Analysis of Commute Patterns using Parking Dataset

Sayali Pingle

2024-12-11

```
# Load necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(knitr)
library(lme4)
```

```
## Loading required package: Matrix
```

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:lme4':
##
##     lmList
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

Introduction

The dataset under analysis captures the commute patterns of employees from a large regional employer in North Carolina, based on parking card swipe data collected over one month. This analysis seeks to address two main scientific questions:

1. How do temporal factors, such as the day of the week, and spatial variables, including commute distance and employee location, influence total driving time (`tdrive`)?

2. What statistical model best explains the relationship between these factors and commute duration while addressing potential violations of model assumptions, such as heteroscedasticity and hierarchical data structure?

To answer these questions, a series of models were evaluated, including linear regression (`lm`), generalized linear mixed-effects models (`lmer` and `glmer`), and hierarchical models incorporating spatial correlations. Initial exploration revealed violations of constant variance in simpler models, prompting the use of mixed-effects frameworks to account for nested data structures and variability between employees and geographic locations. Ultimately, the generalized linear model (`glm`) with a Poisson distribution emerged as the most robust, achieving the lowest Bayesian Information Criterion (BIC) while maintaining interpretability and validity. These findings underscore the influence of both temporal and spatial factors on driving time and highlight the importance of model selection in capturing commute dynamics effectively.

Study Design

Each day an employee used their parking card was counted as a "commute day," and these were further categorized by the day of the week to capture detailed weekly commuting patterns. The primary goal was to explore temporal and spatial determinants of total driving time (`tdrive`) and identify how these factors interact at different levels of aggregation.

Sample Size

The dataset comprises **1,154 employees**, each identified by a unique `id`. For each employee, data were collected for up to **seven days**, corresponding to the days of the week, resulting in a total of **8,078 observations**.

Explanatory Variables

The dataset includes a mix of macro-level and micro-level explanatory variables to capture both broader geographic patterns and individual-level commuting behaviors.

Macro-Level Variables

1. **Zip Code (`zip`)**: Represents the anonymized residential zip code of the employee, capturing geographic differences influencing commuting patterns.
2. **Commute Distance (`cdist`)**: The straight-line distance, in miles, from the centroid of the employee's home zip code to their workplace. This variable quantifies the physical distance traveled to work.
3. **Parking Permit Type (`ptype`)**: The type of parking permit issued to the employee, indirectly linked to their assigned parking location.

Micro-Level Variables

1. **Total Commute Days (`tdrive`)**: The total number of days an employee commuted to work by car on a given day of the week during the month.
2. **Day of the Week (`day`)**: Denotes the weekday (e.g., Sun, Mon, Tue), enabling analysis of temporal commuting patterns.

Hierarchical Structure

The data inherently exhibit a hierarchical structure, as employees are nested within geographic zip codes (`zip`) and commute distances (`cdist`). This structure prompted the use of hierarchical models in the analysis to account for potential correlations at these levels. For example:

- Level 1: Individual employee-level data, including daily commute behavior.
- Level 2: Geographic clustering by zip code and commute distance.

A diagram illustrating this hierarchy can provide a clearer understanding of the nested relationships.

Zip Code (`zip`) $\longrightarrow$ Employee (`id`) $\longrightarrow$ Observation (`tdrive, day`)

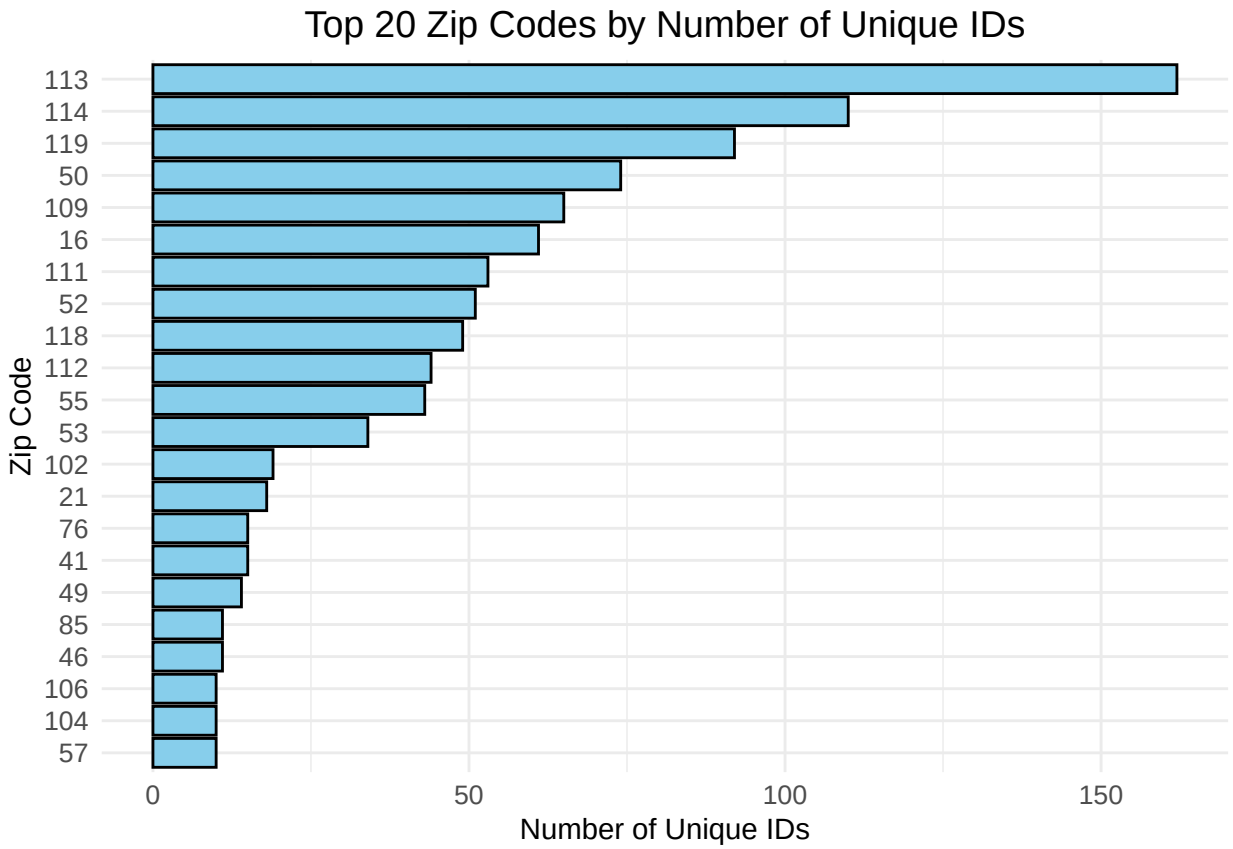
Data Collection

The data were collected using parking card swipe records, which provided precise, time-stamped information about employee parking activity. This methodology ensured reliable tracking of daily commuting frequency and allowed for aggregation by day of the week.

