App stat 2 assignment 1

1

a)

Given that $Y|\theta \sim \text{Poisson}(\mu\theta)$, for the conditional distribution of Y given θ, we have:

$E(Y|\theta) = \mu\theta, \text{Var}(Y|\theta) = \mu\theta$

To find the unconditional expectation and variance of (Y), we use the law of total expectation and the law of total variance:

$E(Y) = E(E(Y|\theta)), \text{Var}(Y) = E(\text{Var}(Y|\theta)) + \text{Var}(E(Y|\theta))$

Substituting the known values for $E(\theta) = 1$ and $\text{Var}(\theta) = \sigma^2$, we get:

$E(Y) = E(\mu\theta) = \mu E(\theta) = \mu$, $\text{Var}(Y) = E(\mu\theta) + \text{Var}(\mu\theta) = \mu E(\theta) + \mu^2\text{Var}(\theta) = \mu + \mu^2\sigma^2$

Thus, the expected value of Y is μ, and the variance of Y is $\mu(1+\mu\sigma^2)$, showing overdispersion relative to the Poisson distribution where the mean and variance are equal.

b)

Assume θ is Gamma distributed with α and β as shape and scale parameters, respectively. Show the unconditional distribution of Y is Negative Binomial.

$\theta \sim \text{Gamma}(\alpha, \beta)$

The moment generating function (MGF) of θ is:

$$M_\theta(t) = (1 - \beta t)^{-\alpha}$$

Given $Y|\theta$ is Poisson distributed, the MGF of Y given θ is:

$$M_{Y|\theta}(t) = \exp\{\mu\theta(e^t - 1)\}$$

To find the MGF of Y, we use the law of total expectation:

$$M_Y(t) = E[e^{tY}] = E[E[e^{tY}|\theta]] = E[\exp\{\mu\theta(e^t - 1)\}] = M_\theta(\mu(e^t - 1))$$

Substitute $M_\theta(t)$ into the above equation:

$$M_Y(t) = \left(\frac{1/(1+\beta\mu)}{1-(1-1/(1+\beta\mu))e^t}\right)^\alpha$$

This is the MGF of a Negative Binomial distribution, which shows that Y is Negative Binomially distributed. We denote this by:

$$Y \sim NB\left(\alpha, \frac{1}{1+\beta\mu}\right)$$

c)

Using the result $Y \sim NB\left(\alpha, \frac{1}{1+\beta}\right)$, in order for $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$, we need to find what α and β equal.

The mean and variance of the Negative Binomial distribution are given by:

$$E(Y) = \frac{\alpha(1 - \frac{1}{1+\beta\mu})}{\frac{1}{1+\beta\mu}} = \mu\alpha\beta$$

$$Var(Y) = \frac{\alpha(1 - \frac{1}{1+\beta\mu})}{\left(\frac{1}{1+\beta\mu}\right)^2} = \alpha\mu\beta(1 + \beta\mu)$$

By setting the mean and variance equal to μ and $\mu(1 + \mu\sigma^2)$ respectively, we have:

$$\mu = \mu\alpha\beta$$

$$\mu(1 + \mu\sigma^2) = \alpha\mu\beta(1 + \beta\mu)$$

Solving this, we get $\alpha = 1/\sigma^2$, $\beta = \sigma^2$

```r
# Load necessary packages
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

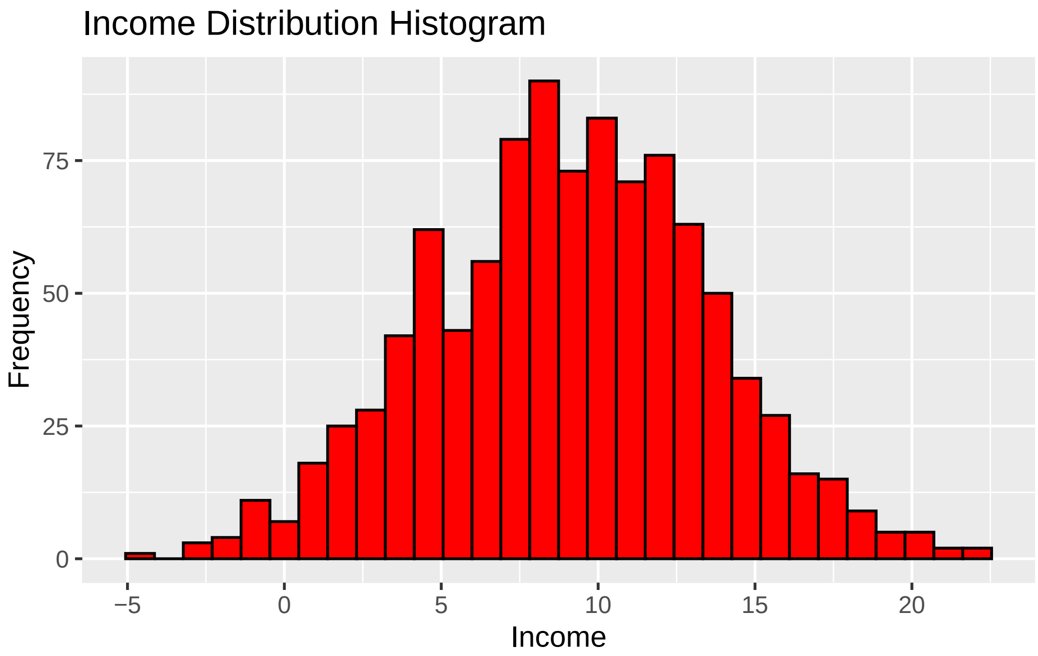
```r
library(janitor)
```

```

Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

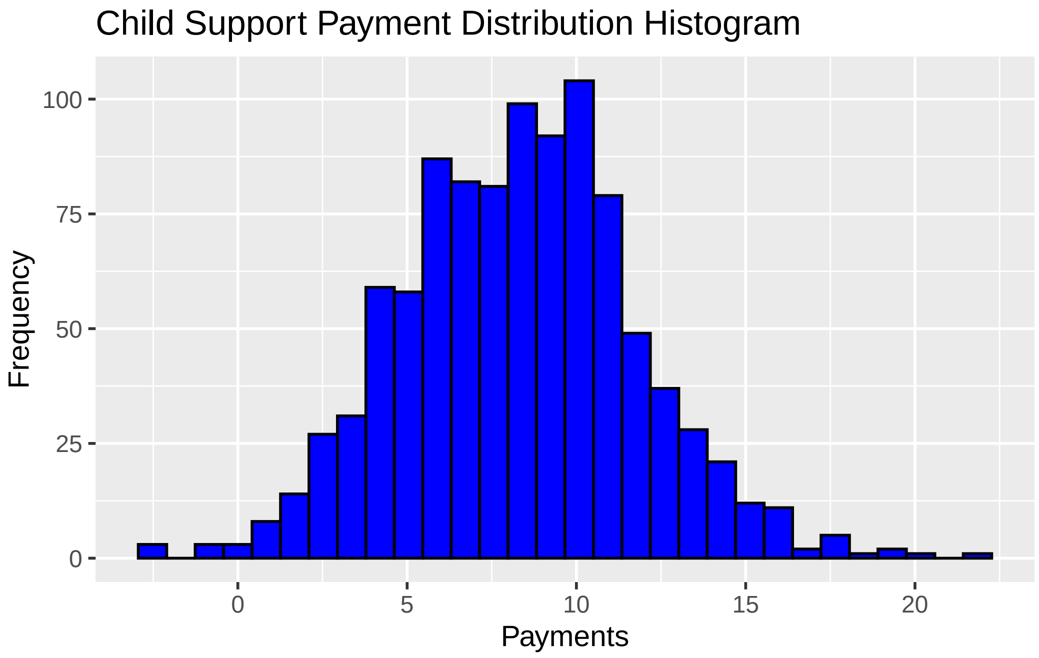
2

a)

```r
# Define sample size
n_samples <- 1000

# Generate log-normal distributed income
average_log_income <- log(10000)
std_dev_log_income <- log(100)
incomes <- rnorm(n_samples, average_log_income, std_dev_log_income)

# Generate log-normal distributed child support payments
average_log_payments <- log(3500)
std_dev_log_payments <- log(30)
child_support_payments <- rnorm(n_samples, average_log_payments, std_dev_log_payments)

# Plotting histograms for both datasets
# Income distribution
ggplot(data.frame(incomes), aes(x=incomes)) +
  geom_histogram(bins=30, fill="red", color="black") +
  labs(title="Income Distribution Histogram",
       x="Income",
       y="Frequency")
```



