Draft Data Analysis and Report

Aubrey Ahern

11/14/2022

Introduction

I took my data from https://data.chhs.ca.gov/dataset/hiv-aids-cases. This work outlines all cause deaths in HIV/AIDS by gender, age, race/ethnicity, and transmission category. This data was collected from 2011-2017 within the state of California in order to collect public health information.

As outlined within the assignment, I will be using the following steps to structure my Data Analysis: find data set -> repair data set -> plot relationships -> summarize focal data -> plot expected relationships in data -> guess expected relationships before fitting model -> decide on a statistical model -> check model assumptions -> run statistical model -> interpret your model ->replot data and model to show key relationships

Getting Started

Set things up and import the necessary packages…

```
rm(list = ls())
library(here)
```

```
## here() starts at /home/aeaher19/Biostatistics/Capstone-Project
```

```
library(ggfortify)
```

```
## Loading required package: ggplot2
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.0.1      v forcats 0.5.1
## v purrr   0.3.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(AICcmodavg)
```

I made sure my data was compatible to work with R by adjusting column names to exclude spaces and special characters and so that categories were organized as columns instead of rows. Throughout this process, I made sure that each cell contained only one piece of information. I would also like to address the limitations of this data set before I beging my main code. While this data set contains a lot of information on demographic information on those who have died from HIV/AIDS, it only includes one piece of information per datum. I have worked with the data as best I could but would potentially use a more extensive data set in the future.

Import my data set:

```
DF <- read.csv(here("Data", "fixed-deaths.csv"), stringsAsFactors = TRUE)
DF$count <- as.numeric(DF$count)
```

Analysis

First, let's consider the relationship between gender identity and death count by HIV/AIDS.

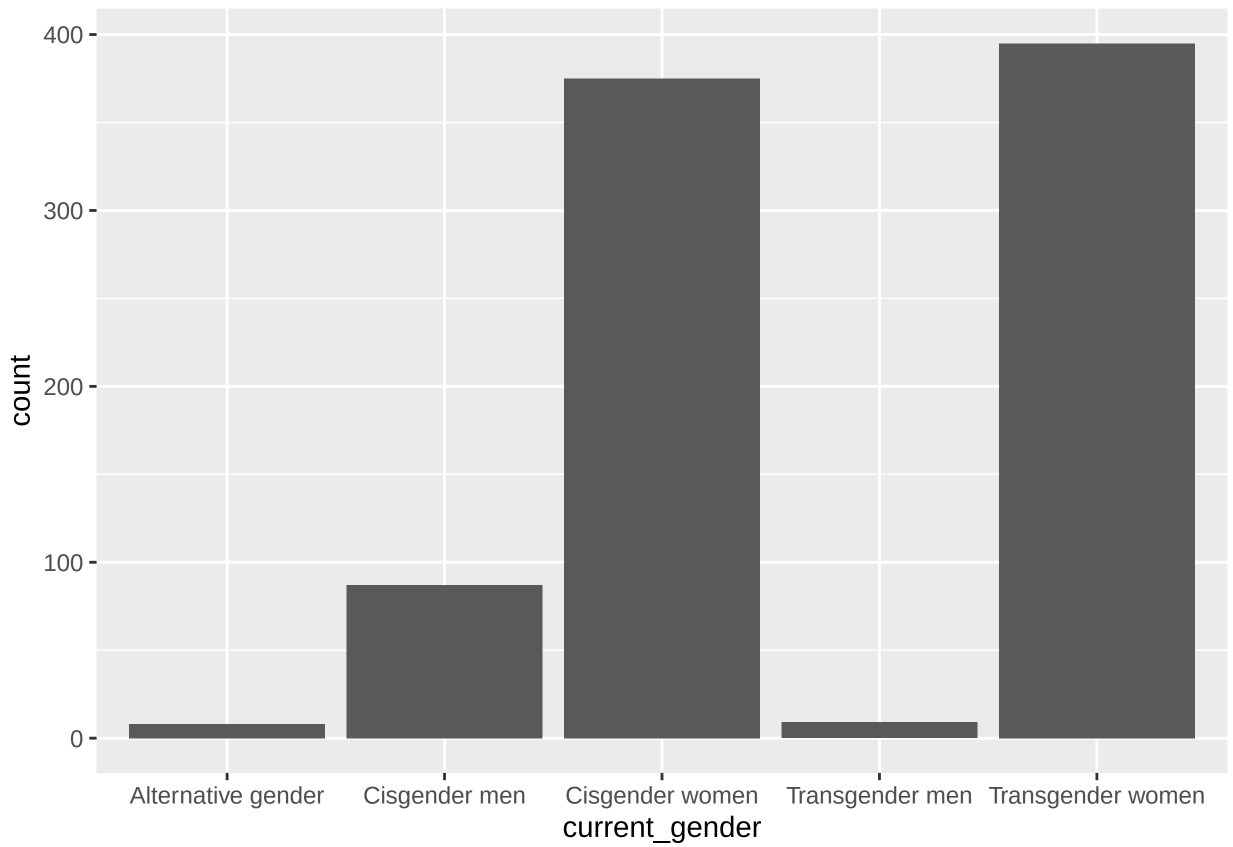
```
DF_gender <- DF %>%
  mutate(count = as.numeric(count)) %>%
  group_by(current_gender) %>%
  summarise(count = sum(count)) %>%
  arrange(desc(count)) %>%
  na.omit()

DF_gender
```

```
## # A tibble: 5 x 2
##   current_gender     count
##   <fct>              <dbl>
## 1 Transgender women    395
## 2 Cisgender women      375
## 3 Cisgender men         87
## 4 Transgender men        9
## 5 Alternative gender     8
```