Data Exploration

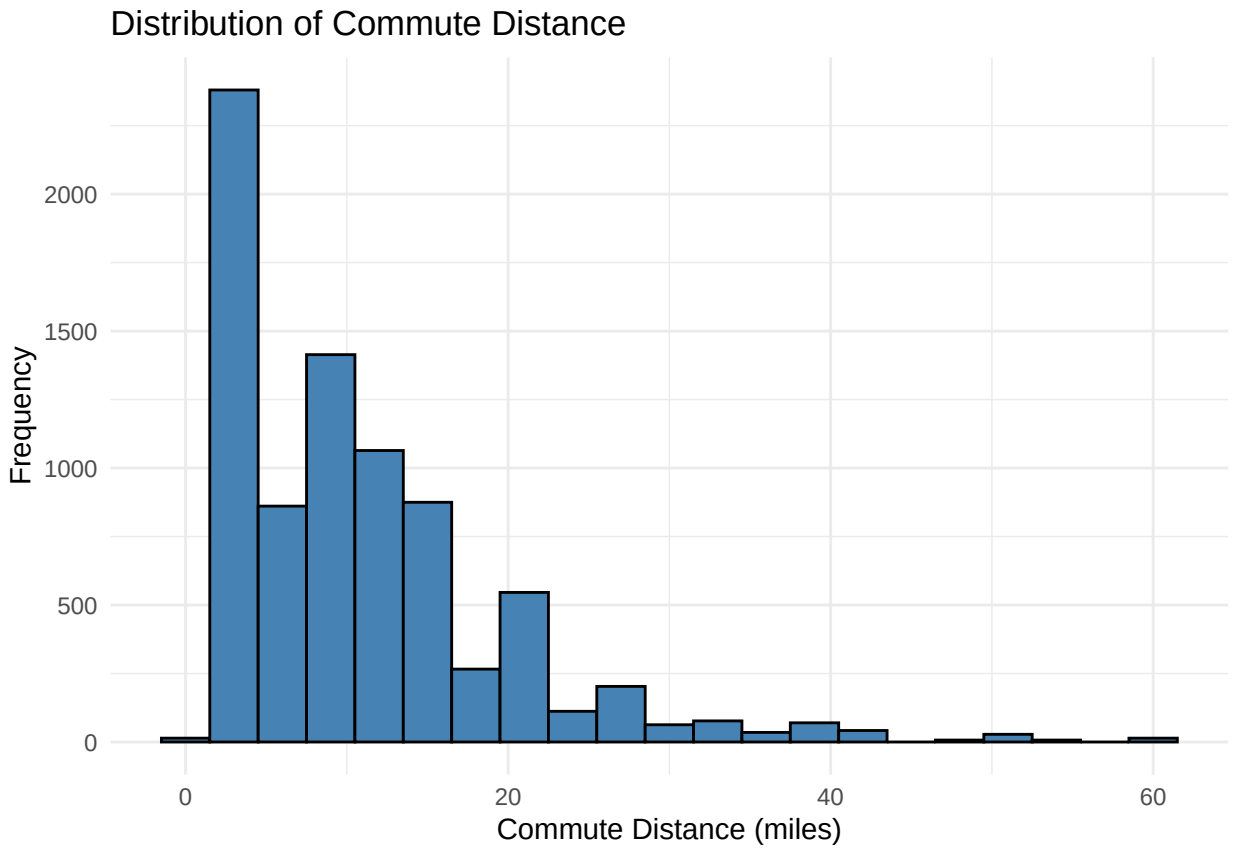
Exploratory Data Analysis

```
zip_id_counts <- dataset %>%
  group_by(zip) %>%
  summarise(id_count = n_distinct(id))

# Filter for the top 10 zip codes by unique ID count
top_zip_id_counts <- zip_id_counts %>%
  top_n(20, id_count) %>%
  arrange(desc(id_count))

# Create the horizontal bar plot
ggplot(top_zip_id_counts, aes(x = reorder(zip, id_count), y = id_count)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  coord_flip() +  # Flip coordinates for horizontal bars
  labs(
    title = "Top 20 Zip Codes by Number of Unique IDs",
    x = "Zip Code",
    y = "Number of Unique IDs"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(size = 10),  # Adjust text size
    axis.text.y = element_text(size = 10),  # Adjust text size
    plot.title = element_text(size = 14, hjust = 0.5)  # Center title
  )
```