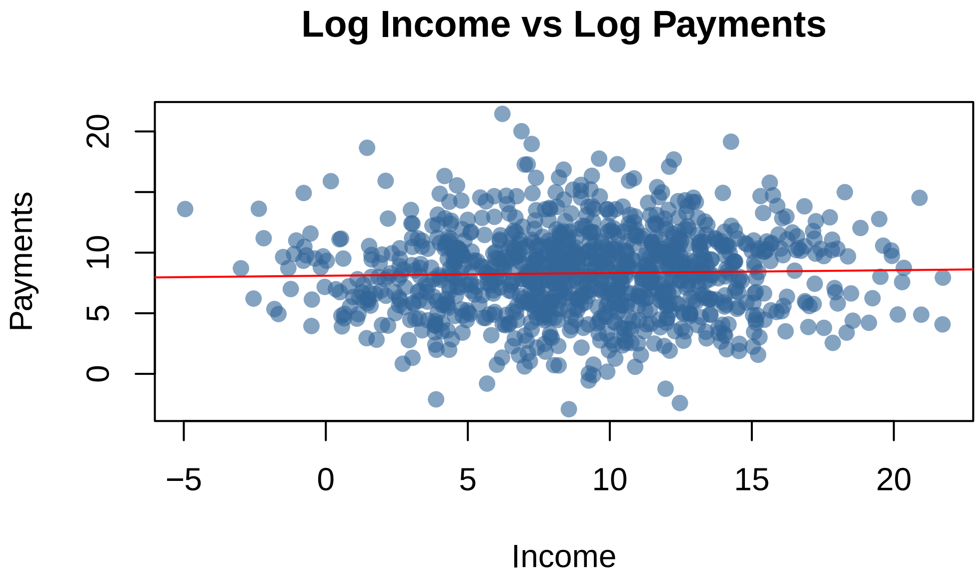
```
# Child support payment distribution
ggplot(data.frame(child_support_payments), aes(x=child_support_payments)) +
  geom_histogram(bins=30, fill="blue", color="black") +
  labs(title="Child Support Payment Distribution Histogram",
       x="Payments",
       y="Frequency")
```



b)

There is a red line that appears to be a line of best fit running horizontally across the plot, suggesting that there is no strong linear relationship between log income and log payments as the slope of the line is close to zero.

```
# Scatter plot of income vs payments
plot(incomes, child_support_payments, main="Log Income vs Log Payments",
     xlab="Income", ylab="Payments", pch=19, col=rgb(0.2,0.4,0.6,0.6))

# Linear regression model
linear_fit <- lm(child_support_payments ~ incomes)
abline(linear_fit, col="red")
```



c)

```
# Standardizing income and 3payments
z_score_income <- scale(incomes)
z_score_payments <- scale(child_support_payments)

# Simulating a survey based on z-scores and random noise
noise <- rnorm(n_samples)
survey <- (z_score_income + z_score_payments + noise) > 0

# Data aggregation and summarization
survey_data <- data.frame(incomes, child_support_payments, survey)
surveyed_means <- aggregate(cbind(incomes, child_support_payments) ~ survey, data=survey_d

# Displaying summary statistics
print(surveyed_means)
```

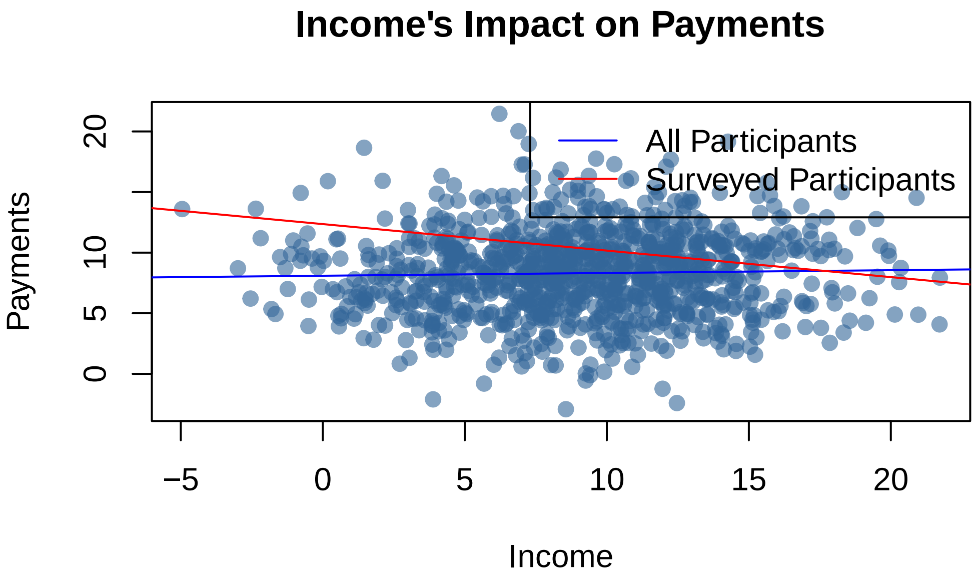
```
  survey   incomes child_support_payments
1  FALSE  6.915957               6.654362
2   TRUE 11.179609               9.911476
```

Surveying process, as modeled by the calculation, is indeed biased towards fathers with higher income and payment levels. In practice, this can mean that if the survey results were to be used to draw conclusions about the entire population of fathers, those conclusions could be skewed by this selection bias. It emphasizes the importance of understanding the sampling process when interpreting survey data.

d)

Since the slope of the red line (surveyed fathers) is steeper than the slope of the blue line (all participants), it suggests that the relationship between income and payments is stronger in the surveyed subset than in the entire population. This could be because of the selection bias in the survey.

```r
# Regression for the entire dataset
total_population_fit <- linear_fit

# Regression for surveyed individuals only
surveyed_subset <- survey_data[survey, ]
fit_for_surveyed <- lm(child_support_payments ~ incomes, data=surveyed_subset)

# Plotting the comparison
plot(incomes, child_support_payments, main="Income's Impact on Payments",
     xlab="Income", ylab="Payments", pch=19, col=rgb(0.2,0.4,0.6,0.6))

# Adding regression lines
abline(total_population_fit, col="blue")
abline(fit_for_surveyed, col="red")

# Including a legend
legend("topright", legend=c("All Participants", "Surveyed Participants"),
       col=c("blue", "red"), lwd=1)
```



e)

When survey data shows discrepancies like the sampling bias we have observed, it's a clear signal to researchers to examine the sampling process and to apply corrective measures in the analysis phase. This ensures that inferences drawn from the data are valid and applicable to the intended population.

3

a)

