petrol price shows a negative correlation with driver fatalities, while kilometers driven has a weaker negative association. These trends indicate that economic factors such as fuel prices influence road fatalities.

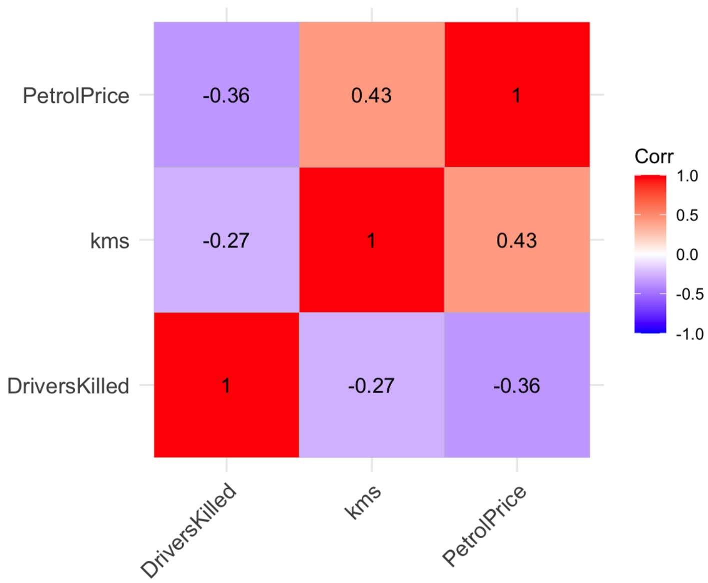

Figure 4: Correlation matrix of driver fatalities, kilometers driven, and petrol price.

These initial observations provide strong motivation for the statistical analyses that follow.

0.4 Statistical Analysis Using Resampling Techniques

To analyze the impact of seatbelt laws and other factors on driver fatalities, we employed resampling techniques, including **permutation tests**, **two-way ANOVA**, and **non-parametric bootstrap methods**. These techniques ensure that our findings are robust without relying on strict parametric assumptions.

Additionally, we tested the normality assumption where necessary using the **Shapiro-Wilk test**. This section details the statistical methods used in our analysis.

0.4.1 Two-Way ANOVA

A two-way analysis of variance (ANOVA) was conducted to examine the effects of **seatbelt laws** and **seasonality (month)** on driver fatalities. ANOVA tests whether mean differences exist across categorical groups by comparing within-group and between-group variance.

- **Factors**: The independent variables in the ANOVA model were:
  - **Seatbelt Law (law)**: A binary categorical variable indicating whether the seatbelt law was in effect.
  - **Month**: A categorical variable representing the month of the year, used to detect seasonal variations in driver fatalities.
- **Interaction Term**: We also tested for an interaction effect between the seatbelt law and month to determine whether the effect of the law varied across different months.
- **Assumptions**: Since `DriversKilled` is a count variable, the normality assumption does not strictly apply. Therefore, to ensure the validity of ANOVA results, we performed a **permutation test on the F-statistics** by randomly shuffling the response variable while keeping the categorical predictors intact.

0.4.2 Permutation Tests

Permutation tests are non-parametric methods used to assess statistical significance by reshuffling data points and recalculating test statistics under the null hypothesis. These tests were employed in two key areas:

- **Permutation Test for Two-Way ANOVA F-Statistic:** To verify the robustness of ANOVA results, we performed a permutation test on the F-statistics. By randomly shuffling the response variable (`DriversKilled`) and recomputing the F-values, we obtained an empirical p-value to assess significance.
- **Comparison of Front and Rear Passenger Fatalities:** We applied the **Shapiro-Wilk normality test** to determine whether the differences followed a normal distribution and got W = 0.97417, p-value = 0.001281. The small p-value (<0.05) suggests that the differences between front and rear fatalities deviate significantly from normality. Since the assumption of normality does not hold, a parametric paired t-test would not be appropriate. Instead, we performed a **permutation test** to assess whether the observed difference in fatalities between front and rear passengers was statistically significant.We used a permutation test by randomly redistributing fatality labels and comparing the observed difference to the permuted differences.