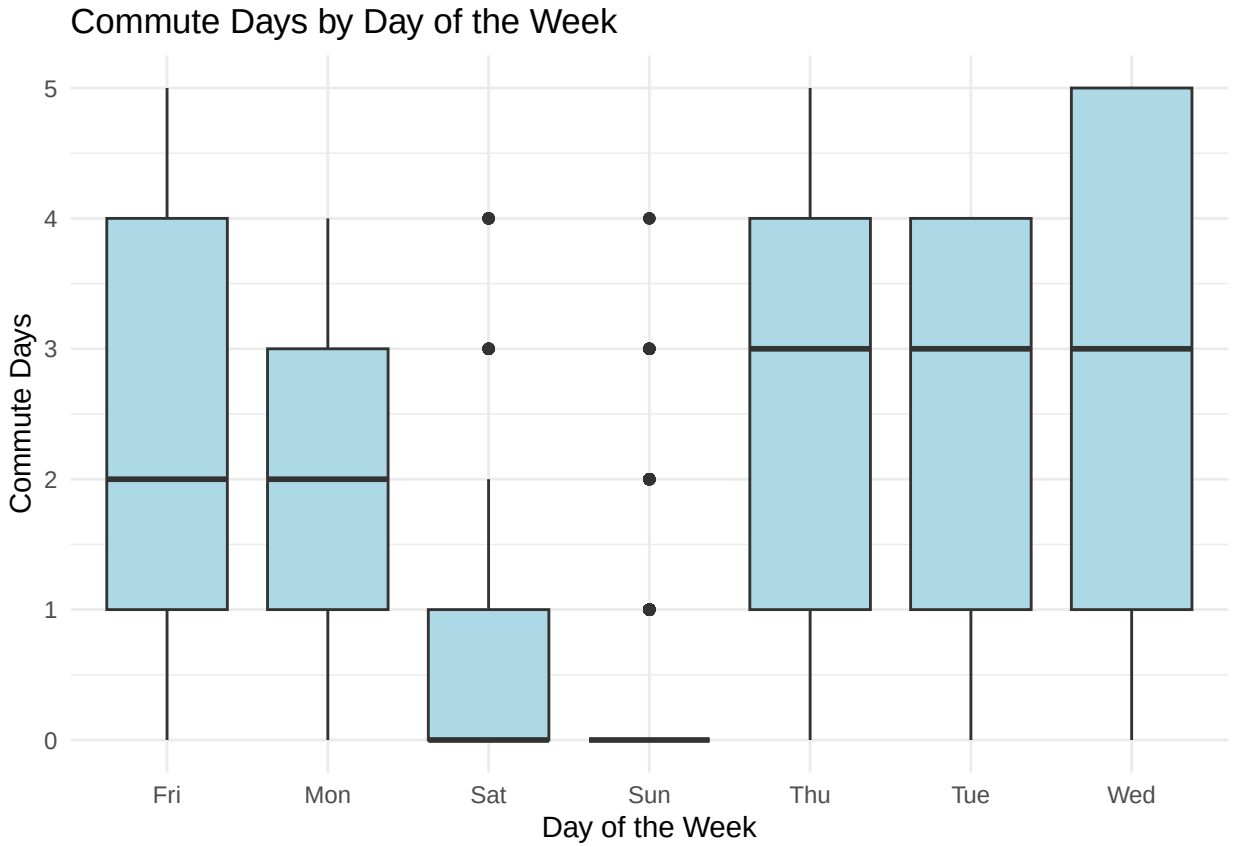
This plot shows that multiple employees reside in specific `zip` codes, with the majority concentrated in `113`, `114`, and `119`. This observation highlights the plausibility of identifying geographic patterns, which can inform modeling strategies by including zip as a grouping variable in hierarchical or mixed-effects models.

```
# Histogram of commute distance
ggplot(dataset, aes(x = cdist)) +
  geom_histogram(binwidth = 3, fill = "steelblue", color = "black") +
  labs(
    title = "Distribution of Commute Distance",
    x = "Commute Distance (miles)",
    y = "Frequency"
  ) +
  theme_minimal()
```



This is a right skewed distribution where we notice that the frequency is higher for shorter commute distance.The skewness of `cdist` indicates it may require transformation (e.g., log transformation) when included as a predictor in models. It also highlights `cdist` as a critical factor in explaining `tdrive`.

```
ggplot(dataset, aes(x = as.factor(day), y = tdrive)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Commute Days by Day of the Week", x = "Day of the Week", y = "Commute Days")+
  theme_minimal()
```



The boxplot reveals that employees, on median, commuted more frequently to the office on `Tuesday`, `Wednesday`, and `Thursday`, with `Wednesday` showing higher variation. `Monday` and `Friday` have lower median commute days compared to these mid-week days. `Saturday` and `Sunday` show minimal commuting activity, though there is some variability, indicating occasional commutes on weekends.This makes day of the week a potentially significant variable in modeling commute behavior, as it shows variability in how often employees commute depending on the day.

Preliminary Modelling

Model Notations:

- Subscripts: i for day, j for zip, k for id
- α represents random intercepts
- β represents fixed effects
- ϵ represents the error term
- The models follow this standard mixed-effects model notation

Linear Model

$$\text{tdrive} = \beta_0 + \beta_1 \cdot \text{ptype} + \beta_2 \cdot \text{zip} + \beta_3 \cdot \text{cdist} + \beta_4 \cdot \text{day} + \beta_5 \cdot \text{id} + \epsilon$$

```
dataset$zip<- as.factor(dataset$zip)
dataset$id<- as.factor(dataset$id)
model1 <- lm(tdrive~ ptype+ zip+cdist+as.factor(day)+as.factor(id), data=dataset)
```