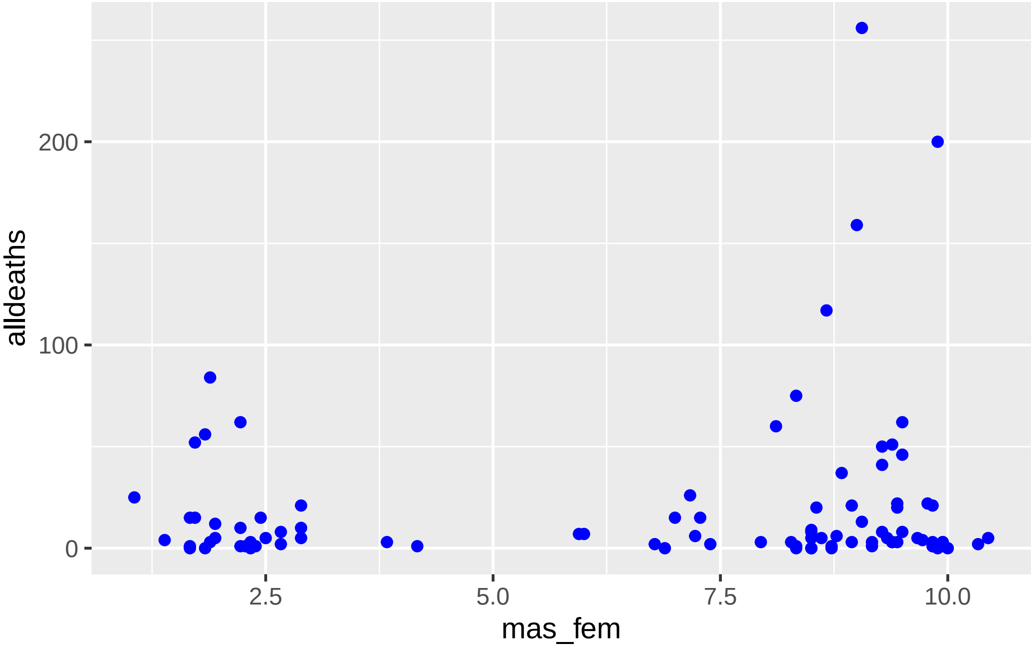
```r
library(readxl)
# Load the dataset
hurricane_data <- read_xlsx("pnas.1402786111.sd01.xlsx")
hurricane_data <- clean_names(hurricane_data)
head(hurricane_data)
```

```
# A tibble: 6 x 14
  year  name     mas_fem min_pressure_before minpressure_updated_2014 gender_mf
  <chr> <chr>      <dbl>               <dbl>                    <dbl>     <dbl>
1 1950  Easy        6.78                 958                      960         1
2 1950  King        1.39                 955                      955         0
3 1952  Able        3.83                 985                      985         0
```

```
4 1953  Barbara     9.83                    987                         987          1
5 1953  Florence    8.33                    985                         985          1
6 1954  Carol       8.11                    960                         960          1
# i 8 more variables: category <dbl>, alldeaths <dbl>, ndam <dbl>,
#   elapsed_yrs <dbl>, source <chr>, z_mas_fem <dbl>, z_min_pressure_a <dbl>,
#   zndam <dbl>
```

```
ggplot(hurricane_data, aes(x=mas_fem, y=alldeaths)) +
  geom_point(color="blue")
```

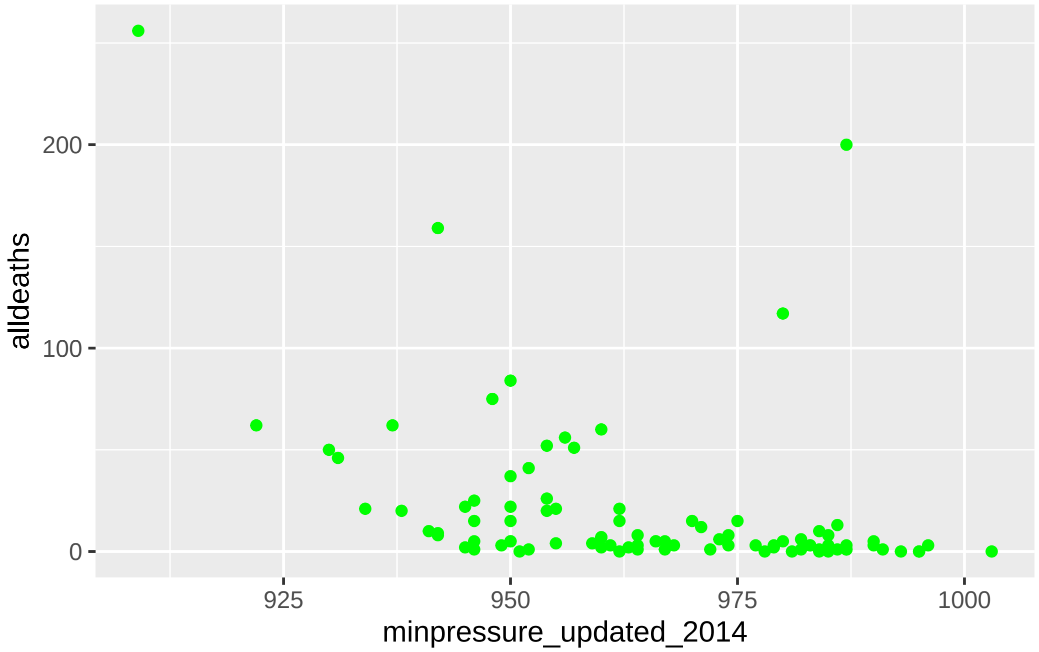
```
Warning: Removed 6 rows containing missing values (`geom_point()`).
```



Analysis reveals that hurricanes with female names show a greater variance in the number of death compared to those with male names. Hurricanes with neutral names demonstrate the lowest fatalities.

```
ggplot(hurricane_data, aes(x=minpressure_updated_2014, y=alldeaths)) +
  geom_point(color="green")
```

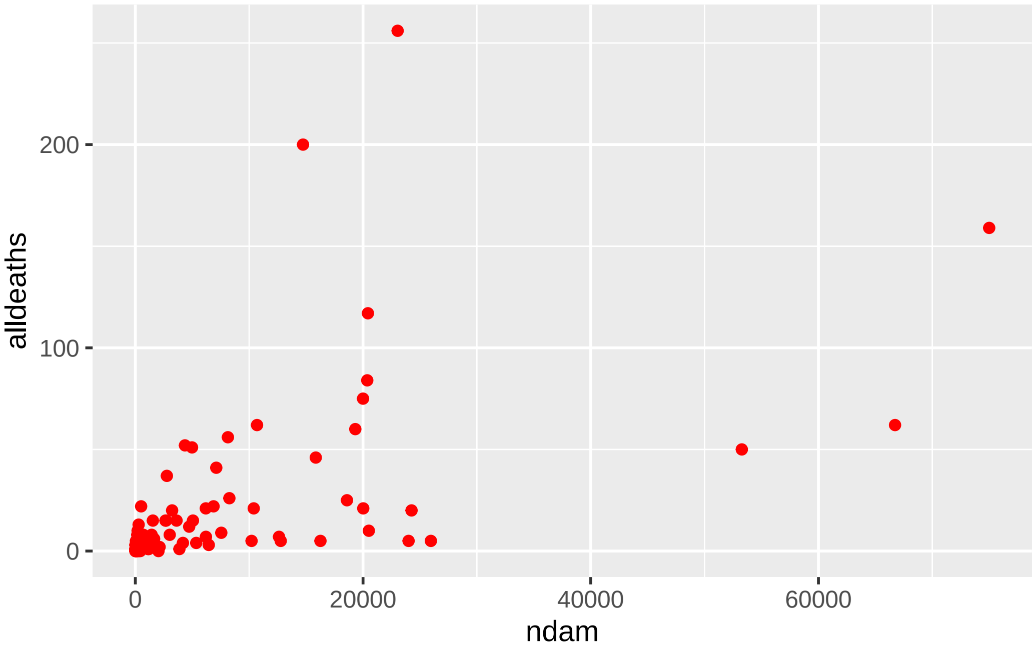
```
Warning: Removed 6 rows containing missing values (`geom_point()`).
```



A trend is observed where hurricanes with lower minimum pressures result in higher fatalities. This is consistent with our understanding that lower pressure indicates a stronger hurricane.

```r
ggplot(hurricane_data, aes(x=ndam, y=alldeaths)) +
  geom_point(color="red")
```

```
Warning: Removed 6 rows containing missing values (`geom_point()`).
```



There is a correlation between higher normalized damage and increased fatalities. This aligns with the idea that normalized damage is a indicator for the fatalities of the hurricane.

b)

The model shows the positive relationship between expected number of death for each unit and the Masculinity-Femininity Index (MFI), suggesting that hurricanes with more feminine names may be more deadly. Coefficient can be interpreted in a way that if each increase of the MFI, the deaths will increase multiplicatively by $e^{0.07387}$.

```
pmodel <- glm(alldeaths ~ mas_fem, data = hurricane_data, family = poisson)
summary(pmodel)$coefficients
```

```
               Estimate  Std. Error   z value     Pr(>|z|)
(Intercept) 2.50036961 0.063297108 39.502114 0.000000e+00
mas_fem     0.07387252 0.007890654  9.362028 7.822065e-21
```

```
# Checking for overdispersion
resid_deviance <- sum(residuals(pmodel, type = "deviance")^2)
df_resid <- pmodel$df.residual
overdispersion_ratio <- resid_deviance / df_resid
```

```
# Display the overdispersion ratio, shows overdispersion since 43 > 1
overdispersion_ratio
```

```
[1] 43.75037
```

```
qpmodel <- glm(alldeaths ~ mas_fem, data=hurricane_data, family = quasipoisson)
summary(qpmodel)
```

```

Call:
glm(formula = alldeaths ~ mas_fem, family = quasipoisson, data = hurricane_data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.50037    0.54371   4.599 1.38e-05 ***
mas_fem      0.07387    0.06778   1.090    0.279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 73.78496)

    Null deviance: 4031.9  on 91  degrees of freedom
Residual deviance: 3937.5  on 90  degrees of freedom
  (6 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 6
```