Now let's visualize this relationship.

```
ggplot(DF_gender, aes(current_gender, count))+
     geom_col()
```



Based on this visualization we can hypothesize that cisgender and transgender women are more likely to die from HIV/AIDS thank those of other genders. Now, let's run a statistical test to see if I can accept or reject the null hypothesis. This relationship includes one piece of categorical data (current gender) and one piece of count data (count). Thus, I will conduct an ANOVA test.

Before I can run my statistical test, I must exclude all pieces of data that do not address gender to make my data compatible for R.

```
DF_model <- DF %>%
  mutate(count = as.numeric(count)) %>%
  slice(106:140)
myMod1 <- lm(count ~ current_gender, data = DF_model)
```

```
anova(myMod1)
```

```
## Analysis of Variance Table
##
## Response: count
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## current_gender  4 21655.5  5413.9  85.598 5.778e-16 ***
## Residuals      30  1897.4    63.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Given my first test, we may see that trans women and cis women experienced more deaths by HIV-AIDS than those of other genders. Thus, let's analyze the relationship between year and deaths for trans and cis women. Before we can run our second statistical test, let's create a new data frame that only contains the data we wish to analyze.

```
DF_transwomen <- DF %>%
  filter(current_gender == "Transgender women")%>%
  mutate(year = as.numeric(year),
         count = as.numeric(count))
DF_transwomen
```

```
##   year age_at_death    current_gender race_ethnicity transmission_category
## 1 2011         <NA> Transgender women           <NA>                  <NA>
## 2 2012         <NA> Transgender women           <NA>                  <NA>
## 3 2013         <NA> Transgender women           <NA>                  <NA>
## 4 2014         <NA> Transgender women           <NA>                  <NA>
## 5 2015         <NA> Transgender women           <NA>                  <NA>
## 6 2016         <NA> Transgender women           <NA>                  <NA>
## 7 2017         <NA> Transgender women           <NA>                  <NA>
##   count
## 1    40
## 2    54
## 3    58
## 4    58
## 5    50
## 6    85
## 7    50
```

Now, we may run a t-test to see if there is a relationship between time (in years) and death count for trans women.