0.4.3 Multiple Linear Regression and Non-Parametric Bootstrap for Uncertainty Estimation

To quantify the relationship between driver fatalities and external factors, we constructed a multiple linear regression model using kilometers driven (`kms`) and fuel prices (`PetrolPrice`) as predictors. The model was specified as:

$$DriversKilled = \beta_0 + \beta_1 \times kms + \beta_2 \times PetrolPrice + \epsilon \tag{1}$$

where:

- **DriversKilled** represents the number of driver fatalities in a given month.
- β_0 is the intercept term, representing the expected number of driver fatalities when all predictors (`kms` and `PetrolPrice`) are zero.

- β_1 is the coefficient associated with `kms`, which represents the change in driver fatalities per unit increase in kilometers driven.
- β_2 is the coefficient associated with `PetrolPrice`, which captures the effect of petrol price fluctuations on driver fatalities.
- ϵ is the error term, accounting for variability in driver fatalities not explained by the independent variables.

Regression models typically assume normality and homoscedasticity of residuals. However, these assumptions may not always hold in real-world datasets. To mitigate potential violations and ensure robust inference, we applied a **non-parametric bootstrap method** to estimate the uncertainty in the regression coefficients.

Bootstrap Confidence Intervals for Regression Coefficients The bootstrap is a resampling technique that repeatedly draws samples with replacement from the observed data to approximate the sampling distribution of a statistic. We used it to estimate confidence intervals for the regression coefficients, allowing us to assess their variability and significance.

- **Bootstrap Confidence Intervals:** We performed non-parametric bootstrap resampling to estimate the variability of the regression coefficients in the multiple linear regression model.
- **Bootstrap-t Confidence Intervals:** Instead of relying on the percentile bootstrap method, we employed the **bootstrap-t method**, which accounts for variability in standard errors by computing t-statistics in each bootstrap iteration.

These resampling techniques ensure that our findings are statistically robust and allows for a comprehensive analysis of the factors influencing driver fatalities.

0.5 Results Analysis

This section presents the results of the statistical analysis performed. The findings are interpreted in the context of prior information and exploratory data analysis (EDA), ensuring robustness through resampling methods.

0.5.1 Effect of Seatbelt Law and Seasonality on Driver Fatalities

The results of the two-way ANOVA conducted to evaluate the impact of the seatbelt law and seasonal patterns on driver indicate:

- A significant main effect of **month** on driver fatalities ($F(11, 168) = 16.007, p < 0.05$), suggesting a seasonal pattern in the data.
- A significant main effect of **the seatbelt law** ($F(1, 168) = 43.011, p < 0.05$), indicating a reduction in fatalities following the introduction of the seatbelt law.
- The interaction effect between month and the seatbelt law was not significant ($F(11, 168) = 0.664, p = 0.771$), suggesting that the seasonal pattern remained consistent before and after the law.

To validate these results, we performed a permutation test on the F-statistics. The empirical p-values from 10,000 permutations were:

- **Month:** $p = 0$
- **Seatbelt Law:** $p = 0$
- **Month-Law Interaction:** $p = 0.7676$

These results confirm the robustness of the ANOVA findings, reinforcing that seasonality and the seatbelt law significantly influenced driver fatalities. The distribution of permuted F-statistics is visualized in Figure 5.