The standard error of the coefficients experienced a significant increase, and the p-value for 'mas_fem' also rose sharply. This change led to 'mas_fem' becoming statistically insignificant, suggesting that the model no longer shows a connection between this index and the number of deaths.

c)

```
library(MASS)
```

```

Attaching package: 'MASS'


The following object is masked _by_ '.GlobalEnv':

    survey


The following object is masked from 'package:dplyr':

    select
```

```
model4 <- glm.nb(alldeaths ~ z_min_pressure_a + zndam + z_mas_fem + z_mas_fem*z_min_pressu
coeffi <- model4$coefficients
summary(model4)$coefficients
```

```
                             Estimate Std. Error   z value     Pr(>|z|)
(Intercept)                 2.4756224  0.1221895 20.260512 2.869655e-91
z_min_pressure_a           -0.5520594  0.1502841 -3.673440 2.393071e-04
zndam                       0.8634616  0.1444778  5.976430 2.280807e-09
z_mas_fem                   0.1723215  0.1238134  1.391784 1.639878e-01
z_min_pressure_a:z_mas_fem  0.3947513  0.1521066  2.595228 9.452809e-03
zndam:z_mas_fem             0.7050953  0.1500591  4.698782 2.617174e-06
```

```
med_pressure <- median(hurricane_data$z_min_pressure_a, na.rm = TRUE)
med_dam <- median(hurricane_data$zndam, na.rm = TRUE)

c(med_pressure, med_dam)
```

```
[1] -0.07239403 -0.43449397
```

```
coeffi["z_mas_fem"] + coeffi["z_min_pressure_a:z_mas_fem"]*med_pressure + coeffi["zndam:z_
```

```
 z_mas_fem
-0.1626157
```

Using median values for minimum pressure and damage, the model suggests that with each unit increase in MFI, expected fatalities decrease by $1 - e^{-0.1626429} \approx 15.01%$, indicating that more masculine-named hurricanes could potentially be more fatal.

d)

```r
hurricane_data <- na.omit(hurricane_data)

new_data <- data.frame(
  z_min_pressure_a =  hurricane_data[hurricane_data$name == "Sandy", "z_min_pressure_a"],
  z_mas_fem = hurricane_data[hurricane_data$name == "Sandy", "z_mas_fem"],
  zndam = hurricane_data[hurricane_data$name == "Sandy", "zndam"]
)
new_data
```

```
  z_min_pressure_a z_mas_fem    zndam
1        -1.027546 0.6876512 5.236567
```

```r
deaths_pred <- predict(model4, newdata = new_data, type = "response")
deaths_pred
```

```
       1
20806.74
```

```r
mean(hurricane_data$alldeaths)
```

```
[1] 20.65217
```

For the specific case of Hurricane Sandy the model predicts the number of death of 20,807. This is inconsistent with the actual number of death of 159. It has low prediction performance.

e)

Strengths:

- The study spans 62 years (1950-2012), offering a comprehensive dataset of hurricanes with varying severity and perceived gender names.
- Utilizing negative binomial regression for data analysis is appropriate given the overdispersed count nature of the data.

Weaknesses:

- The gender perception of hurricane names can be subjective. Surveying more individuals to create a robust Masculinity-Femininity Index (MFI) would have been beneficial. Moreover, limiting the study to U.S. hurricanes restricts its global applicability.
- The dataset is not extensive enough to conclusively establish a link between hurricane fatality and gendered names. Many hurricanes, regardless of their perceived gender, have similar death counts, with only a few female-named storms showing higher fatalities. Important factors like hurricane path, width, and the population density of affected areas were not adequately considered. A more comprehensive approach would involve analyzing the proportionate impact of each hurricane.

f)

The study's result is intriguing but not entirely convincing in its current form. The limited geographic focus on U.S. hurricanes, coupled with a dataset that might not adequately represent the full spectrum of hurricanes, raises questions about the robustness of the findings. Crucial factors like population density, the path and intensity of hurricanes, and evolving public response mechanisms over time were not sufficiently accounted for, potentially skewing the results. The predictive inaccuracy of the model, particularly in cases like Hurricane Sandy, also undermines the study's reliability. To enhance the testing of the hypothesis that hurricane names influence fatality rates, a multifaceted approach incorporating diverse data and analyses is essential. Global hurricane data would provide a broader perspective, assessing if the observed effects are consistent across different cultures.