```
myMod2 <- lm(count ~ year, data = DF_transwomen)
summary(myMod2)
```

```
##
## Call:
## lm(formula = count ~ year, data = DF_transwomen)
##
## Residuals:
##       1       2       3       4       5       6       7
##  -7.429   3.571   4.571   1.571  -9.429  22.571 -15.429
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5985.571   5184.451  -1.155    0.300
## year            3.000      2.574   1.165    0.296
##
## Residual standard error: 13.62 on 5 degrees of freedom
## Multiple R-squared:  0.2136, Adjusted R-squared:  0.05633
## F-statistic: 1.358 on 1 and 5 DF,  p-value: 0.2964
```

This allows us to accept our null hypothesis as there was no observable relationship between year and death count for trans women. Let’s take a look at the same relationship but with cis women.

Again, I will first create a new data frame that only contains the data I wish to analyze.

```
DF_ciswomen <- DF %>%
  filter(current_gender == "Cisgender women")%>%
  mutate(year = as.numeric(year),
         count = as.numeric(count))
DF_ciswomen
```

```
##   year age_at_death  current_gender race_ethnicity transmission_category count
## 1 2011         <NA> Cisgender women           <NA>                  <NA>    56
## 2 2012         <NA> Cisgender women           <NA>                  <NA>    61
## 3 2013         <NA> Cisgender women           <NA>                  <NA>    53
## 4 2014         <NA> Cisgender women           <NA>                  <NA>    52
## 5 2015         <NA> Cisgender women           <NA>                  <NA>    49
## 6 2016         <NA> Cisgender women           <NA>                  <NA>    59
## 7 2017         <NA> Cisgender women           <NA>                  <NA>    45
```

Now, let's run another t-test to see if there is a relationship between time (in years) and death count for cis women.

```
myMod3 <- lm(count ~ year, data = DF_ciswomen)
summary(myMod3)
```

```
##
## Call:
## lm(formula = count ~ year, data = DF_ciswomen)
##
## Residuals:
##      1      2      3      4      5      6      7
## -1.964  4.500 -2.036 -1.571 -3.107  8.357 -4.179
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3002.643   1923.334   1.561    0.179
## year          -1.464      0.955  -1.533    0.186
##
## Residual standard error: 5.053 on 5 degrees of freedom
## Multiple R-squared:  0.3198, Adjusted R-squared:  0.1838
## F-statistic: 2.351 on 1 and 5 DF,  p-value: 0.1858
```

This allows us to accept our null hypothesis as there was no observable relationship between year and death count for cis women.

Let's try to consider what category, race, gender, age, or transmission method, is the best predictor of deth count by HIV/AIDS. First let's create the models necessary for an AIC Table which will indicate to us what the best predictor is.

```
mod_race <- lm(count ~ race_ethnicity, data = DF)
mod_gender <- lm(count ~ current_gender, data = DF)
mod_age <- lm(count ~ age_at_death, data = DF)
mod_transmission <- lm(count ~ transmission_category, data = DF)

death_models <- list(mod_race, mod_gender, mod_age, mod_transmission)
names(death_models) <- c("Race", "Gender", "Age", "Transmission")
```

Now let's run the test…

```
AICdeaths <- aictab(cand.set = death_models, second.ord = TRUE, sort = TRUE)
AICdeaths
```

```
##
## Model selection based on AICc:
##
##                K   AICc Delta_AICc AICcWt Cum.Wt      LL
## Gender         6 254.08       0.00      1      1 -119.54
## Race           8 487.40     233.32      0      1 -233.90
## Transmission 10 1032.61     778.53      0      1 -505.04
## Age          16 1054.35     800.28      0      1 -508.09
```

k = number of parameters in the model AICc - the information score of the model (the lower-case c indicates that the value has been calcuated from the AIC test correct for small sample sizes the smaller the AIC value the better the model fit) the first the model that shows up on the list in the table is the best predictor for death count by HIV/AIDS looking mainly at the delta AIC score - tells you how much worse than the last predictor *look up AIC test - takes a look at the relationship between multiple categories and coun

##ˆ gender is the best predictor for count

#Biological Summary / Challenges