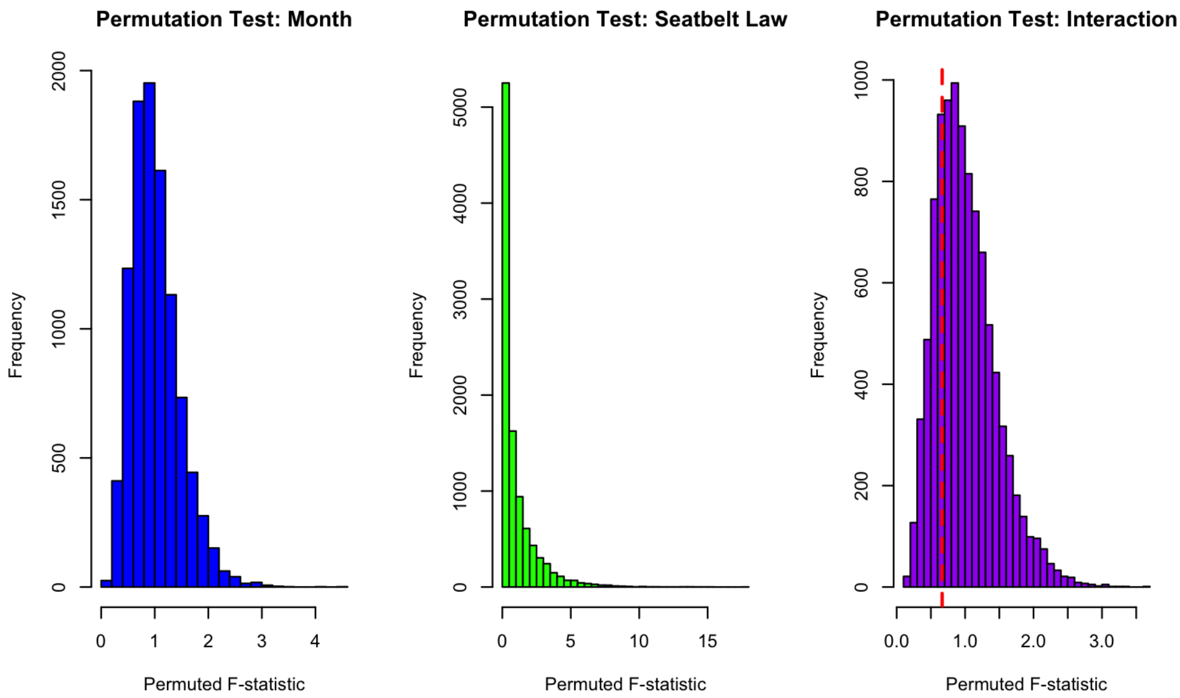

Figure 5: Permutation Test for ANOVA F-Statistics

0.5.2 Comparison of Front and Rear Passenger Fatalities

The permutation test was conducted to assess whether there was a significant difference between front and rear passenger fatalities. The observed mean difference in fatalities was 436.01, and the permutation test yielded a p-value of $p = 0$, confirming a significant difference. The histogram of permuted mean differences is presented in Figure 6. This finding suggests that front-seat passengers were significantly more likely to suffer fatalities compared to rear-seat passengers. The higher fatality rate for front-seat occupants aligns with prior knowledge and vehicle safety research, which shows that front passengers are more exposed to direct impact in collisions, particularly in head-on crashes.

