# Methods_of_Applied_Stat_2_Assignment_1

Rudraharsh Tewary

## Question 1

### a)

Following the given assumptions and law of total expectation, we calculate $E[Y]$ as :

$$E[Y] = E[E[Y|\theta]] = E[\mu\theta] = \mu \times E[\theta] = \mu$$

Similarly, we can calculate Variance as:

$$Var[Y] = E[Var[Y|\theta]] + Var[E[Y|\theta]] = E[\mu\theta] + Var[\mu\theta] = \mu + \mu^2\sigma^2 = \mu(1 + \mu\sigma^2)$$

### b)

Now, to get the distribution of $Y$, we first assume that $\theta \sim Gamma(\alpha, \beta)$, we know the Moment Generating Function (MGF) of $\theta$ is $M_\theta(t) = (1 - \beta t)^{-\alpha}$ and the MGF of $Y|\theta$ is $M_{Y|\theta}(t) = e^{\mu\theta(e^t-1)}$. So, we proceed to calculate as follows:

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

Which is nothing but the Moment Generating Function of negative binomial, so, we can say that $Y \sim NB(\alpha, \frac{1}{1+\mu\beta})$.

### c)

If we want to get $E[Y] = \mu$ and $Var[Y] = \mu(1 + \mu\sigma^2)$. We use the mean and variance of Negative Binomial distribution.

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

And, We know

$$Var[Y] = \frac{\alpha(1 - \frac{1}{1+\mu\beta})}{(\frac{1}{1+\mu\beta})^2} = \mu(1 + \mu\sigma^2)$$

Solving the above equations for $\alpha$ and $\beta$, we get the values for them as :

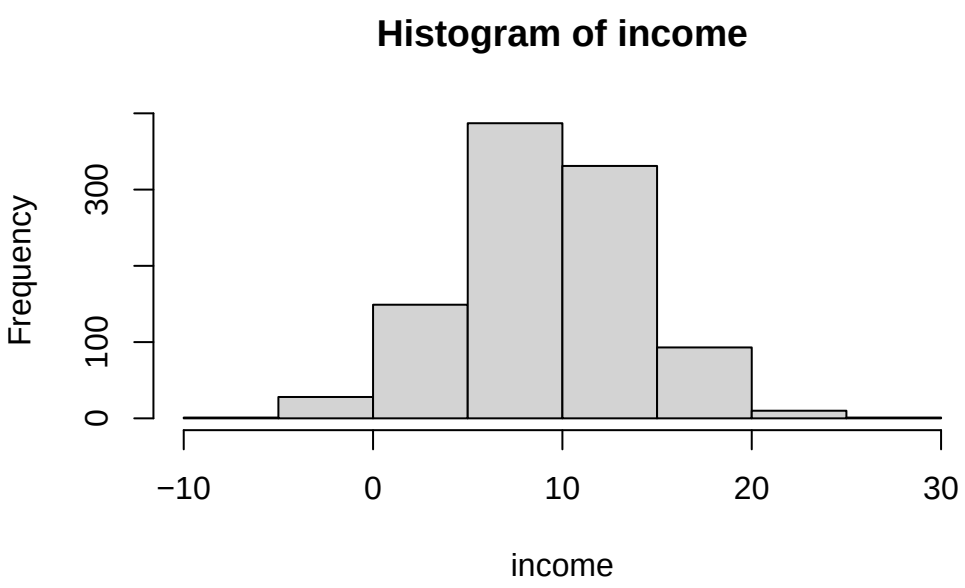
$$\alpha = \frac{1}{\sigma^2}\beta = \sigma^2$$

## Question 2

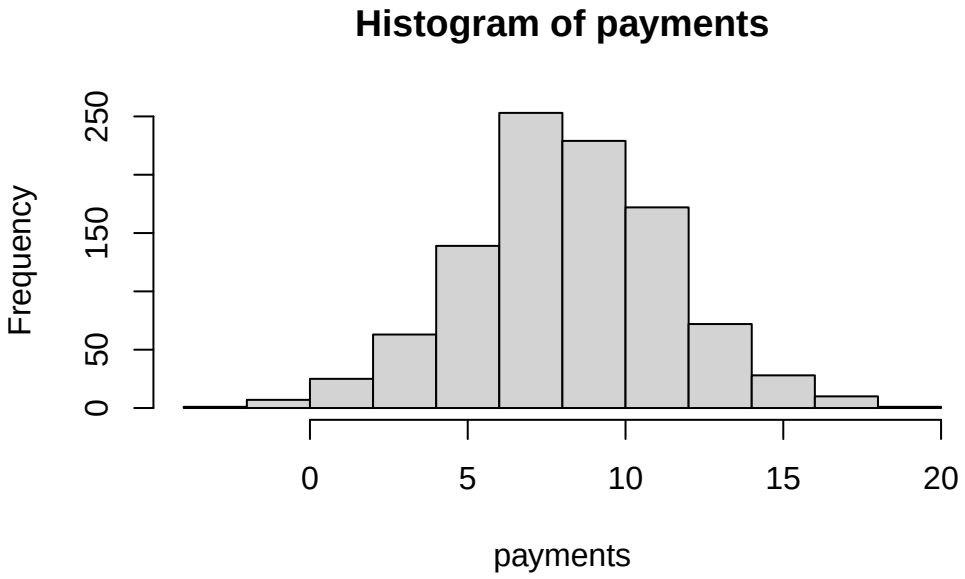
### a)

We generate the data and the plot with the following chunk of code:

```
income <- rnorm(1000, log(10000), log(100))
payments <- rnorm(1000, log(3500), log(30))
hist(income)
```




```
hist(payments)
```

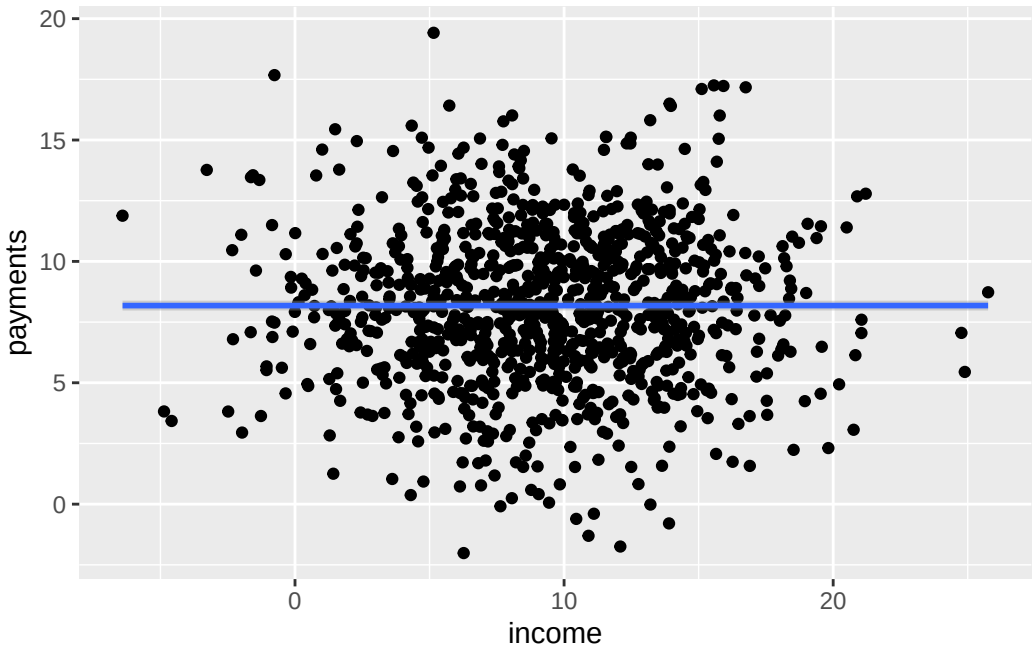


**b)**

We generate the scatter plot with the following chunk of code:

```
library(ggplot2)
data_1 <- data.frame(income,payments)
ggplot(data = data_1, aes(x=income,y=payments))+
  geom_point()+
  geom_smooth()
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




We observe from the following scatter plot that there doesnt seem to exist a correlation between the distribution of income and payments, thats why the line of best fit goes through the middle.

**c)**

We create the survey column in the following code block:

```r
scaled_income <- scale(income)
scaled_payments <- scale(payments)
calculate_survey <- function(z1, z2) {
  random_noise <- rnorm(length(z1))
  sum_z_scores <- z1 + z2 + random_noise
  survey <- sum_z_scores > 0
  return(survey)
}
data_1$survey <- calculate_survey(scaled_income,scaled_payments)
```

We can proceed to calculate the mean income and payment statistics for the surveyed and non_surveyed group of the population in the next code chunk:

```r
head(data_1)
```

```
     income  payments survey
1  8.212684  1.723274  FALSE
2 13.926244  9.936084   TRUE
3  7.948516  9.732813   TRUE
4 10.050426  7.138803   TRUE
5  7.937426 13.327452   TRUE
6  3.727463 10.369961   TRUE
```

```r
mean_income <- mean(data_1$income[data_1$survey == TRUE])
mean_income
```

```
[1] 11.43477
```

```r
mean_payments <- mean(data_1$payments[data_1$survey == TRUE])
mean_payments
```

```
[1] 9.744069
```

```r
mean_income_2 <- mean(data_1$income[data_1$survey == FALSE])
mean_income_2
```

```
[1] 7.181219
```

```r
mean_payments_2 <- mean(data_1$payments[data_1$survey == FALSE])
mean_payments_2
```

```
[1] 6.646929
```

Now, looking at the way we assigned the survey value to the population, we see that our method would be more likely to survey people with a higher income and payment. This leads to us observing that same trend in our summary statistics where in we notice that people who were surveyed tended to have on average a higher income and payment value than the non-surveyed population.

### d)

We make the linear models in the following code chunk:

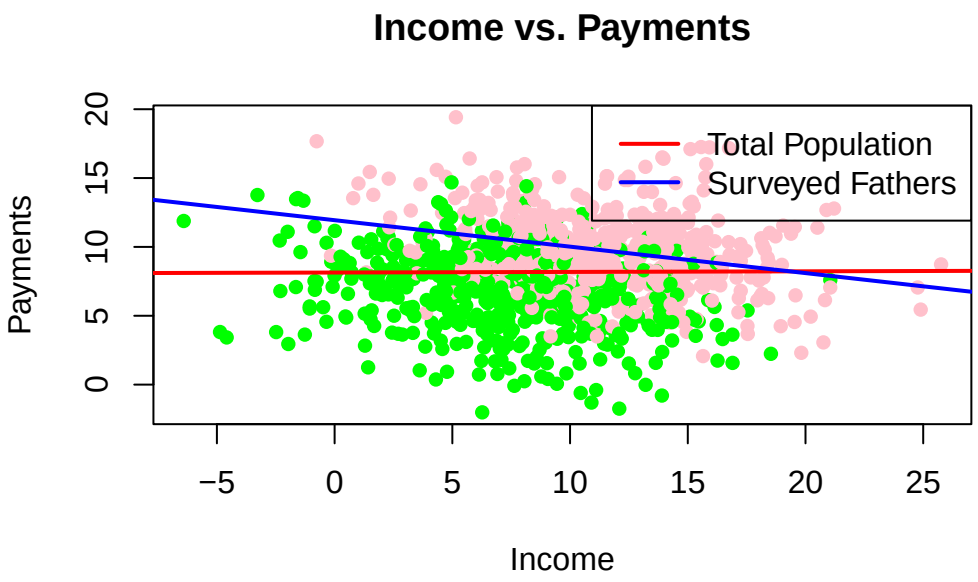
```r
linear_model_1 <- lm(payments ~ income, data = subset(data_1,survey==TRUE))
linear_model_2 <- lm(payments ~ income, data = data_1)
summary(linear_model_1)
```

```
Call:
lm(formula = payments ~ income, data = subset(data_1, survey ==
    TRUE))

Residuals:
    Min      1Q  Median      3Q     Max
-6.8641 -1.8697  0.0193  1.6976  8.4660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.94053    0.36443  32.765  < 2e-16 ***
income      -0.19209    0.03003  -6.396  3.7e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.715 on 493 degrees of freedom
Multiple R-squared:  0.07663,   Adjusted R-squared:  0.07475
F-statistic: 40.91 on 1 and 493 DF,  p-value: 3.702e-10
```

```r
summary(linear_model_2)
```

```
Call:
lm(formula = payments ~ income, data = data_1)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1819  -2.0506  -0.0675   2.1739  11.2551 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.138955   0.227628  35.756   <2e-16 ***
income      0.004421   0.021888   0.202     0.84    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.24 on 998 degrees of freedom
Multiple R-squared:  4.088e-05, Adjusted R-squared:  -0.0009611 
F-statistic: 0.0408 on 1 and 998 DF,  p-value: 0.84
```

```r
plot(data_1$income, data_1$payments, col = ifelse(data_1$survey, "pink", "green"),
     pch = 16, xlab = "Income", ylab = "Payments", main = "Income vs. Payments")

# Adding regression lines
abline(linear_model_2, col = "red", lwd = 2)
abline(linear_model_1, col = "blue", lwd = 2)
legend("topright", legend = c("Total Population", "Surveyed Fathers"),
       col = c("red", "blue"), lty = 1, lwd = 2)
```




We see that the first model which was only fitted on the surveyed population assumes there to be a significant relationship between income level and payment values. But, in the second model fitted on the whole dataset, we see that there is no significant relationship between income level and payments

**e)**

By observing the plots and summary of our models, we begin to see a very clear difference. That is, an ecological fallacy created by the introduction of the 'survey' variable in the dataset. We observed previously that due to the random nature of our generated values, there should be no correlation between payments and income. However, by using the 'survey' value in our model, we see that the linear model gives a very significant p-value to the income covariate, whereas if we use the whole data we see that the model gives an actual non-significant p-value to the income covariate

## Question 3

```
hurricane_data <- read.csv("C:/Users/rudra/Downloads/pnas_1402786111_sd01.csv")
```

```
library(janitor)
```

```
Warning: package 'janitor' was built under R version 4.3.2
```

```
Attaching package: 'janitor'
```

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

    chisq.test, fisher.test
```

```r
hurricane_data <- clean_names(hurricane_data)
head(hurricane_data)
```