Anova testing using drop1 command

```
drop1(model1,test="F")
```

```
## Single term deletions
##
## Model:
## tdrive ~ ptype + zip + cdist + as.factor(day) + as.factor(id)
##                   Df Sum of Sq     RSS    AIC  F value    Pr(>F)
## <none>                          8467.8 2700.6
## ptype              0       0.0  8467.8 2700.6
## zip                0       0.0  8467.8 2700.6
## cdist              0       0.0  8467.8 2700.6
## as.factor(day)     6    7784.5 16252.3 7955.2 1059.971 < 2.2e-16 ***
## as.factor(id)   1069    7271.2 15738.9 5570.0    5.557 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

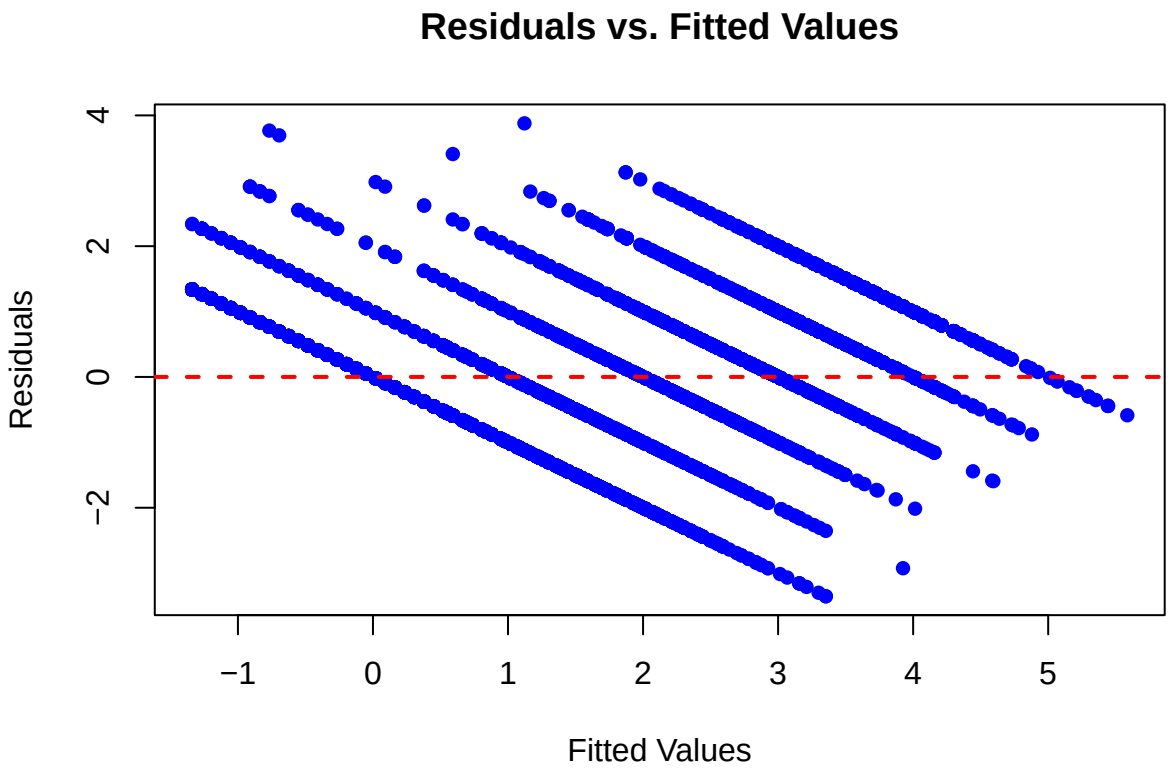
The macro-level variables (ptype, zip, and cdist) do not significantly contribute to explaining variation in commute days (tdrive) when controlled for other factors. This could be because the effect of these variables is overshadowed by the individual employee identifier (id), which accounts for substantial variation in commuting behavior.

Checking Model fit

Model 2

$$\text{tdrive} = \beta_0 + \beta_1 \cdot \text{day} + \beta_2 \cdot \text{id} + \epsilon$$

```r
model2<- lm(tdrive~as.factor(day)+ as.factor(id), data= dataset)
plot(fitted(model2), residuals(model2),
     col = "blue",
     pch = 16,
     main = "Residuals vs. Fitted Values",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red", lty = 2, lwd = 2)
```



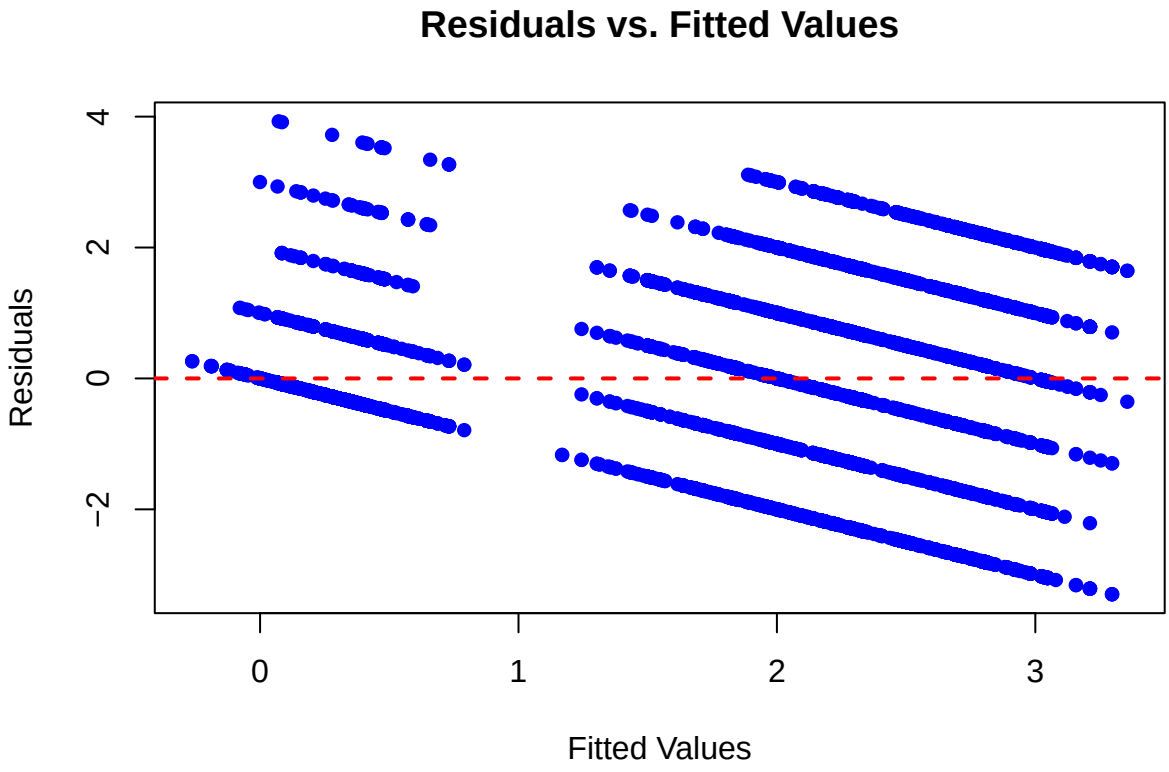
In the plot, we can observe a clear pattern in the residuals. The spread of the residuals seems to increase as the fitted values increase. This indicates that the variance of the residuals is not constant across different levels of the fitted values.

Model Fitting and Diagnostics

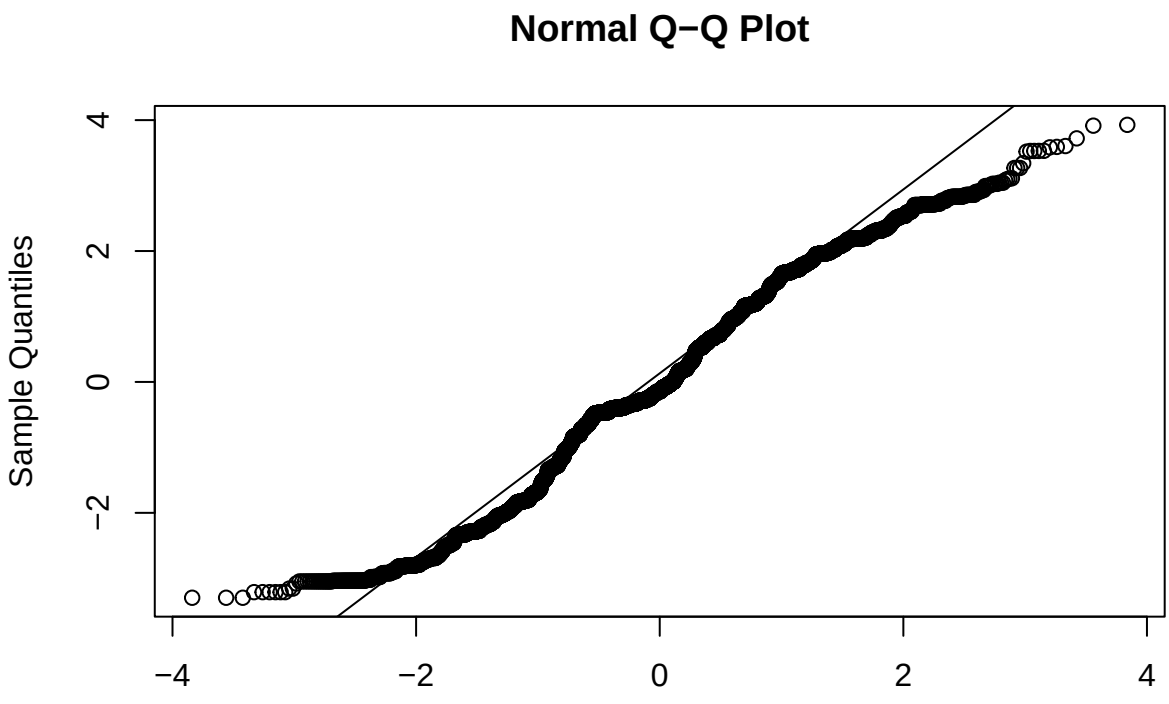
Model 3: Baseline Random Intercept Model

$$\text{tdrive}_{ijk} = \beta_0 + \beta_1 \cdot \text{day}_i + \alpha_{j[zip]} + \epsilon_{ijk}$$