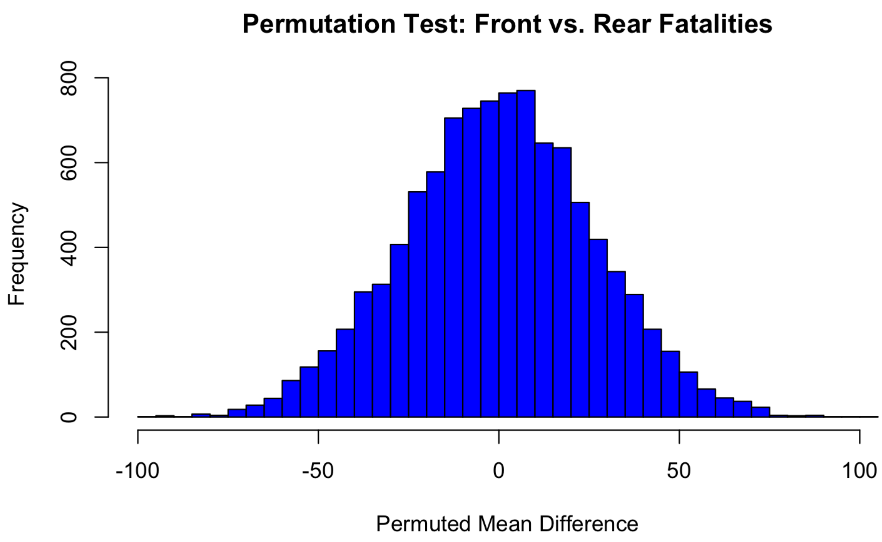

Figure 6: Permutation Test: Front vs. Rear Fatalities

0.5.3 Effect of Distance Driven and Fuel Prices on Driver Fatalities

The predictors in the regression model were transformed to improve interpretability and ensure numerical stability. The variable `mmc` represents the number of kilometers driven per month, centered around its mean, while `pp` is the standardized petrol price, obtained by subtracting the mean and dividing by the standard deviation. These transformations allow for a more intuitive interpretation of the regression coefficients. The estimated model is:

$$DriversKilled = \beta_0 + \beta_1 \times mmc + \beta_2 \times pp + \epsilon \tag{2}$$

The regression results are:

Predictor	Estimate	Std. Error	p-value
Intercept	122.8021	1.6629	$< 2e-16$
mmc	-1.7495	0.6145	0.0049 **
pp	-7.8387	1.8055	$2.3e-05$ ***

Table 1: Multiple Linear Regression Results

To ensure the robustness of these estimates, we applied a non-parametric bootstrap-t method for confidence interval estimation. The results are:

Predictor	Estimate	Bootstrapped Std. Error	95% CI (Lower, Upper)
Intercept	122.8403	1.6584	(119.5052, 126.1015)
mmc	-1.7513	0.6400	(-2.9976, -0.4785)
pp	-7.8551	1.7496	(-11.3237, -4.4845)

Table 2: Bootstrap Confidence Intervals for Regression Coefficients

The bootstrap confidence intervals align closely with the standard OLS regression results, reinforcing the reliability of our findings. The bootstrap distributions of regression coefficients are visualized in Figure 7.

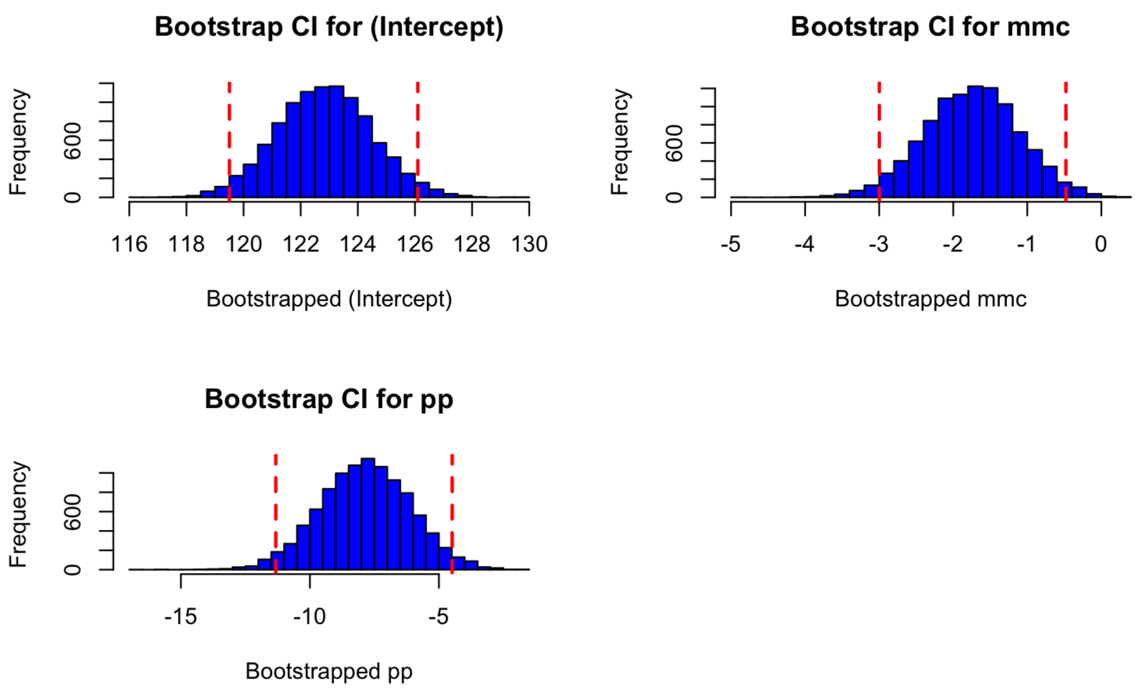

Figure 7: Bootstrap Confidence Intervals for Regression Coefficients

These results suggest:

- The coefficient for `mmc` (kilometers driven) is negative ($\beta_1 = -1.7495$), meaning that an increase in kilometers driven is associated with a slight decrease in driver fatalities. The confidence interval for β_1 does not include zero, indicating its statistically significant.
- The coefficient for `pp` (petrol price) is negative and statistically significant ($\beta_2 = -7.8387$), indicating that higher petrol prices are associated with fewer driver fatalities. This aligns with economic theory and behavioral responses to fuel costs: when petrol prices rise, people tend to reduce non-essential travel, leading to lower traffic volumes and fewer accidents. The confidence interval for β_2 does not include zero, reinforcing the robustness of this finding.

0.5.4 Findings

Findings:

- The introduction of the seatbelt law in 1983 led to a significant reduction in driver fatalities consistent with prior research and regulatory expectations.
- A strong seasonal pattern in driver fatalities was observed, indicating higher risks in certain months.
- There is a significant difference between front and rear passenger fatalities, possibly due to differing impact forces and protection measures.
- Higher fuel prices were associated with fewer driver fatalities, suggesting that economic factors influence driving behavior.

Overall, the findings validate prior expectations and exploratory analysis, demonstrating the effectiveness of resampling techniques in providing robust statistical conclusions.

0.6 Discussion

0.6.1 Limitations of the Study

- Temporal Scope: The dataset spans only from 1968 to 1984, excluding recent advancements in road safety, vehicle technology, and driving behavior.
- Geographical Restriction: Data is limited to Great Britain, making generalization to other regions challenging due to differences in road conditions and policies.
- Potential Confounders: The analysis does not explicitly control for factors such as weather conditions, road infrastructure improvements, or law enforcement efforts.
- Causal Limitations: While a reduction in fatalities is observed post-law, the effect cannot be solely attributed to seatbelt legislation, as other safety measures and economic conditions may have contributed.

Despite its limitations, this analysis remains relevant for road safety policies and public health research.

0.6.2 Policy and Practical Implications:

- Policymakers should continue reinforcing seatbelt laws and promoting safety regulations to further reduce fatalities.
- Seasonal patterns in fatalities highlight the need for targeted safety campaigns during high-risk months.
- The analysis of front and rear passenger fatalities further highlights that front-seat passengers face a higher risk of fatality. This underscores the importance of safety measures such as airbags and advanced restraint systems. Policymakers and automobile manufacturers should prioritize improvements in front-seat safety features to further reduce fatalities.

0.6.3 Future Research Directions

Building upon the findings of this study, future research could explore the following avenues:

- **Extension to Modern Data:** Analyzing more recent traffic fatality data would help assess whether similar trends persist with modern vehicle safety technologies and road infrastructure.
- **Machine Learning Approaches:** Advanced predictive modeling techniques, such as time-series forecasting and machine learning algorithms, could enhance accident risk prediction and inform data-driven safety interventions.
- **Regional Comparisons:** Expanding the study to include data from multiple countries could provide comparative insights into how different safety regulations impact road fatalities.

0.7 Conclusion

This study examined driver fatalities in Great Britain from 1969 to 1984 using the Seatbelts dataset, employing resampling techniques such as permutation tests and non-parametric bootstrap methods to evaluate the effects of seatbelt laws, seasonal patterns, and economic factors. The findings confirm that the introduction of the seatbelt law in 1983 significantly reduced driver fatalities, reinforcing its effectiveness as a safety regulation. Seasonal trends indicated higher fatalities in winter months, suggesting that adverse weather conditions and increased travel contribute to accident risks. Additionally, front-seat passengers experienced significantly higher fatality rates than rear-seat occupants, underscoring the need for enhanced front-seat safety measures such as airbags and improved restraint systems. Economic factors also played a role, with higher petrol prices associated with fewer fatalities, likely due to reduced travel. While the study is limited by its historical scope and geographic focus, it provides robust statistical validation of key road safety policies and offers valuable insights for future transportation safety strategies.