4

a)

```r
# Data load
immigrant_data <- read.csv("98100468.csv")

head(immigrant_data)
```

```
  REF_DATE    GEO           DGUID    Visible.minority..15.   Age..15A.
1     2021 Canada 2021A000011124 Total - Visible minority Total - Age
2     2021 Canada 2021A000011124 Total - Visible minority Total - Age
3     2021 Canada 2021A000011124 Total - Visible minority Total - Age
4     2021 Canada 2021A000011124 Total - Visible minority Total - Age
5     2021 Canada 2021A000011124 Total - Visible minority Total - Age
6     2021 Canada 2021A000011124 Total - Visible minority Total - Age
```

```
     Gender..3. Statistics..3.             Main.mode.of.commuting..11A.
1 Total - Gender          Count            Total - Main mode of commuting
2 Total - Gender          Count                           Car, truck or van
3 Total - Gender          Count           Driver (only worker in vehicle)
4 Total - Gender          Count        Passenger (only worker in vehicle)
5 Total - Gender          Count 2 or more persons shared the ride to work
6 Total - Gender          Count              Driver with 1 or more workers
   Coordinate
1 1.1.1.1.1.1
2 1.1.1.1.1.2
3 1.1.1.1.1.3
4 1.1.1.1.1.4
5 1.1.1.1.1.5
6 1.1.1.1.1.6
  Immigrant.status.and.period.of.immigration..11..Total...Immigrant.status.and.period.of.immi
1
2
3
4
5
6
  Symbol Immigrant.status.and.period.of.immigration..11..Non.immigrants.2.
1                                                                  9420575
2                                                                  8176110
3                                                                  7250370
4                                                                   292675
5                                                                   633070
6                                                                   358505
  Symbol.1 Immigrant.status.and.period.of.immigration..11..Immigrants.3.
1                                                               3205590
2                                                               2549400
3                                                               2150105
4                                                                115025
5                                                                284270
6                                                                162155
  Symbol.2 Immigrant.status.and.period.of.immigration..11..Before.1980.4.
1                                                                  302620
2                                                                  256925
3                                                                  231705
4                                                                    6570
5                                                                   18655
6                                                                   12400
  Symbol.3 Immigrant.status.and.period.of.immigration..11..1980.to.1990.5.
```

```
1                                                                   357515
2                                                                   304290
3                                                                   266580
4                                                                     9295
5                                                                    28415
6                                                                    18535
  Symbol.4 Immigrant.status.and.period.of.immigration..11..1991.to.2000.6.
1                                                                   658175
2                                                                   545640
3                                                                   474020
4                                                                    19100
5                                                                    52520
6                                                                    32700
  Symbol.5 Immigrant.status.and.period.of.immigration..11..2001.to.2010.7.
1                                                                   894415
2                                                                   714400
3                                                                   603070
4                                                                    33150
5                                                                    78180
6                                                                    45160
  Symbol.6 Immigrant.status.and.period.of.immigration..11..2011.to.2021.8.
1                                                                   992860
2                                                                   728140
3                                                                   574735
4                                                                    46920
5                                                                   106495
6                                                                    53355
  Symbol.7 Immigrant.status.and.period.of.immigration..11..2011.to.2015.9.
1                                                                   497625
2                                                                   383760
3                                                                   310360
4                                                                    22445
5                                                                    50955
6                                                                    27035
  Symbol.8 Immigrant.status.and.period.of.immigration..11..2016.to.2021.10.
1                                                                    495235
2                                                                    344390
3                                                                    264375
4                                                                     24470
5                                                                     55540
6                                                                     26320
  Symbol.9
1
```

```
2
3
4
5
6
  Immigrant.status.and.period.of.immigration..11..Non.permanent.residents.11.
1                                                                      422340
2                                                                      225240
3                                                                      165990
4                                                                       15025
5                                                                       44225
6                                                                       19475
  Symbol.10
1
2
3
4
5
6
```

```r
library(tidyverse)

# Cleaning column names and removing specific columns
immigrant_data <- clean_names(immigrant_data)
drops <- c('symbol','symbol_1','symbol_2','symbol_3','symbol_4','symbol_5','symbol_6','sym
immigrant_data <- immigrant_data[ , !(names(immigrant_data) %in% drops)]
head(immigrant_data)
```

```
     geo          dguid      visible_minority_15     age_15a       gender_3
1 Canada 2021A000011124 Total - Visible minority Total - Age Total - Gender
2 Canada 2021A000011124 Total - Visible minority Total - Age Total - Gender
3 Canada 2021A000011124 Total - Visible minority Total - Age Total - Gender
4 Canada 2021A000011124 Total - Visible minority Total - Age Total - Gender
5 Canada 2021A000011124 Total - Visible minority Total - Age Total - Gender
6 Canada 2021A000011124 Total - Visible minority Total - Age Total - Gender
  statistics_3                 main_mode_of_commuting_11a  coordinate
1        Count            Total - Main mode of commuting 1.1.1.1.1.1
2        Count                         Car, truck or van 1.1.1.1.1.2
3        Count           Driver (only worker in vehicle) 1.1.1.1.1.3
4        Count        Passenger (only worker in vehicle) 1.1.1.1.1.4
5        Count 2 or more persons shared the ride to work 1.1.1.1.1.5
```

```
6        Count             Driver with 1 or more workers 1.1.1.1.1.6
  immigrant_status_and_period_of_immigration_11_total_immigrant_status_and_period_of_immigra
1                                                                                         130
2                                                                                         109
3                                                                                          95
4                                                                                           4
5                                                                                           9
6                                                                                           5
  immigrant_status_and_period_of_immigration_11_non_immigrants_2
1                                                        9420575
2                                                        8176110
3                                                        7250370
4                                                         292675
5                                                         633070
6                                                         358505
  immigrant_status_and_period_of_immigration_11_immigrants_3
1                                                    3205590
2                                                    2549400
3                                                    2150105
4                                                     115025
5                                                     284270
6                                                     162155
  immigrant_status_and_period_of_immigration_11_before_1980_4
1                                                      302620
2                                                      256925
3                                                      231705
4                                                        6570
5                                                       18655
6                                                       12400
  immigrant_status_and_period_of_immigration_11_1980_to_1990_5
1                                                       357515
2                                                       304290
3                                                       266580
4                                                         9295
5                                                        28415
6                                                        18535
  immigrant_status_and_period_of_immigration_11_1991_to_2000_6
1                                                       658175
2                                                       545640
3                                                       474020
4                                                        19100
5                                                        52520
6                                                        32700
```

```
  immigrant_status_and_period_of_immigration_11_2001_to_2010_7
1                                                       894415
2                                                       714400
3                                                       603070
4                                                        33150
5                                                        78180
6                                                        45160
  immigrant_status_and_period_of_immigration_11_2011_to_2021_8
1                                                       992860
2                                                       728140
3                                                       574735
4                                                        46920
5                                                       106495
6                                                        53355
  immigrant_status_and_period_of_immigration_11_2011_to_2015_9
1                                                       497625
2                                                       383760
3                                                       310360
4                                                        22445
5                                                        50955
6                                                        27035
  immigrant_status_and_period_of_immigration_11_2016_to_2021_10
1                                                        495235
2                                                        344390
3                                                        264375
4                                                         24470
5                                                         55540
6                                                         26320
  immigrant_status_and_period_of_immigration_11_non_permanent_residents_11
1                                                                   422340
2                                                                   225240
3                                                                   165990
4                                                                    15025
5                                                                    44225
6                                                                    19475
```

```
# Renaming specific columns
detach("package:MASS", unload=TRUE)
immigrant_data <- immigrant_data |>
    select(region = geo,
           minority_group = visible_minority_15,
           age = age_15a,
           gender = gender_3,
           statistics = statistics_3,
           commute = main_mode_of_commuting_11a,
           non_immigrants = immigrant_status_and_period_of_immigration_11_non_immigrants_2
           immigrants = immigrant_status_and_period_of_immigration_11_immigrants_3,
           non_permanent_residents = immigrant_status_and_period_of_immigration_11_non_per


# Filter for Census Metropolitan Areas and remove 'Total' categories
immigrant_data <- immigrant_data |>
  filter(grepl("(CMA)", region)) |>
  filter(!grepl("Total", minority_group) &
         !grepl("Total", age) &
         !grepl("Total", gender)) |>
  filter(statistics == "Count") |>
  filter(commute %in%
           c("Car, truck or van", "Public transit", "Active transportation", "Other method

# Create categories
immigrant_data$immigrants_total <- immigrant_data$immigrants + immigrant_data$non_permanen
immigrant_data <- immigrant_data |>
  select(
    region,
    minority_group,
    age,
    gender,
    commute,
    non_immigrants,
    total_immigrants = immigrants_total
  )

immigrant_modified <- immigrant_data |>
  mutate(transportation_type = ifelse(commute == "Public transit", "Public", "Non-Public")
  group_by(region, minority_group, age, gender, transportation_type) |>
  summarise(
    native_count = sum(non_immigrants),
    immigrant_count = sum(total_immigrants)
  )
```

```
`summarise()` has grouped output by 'region', 'minority_group', 'age',
'gender'. You can override using the `.groups` argument.
```

```r
# Pivoting and rearranging the dataset 'immigrant_long_form'
immigrant_long_form <- immigrant_modified |>
  pivot_longer(
    cols = c('immigrant_count', 'native_count'),
    names_to = "population_type",
    values_to = "population_count"
  ) |>
  pivot_wider(
    names_from = "transportation_type",
    values_from = "population_count"
  )

# Renaming columns and calculating total population
commuting_data <- immigrant_long_form |>
  rename(public_transport = Public, other_transport = `Non-Public`)
commuting_data$total_population = commuting_data$public_transport + commuting_data$other_t

# Display the final dataset
commuting_data
```

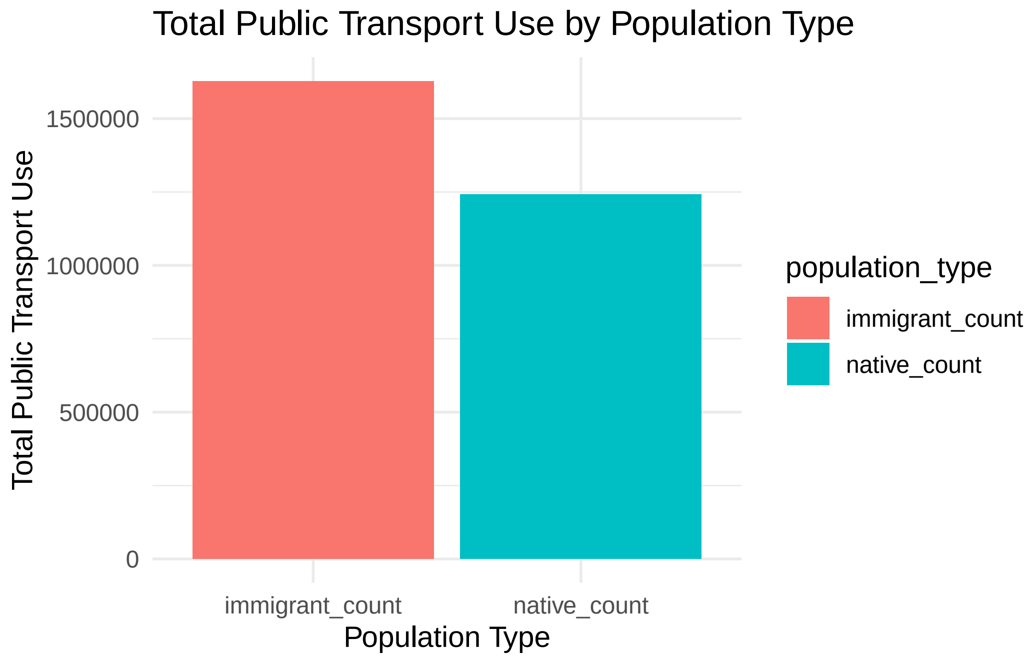
```
# A tibble: 31,304 x 8
# Groups:   region, minority_group, age, gender [15,652]
   region             minority_group age   gender population_type other_transport
   <chr>              <chr>          <chr> <chr>  <chr>                     <int>
 1 Abbotsford - Mis~ Arab           15 t~ Men+   immigrant_count              20
 2 Abbotsford - Mis~ Arab           15 t~ Men+   native_count                  0
 3 Abbotsford - Mis~ Arab           15 t~ Women+ immigrant_count              10
 4 Abbotsford - Mis~ Arab           15 t~ Women+ native_count                  0
 5 Abbotsford - Mis~ Arab           15 t~ Men+   immigrant_count              35
 6 Abbotsford - Mis~ Arab           15 t~ Men+   native_count                 10
 7 Abbotsford - Mis~ Arab           15 t~ Women+ immigrant_count              20
 8 Abbotsford - Mis~ Arab           15 t~ Women+ native_count                  0
 9 Abbotsford - Mis~ Arab           20 t~ Men+   immigrant_count              10
10 Abbotsford - Mis~ Arab           20 t~ Men+   native_count                  0
# i 31,294 more rows
# i 2 more variables: public_transport <int>, total_population <int>
```

b)