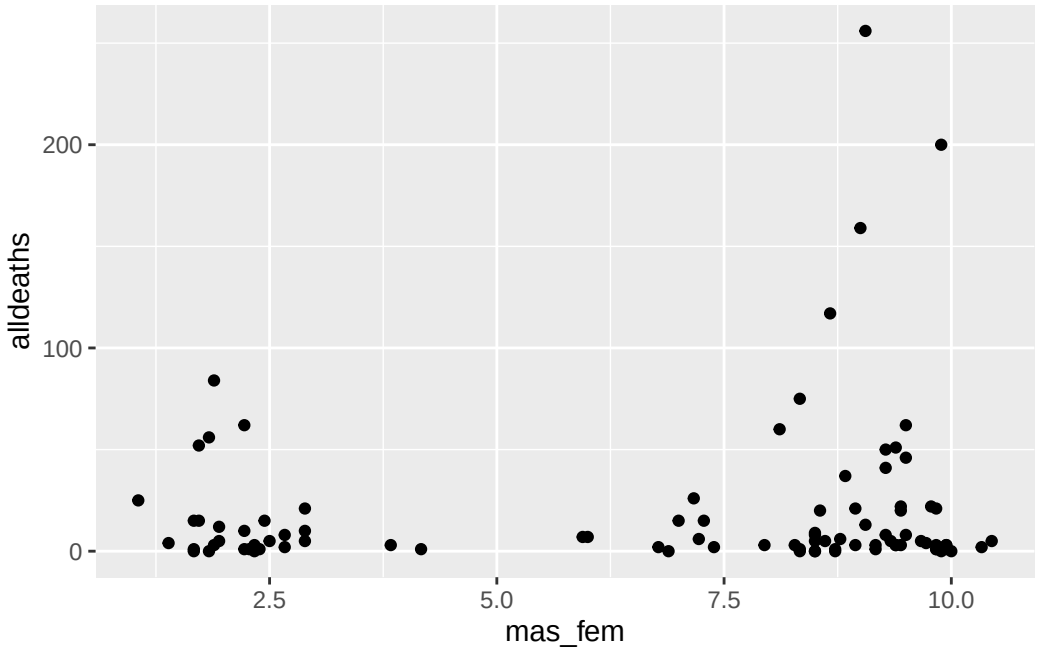
```
  year     name mas_fem min_pressure_before minpressure_updated_2014 gender_mf
1 1950     Easy 6.77778                 958                       960         1
2 1950     King 1.38889                 955                       955         0
3 1952     Able 3.83333                 985                       985         0
4 1953  Barbara 9.83333                 987                       987         1
5 1953 Florence 8.33333                 985                       985         1
6 1954    Carol 8.11111                 960                       960         1
  category alldeaths  ndam elapsed_yrs source z_mas_fem z_min_pressure_a
1        3         2  1590          63    MWR  -0.00094         -0.35636
2        3         4  5350          63    MWR  -1.67076         -0.51125
3        1         3   150          61    MWR  -0.91331          1.03765
4        1         1    58          60    MWR   0.94587          1.14091
5        1         0    15          60    MWR   0.48108          1.03765
6        3        60 19321          59    MWR   0.41222         -0.25310
     zndam
1 -0.43913
2 -0.14843
3 -0.55047
4 -0.55758
5 -0.56090
6  0.93174
```

**a)**

We create the graphs in the following code chunk:

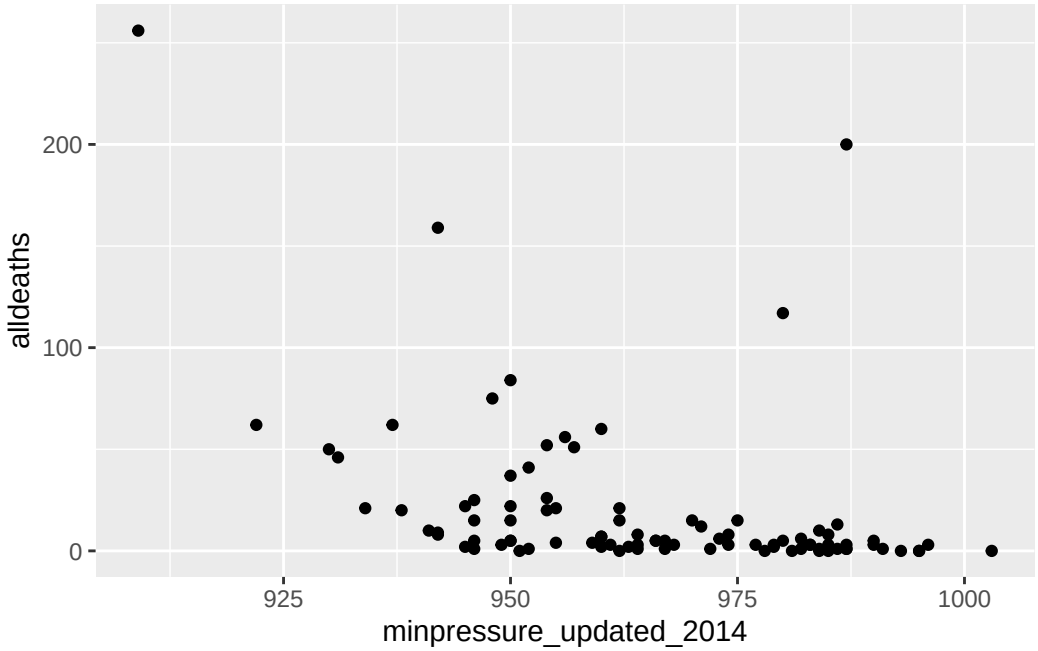
```r
ggplot(data = hurricane_data, aes(x=mas_fem,y=alldeaths))+
  geom_point()
```

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



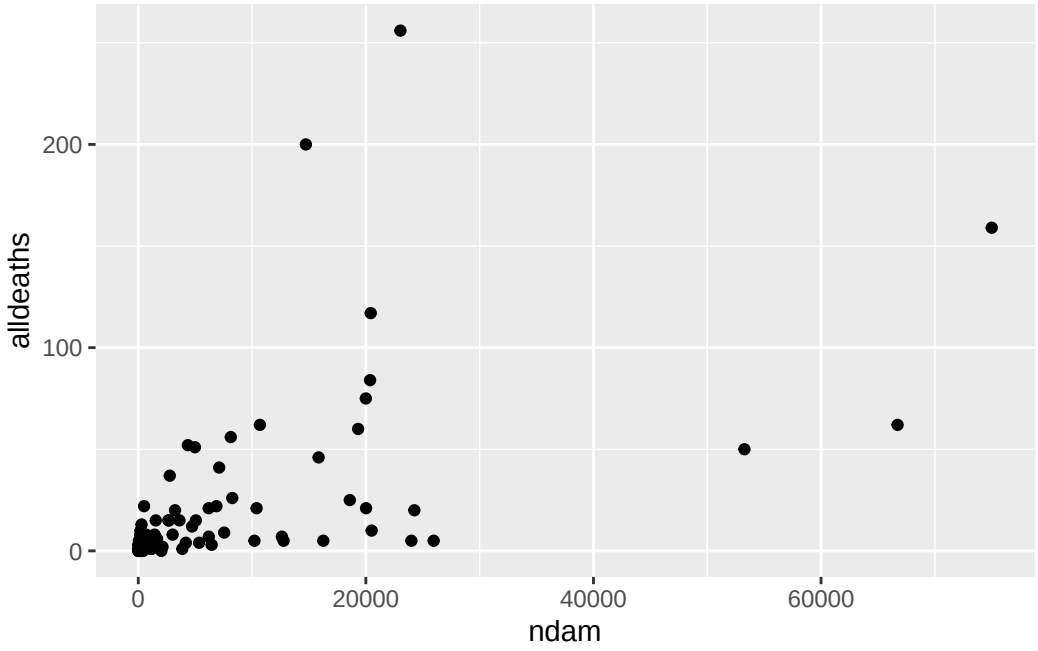

```
ggplot(data = hurricane_data, aes(x=minpressure_updated_2014,y=alldeaths))+
  geom_point()
```

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




```
ggplot(data = hurricane_data,aes(x=ndam,y=alldeaths))+
  geom_point()
```

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




1. In the first graph, we observe that the distribution of deaths caused by hurricanes is pretty even for both extremes of the mas_fem index, with hurricanes having feminine/masculine names causing nearly equivalent death counts. Though there is a presence of few outliers in death counts in the feminine side of the graph.
2. In the second graph, we notice that the air pressure of the hurricane has no visible impact on the number of deaths caused by it.
3. In the third graph, we notice that most hurricanes caused damage between ndam 0 to 20,000. With some hurricanes causing major damage, we also noticed that the death count is not highly correlated with damage, with hurricanes that caused extreme infrastructural damage having lower death counts than some hurricanes which had lower damage and higher death counts.