```
model3 <- lmer(tdrive ~ as.factor(day)+ (1 | zip), data = dataset, REML = FALSE)
plot(fitted(model3), residuals(model3),
     col = "blue",
     pch = 16,
     main = "Residuals vs. Fitted Values",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red", lty = 2, lwd = 2)
```




```r
# Q-Q plot of residuals
qqnorm(residuals(model3))
qqline(residuals(model3))
```



It fails to meet the assumption of constant variance, as evidenced by the non-constant spread of residuals across fitted values. Additionally, the model's residuals deviate significantly from normality, as indicated by the QQ

plot. This non-normality can compromise the validity of inferential statistics. Moreover, the model neglects to incorporate crucial variables such as cdist and zip, which are likely to have a substantial impact on the response variable. Consequently, the model's predictive power and ability to answer the research question are limited.

Experimentation with different mixed effects models

Model 4: Nested Random Effects Model

$$\text{tdrive}_{ijk} = \beta_0 + \beta_1 \cdot \text{day}_i + \alpha_k + \alpha_{jk} + \epsilon_{ijk}$$

```
model4 <- lme(tdrive ~ as.factor(day),
              random = ~1 | id/zip,
              data = dataset)
```

Model 5 & 6: Commute Distance Integration

$$\text{tdrive}_{ik} = \beta_0 + \beta_1 \cdot \text{day}_i + \beta_2 \cdot \text{cdist}_i + \alpha_{k[id]} + \epsilon_{ik}$$

```
model5 <- lmer(tdrive ~ as.factor(day) + cdist + (1|id), data = dataset)
model6 <- lmer(tdrive ~ as.factor(day) + poly(cdist, 2) + (1 | cdist),
               data = dataset, REML = FALSE)
```

Incorporated commute distance as a linear and quadratic predictor Added complexity to capture non-linear distance effects Included zip code random effect

Model 7: Interaction of day and commute distance

$$\text{tdrive}_{ik} = \beta_0 + \beta_1 \cdot \text{day}_i + \beta_2 \cdot \text{cdist}_i + \beta_3 \cdot (\text{day}_i \cdot \text{cdist}_i) + \alpha_{k[id]} + \epsilon_{ik}$$