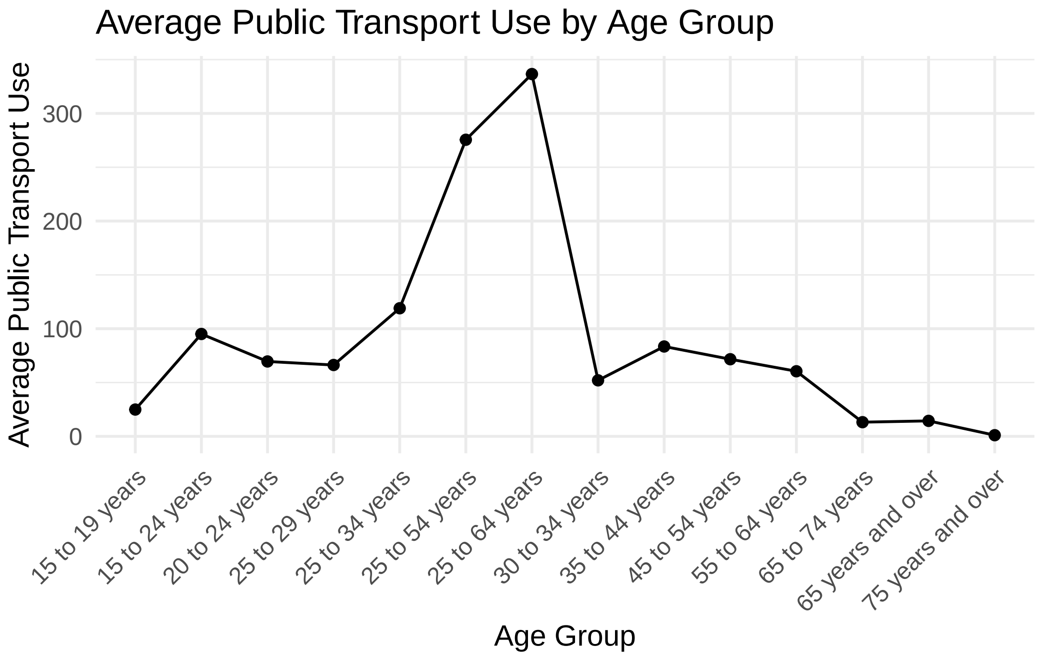
- Total Public Transport Use by Population Type: The bar chart indicates that immigrants have a higher total use of public transport compared to non-immigrants.

- Average Public Transport Use by Age Group: The line plot reveals that the average public transport use is higher among the middle age groups, with a peak in the 25 to 64 years category.
- Average Public Transport Use by Gender: There's a notable decline in public transport use among the older age group of 65 years and over. The bar chart demonstrates that on average, the Women+ category uses public transport more than the Men+ category.

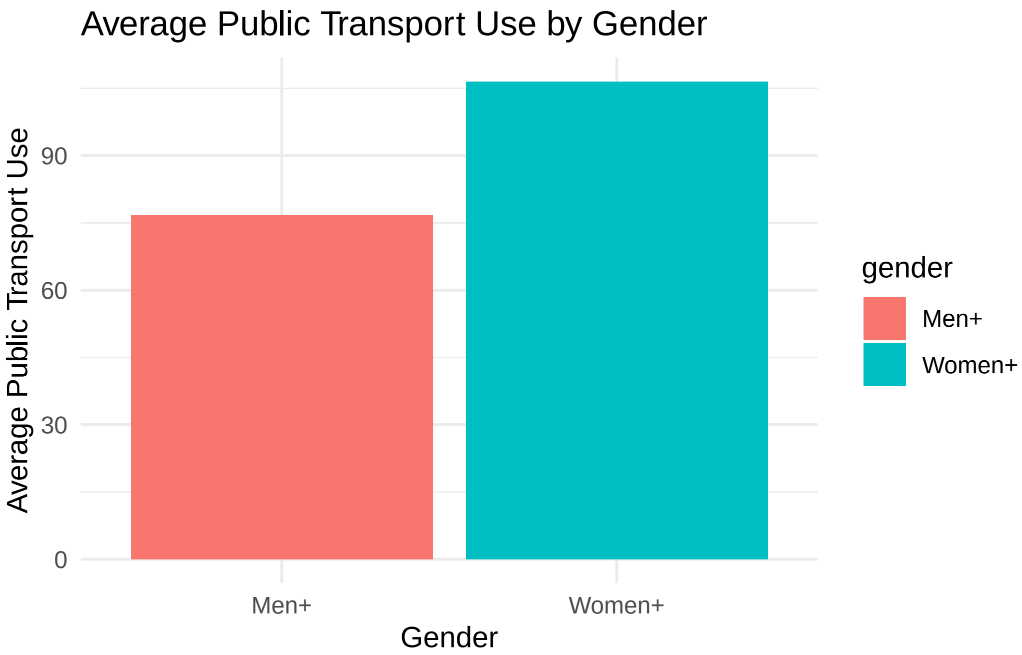
```
# Total Public Transport Use by Population Type
public_transport_by_type <- commuting_data %>%
  group_by(population_type) %>%
  summarise(TotalPublicTransport = sum(public_transport))

ggplot(public_transport_by_type, aes(x = population_type, y = TotalPublicTransport, fill =
  geom_col() +
  labs(title = "Total Public Transport Use by Population Type",
       x = "Population Type",
       y = "Total Public Transport Use") +
  theme_minimal()
```



```
# Average Public Transport Use by Age Group
public_transport_by_age <- commuting_data %>%
  group_by(age) %>%
  summarise(AveragePublicTransport = mean(public_transport))

ggplot(public_transport_by_age, aes(x = age, y = AveragePublicTransport)) +
  geom_line(group = 1) +
  geom_point() +
  labs(title = "Average Public Transport Use by Age Group",
       x = "Age Group",
       y = "Average Public Transport Use") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# Average Public Transport Use by Gender
public_transport_by_gender <- commuting_data %>%
  group_by(gender) %>%
  summarise(AveragePublicTransport = mean(public_transport))

ggplot(public_transport_by_gender, aes(x = gender, y = AveragePublicTransport, fill = gend
  geom_col() +
  labs(title = "Average Public Transport Use by Gender",
       x = "Gender",
          y = "Average Public Transport Use") +
    theme_minimal()
```



c)