**b)**

We can define the poisson model in the following code chunk:

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

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

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.500369   0.063297  39.502   <2e-16 ***
mas_fem     0.073873   0.007891   9.362   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    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: 4266.4

Number of Fisher Scoring iterations: 6
```

We get a low and positive coefficient estimate for the mas_fem index, with a significant p-value, which means the model thinks that, for the mas_fem index the following relation holds $e^{(2.5+x\times 0.073873)}$. Which means that for each increase of the mas_fem index, the deaths will increase multiplicatively by $e^{(0.073873)} \approx 1.07$.

We test for overdispersion in the following code block:

```
library(AER)
```

```
Warning: package 'AER' was built under R version 4.3.2

Loading required package: car

Warning: package 'car' was built under R version 4.3.2

Loading required package: carData

Warning: package 'carData' was built under R version 4.3.2

Loading required package: lmtest

Loading required package: zoo
```

```
Warning: package 'zoo' was built under R version 4.3.2


Attaching package: 'zoo'

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

    as.Date, as.Date.numeric

Loading required package: sandwich

Warning: package 'sandwich' was built under R version 4.3.2

Loading required package: survival
```

```r
dispersiontest(poisson_model,trafo=1)
```

```
    Overdispersion test

data:  poisson_model
z = 2.4631, p-value = 0.006887
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
71.17896
```

We see that 'dispersiontest' gives us $\alpha = 71.7$ which means the data is heavily overdispersed. To account for that, we proceed to fit a quasipoisson model.

We define a quasipoisson model in the following code chunk:

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

```
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 73.78495)

    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
```

We see that the interpretation drastically changed with the quasipoisson model, although the coefficient estimates stay mostly the same. Our standard error for the coefficients increased by an order of a magnitude, also the p-value for 'mas_fem' heavily jumped, causing it to become non-significant and the model suggesting there is no correlation between that index and deaths.

### c)

We reproduce model 4 given in table S2 using the following code:

```r
library(MASS)
model_4 <- glm.nb(alldeaths ~ z_min_pressure_a + z_mas_fem + zndam + z_mas_fem*z_min_press
summary(model_4)
```

```
Call:
glm.nb(formula = alldeaths ~ z_min_pressure_a + z_mas_fem + zndam +
    z_mas_fem * z_min_pressure_a + z_mas_fem * zndam, data = hurricane_data,
    init.theta = 0.8112505161, link = log)

Coefficients:
```

```
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                  2.4756     0.1222  20.261  < 2e-16 ***
z_min_pressure_a            -0.5521     0.1503  -3.673 0.000239 ***
z_mas_fem                    0.1723     0.1238   1.392 0.163992
zndam                        0.8635     0.1445   5.976 2.28e-09 ***
z_min_pressure_a:z_mas_fem   0.3947     0.1521   2.595 0.009453 **
z_mas_fem:zndam              0.7051     0.1501   4.699 2.62e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.8113) family taken to be 1)

    Null deviance: 184.86  on 91  degrees of freedom
Residual deviance: 102.83  on 86  degrees of freedom
  (6 observations deleted due to missingness)
AIC: 658.09

Number of Fisher Scoring iterations: 1

              Theta:  0.811
          Std. Err.:  0.124

 2 x log-likelihood:  -644.091
```

We calculate the effect of femininity on deaths given median pressure and damage using the following code chunk:

```r
median_dam <- median(hurricane_data$zndam,na.rm = TRUE)
median_press <- median(hurricane_data$z_min_pressure_a,na.rm = TRUE)

coefs <- model_4$coefficients

coefs["z_mas_fem"] + coefs["z_min_pressure_a:z_mas_fem"]*median_press + coefs["z_mas_fem:z
```

```
 z_mas_fem
-0.1626146
```

We get that for median pressure and damage, given a unit increase in z_mas_fem, the average of the logarithm of deaths decreases by 0.1626146

**d)**

We use the following code chunk to predict the deaths caused by hurricane sandy estimated using model_4:

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.2
```

```
Warning: package 'readr' was built under R version 4.3.2
```

```
Warning: package 'forcats' was built under R version 4.3.2
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
x dplyr::recode() masks car::recode()
x dplyr::select() masks MASS::select()
x purrr::some()   masks car::some()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
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"]
)

pred_death <- predict(model_4,newdata = new_data,type = "response")
pred_death
```

```
       1
20806.95
```

```r
actual_death <- hurricane_data |>
                filter(name == "Sandy")
actual_death$alldeaths
```

```
[1] 159
```

We see that the model is horribly off predicting the number of deaths to be 20,806. Which is way off the actual death count of 159. Which means the model has very bad prediction capabilities.

### e)

Upon reading the archival study of the paper, I gleamed the following strengts and weaknesses of that section:

Strengths

1. One strength was the time period of their data, taking samples in 62 year span of 1950-2012 would give a comprehensive range of hurricanes with varying levels of severity and femininity.
2. The idea of using a negative binomial regression for modelling the data was also a good choice due to the data being count data and intrinsically overdispersed.

Weaknesses

1. Except some very obvious names, the perceived gender of a name can be quite subjective, so to make the index more robust, more people should have been surveyed. Not to mention, restricting the analysis to hurricanes in the United States is a weakness, to make the analysis more robust, global hurricanes which made landfall should have been considered.
2. Again, the authors aim to establish a correlation between the lethality of a hurricane and its femininity, but the data used in the archival study is not large enough to confidently establish such a link, most male and female hurricanes in the dataset have similar death counts, with only some female hurricanes having higher outlier level death counts. Not to mention, multiple factors which affect a hurricane's death count like path, width, population density of area which they hit, are important factors which haven't been talked about. It might be much better to take a proportion estimate of the damage/deaths caused by a hurricane

### f)

No, I am not convinced by the analysis and results of the authors in their study. Their claim of establishing a strong link between the MFI (Mas-Fem Index) of a hurricane's name and its death count is flimsy at best. First of all, the trend itself is not uniform, the authors claim that the relation only establishes itself in the case of severe hurricanes, even then, the veracity of their claims is questionable due to them only using data of hurricanes in the United States. Then, the statistical models used for their archival study are not great at explaining the trends in the data, their first 3 models have outright bad summary statistics, and the final model which they use, a negative binomial model with standardized covariates and interaction terms, is also not good enough. Although the model has decent summary statistics and low overdispersion. It has horrible prediction accuracy, being off by a multiple of 133 when asked to predict the death counts of hurricane sandy. This would indicate that the model might be cherry picked. Again, the question asked by the authors of the paper that the perceived femininity of a hurricane might cause people to take lesser protective action can be a good social science question, but there are a lot of factors that havent been considered which impede this analysis. One, hurricanes themselves are not highly frequent events, much less hurricanes which form, make landfall and cause damage to public property. Secondly, the authors dont take into account that a hurricane might be much more lethal simply due to the population density of an area, or the development level of the area that the hurricane hits. Finally, prevailing social sentiments of a locality also matter, in some places people might be just that much lesser inclined to evacuate, causing more deaths to occur if a hurricane hits.