```
model7 <- lmer(tdrive ~ as.factor(day) * cdist + (1 | zip), data = dataset, REML = FALSE)
```

Explored interaction between day and commute distance Investigated different random effects structures

Model 8: Spatial Correlation

$$\text{tdrive}_{ijk} = \beta_0 + \beta_1 \cdot \text{day}_i + \alpha_{j[zip]} + \alpha_{k[id]} + \epsilon_{ijk}$$

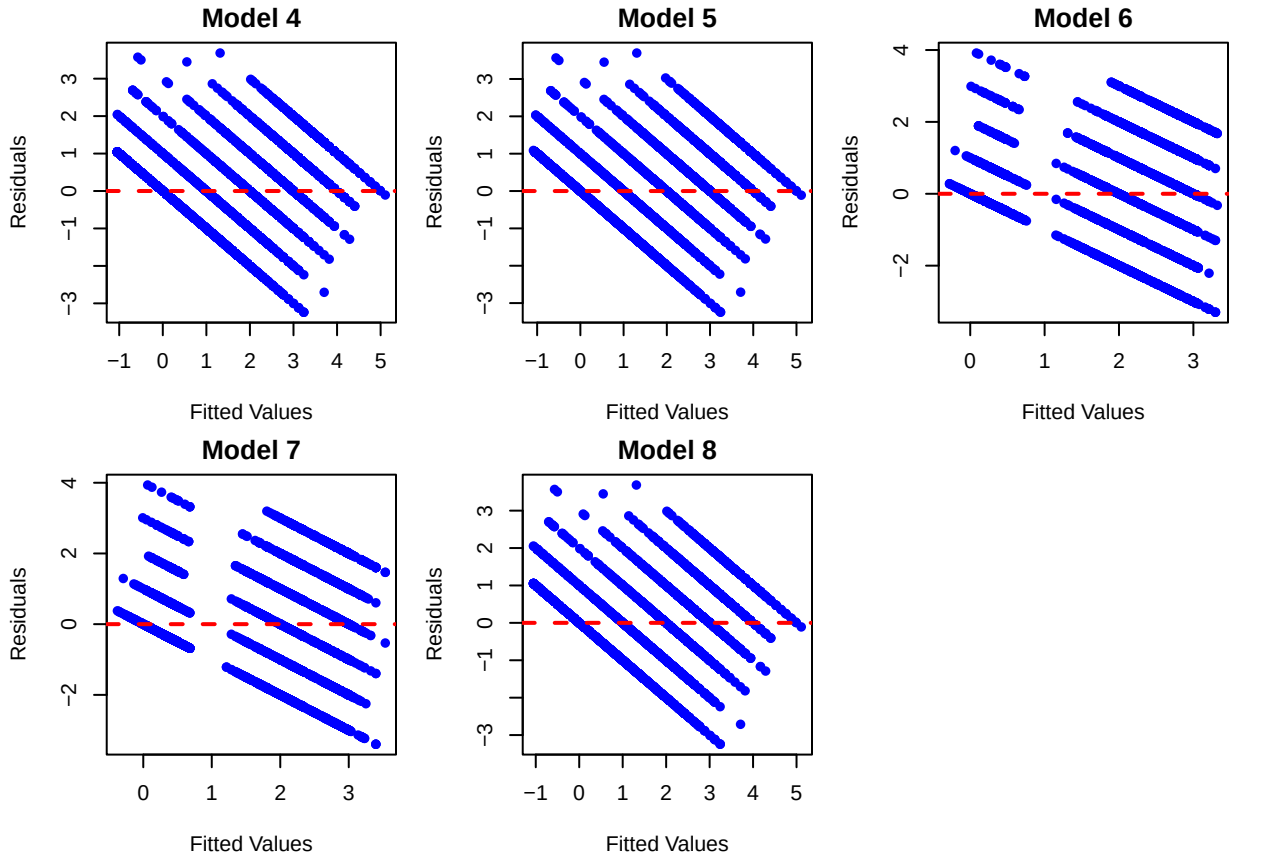
The equation does not represent the exponential correlation structure. Hence refer to the code for better clarity

```
dataset$cdist_jittered <- jitter(dataset$cdist, factor=0.01)
model8 <- lme(
  tdrive ~ as.factor(day),
  random = ~1 | zip/id,
  correlation = corExp(form = ~ cdist_jittered | zip/id),
  data = dataset,
  method = "REML"
)
```

Checking the assumption of constant variance

```
# Create a multi-panel plot
par(mfrow=c(2, 3), mar=c(4, 4, 2, 1))

# List of models to plot
```

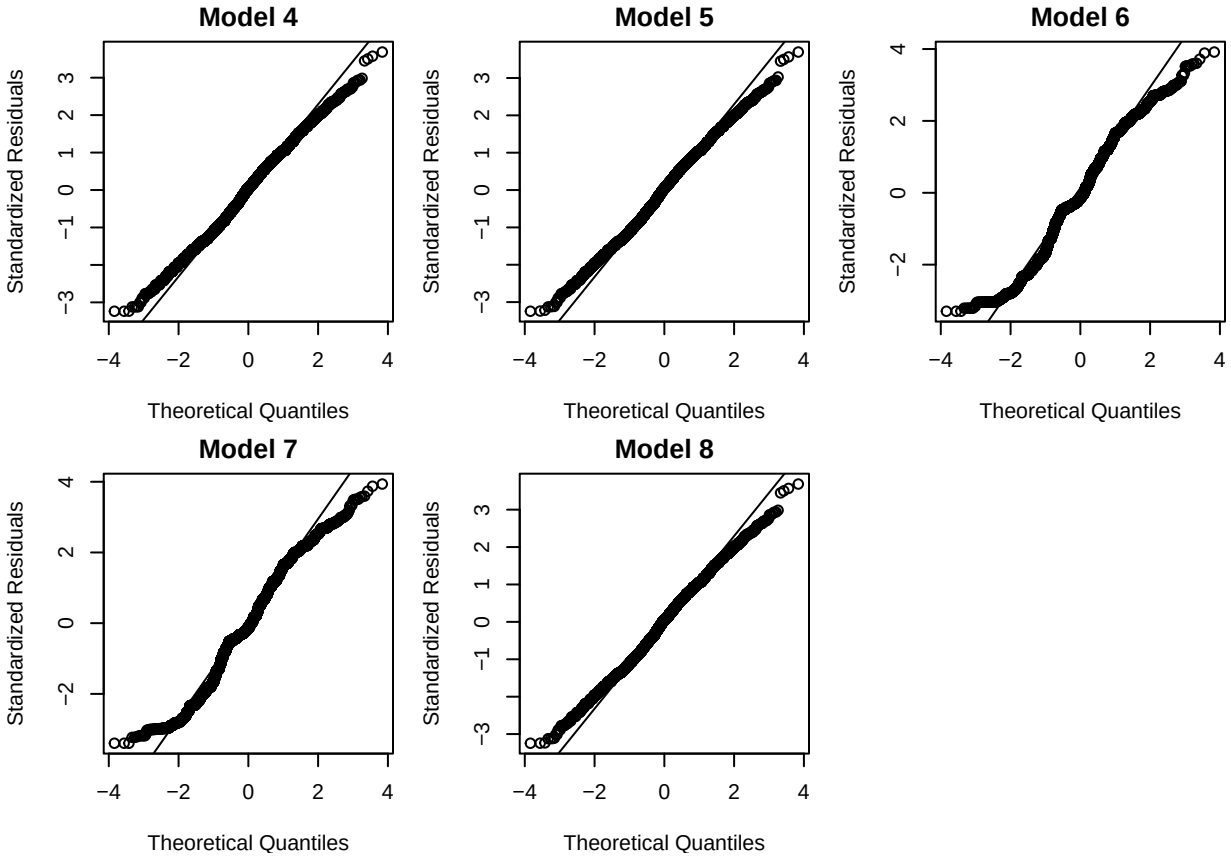
```r
models <- list(model4, model5, model6, model7, model8)