Bibliography

[1] McCarthy, M. (1989). The benefit of seat belt legislation in the United Kingdom. *Journal of Epidemiology and Community Health*, **43**(3), 218-222. https://doi.org/10.1136/jech.43.3.218.

[2] Harvard Law. *The Impact of Seat Belt Legislation on Road Safety*. http://www.law.harvard.edu/programs/olin$_{c}enter/papers/pdf/$341.pdf.

[3] Actuarial Post. (2021). 15 percent increase in accidents in winter versus summer. https://www.actuarialpost.co.uk/article/15-percent-increase-in-accidents-in-winter-versus-summer-17825.htm.

[4] National Claims. Road traffic accidents dependent on the time of year. https://national-claims.co.uk/road-traffic-accidents-dependent-on-the-time-of-year/.

[5] Grist, V. (2017). Analysis of Seatbelt Data. RPubs. https://rpubs.com/Vikki$_{G}rist/$299410.

[6] GraphPad. How to think about results from two-way ANOVA. https://www.graphpad.com/guides/prism/latest/statistics/how$_{t}o_{t}hink_{a}bout_{r}esults_{f}rom_{t}wo - way_{a}nova.htm$.

[7] MIT. (n.d.). Understanding the Shapiro-Wilk test for normality. https://math.mit.edu/ rmd/465/shapiro.pdf.

[8] Good, P. I. (2005). *Resampling Methods: A Practical Guide to Data Analysis*. Birkhäuser.

.1 Appendix: R Code

~~~~
---
title: "Ex4"
author: "Vishal Nair 1740105"
date: "2025-02-09"
output: pdf_document
---

### Vishal Nair 1740105

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
# Load necessary libraries
library(tidyverse)
library(ggcorrplot)
library(leaps)
```

### Data Loading and Pre-Processing
```{r}
# Load and format the dataset
data(Seatbelts)
df <- data.frame(Year = floor(time(Seatbelts)),
                 Month = factor(cycle(Seatbelts), labels = month.abb),
                 Seatbelts)

# Convert 'law' to a factor (0 = Before Law, 1 = After Law)
df$law <- as.factor(df$law)

```

### Exploratory Data Analysis
```{r}
summary(df)
```

```{r}
ggplot(df, aes(x = law, y = DriversKilled, fill = law)) +
  geom_boxplot() +
  ggtitle("Effect of Seatbelt Law on Driver Fatalities") +
  xlab("Seatbelt Law (0 = Before, 1 = After)") + ylab("Driver Fatalities") +
  theme_minimal()

```


```{r}
ggplot(df, aes(x = Month, y = DriversKilled, fill = Month)) +
  geom_boxplot() +
  ggtitle("Seasonality in Driver Fatalities") +
  xlab("Month") + ylab("Driver Fatalities") +
  theme_minimal()

```




```{r}
ggplot(df, aes(x = "Front", y = front)) +
  geom_boxplot(fill = "blue") +
  geom_boxplot(aes(x = "Rear", y = rear), fill = "red") +
  ggtitle("Comparison of Front vs. Rear Fatalities") +
  xlab("Seat Position") + ylab("Number of Fatalities") +
  theme_minimal()
~~~~

````

```
```{r}
ggplot(df, aes(x = PetrolPrice, y = DriversKilled)) +
  geom_point(color = "darkgreen") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  ggtitle("Relationship Between Petrol Price and Driver Fatalities") +
  xlab("Petrol Price") + ylab("Driver Fatalities") +
  theme_minimal()

```

```{r}
ggplot(df, aes(x = kms, y = DriversKilled)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  ggtitle("Kilometers Driven vs. Driver Fatalities") +
  xlab("Kilometers Driven (Thousands)") + ylab("Driver Fatalities") +
  theme_minimal()