## Question 4

### a)

In the following series of code chunks, we will tidy up the dataset:

```
immigrant <- read.csv("C:/Users/rudra/Documents/98100468.csv")
```

```
immigrant <- clean_names(immigrant)
drops <- c('symbol','symbol_1','symbol_2','symbol_3','symbol_4','symbol_5','symbol_6','sym
immigrant <- immigrant[ , !(names(immigrant) %in% drops)]
head(immigrant)
```

```
     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_immigrat
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
```

```r
immigrant <- immigrant|>
  select(geo,
         visible_minority = 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_perma
```

```r
immigrant <- immigrant |>
  filter(grepl("(CMA)", geo)) |>
  filter(!grepl("Total", visible_minority) &
         !grepl("Total", age) &
         !grepl("Total", gender))|>
  filter(statistics == "Count")|>
  filter(commute %in%
           c("Car, truck or van", "Public transit", "Active transportation", "Other method
```

```r
immigrant$immigrants_count <- immigrant$immigrants + immigrant$non_permanent_residents
immigrant <- immigrant |>
  select(
    geo,
    visible_minority,
    age,
    gender,
    commute,
    non_immigrants,
    immigrants = immigrants_count)
head(immigrant)
```

```
                   geo visible_minority           age gender
1 St. John's (CMA), N.L.      South Asian 15 to 24 years   Men+
2 St. John's (CMA), N.L.      South Asian 15 to 24 years   Men+
3 St. John's (CMA), N.L.      South Asian 15 to 24 years   Men+
4 St. John's (CMA), N.L.      South Asian 15 to 24 years   Men+
5 St. John's (CMA), N.L.      South Asian 15 to 24 years Women+
6 St. John's (CMA), N.L.      South Asian 15 to 24 years Women+
               commute non_immigrants immigrants
1    Car, truck or van             10        220
```

```
2        Public transit              0        125
3 Active transportation              0         80
4          Other method              0          0
5     Car, truck or van             15         90
6        Public transit              0         80
```

```
immigrant_new <- immigrant |>
  mutate(commuting = ifelse(commute == "Public transit", "Public", "Non-Public")) |>
  group_by(geo, visible_minority, age, gender, commuting) |>
  summarise(non_immigrant = sum(non_immigrants), immigrant = sum(immigrants))
```

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

```
head(immigrant_new)
```

```
# A tibble: 6 x 7
# Groups:   geo, visible_minority, age, gender [3]
  geo            visible_minority age   gender commuting non_immigrant immigrant
  <chr>          <chr>            <chr> <chr>  <chr>             <int>     <int>
1 Abbotsford - ~ Arab             15 t~ Men+   Non-Publ~             0        20
2 Abbotsford - ~ Arab             15 t~ Men+   Public                0         0
3 Abbotsford - ~ Arab             15 t~ Women+ Non-Publ~             0        10
4 Abbotsford - ~ Arab             15 t~ Women+ Public                0         0
5 Abbotsford - ~ Arab             15 t~ Men+   Non-Publ~            10        35
6 Abbotsford - ~ Arab             15 t~ Men+   Public                0         0
```

```
immigrant_long <- immigrant_new |>
  pivot_longer(cols = c('immigrant', 'non_immigrant'),
               names_to = "immigrant_status",
               values_to = "pop_count")|>
  pivot_wider(names_from = "commuting",values_from = "pop_count")

immigrant_long <- immigrant_long|>
                  rename(public = Public, non_public = `Non-Public`)
immigrant_long$total = immigrant_long$public + immigrant_long$non_public

head(immigrant_long)
```

```
# A tibble: 6 x 8
# Groups:   geo, visible_minority, age, gender [3]
  geo     visible_minority age   gender immigrant_status non_public public total
  <chr>   <chr>            <chr> <chr>  <chr>                 <int>  <int> <int>
1 Abbots~ Arab             15 t~ Men+   immigrant                20      0    20
2 Abbots~ Arab             15 t~ Men+   non_immigrant             0      0     0
3 Abbots~ Arab             15 t~ Women+ immigrant                10      0    10
4 Abbots~ Arab             15 t~ Women+ non_immigrant             0      0     0
5 Abbots~ Arab             15 t~ Men+   immigrant                35      0    35
6 Abbots~ Arab             15 t~ Men+   non_immigrant            10      0    10
```

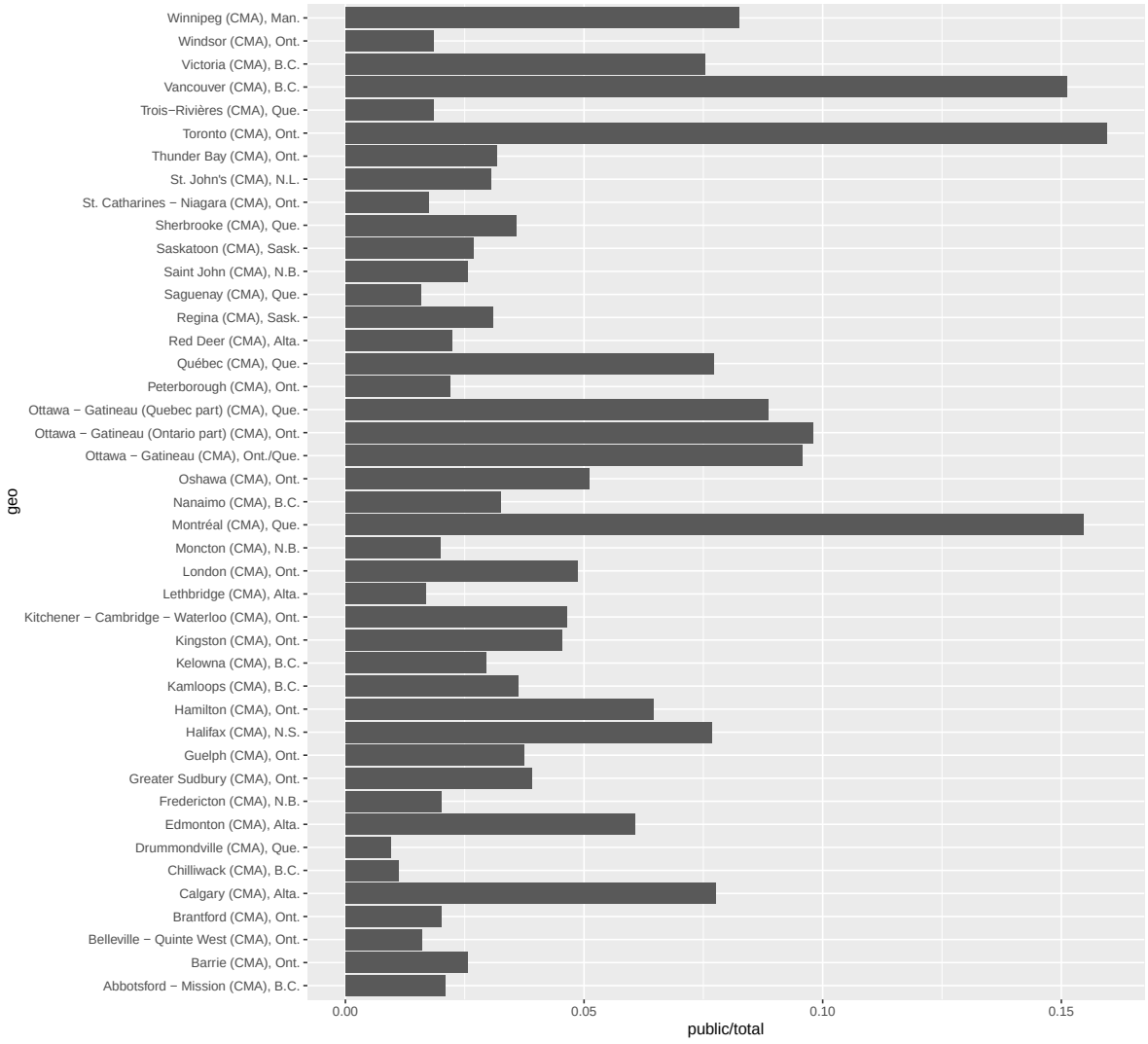
```
library(stringr)

immigrant_long$province <- str_extract(immigrant_long$geo, "\\(CMA\\),\\s(.+)$")
```

**b)**

In the following code chunks, we generate plots for EDA of the dataset:

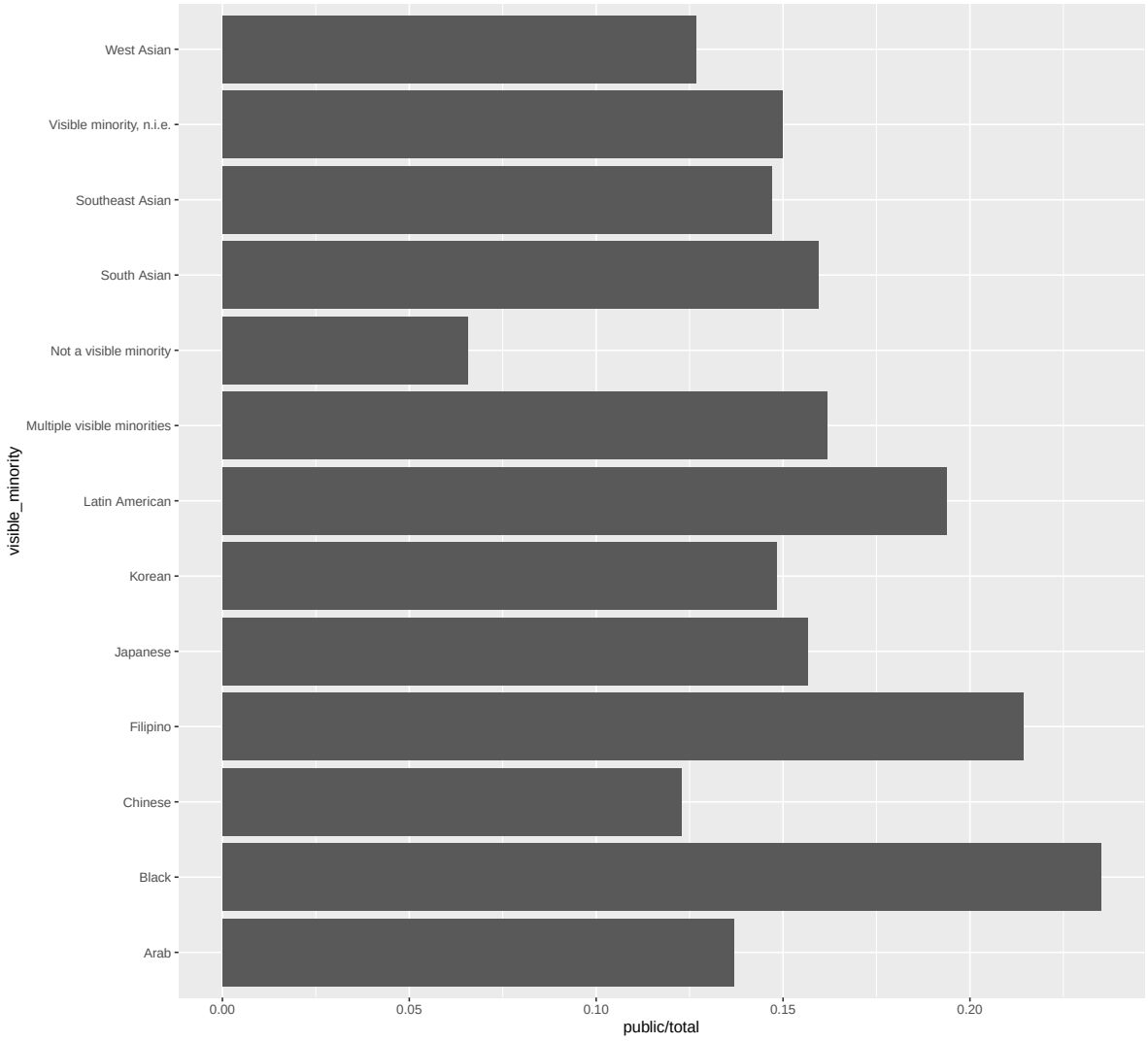
```
immigrant_long |>
  group_by(geo)|>
  summarise(public = sum(public), total = sum(total))|>
ggplot(aes(x=public/total,y=geo))+
  geom_col()+
  geom_smooth()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



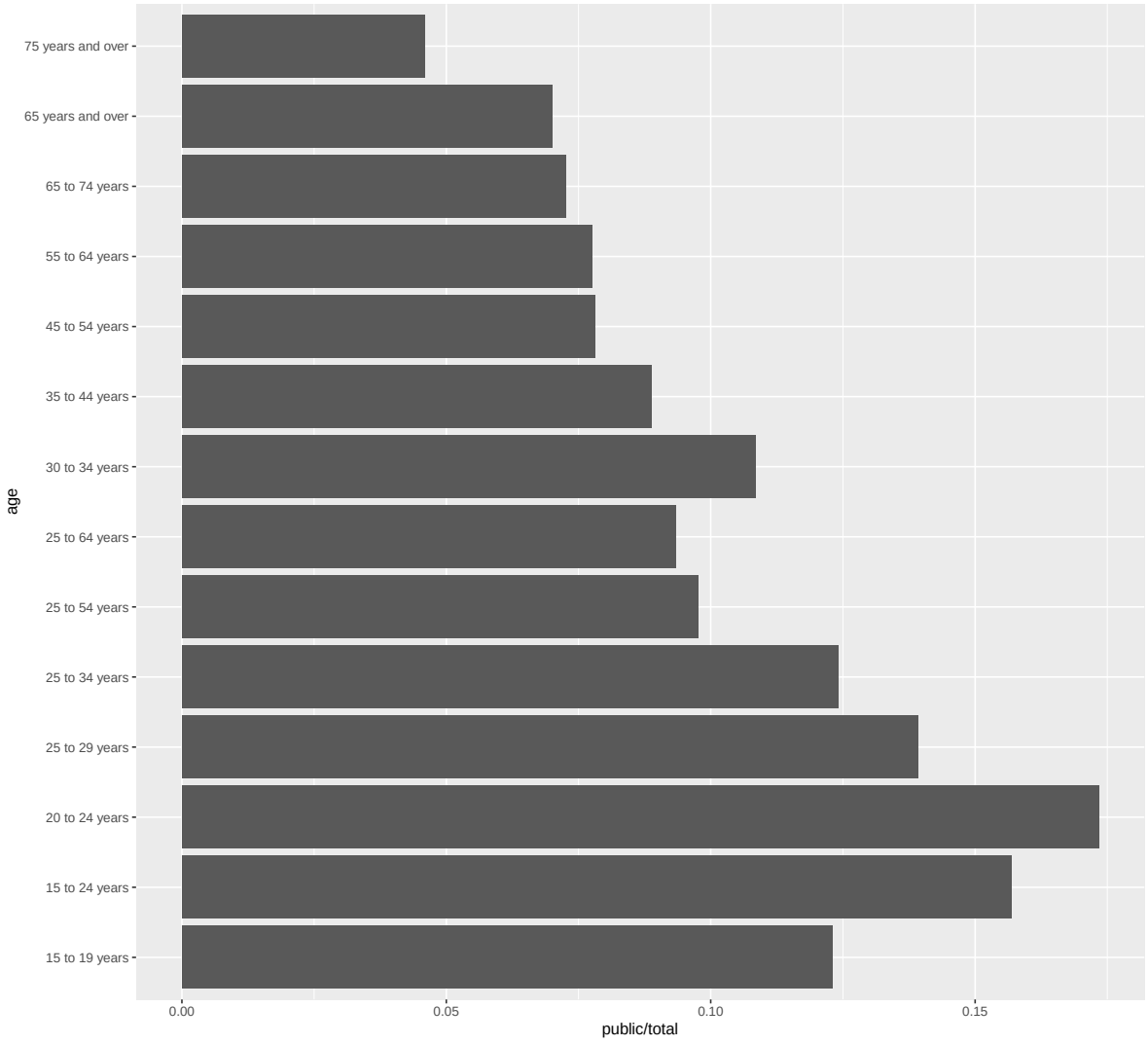