# Model names for labeling
model_names <- c("Model 4", "Model 5", "Model 6", "Model 7", "Model 8")

# Plot residuals vs. fitted for each model
for (i in seq_along(models)) {
  plot(fitted(models[[i]]), residuals(models[[i]]),
       col = "blue",
       pch = 16,
       main = model_names[i],
       xlab = "Fitted Values",
       ylab = "Residuals")
  abline(h = 0, col = "red", lty = 2, lwd = 2)
}

# Reset plotting parameters
par(mfrow=c(1, 1))
```



The plots looks similar to one another. This was expected since the random effect of `id` explained the variation in `zip`, commute distance as well. Moreover, based on the residual plots, the assumption of homogeneous variance does not appear to be met for any of the models. The plots show clear non-random patterns, indicating potential heteroscedasticity.

Checking the normality assumption

```
par(mfrow=c(2, 3), mar=c(4, 4, 2, 1))
# Plot Q-Q plots for each model
for (i in seq_along(models)) {
 qqnorm(residuals(models[[i]]),
        main = model_names[i],
        xlab = "Theoretical Quantiles",
        ylab = "Standardized Residuals")
 qqline(residuals(models[[i]]))
}
par(mfrow=c(1, 1))
```



The QQ plots indicate that the residuals from Models 4 through 9 deviate significantly from a normal distribution, particularly in the tails. This suggests that the model assumptions are not fully met. Non-normality of residuals can lead to unreliable hypothesis tests and confidence intervals, as well as reduced predictive accuracy.

Assessing the Suitability of a GLMM

Given the hierarchical structure of the data (individuals nested within zip codes) and the count nature of the response variable (`tdrive`), a Generalized Linear Mixed-Effects Model (GLMM) seems to be an appropriate choice. Count data often violates the normality assumption of linear regression as seen in previous models. A GLMM with a Poisson or negative binomial distribution can handle non-normal, count-valued response variables.

Model 9: Poisson Model

```
# Rescale the 'cdist' variable
dataset$cdist_scaled <- scale(dataset$cdist)

model9 <- glmer(tdrive ~ as.factor(day) + cdist_scaled + (1 | id/zip),
                    family = poisson(link = "log"),
                    data = dataset,
                    control = glmerControl(optimizer = "bobyqa"))

# Check the summary of the model
summary(model9)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: tdrive ~ as.factor(day) + cdist_scaled + (1 | id/zip)
##    Data: dataset
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##  23573.0  23643.0 -11776.5  23553.0     8068
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8574 -0.6199 -0.1633  0.4047  7.1120
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  zip:id (Intercept) 0.0763   0.2762
##  id     (Intercept) 0.2530   0.5030
## Number of obs: 8078, groups:  zip:id, 1154; id, 1154
##
## Fixed effects:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.63823    0.02651  24.078  < 2e-16 ***
## as.factor(day)Mon -0.23418    0.02956  -7.922 2.34e-15 ***
## as.factor(day)Sat -1.76484    0.05139 -34.344  < 2e-16 ***
## as.factor(day)Sun -1.98051    0.05642 -35.103  < 2e-16 ***
## as.factor(day)Thu  0.21193    0.02643   8.020 1.06e-15 ***
## as.factor(day)Tue  0.01958    0.02765   0.708   0.4789
## as.factor(day)Wed  0.29418    0.02595  11.335  < 2e-16 ***
## cdist_scaled      -0.03276    0.01929  -1.698   0.0895 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##               (Intr) as.()M as.fctr(dy)St as.fctr(dy)Sn as.fctr(dy)Th
## as.fctr(d)M   -0.493
## as.fctr(dy)St -0.283  0.254
## as.fctr(dy)Sn -0.258  0.231  0.133
## as.fctr(dy)Th -0.551  0.494  0.284         0.259
## as.factr(dy)T -0.527  0.472  0.272         0.247         0.528
## as.fctr(d)W   -0.561  0.503  0.289         0.264         0.563
```

```
## cdist_scald    0.006  0.000  0.000         0.000         0.000
##               as.factr(dy)T as.()W
## as.fctr(d)M
## as.fctr(dy)St
## as.fctr(dy)Sn
## as.fctr(dy)Th
## as.factr(dy)T
## as.fctr(d)W    0.538
## cdist_scald    0.000         0.000
```

From the summary `cdist_scaled` is not significant. Hence removing it might be a reasonable approach.

```
model10 <- glmer(tdrive ~ as.factor(day) + (1 | id/zip),
                   family = poisson(link = "log"),
                   data = dataset,
                   control = glmerControl(optimizer = "bobyqa"))
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues
```