```
```{r}
cor_data <- df %>% select(DriversKilled, kms, PetrolPrice)

# Compute Spearman correlation matrix
cor_matrix <- cor(cor_data, use = "complete.obs", method = "spearman")

# Visualize correlation matrix as a heatmap
ggcorrplot(cor_matrix,lab=TRUE)

```


### Statistical Analysis with Resampling Techniques

Permutation Test for Anova F-statistic
```{r}
set.seed(123)

nr <- 10000  # Number of permutations
perm_F_month <- numeric(nr)  # Store permuted F-statistics for Month
perm_F_law <- numeric(nr)    # Store permuted F-statistics for Law
perm_F_interaction <- numeric(nr)  # Store permuted F-statistics for interaction

# Extract relevant variables
Month <- df$Month
law <- df$law
DriversKilled <- df$DriversKilled

# Compute observed ANOVA model
fit <- lm(DriversKilled ~ Month + law + Month:law)
anova_obs <- anova(fit)

# Extract observed F-statistics
obs_F_month <- anova_obs$`F value`[1]   # F-statistic for Month
obs_F_law <- anova_obs$`F value`[2]     # F-statistic for Law
obs_F_interaction <- anova_obs$`F value`[3]  # F-statistic for interaction

# Perform permutation test
for (i in 1:nr) {
  nDriversKilled <- sample(DriversKilled)  # Shuffle response variable
  fit_perm <- lm(nDriversKilled ~ Month + law + Month:law)
  anova_perm <- anova(fit_perm)

  # Store permuted F-statistics
  perm_F_month[i] <- anova_perm$`F value`[1]
  perm_F_law[i] <- anova_perm$`F value`[2]
  perm_F_interaction[i] <- anova_perm$`F value`[3]
}

````

````r
# Compute empirical p-values
p_value_month <- mean(perm_F_month >= obs_F_month)
p_value_law <- mean(perm_F_law >= obs_F_law)
p_value_interaction <- mean(perm_F_interaction >= obs_F_interaction)

# Print results
cat("Permutation Test p-value for Month:", p_value_month, "\n")
cat("Permutation Test p-value for Seatbelt Law:", p_value_law, "\n")
cat("Permutation Test p-value for Month:Law Interaction:", p_value_interaction, "\n")

# Plot permutation distributions
par(mfrow = c(1, 3))  # Arrange plots side by side

hist(perm_F_month, breaks = 30, col = "blue", main = "Permutation Test: Month",
     xlab = "Permuted F-statistic", ylab = "Frequency")
abline(v = obs_F_month, col = "red", lwd = 2, lty = 2)

hist(perm_F_law, breaks = 30, col = "green", main = "Permutation Test: Seatbelt Law",
     xlab = "Permuted F-statistic", ylab = "Frequency")
abline(v = obs_F_law, col = "red", lwd = 2, lty = 2)

hist(perm_F_interaction, breaks = 30, col = "purple", main = "Permutation Test: Interaction",
     xlab = "Permuted F-statistic", ylab = "Frequency")
abline(v = obs_F_interaction, col = "red", lwd = 2, lty = 2)

par(mfrow = c(1, 1))  # Reset plot layout