```
immigrant_long |>
  group_by(visible_minority)|>
  summarise(public = sum(public), total = sum(total))|>
ggplot(aes(x=public/total,y=visible_minority))+
  geom_col()+
  geom_smooth()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




```
immigrant_long |>
  group_by(age)|>
  summarise(public = sum(public), total = sum(total))|>
ggplot(aes(x=public/total,y=age))+
  geom_col()+
  geom_smooth()
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




For each of our graphs, we inspect the proportion 'public/total' which tells us the proportion of users of public transit over the total, we make the following inferences from the data:

- In most metropolitan areas the proportion of people who use public transit is significantly lower than the biggest centers of population in the country, namely Montreal,Toronto and Vancouver.
- If we evaluate by ethnicity, the minority population of black people have the highest ratio of use of public transit, with filipinos following close in second place. 'Not a Visible Minority' shows us that most people who are not part of a minority have the lowest ratio of public transit use.

- We also see the trend the public transit is most frequently used by young people, with a high density of users being the age groups of 15-34 years old. This is unsurprising due to the physical effort required to use public transport and that younger people would have lesser access to private transportation.

### c)

We can define the model in the following code chunk, we use a poisson glm as public transit users are discrete count values, secondly, as we are interested in the proportion of public transit users to the total, we offset the model by the total:

```
library(MASS)
library(brglm)
```

```
Warning: package 'brglm' was built under R version 4.3.2
```

```
Loading required package: profileModel
```

```
Warning: package 'profileModel' was built under R version 4.3.2
```

```
'brglm' will gradually be superseded by the 'brglm2' R package (https://cran.r-project.org/pa
 Methods for the detection of separation and infinite estimates in binomial-response models a
```

```
library(logistf)
```

```
Warning: package 'logistf' was built under R version 4.3.2
```

```
immigrant_long <- immigrant_long|>
                  filter(total!= 0)

immigrant_model <- glm(public ~ geo + visible_minority + age + gender + immigrant_status +
summary(immigrant_model)
```

```
Call:
glm(formula = public ~ geo + visible_minority + age + gender +
    immigrant_status + province, family = poisson, data = immigrant_long,
```

```
    offset = log(total))

Coefficients: (9 not defined because of singularities)
                                                 Estimate Std. Error  z value
(Intercept)                                    -3.496096   0.015413 -226.833
geoBarrie (CMA), Ont.                           0.359712   0.019526   18.422
geoBelleville - Quinte West (CMA), Ont.         0.013317   0.027840    0.478
geoBrantford (CMA), Ont.                        0.154614   0.023324    6.629
geoCalgary (CMA), Alta.                         1.222331   0.014671   83.315
geoChilliwack (CMA), B.C.                      -0.417242   0.030872  -13.515
geoDrummondville (CMA), Que.                   -0.486453   0.032870  -14.799
geoEdmonton (CMA), Alta.                        1.022773   0.014739   69.393
geoFredericton (CMA), N.B.                      0.191745   0.024784    7.737
geoGreater Sudbury (CMA), Ont.                  0.892058   0.018799   47.452
geoGuelph (CMA), Ont.                           0.648334   0.019092   33.959
geoHalifax (CMA), N.S.                          1.453387   0.015313   94.915
geoHamilton (CMA), Ont.                         1.191285   0.015103   78.877
geoKamloops (CMA), B.C.                         0.758566   0.020450   37.095
geoKelowna (CMA), B.C.                          0.510468   0.018577   27.479
geoKingston (CMA), Ont.                         0.995838   0.018557   53.662
geoKitchener - Cambridge - Waterloo (CMA), Ont. 0.831728   0.015650   53.146
geoLethbridge (CMA), Alta.                     -0.091766   0.025058   -3.662
geoLondon (CMA), Ont.                           0.927728   0.015693   59.119
geoMoncton (CMA), N.B.                          0.135775   0.022353    6.074
geoMontréal (CMA), Que.                         2.006484   0.014435  138.998
geoNanaimo (CMA), B.C.                          0.640553   0.021469   29.836
geoOshawa (CMA), Ont.                           0.959919   0.016042   59.836
geoOttawa - Gatineau (CMA), Ont./Que.           1.550289   0.014674  105.652
geoOttawa - Gatineau (Ontario part) (CMA), Ont. 1.547834   0.014772  104.785
geoOttawa - Gatineau (Quebec part) (CMA), Que.  1.550893   0.015516   99.953
geoPeterborough (CMA), Ont.                     0.309837   0.024460   12.667
geoQuébec (CMA), Que.                           1.535031   0.014883  103.137
geoRed Deer (CMA), Alta.                        0.087636   0.024415    3.590
geoRegina (CMA), Sask.                          0.394437   0.018012   21.899
geoSaguenay (CMA), Que.                         0.052681   0.023545    2.237
geoSaint John (CMA), N.B.                       0.483934   0.022187   21.811
geoSaskatoon (CMA), Sask.                       0.302187   0.017490   17.277
geoSherbrooke (CMA), Que.                       0.757809   0.017817   42.532
geoSt. Catharines - Niagara (CMA), Ont.        -0.003499   0.018793   -0.186
geoSt. John's (CMA), N.L.                       0.639660   0.018654   34.290
geoThunder Bay (CMA), Ont.                      0.660941   0.021425   30.850
geoToronto (CMA), Ont.                          1.834173   0.014416  127.232
geoTrois-Rivières (CMA), Que.                   0.134491   0.022816    5.895
```

```
geoVancouver (CMA), B.C.                          1.850565   0.014459  127.986
geoVictoria (CMA), B.C.                           1.412375   0.015454   91.390
geoWindsor (CMA), Ont.                           -0.002772   0.018610   -0.149
geoWinnipeg (CMA), Man.                           1.262737   0.014819   85.211
visible_minorityBlack                             0.602241   0.003868  155.701
visible_minorityChinese                          -0.078115   0.004383  -17.822
visible_minorityFilipino                          0.558233   0.004083  136.708
visible_minorityJapanese                          0.317657   0.010070   31.546
visible_minorityKorean                            0.086923   0.006841   12.706
visible_minorityLatin American                    0.385269   0.004466   86.262
visible_minorityMultiple visible minorities       0.274753   0.006149   44.680
visible_minorityNot a visible minority           -0.131542   0.003805  -34.574
visible_minoritySouth Asian                       0.215212   0.003890   55.320
visible_minoritySoutheast Asian                   0.192764   0.005427   35.521
visible_minorityVisible minority, n.i.e.          0.187102   0.007283   25.689
visible_minorityWest Asian                       -0.090669   0.005972  -15.183
age15 to 24 years                                 0.143159   0.004766   30.036
age20 to 24 years                                 0.195185   0.004948   39.448
age25 to 29 years                                -0.024090   0.004982   -4.835
age25 to 34 years                                -0.145383   0.004675  -31.098
age25 to 54 years                                -0.431362   0.004449  -96.965
age25 to 64 years                                -0.467086   0.004417 -105.740
age30 to 34 years                                -0.286836   0.005169  -55.488
age35 to 44 years                                -0.541601   0.004852 -111.616
age45 to 54 years                                -0.686959   0.004949 -138.808
age55 to 64 years                                -0.621103   0.005067 -122.573
age65 to 74 years                                -0.653010   0.007237  -90.235
age65 years and over                             -0.677888   0.007039  -96.310
age75 years and over                             -1.004168   0.020969  -47.889
genderWomen+                                      0.448150   0.001201  373.218
immigrant_statusnon_immigrant                    -0.490020   0.001660 -295.134
province(CMA), B.C.                                     NA         NA       NA
province(CMA), Man.                                     NA         NA       NA
province(CMA), N.B.                                     NA         NA       NA
province(CMA), N.L.                                     NA         NA       NA
province(CMA), N.S.                                     NA         NA       NA
province(CMA), Ont.                                     NA         NA       NA
province(CMA), Ont./Que.                                NA         NA       NA
province(CMA), Que.                                     NA         NA       NA
province(CMA), Sask.                                    NA         NA       NA
                                                 Pr(>|z|)
(Intercept)                                       < 2e-16 ***
geoBarrie (CMA), Ont.                             < 2e-16 ***
```