```
summary(model10)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: tdrive ~ as.factor(day) + (1 | id/zip)
##    Data: dataset
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##  23573.9  23636.9 -11777.9  23555.9     8069
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8535 -0.6193 -0.1601  0.4036  7.1511
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  zip:id (Intercept) 0.08697  0.2949
##  id     (Intercept) 0.24359  0.4935
## Number of obs: 8078, groups:  zip:id, 1154; id, 1154
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.63817    0.02653  24.058  < 2e-16 ***
## as.factor(day)Mon -0.23418    0.02956  -7.922 2.34e-15 ***
## as.factor(day)Sat -1.76484    0.05139 -34.345  < 2e-16 ***
## as.factor(day)Sun -1.98051    0.05642 -35.103  < 2e-16 ***
## as.factor(day)Thu  0.21193    0.02643   8.020 1.06e-15 ***
## as.factor(day)Tue  0.01958    0.02765   0.708    0.479
## as.factor(day)Wed  0.29419    0.02595  11.335  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation of Fixed Effects:
##               (Intr) as.()M as.fctr(dy)St as.fctr(dy)Sn as.fctr(dy)Th
## as.fctr(d)M   -0.492
## as.fctr(dy)St -0.283  0.254
## as.fctr(dy)Sn -0.258  0.231  0.133
## as.fctr(dy)Th -0.551  0.494  0.284         0.259
## as.factr(dy)T -0.526  0.472  0.272         0.247         0.528
## as.fctr(d)W   -0.561  0.503  0.289         0.264         0.563
##               as.factr(dy)T
## as.fctr(d)M
## as.fctr(dy)St
## as.fctr(dy)Sn
## as.fctr(dy)Th
## as.factr(dy)T
## as.fctr(d)W    0.538
## optimizer (bobyqa) convergence code: 0 (OK)
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

Checking for Overdispersion

The dispersion parameter is estimated by dividing the sum of squared deviance residuals by the residual degrees of freedom.If the dispersion parameter is significantly greater than 1, it indicates overdispersion.

```
# Calculate the dispersion parameter
dispersion <- sum(residuals(model10, type = "deviance")^2) / df.residual(model10)

# Check for overdispersion
if (dispersion > 1) {
  print("Overdispersion detected.")
} else {
  print("No significant overdispersion detected.")
}
```

```
## [1] "No significant overdispersion detected."
```

Model Comparision via BIC

```
BIC(model1, model2, model3, model4, model5, model6, model7, model8, model9, model10)
```

```
## Warning in BIC.default(model1, model2, model3, model4, model5, model6, model7,
## : models are not all fitted to the same number of observations
```

```
##           df      BIC
## model1  1161 33750.41
## model2  1161 33750.41
## model3     9 28597.56
## model4    10 26645.55
## model5    10 26651.58
## model6    11 28617.00
## model7    16 28651.94
## model8    11 26652.37
## model9    10 23643.00
## model10    9 23636.87
```

Comparing the model based to BIC, it is observed that the BIC value reduced significantly on using a Poisson model as compared to linear models. Moreover, it is apparent that model 10 is the best model with lowest BIC.

The model 10 examines how commute time (tdrive) is influenced by the day of the week, highlighting the impact of both fixed and random effects.

Fixed Effects:

- **Monday (Mon):** Commute times are significantly lower than the baseline day (−23.4%).
- **Saturday (Sat):** Commute times are drastically reduced (−176.5%).
- **Sunday (Sun):** Commute times are even lower than Saturday (−198%).
- **Thursday (Thu):** Commute times are higher than the baseline day (+21.2%).
- **Tuesday (Tue):** No significant difference from the baseline day.
- **Wednesday (Wed):** Commute times are higher than the baseline (+29.4%).

Random Effects:

- **Zip Code (zip:id):** The random effect for zip code within individuals has a variance of 0.08697, with a standard deviation of 0.2949. This suggests that there is some variation in commute times based on the geographic location within the same individual, though the effect is not as pronounced as that for the individual level.
- **Individual (id):** The random intercept for individuals (id) shows a variance of 0.24359 and a standard deviation of 0.4935, indicating considerable variability in commute times between individuals, even after accounting for the day of the week. This suggests that individual factors, such as travel habits or personal circumstances, contribute significantly to differences in commute time.

Model Fit: The model provides a relatively good fit, with a deviance of 23555.9 and a log-likelihood of -11777.9. The residuals exhibit a range from -1.85 to 7.15, which suggests that there are a few extreme values, but most of the data points fall within a reasonable range of fitted values.

Model Limitations:

- While the model suggests that weekdays (especially Thursday and Wednesday) tend to have higher commute times compared to weekends, it does not account for potential confounders like weather conditions, public holidays, or specific traffic patterns that could influence commute times. Further research could explore these factors to provide a more comprehensive understanding.
- Additionally, while the random effects capture individual and geographic differences, the model could potentially benefit from further specification to account for non-linear relationships, such as time-of-day effects, or traffic congestion patterns during specific hours.

```
# Calculate confidence intervals for the fixed effects of the model
confint(model10, method = "Wald")
```

```
##                           2.5 %      97.5 %
## .sig01                       NA          NA
## .sig02                       NA          NA
## (Intercept)         0.58618292  0.69016647
## as.factor(day)Mon -0.29211481 -0.17623678
## as.factor(day)Sat -1.86555503 -1.66412456
## as.factor(day)Sun -2.09108830 -1.86992687
## as.factor(day)Thu  0.16013898  0.26372399
## as.factor(day)Tue -0.03461414  0.07377192
## as.factor(day)Wed  0.24331689  0.34505602
```

Conclusion

In this analysis, I explored the factors influencing total commute days (tdrive) in a month, focusing on the effects of the day of the week. Initially, linear models were used to examine the relationship between the day of the week and commute time, but these models were discarded due to several limitations. The plots from the earlier models showed similar patterns, which was expected, as the random effect of id accounted for much of the variation in zip and commute distance. However, residual analysis revealed clear signs of heteroscedasticity, indicating that the assumption of homogeneous variance was not met in any of the models. Additionally, the Q-Q plots showed that residuals deviated significantly from a normal distribution, particularly in the tails, further suggesting that linear models were not appropriate for the data.

In contrast, the generalized linear mixed model (GLMM) with a Poisson distribution and a log link function proved more effective in capturing the relationship between commute days and the day of the week. By appropriately modeling both fixed effects (e.g., days of the week) and random effects (e.g., variations by id and zip), the GLMM provided a more accurate representation of the data, accounting for the count nature of commute times and repeated measures. This model revealed that certain weekdays, such as Monday, Wednesday, and Thursday, were associated with significantly higher commute times compared to weekends (Saturday and Sunday).

Ultimately, the GLMM approach offered a more reliable method for analyzing the data, overcoming the limitations of earlier linear models and providing better insights into how commute days are influenced by the day of the week.