```

Bootstrap CI for best regression coefficients

```{r}
# Transform variables
seatbeltstransformd <- df %>%
  mutate(
    pp = (PetrolPrice - mean(PetrolPrice)) / sd(PetrolPrice),  # Standardize petrol price
    mm = kms / 1000,  # Convert kms to thousands
    mmc = mm - mean(mm)  # Centering kms
  )

# Fit the regression model using transformed variables
fit_transformed <- lm(DriversKilled ~ mmc + pp, data = seatbeltstransformd)

# Display regression summary
summary(fit_transformed)

set.seed(123)

# Number of bootstrap resamples
nr <- 10000
boot_t <- matrix(NA, nrow = nr, ncol = length(coef(fit_transformed)))  # Store bootstrap t-values
boot_theta <- matrix(NA, nrow = nr, ncol = length(coef(fit_transformed)))  # Store bootstrap coefficients

# Perform bootstrap resampling
for (i in 1:nr) {
  boot_indices <- sample(1:nrow(seatbeltstransformd), replace = TRUE)  # Resample with replacement
  boot_sample <- seatbeltstransformd[boot_indices, ]  # Get bootstrap sample
  boot_fit <- lm(DriversKilled ~ mmc + pp, data = boot_sample)  # Fit model

  boot_theta[i, ] <- coef(boot_fit)  # Store bootstrapped coefficients
  boot_t[i, ] <- (coef(boot_fit) - coef(fit_transformed)) / summary(boot_fit)$coefficients[, 2]  # Compute t-values
}

# Compute Bootstrap Estimates and Standard Errors
````

````
bootstrap_mean <- colMeans(boot_theta)  # Mean of bootstrapped estimates
bootstrap_se <- apply(boot_theta, 2, sd)  # Standard deviation (SE) of bootstrapped estimates

# Compute 95% CI using bootstrap-t method
ci_lower <- coef(fit_transformed) + summary(fit_transformed)$coefficients[, 2] * apply(boot_t, 2, function(x) quantile(x, 0.025))
ci_upper <- coef(fit_transformed) + summary(fit_transformed)$coefficients[, 2] * apply(boot_t, 2, function(x) quantile(x, 0.975))


# Print results
cat("Bootstrap Results:\n")
results_df <- data.frame(
  Estimate = bootstrap_mean,
  Std_Error = bootstrap_se,
  CI_Lower = ci_lower,
  CI_Upper = ci_upper
)
print(results_df)
# Plot histograms for each regression coefficient
par(mfrow = c(2, 2))  # Arrange plots in a grid

for (i in 1:length(coef(fit_transformed))) {
  hist(boot_theta[, i], breaks = 30, col = "blue", main = paste("Bootstrap CI for", names(coef(fit_transformed))[i]),
       xlab = paste("Bootstrapped", names(coef(fit_transformed))[i]), ylab = "Frequency")
  abline(v = c(ci_lower[i], ci_upper[i]), col = "red", lwd = 2, lty = 2)  # Add CI lines
}

par(mfrow = c(1, 1))  # Reset plot layout

```

```{r}
residuals <- resid(your_anova_model)

```


```{r}
# Check normality of differences
diff_fatalities <- df$front - df$rear
shapiro.test(diff_fatalities)  # Shapiro-Wilk test for normality
# Q-Q plot
qqnorm(diff_fatalities)
qqline(diff_fatalities)
```


```{r}
set.seed(123)  # For reproducibility

nr <- 10000  # Number of permutations
perm_diffs <- numeric(nr)  # Store permuted mean differences
cnt <- 0  # Initialize counter

# Compute observed mean difference
obs_diff <- mean(df$front) - mean(df$rear)

# Combine both groups into a single vector
vect <- c(df$front, df$rear)
n1 <- length(df$front)
n2 <- length(df$rear)
total <- n1 + n2

# Perform permutation test
for (i in 1:nr) {
  d <- sample(vect, total)  # Shuffle the combined dataset
  ne <- d[1:n1]  # First half assigned to front fatalities
````

```
  co <- d[(n1+1):total]  # Second half assigned to rear fatalities
  perm_diffs[i] <- mean(ne) - mean(co)  # Compute new mean difference

  if (perm_diffs[i] > obs_diff) {
    cnt <- cnt + 1  # Count how many times permuted diff is greater than observed
  }
}

# Compute empirical p-value
p_value_perm <- cnt / nr

# Print results
cat("Observed Mean Difference:", obs_diff, "\n")
cat("Permutation Test p-value:", p_value_perm, "\n")

# Plot permutation results
hist(perm_diffs, breaks = 30, col = "blue", main = "Permutation Test: Front vs. Rear
    Fatalities",
     xlab = "Permuted Mean Difference", ylab = "Frequency")
abline(v = obs_diff, col = "red", lwd = 2, lty = 2)

```
```

Listing 1: R Code for Data Analysis
```