```
geoBelleville - Quinte West (CMA), Ont.         0.632410
geoBrantford (CMA), Ont.                        3.38e-11 ***
geoCalgary (CMA), Alta.                          < 2e-16 ***
geoChilliwack (CMA), B.C.                        < 2e-16 ***
geoDrummondville (CMA), Que.                     < 2e-16 ***
geoEdmonton (CMA), Alta.                         < 2e-16 ***
geoFredericton (CMA), N.B.                      1.02e-14 ***
geoGreater Sudbury (CMA), Ont.                   < 2e-16 ***
geoGuelph (CMA), Ont.                            < 2e-16 ***
geoHalifax (CMA), N.S.                           < 2e-16 ***
geoHamilton (CMA), Ont.                          < 2e-16 ***
geoKamloops (CMA), B.C.                          < 2e-16 ***
geoKelowna (CMA), B.C.                           < 2e-16 ***
geoKingston (CMA), Ont.                          < 2e-16 ***
geoKitchener - Cambridge - Waterloo (CMA), Ont.  < 2e-16 ***
geoLethbridge (CMA), Alta.                      0.000250 ***
geoLondon (CMA), Ont.                            < 2e-16 ***
geoMoncton (CMA), N.B.                          1.25e-09 ***
geoMontréal (CMA), Que.                          < 2e-16 ***
geoNanaimo (CMA), B.C.                           < 2e-16 ***
geoOshawa (CMA), Ont.                            < 2e-16 ***
geoOttawa - Gatineau (CMA), Ont./Que.            < 2e-16 ***
geoOttawa - Gatineau (Ontario part) (CMA), Ont.  < 2e-16 ***
geoOttawa - Gatineau (Quebec part) (CMA), Que.   < 2e-16 ***
geoPeterborough (CMA), Ont.                      < 2e-16 ***
geoQuébec (CMA), Que.                            < 2e-16 ***
geoRed Deer (CMA), Alta.                        0.000331 ***
geoRegina (CMA), Sask.                           < 2e-16 ***
geoSaguenay (CMA), Que.                         0.025256 *
geoSaint John (CMA), N.B.                        < 2e-16 ***
geoSaskatoon (CMA), Sask.                        < 2e-16 ***
geoSherbrooke (CMA), Que.                        < 2e-16 ***
geoSt. Catharines - Niagara (CMA), Ont.         0.852299
geoSt. John's (CMA), N.L.                        < 2e-16 ***
geoThunder Bay (CMA), Ont.                       < 2e-16 ***
geoToronto (CMA), Ont.                           < 2e-16 ***
geoTrois-Rivières (CMA), Que.                   3.75e-09 ***
geoVancouver (CMA), B.C.                         < 2e-16 ***
geoVictoria (CMA), B.C.                          < 2e-16 ***
geoWindsor (CMA), Ont.                          0.881603
geoWinnipeg (CMA), Man.                          < 2e-16 ***
visible_minorityBlack                            < 2e-16 ***
visible_minorityChinese                          < 2e-16 ***
```

```
visible_minorityFilipino                          < 2e-16 ***
visible_minorityJapanese                          < 2e-16 ***
visible_minorityKorean                            < 2e-16 ***
visible_minorityLatin American                    < 2e-16 ***
visible_minorityMultiple visible minorities       < 2e-16 ***
visible_minorityNot a visible minority            < 2e-16 ***
visible_minoritySouth Asian                       < 2e-16 ***
visible_minoritySoutheast Asian                   < 2e-16 ***
visible_minorityVisible minority, n.i.e.          < 2e-16 ***
visible_minorityWest Asian                        < 2e-16 ***
age15 to 24 years                                 < 2e-16 ***
age20 to 24 years                                 < 2e-16 ***
age25 to 29 years                                1.33e-06 ***
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                              < 2e-16 ***
age75 years and over                              < 2e-16 ***
genderWomen+                                      < 2e-16 ***
immigrant_statusnon_immigrant                     < 2e-16 ***
province(CMA), B.C.                                    NA
province(CMA), Man.                                    NA
province(CMA), N.B.                                    NA
province(CMA), N.L.                                    NA
province(CMA), N.S.                                    NA
province(CMA), Ont.                                    NA
province(CMA), Ont./Que.                               NA
province(CMA), Que.                                    NA
province(CMA), Sask.                                   NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1995154  on 18401  degrees of freedom
Residual deviance:  276728  on 18332  degrees of freedom
AIC: 327013
```

```
Number of Fisher Scoring iterations: 5
```

Looking at the summary of the model and the obtained coefficient values, we see that the model's estimates are similar to our EDA, the highest positive coefficients are offered to the biggest metro cities of Toronto, Vancouver and Montreal. Similarly, the highest positive coefficients in other categories are Black people for minorities, and young people in the age group of 15-24 years old.

### d)

We use our model to get predictions of the given data in the following code chunk:

```
new_data <- data.frame(geo = "Edmonton (CMA), Alta.",
                       visible_minority = "Not a visible minority",
                       age = "35 to 44 years",
                       gender = "Men+",
                       immigrant_status = "non_immigrant",
                       province = '(CMA), Alta.',
                       total=1
                       )
print(new_data)
```

```
                  geo       visible_minority            age gender
1 Edmonton (CMA), Alta. Not a visible minority 35 to 44 years   Men+
  immigrant_status     province total
1    non_immigrant (CMA), Alta.     1
```

```
prediction <- predict(immigrant_model,newdata = new_data,type = "response")
prediction
```

```
         1
0.02634472
```

```
new_data_2 <- data.frame(geo = "Toronto (CMA), Ont.",
                       visible_minority = "Not a visible minority",
                       age = "35 to 44 years",
                       gender = "Men+",
                       immigrant_status = "non_immigrant",
                       province = '(CMA), Ont.',
```

```
                        total = 1
                        )
prediction_2 <- predict(immigrant_model, newdata = new_data_2, type = 'response')
prediction_2
```

```
         1
0.05930348
```

We see that the predictions are off by just a bit to the real values, with 3 being the proportion of public transit users in Edmonton, and 8 being the proportion of public transit users in Toronto. Although the predicitons are not perfect, we see that the model is able to account for the effects of our covariates and is able to get an idea of how much public transit will be used given the demographics and structure and location of a city.

**e)**

In our analysis, we took the transportation dataset from StatCan and after extensive clean-up and editing of the dataset, we started analyzing the trends of public transit use in different metropolitan areas across the country. We managed to uncover trends of public transport use depending on age, ethnicity and the city itself. We created a poisson model with an offset to analyze and predict the impact of these covariates on the proportion of public transit use. Our model has a few limitations, namely it having not a great fit as we can see from the deviance scores, secondly its prediction accuracy is also not great compared to the actual values in the dataset. For analysis, some more variables that would be of more interest are the inclusion of time, that is year when the proportions were counted/recorded so that we can analyze time series trends. Secondly, an inclusion of the economy of a city may also be decent to assess the quality of public infrastructure present. Finally, population density will also give a good idea because a higher population might neccesitate need of public transport.