The outcome variable is public_transport, which represents the count of individuals using public transportation. A count data outcome is appropriate for Poisson or Quasi-Poisson regression models, as these models are designed for non-negative integer outcomes. The initial assumption was that the outcome variable followed a Poisson distribution, which assumes the mean and variance of the count data are equal. However, the overdispersion ratio (residual deviance divided by the degrees of freedom) was found to be significantly greater than 1, indicating overdispersion. This justifies the use of a Quasi-Poisson model, which relaxes the Poisson assumption by allowing the variance to be a function of the mean. The model includes covariates such as region, age, gender, population type (tentative count), and minority group. These covariates were likely chosen based on exploratory data analysis (EDA) findings that suggested these factors could influence the use of public transport. For example, different age groups may have different commuting patterns, gender may influence transportation preferences or needs, population type could capture the differences between immigrants and non-immigrants, and minority group status may reflect socio-economic factors affecting public transport usage.

```
poisson_model <- glm(public_transport ~ region + age + gender + population_type + minority
                     family = "poisson",
                     data = commuting_data)

# Checking for overdispersion
resid_deviance <- sum(residuals(poisson_model, type = "deviance")^2)
df_resid <- poisson_model$df.residual
overdispersion_ratio <- resid_deviance / df_resid

# Display the overdispersion ratio, shows overdispersion since 85 > 1
overdispersion_ratio
```

```
[1] 85.04057
```

```
quasi_poisson_model <- glm(public_transport ~ age + gender + population_type + minority_gr
                           family = "quasipoisson",
                           data = commuting_data)

# Summary of the Quasi-Poisson model
summary(quasi_poisson_model)
```

```

Call:
glm(formula = public_transport ~ age + gender + population_type + 
    minority_group + region, family = "quasipoisson", data = commuting_data)

Coefficients:
                                                   Estimate Std. Error t value
(Intercept)                                        -0.44265    0.15143  -2.923
age15 to 24 years                                   1.34071    0.04689  28.592
age20 to 24 years                                   1.02799    0.04864  21.133
age25 to 29 years                                   0.97981    0.04896  20.013
age25 to 34 years                                   1.56496    0.04590  34.092
age25 to 54 years                                   2.40474    0.04359  55.167
age25 to 64 years                                   2.60478    0.04326  60.210
age30 to 34 years                                   0.73771    0.05076  14.534
age35 to 44 years                                   1.21046    0.04756  25.450
age45 to 54 years                                   1.05764    0.04846  21.827
age55 to 64 years                                   0.88803    0.04960  17.905
age65 to 74 years                                  -0.63671    0.07097  -8.971
```

```
age65 years and over                               -0.54924    0.06900  -7.960
age75 years and over                               -3.15373    0.20631 -15.286
genderWomen+                                        0.32829    0.01178  27.861
population_typenative_count                        -0.27092    0.01173 -23.090
minority_groupBlack                                 1.49391    0.03791  39.404
minority_groupChinese                               0.66079    0.04219  15.662
minority_groupFilipino                              1.15853    0.03927  29.499
minority_groupJapanese                             -1.98159    0.09843 -20.131
minority_groupKorean                               -1.02325    0.06663 -15.357
minority_groupLatin American                        0.45361    0.04381  10.353
minority_groupMultiple visible minorities          -0.72613    0.06000 -12.102
minority_groupNot a visible minority                2.69693    0.03540  76.190
minority_groupSouth Asian                           1.65434    0.03739  44.240
minority_groupSoutheast Asian                      -0.33550    0.05306  -6.323
minority_groupVisible minority, n.i.e.             -1.19701    0.07113 -16.828
minority_groupWest Asian                           -0.64308    0.05837 -11.018
regionBarrie (CMA), Ont.                            0.16777    0.19217   0.873
regionBelleville - Quinte West (CMA), Ont.         -1.01263    0.27403  -3.695
regionBrantford (CMA), Ont.                        -0.49107    0.22958  -2.139
regionCalgary (CMA), Alta.                          3.18247    0.14436  22.045
regionChilliwack (CMA), B.C.                       -1.28528    0.30389  -4.229
regionDrummondville (CMA), Que.                    -1.44254    0.32347  -4.460
regionEdmonton (CMA), Alta.                         2.97176    0.14503  20.490
regionFredericton (CMA), N.B.                      -0.67982    0.24392  -2.787
regionGreater Sudbury (CMA), Ont.                   0.34265    0.18497   1.852
regionGuelph (CMA), Ont.                            0.26814    0.18792   1.427
regionHalifax (CMA), N.S.                           2.00682    0.15066  13.320
regionHamilton (CMA), Ont.                          2.26580    0.14861  15.246
regionKamloops (CMA), B.C.                         -0.02402    0.20126  -0.119
regionKelowna (CMA), B.C.                           0.40029    0.18281   2.190
regionKingston (CMA), Ont.                          0.40581    0.18261   2.222
regionKitchener - Cambridge - Waterloo (CMA), Ont.  1.68360    0.15403  10.930
regionLethbridge (CMA), Alta.                      -0.71295    0.24664  -2.891
regionLondon (CMA), Ont.                            1.65167    0.15442  10.696
regionMoncton (CMA), N.B.                          -0.34977    0.22000  -1.590
regionMontréal (CMA), Que.                          4.95704    0.14195  34.920
regionNanaimo (CMA), B.C.                          -0.20804    0.21130  -0.985
regionOshawa (CMA), Ont.                            1.40445    0.15787   8.896
regionOttawa - Gatineau (CMA), Ont./Que.            3.19147    0.14433  22.112
regionOttawa - Gatineau (Ontario part) (CMA), Ont.  2.89495    0.14532  19.922
regionOttawa - Gatineau (Quebec part) (CMA), Que.   1.81118    0.15258  11.870
regionPeterborough (CMA), Ont.                     -0.63990    0.24074  -2.658
regionQuébec (CMA), Que.                            2.65652    0.14634  18.153
```

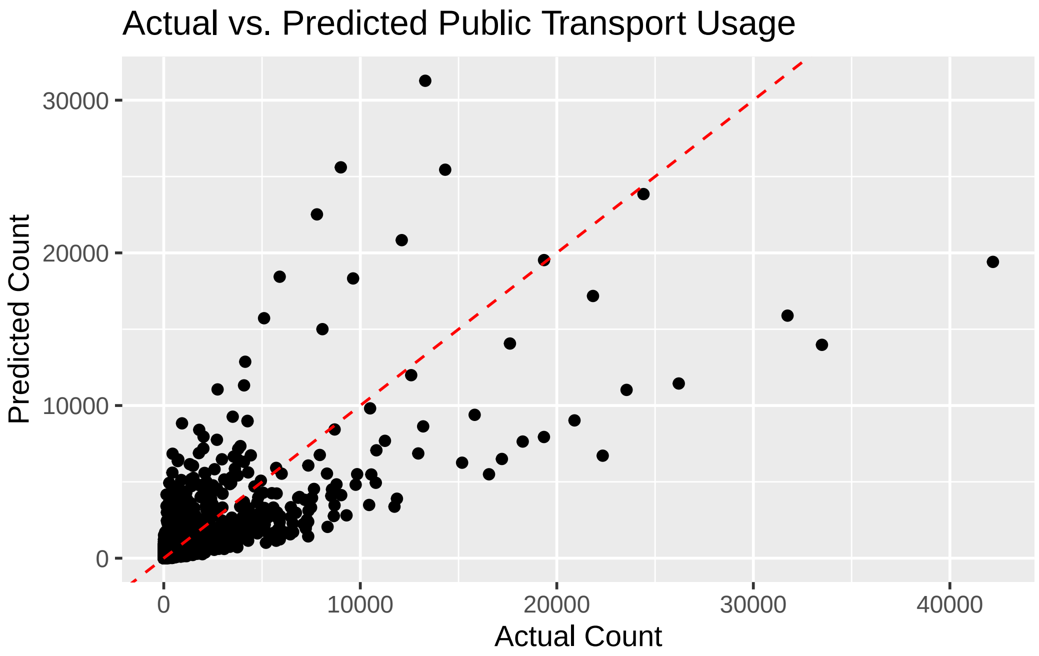
```
regionRed Deer (CMA), Alta.                    -0.63404    0.24028  -2.639
regionRegina (CMA), Sask.                       0.56094    0.17728   3.164
regionSaguenay (CMA), Que.                     -0.52016    0.23163  -2.246
regionSaint John (CMA), N.B.                   -0.32376    0.21833  -1.483
regionSaskatoon (CMA), Sask.                    0.73261    0.17213   4.256
regionSherbrooke (CMA), Que.                    0.62537    0.17526   3.568
regionSt. Catharines - Niagara (CMA), Ont.      0.34338    0.18494   1.857
regionSt. John's (CMA), N.L.                    0.38074    0.18353   2.075
regionThunder Bay (CMA), Ont.                  -0.20044    0.21085  -0.951
regionToronto (CMA), Ont.                       5.16317    0.14186  36.396
regionTrois-Rivières (CMA), Que.               -0.41793    0.22449  -1.862
regionVancouver (CMA), B.C.                     4.47533    0.14226  31.459
regionVictoria (CMA), B.C.                      1.86183    0.15205  12.245
regionWindsor (CMA), Ont.                       0.39126    0.18314   2.136
regionWinnipeg (CMA), Man.                      2.77709    0.14579  19.049
                                               Pr(>|t|)
(Intercept)                                     0.00347 **
age15 to 24 years                               < 2e-16 ***
age20 to 24 years                               < 2e-16 ***
age25 to 29 years                               < 2e-16 ***
age25 to 34 years                               < 2e-16 ***
age25 to 54 years                               < 2e-16 ***
age25 to 64 years                               < 2e-16 ***
age30 to 34 years                               < 2e-16 ***
age35 to 44 years                               < 2e-16 ***
age45 to 54 years                               < 2e-16 ***
age55 to 64 years                               < 2e-16 ***
age65 to 74 years                               < 2e-16 ***
age65 years and over                           1.78e-15 ***
age75 years and over                            < 2e-16 ***
genderWomen+                                    < 2e-16 ***
population_typenative_count                     < 2e-16 ***
minority_groupBlack                             < 2e-16 ***
minority_groupChinese                           < 2e-16 ***
minority_groupFilipino                          < 2e-16 ***
minority_groupJapanese                          < 2e-16 ***
minority_groupKorean                            < 2e-16 ***
minority_groupLatin American                    < 2e-16 ***
minority_groupMultiple visible minorities       < 2e-16 ***
minority_groupNot a visible minority            < 2e-16 ***
minority_groupSouth Asian                       < 2e-16 ***
minority_groupSoutheast Asian                  2.60e-10 ***
minority_groupVisible minority, n.i.e.          < 2e-16 ***
```

```
minority_groupWest Asian                            < 2e-16 ***
regionBarrie (CMA), Ont.                            0.38266    
regionBelleville - Quinte West (CMA), Ont.          0.00022 ***
regionBrantford (CMA), Ont.                         0.03244 *  
regionCalgary (CMA), Alta.                          < 2e-16 ***
regionChilliwack (CMA), B.C.                       2.35e-05 ***
regionDrummondville (CMA), Que.                    8.24e-06 ***
regionEdmonton (CMA), Alta.                         < 2e-16 ***
regionFredericton (CMA), N.B.                       0.00532 ** 
regionGreater Sudbury (CMA), Ont.                   0.06397 .  
regionGuelph (CMA), Ont.                            0.15362    
regionHalifax (CMA), N.S.                           < 2e-16 ***
regionHamilton (CMA), Ont.                          < 2e-16 ***
regionKamloops (CMA), B.C.                          0.90499    
regionKelowna (CMA), B.C.                           0.02856 *  
regionKingston (CMA), Ont.                          0.02627 *  
regionKitchener - Cambridge - Waterloo (CMA), Ont.  < 2e-16 ***
regionLethbridge (CMA), Alta.                       0.00385 ** 
regionLondon (CMA), Ont.                            < 2e-16 ***
regionMoncton (CMA), N.B.                           0.11187    
regionMontréal (CMA), Que.                          < 2e-16 ***
regionNanaimo (CMA), B.C.                           0.32485    
regionOshawa (CMA), Ont.                            < 2e-16 ***
regionOttawa - Gatineau (CMA), Ont./Que.            < 2e-16 ***
regionOttawa - Gatineau (Ontario part) (CMA), Ont.  < 2e-16 ***
regionOttawa - Gatineau (Quebec part) (CMA), Que.   < 2e-16 ***
regionPeterborough (CMA), Ont.                      0.00786 ** 
regionQuébec (CMA), Que.                            < 2e-16 ***
regionRed Deer (CMA), Alta.                         0.00832 ** 
regionRegina (CMA), Sask.                           0.00156 ** 
regionSaguenay (CMA), Que.                          0.02473 *  
regionSaint John (CMA), N.B.                        0.13812    
regionSaskatoon (CMA), Sask.                       2.09e-05 ***
regionSherbrooke (CMA), Que.                        0.00036 ***
regionSt. Catharines - Niagara (CMA), Ont.          0.06337 .  
regionSt. John's (CMA), N.L.                        0.03804 *  
regionThunder Bay (CMA), Ont.                       0.34180    
regionToronto (CMA), Ont.                           < 2e-16 ***
regionTrois-Rivières (CMA), Que.                    0.06266 .  
regionVancouver (CMA), B.C.                         < 2e-16 ***
regionVictoria (CMA), B.C.                          < 2e-16 ***
regionWindsor (CMA), Ont.                           0.03266 *  
regionWinnipeg (CMA), Man.                          < 2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 96.94978)

    Null deviance: 17965225  on 31303  degrees of freedom
Residual deviance:  2656157  on 31234  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Different age groups have varying effects on the use of public transport. For example, age groups "25 to 54 years" and "25 to 64 years" show a positive association, indicating higher usage rates among these age brackets compared to the reference group (likely "15 to 24 years" based on the coefficients provided). The positive coefficient for "genderWomen+" suggests that women are more likely to use public transport compared to men, holding other factors constant. The negative coefficient for "population_typetentative_count" could indicate that a higher tentative count (possibly non-immigrants) is associated with lower public transport usage. The coefficients for minority groups indicate that some minority groups are more or less likely to use public transport compared to the reference group (likely "Not a visible minority"). For instance, the "minority_groupNot a visible minority" has a significantly high positive coefficient, suggesting a much higher likelihood of using public transport compared to visible minorities.

```r
# Plotting actual vs. predicted counts
actual <- commuting_data$public_transport
predicted <- predict(quasi_poisson_model, type = "response")

ggplot(commuting_data, aes(x = actual, y = predicted)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(title = "Actual vs. Predicted Public Transport Usage",
       x = "Actual Count",
       y = "Predicted Count")
```



The variability in predictions at higher counts suggests that there may be other forms of nonlinearity or variability not accounted for by the model.

d)

```
# Filter the dataset for the specified demographic group in Edmonton and Toronto
edmonton_data <- commuting_data %>%
  filter(region == "Edmonton (CMA), Alta.",
         age == "35 to 44 years",
         gender == "Men+",
         population_type == "native_count",
         minority_group == "Not a visible minority")

toronto_data <- commuting_data %>%
  filter(region == "Toronto (CMA), Ont.",
         age == "35 to 44 years",
         gender == "Men+",
         population_type == "native_count",
         minority_group == "Not a visible minority")

# Predict the number of individuals using public transit
edmonton_pred <- predict(quasi_poisson_model, newdata = edmonton_data, type = "response")
```

```
toronto_pred <- predict(quasi_poisson_model, newdata = toronto_data, type = "response")

# Calculate the proportions for Edmonton and Toronto
edmonton_proportion <- sum(edmonton_pred) / sum(edmonton_data$total_population)
toronto_proportion <- sum(toronto_pred) / sum(toronto_data$total_population)

# Print the predicted proportions
print(paste("Predicted proportion for Edmonton:", edmonton_proportion))
```

```
[1] "Predicted proportion for Edmonton: 0.0113772365011226"
```

```
print(paste("Predicted proportion for Toronto:", toronto_proportion))
```

```
[1] "Predicted proportion for Toronto: 0.0677782753402027"
```

The result of the Quasi-Poisson regression model predicts that approximately 1.1% of men in Edmonton who are aged 35-44, are not from a visible minority, and are not immigrants use public transit. In comparison, for the same demographic in Toronto, the predicted proportion is approximately 6%. The predicted proportions are within a plausible range for public transit usage, which suggests that the model is providing reasonable estimates.

e)

This model could be used for:

- Policy Development: To help policymakers understand public transit utilization patterns and develop targeted initiatives to encourage public transit use.
- Social Research: To study the mobility patterns of various population segments, contributing to research on social inclusion and urban mobility.

Limitations of the Model The model has several limitations:

- Omitted Variables: Important predictors such as income, occupation, distance to transit stops, and service quality were not included. These could significantly influence public transit use.
- CMA-Level Aggregation: The analysis was conducted at the CMA level, which may mask local variations within CMAs.
- Causality: The regression model identifies correlations but does not establish causality.

Variables of Interest for Future Investigation Future models could benefit from including variables such as:

- Socioeconomic Factors: Income levels, employment status, or car ownership could influence public transit usage.
- Geographic Data: Distance to the nearest transit stop or the density of the transit network might affect usage patterns.
- Service Attributes: Frequency, reliability, and coverage of public transit services can play a significant role.

By incorporating these additional variables, future models could provide a more comprehensive understanding of public transit usage patterns and allow for more accurate predictions and better-informed decision-making.